PROTEINS: Structure, Function, and Bioinformatics 54:671–680 (2004)

Recovering the True Targets of Specific Ligands by Virtual Screening of the Protein Data Bank

Nicodéme Paul, Esther Kellenberger, Guillaume Bret, Pascal Müller, and Didier Rognan*
Bioinformatics Group, Laboratoire de Pharmacochimie de la Communication Cellulaire, CNRS UMR 7081, Illkirch, France

ABSTRACT **The Protein Data Bank (PDB) has been processed to extract a screening protein library (sc-PDB) of 2148 entries. A knowledge-based detection algorithm has been applied to 18,000 PDB files to find regular expressions corresponding to either protein, ions, co-factors, solvent, or ligand atoms. The sc-PDB database comprises high-resolution X-ray structures of proteins for which (i) a well-defined active site exists, (ii) the bound-ligand is a small molecular weight molecule. The database has been screened by an inverse docking tool derived from the GOLD program to recover the known target of four unrelated ligands. Both the database and the inverse screening procedures are accurate enough to rank the true target of the four investigated ligands among the top 1% scorers, with 70–100 fold enrichment with respect to random screening. Applying the proposed screening procedure to a small-sized generic ligand was much less accurate suggesting that inverse screening shall be reserved to rather selective compounds. Proteins 2004; 54:671–680.** © **2004 Wiley-Liss, Inc.**

Key words: docking; PDB; protein library; virtual screening

INTRODUCTION

High-throughput docking of large chemical libraries[1] has recently established as a promising tool for identifying new hits from protein three-dimensional (3D) structures coming mostly from X-ray diffraction data[2] but also from homology modeling.[3] Discovering which ligands, out of a large library, are likely to bind to a protein of interest is slowly turning into routine computational chemistry.[4] Surprisingly, the opposite question is still an issue. Given a known ligand, is it possible to recover its most likely target(s)? Answering this question using the above-mentioned docking approach implies first the development of a collection of protein active sites, and second the use of an inverse docking tool able to dock a single ligand to multiple macromolecules. As a database of choice to develop the inverse screening procedure, we have chosen the Protein Data Bank (PDB)[5] as it is the major 3D protein database for which experimentally determined protein coordinates are available. Several protein–ligand databases derived from the PDB have already been recently described.[6–11] Relibase[6] easily allows retrieval of protein–ligand complexes from an user-defined query focusing on specific molecular interactions. The LPDB[7] stores 195 high-resolution protein–ligand complexes and related physicochemical descriptors as well as binding constants. Its main purpose, as well as other related protein–ligand datasets[8,9] is to provide reliable 3D information for calibrating docking algorithms and scoring functions. The ProLINT database[10] contains about 20,000 interaction data for two protein families (kinases, proteases) with attached information about the ligand, the protein, experimental binding constants, and published literature. It has been used to derive structure-activity relationships and predict binding constants. LigBase[11] is a database of ligand binding sites aligned with related protein structures and sequences containing 50,000 binding sites for heterogeneous ligands (ions, solvent, cofactors, inhibitors, etc. . .).

None of the above-mentioned databases are directly usable to generate a collection of protein active sites customized to accommodate small-molecular-weight "drug-like" ligands. The only report of ligand–protein inverse docking[12] is derived from the well-known DOCK program[13] to generate a collection of 2,700 protein cavities for searching potential targets of two known drugs. Although very promising, the reported INVDOCK procedure[12] presents three major disadvantages: (i) it uses a shape-based docking algorithm and a force-field scoring function whose accuracy lies significantly below recently-described docking methods;[14] (ii) the cpu time required for searching all entries (8 to 20 days/ligand, depending on the hardware architecture) prohibits its general use; (iii) the active sites, defined from overlapping spheres filling automatically-detected cavities, are not defined as molecular entities.

Here we present a similar method, which uses one of the most precise docking engines,[15] is fast enough (64h/ligand) to be used at a high throughput, and can be linked to a relational database that may be browsed from chemically well-characterized fields (protein, active-site, or ligand).

COMPUTATIONAL METHODS

Protein Library Set-Up

The Protein Data Bank[5] (December 2001 release) was analyzed by a series of "in-house" perl scripts (Fig. 1) for retaining the entries suitable for molecular docking of a small-molecular-weight ligand. Every entry for which any of three keywords ("complex," "inhibitor," "with") could be

*Correspondence to: Didier Rognan, Bioinformatics Group, Laboratoire de Pharmacochimie de la Communication Cellulaire, CNRS UMR 7081, 74 route du Rhin, B.P.24, F-67401 Illkirch, France. E-mail: didier.rognan@pharma.u-strasbg.fr

