Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

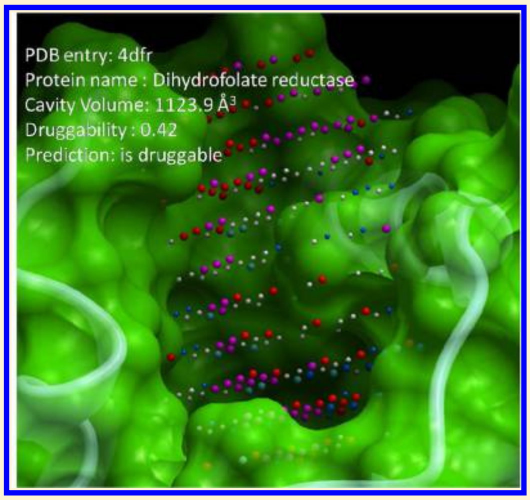
Comparison and Druggability Prediction of Protein−Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes

Jérémy Desaphy, Karima Azdimousa, Esther Kellenberger, and Didier Rognan*

Laboratory of Therapeutic Innovation, UMR 7200 Université de Strasbourg/CNRS, Medalis Drug Discovery Center, F-67400 Illkirch, France

Supporting Information

ABSTRACT: Estimating the pairwise similarity of protein−ligand binding sites is a fast and efficient way of predicting cross-reactivity and putative side effects of drug candidates. Among the many tools available, three-dimensional (3D) alignment-dependent methods are usually slow and based on simplified representations of binding site atoms or surfaces. On the other hand, fast and efficient alignment-free methods have recently been described but suffer from a lack of interpretability. We herewith present a novel binding site description (VolSite), coupled to an alignment and comparison tool (Shaper) combining the speed of alignment-free methods with the interpretability of alignment-dependent approaches. It is based on the comparison of negative images of binding cavities encoding both shape and pharmacophoric properties at regularly spaced grid points. Shaper approximates the resulting molecular shape with a smooth Gaussian function and aligns protein binding sites by optimizing their volume overlap. Volsite and Shaper were successfully applied to compare protein−ligand binding sites and to predict their structural druggability.



INTRODUCTION

Although the pace of protein structure determination is by far inferior to that of protein sequencing, outstanding efforts of structural genomics consortia[1,2] and methodological advances in structural biology[2−4] contribute to considerably change our understanding of the structural proteome. For example, the most important family of drug targets (G Protein-coupled receptors) has long been described by a single representative[5] in the Protein Data Bank (PDB),[6] but has been supplemented by 14 novel receptors and 30 new receptor−ligand complexes in the last 4 years.[7,8] A comprehensive coverage of UniProt targets by the PDB is therefore anticipated in ca. 15 years.[9] Among the applications that will benefit from a better structural coverage of biological space is the prediction of off-targets and resulting side effects for known drug candidates.[10,11] Hence, if one assumes that similar protein pockets accommodate similar ligands,[12] it is possible to predict ligand cross-reactivity to targets sharing similar ligand-binding sites.[13−16] To achieve this goal, a pocket detection algorithm must first be precise enough to focus on druggable binding sites only, quantitative comparison should then remain fuzzy enough to accommodate ligand-induced structural changes.[17] A key issue is the ability to detect a three-dimensional (3D) similarity for binding sites of unrelated proteins in absence of fold conservation. In most cases, cavity comparison tools rely on the prior 3D alignment of a set of representative protein atoms.[17] These methods are relatively slow (1−10 comparisons/min) and highly dependent on atomic coordinates but are easy to interpret and to couple with other computational methods like ligand docking.[10] A wrong 3D alignment of two binding sites, whatever the reason, will however underestimate their true similarity.[18] Alignment-independent methods,[18−22] amenable to a very high throughput (up to 1000 comparisons/s) have thus been recently described to quantify pocket similarities but still suffer from a lack of interpretability since 3D information is often loss upon converting cavity properties into simple 1D fingerprints.

We herewith introduce a novel cavity description and comparison method combining the advantages of alignment-dependent methods with the speed of alignment-independent methods. By contrast to many existing tools, the protein−ligand binding pocket is not represented by either protein atoms or surface points but by regularly spaced pharmacophoric grid points defining its inverse image from a protein−ligand interaction point of view. We next used a shape-based alignment method[23] using a smooth Gaussian function approximating molecular volumes, to align cavities by optimizing the volume overlap of their pharmacophore-annotated shapes. The method is simple, fast, and particularly efficient in detecting binding site similarity in the absence of sequence and fold conservation. It can be used for three main applications: (i) infer the function of a protein by measuring the similarity of its known or potential ligand-binding pockets to a collection of functionally annotated binding sites, (ii)

Received: April 13, 2012
Published: July 26, 2012

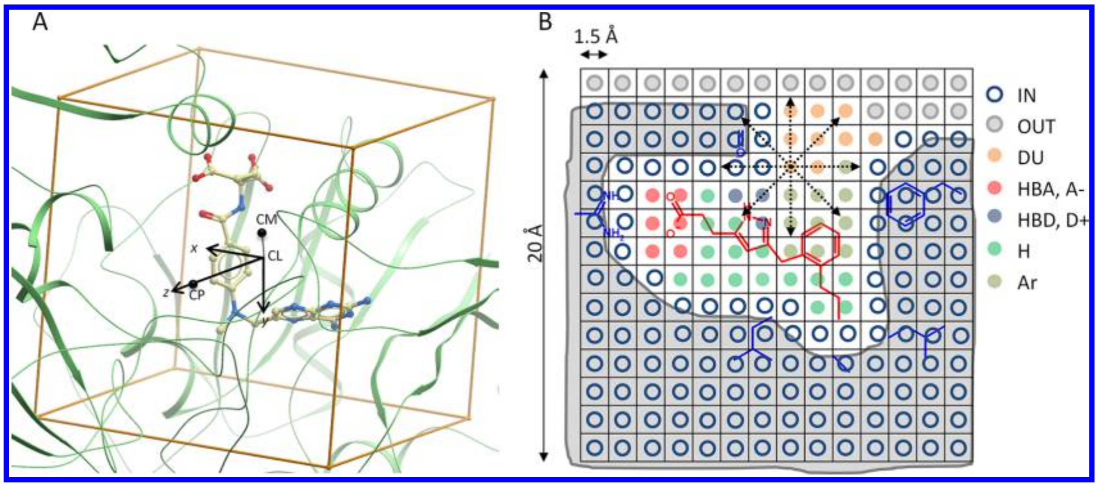


Figure 1. Orientation and definition of the grid lattice from atomic coordinates of a protein−ligand complex. (A) From the center of mass of the ligand (CL, grid center), and the center of mass of the protein (CP), coordinates of the reference inertia point CM are computed ($\mathrm{CM} = \mathrm{CL} + 1/N_l[\sum_{i=1}^{N_l} w_i(a_i - \mathrm{CL})^2]^{1/2}$, N_l number of ligand heavy atoms, w_i mass of atom i, a_i coordinates of atom i). The three main axes defining the grid lattice are deduced computing first the vector normal to $\overrightarrow{\mathrm{CLCP}}$ and $\overrightarrow{\mathrm{CLCM}}$ (x axis), then the vector normal to $\vec{x}$ and $\overrightarrow{\mathrm{CLCP}}$ (y axis), and last the vector normal to $\vec{x}$ and $\vec{y}$ (z axis). (B) A 3D lattice of size 20 Å and resolution 1.5 Å is centered on the ligand (red sticks) center of mass. A site point is placed at the center of each of the 2730 cells and assigned a property according to its location with respect to the protein (gray solid surface) active site residues (blue sticks): IN (cell intersecting any protein atom or site point closer than 2.5 Å to any protein atom), OUT (site point outside the cavity), DU (site point inside the cavity but farther than 4.0 Å from any protein atom), HBA (site point inside the cavity close to a protein H-bond donor), HBD (site point inside the cavity close to a protein H-bond acceptor), D+ (site point inside the cavity close to a negative ionizable protein atom), A− (site point inside the cavity close to a positive ionizable protein atom), HYD (site point inside the cavity close to a protein aliphatic apolar atom), AR (site point inside the cavity close to a protein aromatic atom). Assigning site points inside or outside the cavity is decided with respect to the proportion of 8 Å-long rays (dotted arrows) projected from every point intersecting an IN cell.

classify targets according to the similarity of their binding sites, (iii) predict the structural druggability (ligandability) of a binding site from the properties of its pharmacophoric shape.

■ METHODS

Pharmacophoric Annotation of Cavity Grid Points (VolSite). Starting from atomic coordinates of a protein−ligand complex, a three-dimensional cube of 20 Å-edge is centered on the center of mass of the bound ligand and filled with a 1.5 Å-resolution grid defining 2370 cells of 3.375 Å^3 volume each (Figure 1). To each cell is associated a grid point and a property at its center. If the corresponding cell comprises a protein atom or if its center is less than 2.5 Å away from any protein atom, the grid point is given the "IN" property. Any other point is then checked for buriedness by generating, from its coordinates, a set of 120 regularly spaced rays of 8 Å length. If the number of rays intersecting an IN cell (N_{ri}) is smaller than a user-defined threshold (by default 40), the corresponding point is considered to be outside the enclosing cavity and is assigned the "OUT" property. Remaining points are considered to encompass the cavity and checked for direct neighborhood with other cavity points. If isolated (less than three neighbors in adjacent cells), cavity points are deleted. Site points closer than 4.0 Å to a protein atom are assigned one of seven possible pharmacophoric properties (H-bond acceptor, H-bond donor, H-bond acceptor and donor, negative ionizable, positive ionizable, hydrophobic, aromatic) complementary to that of the closest protein atom using standard interaction rules[24] (Table 1). Points with no protein atoms within 4 Å are assigned a null property. The pharmacophoric properties of protein atoms are detected on-the-fly from their names (PDB input) or atom types (MOL2 input) thus enabling in the latter case to consider additional molecules (ions, cofactors, water, prosthetic groups, nucleic acid) as part of the protein. Once every point has been assigned a pharmacophoric property, five sets of cavity points are defined with respect to their largest distance (4, 6, 8, 12 Å, any) to any protein-bound ligand heavy atom. From here on, we will refer these binding sites of increasing sizes to cavities truncated at 4, 6, 8, and 12 Å, respectively. The full cavity is defined when no truncation is applied.

