

Ranking Targets in Structure-Based Virtual Screening of Three-Dimensional Protein Libraries: Methods and Problems

Esther Kellenberger,* Nicolas Foata, and Didier Rognan

Bioinformatics of the Drug, UMR 7175 CNRS-ULP (Université Louis Pasteur-Strasbourg I),
F-67400 Illkirch, France

Received January 18, 2008

Structure-based virtual screening is a promising tool to identify putative targets for a specific ligand. Instead of docking multiple ligands into a single protein cavity, a single ligand is docked in a collection of binding sites. In inverse screening, hits are in fact targets which have been prioritized within the pool of best ranked proteins. The target rate depends on specificity and promiscuity in protein–ligand interactions and, to a considerable extent, on the effectiveness of the scoring function, which still is the Achilles' heel of molecular docking. In the present retrospective study, virtual screening of the sc-PDB target library by GOLD docking was carried out for four compounds (biotin, 4-hydroxy-tamoxifen, 6-hydroxy-1,6-dihydropurine ribonucleoside, and methotrexate) of known sc-PDB targets and, several ranking protocols based on GOLD fitness score and topological molecular interaction fingerprint (IFP) comparison were evaluated. For the four investigated ligands, the fusion of GOLD fitness and two IFP scores allowed the recovery of most targets, including the *rare* proteins which are not readily suitable for statistical analysis, while significantly filtering out most false positive entries. The current survey suggests that selecting a small number of targets (<20) for experimental evaluation is achievable with a pure structure-based approach.

INTRODUCTION

The pace at which novel chemogenomic approaches¹ are currently developed is likely to strongly influence the design of bioactive molecules in the next decade. Notably, it is now commonplace to virtually and/or physically profile a ligand toward an heterogeneous set of macromolecular targets.² Exhaustive coverage of the chemogenomic space for either a ligand or a gene family cannot be addressed at the experimental level such as new virtual screening techniques which are needed to fill chemogenomic matrices. Up to now, most approaches to ligand profiling rely on ligand-based in silico screening methods by querying biologically characterized ligands with either 2D fingerprints^{3–7} or 3D pharmacophores.⁸ Structure-based approaches in which a single compound is docked into several protein structures have received less attention for the simple reason that biologically annotated compound libraries are easier to set up than collections of protein binding sites. However, both retrospective^{9–11} and prospective structure-based target identification^{12,13} have been reported recently, illustrating the fact that automating binding site setup has become a mature field.^{14,15} A clear drawback of structure-based approaches to target profiling is the acknowledged low accuracy of scoring functions¹⁶ used to prioritize targets from a list of protein–ligand complexes, notably for highly polar binding sites (e.g., metalloenzymes)¹³ and the problematic statistical treatment of proteins present in low copy number in the target database.¹²

The present study aims at overcoming the above-cited hurdles by applying a robust and simple postdocking process

combining two different scoring strategies. On the one hand, a classical energy-based scoring function is used to filter out the very unlikely targets. On a second hand, a topological scoring function based on protein–ligand interaction fingerprints¹⁷ is selected for retrieving ligand poses resembling protein–ligand interactions observed in crystal structures. Interaction fingerprints (IFPs) encode ligand–protein interactions (principally hydrophobic and ionic interactions, aromatic stacking, and H-bonds) into bit strings. The similarity of the binding mode of two ligands to a common ligand-binding site is thus measured as a distance between two bit strings. Scoring docking poses by similarity of interaction fingerprints to a given reference was shown to be statistically superior to conventional scoring functions in ranking drug-like compounds^{18,19} and low-molecular-weight fragments.¹⁷ We here apply the same concept to inverse screening by systematically comparing docking poses of a unique ligand into PDB active sites to the binding mode of the corresponding cocrystallized ligands. Various scoring schemes were applied to rate the ability of 4300 ligand-binding sites to accommodate four different ligands (biotin, methotrexate, 4-hydroxy-tamoxifen, and 6-hydroxy-1,6 dihydropurine ribonucleoside). Combining docking and IFP scores is the best choice for any of the four ligands in order to retrieve their true targets within a reasonably small number (<20) of candidates.

COMPUTATIONAL METHODS

sc-PDB Target Library Setup. The 2006 release of the sc-PDB target library¹⁴ was screened throughout this study. It is a collection of 4300 druggable ligand-binding sites found in the PDB in which an active site is automatically detected from the coordinates of the bound pharmacological ligand.

* Corresponding author. Mailing address: Bioinformatics of the Drug, UMR 7175 CNRS-ULP (Université Louis Pasteur-Strasbourg I), 74 route du Rhin, B. P. 24, F-67400 Illkirch, France. Phone: +33-3-90244224. Fax: +33-3-90244310. E-mail: esther.kellenberger@pharma.u-strasbg.fr.

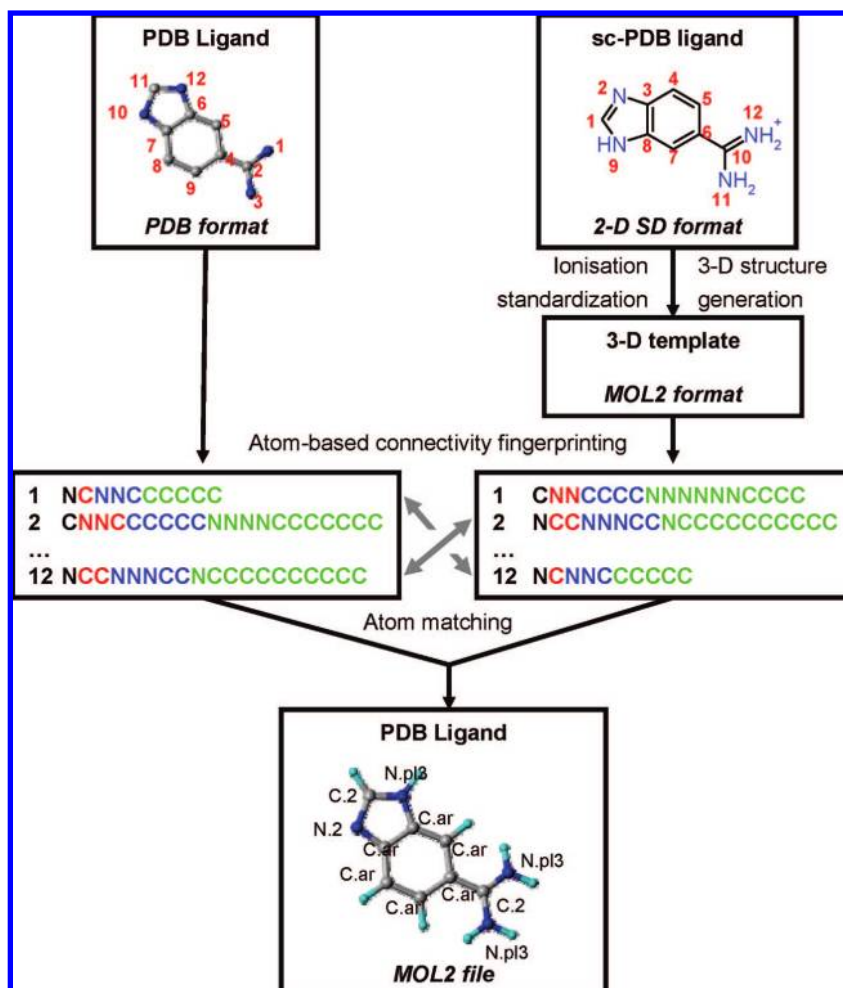


Figure 1. PDB to MOL2 file format conversion of sc-PDB ligands. A ligand MOL2 file is obtained by merging original PDB coordinates with atom and bond types of a MOL2 template file. Atom matching is performed by pairwise comparison of atom fingerprints. For each atom, a unique fingerprint is generated by concatenating atom elements (sorted by decreasing element number) of its nearest neighbors (from 6 to 12 bonds apart). In the figure, fingerprint generation is exemplified up to three iterations. The atom under consideration (top left column) is colored in black, the nearest neighbors (1st generation), in red, the neighbors of the 1st generation atoms (2nd generation), in blue, and the neighbors of 2nd generation atoms (3rd generation), in green. Up to 12 iterations are necessary to unambiguously match atoms of PDB- and MOL2-generated fingerprints.