Received 22 May 2003; Accepted 24 August 2003

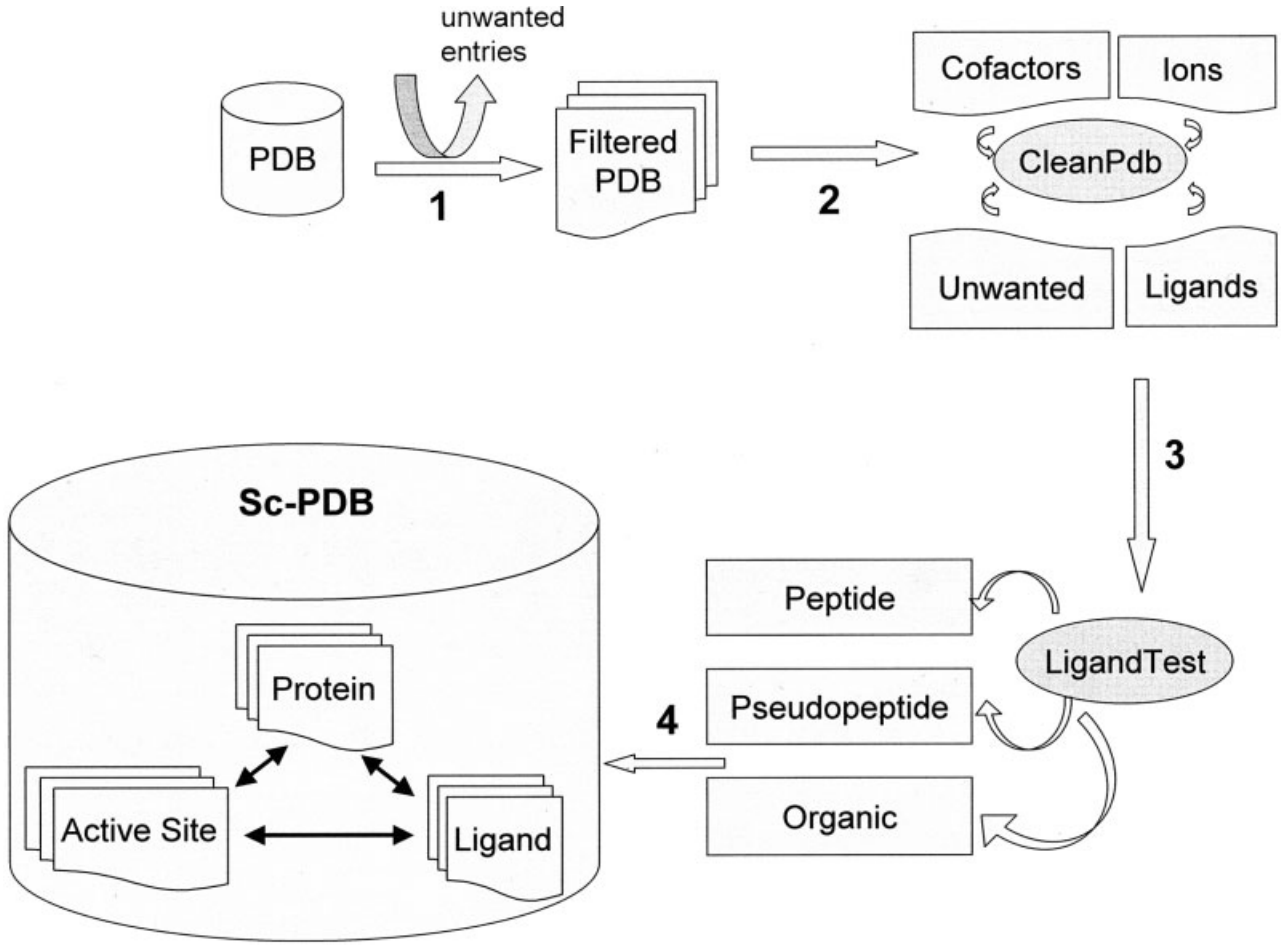

Fig. 1. Flow chart of the protein library set-up. (1) Filtering out undesirable entries (low resolution structures, apoenzymes, etc.) from the PDB, (2) defining building lists (cofactors, ions, ligands, unwanted molecules), (3) assigning the ligand chemotype, (4) merging the remaining entries into a relational database of 2148 entries (sc-PDB)

detected was first saved. In a second step, the remaining files were browsed for eliminating unwanted entries: (i) low resolution structures (< 3Å), superseded entries (e.g., differing by the real number preceding the three-letters code), entries with high-molecular-weight ligands (oligonucleotides, carbohydrates, peptides longer than ten amino acids), unwanted macromolecules (DNA, RNA, carbohydrates). The remaining PDB files were then checked to assign a molecule to each atom. A complete list of all PDB ligands[16] was used for assigning every "HET" card (non-protein atom) found in selected PDB files. Possible molecules, stored in separate lists include cofactors (e.g., NAD, NADP), ions to keep (Ca^{2+}, Fe^{2+}, Mg^{2+}, Mn^{2+}, Zn^{2+}), undesirable molecules (other monoatomic ions, metals, solvent). A ligand was defined as any molecule not belonging to the above-described chemotypes. In a fourth step, the molecular nature of the ligand (peptide, pseudopeptide, organic) was determined according to the corresponding "HET" cards. A ligand was determined as organic if it is referenced by a "HET" card not being defined in the "SEQRES" entry. If the "HET" card is also present in a SEQRES entry it is defined as a modified peptide (herein pseudopeptide chemotype). If no "HET" cards other than those characterizing non-ligand atoms (cofactor, ion to keep, unwanted molecule) could be found, the ligand was defined as a peptide (separate entry in the "SEQRES" definition) or simply missing (no SEQRES entry in addition to that of the protein). In the later case (apoenzyme), the PDB entry was eliminated from the target list. In a last step, all entries (total of 2148 files) for which the molecular nature of the ligand could be unambiguously assigned were merged into three databases. The first one comprises the atomic coordinates of the target protein, the second one the 3-D coordinates of the ligand. The last one compiles atomic coordinates of the active site. An active site is here defined as the set of residues that intersect a sphere S(O, r) where O is the center of mass of the ligand and $r = 12$ Å. The center of mass of one active site per database entry was stored in a single ASCII file. In cases where the ligand is present in multiple copies, the first copy of the ligand was selected to define the active site.

Development of a Reverse-Screening Tool

GOLD was chosen for setting up a reverse-screening tool for two main reasons: (i) it has been extensively tested over protein-ligand complexes from the PDB,[9] (ii) among sev-

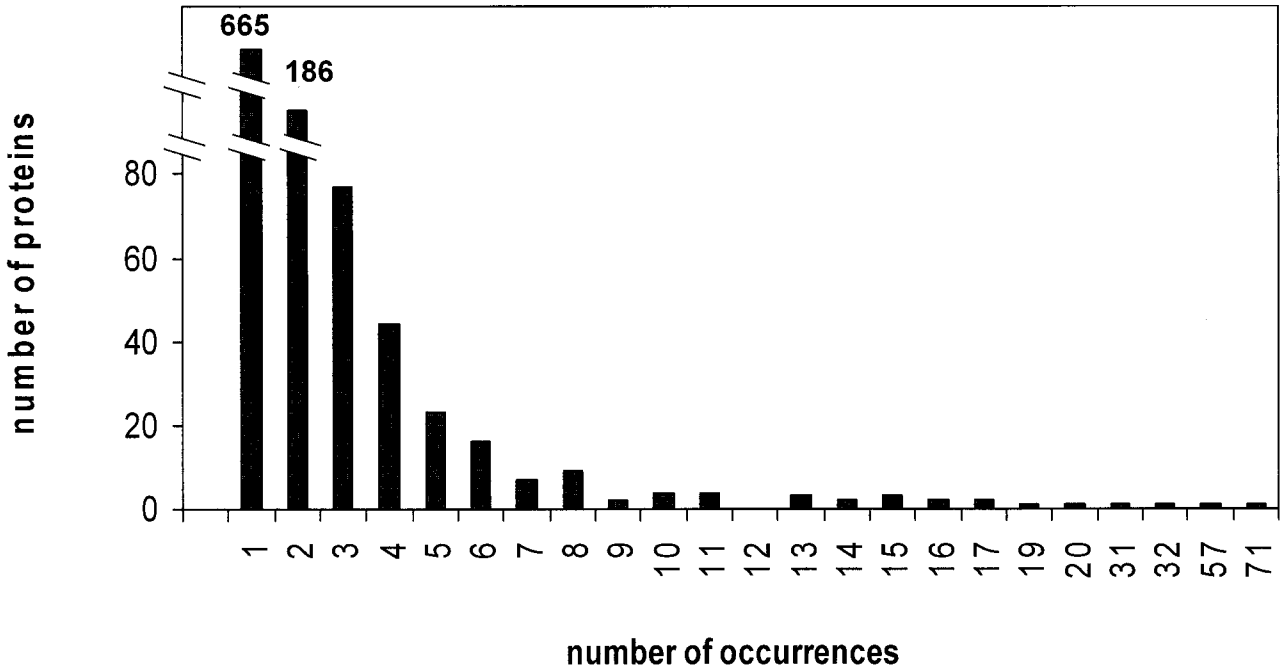

Fig. 2. Observed redundancy of the sc-PDB entries. Vertical bars represents the number of proteins (SwissProt/TrEMBL entries) having from 1 to a maximum of 71 copies. For sake of clarity, the scale of the Y axis is broken twice to allow compact display of extreme values (indicated between brackets).

eral docking tools, it reproduce the best known X-ray structures of 100 protein-bound ligands,[14] (ii) it requires a single input configuration file and is thus well suited for reverse screening. For each entry of the sc-PDB database, initial PDB coordinates of the protein were converted in TRIPOS mol2 format[17] while adding all hydrogen atoms, using the SYBYL6.8 package. In addition, a configuration file (gold.conf) was defined for each entry, including the absolute path of the corresponding protein mol2 file and the 3-D coordinates of the active site center. Fast virtual screening settings[15] were chosen for launching a separate job per entry. In the current study, 32 processors of a SGI Origin3800 supercomputer were used to screen the full database in ca. two hours (total of 64 cpu-hours).

