PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

A simple and fuzzy method to align and compare druggable ligand-binding sites

Claire Schalon, Jean-Sébastien Surgand, Esther Kellenberger, and Didier Rognan*

Bioinformatics of the Drug, Institut Gilbert Laustriat, CNRS UMR 7175-LC1, 74 route du Rhin, F-67400 Illkirch

ABSTRACT

A novel method to measure distances between druggable protein cavities is presented. Starting from user-defined ligand binding sites, eight topological and physicochemical properties are projected from cavity-lining protein residues to an 80 triangle-discretised sphere placed at the centre of the binding site, thus defining a cavity fingerprint. Representing binding site properties onto a discretised sphere presents many advantages: (i) a normalised distance between binding sites of different sizes may be easily derived by summing up the normalised differences between the 8 computed descriptors; (ii) a structural alignment of two proteins is simply done by systematically rotating/translating one mobile sphere around one immobile reference; (iii) a certain degree of fuzziness in the comparison is reached by projecting global amino acid properties (e.g., charge, size, functional groups count, distance to the site centre) independently of local rotameric/tautomeric states of cavity-lining residues. The method was implemented in a new program (SiteAlign) and tested in a number of various scenarios: measuring the distance between 376 related active site pairs, computing the cross-similarity of members of a protein family, predicting the targets of ligands with various promiscuity levels. The proposed method is robust enough to detect local similarity among active sites of different sizes, to discriminate between protein subfamilies and to recover the known targets of promiscuous ligands by virtual screening.

Proteins 2008; 71:1755–1778.

Key words: alignment; binding site; chemogenomics; ligand; similarity.

INTRODUCTION

High-resolution three-dimensional (3D) protein structures are now widely used in structure-based drug discovery, usually by modelling the protein–ligand interactions. Many success stories of hit finding and hit optimization for a validated target have been reported.[1] Stimulated by chemogenomic[2] and structural genomic projects,[3] structure-based design of drug candidates has progressively evolved from single target to full protein subfamily-biased approaches. Within this context, the rapid growing of protein structure repositories have opened up new perspectives in computed-aided drug design to capture selectivity/promiscuity of ligand binding events by comparing protein binding sites. Several tools for comparing 3D protein binding sites have been proposed in the last 5 years.[4–10] All these approaches are sequence and fold independent and primarily aimed at the automated functional annotation of query proteins by browsing a collection of protein cavities of known 3D structure and function. Pair-wise cavity comparisons usually imply three steps: (i) converting a protein cavity into a simplified representation; (ii) searching a database of known cavities for local similarities; (iii) ranking the most likely hits. Protein binding sites are represented either by a simplified molecular surface (triangle mesh vertices with electrostatic potential and curvature,[4] selected chemically important surface points often called pseudo-centres),[5,6,8] by a discretised 3D image (e.g., Cartesian grid points labelled according to neighbouring atoms)[11] or by selected atomic coordinates of ligand-proximal residues.[9,10] Local similarities between protein representations can be detected by an exhaustive search for equivalent triplets of pseudo-centres,[6] by clique detection algorithms[4,5] (i.e., finding the maximal subgraph isomorphism between the two graphs made of vertices or pseudo-centres connected in a distant-dependent manner), geometric hashing,[8] geometric matching,[11] or iterative conformational search to find the best 3D alignment.[9,10] All methods provide the geometrical operators to superimpose the similar patches of the two input proteins, hence, allowing the evaluation of spatial overlap. Similarity of two cavities is evaluated by counting equivalent objects (e.g., pseudo-centres) and computing their root mean square deviation (rmsd). The utility of all described methods was demonstrated for diverse applications: biochemical functional assignment of a query protein,[8–10,12,13] alignment of conserved active residues in unrelated enzymes catalyzing the

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsor: French Ministry of Research.

*Correspondence to: Didier Rognan, Bioinformatics of the Drug, Institut Gilbert Laustriat, CNRS UMR7175-LC1, 74 route du Rhin, F-67400 Illkirch. E-mail: didier.rognan@pharma.u-strasbg.fr

Received 17 August 2007; Revised 28 September 2007; Accepted 4 October 2007
Published online 3 January 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21858

same reaction (e.g., catalytic triad of serine proteases),[5,6,8,9,13] finding/classifying binding sites accommodating the same ligand in absence of overall sequence or fold similarity of the compared proteins.[8–11,13]

However, some hurdles still need to be overcome. Among the most critical are (i) the high-dependency to high-resolution X-ray or NMR coordinates (rotameric/tautomeric states, ligand-induced backbone shifts), (ii) the lack of a generic similarity score for comparing binding sites as it is quite common for comparing ligands,[14] (iii) the difficulty to compare binding sites of different sizes, and (iv) the comparison of irrelevant binding site cavities (for ions, solvents, detergents). To address the above issues, we herewith present a novel method (SiteAlign) for aligning and measuring distances between druggable binding sites, irrespective of their dimensions. We took a particular attention in selecting only structurally-druggable ligand-binding sites[15] and to consider a ligand from a pharmacological and not a structural point of view. Hence, druggable protein cavities have recognizable physicochemical properties[16,17] and usually consist in deep pockets rather than shallow sites. Observation of experimental drug binding modes reveals that the bound ligand rarely fills completely the protein cavity,[16,18,19] that proteins recognizing similar ligands only have similar geometrical properties when their cognate ligands are not flexible,[18] and that proteins may experience significant structural rearrangements upon ligand binding.[20]

Our method is fuzzy enough to tolerate moderate protein flexibility, applicable to a wide array of protein 3D structures (from high-resolution X-ray structures to low-resolution homology models), insensitive to the definition of the ligand-binding site, and provides a normalised global similarity score that is easy to interpret in terms of biological relevance.