Like its mother PDB database, the sc-PDB archive contains redundant information. Duplicate ligand-binding sites were cleared from the database upon pairwise comparison of entries using both ligand SMILES strings²⁰ and protein UNIPROT²¹ identifiers. To better focus on protein druggability, the ligand selection was narrowed with respect to the previous sc-PDB release by increasing the minimal molecular weight (non-hydrogen atoms only) from 70 to 150 and discarding small-sized ligand-buried cavities. The new sc-PDB data set consists in 4300 entries describing 1550 different proteins in variable copy number (from 1 to 87).

The structural characterization of sc-PDB ligands, prosthetic groups and cofactors was improved by defining for each molecule a correct MOL2 file format²² with native X-ray coordinates. For this purpose, we developed an in-house Python script (*AtomTyper*) converting HET groups from PDB²³ to MOL2 file formats (Figure 1). AtomTyper merges coordinates of the PDB file with atom and bond types from the corresponding MOL2 template. The MOL2 template was obtained for each HET group from the previous 2D structural data (SD) representation,¹⁴ by ionization with Filter v2.0.1,²⁴ structure standardization with JChem v3.2.3,²⁵ and 3D coordinate generation with Corina v3.4.²⁶ None of the

PDB and the MOL2 template input files contain hydrogen atoms. The AtomTyper algorithm reads the two input files using the OEChem1.4.2 library,²⁷ verifies that both molecules have a unique number of atoms, and then transforms each molecule into a series of atomic fingerprints. Practically, the fingerprint encodes for each atom its neighborhood using atom element information and successive iterative steps (Figure 1). The chemical element (C, N, O, P, S, Br, Cl, or F) of the atom determines the first member of the vector. Every iteration then uses the atom identifiers from the previous iteration and outputs the chemical elements of all covalently bound atoms, sorted by decreasing atomic number. Depending on the molecular size and complexity, up to 12 iterations are required to unambiguously define a unique fingerprint per atom taking into account possible symmetry information. Once fingerprints have been defined for both the PDB and the MOL2 files, all pairwise fingerprint comparisons are performed to match atoms of the two input files. Hydrogen atoms were finally added to the ligand MOL2 file assuming standard geometries as defined in SYBYL 7.3.²²

Generating a Molecular Interaction Fingerprint for All sc-PDB Protein–Ligand Complexes. Atomic coordinates of polar hydrogen atoms in both protein and ligand/

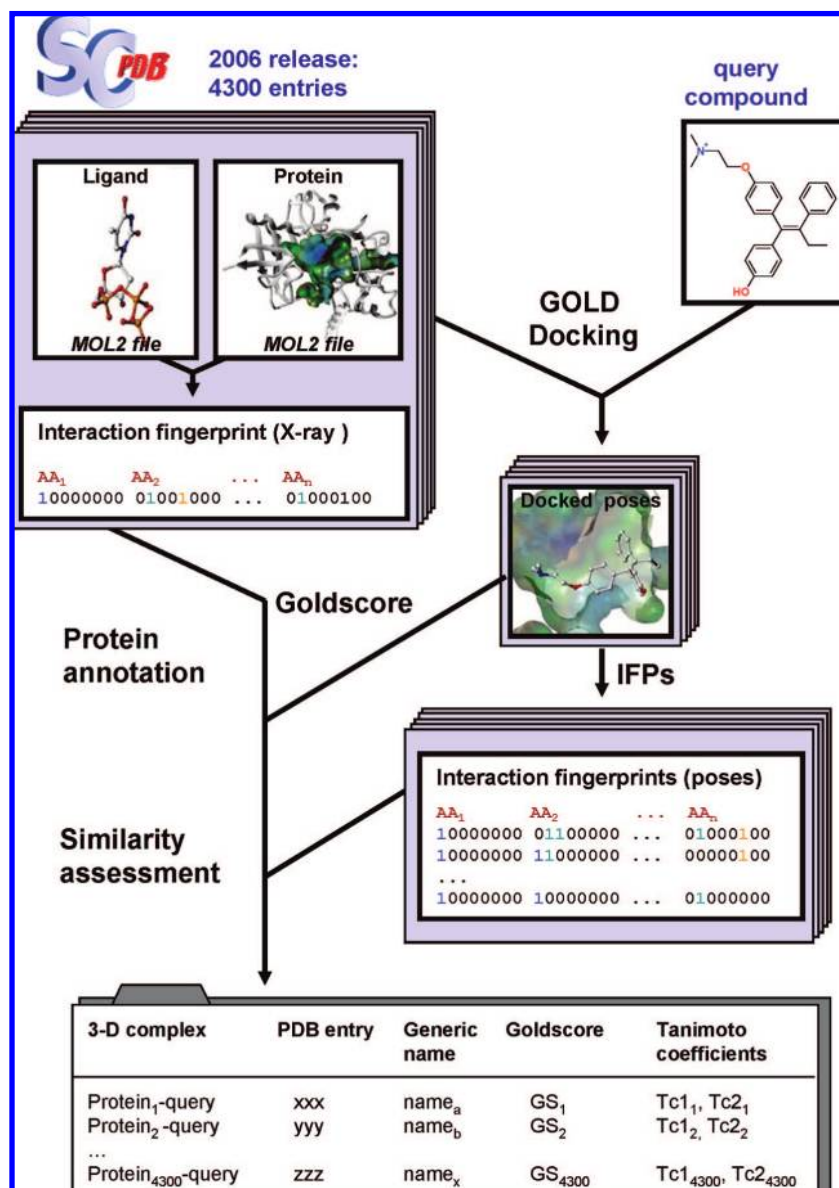


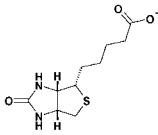
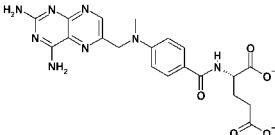
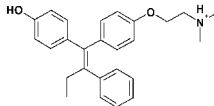
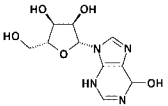
Figure 2. Flowchart for screening the sc-PDB target library. A ligand is docked in a sc-PDB binding site using the GOLD v3.1 program²⁸ and scored both according to the GOLD fitness function and similarity of interaction fingerprints (IFPs)¹⁷ to the corresponding sc-PDB entry (X-ray coordinates of the protein–ligand complex). All virtual complexes are stored in a table registering the protein active site, the corresponding protein name, average fitness, and IFP Tanimoto coefficients (Tc1 and Tc2 scores). Tc1 describes the similarity to the X-ray pose using a full bit string, whereas Tc2 registers similarity to a fingerprint truncated to polar interactions only.

cofactor molecules have been optimized by constrained docking using the 7–8 times speed-up setting of the GOLD 3.1 program.²⁸ To lock the heavy atom positions, a strong similarity constraint must be satisfied (shape overlap to the input MOL2 ligand coordinates, constraint weight of 100). For the full sc-PDB, the root-mean-square-deviation (computed on all atom positions) of the refined ligand from the original structure was always below 1 Å. Polar hydrogens of the protein which have been rotated to optimize H-bonding were saved and stored in the final set of protein coordinates.