Finding the Macromolecular Target of Test Ligands

Starting from an IsisDraw[18] 2-D sketch, 3-D coordinates of test ligands were obtained using Concord4.0,[17] minimized in SYBYL using the TRIPOS force field[19] and saved in mol2 format. For each reference ligand, the sc-PDB database was screened using the above-described GOLD procedure. Ten independent jobs were submitted and the average fitness score for each PDB entry saved in an ASCII file.

RESULTS AND DISCUSSION

Analysis of sc-PDB Entries

The current database contains 2148 ligand-binding sites for peptides or small molecules; 2021 are formed by one single protein, 116 are located at the interface of a protein binary complex and 11 consists of a ternary complex of proteins. In total, the database refers to 2286 proteins. We assigned a unique SwissProt or TrEMBL[20] accession number to each protein, thereby identifying 1045 different proteins in the database. Figure 2 gives an overview of the redundancy of current database entries. In most cases, less than ten copies of an active site corresponding to a given protein are available in the database. Sixty-three percent (63%) of the proteins are observed only once in the database. About 2% of the proteins are highly repeated (more than ten copies) and represent 20% of the database entries. The most frequent proteins are the HIV-1 protease (71 occurrences), the human thrombin (57 occurrences), the human carbonic anhydrase II (32 occurrences) and the bovine trypsinogen (31 occurrences). The uneven protein entries distribution, which reflects the intrinsic PDB redundancy, is of great interest for application like inverse virtual screening. Indeed, conformational differences between several copies of an active site reflect the local protein flexibility.

Additional information from both SwissProt/TrEMBL and PDB databanks was collected to obtain the source organism and the biological function of each protein. Resulting annotation of entries indicates that 291 organisms are represented in the database. Mammals, prokaryotes/archebacteria, and virus constitute 39%, 36%, and 8% of the species in the database, respectively. It is noteworthy that human proteins represent more than the half of mammalian entries.

A functional classification of the database entries is shown in Figure 3. Entries were separated into two superfamilies, namely enzymatic and non-enzymatic proteins. Out of the 2286 different proteins of the database 1741 proteins are enzymes. The set of enzymatic proteins was organized into six families, according to EC (Enzyme Commission) number.[21] The distribution of enzyme fami-

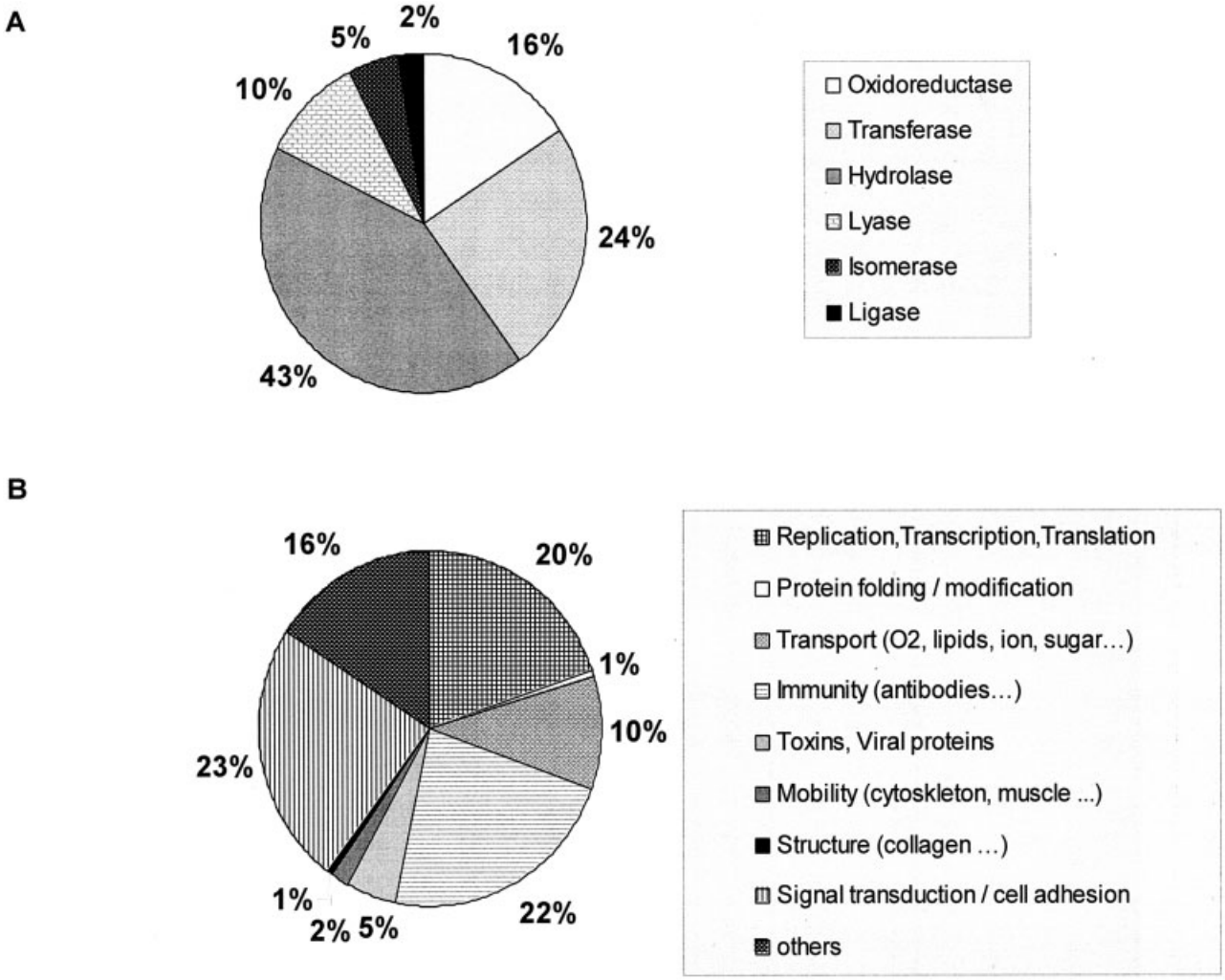

Fig. 3. Functional classification of the sc-PDB entries. **A**: Distribution of the 1741 enzymatic proteins. **B**: Distribution of the 545 non-enzymatic proteins.

lies displayed in Figure 3(A) reveals that the most populated family is that of hydrolases (43% of the enzymes). This is correlated to the high number of proteases in the sc-PDB database. The diverse functions of the 545 non-enzymatic proteins were clustered into families corresponding to the different cellular processes [Fig. 3(B)]. Four major classes (including replication/transcription/translation, molecular transport, immunity, and signal transduction/cell adhesion) make up 75% of the nonenzymatic proteins set.