METHODS

Description of the SiteAlign algorithm

The basic idea behind the herein presented methodology is to map binding site properties onto a discretised sphere placed at the centre of the ligand-binding site. Binding site attributes are therefore not described by a variable number of pseudo-centres with properties and Cartesian coordinates, but by a fixed-length cavity fingerprint which enables an easier comparison of ligand-binding sites and a straightforward alignment method. Each step of the structural alignment method will be presented from hereon.

Ligand-binding site definition

Ligand-binding sites were extracted from the sc-PDB dataset[15] and defined as a collection of amino acids (cofactors, solvent, detergents, and metal ions are not taken into account) for which any heavy atom is closer than 6.5 Å from any ligand heavy atom. Atomic coordinates of each sc-PDB entry are derived from those in the PDB, excepted that only protein atomic coordinates are kept. Because chain name is not read in the current version of SiteAlign, amino acids from the second chain should be renumbered to avoid residue number duplication. No explicit surface description is required.

Sphere representation

A 1-Å radius sphere is placed at the centre of the active site defined by the centre of mass of Cα-carbons from cavity-lining residues. The sphere is discretised in 80 triangles of roughly similar dimensions starting from a regular icosahedron (20 identical faces characterized by 12 vertices and 30 edges) and further dividing each triangle into four new ones by joining the three former edge centres by three new edges [Fig. 1(A)]. Because the new vertices are no more located on the sphere, their coordinates are slightly modified to belong to the sphere as follows: If $P'(x,y,z)$ describes a point P' of Cartesian coordinates x,y,z to project on the sphere of radius r, the new coordinates of the projected point P will be $x\frac{r}{r'}, y\frac{r}{r'}, z\frac{r}{r'}$ where $r' = \sqrt{(x^2 + y^2 + z^2)}$.

All 80 triangles are therefore not exactly of the same dimension. The 60 triangles with vertices originating from the starting icosahedron are slightly smaller than the 20 new ones whose vertices have been projected on the sphere in the final discretisation step. The ratio between the surface of a "big" triangle to the surface of a "small" triangle is 0.891 and therefore insignificant considering the global fuzziness of our method. Each triangle is indexed with a number from 1 to 80 which is independent of the relative position of the corresponding sphere in its active site.

Cavity descriptors

Three topological and five chemical descriptors are used to characterise user-selected cavity-lining residues (Table I). The first descriptor reports the distance from the Cβ atom of cavity-lining residues to the sphere centre, excepted for glycine residues for which the Cα atom is used. The distance in Å is discretised in a series of 30 bins of 0.5 Å (from 0 to 15 Å) and the bin number corresponding to the current distance is finally reported. The second descriptor checks whether the side chain of the cavity residue is pointing inward or outward the sphere centre and outputs either a "1" or a "2" value according to the answer to the above question. The last topological descriptor reports the size of the side chain of the cavity residue. Natural amino acids have been classified into three groups according to the number of heavy atoms (<4 heavy atoms: Ala, Cys, Gly, Pro, Ser, Thr, Val; 4–6 heavy atoms: Asn, Asp, Gln, Glu, His, Ile,

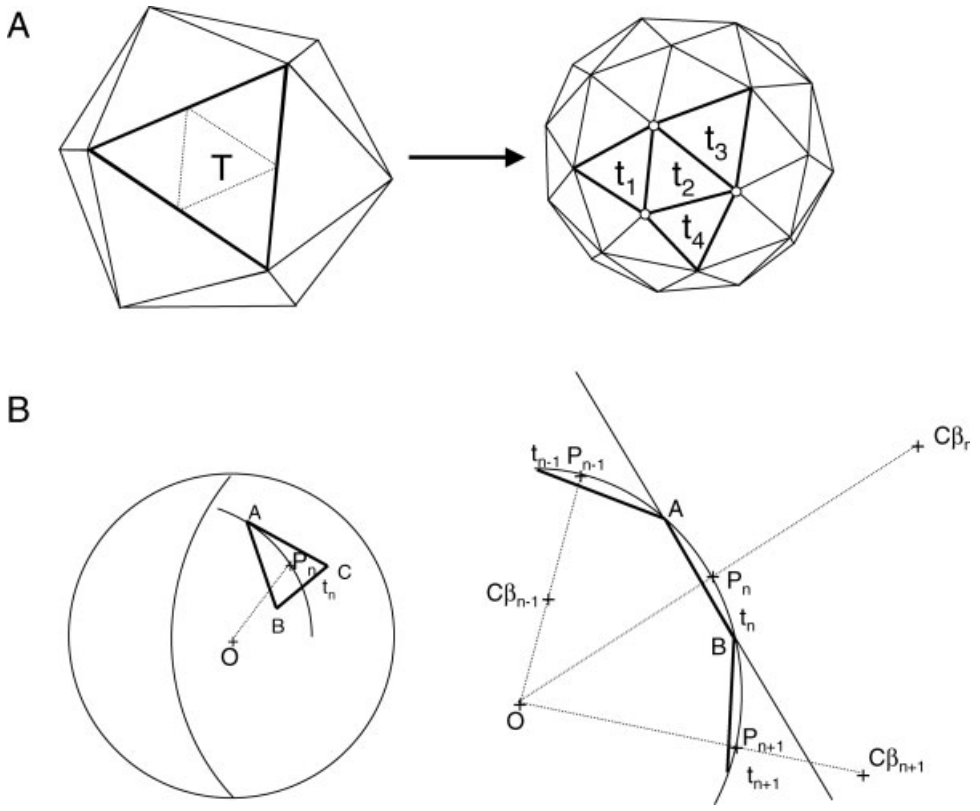


Figure 1

*Sphere construction and projection of cavity descriptors. (**A**) Discretisation of an icosahedron (20 triangular faces, 12 vertices, and 30 edges) into an 80-triangle sphere (80 faces, 42 vertices, 120 edges). Each of the initial 20 triangles T of the icosahedron are divided into four smaller triangles t_1, t_2, t_3, and t_4 by joining the former edges centres to build three new edges (dotted lines in icosahedron) per triangle. The three new vertices (white circles) are projected onto the sphere such that the middle triangle t_2 is slightly larger than triangles t_1, t_3, and t_4 whose vertices originate from the icosahedron. (**B**) Projection of descriptors in one of the 80 triangles of the sphere. A triangle t_n characterized by vertices A, B, C is hit by the projection P_n from C_β of residue n ($C\beta_n$) on a sphere of centre O if $\lfloor \overrightarrow{AO}, \overrightarrow{AB}, \overrightarrow{AC} \rfloor \times \lfloor \overrightarrow{AP}, \overrightarrow{AB}, \overrightarrow{AC} \rfloor \leq 0$. This condition is not verified for neighbouring residues ($C\beta_{n-1}$, $C\beta_{n+1}$) which project on neighboring triangles t_{n-1}, t_{n+1} at projection points P_{n-1}, P_{n+1}.*