Table 1. Cavity Point Properties and Pharmacophore Matching Rules

property	name	residue	closest protein atom
hydrophobic (HYD)	CA	Gly	hydrophobic
aromatic (AR)	CZ	Phe	aromatic
acceptor (HBA)	O	Ala	donor
donor (HBD)	N	Ala	acceptor
acceptor/donor (HBAD)	OG	Ser	acceptor/donor
positive ionizable (D+)	NZ	Lys	negative ionizable
negative ionizable (A-)	OD1	Asp	positive ionizable
null (DU)	DU	Cub	none

Since every cell has a fixed and unique volume (3.375 Å^3), the total number of pharmacophore-annotated cells approximates the global cavity volume. In absence of cocrystallized ligands (apo-proteins), any set of atomic coordinates (e.g., center of a cavity detected by a third-party software) can be given as input to define the 3D grid and generate cavity points. To enable their visualization by any software, standard protein atom names with corresponding pharmacophoric properties are given to each cavity point (Table 1).

Druggability Prediction. The recently described nonredundant set of druggable and less druggable binding sites (NRDLD)[25] describing 113 cavities (71 druggable, 42 undruggable) was utilized for assessing the suitability of Volsite attributes to predict the structural druggability (or ligandability) from protein X-ray structures. The corresponding protein−

ligand complexes were retrieved from the sc-PDB[26] or the Protein Data Bank;[6] cavity points were generated using standard parameters and no truncation to consider the entire cavity. The data set was split, as originally proposed[25] into a training set of 76 entries (48 druggable, 28 undruggable) and a test set of 37 entries (23 druggable, 14 undruggable).

For each entry, 73 VolSite descriptors (Supporting Information Table 1) were read as input values for a binary classification model using a support vector machine algorithm (SVM), as implemented in SVM$^{\text{light}}$.[27]

These descriptors encode the volume of the cavity as the total number of cavity points (descriptor no. 1), the proportion (expressed in percent) of points having each of the eight pharmacophoric types (in other words, hydrophobicity, aromaticity, and polarity; descriptor nos. 2–9), and the accessibility of every site point (descriptor nos. 10–73) expressed for each of the eight pharmacophoric types, as the number of "IN" cell-intersecting rays (N_{ri}) within eight ranges (40–50, 50–60, 60–70, 80–90, 90–100, 100–110, 110–120).

Optimal values for gamma and c parameters were found after systematic variation of both parameters in a 5-fold cross-validation procedure, using the rbf kernel, applied to the entire training set. We first iterate gamma from 0 to 1 using a 0.1 increment and c from 0 to 100 with a step of 1. The best average F-measure of the 5 folds gave gamma, and c values around which a novel systematic variation was repeated using a tighter range and increment (10% of the previous ones) until no gain in the F-measure was observed. The model leading to the best F-measure (gamma = 4×10^{-6} and $c = 100$) was finally selected for external predictions.

The predicted druggability of all cavities in the external test set was compared to that reported for three state-of-the-art methods, DrugPred,[25] Fpocket,[28] and SiteMap.[29] DrugPred and Fpocket values were directly taken from the literature.[25] In SiteMap v.2.2,[30] protein and ligand files were first extracted from the Protein DataBank and transformed in mol2 format in SybylX1.3.[31] After adding hydrogen atoms and manually optimizing intermolecular hydrogen bonds, protein and ligand coordinates were converted in mae file format using Maestro v8.5.[30] If the cavity contains a metal ion, the "Protein preparation wizard" in Maestro was used to verify and manually correct whenever necessary atom types. The bound ligand was used as constraint to detect the cavity boundaries within 6 Å of any ligand atom.

Binding Site Alignment (Shaper). *Method.* The alignment tool (Shaper) relies on OEChem and OEShape toolkits.[32] The main advantage of these toolkits is the possibility to describe molecular shapes by a smooth Gaussian function and to align two molecules by optimizing the overlap of their corresponding volumes.[23,33,34] During the alignment, a reference set of cavity points is kept rigid while the set of cavity points to fit (fit object) undergoes rigid body rotations and translations. To speed-up calculations, the "Grid" volume overlap method was chosen to represent the volume of the target molecule and all atom radii were set to that of carbon (1.7 Å). Once the best shape alignment has been achieved, it is scored by a "Color Force Field" (a color being a pharmacophoric feature) similar to that used by the ligand matching tool ROCS[32] to account for pharmacophoric properties matching. The force field (Supporting Information Table 2) consists in SMARTS patterns for six pharmacophoric properties (H, Ar, HBA, HBD, A−, D+) and six pattern matching rules (H to H, Ar to Ar, HBA to HBA, HBD to HBD, A− to A−, D+ to D+) to score the shape-based alignment by pharmacophoric similarity. Two pharmacophoric properties (HBAD, DU) were not considered for the color alignment since the first one is implicitly taken into account by either acceptor or donor SMART patterns and because the latter was not found relevant in preliminary trials. Color matches were considered for cavity points up to 1.5 Å apart with a single weight for all matching rules (Supporting Information Table 2).

Similarity Metric and Statistical Evaluations. The similarity $S_{\text{A,B}}$ between cavities A (reference) and B (fit) was calculated by a Tversky index as follows:

$$S_{\text{A,B}} = \frac{O_{\text{A,B}}}{\alpha I_{\text{A}} + \beta I_{\text{B}} + O_{\text{A,B}}}$$

where $O_{\text{A,B}}$ is the overlap between colors of cavities A and B, and I is non-overlapped colors of each entity A and B. By contrast to a Tanimoto index ($\alpha = \beta = 1$), the Tversky index gives more importance to either the reference or the fit object by assigning different weights ($\alpha \neq \beta$, $\alpha + \beta = 1$) to the self-color nonoverlap I_{A} and I_{B} values. The metric is asymmetric and varies between 0 and 1. Preliminary trials indicated that the peak performance was reached with $\alpha = 0.95$ and $\beta = 0.05$ (from hereon *Ref Tversky metric*). Classification models based on pairwise similarity values were assessed by computing the area under the receiver operating characteristic (ROC) curve,[35] the *F*-measure, the accuracy, and the Matthew's correlation coefficient (MCC) as follows:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$F\text{-measure} = \frac{2(\text{recall})(\text{precision})}{\text{precision} + \text{recall}}$$

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}$$

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

where TP are true positives, FN, false negatives, FP, false positives, and FN, false negatives. The best similarity threshold is found by the maximum of the *F*-measure curve when the threshold was varied from 0 to 1 with an increment of 0.01.

Parameter Selection and Optimization. To select the most appropriate parameters and their values, a previously reported data set of 1538 pairs of ligand-binding sites[18] was used for benchmarking. The data set comprises 769 pairs of known similar sites (similar protein binding sites cocrystallized with two different ligands) and 769 pairs of sites randomly chosen among known dissimilar sites.[18] The systematic variation of 31 VolSite and Shaper parameters (Supporting Information Table 3) yielded 16 384 different parameter sets and as many lists of 1538 pairs sorted by decreasing similarity score. The binary classification of pairs (similar/dissimilar) allowed the calculation of the area under the ROC curve for all lists. The standard Shaper parameters were thus defined as those maximizing the value of the area under the ROC curve.

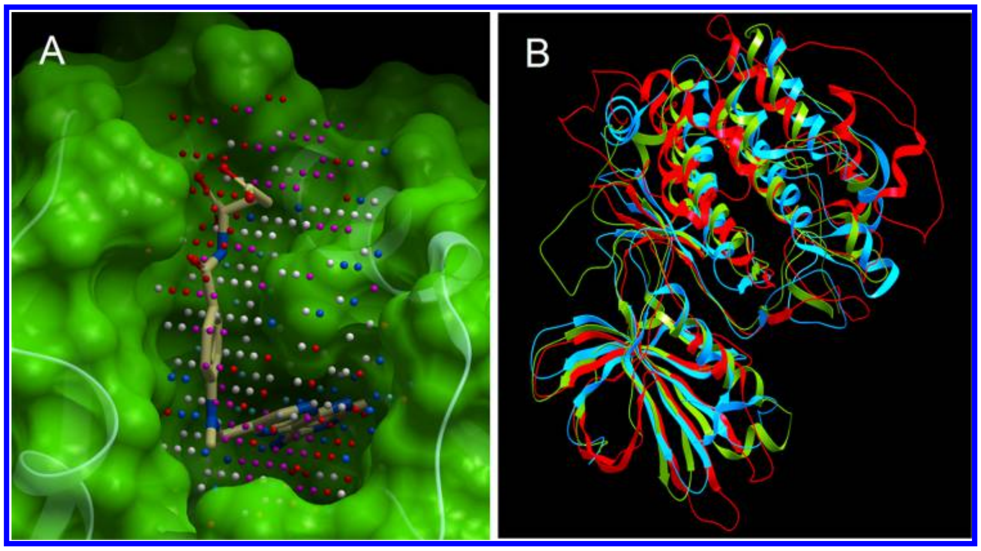


Figure 2. Cavity description and alignment. (A) Detection and pharmacophore annotation of all cavity points in the X-ray structure of *L. bacillus* dihydrofolate reductase (PDB code 4dfr). The cognate ligand (methotrexate, sticks) is shown in the binding site of the protein (green transparent surface). Cavity points are colored by pharmacophoric properties (H-bond acceptor and negative ionizable, red; H-bond donor and positive ionizable, blue; hydrophobe, white; aromatic, cyan; null, magenta). (B) Site points-based alignment of three protein kinases (Pim-1, cyan, PDB entry 3cy3; Rac-β, red, PDB entry 1uv5; Akt-2, green, PDB entry 2jdr).