In Silico Reverse Pharmacology: Virtual Screening of the sc-PDB

To validate both the sc-PDB database setup and the inverse docking protocol, we recovered the most likely target of four ligands (biotin, 4-hydroxy-tamoxifen, 6-hydroxyl-1,6-dihydropurine ribonucleoside, methotrexate; Fig. 4) known to specifically bind to a well-defined target (streptavidin,[22] estrogen receptor α,[23] adenosine deaminase,[24] and dihydrofolate reductase,[25] respectively). These four ligands were chosen for (i) their chemical diversity, (ii)

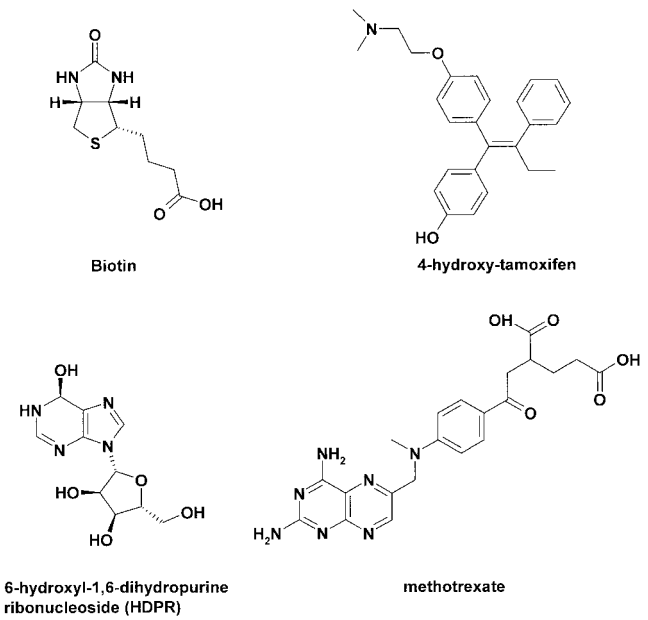

Fig. 4. Chemical structure of ligands taken for validating the inverse screening procedure.

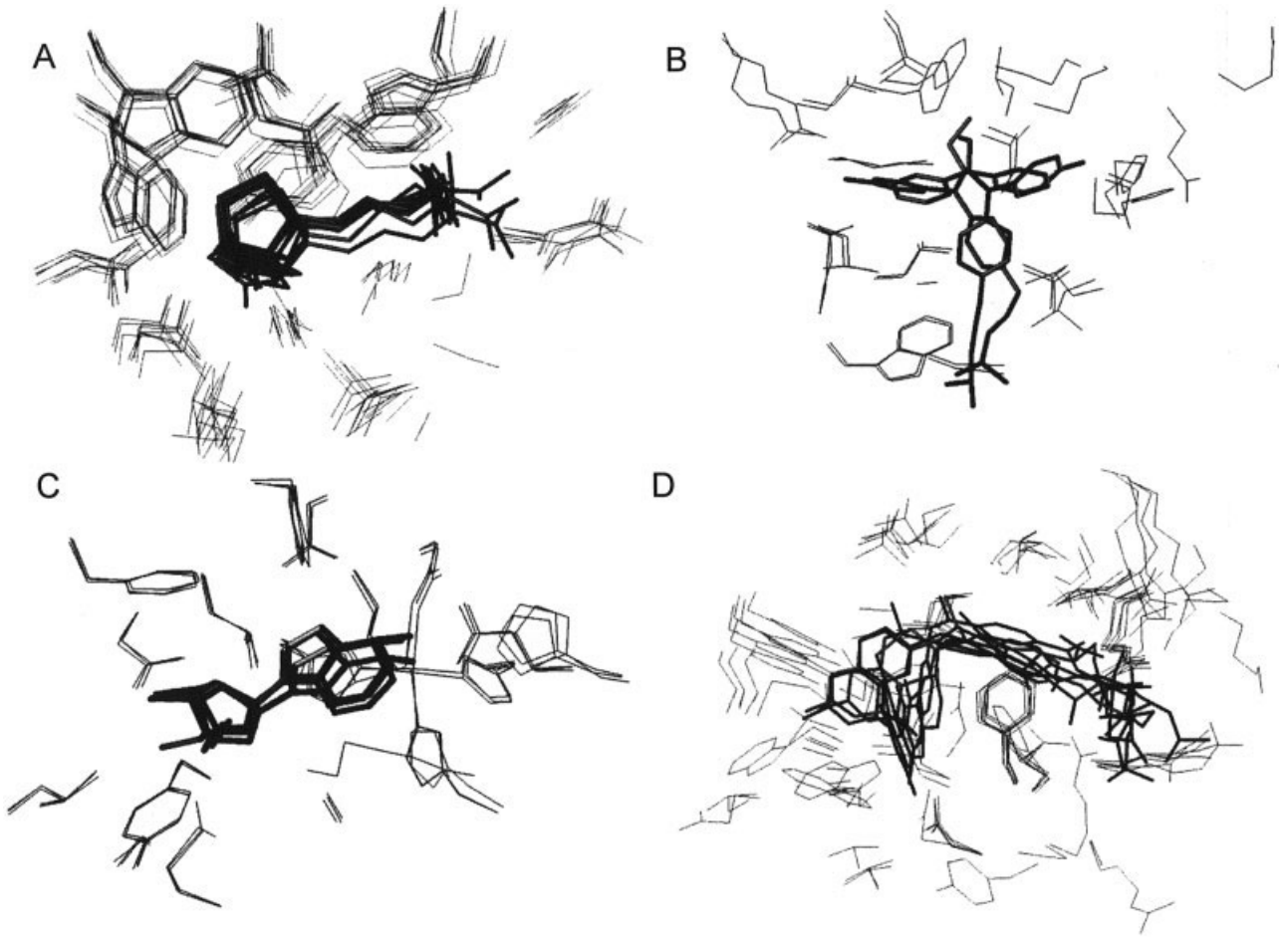

Fig. 5. Overlay of the GOLD poses (dark sticks) bound to all sc-PDB copies of the true target (thin sticks) for (**A**) biotin , (**B**) 4-hydroxy tamoxifen , (**C**) 6-hydroxyl-1,6-dihydroxypurine ribonucleoside , (**D**) methotrexate. For sake of clarity, only sc-PDB entries for which methotrexate is the native ligand are shown here.

their different rotational degree of freedom (from 2 to 9 rotatable bonds), (iii) the diverse number of sc-PDB entries corresponding to the true target (from 2 to 25).