Leu, Lys, Met; >6 heavy atoms: Arg, Phe, Trp, Tyr) and three values ("1," "2," "3") are outputted according to the group to which the current residues belong to (Table I). Three out of the five chemical descriptors feature the molecular interaction capacity of each side chain and reports the number of H-bond acceptors, donors, and

Table I

Eight Descriptors Used to Encode Properties of Cavity-Lining Residues

Descriptor	Type	Possible values
Distance C_β-sphere centre	Topological	Any integer between 1 and 30
Side chain orientation	Topological	1 (inward the active site) 2 (outward the active site)
Size	Topological	1 (<4 heavy atoms) 2 (4–6 heavy atoms) 3 (>6 heavy atoms)
H-bond donor count	Chemical	0, 1, 2, 3
H-bond acceptor count	Chemical	0, 1, 2
Aromatic character	Chemical	0 (all but aromatic) 1 (aromatic: His, Hid, Hie, Phe, Tyr, Trp)
Aliphatic character	Chemical	0 (all but aliphatic) 1 (aliphatic: Ala, Cys, Ile, Leu, Lys, Met, Pro, Thr, Val)
Charge	Chemical	−1 (Asp, Glu) 0 (all but charged) 1 (Lys, Arg, Hip)

the formal charge. The last two descriptors account for the hydrophobic character or aromatic character of the residue under investigation (Table I). If no projections occur on a triangle, all descriptors are set to 0. A table of chemical descriptors for all 20 natural amino acids is given as supplementary material.

Projection of cavity descriptors on the sphere

A geometrical vector is derived from the Cα carbon of each residue of the active site to the sphere centre and the corresponding eight descriptor values assigned to the sphere triangle hit by the projection [Fig. 1(B)]. Preliminary tests ascertain that using 80 triangles for a prototypical druggable ligand binding site[15] prevents targeting the same triangle for two different residues in 90% of the projections. Increasing the number of triangles in an additional discretisation step to 320 (80*4) would reduce the number of duplicate projections to 1% but at the cost of an increased computing time. We therefore kept the number of triangles to 80 but register residues projecting towards the same triangle.

In theory, the Cα-sphere projection may also hit either one edge or one vertex and, thus, be assigned to multiple triangles. When applied to the 6415 active sites from the sc-PDB dataset,[15] hitting an edge was very rare (one case every in 10,000 projections) and hitting a vertex even less frequent. Therefore, such projections are not recorded when they appear and unlikely to bias our results. Projections are operated for each residue of the binding site thus defining a vector of eight integers by triangle and a final fingerprint of 640 (8×80) integers for each active site.

Aligning ligand-binding sites

Aligning two active sites i and j is done by finding the highest possible similarity between their respective fingerprints F_i, F_j. Considering an immobile reference site i with its corresponding sphere s_i, the sphere s_j is moved in its site j by user-specified systematic rotations/translations. At each move, cavity descriptors are projected to the updated triangle positions (the indexation from 1 to 80 is kept invariant) and a new fingerprint F_j is generated and compared to the reference F_i. In a first step, a low resolution search is undertaken and the three best solutions are stored. The rotation (Ri) and translation (Ti) increments are derived from user-defined rotation (Rop) and translation (Top) intensities, and user-defined number of rotations (Rn) and translations (Tn) such as:

$$\mathrm{Ri} = \mathrm{Rop}/\mathrm{Rn}$$

$$\mathrm{Ti} = \mathrm{Top}/(\mathrm{Tn} - 1)$$

A local search is then performed around each of the best three initial solutions, by reducing the rotation/translation space and decreasing the rotation/translation increments such as:

$$\mathrm{Rop_{opt}} = 2 \times \mathrm{Ri}; \quad \mathrm{Rn_{opt}} = \mathrm{Rn}/2; \quad \mathrm{Ri_{opt}} = \mathrm{Rop_{opt}}/\mathrm{Rn_{opt}}$$

$$\mathrm{Top_{opt}} = 2 \times \mathrm{Ti}; \quad \mathrm{Tn_{opt}} = \mathrm{Tn}; \quad \mathrm{Ti_{opt}} = \mathrm{Top_{opt}}/(\mathrm{Tn_{opt}} - 1)$$

Scoring without aligning is simply done by assigning null values to Rop and Top parameters. Otherwise, the alignment giving the highest similarity between fingerprints F_i and F_j is stored as final solution.

Scoring the alignment

The similarity score is calculated from a sum of normalized differences for each descriptor of each triangle. Starting from the first indexed triangle t, a normalized difference $S^d_{t,ij}$ is calculated for every descriptor d between sites i and j as:

$$S^d_{t,ij} = 1 - \frac{|v^d_{t,i} - v^d_{t,j}|}{v^d_{max} - v^d_{min}} \tag{1}$$

where $v^d_{t,i}$ is the current value of descriptor d for triangle t in site i; $v^d_{t,j}$ is the current value of descriptor d for triangle t in site j; v^d_{max} is the largest possible value for descriptor d; v^d_{min} is the smallest possible value for descriptor d.

In cases where several residues target the same triangle for a single site, the one which maximises the $S^d_{t,ij}$ value is finally kept. If such a case exists in both sites, the best pair is conserved.