For a refined sc-PDB entry, a molecular interaction fingerprint (IFP) was computed as previously reported¹⁷ to describe the *reference* binding mode. The interaction fingerprint comprises 8 bits per binding site residue encoding for hydrophobic contacts, aromatic interactions (face to face, edge to face), H-bonds (donor to acceptor, acceptor to donor), ionic interactions (positive to negative, negative to positive), and metal complexation.

sc-PDB Library Screening. The overall screening strategy is described in Figure 2. The sc-PDB target library was screened by high-throughput docking of a ligand of interest. Ligand 3D coordinates were generated from the corresponding sc-PDB SD files using Corina v3.4, the ionization state at pH 7 was manually checked and no formal charges specified in the MOL2 output file. For docking, the 2 times speed-up settings of the GOLD 3.1 genetic algorithm were used. Ten independent jobs were submitted for each ligand, and poses were scored with the Gold fitness score (*Goldscore*) and saved if the Goldscore was positive. For each sc-PDB entry, the docking output was processed to extract the average fitness score⁹ and to generate the *docked IFPs*, i.e. the interaction fingerprint between the protein and saved ligand poses. Again, coordinates of polar hydrogens of the target which have been rotated to score each pose were saved in temporary files. The similarity of the binding mode of docked poses to the reference pose (refined protein–ligand

Table 1. Properties of Four Ligands Used in the Present Study

Ligand	BTN	MTX	OHT	HDPR
Structure				
MW ^a	244	454	387	270
Nrot ^b	5	9	8	2
PSA ^c	78	211	33	132
XlogP ^d	0.5	-2.2	6.7	-2
HBA ^e	4	5	2	7
HBD ^f	2	3	2	5
Primary ^g targets	Streptavidin (13)	Dihydrofolate reductase (22)	Estrogen receptor α (17) Estrogen receptor β (12) Estrogen-related receptor γ (3)	Adenosine deaminase (16)
Secondary targets ^g	Avidin-related protein 2 (1) Avidin-related protein 4/5 (2) Avidin (2) BirA protein (2) Hypothetical biotin-ligase (3)	Pterin reductase (2) Thymidilate synthase (15)	Protein kinase C (1) Prostaglandin synthase I (5) Quinone oxidoreductase (2)	

^a Molecular weight. ^b Number of rotatable bonds. ^c Polar surface area. ^d Predicted octanol/water partition coefficient. ^e H-bond acceptor count. ^f H-bond donor count (all properties calculated in Pipeline Pilot²⁹). ^g The copy number for each target in the sc-PDB is indicated in brackets.

X-ray structure) was measured using the Tanimoto coefficient derived from corresponding interaction fingerprints. Two Tanimoto coefficients were computed by comparing either the full fingerprints (*Tc1* score, 8 bits/residue) or truncated fingerprints (*Tc2* score, 5 bits/residue coding for polar interactions only) of reference and docked IFPs. A final table was generated to summarize all results describing the sc-PDB entry, the protein name, pose number, average Goldscore, and *Tc1* and *Tc2* values.

Target Ranking. The Pipeline Pilot protocol²⁹ was designed to postprocess the docking output and iteratively compute sensitivity (Se) and specificity (Sp) for each protein name stored in the database.

$$Se = \frac{TP}{TP + FN}$$

$$Sp = \frac{TP}{TP + FP}$$

where TP are true positives, TN are true negatives, FP are false positives, and FN are false negatives (actives standing for all sc-PDB copies of the considered protein).

The receiver operating characteristic (ROC) curve is a graphical plot of the sensitivity (true positive rate) vs 1 – specificity (false positive rate) for a binary classifier system (active or inactive) as its discrimination threshold is varied.

Sensitivity, specificity, and area under ROC curve (ROC score) were computed for three different ranking schemes (ranking by average Goldscore, *Tc1*, and *Tc2*).

RESULTS AND DISCUSSION

Serial Ligand Docking in sc-PDB Binding Sites. Four independent screens of the sc-PDB target library were carried

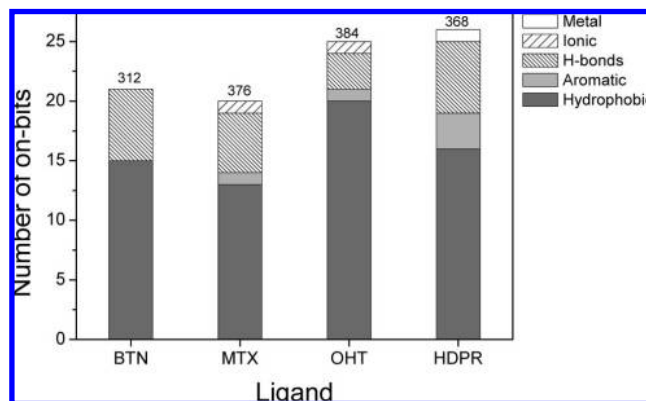


Figure 3. Distribution of molecular interactions in representative interaction fingerprints (IFPs) of four ligands cocrystallized with one of their primary targets (BTN–streptavidin 1luq entry; MTX–dihydrofolate reductase 4dfr entry; OHT–estrogen receptor α 3ert entry; HDPR–adenosine deaminase: 1a4m entry). The total IFP lengths are indicated on the top of the columns.

out by high-throughput docking. The ligands of interest were biotin (BTN), methotrexate (MTX), 4-hydroxy-tamoxifen (OHT), and 6-hydroxy-1,6-dihydropurine ribonucleoside (HDPR) (Table 1). Each of these ligands has registered primary and secondary targets in the sc-PDB (Table 1). By primary targets, we here mean the proteins for which the current ligand shows the best in vitro affinity. Secondary targets also bind to the ligand but usually with a lower affinity. The latter were selected either from existing literature⁹ or the occurrence of cocrystal structures in the sc-PDB.

All four ligands are chemically diverse and exhibit different molecular properties (Table 1) and different binding modes to their primary targets (Figure 3). Although hydro-

Table 2. Statistical analysis of inverse screening runs for four ligands

ligand	BTN	MTX	OHT	HDPR
percent of successful dockings ^a	100	98.7	96.7	100
distribution of average Goldscores ^b				
mean \pm std ^c	43.4 \pm 7.0	50.8 \pm 11.5	27.0 \pm 10.0	46.5 \pm 6.5
Z maximal value ^d	4.5	3.5	3.4	5.2
percent of entries with $-1.2 < Z < 1.2$	79.1	87.1	80.7	79.0
percent of entries with $Z \geq 1.2$	11.7	6.5	10.1	12.0
distribution of Tc1 scores ^e				
mean \pm std	0.448 \pm 0.139	0.494 \pm 0.127	0.424 \pm 0.152	0.457 \pm 0.143
Z maximal value	3.9	4.0	3.8	3.8
percent of entries with $-1.2 < Z < 1.2$	77.7	81.2	81.1	77.8
percent of entries with $Z \geq 1.2$	11.8	10.7	10.7	12.0
distribution of Tc2 scores ^f				
mean \pm std	0.289 \pm 0.221	0.301 \pm 0.216	0.103 \pm 0.149	0.301 \pm 0.202
Z ^g maximal value	3.2	3.2	6.0	3.5
percent of entries with $-1.2 < Z < 1.2$	71.0	73.0	90.0	78.6
percent of entries with $Z \geq 1.2$	11.8	11.7	10.0	10.0

^a Percentage of sc-PDB entries for which at least one docking pose could be registered. ^b GOLD fitness score averaged over a maximum of 10 independent runs. ^c Mean value \pm standard deviation. ^d $Z = (\text{score} - \text{mean})/\text{std}$. ^e Similarity on full length IFPs. ^f Similarity on truncated IFPs (see Computational Methods).

phobic contacts are always predominant in the corresponding protein–ligand IFPs and mainly encode for shape recognition, aromatic stacking is absent in BTN recognition, ionic interactions only occur in two test cases, and metal coordination only appears for HDPR binding. As a consequence, the number of successful dockings is slightly different in the four runs (Table 2) and is typically due to failures in docking large ligands into the smallest binding pockets. GOLD²⁸ fitness (GS) values depend on the ligand (the more hydrophobic, the lower the GS score). The overall shapes of the four GS distributions are nearly identical (Figure 4, Table 2). About 80% of sc-PDB database entries are ranked within one standard deviation of the average GS mean, and ca. 10% of them show significantly higher GS values. Similar observations can be drawn by inspecting IFP similarity distributions (Tc1 and Tc2 scores). Therefore, fusing GS and Tc1/Tc2 values in a consensus scoring should be most efficient by selecting only the top 10% best-ranked entries for each individual target list.