Screening for biotin targets

We first looked at the docking accuracy of our inverse screening procedure by comparing predicted with X-ray poses of biotin bound to either wild type or mutant streptavidin entries present in the sc-PDB dataset. The computed binding mode of biotin to wild type strepatavidin is indeed very close to the X-ray pose (1stp) with RMS deviations on heavy atoms of 0.60 Å [Fig. 5(A)]. For protein mutations that do not affect the binding mode of biotin (1df8, 1swg, 1swr, 1swt), low RMSD are also observed (Table I). Thus, both the inverse docking protocol and the automatically-generated input files seem accurate enough to reproduce X-ray structures. We next looked at the propensity of the scoring function to discriminate true target entries from other sc-PDB entries. When screening our database for potential targets of biotin, 7 out of the 10 streptavidin entries present in the sc-PDB are ranked at the top eight positions with very good averaged fitness scores [Table I, Fig. 6(A)]. Interestingly, the three streptavidin copies with lower rankings (90th, 195th, 315th) correspond to either an active site for which a key amino acid (Asp128) has been mutated (1swt) or alternative binding sites (peptide binding sites for 1vwr and 1rsu). Altogether,

TABLE I. Rank and Average Fitness Score of Streptavidin Entries (Biotin Screening)

PDB entry	Native ligand	Rank	Fitness	RMSD, Å[a]
2izl	2-Iminobiotin	1	55.36	—
1i9h	Biotinyl p.nitroaniline	2	55.33	—
2rtr	2-Iminobiotin	3	55.11	—
1df8	Biotin	4	54.83	0.50
1stp	Biotin	5	54.01	0.60
1swg	Biotin	6	52.54	2.20
1swr	Biotin	8	51.24	0.42
1swt	Biotin	90	38.86	0.73
1vwr	Peptide ligand[b]	135	35.80	—
1rsu	SREP tagII peptide[c]	315	34.08	—

[a]RMSD of the GOLD pose from the X-ray coordinates (heavy atoms only).
[b]Cyclo-[5-S-valeramide-HPQGPPC]K-NH2.
[c]SNWSHPQFEK.

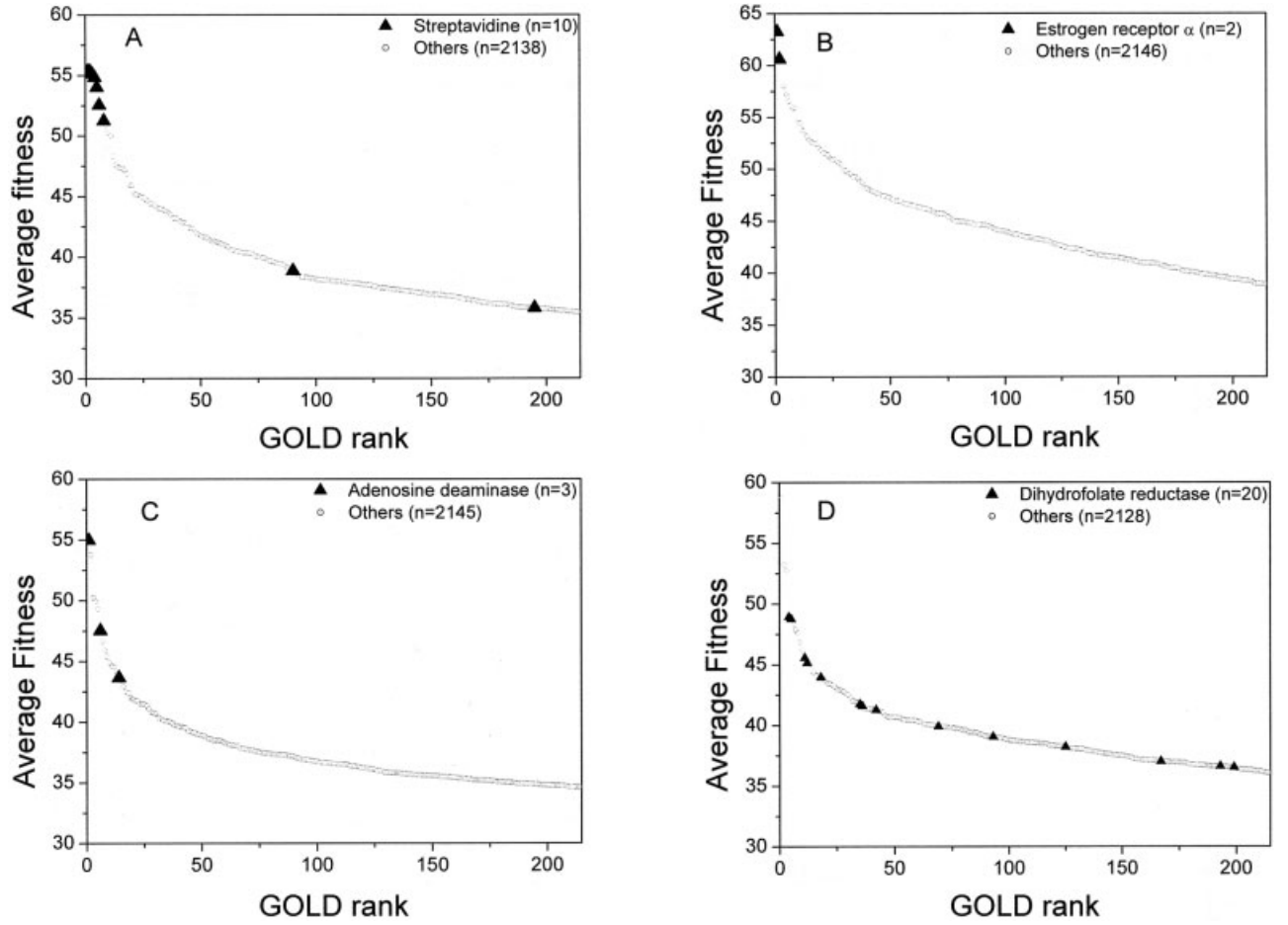

Fig. 6. Inverse screening of the sc-PDB database for finding the target of four small molecular weight ligands: (**A**) biotin, (**B**) 4-hydroxy tamoxifen, (**C**) 6-hydroxyl-1,6-dihydropurine ribonucleoside, (**D**) methotrexate. Filled triangles indicate the scores obtained by the different sc-PDB entries of the true target (A: streptavidin, B: estrogen receptor α, C: adenosine deaminase, D: dihydrofolate reductase).

the proposed inverse screening protocol is able to unambiguously rank streptavidin as the most likely target for biotin with a percentage of coverage of 70% (7 out of 10) among the top 10 (5%) positions. Several copies of two other enzymes (tyrosine phosphatase 1B, Influenza virus neuraminidase) are also ranked among the top 25 positions. However, no experimental data yet support the hypothesis that biotin may bind to the latter two proteins.