The procedure is iterated for each of the eight descriptors and a similarity score $S_{t,ij}$ for triangle t is calculated as:

$$S_{t,ij} = \frac{1}{8}\sum_{d=1}^{S} S^d_{t,ij} \tag{2}$$

After iterating the procedure for each of the 80 triangles from the first indexed triangle t_1 to the last one t_{80}, two similarity scores S_1 and S_2 are calculated as follows:

$$\begin{aligned} S_1 &= \frac{1}{N_1}\sum_t S_t \\ S_2 &= \frac{1}{N_2}\sum_t S_t \end{aligned} \tag{3}$$

where N_1 is the number of triangles with non-null values for either sites i or j and N_2 is the number of triangles with non-null values for both sites i and j.

To avoid comparing binding sites with too few common projections, a minimal threshold of seven common projections ($N_2 = 7$) has been used to consider an alignment. Otherwise, the S_2 score has been set to 0.

Table II
Default Rotation/Translation Parameters Used in SiteAlign

Parameter	Global search	Optimization
Rotation intensity (deg)	Rop = 360	Rop_{opt} = 45
Rotation moves	Rn =16	Rn_{opt} = 8
Rotation increment (deg)	Ri = 22.5	Ri_{opt} = 5.625
Translation intensity (Å)	Top =4	Top_{opt} = 2
Translation moves	Tn = 5	Tn_{opt} = 5
Translation increment (Å)	Ti = 1	Ti_{opt} = 0.5

Throughout the manuscript, distances ($D_1 = 1 - S_1$; $D_2 = 1 - S_2$) will be used instead of the similarity scores.

In addition to the quantification of the alignment, the geometrical operators to fit site j to i according to the best alignment score are given to visualise the proposed alignment. For typical binding sites of about 30 residues, an alignment is obtained in about 60 s. on a standard PC/Linux (e.g., Opteron 2.8 GHz with 1 GB RAM).

SiteAlign comparisons

For each entry to align, SiteAlign requires a pdb file and a list of residue numbers defining the ligand-binding site. Last, a third input file specifies which proteins to align, the first one being the reference. In all but one *in silico* screen, default SiteAlign parameters (Table II) were used. For comparing sc-PDB entries to the staurosporine-binding sites in protein kinases, only 15 rotations and 3 translations (Rn = 15, Tn = 3) were done to fasten the computation. Other parameters were left unchanged.

Iterative computation of ROC scores

A Pipeline Pilot protocol[21] was set-up to compute, for each different protein name stored in the sc-PDB database, the area under the ROC curve (ROC score) of all classifications. sc-PDB entries were ranked according to their D_2 distance to the reference first, and according to the D_1 distance for equivalent D_2 scores. Please note that D_2 was arbitrarily set to 1 when the corresponding D_1 distance was higher than 0.6. Proteins were ranked by decreasing ROC scores and selected for analysis at the condition that (i) the ROC score was above 0.75, (ii) the lowest D_2 score for at least one entry of that protein was lower than 0.2.

Setting-up a collection of 376 nonredundant related ligand-binding site pairs

The sc-PDB database[15] was first filtered to remove entries lacking an E.C. annotation,[22] mutant proteins and binding sites with high thermal motion (mean B-factor of cavity-lining heavy atoms $\geq$ 30 Å^2). The 4108 remaining enzyme entries describe 730 functional classes as indicated by the fourth level of the EC number. According to the number of entries belonging to each functional class, three different cases were further investigated.

Classes populated by only one entry were first discarded. If the class included exactly two entries (141 classes), the identity of the concatenated sequences of the ligand-binding sites was calculated using the GCG10 Bestfit routine.[23] If the length of the alignment was smaller than half of the shortest sequence, the corresponding two ligand-binding sites were considered different. If the length of the alignment was bigger than half of the smallest concatenated sequence and if the identity was sufficient (identity score $>$ 70% or at least eight identical residues), the two sites were kept and considered similar. Otherwise, the sites were considered different and discarded. At this step, a total of 118 pairs of related binding sites were selected.

Last, for classes populated by more than two entries, two of them were selected as follows. As before, the active site similarity has been evaluated by comparison of the concatenated sequences using the multiple alignment program ClustalW.[24] All concatenated sequences having a similarity score better than 50% were grouped together thus defined a fifth level in the E.C. classification and the entries cocrystallized with the lowest and the highest molecular-weight ligands were selected in each newly-defined class. Concatenated sequences with a similarity score lower than 50% were clustered in subgroups until the minimum similarity score within each subgroup was above 50%. Two entries by subgroup were then selected as stated above. In total 258 pairs originating from E.C. classes populated by more than three entries were retrieved to yield a final dataset of 376 pairs of entries with similar ligand-binding sites.

RESULTS

Quantifying the similarity of related ligand-binding sites

To define similarity/distance thresholds for discriminating similar from dissimilar ligand-binding sites, a dataset comprising 376 pairs of related binding sites was set-up from the sc-PDB repository of druggable binding sites.[15] To facilitate the biological annotation of selected binding sites, only enzymes cocrystallized with a drug-like ligand were retrieved first. Two entries were then extracted for every occurrence of a new E.C. number.[22] To guarantee the selection of real diverse binding sites, the lowest and the largest molecular weight ligands of all entries sharing the same E.C. number were used as filters to define a pair of corresponding binding sites. For a given pair AB, every binding site was aligned to its congeneric entry (A vs. B, B vs. A) and the D_1 and D_2 distances of the best possible alignment recorded. The D_1 distance distribution over 752 comparisons shows a