Virtual Screening of the sc-PDB Database. The similarity of 5952 sc-PDB (v. 2009) binding sites to the inhibitor-binding site in bovine tryspin (PDB entry 1aq7) was computed from the corresponding cavity points with standard VolSite parameters. sc-PDB entries were classified by fold and substrate cleavage specificity in five groups according to the CATH protein structure classification[36] and the CutDB proteolytic event database.[37] The first group is composed of 271 serine endopeptidase entries sharing a trypsin-like fold and a trypsin substrate cleavage specificity. The second group is composed of 17 other serine endopeptidase entries presenting a trypsin-like fold but a substrate cleavage specificity different from that of trypsin. The third group is composed of 11 serine endopeptidase entries with a subtilisin-like fold. The fourth group is composed of 13 entries with a α/β hydrolase fold. The last class is composed by the 5640 remaining sc-PDB entries. All entries were ranked by decreasing RefTversky similarity score, and the rank list used to compute the area under the ROC curve for a binary classification model considering iteratively each of the group as positive instances.

A second virtual screening for similarity to a structurally different ligand-binding site (ATP-competitive inhibitor-binding site in Pim-1 kinase) was undertaken, using three reference sites of the same enzyme cocrystallized by three inhibitors of different sizes (PDB codes 1hys, 3cy3, 1yi4). In this second screen, 9877 sc-PDB entries (v. 2011) were classified in 4 groups according to their E.C. number (protein kinases, other kinases, other ATP/ADP-binding sites, other sc-PDB entries) as previously described.[18] All entries were ranked by decreasing RefTversky similarity score to each of the three references, and the rank lists used to compute the area under the ROC curve for a binary classification model considering iteratively each of the group as positive instances.

Data Set of Promiscuous Protein–Ligand Complexes. A data set of promiscuous ligands was setup by parsing the sc-PDB database (v. 2010) for ligands cocrystallized with at least 2 different proteins, according to their sc-PDB name.[26] The sc-PDB name is derived from the UniProt recommended name, without indications for source organism, cellular location, or maturation state. Remaining ligands were then manually filtered to remove oligomeric compounds (nucleic acids, peptides, oligosaccharides) and lipids. Then, 247 promiscuous pharmacological ligands were finally identified, bound to 401 different proteins in 689 unique sc-PDB entries.

The corresponding protein sequences in fasta format were downloaded from the RCSB PDB.[6] The all-against-all comparison of sequences was performed for all the targets of each promiscuous ligand using default parameters of the Needle routine for global sequence alignment in the EMBOSS package.[38] Only protein chains involved in the ligand binding site were considered. If several comparisons were made for a given pair of proteins, only the highest sequence identity value was retained. A sequence identity above 30% is a good indicator of protein homology.[39] In the present analysis, we consider that an evolutionary link exists between two proteins aligned over more than 100 residues with a sequence identity above 25%.

The structures of complexes in PDB format were downloaded from the RCSB PDB. The all-against-all comparison of structures was performed for all the targets of each promiscuous ligand using default parameters of the CE program.[40] Only protein chains involved in the ligand binding site were considered. If several comparisons were made for a given pair of proteins, only the result with the highest Z-score was retained. A Z-score value lower than 3.7 indicates that the similarity is of low significance. A Z-score value higher than 4.5 denotes conservation of the overall fold. In the 3.7−4.0 range, CE Z-score values define a twilight zone. In the present analysis, we consider that two protein chains aligned with a Z-score higher than 4 share a similar fold.

The similarity of 1070 binding site pairs sharing the same ligand was estimated with three different tools (SiteAlign,[41] FuzCav,[18] Shaper) using default parameters of each program, and previously defined similarity thresholds[18,41] (SiteAlign: d1 < 0.6 and d2 < 0.2; FuzCav: score > 0.16; Shaper: score > 0.35)

All-against-All Comparaison of sc-PDB Druggable Binding Sites. Cavity points were computed with standard VolSite

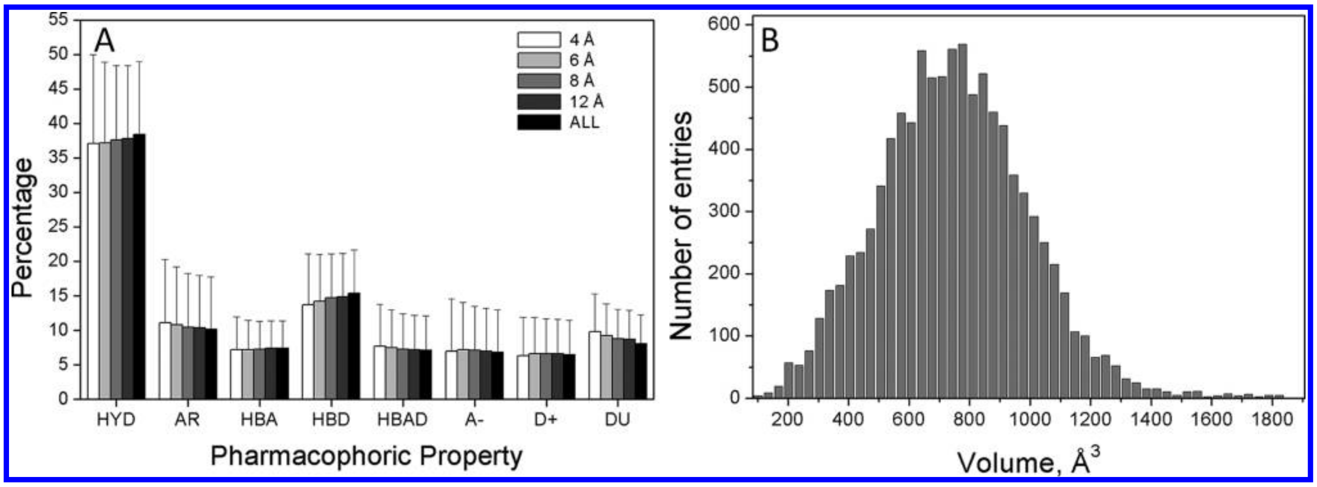


Figure 3. Physicochemical properties of VolSite cavities. (A) Distribution of pharmacophoric properties among cavity points truncated at various distances of the bound ligand. (B) Distribution of the overall volume of sc-PDB cavities.

parameters (no truncation) for the 9877 binding sites of the current sc-PDB database (v. 2011) and further compared with Shaper using a RefTversky metric to generate a full similarity matrix out of which 300 000 values were randomly chosen to select a statistically relevant sample for further analysis.

Classification of GPCR X-ray Structures. A set of 30 X-ray structures of G protein-coupled receptors cocrystallized with low molecular weight ligands was retrieved from the RCSB PDB. This collection comprises 14 different receptors (adenosine A2a, β1 and β2 adrenergic receptors, chemokine CXCR4, dopamine D3, histamine H1, muscarinic M2, muscarinic M3, delta opiate, kappa opiate, mu opiate, nociceptin, rhodopsin, and sphingosine-1 phosphate S1P1) sharing the same fold and cocrystallized with ligands exhibiting different functional effects (full agonists, inverse agonists, neutral antagonists; Supporting Information Table 4). For each entry, all molecules were removed with the exception of protein and pharmacological ligand whose coordinates were separately saved in mol2 file format. Cavities around the bound ligand were computed with default VolSite settings. All pairwise cavity comparisons were done with default Shaper settings and similarities expressed by the RefTversky score. The full similarity table was filtered to remove pairs with a similarity below 0.35 and then imported into Cytoscape v2.8.2[42] to depict a network rendered with a force directed layout using similarity-dependent edge lengths.

■ RESULTS AND DISCUSSION

Binding Site Description. When applied to a typical protein–ligand cavity (e.g., dihydrofolate reductase), site points defined by VolSite encompass the bound ligand but also ligand-unexplored regions of the pocket and stops at soon as accessibility is too high (Figure 2A). The pharmacophoric mapping of cavity points is in overall agreement with the known inhibitor binding mode, cavity, and ligand pharmacophoric features matching well (Figure 2A). Applying the VolSite algorithm to the entire sc-PDB data set (9877 entries) shows that hydrophobic points are the most frequent, whatever the ligand-binding site definition (Figure 3). Interestingly, the respective proportions of pharmacophoric types are independent of the binding site definition (Figure 3A). The distribution of entire cavity volumes, computed over all sc-PDB binding sites is centered at the value of 735 Å^3 (Figure 3B), inferior to that of 930 Å^3 previously reported for a much smaller subset of 99 drug-binding sites.[43]

We next confirmed that aligning cavity site points indeed leads to a correct alignment of the corresponding protein 3D structures. We chose as a prototypical example, the ATP-binding site of three serine/threonine protein kinases: Pim-1, Rac-β, Akt-2. Their binding site-based alignment is problematic due to the protein flexibility (diverse conformations of the loop connecting the N- and C-terminal domains). Moreover, the presence of a few additional charged residues in one of the proteins (Akt-2) significantly modifies the pharmacophoric description of the sites and was shown to cause the failure of alignment-free comparisons.[18] As illustrated in Figure 2B, Shaper proposes a very reliable alignment of the three structures with root-mean-square deviations to the Pim-1 structure below 1.5 Å (Cα atoms of the catalytic site only).

Many tools are available to detect protein cavities from 3D atomic coordinates.[44] The herein presented approach should not be considered as a cavity detection utility since it relies on user-defined atomic coordinates (that of the bound ligand) and does not scan the entire target surface. VolSite uses a standard grid-based cavity detection algorithm similar in spirit to the recently published SiteMap method[29] and presents the advantage of nicely delimiting pocket boundaries by projecting exit vectors from grid points and counting those intersecting protein atoms. Importantly, the computation of the cavity volume requires a reliable way of detecting its boundaries, notably at the interface to solvent. We herein apply a "shrink-wrap" protocol by generating 120 uniformly spaced rays from accessible site points and retaining only those which are buried enough. Visual inspection of many cavities suggests a threshold of 40 for the $N_{\rm ri}$ parameter (see Methods).

VolSite is thus robust in detecting numerous cavity shapes (grooves, clefts, needles). Our algorithm presents however some peculiarities with respect to existing methods. First, Shaper just focuses at the neighborhood of protein-bound ligands in the sc-PDB archive of druggable binding sites, and not at all possible binding cavities. Known pocket occupancy by druggable ligands is a restrictive but safe pocket selection filter. The method could be applied to de novo cavity detection, but we wanted here to restrict our analysis to a well-defined repertoire of binding pockets with strong physicochemical and biological annotations. Second, site points are labeled with pharmacophoric properties complementary to that of the closest protein atom. This labeling procedure enables a chemically relevant description of the pocket as a 3D-lattice of pharmacophoric features characteristic of potential ligands of this pocket. Since the default grid resolution (1.5 Å) is close to common bond lengths in organic compounds, pseudoatoms describing the protein cavity should feature true ligand atoms of these pockets. Therefore, typical ligand-based alignment methods may also be applied to the cavity points. Third, we store for each target 5 ligand-binding sites of increasing size. Hence, a ligand-binding site (immediate vicinity of the bound ligand) may or may not correspond to the entire protein cavity regarding their respective size and ligand buriedness. The method can thus be applied to compare binding sites (ligand-dependent object) but also to assess the structural druggability of the corresponding cavity (ligand-independent object).