TABLE II. Rank and Average Fitness Score of Estrogen Receptor α Entries (4-Hydroxy Tamoxifen Screening)

PDB entry	Native ligand	Rank	Fitness	RMSD, Å[a]
1qkt	Estradiol	1	63.26	—
3ert	4-Hydroxy tamoxifen	2	60.61	1.36

[a]RMSD of the GOLD pose from the X-ray coordinates (heavy atoms only).

Screening for 4-hydroxy tamoxifen targets

Only two copies of the estrogen receptor α (pdb entries 1qkt, 3ert) are present in the current sc-PDB database. As previously described for biotin, 4-hydroxytamoxifen has been accurately docked in the ligand-binding domain of the ERα receptor [Table II, Fig. 5(B)] with RMSD values from the X-ray pose (3ert) of 1.36 Å. GOLD remarkably ranks both entries at the top two positions [Table II, Fig. 6(B)] with very high average fitness scores (above 60). Interestingly, three other targets are at least ranked twice among the top 25 scorers. Although there are presently no data suggesting that the first one (human coagulation factor Xa) is a real target of 4-hydroxy tamoxifen, several reports suggest direct binding of antiestrogens to the other two putative targets (NADP[H] quinone oxidoreductase, 3α-hydroxysteroid dehydrogenase). Hence, tamoxifen is reported to bind to the NADP[H] quinone oxidoreductase, a carninogen metabolizing enzyme, with a IC_{50} of 5.9 μM.[26] Antiestrogens block the enzymatic activity of 3α-hydroxysteroid dehydrogenase (HSD)[27] and have been reported to bind to 3-β HSD and 17β-HSD,[28–29] two enzymes closely related to 3α-HSD. Therefore, inverse screening of target databases could also be viewed as a computational filter to roughly predict the selectivity profile of a given ligand and thus putative site effects.

Screening for HDPR targets

HDPR (6-hydroxyl-1,6-dihydropurine ribonucleoside) is a known inhibitor of adenosine deaminase. GOLD accu-

rately docks HDPR to sc-PDB entries of the latter enzyme [Table III, Fig. 5(C)] with low RMSD (< 0.7 Å) to the X-ray pose, and places the three copies of adenosine deaminase stored in the sc-PDB database at ranks 1, 6, and 14, respectively [Table III, Fig. 6(B)]. The current inverse screening thus leads to a percentage of coverage of adenosine deaminase of 100% in the top 0.65% scorers. Two other enzymes (thymidine kinase, orotidine 5′-phosphate decarboxylase) also implicated in nucleoside metabolism and recognizing ligands chemically similar to HDPR are found at the top ranked positions. A putative binding of HDPR to these two enzymes yet remains to be experimentally demonstrated.

Screening for methotrexate targets

Methotrexate was chosen as the last ligand to be screened against the sc-PDB database. It is supposed to be a very hard test as methotrexate is a flexible molecule (nine rotatable bonds) that is partially buried in the binding site of dihydrofolate reductase (DHFR) in which additional molecules (water, cofactor) mediate flexible protein-ligand intermolecular interactions. Nevertheless, methotrexate is accurately docked in the catalytic site of DHFR [Table III, Fig. 5(D)] with low RMSD deviations to known positions of methotrexate in complex with DHFR (pdb entries 1dls, 1rb3, 1rg7, 1df7). Out of the 25 copies of DHFR present in the datatabase, six are placed in the top 1% scorers and 11 in the top 5% scorers [Fig. 6(D)]. The low ranking (above position 100) of other DHFR entries can be easily explained by (i) conformation changes of the Met-20 loop frequently occurring in the neighborhood of the bound cofactor[30] (pdb entries 1jom, 1hfr, 1boz, 1dg8, 1ia4, 1rh3, 2cd2), (ii) the absence of accessory molecules (e.g. glycerol) in the screened protein coordinates (1d8r, 1df7), (iii) a closed binding cavity due to the absence of co-crystallized inhibitor or substrate analog (1dg8, 1drh, 1ra9, 1rx9).

Among the top-ranked scorers, adenosine deaminase (1add, rank 3) and acetyl cholinesterase (1eve, rank 7; 1dx6, rank 8) are confirmed targets of methotrexate.[31–32] The present results demonstrate the capability of our screening setup to predict known multiple targets of a given ligand.

Applying the herein described screening protocol to the current sc-PDB database clearly allowed to unambiguously recover the true target of four unrelated ligands. When compared to random screening, a significant enrichment in the true target is observed among the top scorers (Fig. 7; Table V). Analyzing both the enrichment factor and

TABLE III. Rank and Average Fitness Score of Adenosine Deaminase Entries (HDPR[a] Screening)

PDB entry	Native ligand	Rank	Fitness	RMSD, Å[b]
2ada	HDPR	1	54.96	0.68
1add	1-Deaza adenosine	6	47.52	—
1a4m	HDPR	14	43.65	0.54

[a]6-Hydroxyl-1,6-dihydroxypurine ribonucleoside.
[b]RMSD of the GOLD pose from the X-ray coordinates (heavy atoms only).

TABLE IV. Rank and Average Fitness Score of Dihydrofolate Reductase Entries (Methotrexate Screening)

PDB entry	Native ligand	Rank	Fitness	RMSD, Å[a]
1rf7	Dihydrofolate	4	48.92	—
1aoe	GW345 inhibitor	5	48.77	—
1drf	Folate	11	45.56	—
1dls	Methotrexate	12	45.14	1.20
1dds	Methotrexate	18	43.95	3.13
4cd2	Folate	19	43.89	—
1rc4	5,10-Dideazatetrahydrofolate	35	41.74	—
2dhf	5-Deazafolate	36	41.60	—
1rb3	Methotrexate	42	41.27	1.34
1rg7	Methotrexate	69	39.91	1.20
1daj	Furo[2,3D]pyridimidine antifolate	93	39.05	—
1d8r	TAB inhibitor	125	38.22	—
1df7	Methotrexate	167	37.02	1.75
1jom	Folinic acid	193	36.62	—
1dyj	5,10-Dideazatetrahydrofolate	199	36.53	—
1hfr	Furo[2,3D]pyridimidine antifolate	253	35.30	—
1boz	Quinazoline inhibitor	333	33.91	—
1dg8	—[b]	346	33.64	—
1ia4	TQ6 inhibitor	420	32.78	—
1rh3	Methotrexate	423	32.74	2.34
1drh	—[b]	458	32.34	
1ra9	—[b]	783	28.9	
1rx9	—[b]	1127	25.66	
2cd2	folate	1277	25.60	
1dr7	covalent ligand	1912	9.06	

[a]RMSD of the GOLD pose from the X-ray coordinates (heavy atoms only).
[b]Cofactor only.