Neither Fitness nor IFP Similarity Scores Can Alone Ensure an Efficient Ranking of True Primary Targets.

BTN's true primary target is bacterial streptavidin. The sc-PDB database contains 13 streptavidin entries, which share similar ligand-binding sites, all being suitable for biotin binding. GOLD accurately docked BTN into all sc-PDB streptavidin entries; the predicted binding modes are highly similar to that observed in the cocrystallized biotin–streptavidin complex (Figure 5, Supporting Information Table S1). In the screening for BTN targets, the ranking of the 4300 sc-PDB entries using Goldscore (GS) allowed the recovery of 12 true primary targets among the top 10% scorers (Figure 4A). The Tc1 score clearly underperformed the GS score in ranking true primary targets, and only placed seven of them among the top 10% scorers, despite the good docking poses. This observation is a direct consequence of the chemical heterogeneity of the reference ligands cocrystallized with the 13 streptavidin entries, from which the Tc1 score is derived. Although all streptavidin ligand-binding sites describe the same protein cavity, the cocrystallized ligands belong to three distinct chemotypes (biotin analogues, benzoate derivatives, and a peptide) corresponding to three different binding modes.

Similar observations could be drawn from the sc-PDB screening for HDPR targets. HDPR docking into its true primary target (adenosine deaminase) was accurate (Supporting Information Table S2), and GS efficiently ranked adenosine deaminase entries among the top scorers; however, IFP scoring by Tc1 values discriminated adenosine deaminase binding sites cocrystallized with nucleoside and non-nucleoside inhibitors (Figure 4B). In both BTN and HDPR “easy docking” cases, GS efficiently prioritized the true primary target whereas Tc1 ranking was less precise, due to several possible binding modes observed for the reference cocrystallized ligands.

Let us consider the more difficult example of MXT docking into its true primary target dihydrofolate reductase (DHFR). MXT is a flexible ligand, which is only partly buried upon binding to DHFR (Figure 6A). Moreover, the experimentally determined binding mode involves a cofactor (NADP) and water molecules. The 22 DHFR sc-PDB entries consist of an extremely heterogeneous set of druggable sites; in addition to conformational changes due to species specificity or protein flexibility, multiple ligand-binding site definition arises from the diversity of cocrystallized ligands (folate analogues, pyrimidine or quinazoline derivatives in presence/absence of cofactor, cofactor analogues, see Supporting Information Table S3). Docking of MXT into DHFR resulted in a wide range of poses (Figure 6B), whose accuracy was not well evaluated by the GS score (Figure 6C). Typically, the GS score failed to discriminate poses of MTX into the DHFR substrate binding site (e.g., MTX poses in 1dyh and 1dr1 entries got almost identical GS scores although only docking in 1dyh was accurate). The GS score could even not distinguish MTX docked into the cofactor site (1dr7 entry) from MTX reasonably well redocked into the substrate site of DHFR (4dfr entry). By contrast, IFP similarity scores efficiently captured the key interactions that involved important protein residues. Despite the flexibility of MTX solvent exposed moiety, these interactions were mainly fulfilled, consequently yielding to high Tc1 and Tc2 values. In the sc-PDB screening for MTX targets, Tc1 consequently outperformed GS in ranking true primary targets (Figure 4C). The first DHFR entry was ranked in

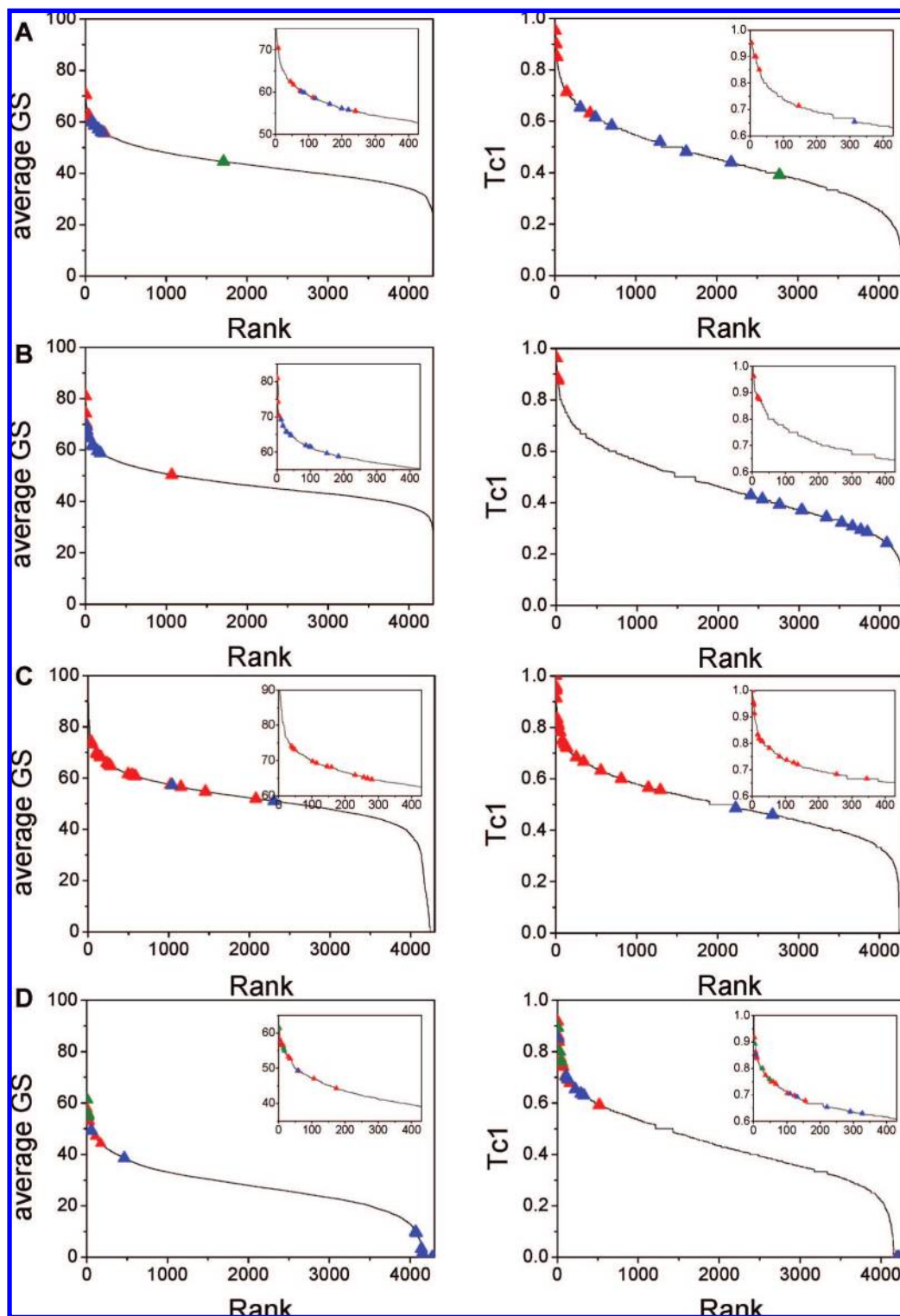


Figure 4. Ranking sc-PDB entries by Gold fitness (GS) and fingerprint similarity (Tc1) scores. GS and Tc1 distributions are presented on the left and right panels, respectively. (A) BTN screen, streptavidin sites derived from biotin-like compounds (red), benzoate derivatives (blue), and peptide (green). (B) HDPR screen, adenosine deaminase sites derived from nucleoside derivatives (red) and non-nucleoside inhibitors (blue). (C) MTX screen, dihydrofolate reductase substrate (red) and NADP cofactor (blue) sites, (D) OHT screen, estrogen receptor α (red), estrogen receptor β (blue), and estrogen-related receptor γ (green). The top 10% entries are displayed on the top right corner of each plot.

position 39 of sc-PDB entries sorted by GS, whereas nine DHFR were found among the top 39 scorers (ca. top 1% scorers) when ranked by the Tc1 value.