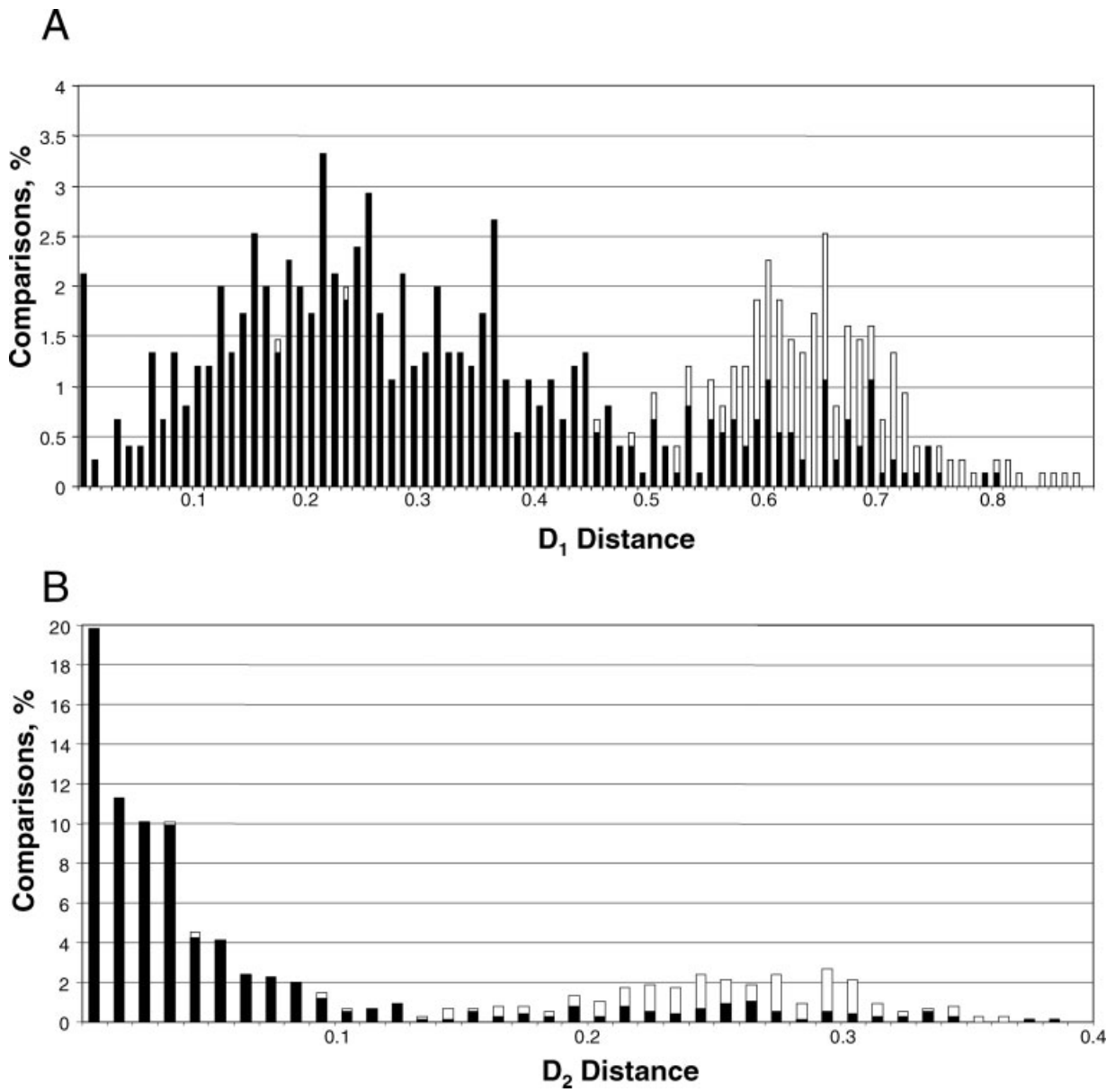


Figure 2

Distribution of ligand binding site distances (D_1 distance, panel A; D_2 distance, panel B) on 376 pairs of related sc-PDB entries (two entries by E.C. number). The quality of the alignment was checked irrespectively of the computed distances, by computing the volume overlap between both ligands for each pair of aligned binding sites. Dark bars indicate a percentage of volume overlap higher or equal to 30%, white bars indicate a percentage of volume overlap below 30%.

rather bimodal distribution with optimal D_1 values at 0.22 and 0.66, respectively [Fig. 2(A)]. Distribution of the D_2 distance is biased towards lower values (0.0–0.1) with still another minor distribution centred around higher values of about 0.3 [Fig. 2(B)].

To establish relationships between the D_1, D_2 distance values and the quality of the alignment, a metric was designed to distinguish "good" from "bad" alignments. We found that computing the percentage of volume overlap (VO) of SiteAlign fitted protein-bound ligands (whose coordinates can be easily reconstructed after merging the X-ray protein coordinates to the fitted protein coordinates) enables a good distinction of good from bad alignments which have been visually checked, one by one, and compared to a full protein sequence-guided structural alignment.[25] A VO threshold of 30% seems adequate to differentiate between acceptable from unacceptable fits. This enable us to propose distance thresholds for related binding sites since 80% of related binding site pairs are characterised by a D_1 distance

Table III
Percentage of Related Ligand-Binding Site Pairs Recovered by Using Various Cut-Off Distances on a Dataset of 376 Binding Site Pairs

Cut-off distance	Pairs recovered (%)	Good alignment (%)	Bad alignment (%)
$D_1 \leq 0.6$	79.8	75.3	4.5
$D_2 \leq 0.2$	79.8	75.8	4.0
$D_1 \leq 0.6$ and $D_2 \leq 0.2$	75.5	73.9	1.6

The quality of the alignment for each pair was checked by visualising fitted protein coordinates and by computing the percentage of volume overlap (VO) of the two protein-bound ligands (good alignment if VO $\geq$ 30%; bad alignment if VO $<$ 30%).

lower than 0.6 or a D_2 distance lower than 0.2 (Fig. 2, Table III). Among those pairs which are found similar, nearly all of them were well aligned after visual check and computing the VO value. Applying a consensus cut-off ($D_1 \leq 0.6$ and $D_2 \leq 0.2$), slightly limits the percentage of related pairs which are recovered (75.5%) but decreases the percentage of false positives (good score and bad alignment) to a marginal value of 1.6% (Table III). From here on, the conditions to consider two ligand-binding sites similar will be thus defined by the above-defined consensus cut-off.