Prediction of Structural Druggability. Predicting the druggability of a given target from its three-dimensional structure is an intense field of research in order to reduce attrition rates in pharmaceutical discovery.[45] As druggability is by far more complex than the simple propensity of a particular protein cavity to accommodate high-affinity drug-like compounds, other terms like "ligandability"[45] or "bindability"[46] have recently been proposed since they better capture target property ranges (cavity volume, polarity, and buriedness) known to be important for druggable targets.[28,29,46−48] Since those important properties are theoretically encoded in the herein-described cavity points, we investigated whether the present cavity descriptors might be suitable for predicting the druggability of cavities from their 3D structures. A recently described training set (NRDLD) of 76 cavities (48 druggable, 28 undruggable) was retrieved from literature,[25] and the distribution of site-point properties was given as input for a support vector machine (SVM) classifier. The best 5-fold cross-validated classification model achieves an accuracy of 0.80 and a Matthew's correlation coefficient (MCC) of 0.60. It was further challenged to predict the druggability of 37 novel cavities (23 druggable, 14 undruggable) still from the NRDLD data set (Table 2). Our SVM classifier exhibits a significantly better performance than two state-of-the-art druggability prediction methods (SiteMap, Fpocket) and an accuracy similar to that of DrugPred,[25] one of the best methods reported up to now.

Table 2. Accuracy of 4 Computational Methods in Predicting the Structural Druggability of 37 Cavities (23 Druggable, 14 Undruggable) of Known X-ray Structure[25]

method	VolSite[a]	SiteMap[b]	Fpocket[c]	DrugPred[d]
accuracy	0.89	0.65	0.73	0.89
MCC	0.77	0.24	0.39	0.77

[a] Druggable if score > 0. [b] Druggable if SiteScore > 0.8.[29] [c] Druggable if DG score > 0.50.[28] [d] Druggable if DrugPred score > 0.50.[25]

The current SVM model mispredicts only two druggable proteins of the data set (DNA gyrase B, thymidine phosphorylase) as undruggable (Table 3). DNA gyrase is a clear false negative since our approach also failed in predicting other inhibitor-bound DNA gyrase PDB entries (e.g., 3ttz, 3g75) as druggable. We suspected that the main reason lies in the open and polar binding site of this enzyme, despite a deeply buried subsite. All methods used herein failed in predicting human thymidine phosphorylase as druggable, and this protein should probably be defined as weakly or not druggable at all, as recently suggested.[25] Conversely, only two false positives (glutamate racemase, dialkyglycine dicarboxylase) were observed out of the 14 nondruggable entries of the test set (Table 3). This is less than Fpocket (five false positives) and far less than SiteMap which tends to mispredict almost all nondruggable entries.

The good performance of our druggability prediction model can be explained by an extended data set of druggable and undruggable cavities,[25] and by the use of a machine learning algorithm as predictor. Training a support vector machine on global and local pocket descriptors has also been reported to yield excellent results (90% of correct classification), although on a different training set.[49] A direct comparison of all druggability prediction methods is however difficult since they usually rely on different principles to define pocket boundaries (ligand-based method[25,29] or de novo pocket detection[49,50]) and are applied to different training and test sets. If initiatives to harmonize data sets of druggable and undruggable sites should be acknowledged,[25,28] a uniform definition of binding pockets (e.g., list of cavity lining residues including or not accessory molecules like ions, cofactors) would be desirable. Concluding, we can safely estimate that our approach is fast (ca. 10 s for cavity detection and druggability prediction) and very competitive with the yet best available methods.[25,49]

Determining a Robust Similarity Threshold for Parwise Comparison of Binding Sites. To determine a reliable metric and similarity threshold for distinguishing similar from dissimilar binding sites, we measured the pairwise similarity of 1538 pairs of protein−ligand binding sites[18] split in two categories, 769 pairs of known similar and 769 pairs of known dissimilar sites. Among the 16 384 possible classifications obtained by systematically varying 31 VolSite and Shaper parameters (Supporting Information Table 3), the best classification was obtained using ligand-binding sites truncated at 6 Å and discarding "null" pharmacophoric points from matching rules. The corresponding binary classification model gives an area under the ROC curve of 0.983 and nicely segregates similar from dissimilar pairs (Figure 4A). This result is of course to be expected since pairs of similar/dissimilar binding sites were chosen on pupose. However, it enables the automated determination of the optimal similarity threshold, found for a RefTversky similarity value of 0.35 that gives excellent F-measure, recall and precision values of 0.932 (Figure 4B). Alternatively, a more conservative cutoff of 0.44 could also be chosen, enabling a perfect prediction of all true positives (100% precision). To ascertain that the two above-mentioned thresholds are not data set-dependent, we generated a full similarity matrix from all current sc-PDB entries and examined the distribution of similarity scores from two randomly chosen subsets of 300 000 pairs (Figure 4C). First, the distribution of scores was identical (according to the Kruskall−Wallis test) whatever the sample chosen, suggesting that the selected number of 300 000 pairs was statistically significant. According to the Kolmogorov−Smirnov test, it follows a generalized extreme value distribution (test statistic D = 0.03281, P-value = 0.64132, α = 0.05) with a probability density function of the type:

$$f(x) = \frac{1}{\sigma}\exp(-(-1 + kz)^{-1/k})(1 + kz)^{-1-k/z}$$

with $k = -0.24296$; $\sigma = 0.05309$ (standard deviation); $\mu = 0.24827$ (mean value); $Z = (x - \mu)/\sigma$. The significance level p of the detected similarity represents the probability of obtaining

Table 3. Predicted Druggability Values for a Test Set of 37 Entries[a]

		method			
PDB entry	name	DrugPred[b]	SiteMap[c]	Fpocket[d]	VolSite[e]
druggable					
1e66	acetylcholinesterase	0.81	1.14	0.75	0.80
1fk9	HIV reverse transcriptase	0.79	1.27	0.84	1.07
1kzn	DNA gyrase	0.81	1.04	0.75	−0.48
1lox	15-lipoxygenase	1.15	1.13	0.76	1.01
1oq5	carbonic anhydrase II	0.77	1.00	0.10	0.48
1owe	urokinase plasminogen activator	0.40	0.93	0.25	0.28
1pmn	C-Jun kinase	0.93	1.09	0.88	1.25
1pwm	aldose reductase	0.94	0.97	0.86	0.67
1q41	glycogen synthase kinase 3	0.55	1.09	0.46	1.15
1r55	ADAM33	0.69	0.89	0.08	0.07
1sqn	progesterone receptor	1.11	1.28	0.95	1.99
1t46	c-Kit kinase	1.17	1.12	0.84	1.37
1unl	cyclin-dependent kinase 5	0.56	1.06	0.12	0.47
1uou	thymidine phosphorylase	0.28	nd[f]	0.40	−0.65
1xoz	phosphodiesterase 5A	1.14	1.10	0.81	1.06
2aa2	mineralocorticoid receptor	1.02	1.24	0.92	1.16
2cl5	catechol-O-methyltransferase	0.82	1.19	0.70	1.48
2i1m	FMS kinase	0.74	1.10	0.82	0.70
3b68	androgen receptor	1.13	1.29	0.95	2.09
3etr	xanthine oxidase	0.85	1.13	0.67	1.01
3f0r	histone deacetylase 8	0.89	1.12	0.59	0.96
3f1q	dihydroorotate dihydrogenase	1.15	1.20	0.90	1.73
3ia4	dihydrofolate reductase	0.79	1.07	0.65	0.63
undruggable					
1ajs	aspartate aminotransferase	0.49	1.14	0.60	−0.60
1b74	glutamate racemase	0.41	1.05	0.56	0.60
1bls	Beta-lactamase	0.34	1.04	0.26	−0.01
1bmq	interleukin-1beta-converting enzyme	0.38	0.79	0.01	−0.05
1ec9	D-glucarate dehydratase	−0.31	1.03	0.15	−1.28
1g98	phosphoglucose isomerase	0.09	1.14	0.03	−0.57
1kc7	pyruvate phosphate dikinase	0.01	0.86	0.01	−1.61
1m0n	dialkylgycine decarboxylase	0.50	0.98	0.76	0.46
1mai	phospholipase C	0.09	0.90	0.03	−1.93
1od8	xylanase	0.06	0.79	0.05	−1.01
1px4	beta-galactosidase	0.55	1.06	0.13	−0.03
1v16	α-keto acid dehydrogenase	0.41	1.08	0.02	−0.57
1wvc	CDP-D-glucose synthase	0.65	1.03	0.67	−0.52
3jdw	L-arginine:glycine amidinotransferase	0.17	1.06	0.09	−0.18

[a] False predictions are underlined. [b] Druggable if DrugPredScore > 0.50. [c] Druggable if SiteScore > 0.8. [d] Druggable if DGscore > 0.5. [e] Druggable if score > 0. [f] No cavity detected.

the same or higher similarity score $Z > z$ by chance is the following:

$$p(Z > z) = 1 - \exp(-e^{-1.282z-0.5772})$$

Plotting the probability p-value against the raw similarity scores from the random distribution (Figure 4D) indicates that the two thresholds previously used (0.35 and 0.44) correspond to very low probabilities of 0.05 and 0.005, respectively.

By opposition to the Tanimoto similarity index, The Tversky index used by Shaper presents the advantage to delineate similarity among ligand-binding sites of different sizes (e.g., monomer vs homodimer-lining sites, site cocrystallized with two ligands of very different sizes). In this case, only the Tversky index detects similarity between both entries when the smallest of both sites is used as a reference.