The last example we investigated concerns the screening for OHT targets. OHT true primary targets consist of the estrogen receptor α (ER α), the estrogen receptor β (ER β),

and the estrogen-related receptor γ (ERR γ). OHT is an antagonist thereby exclusively binds to the antagonist state of estrogen receptors. Activation of receptor is accompanied by large scale conformational changes preventing OHT entry into its binding site (Figure 7). In the sc-PDB database, the percentage of antagonized protein structures is 58.8%, 16.7%,

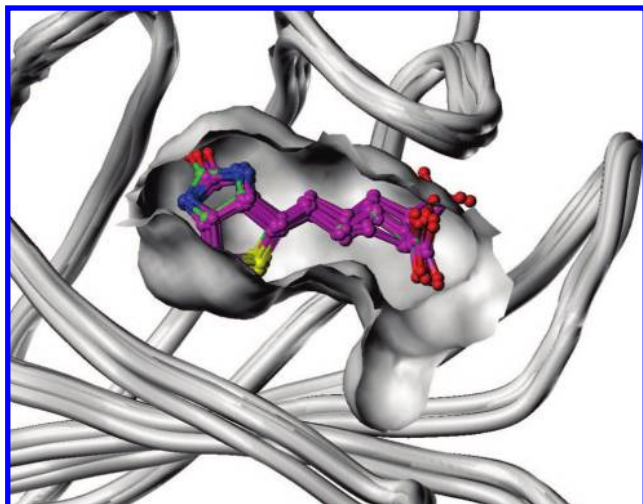


Figure 5. Docked poses of biotin in streptavidin. Docked conformations (magenta) after best-fit C α superposition of streptavidin coordinates used for docking and the 1luq entry (cocrystallized biotin in green). White wires represent the backbones of fitted proteins. The white Connolly surface shows the binding cavity in the 1luq entry. The rmsd of docked biotin heavy atoms from the X-ray coordinates ranges from 0.33 to 2.03 Å.

and 100% for ER α , ER β , and ERR γ , respectively. Docking of OHT failed in the agonist state of estrogen receptors and mainly succeeded in the antagonist state, leading to a bimodal distribution of OHT primary true targets in both GS and Tc1 ranking lists (Figure 4D, Supporting Information Table S4). Note that inaccurate OHT poses in agonist state of ER β yielded Tc1 scores higher than 0.6 (which is the top 10% threshold) and Tc2 scores close or equal to 0. A common feature of very hydrophobic binding sites, like that of estrogen receptors, is overestimation of Tc1 due to preponderance of non directional apolar interactions in the fingerprint. Tc2 is thus required to unambiguously identify inaccurate docking poses.

As previously reported,⁹ we thus observed that GOLD is able to accurately predict the binding mode of the ligands of BTN, MTX, OHT, and HDPR into their primary true targets, provided that the selected ligand-binding site is the suitable one. Here, we showed that good poses do not necessarily correspond to high GS values. IFP Tc1/Tc2 scores help to correct this situation at the condition that the query (ligand under investigation) and the reference ligands (compound cocrystallized with the binding site of interest) share similar pharmacophoric features.

Is the ROC Score a Better Metric than Enrichment to Rank Targets? Since none of the scoring functions can efficiently rank targets for all four investigated ligands by applying a simple threshold, we investigated the possibility of applying a thorough statistical treatment of GS and Tc1 scores for target selection. Unfortunately and as already mentioned in previous reports,^{9,12} the benefit of such an approach varies with the number of registered entries describing each protein stored in the database. Multiple ranking strategies using enrichment-based or ROC-based curves³⁰ can be applied at the condition that a precise protein is sampled enough throughout the curve. Annotation of sc-PDB entries by a protein generic name allows to discriminate *populated* proteins (at least five copies in the data set) from rare proteins (less than five entries) and consequently apply different statistical treatments to both

protein subsets. Among the 1434 different proteins present in the sc-PDB, 189 belong to the category of populated proteins for which sensitivity (percentage of the protein copies retrieved at a defined threshold) and ROC curves can be iteratively computed.

Four composite ranking protocols were thus tested for their ability to optimize both sensitivity (percentage of the true targets retrieved) and precision (percentage of true targets among selected proteins). In a first protocol (protocol 1, Table 3) a target was selected if 50% of their corresponding sc-PDB entries (sharing the same generic protein name) were present among the top 10% scorers in three target lists ranked by decreasing GS, Tc1, and Tc2 scores. The loose top 10% threshold was selected for its ability to retrieve entries with Z scores significantly above 1 whatever the metric used (Table 2). The ranking protocol successfully identified a true target in two cases (MTX, OH) without generating false positives, but the method is too drastic and prone to generate false negatives, especially in the BTN and HDPR screens.

The observed low target recovery rates are not necessarily related to docking failures nor to inaccurate scoring, but can result from the heterogeneity of ligand-binding sites for a single target, in which the most suitable one for the ligand under investigation is insufficiently represented. The OHT screening is an illustrative example of multiple binding site definitions. Among the 12 sc-PDB copies of ER β , only two correspond to the antagonist state of the receptor to which OHT binds to. Consequently, docking OHT into ER β binding sites was unproductive 10 times and was successful only twice, resulting in overall poor statistics for this receptor (Supporting Information Table S4). Out of the 17 sc-PDB copies of ER α , a majority (10) describe the relevant antagonist state thus explaining why this receptor is selected by the ranking protocol 1.

The area under ROC curve (ROC score) is insensitive to early recognition³⁰ and has been demonstrated to correspond to a linear scale average of the protein ranks. In the present application, it may account for the heterogeneity of ligand-binding sites. Two ranking protocols (protocols 2 and 3, Table 3) were thus used to prioritize targets exhibiting a ROC score above a defined threshold (0.7 in the current study). Whereas protocol 2 ranked targets according to their GS score, protocol 3 was used to evaluate the Tc1 metric in a similar ranking scheme (note that we applied a Tc2-dependent correction to Tc1 in order to take into account bias introduced by hydrophobic binding sites). Both protocols are more efficient than the enrichment-based protocol 1 in selecting the true targets in the target lists (sensitivity) but at the expense of a much lower precision (Table 3).

ROC-based selection illustrates two potential weaknesses in virtual screening. First, the ROC score is highly dependent on the score distribution mode and systematically misses true targets if few of the corresponding sc-PDB entries are retrieved at the very beginning of the rank-order list and the remaining entries at the end as exemplified by ER β in OHT screen (GS and Tc1 metric) and for adenosine deaminase in HDPR screen (Tc1 metric, Table 3). Second, permissive ligand-binding sites in term of molecular interactions will be systematically selected by any ROC-derived protocol as far as scores are neither evenly distributed nor biased toward low values. This explains the higher number of false positives