Quantifying binding site similarity across a protein family

A first validation of our structural alignment method has been the estimation of inhibitor-binding site similarity for protein entries of a same gene family. A prototypical ligand-binding site in bovine trypsin (pdb entry 1aq7) was chosen as reference and compared to all 6415 sc-PDB entries. Active site distances were evaluated with the D_2 score assuming preceeding results suggesting that two active sites can only be considered as similar if the D_1 distance is lower than 0.6 and the D_2 distance lower than 0.2. Therefore, active sites with D_1 scores higher than 0.6 were arbitrarily assigned a D_2 score of 1. Out of 245 active sites selected by the $D_1 - D_2$ consensus threshold, 226 (92%) originate from a protein sharing with the 1aq7 reference a trypsin fold and a trypsin-like substrate cleavage specificity (Fig. 3). Interestingly, 10 active sites (4%) were selected as similar to that of bovine trypsin although they share the trypsin fold but not the substrate cleavage specificity (e.g., elastase, chymotrypsinogen). One serine protease entry (1p7v) having another fold than trypsin was nevertheless found similar to 1aq7, regarding the ligand binding site. Last, only eight out of the 6058 nonserine proteases stored in the sc-PDB database were retrieved among the predicted similar binding sites. The sensitivity and specificity of the *in silico* comparison was evaluated by computing a Receiver-Operating Characteristic (ROC) plot[26] for each protein name and calculating the area under the ROC curve (ROC score, Table IV). The ROC scores were obtained iteratively, for each different protein name of the sc-PDB dataset, by ranking all entries by decreasing D_2 score and looking at the rank of all entries of the protein name under investigation with respect to all other entries. A ROC score higher than 0.5 and close to 1 indicate a selective and specific enrichment among those protein name entries in top-scored binding sites (low D_2 distance) whereas a ROC score close to 0.5 indicate no sensibility/specificity of the scoring method (random selection).

After ranking all proteins by decreasing ROC score, all selected proteins with active sites predicted similar to that of bovine trypsin are indeed serine proteases. Interestingly, two of them (chymotrypsinogen A, elastase) although sharing the trypsin fold with the 1aq7 reference do not share the same substrate cleavage specificity. Proteinase K was also selected even if it presents a subtilisin fold different from the trypsin one. This observation prompted us to compute ROC scores for five categories of proteins classified according to their fold and substrate cleavage specificity (Table V). As expected, proteins with a trypsin-like fold and trypsin substrate cleavage specificity get the highest ROC score. Keeping the same fold but changing the cleavage specificity results in a lower ROC score. However, serine proteases with a subtilisin fold quite different from that of trypsin are still well ranked

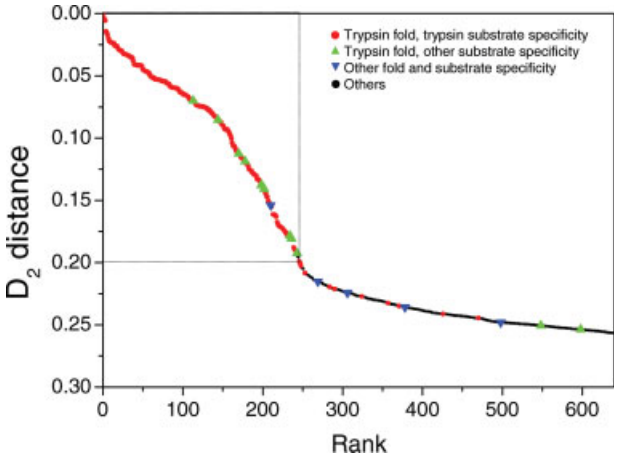


Figure 3
D_2 distance between the ligand binding site of bovine trypsin (1aq7 entry, ligand HET code: AEB) and the top 10% ranked sc-PDB entries. Protein entries are ranked by decreasing D_2 distance and classified in four groups: serine proteases with trypsin fold and trypsin substrate specificity (red circles), serine proteases with trypsin fold and other substrate specificity (green up triangles), serine protease with other fold and substrate specificity (blue down triangles), non serine proteases (black circles). A dashed rectangle indicates the D_2 cut-off distance (0.2) used to discriminate similar from dissimilar binding sites. Alignments have been generated using default settings of SiteAlign (See Methods). The D_2 distance was set to 1 for any binding site whose D_1 distance was higher than 0.6.

Table IV
Similarity of 6415 Ligand Binding Sites, Annotated by Protein Name, to a Trypsin Ligand-Binding Site (pdb Entry 1aq7, Ligand HET Code: AEB)

Protein	Rank	ROC score	n	D_2^{mean}	D_2^{min}	D_2^{max}	D_2^{sd}
Trypsin	1	0.970	165	0.12	0.00	1.00	0.23
Coagulation factor X	2	0.968	19	0.15	0.12	0.19	0.01
Urokinase-type plasminogen activator	3	0.936	28	0.16	0.05	1.00	0.23
Proteinase K	4	0.889	5	0.36	0.15	1.00	0.31
Chymotrypsinogen A	5	0.878	6	0.33	0.13	1.00	0.30
Elastase	6	0.823	9	0.43	0.07	1.00	0.40
Coagulation factor VII	7	0.792	6	0.55	0.06	1.00	0.44

D_2 score is used to rank entries by decreasing distance to the 1aq7 reference. If the D_1 score of the alignment is higher than 0.6, the D_2 distance was arbitrary set to 1, and corresponding sc-PDB entries ranked by decreasing original D_2 score and if necessary, by decreasing D_1 score. A Receiver-Operating Characteristic (ROC) score (area under the curve of the ROC plot) is computed, within an in-house Pipeline Pilot workflow,[21] for each occurrence of a protein name represented by at least five entries in the sc-PDB dataset. Proteins were filtered according to the minimal D_2 distance observed for all entries of that protein ($D_2^{min} \leq 0.20$) in the sc-PDB dataset and ranked by decreasing ROC score. Only proteins with ROC scores higher than 0.75 were considered as sufficiently close to the reference. n is the number of sc-PDB entries for a particular protein. D_2^{mean}, D_2^{min}, D_2^{max}, and D_2^{sd} are mean, minimum, maximum and standard deviation values for the D_2 distance of all sc-PDB entries of a given protein binding site to the reference.