Influence of Various Parameters (Grid Resolution and Orientation, Atomic Coordinates) on Similarity Measurements. Since the ligand-binding site is discretized at regularly spaced grid points, we first investigated whether changing either the grid resolution or the grid center alters the description and comparison of ligand-binding sites. When applied to the classification of the above-described pairs of similar and dissimilar ligand-binding sites, obtained results slightly varies with the grid resolution (Table 4). As to be expected, a tighter grid spacing (1.0 Å) enabling a better distinction of the two sets of binding sites (AUC = 0.991) whereas a smoother resolution (2.0 Å) deteriorates the binary classification (AUC = 0.947). Interestingly, the optimal similarity threshold (RefTversky index) and the classification accuracy (F-measure) slightly decrease when the grid spacing increases (Table 4). However, increasing the grid resolution significantly increases the CPU time necessary for defining the cavity points (Table 4). The intermediate resolution of 1.5 Å was therefore chosen as default since it yields the best

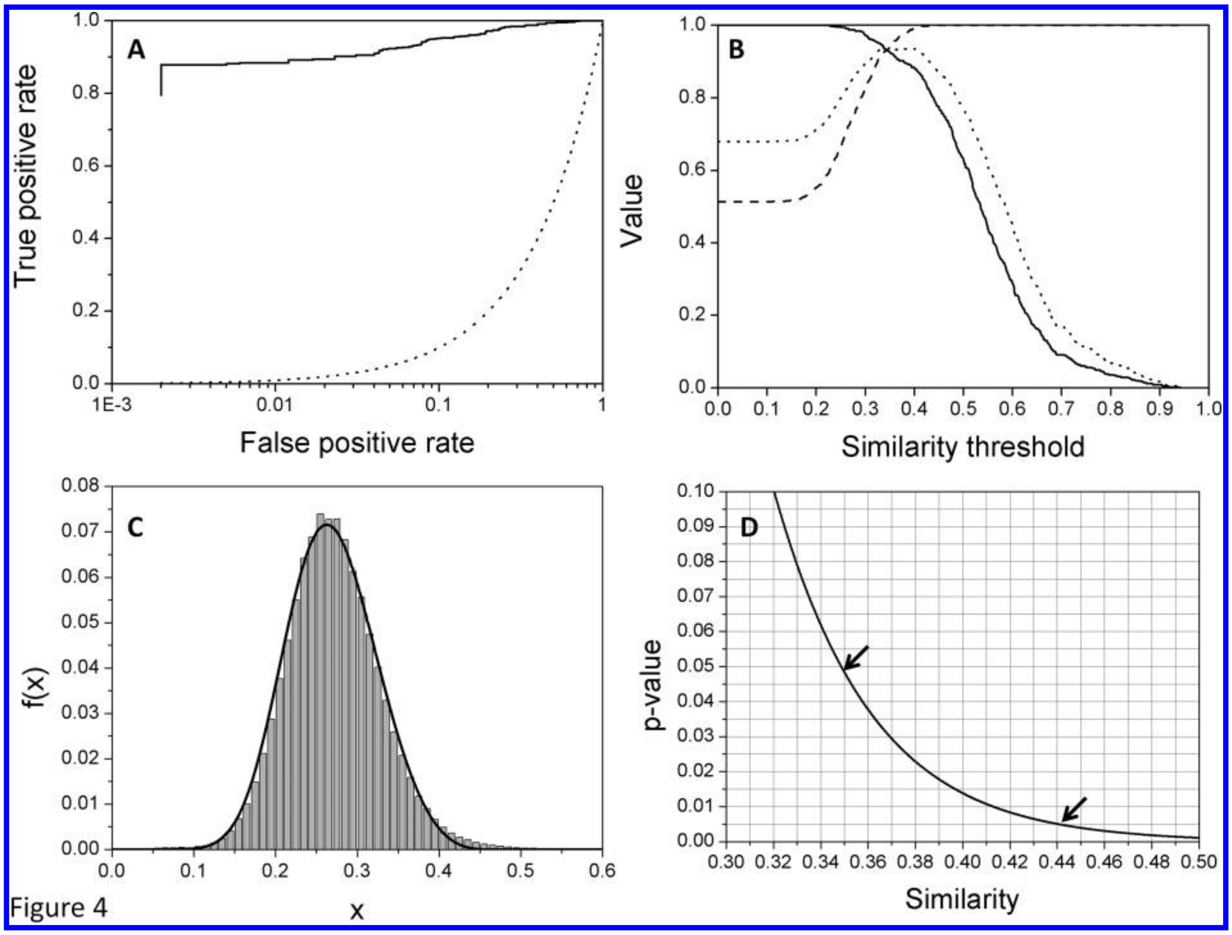


Figure 4. Statistical evaluation of Shaper similarity scores. (A) ROC plot (solid line) obtained by sorting 1538 sc-PDB pairs of binding sites by decreasing RefTversky similarity values. True positives (n = 769) are pairs of similar binding sites predicted similar whereas true negatives (n = 769) are pairs of dissimilar sites predicted dissimilar. Accuracy of random picking is represented by a dotted line. (B) Variation of statistical parameters (recall, solid line; precision, dashed line; F-measure, dotted line) of a binary classification model (similar/dissimilar) for increasing Tversky similarity score thresholds. (C) Distribution of Shaper similarity scores for a randomly chosen population of 300 000 scores retrieved from the all-against-all comparison of 9 877 sc-PDB binding sites (97 555 129 comparisons in total). The fit (bold line) to a generalized extreme value distribution represents the ideal probability density $f(x)$ for a similarity value of x. (D) Decay of the p-value (probability to get by chance a similarity score $Z > x$) as a function of the observed similarity value.

Table 4. Influence of the Grid Resolution on the Correct Classification of 1538 Pairs of Ligand-Binding Sites[a]

	grid resolution, Å		
	1.0	1.5	2.0
AUC[b]	0.991	0.983	0.947
threshold[c]	0.422	0.350	0.272
F-measure[d]	0.956	0.932	0.870
CPU[e]	46.29	8.18	2.36

[a] 769 similar and 769 dissimilar pairs; see Methods for data set description. [b] Area under the ROC curve for classifying the set of 769 similar and 769 dissimilar pairs, according to the measured similarity value (RefTversky metric). [c] Optimal similarity threshold (RefTversky metric) for discriminating 769 similar from 769 dissimilar pairs. [d] F-measure of the classification at the optimal similarity threshold. [e] CPU time in seconds (3.40 GHz Intel Pentium D processor with 2 Go RAM) for computing pharmacophoric site points for a ligand-binding site (PDB entry 1bji, ligand HET code: DPC) of 735 Å^3 (average volume among 9877 sc-PDB binding sites)

compromise between speed and accuracy. This parameter is however user-tunable in Volsite.

In a second computational experiment, the center of the grid encompassing the ligand-binding site was systematically translated from the origin (center of mass of the bound ligand) by 0.1 Å increments up to 3.0 Å (1.5 Å in each direction from the origin) along the three main axis, therefore leading to 26 999 (30^3 −1) alternative grid lattices for a unique binding site. The pairwise similarity of all these representations to the native one is far above the previously defined similarity threshold of 0.350 (Figure 5A), therefore demonstrating that both VolSite and Shaper are insensitive to overall translations of the grid lattice up to 1.5 Å (the default grid resolution). It should be noted that the druggability score predicted by the herein presented SVM model is also relatively insensitive to grid translations (Figure 5B).

In a last control experiment, we investigated how much the Volsite druggability and Shaper similarity scores vary with moderate variations in atomic coordinates of the cavity under investigation. A data set of 14 proteins for which ligand-free and ligand-bound X-ray structures are available was investigated for Shaper pairwise comparisons (Figure 5C). In 12 out of 14 cases, despite significant changes occurring at the binding site (up to 4.6 Å rmsd on binding site heavy atoms), the computed similarity was above similarity threshold of 0.35. Likewise,