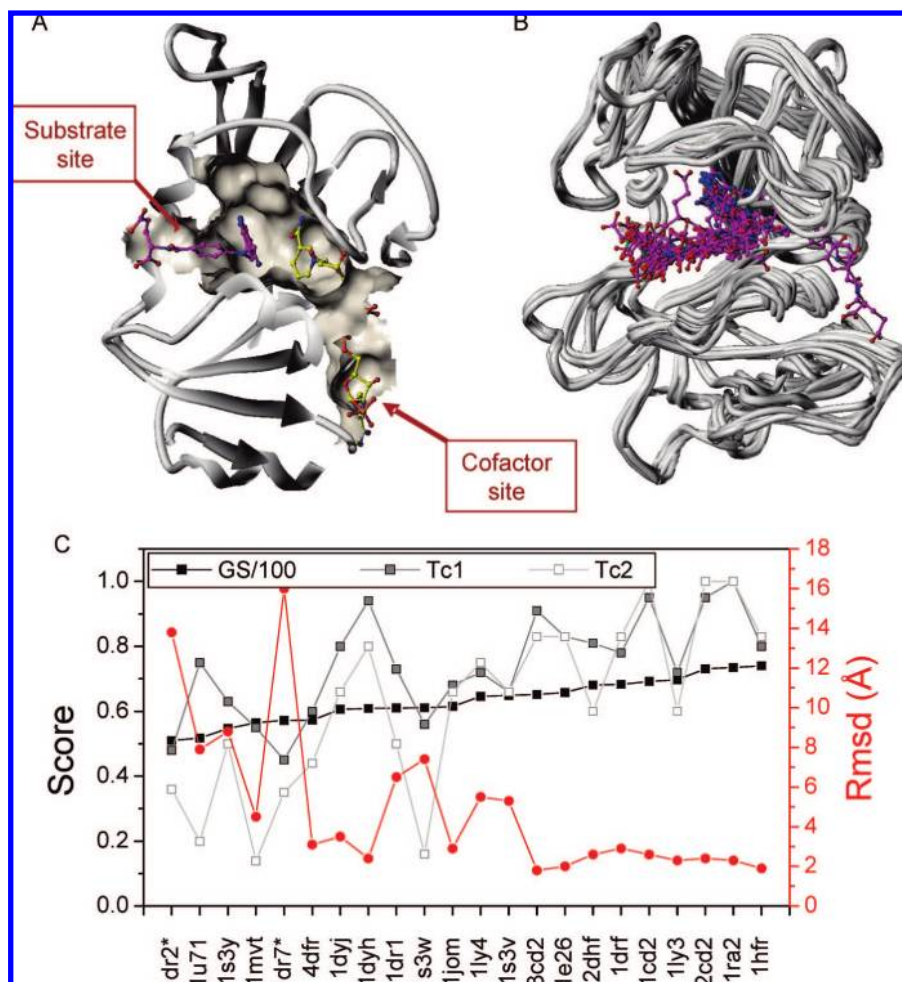


Figure 6. Three-dimensional structures of the MTX–DHFR complex. (A) Substrate (pdb entry 4dfr) and cofactor (pdb entry 1ra2) binding sites in DHFR from *E. coli*. Both protein coordinates have been matched on C α atoms. Protein-bound coordinates of MTX (magenta sticks) in the substrate-binding site and of NADP (yellow sticks) in the contiguous cofactor-binding site are shown in the protein cleft rendered as a MOLCAD³² surface channel. (B) Docked poses in all 22 DHFR entries (entries have been matched to a X-ray pose (in 4dfr, green) and overlaid after best-fit superposition of DHFR C α coordinates. White wires represent the backbones of fitted protein coordinates. (C) GS, Tc1, and Tc2 scores of MTX/DHFR docked complexes related to the root-mean-square deviation in angstroms of MTX heavy atoms in the 4dfr entry from those in the docked complex after bestfit superposition of DHFR C α coordinated. For the sake of clarity, GS values have been divided by 100 and entries were sorted by increasing GS. The two HDPR cofactor sites, in 1dr2 and 1dr7 entries, are marked by a star.



Figure 7. Conformational changes upon ER α receptor activation. The agonist state of ER α bound to the agonist estradiol (left, 1qkt entry) is compared to the antagonist state of ER α bound to the antagonist OHT (right, 3ert entry). Protein backbones are depicted using white ribbons, excepted for the activation helix 12 which is colored in green and magenta in the agonist and antagonist forms of the receptor, respectively. In the ball and stick representation of ligands, oxygen atoms are colored in red, nitrogen atoms in blue, and carbon atoms either in green (estradiol) or magenta (OHT). The binding cavity indicated by a white Connolly surface is significantly smaller and more closed in the activated state than in the inactivated state of the receptor.

generated by protocols 2 and 3 (20–26% of all sc-PDB populated proteins upon fitness ranking and 9–18% upon Tc1 ranking, Table 3). Interestingly, none of the false positives selected by Tc1-based ROC curves are common to the four screens, whereas four frequent hitters were observed upon

the GS-based ROC selection (5'-deoxyribonucleotidase, isopenicillin *N* synthetase, Pentaerythritol tetranitrate reductase, thermolysin). For these four proteins sharing very polar active sites, all sc-PDB copies are scored in the upper part of the GS ranking list (Z score of ca. 1).

Table 3. Composition of Target Lists from the Set of Populated Proteins

ligand	BTN	MTX	OHT	HDPR
ranking protocol 1 ^a				
true positives ^b	0	1	1	0
false positives ^c	0	0	0	1
sensitivity, % ^d	0	50	33	0
precision, % ^e	0	100	100	0
ranking protocol 2 ^f				
true positives	1	1	1	1
false positives	37	36	41	47
sensitivity, %	100	50	33	100
precision, %	2.6	2.7	2.3	2.1
ranking protocol 3 ^g				
true positives	1	1	1	0
false positives	31	18	15	34
sensitivity, %	100	50	33	0
precision, %	3.1	5.2	6.25	0
ranking protocol 4 ^h				
true positives	1	1	1	0
false positives	12	7	3	12
sensitivity, %	100	50	33	0
precision, %	7.7	12.5	25	0

^a Hits are proteins with at least 50% of their corresponding entries exhibiting GS, Tc1, and Tc2 values among the top 10% scoring sc-PDB entries ranked by decreasing GS, Tc1, and Tc2 scores, respectively. ^b Number of true targets in the populated target list (see exhaustive list of true targets in Table 1). ^c Number of targets not known to bind to the ligand but present in the populated target list. ^d Percentage of true targets recovered in the populated target list: (true positives/true targets) \times 100. ^e Percentage of true targets in the populated target list: (true positives/positives) \times 100. ^f Hits are proteins with a ROC score (area under the ROC curve) higher than 0.7, when entries are sorted by decreasing GS values. ^g Hits are proteins with a ROC score (area under the ROC curve) higher than 0.7, when entries are sorted by decreasing Tc1 values (Tc1 set to 0 if Tc2 < 0.15). ^h Hits are proteins common to the two lists defined by ROC-based protocols 2 and 3.

Combination of the two ROC-based target lists (consensus protocol 4) significantly decreases the number of false positives to a reasonable low amount of targets (from 3 to 12 which means less than 6.5% of all sc-PDB populated proteins) without affecting the target sensitivity. It would still miss the adenosine deaminase as true target of HDPR because of the inaccuracy of the Tc1 metric in this precise case. In fact, the medium accuracy of ROC-based selection using either GS or the Tc1 metric is not due to the metric itself but to imperfect annotation of ligand-binding sites. In all four cases, an additional level of manual annotation based on the precise location of the ligand-binding site and protein conformation (e.g., active vs inactive state of estrogen receptors) drastically increases ROC scores (Figure 8) for true targets of the four reference ligands, whatever the metric used. It is worth noticing that the GS score is accurate enough to discriminate true from wrong binding sites in all four cases. Although it requires an additional postprocessing step, the Tc1 metric is particularly well suited for the precise binding site triage, once the annotation level is high enough.

How to Handle Rare Proteins. The ROC statistics can only be used if a property (protein name in the present case) is sufficiently sampled in the data set. We here made the choice to apply it to proteins present in at least five copies in the database. However, the sc-PDB contains much more

rare proteins ($n = 1245$) in low copy number (<5) than populated proteins ($n = 189$) in high copy number (≥ 5). Rare proteins thus deserve a special statistical treatment. Among the various possibilities based on enrichment calculations, the ranking protocol 1 (Table 3) which prioritize proteins common to the top 10% scorers (GS, Tc1, and Tc2 scores) appears the best (Table 4). Corresponding target lists are very small (from 2 to 13 targets) and contains few false positives. However, the ranking scheme is far from perfect. It recovers four rare true targets out of five for biotin (hypothetical biotin ligase is omitted) but misses pterin reductase as MTX target and three rare targets (protein kinase C, prostaglandin synthase I, quinone oxidoreductase) out of four in the OHT screen.