(ROC score = 0.819). Serine proteases with the α/β hydrolase fold have, in our comparison, a ligand-binding site quite different from that of bovine trypsin. Last, nonserine proteases are logically found to be distant from bovine trypsin. Seven entries from the latter category (1lvu, 1ocl, 1v6l, 1obl, 1zdf, 1amk, 1uou, 1wbl), however, pass the $D_1 - D_2$ score similarity threshold. Because ROC scores show only statistical significance for proteins present in multiple copies ($n \geq 5$) in the starting dataset, retrieving less populated proteins ($n < 5$) was done by simply computing enrichments in high scored entries ($D_1 \leq 0.6$ and $D_2 \leq 0.2$) among sc-PDB active sites. Only two proteins (hepsin, malonamidase E2) show enrichment higher than 50%. Hepsin is a serine endopeptidase and malonamidase E2 is a linear peptide amidase with a Ser-cisSer-Lys catalytic triad.

Predicting binding sites for ligands with different promiscuity levels

Degeneracy in ligand binding can be achieved when two different proteins exhibit sufficient similarity in their binding sites for accommodating the same compound. Our active site comparison method offers a good opportunity to verify this statement by trying to recover the known targets of ligands with various promiscuity levels.

We first examined the possible binding sites of 4-hydroxy tamoxifen (4-OHT), a ligand known to primarily bind to both estrogen receptor (ER) subtypes (α, β) and to the estrogen-related receptor (ERR) γ subtype,[29] all present in variable copy numbers in the sc-PDB active site database. A ligand-binding site of each of these three nuclear hormone receptors was, thus, used as a reference to compute active site distances to all sc-PDB entries (Fig. 4). Seventy-nine actives sites were found similar to the 4-OHT binding site in the ERα receptor [Fig. 4(A)]. Twelve out of the 13 ERα binding sites and four out of the seven ERβ binding sites present in the sc-PDB are present among the selected entries. None of the two ERRγ was retrieved although one entry (1s9q) has been cocrystallized with 4-OHT itself. A significant increase of the computed D_2 distance can be observed between the first group of estrogen receptor entries ($D_2 < 0.06$) and the next group of targets ($D_2 > 0.15$) which contains almost only nonestrogen receptors. The same observations can be done if the reference originates from the estrogen receptor β with 7/7 ERβ entries and 5/14 ERα entries retrieved among the most 68 similar sc-PDB active sites [Fig. 4(B)]. Interestingly, both ERRγ entries have been now retrieved and delimit the group of estrogen receptors ($D_2 < 0.15$) from other targets ($D_2 > 0.15$). Considering then the 4-OHT binding site of the ERRγ receptor as reference, 64 close active sites were selected [Fig. 4(C)] consisting in three separated groups: both ERRγ sites ($D_2 < 0.03$), other estrogen receptors (10 ERα, 7 ERβ entries; $0.06 < D_2 < 0.12$), and nonestrogen receptors ($D_2 > 0.12$). Area under the ROC curves were calculated for all three virtual screens (Tables VI–VIII) to distinguish the most statistically relevant targets by considering both the sensitivity (ability to highly rank true positive entries) and the specificity (ability to badly rank true negative entries) of the distance scoring function. Whatever the reference used, both estrogen receptors were ranked at the top two positions. Interestingly, several other targets present in multiple copies ($n \geq 5$) and enriched in ligand-binding sites similar to the reference, have been found (e.g., HIV-1 and HIV-2 protease, p38 MAP kinase) out of which at least one of them (MAP kinase) is a true target of 4-OHT.[30] Among less populated proteins (<5) in our starting dataset, a few

Table V
Similarity of 6415 Ligand Binding Sites, Annotated by Fold, to a Bovine Trypsin Ligand-Binding Site (pdb Entry 1aq7, Ligand HET code: AEB)

Fold	ROC score	n	D_2^{mean}	D_2^{min}	D_2^{max}	D_2^{sd}
Trypsin fold and Trypsin specificity	0.877	314	0.28	0.00	1.00	0.3
Trypsin fold and other specificity	0.786	22	0.44	0.07	1.00	0.38
Subtilisin fold	0.819	11	0.44	0.15	1.00	0.34
α/β hydrolase fold	0.309	10	0.93	0.33	1.00	0.20
Others	0.144	6058	0.79	0.15	1.00	0.32

D_2 score is used to rank entries by decreasing distance to the 1aq7 reference. If the D_1 score of the alignment is higher than 0.6, the D_2 distance was arbitrary set to 1, and corresponding sc-PDB entries ranked by decreasing original D_2 score and if necessary, by decreasing D_1 score. N, D_2^{mean}, D_2^{min}, D_2^{max}, and D_2^{sd} are defined as in Table IV. Classification by fold and substrate cleavage specificity was done according to the CATH protein structure classification[27] and CutDB proteolytic event database.[28]

proteins were also found enriched in similar ligand-binding site ($D_1 \leq 0.6$ and $D_2 \leq 0.2$). Two of them are nuclear hormone receptors (RXR-α receptor and of course the ERR-γ receptor), two others are dioxygenases (seed lipoxygenase-3, catechol 1,2-dioxygenase), the last two targets being major urinary proteins-1 and 6, respectively (Table IX).