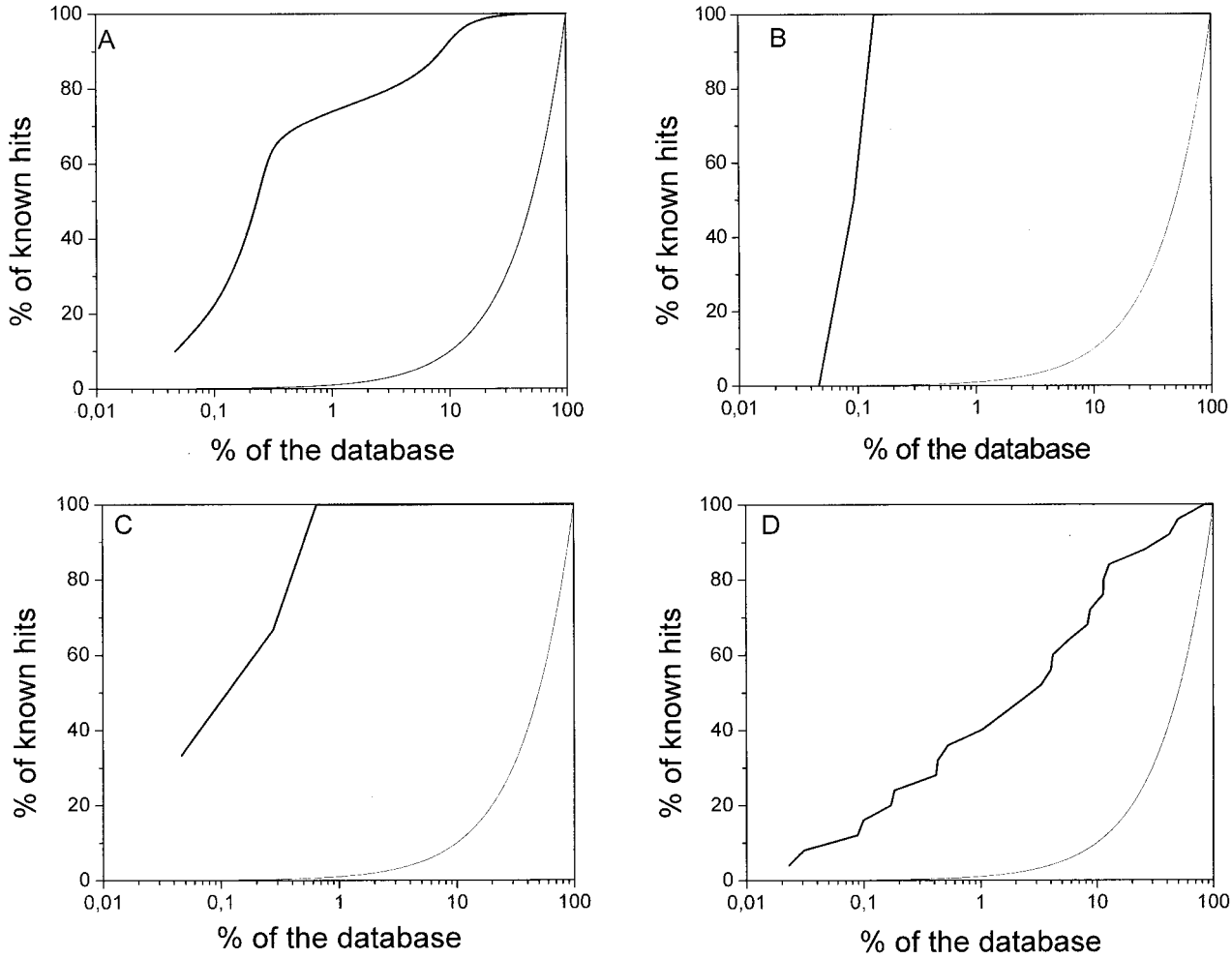

Fig. 7. Percentage of recovery of known targets as a function of the top scoring fraction extracted for analysis by inverse screening (bold line) and random picking (thin line). The percentage of coverage of known targets is the ratio in percentage between the number of true target entries recovered by inverse screening at a defined top scoring fraction and the total number of true target entries in the sc-PDB dataset.

TABLE V. Enrichment Factor in the True Target (Virtual Versus Random Screening)[†]

Top scoring fraction, %	Biotine	4-OH tamoxifen	HDPR	Methothrexate
0.5	150	215	143	21
1	71	102	102	26
2	35	50	50	20
5	16	20	20	10
10	9	10	10	7

[†]The enrichment factor (EF) is calculated as the ratio between hit rates obtained by virtual screening and that obtained by random screening.

the percentage of coverage of known targets indicates that the best compromise can be reached by selecting a very small fraction (0.5%) of the sc-PDB database. Even for the rather difficult case of methotrexate, selecting the top 2.6% scorers would allow to select 40% of all dihydrofolate reductase entries with a 15-fold enrichment with respect to random screening. From the present data, we recommended to select the top 1% or 2 % scorers to be sure to significantly enrich the hit list in the true target. To determine the accuracy of the proposed inverse screening procedure, we next tried to recover the true targets of a small-sized generic ligand (adenosine monophosphate or AMP). Out of the 15 sc-PDB entries co-crystalized with AMP, six are ranked among the top 10% scorers and three among the top 5% scorers (data not shown). The lower accuracy of our screening protocol in this peculiar case cannot be explained by poor docking of AMP because RMS deviations of ligand-heavy atoms to known X-ray poses was in most cases below 2 Å. Four entries were poorly ranked because of the absence of an accessory molecule (inorganic phosphate, ANP, phenylalanine, TIA) in the protein coordinates set-up. In most cases, AMP was not very well scored (average fitness lower than 45) in the binding site of known targets.

The present result suggests that inverse screening should be reserved to rather selective ligands and that deciphering the numerous targets of very promiscuous compounds remains an issue.

With regard to the previously-described InvDOCK program,[13] the present method presents multiple adavan-

tages: (i) it is fast enough to be routinely applied as 64 cpu-hours only are necessary to browse the entire database, (ii) the functional classification of the sc-PDB database allows both 3-D searching and data analysis on focussed target libraries (e.g., protein kinases), (iii) the accurate definition of binding sites allows to search cavities for "drug-like" molecules only.

FUTURE IMPROVEMENTS

Several directions can be followed to improve the current sc-PDB database. First, some entries (e.g., toxins, antibodies) could be removed in order to enhance the therapeutical potential of stored entries. Second, applying a cavity detection algorithm[33–34] on PDB entries describing ligand-free enzymes will supply the sc-PDB database not only with new entries but also new proteins and thus increase its molecular diversity. The use of a minimal volume threshold need however to be used in order to avoid saving small cavities. Alternatively, the biggest predicted cavity may simply be stored. Last, the most significant improvement needs to be the automated update of the sc-PDB database. About 250–300 new entries are deposited every month in the PDB and this perpetual upgrade must be transferred to the sc-PDB. A clear limitation to the automated upgrade of the sc-PDB is the knowledge-based use of building lists (ions, cofactors, solvent, etc. . .) based on very heterogeneous PDB heteroatoms cards. As far as already-stored "HET" cards are detected, the nature of the corresponding ligand will be unambiguously assigned. Of course, this implies a conserved usage of HET acronyms for describing the same molecule. In cases where new HET cards are found, browsing the PDBsum database[16] allows to name the corresponding molecule and assign it to one of our building lists.