Customized Protocol to Efficiently Retrieve True Targets in All Scenarios. Having selected relatively robust ranking protocols for the two populations of targets (populated, rare) present in the sc-PDB, we next combined them into a unique workflow to address all targets at the same time. Protocol A (Table 5) combines the ROC-based protocol 4 for populated targets (Table 3) and the enrichment-based protocol 1 for rare proteins (Table 4). Altogether, 8 out of the 16 true targets of the four ligands investigated herein are effectively recovered. As expected, primary targets (for which the ligands exhibit the highest affinity) are easier to recover than secondary targets. However, the ranking protocol A still misses two primary targets (ER β for OHT, adenosine deaminase for HDPR) mainly because of the current inaccuracy of binding site definition which merges agonist and antagonist states (ER β) and nucleoside and non-nucleoside binding sites (adenosine deaminase). We thus decided to apply a fuzzier selection scheme (protocol B) consisting in selecting any protein for which at least one entry is common to three lists (top 10% scorers) of sc-PDB entries ranked by decreasing GS, Tc1, and Tc2 scores, respectively. Protocol B selects 12 out of the 16 true targets and misses no primary target for any of the four ligands (Table 5). Although protocol B is less drastic than protocol A, it does not select significantly more false positives than protocol A (Table 6). The obtained target lists contain a proportion of rare vs populated proteins similar to that obtained with protocol A. The sensitivity is notably enhanced. In the BTN screen, all primary and secondary targets were identified, especially the hypothetical biotin-ligase which was lost by protocol A. In the MTX screen, the secondary target thymidilate synthase was indeed selected, whereas it was not omitted in all previously investigated ranking scenarios. In the OHT screen, ER β was successfully identified. It is true that none of the three known secondary targets of OHT were retrieved. However, there are no structural data describing the corresponding binding site location which may significantly differ from the canonical hormone binding site. Last, adenosine deaminase was successfully retrieved in the HDPR screen, along with other ribonucleoside-binding proteins (e.g., cytidine deaminase) which were considered here as false negatives but could in fact be true targets of ADPR due to their binding site similarity to that of adenosine deaminase. The overall precision is significantly higher (from 7 to 30%) than that expected from random picking (from 0.07% for HDPR targets to 0.4% for biotin and OHT targets) and the target lists importantly small enough (from 10 to 26 proteins)

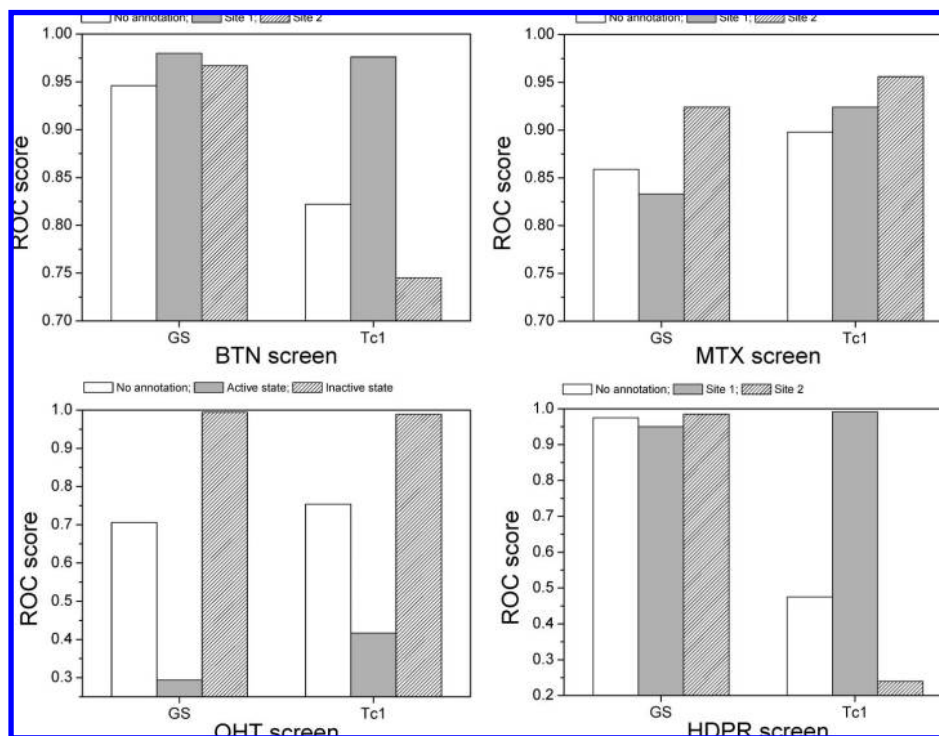


Figure 8. Variation of ROC scores for populated true primary targets of BTN, MTX, OHT, and HDPR, according to binding site annotation. No annotations means that all sc-PDB entries sharing the generic protein name of a true populated target are considered as true positives. In the BTN screen, site 1 and site 2 are the biotin and benzoate-binding sites in streptavidin, respectively. In the MTX screen, site 1 and site 2 are the substrate and substrate/cofactor-binding sites in DHFR, respectively. In the OHT screen, only ER α binding sites have been considered in their active and inactive state, respectively. In the HDPR screen, site 1 and site 2 are nucleoside- and non-nucleoside-binding sites in adenosine deaminase, respectively.

Table 4. Composition of Target Lists^a from the Set of Rare Proteins

ligand	BTN	MTX	OHT	HDPR
true positives ^b	4	0	1	0
false negatives ^c	1 ^f	1	3	0
false positives ^d	9	6	1	6
sensitivity, % ^e	80	0	25	na ^g
precision, % ^f	31	0	50	na

^a Hits are rare proteins with at least 50% of their corresponding entries exhibiting GS, Tc1, and Tc2 values among the top 10% scoring sc-PDB entries ranked by decreasing GS, Tc1, and Tc2 scores, respectively. ^b Number of true targets in the rare target list (see exhaustive list of targets in Table 1). ^c Number of true targets missed by the rare target list. ^d Number of targets not known to bind to the ligand but present in the rare target list. ^e Percentage of true targets recovered in the rare target list: (true positives/true targets) \times 100. ^f Percentage of true targets in the rare target list: (true positives/positives) \times 100. ^g Not applicable since there are no rare targets reported for HDPR.

to initiate an exhaustive experimental evaluation of in silico selected targets.

The best possible protein ranking strategy for the four ligands is thus to combine both an energy-based and a topology-based scoring scheme and select any target for which at least one entry is present among the top 10% scorers using three separate ranking lists based on a docking value (GS) and two topological scores (Tc1, Tc2). This strategy, which is derived only from four ligands covering only 16 out of the 1434 targets registered in the sc-PDB, may be however not directly applicable to any new prospective screen. A systematic docking of all nonredundant sc-PDB ligands in all 4300 sc-PDB entries should enable to derive

more general rules for enhancing the accuracy of target ranking with a thorough statistical control.