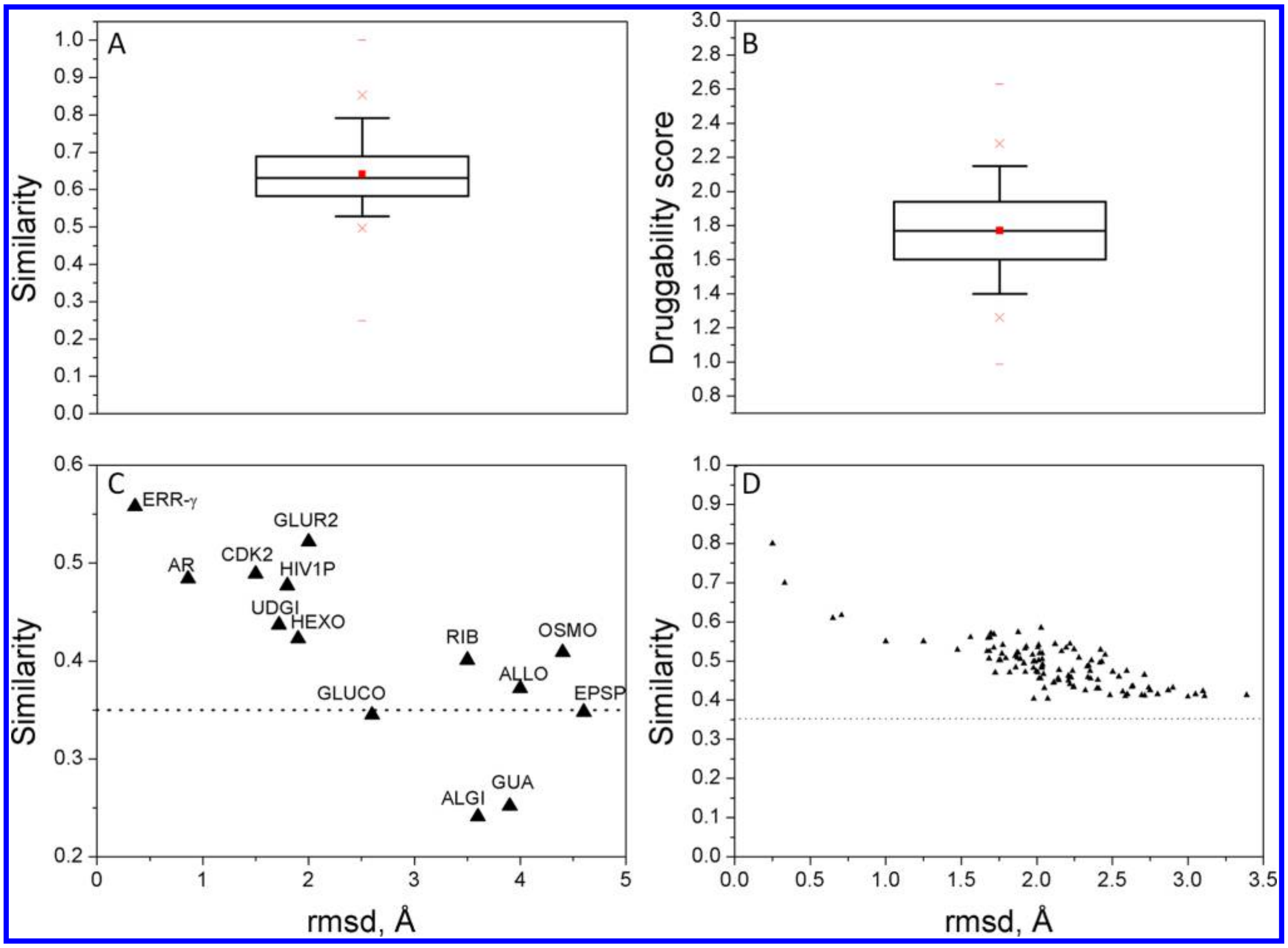


Figure 5. Sensitivity of Volsite and Shaper to input data. (A) Distribution of the pairwise similarity for an entire ligand-binding site (PDB entry 3k3i) after $30^3 - 1$ systematic translations (increment: 0.1 Å, range 3.0 Å) of the grid center along the three main axes. The box delimit the 25th and 75th percentiles, the whiskers delimit the 5th and 95th percentiles. The median and mean values are indicated by a horizontal line and a filled square in the box. Crosses delimit the 1st and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash. (B) Distribution of the predicted druggability score for the 27 000 representations of the 3k3i ligand-binding site. The box delimits the 25th and 75th percentiles, the whiskers delimit the 5th and 95th percentiles. The median and mean values are indicated by a horizontal line and a filled square in the box. Crosses delimit the 1st and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash. (C) Sensitivity to atomic coordinates: Shaper similarity versus rms deviations of the holo from the apo structure (active site only) of 14 targets: uracil DNA−glycosylase inhibitor (UDGI, 36 residues, pdb identifier 1udi vs 1ugi), cell division protein kinase 2 (CDK2, 36 residues, 1dm2 vs 2jgz), HIV-1 protease (HIV-1p, 52 residues, 1qbs vs 1hhp), estrogen-related receptor gamma (ERRγ, 27 residues, 2zkc vs 2zbs), aldose reductase (AR, 24 residues, 1ads vs 2nvd), glutamate receptor subunit 2 (GLUR2, 22 residues, 1ftm vs 1fto), DNA-β-glucosyltransferase (GLUCO, 28 residues, 1jg6 vs 1jej), D-allose binding protein (ALLO, 21 residues, 1rpj vs 1gud), D-ribose binding protein (RIB, 20 residues, 2dri vs 1urp), 5-enolpyruvylshikimate-3-phosphate synthase (EPSP, 33 residues, 1rf4 vs 1rf5), Osmo-protection protein (OSMO, 19 residues, 1sw2 vs 1sw5), guanylate kinase (GUA, 24 residues, 1ex7 vs 1ex6), hexokinase (HEX, 25 residues, 2e2o vs 2e2n), and alginate-binding protein (ALGI, 19 residues, 1y3n vs 1y3q). The active site was defined from the holostructure by considering any amino acid with at least one heavy atom present in a 6.5 Å-radius sphere centered on the center of mass of the bound ligand. (D) Sensitivity to atomic coordinates: Shaper similarity versus rms deviations of active site heavy atoms to the X-ray structure for a 1-ns molecular dynamics (MD) simulation of the water solvated cyclin-dependent kinase type 2 (PDB entry 1dm2). Rmsd for 100 MD snapshots are displayed. The horizontal dotted line represents the similarity threshold (0.35) used throughout this study to discriminate similar from dissimilar protein−ligand binding sites.

molecular dynamics snapshots of a typical druggable cavity (ATP-binding site of cyclin dependent kinase type 2) were still considered similar enough (RefTversky similarity >0.35) to the native crystal structure (Figure 5D). Importantly, the VolSite druggability score (1.15 ± 0.21) did not vary either all along the MD trajectory. Both tools are therefore relatively insensitive to moderate variation in protein coordinates up to 3.0−3.5 Å rmsd.

Virtual Screening for Similarity to a Known Cavity. Predicting the function of a protein from the similarity of its cavities to functionally annotated binding sites may help the annotation of novel genomic structures. To address this issue, we measured the pairwise similarity between the inhibitor-binding site in bovin trypsin (PDB entry 1aq7) and 5952 sc-PDB binding sites. The canonical inhibitor binding site in trypsin was chosen here as a template for two main reasons: (i) trypsin belongs to the family of serine endopeptidases which presents the interest to share the same catalytic activity, a common catalytic triad, but different 3-D folds (trypsin, subtilisin, $\alpha\beta$ hydrolase); (ii) the same query has already been conducted by us with two other binding site similarity search programs (SiteAlign[41] and FuzCav[18]) thus enabling to compare the herein presented approach to two state-of-the-art methods.

When comparing the close proximity of bound ligands (cavities truncated at 4 Å), Shaper clearly discriminate endopeptidases from all other entries (Table 5). As expected, the highest ROC score is observed for the group of entries (group 1) sharing the fold and substrate specificity with the trypsin query. However, the second (trypsin fold, other cleavage specificity), third (subtilisin fold), and fourth (*αβ* hydrolase fold) groups are also statistically enriched in binding sites found similar to that of bovine trypsin. Enlarging the binding site definition (truncations at 6, 8, 12 Å; no truncation at all) changes the scope of the search to retrieve binding sites which are no more locally but globally similar to the query. Is is therefore no surprise that the ROC score increases for the closest group (group 1) and decreases for groups exhibiting only local similarity at the catalytic side (groups 3 and 4).

Table 5. Area under the ROC Curve for Classification Models of 5952 sc-PDB Entries According to Shaper Similarity to the 1aq7 PDB Entry (Bovine Trypsin in Complex with Inhibitor Aeruginosin 98-B)

	ligand-binding site truncation				
group	4 Å	6 Å	8 Å	12 Å	none
1[a]	0.868	0.914	0.962	0.964	0.965
2[b]	0.771	0.684	0.764	0.763	0.763
3[c]	0.779	0.667	0.634	0.630	0.634
4[d]	0.684	0.618	0.593	0.606	0.606
5[e]	0.151	0.109	0.063	0.062	0.062

[a] Serine proteases with trypsin fold and trypsin substrate specificity. [b] Serine proteases with trypsin fold and other substrate specificity. [c] Serine protease with subtilisin fold. [d] Serine protease with *αβ* hydrolase fold. [e] Any other sc-PDB entry.

A second screen against a reference cavity from a structurally different class (protein kinase Pim-1) confirmed these results (Table 6). In this example, we investigated the effect of changing the reference cavity (same binding site but cocrystallized with three chemically different inhibitors) on the screening results. As noted for the previous trypsin screening, a good classification of protein kinases from other protein classes is obtained, whatever the binding site truncation method and the reference cavity (Table 6). As previously observed,[18,41,51] we confirm that ATP-binding sites in protein kinases dot not ressemble neither ATP-binding sites in other kinases, nor generic ATP/ADP-binding sites in general (Table 6).

When compared to previously reported benchmarks, Shaper was shown to perform as well as SiteAlign[41] and FuzCav[18] in discriminating the different endopeptidase groups from decoy binding sites (Table 7). As initially requested, the method is therefore able to combine the advantage of an alignment-dependent method (visualization and interpretation of matched structures) with the speed of an alignment-free method (ca. 10 comparisons/sec on an Intel Xeon E5504 processor)

Table 7. Comparison of Three Binding Site Comparison Methods, Expressed by Area under the ROC Curve for Binary Classification Models of 5 882 sc-PDB Entries According to Shaper Similarity to the 1aq7 PDB Entry (Bovine Trypsin in Complex with Inhibitor Aeruginosin 98-B)

group	SiteAlign[a]	FuzCav[b]	Shaper[c]
1[d]	0.939	0.906	0.914
2[e]	0.604	0.780	0.684
3[f]	0.462	0.649	0.779
4[g]	0.486	0.662	0.618
5[h]	0.109	0.114	0.109

[a] 3D alignment-based, slow (2 comparisons/min).[41] [b] 3D alignment-free, ultrafast (1000 comparison/s).[18] [c] 3D alignment based, fast (10 comparisons/s). [d] Serine proteases with trypsin fold and trypsin substrate specificity. [e] Serine proteases with trypsin fold and other substrate specificity. [f] Serine protease with substilisin fold. [g] Serine protease with *αβ* hydrolase fold. [h] Any other sc-PDB entry.

Binding Site Similarity Detection As a Function of Target Sequence and Structure Conservation. The previous application illustrated the ability of Shaper to detect local and global binding site similarities among a class of proteins sharing the same catalytic activity. Detecting remote ligand-binding site similarity among unrelated proteins is undoubtley more difficult. We specifically designed a data set of promiscuous ligand-binding sites to address this issue and challenged our approach for difficult cases of shared ligand binding irrespective of sequence and structure conservation. We have identified 1 070 pairs of protein−ligand complexes in which the same ligand has been cocrystallized with different proteins. By computing both sequence and structure conservation of the corresponding targets (see Methods), we have classified these pairs in three categories: (i) easy, which means that the ligand is shared by proteins exhibiting both sequence (sequence identity >25%) and structure conservation (CE *Z* score >4); (ii) medium, the ligand being shared by proteins showing structure but not sequence conservation; (iii) difficult, the ligang being shared by proteins exhibiting neither sequence nor structure conservation.