CONCLUSIONS

A 3D protein library comprising 2148 entries has been set-up from the Protein Data Bank for reverse docking purpose. To validate both the protein library set-up and the reverse docking method, we tried to recover "in silico" the most likely receptor of four unrelated ligands. In all cases, the top 1% scorers were significantly enriched in the true target with enrichment factors of 70–100 over random screening. The current protein library can be fully screened against a single ligand within 64 cpu-hrs, and is suitable for "in silico" target prediction and "virtual selectivity" profiling of any ligand of interest. Furthermore, as several copies of the same protein are saved in our dataset, the current reverse screening protocol is an alternative to existing flexible protein/flexible ligand docking approaches. It can be applied to any protein library (e.g., homology 3D models, molecular dynamics snaphots of a single target) and thus extend the inverse screening to proteins for which X-ray structures are still missing.

ACKNOWLEDGMENTS

This work is supported by grants from the Fondation pour la Recherche Médicale (Paris, France) and the French Ministry of Research and Technology. The Centre Informatique National de l'Enseignement Supérieur (CINES, Montpellier, France) is acknowledged for allocation of computing time on the Origin3800 supercomputer. The sc-PDB database (proteins, active sites, ligands) is available for non-commercial academic research use upon request to the authors.

REFERENCES

1. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins 2002;47:409–443.
2. Abagyan R, Totrov M. High-throughput docking for lead generation Curr Opin Chem Biol 2001;5:375–382.
3. Bissantz C, Bernard P, Hibert M, Rognan D. Protein-based virtual screening of chemical databases. II. Are homology models of G-protein coupled receptors suitable targets? Proteins 2003;50:5–25.
4. Schneider G, Böhm HJ. Virtual screening and fast automated docking methods Drug Discov Today 2002;7:64–70.
5. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
6. 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.
7. Roche O, Kiyama R, Brooks III CL. Ligand-Protein DataBase: Linking protein-ligand complex structures to binding data. J Med Chem 2001;44:3592–3598.
8. Kramer B, Rarey M, Lengauer T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. Proteins 1999;37:228–241.
9. Nissink JWM, Murray C, Hartshorn M, Verdonk ML, Cole JS., Taylor R. A new test set for validating predictions of protein-ligand interaction. Proteins 2002;49:457–471.
10. Kitajima K, Ahmad S, Selvaraj S, Kubodera H, Sunada S, An J, Sarai A. Development of a protein-ligand interaction database, ProLINT, and its application to QSAR analysis. Genome Informatics 2002;13:498–499.
11. Stuart AC., Ilyin VA., Sali A. LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. Bioinformatics 2002;18:200–201.
12. Kuntz ID, Blaney JM, Oaley SJ, Langridge R, Ferrin, T.E. A geometric apporach to maromolecule-ligand recognitions. J Mol Biol 1982;161:269–288.
13. Chen YZ, Zhi DG. Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. Proteins 2001;43:217–226.
14. Paul N, Rognan D. ConsDock: A new program for the consensus analysis of protein-ligand interactions. Proteins 2002;47:521–533.
15. Jones G, Wilett P, Glen RC, Leach AR, Taylor R Development and validation of a genetic algorithm for flexible docking. J Mol Biol 1997;267:727–748.
16. Laskowski RA. PDBsum: summaries and analyses of PDB structures. Nucleic Acids Res 2001;221–222.
17. Tripos Assoc., Inc., St. Louis, MO 63144.
18. MDL Information Systems, Inc., San Leandro, CA 94577.
19. Clark M, Cramer III RD, Van Opdenbosch N. Validation of the general purpose TRIPOS 5.2 force field. J Comp Chem 1989;10: 982–1012.
20. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, 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.
21. Bairoch A. The ENZYME database in 2000. Nucleic Acids Res 2000;28:304–305.
22. Weber PC, Ohlendorf DH, Wendoloski JJ, Salemme FR. Structural origins of high-affinity biotin binding to streptavidin. Science 1989;243:85–89.
23. Shiau AK, Barstad DB, Loria PM, Cheng L, Kushner PJ., Agard DA, Greene GL. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. Cell 1998;95:927–937.
24. Wilson DK, Rudolph FB, Quiocho, FA. Atomic structure of adenosine deaminase complexed with a transition-state analog: understanding catalysis and immunodeficiency mutations. Science 1991; 252:1278–1284.

25. Filman DJ., Matthews DA., Bolin JT, Kraut J. Crystal structures of Escherichia coli and Lactobacillus casei dihydrofolate reductase refined at 1.7 Å resolution. I. General features and binding of methotrexate. J Biol Chem 1982;257:13650–13662.
26. Gerhauser C, Klimo K, Heiss E, Neumann I, Gamal-Eldeen A, Knauft J, Liu GY, Sitthimonchai S., Frank, N. Mechanism-based in vitro screening of potential cancer chemopreventive agents. Mutat Res 2003;523–524:163–172.
27. Lax ER, Rumstadt F, Plascyck H, Peetz A, Schriefers, H. Antagonistic action of estrogens, flutamide, and human growth hormone on androgen-induced changes in the activities of some enzymes of hepatic steroid metabolism in the rat. Endocrinol 1983;113:1043–1055.
28. Le Lain R, Barrell KJ, Saeed GS, Nicholls PJ, Simons C, Kirby A, Smith H. Some coumarins and triphenylethene derivatives as inhibitors of human testes microsomal 17beta-hydroxysteroid dehydrogenase (17beta-HSD type 3): further studies with tamoxifen on the rat testes microsomal enzyme. J Enzyme Inhib Med Chem 2002;17:93–100.
29. Le Bail JC, Champavier Y, Chulia AJ, Habrioux G. Effects of phytoestrogens on aromatase, 3beta and 17beta-hydroxysteroid dehydrogenase activities and human breast cancer cells. Life Sci 2002;25:1281–1291.
30. Sawaya, M.R. and Kraut, J. Loop and subdomain movements in the mechanism of action of Escherichia coli dihydrofolate reductase: crystallographic evidence. Biochemistry 36 (1997), 586–603.
31. Al-Jafari A, Al-Khwyter F, Kamal MA., Alhomida, S.A. Kinetics for camel (Camelus dromedarius) retina acetylcholinesterase inhibition by methotrexate in vivo. Jpn J Pharmacol 1996;72:49–55.
32. Cronstein BN, Naime D, Ostad, E. The antiinflammatory mechanism of methotrexate. Increased adenosine release at inflamed sites diminishes leukocyte accumulation in an in vivo model of inflammation. J Clin Invest 1993;92:2675–2682.
33. Hendlich M, Rippman F, Barnickel G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph 1997;15:359–363.
34. Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 1995;13:323–330.