CONCLUSIONS

The current study illustrates the limitation of single scoring functions to efficiently prioritize targets in inverse screening campaigns. As to be expected, docking scores are not accurate enough, select numerous false positives and cannot be applied for rare proteins where there is no rationale to applying a particular cutoff (only Goldscore has been investigated herein but the same conclusions are likely to be drawn for any energy-based scoring function). Using interaction fingerprints in a topological scoring function helps rescuing a lot of situations inadequately treated by docking scores but requires a very precise binding site annotation at the structural level to prevent comparing overlapping or distinct ligand-binding cavities. This annotation level is currently missing in the sc-PDB database. This study also revealed limitations that have to be overcome. First, the Tc1 score would gain in using multiple IFP references originating from all ligands cocrystallized with a single protein, then fusing data. This would avoid poor scoring due to different binding modes. This improvement is however quite hard to set up, because in present IFPs, the size as well as the bit ordering are binding site dependent and not necessarily consistent for sc-PDB entries describing a single protein. The current sc-PDB annotation by protein generic name is not sensitive to significant conformational change (e.g., agonist/antagonist states), multiple druggable sites (e.g., cofactor and substrate sites), or interspecies residue variability. An

Table 5. Effectiveness of Two Customized Protocols for Ranking All True Targets

Target	Protocol A ^a				Protocol B ^b			
	BTN	MTX	OHT	HDPR	BTN	MTX	OHT	HDPR
Streptavidin	✓				✓			
<i>Avidin-related protein 2</i>	✓				✓			
<i>Avidin-related protein 4/5</i>	✓				✓			
<i>Avidin</i>	✓				✓			
<i>BirA protein</i>	✓				✓			
<i>Hypothetical biotin ligase</i>	✗				✓			
Dihydrofolate reductase		✓				✓		
<i>Pterine reductase</i>		✗				✗		
<i>Thymidilate synthase</i>		✗				✓		
Estrogen receptor α			✓				✓	
Estrogen receptor β			✗				✓	
Estrogen-related receptor γ			✓				✓	
<i>Protein kinase C</i>			✗				✗	
<i>Prostaglandin synthase I</i>			✗				✗	
<i>Quinone oxidoreductase</i>			✗				✗	
Adenosine deaminase				✗				✓

^a Ranking protocol 4 for populated targets (Table 3) and ranking protocol 1 for rare targets (Table 4). ^b Selection of any protein for which at least one sc-PDB entry is scored among the top 10% scorers of three sc-PDB entry lists ranked by decreasing GS, Tc1, and Tc2 scores. Rare proteins are displayed in italic. A green icon means that the target has been selected whereas a red cross means that the target has been missed by the corresponding ranking protocol.

Table 6. Composition of Target Lists from All sc-PDB Proteins

target	protocol A ^a				protocol B ^b			
	BTN	MTX	OHT	HDPR	BTN	MTX	OHT	HDPR
targets selected	23	14	6	18	26	19	10	15
true positives	5	1	2	0	6	2	3	1
false positives	18	13	4	18	20	17	7	14
false negatives	1	2	4	1	0	1	3	0
sensitivity, %	83	33	33	0	100	66	50	100
precision, %	22	7	33	0	23	10	30	7

^a Ranking protocol 4 for populated targets (Table 3) and ranking protocol 1 for rare targets (Table 4). ^b Selection of any protein for which at least one sc-PDB entry is scored among the top 10% scorers of three sc-PDB entry lists ranked by decreasing GS, Tc1, and Tc2 scores.

improvement of the structural annotation of ligand-binding sites should be reached by a systematic pairwise comparison of registered binding sites for a single generic protein name thus affording to cluster binding sites according to their 3D similarity.³¹

ACKNOWLEDGMENT

We thank the Centre Informatique National de la Recherche Scientifique (CINES, Montpellier, France) for allocation of computing time on the Linux cluster.

Supporting Information Available: Table S1 (Streptavidin ranking in BTN screen), Table S2 (Adenosine deami-

nase ranking in HDPR screen), Table S3 (Dihydrofolate reductase ranking in MTX screen), and Table S4 (Estrogen receptors ranking in OHT screen). This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES AND NOTES

- (1) Rognan, D. Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* **2007**, *152*, 38–52.
- (2) Hall, S. E. Chemoproteomics-driven drug discovery: addressing high attrition rates. *Drug Discov. Today* **2006**, *11*, 495–502.
- (3) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* **2006**, *46*, 1124–1133.
- (4) Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging chemical and biological space: “target fishing” using 2D and 3D molecular descriptors. *J. Med. Chem.* **2006**, *49*, 6802–6810.
- (5) Bender, A.; Jenkins, J. L.; Glick, M.; Deng, Z.; Nettles, J. H.; Davies, J. W. “Bayes affinity fingerprints” improve retrieval rates in virtual screening and define orthogonal bioactivity space: when are multitarget drugs a feasible concept. *J. Chem. Inf. Model.* **2006**, *46*, 2445–2456.
- (6) Mestres, J.; Martin-Couce, L.; Gregori-Puigjane, E.; Cases, M.; Boyer, S. Ligand-based approach to in silico pharmacology: nuclear receptor profiling. *J. Chem. Inf. Model.* **2006**, *46*, 2725–2736.
- (7) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25*, 197–206.
- (8) Steindl, T. M.; Schuster, D.; Laggner, C.; Langer, T. Parallel screening: a novel concept in pharmacophore modeling and virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 2146–2157.
- (9) 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: Struct., Funct., Bioinf.* **2004**, *54*, 671–680.

- (10) Chen, Y. Z.; Zhi, D. G. Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins: Struct., Funct., Bioinf.* **2001**, *43*, 217–226.
- (11) Li, H.; Gao, Z.; Kang, L.; Zhang, H.; Yang, K.; Yu, K.; Luo, X.; Zhu, W.; Chen, K.; Shen, J.; Wang, X.; Jiang, H. TarFisDock: a web server for identifying drug targets with docking approach. *Nucleic Acids Res.* **2006**, *34*, W219–W224.
- (12) Muller, P.; Lena, G.; Boilard, E.; Bezzine, S.; Lambeau, G.; Guichard, G.; Rognan, D. In silico-guided target identification of a scaffold-focused library: 1,3,5-triazepan-2,6-diones as novel phospholipase A2 inhibitors. *J. Med. Chem.* **2006**, *49*, 6768–6778.
- (13) Zahler, S.; Tietze, S.; Totzke, F.; Kubbutat, M.; Meijer, L.; Vollmar, A. M.; Apostolakis, J. Inverse in silico screening for identification of kinase inhibitor targets. *Chem. Biol.* **2007**, *14*, 1207–1214.
- (14) Kellenberger, E.; Muller, P.; Schalon, C.; Bret, G.; Foata, N.; Rognan, D. sc-PDB: an Annotated Database of Druggable Binding Sites from the Protein Data Bank. *J. Chem. Inf. Model.* **2006**, *46*, 717–727.
- (15) Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. Binding MOAD (Mother Of All Databases). *Proteins: Struct., Funct., Bioinf.* **2005**, *60*, 333–340.
- (16) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. 3rd, Assessing scoring functions for protein-ligand interactions. *J. Med. Chem.* **2004**, *47*, 3032–3047.
- (17) Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195–207.
- (18) Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFT): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **2004**, *47*, 337–344.
- (19) Mpamhanga, C. P.; Chen, B.; McLay, I. M.; Willett, P. Knowledge-based interaction fingerprint scoring: a simple method for improving the effectiveness of fast scoring functions. *J. Chem. Inf. Model.* **2006**, *46*, 686–698.
- (20) 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.
- (21) 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*, D115–119.
- (22) Sybyl, version 7.3; TRIPOS: St. Louis, MO, 2007.
- (23) 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.
- (24) Filter, version 2.0.1; OpenEye Scientific software: Santa Fe, NM, 2007.
- (25) JChem, version 3.2.3; ChemAxon: Budapest, Hungary, 2007.
- (26) Corina, version 3.4; Molecular Networks GmbH: Erlangen, Germany, 2007.
- (27) OEChem, version 1.4.2; OpenEye Scientific software: Santa Fe, NM, 2007.
- (28) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins: Struct., Funct., Bioinf.* **2003**, *52*, 609–623.
- (29) Pipeline Pilot, version 6.1; SciTegic: San Diego, CA, 2007.
- (30) Truchon, J. F.; Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the “early recognition” problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.
- (31) Schalon, C.; Surgand, J.-S.; Kellenberger, E.; Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins: Struct., Funct., Bioinf.*, available online March 4, 2008, <http://dx.doi.org/10.1002/prot.21858>.
- (32) Exner, T. E.; Keil, M.; Moeckel, G.; Brickmann, J. Identification of Substrate Channels and Protein Cavities. *J. Mol. Model.* **1998**, *4*, 340–343.

CI800023X