Comparison of Shaper with respect to SiteAlign and FuzCav in detecting binding site similarity across these 1070 pairs show

Table 6. Area under the ROC Curve for Classification Models of 10 435 sc-PDB Entries According to Shaper Similarity to the ATP-Binding Site of Pim-1 Kinase[a]

	ligand-binding site truncaction				
group	4 Å	6 Å	8 Å	12 Å	none
1[b]	0.883 ± 0.007	0.888 ± 0.008	0.892 ± 0.007	0.886 ± 0.007	0.886 ± 0.010
2[c]	0.428 ± 0.008	0.430 ± 0.008	0.422 ± 0.004	0.419 ± 0.006	0.430 ± 0.003
3[d]	0.460 ± 0.012	0.470 ± 0.003	0.473 ± 0.010	0.474 ± 0.008	0.440 ± 0.013
4[e]	0.269 ± 0.002	0.263 ± 0.004	0.261 ± 0.002	0.265 ± 0.004	0.266 ± 0.009

[a] Values are means and standard deviations for three independent screens against three Pim-1 kinase binding sites (1hys, 3cy3, 1yi4). [b] Protein kinases (*n* = 1138). [c] Other kinases (*n* = 294). [d] Other ATP/ADP-binding sites (*n* = 423). [e] Other sc-PDB entries (*n* = 8580).

three clear trends (Figure 6): (i) in easy cases, all programs perform very well and recover ca. 85% of the pairs as truly similar; (ii) in cases of medium difficulty, the success rate drops drastically for SiteAlign (52%) and FuzCav (36%) but not for Shaper (76%); (iii) in really difficult cases, only Shaper provides a good performance (46% of pairs recovered) whereas SiteAlign and FuzCav fails in 90% of the cases (Figure 6). Out of the three tools, Shaper is clearly the one that recovers the highest proportion of similar binding sites (69%). The noticeable advantage of Shaper to detect binding site similarity in cases of medium and high difficulty is explained by a better description of binding site attributes, notably the shape of the cavity that is a known important prerequisite for ligand binding. SiteAlign and FuzCav both encode the same pharmacophoric properties than Shaper but only at the Cα atom of the cavity-lining residue. Conversely, Shaper places multiple pharmacophoric points at the vicinity of all ligand-accessible binding site atoms and therefore gives more importance to ligand-accessible than to buried protein atoms.

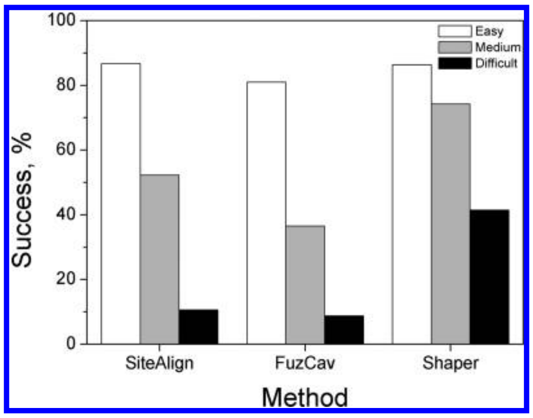


Figure 6. Comparative ability of three binding site comparison programs (SiteAlign, FuzCav, Shaper) in detecting similarity for ligand-binding sites from different proteins but sharing the same ligand. Success is defined when the pairwise similarity is above each program-specific similarity threshold[18,41] (SiteAlign: d1 <0.6 and d2 <0.2; FuzCav: score >0.16; Shaper: score >0.35). Binding sites are classified in three categories: easy (white bars, conserved sequence and structure), medium (gray bars, different sequence but conserved structure), and difficult (dark bars, different sequence and structure).

Network of Binding Sites for a Protein Family. The recent X-ray structure determination of several GPCRs in complex with noncovalent drug-like compounds[7] offered us the opportunity to measure pairwise similarities of 30 ligand-binding sites from 14 different receptors. We particularly paid attention to (i) the comparison of different entries of the same receptor cocrystallized with ligands exhibiting different functional effects (agonist, inverse agonist, antagonist); (ii) the influence of the binding site definition (immediate vicinity of the ligand, full cavity) on the obtained networks.

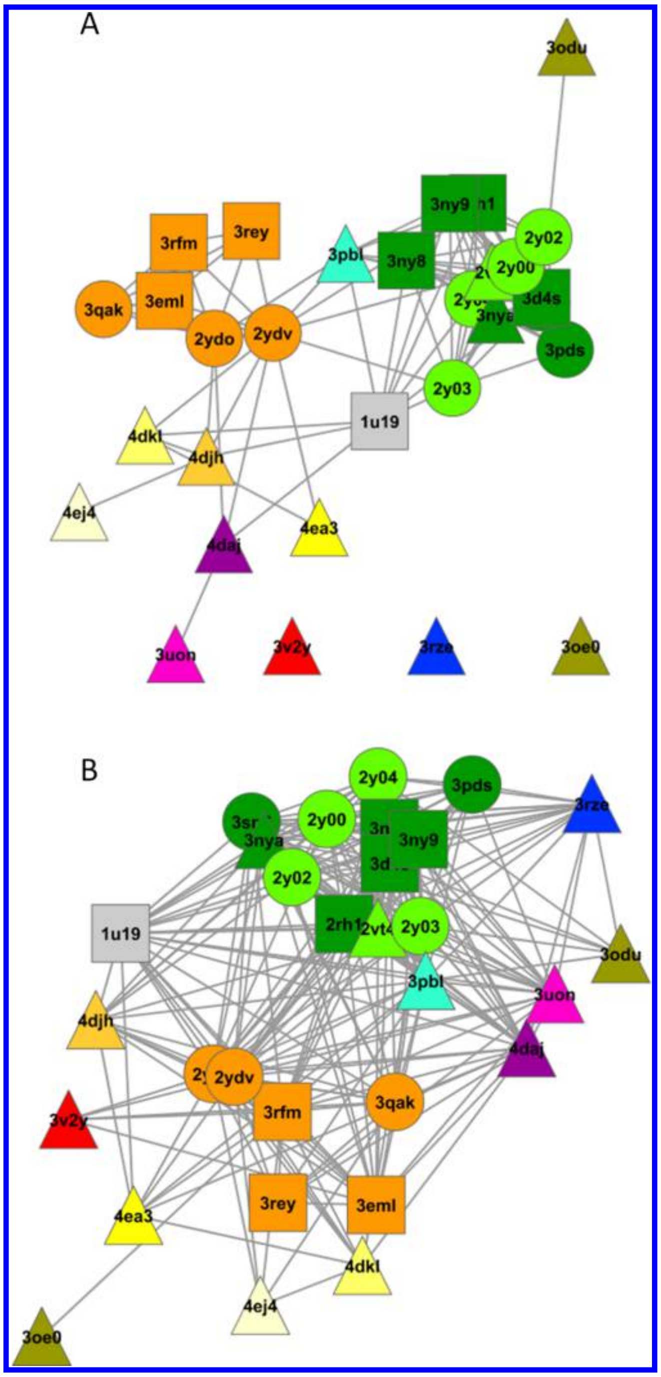


Figure 7. Network of G protein-coupled receptor binding sites. (A) Network of full cavities. (B) Network of binding sites truncated at 4 Å of the bound ligand. Nodes are colored by receptor name (Adenosine A2a, orange; β1 adrenergic, light green; β2-adrenergic, dark green; chemokine CXCR3, olive; Dopamine D2, cyan; Histamine H1, blue; Muscarinic M2, light purple; Muscarinic M3, purple; delta opiate, light yellow; kappa opiate, light orange; mu opiate, yellow; nociceptin, bright yellow; rhodopsin, gray; sphingolipid S1P1, red) and shaped according the functional effect of the bound ligand (agonist, circle; inverse agonist, rectangle; antagonist, triangle).

The network obtained from full cavities is very homogeneous and separate very well entries by receptor name (Figure 7A), with the exception of the very similar β1 and β2 adrenergic receptors which are grouped together. The dopamine D3 receptor cavity links the adenosine A2a receptor group to the adrenergic receptor entries. Three receptors (S1P1, Histamine H1, CXCR4) exhibit unique binding site properties and thus are represented as singletons. It is noteworthy that the solvent-exposed peptide-binding site in CXCR4 (3oe0) is not related to the transmembrane nonpeptide binding site of the same receptor (3odu). As to be expected from the fine observation of all X-ray structures,[7] it is not possible to distinguish agonist from antagonist (or inverse agonist) binding sites at the default Shaper resolution. However, it is interesting to notice the much more complex network derived from binding sites truncated at a maximal 4 Å-distance from the bound ligand (Figure 7B). The later network, although still grouping entries by receptor names offers much more edges (244 vs 111 for the previous network) between different receptors, mainly those of biogenic amines (β1 and β2 adrenergic, dopamine D3, histamine H1, muscarinic M2 and M3). This observed difference is in agreement with the recently established evidence that fine receptor selectivity is principally gained from interaction of ligand moieties at the periphery of transmembrane binding sites (notably close to the extracellular loops) and not within the ancestral retinal binding site.[7] The possible definition of binding sites of increasing sizes (from the ligand center of mass) in VolSite permits to delineate the presence or absence of ligand-proximal or more distal relationships, and therefore to estimate whether selective ligands could be designed for receptors from the same family.

CONCLUSIONS

We herewith present two complementary methods for detecting and comparing protein−ligand binding sites. VolSite first describes cavities from the known position of bound ligands and represent the binding site by grid points bearing pharmacophoric properties complementary to that of the nearest protein atom. VolSite has been applied to pockets occupied by known ligands but could be easily used to systematically scan an entire protein surface and rank detected cavities by decreasing ligandability. A further line of improvement lies in the choice for assigning pharmacophoric properties to cavity points. The current protocol uses simple distance criteria (closest protein atom) to match a pharmacophore feature onto site points. A probabilistic approach, taking into account the density of the different protein atom types at the close vicinity of the cavity point, could avoid polarity mismatches (e.g., hydrophobic point in a very polar environment or vice versa) or incorrect assignments arising from local uncertainties in protein atomic coordinates. Beside discribing cavities, VolSite descriptors can be directed read by a machine learning classifier (SVM in the present study) for predicting the ligandability (structural druggability) of the corresponding pocket, with accuracy at least similar to that of the best methods reported yet.

The second tool (Shaper) aligns and measures the similarity of two pockets by approximating site points with Gaussians thus enabling to quickly align two sites by optimizing their volume overlap. Shaper was shown to combine the pace of alignment-free site comparison tools and the accuracy and interpretability of alignment-dependent methods. It can be used for clustering a set of binding sites according to their physicochemical properties as well as screening a collection of binding sites for similarity to a query. Interestingly, Shaper was particularly efficient in detecting binding site similarity in absence of sequence of fold conservation. Both methods are relatively insensitive to variations in atomic coordinates and definition of the grid box.

ASSOCIATED CONTENT

Supporting Information

List of 73 VolSite descriptors used for predicting structural druggability, color force-field to postprocess shape matching in Shaper, list of VolSite and Shaper parameters, and GPCR X-ray structures along with ligand names and their functional effects. This material is available free of charge via the Internet at http://pubs.acs.org

AUTHOR INFORMATION

Corresponding Author

*Phone: +33 3 68 85 42 35. Fax: +33 3 68 85 43 10. E-mail: rognan@unistra.fr.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The Calculation Center of the IN2P3 (CNRS, Villeurbanne, France) and the GENCI (Project x2010075024) are acknowledged for allocation of computing time. We sincerely thank Dr. R. Brenk (University of Dundee, Scotland) for sharing DrugPred data with us.

REFERENCES

(1) Dessailly, B. H.; Nair, R.; Jaroszewski, L.; Fajardo, J. E.; Kouranov, A.; Lee, D.; Fiser, A.; Godzik, A.; Rost, B. Orengo, C. PSI-2: structural genomics to cover protein domain family space. *Structure* **2009**, *17*, 869−881.

(2) Joachimiak, A. High-throughput crystallography for structural genomics. *Curr. Opin. Struct. Biol.* **2009**, *19*, 573−584.

(3) Svergun, D. I. Small-angle X-ray and neutron scattering as a tool for structural systems biology. *Biol. Chem.* **2010**, *391*, 737−743.

(4) Montelione, G. T.; Szyperski, T. Advances in protein NMR provided by the NIGMS Protein Structure Initiative: impact on drug discovery. *Curr. Opin. Drug Discov. Dev.* **2010**, *13*, 335−349.

(5) Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C. A.; Motoshima, H.; Fox, B. A.; Le Trong, I.; Teller, D. C.; Okada, T.; Stenkamp, R. E.; Yamamoto, M.; Miyano, M. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **2000**, *289*, 739−745.

(6) 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.

(7) Congreve, M.; Langmead, C. J.; Mason, J. S.; Marshall, F. H. Progress in structure based drug design for G protein-coupled receptors. *J. Med. Chem.* **2011**, *54*, 4283−4311.

(8) Manglik, A.; Kruse, A. C.; Kobilka, T. S.; Thian, F. S.; Mathiesen, J. M.; Sunahara, R. K.; Pardo, L.; Weis, W. I.; Kobilka, B. K.; Granier, S. Crystal structure of the micro-opioid receptor bound to a morphinan antagonist. *Nature* **2012**, *485*, 400−404.

(9) Nair, R.; Liu, J.; Soong, T. T.; Acton, T. B.; Everett, J. K.; Kouranov, A.; Fiser, A.; Godzik, A.; Jaroszewski, L.; Orengo, C.; Montelione, G. T.; Rost, B. Structural genomics is the largest contributor of novel structural leverage. *J. Struct. Funct. Genomics* **2009**, *10*, 181−191.

(10) Xie, L.; Bourne, P. E. Structure-based systems biology for analyzing off-target binding. *Curr. Opin. Struct. Biol.* **2011**, *21*, 189−199.

(11) Rognan, D. Structure-based approaches to target fishing and ligand profiling. *Mol. Inf.* **2010**, *29*, 176−187.

(12) Rognan, D. Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* **2007**, *152*, 38−52.

(13) Kinnings, S. L.; Liu, N.; Buchmeier, N.; Tonge, P. J.; Xie, L.; Bourne, P. E. Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput. Biol.* **2009**, *5*, e1000423.

(14) Stauch, B.; Hofmann, H.; Perkovic, M.; Weisel, M.; Kopietz, F.; Cichutek, K.; Munk, C.; Schneider, G. Model structure of APOBEC3C reveals a binding pocket modulating ribonucleic acid interaction required for encapsidation. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 12079−12084.

(15) Defranchi, E.; Schalon, C.; Messa, M.; Onofri, F.; Benfenati, F.; Rognan, D. Binding of protein kinase inhibitors to synapsin I inferred from pair-wise binding site similarity measurements. *PLoS One* **2010**, *5*, e12214.

(16) Xie, L.; Wang, J.; Bourne, P. E. In silico elucidation of the molecular mechanism defining the adverse effect of selective estrogen receptor modulators. *PLoS Comput. Biol.* **2007**, *3*, e217.

(17) Kellenberger, E.; Schalon, C.; Rognan, D. How to measure the simialrity between protein ligand-binding sites. *Curr. Comput. Aided Drug. Des.* **2008**, *4*, 209−220.

(18) Weill, N.; Rognan, D. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *J. Chem. Inf. Model.* **2010**, *50*, 123−135.

(19) Yeturu, K.; Chandra, N. PocketMatch: a new algorithm to compare binding sites in protein structures. *BMC Bioinf.* **2008**, *9*, 543.

(20) Yin, S.; Proctor, E. A.; Lugovskoy, A. A.; Dokholyan, N. V. Fast screening of protein surfaces using geometric invariant fingerprints. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 16622−16626.

(21) Das, S.; Kokardekar, A.; Breneman, C. M. Rapid comparison of protein binding site surfaces with property encoded shape distributions. *J. Chem. Inf. Model.* **2009**, *49*, 2863−2872.

(22) Xiong, B.; Wu, J.; Burk, D. L.; Xue, M.; Jiang, H.; Shen, J. BSSF: a fingerprint based ultrafast binding site similarity search and function analysis server. *BMC Bioinf.* **2010**, *11*, 47.

(23) Grant, J. A.; Gallardo, M.; Pickup, B. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17*, 1653−1666.

(24) Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195−207.

(25) Krasowski, A.; Muthas, D.; Sarkar, A.; Schmitt, S.; Brenk, R. rugPred: A Structure-Based Approach To Predict Protein Druggability Developed Using an Extensive Nonredundant Data Set. *J. Chem. Inf. Model.* **2011**, *51*, 2829−2842.

(26) Meslamani, J.; Rognan, D.; Kellenberger, E. sc-PDB: a database for identifying variations and multiplicity of "druggable" binding sites in proteins. *Bioinformatics* **2010**, *24*, 1324−1326.

(27) Joachims, T. *Learning To Classify Text Using Support Vector Machines, Methods, Theory And Algorithms*; Springer-Verlag, LLC: New York, 2002.

(28) Schmidtke, P.; Barril, X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.* **2010**, *53*, 5858−5867.

(29) Halgren, T. A. Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.* **2009**, *49*, 377−389.

(30) *SiteMap*, v. 2.2; Schrodinger Inc: New York, 2011.

(31) *SybylX1.3*; Tripos: St. Louis, MO, 2011.

(32) *OEChem and OEShape toolkit*; OpenEye Scientific Software: Santa Fe, NM, 2011.

(33) Grant, J. A.; Pickup, B. A Gaussian description of molecular shape. *J. Phys. Chem.* **1995**, *99*, 3503−3510.

(34) Nicholls, A.; Grant, J. A. Molecular shape and electrostatics in the encoding of relevant chemical information. *J. Comput. Aided Mol. Des.* **2005**, *19*, 661−686.

(35) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005**, *48*, 2534−2547.

(36) Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. CATH−a hierarchic classification of protein domain structures. *Structure* **1997**, *5*, 1093−1108.

(37) Igarashi, Y.; Eroshkin, A.; Gramatikova, S.; Gramatikoff, K.; Zhang, Y.; Smith, J. W.; Osterman, A. L.; Godzik, A. CutDB: a proteolytic event database. *Nucleic Acids Res.* **2007**, *35*, D546−D549.

(38) Mullan, L. J.; Bleasby, A. J.; Short, EMBOSS User Guide. European Molecular Biology Open Software Suite. *Brief Bioinform* **2002**, *3*, 92−94.

(39) Rost, B. Twilight zone of protein sequence alignments. *Protein Eng.* **1999**, *12*, 85−94.

(40) Shindyalov, I. N.; Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **1998**, *11*, 739−747.

(41) Schalon, C.; Surgand, J. S.; Kellenberger, E.; Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins* **2008**, *71*, 1755−1778.

(42) Cytoscape: An Open Source Platform for Complex Network Analysis and Visualization. http://www.cytoscape.org/.

(43) Nayal, M.; Honig, B. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins* **2006**, *63*, 892−906.

(44) Perot, S.; Sperandio, O.; Miteva, M. A.; Camproux, A. C.; Villoutreix, B. O. Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discov. Today* **2010**, *15*, 656−667.

(45) Edfeldt, F. N.; Folmer, R. H.; Breeze, A. L. Fragment screening to predict druggability (ligandability) and lead discovery success. *Drug Discov. Today* **2011**, *16*, 284−287.

(46) Sheridan, R. P.; Maiorov, V. N.; Holloway, M. K.; Cornell, W. D.; Gao, Y. D. Drug-like density: a method of quantifying the "bindability" of a protein target based on a very large set of pockets and drug-like ligands from the Protein Data Bank. *J. Chem. Inf. Model.* **2010**, *50*, 2029−2040.

(47) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **2007**, *25*, 71−75.

(48) Hajduk, P. J.; Huth, J. R.; Fesik, S. W. Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.* **2005**, *48*, 2518−2525.

(49) Volkamer, A.; Kuhn, D.; Grombacher, T.; Rippmann, F.; Rarey, M. Combining global and local measures for structure-based druggability predictions. *J. Chem. Inf. Model.* **2012**, *52*, 360−372.

(50) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinf.* **2009**, *10*, 168.

(51) Kahraman, A.; Morris, R. J.; Laskowski, R. A.; Thornton, J. M. Shape variation in protein binding pockets and their ligands. *J. Mol. Biol.* **2007**, *368*, 283−301.