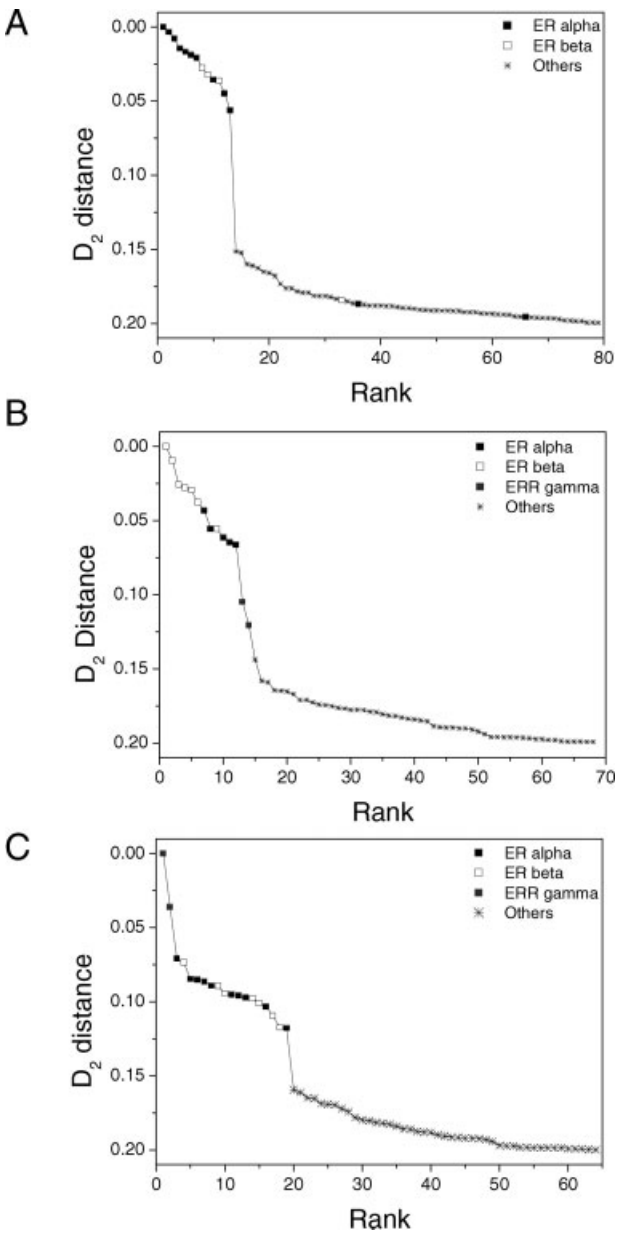


Figure 4

*D_2 distance between the 4-hydroxy tamoxifen binding site in human estrogen receptors and 6415 sc-PDB entries. Protein entries are ranked by decreasing D_2 distance and classified in four groups: estrogen receptor α (filled squares), estrogen receptor β (empty squares), estrogen-related receptor γ (gray squares), other targets (crosses). Alignments have been generated using default settings of SiteAlign (See Methods). The D_2 distance was set to 1 for any binding site whose D_1 distance was higher than 0.6. (**A**) estrogen receptor α (pdb entry: 3ert, ligand HET code: OHT) as reference, (**B**) estrogen receptor β (pdb entry: 1x7b, ligand HET code: O41) as reference, (**C**) estrogen-related receptor γ (pdb entry: 1s9q, ligand HET code: OHT) as reference.*

Table VI

Similarity of 6415 Ligand Binding Sites, Annotated by Protein Name, to an Estrogen Receptor α Ligand-Binding Site (pdb Entry: 3ert, Ligand HET code: OHT)

Protein	Rank score	ROC	n	D_2^{mean}	D_2^{min}	D_2^{max}	D_2^{sd}
Estrogen receptor α	1	0.992	13	0.06	0.00	0.23	0.08
Estrogen receptor β	2	0.983	7	0.13	0.02	0.22	0.08
L-isoaspartate O-methyltransferase	3	0.917	8	0.22	0.18	0.27	0.02
HIV-2 protease	4	0.894	10	0.22	0.15	0.27	0.04
HIV-1 protease	5	0.887	140	0.25	0.18	1.00	0.11
Phosphodiesterase 5A	6	0.830	7	0.33	0.18	1.00	0.27
p38 MAP kinase 14	7	0.820	26	0.35	0.18	1.00	0.28
Catabolite gene activator	8	0.780	6	0.36	0.19	1.00	0.28
cAMP-dependent protein kinase	9	0.766	14	0.35	0.19	1.00	0.26
67-Dimethyl-8-ribityllumazine synthase	10	0.765	9	0.33	0.19	1.00	0.23

D_2 score is used to rank entries by decreasing distance to the 3ert reference. If the D_1 score of the alignment is higher than 0.6, the D_2 distance was arbitrary set to 1, and corresponding sc-PDB entries ranked by decreasing original D_2 score and if necessary, by decreasing D_1 score. A ROC score is computed, within an in-house Pipeline Pilot workflow,[21] for each occurrence of a protein name represented by at least five entries in the sc-PDB dataset. Proteins were filtered according to the minimal D_2 distance observed for all entries of that protein ($D_2^{min} \leq 0.20$) in the sc-PDB dataset and ranked by decreasing ROC score. Only proteins with ROC scores higher than 0.75 were considered as sufficiently close to the reference. n is the number of sc-PDB entries for a particular protein. N, D_2^{mean}, D_2^{min}, D_2^{max}, and D_2^{sd} are defined as in Table IV.

We next examined the case of a ligand showing a broader target spectrum but limited to a protein family. Staurosporine (STAU) was a perfect candidate since it is known to bind to the ATP-binding site of most protein kinases[31] but not to targets from other gene families.* Six STAU-binding sites originating from either serine/threonine-protein kinases (Aurora-A, cyclin-dependant kinase 2, Pim-1, Protein kinase C) or tyrosine-protein kinases (Lck, interleukin-2 tyrosine kinase) were, thus, compared to the entire collection of sc-PDB active sites. From 96 to 136 sites were found similar to one of the six references, a large majority of them (81–94%) being ATP-binding sites from protein kinases (Fig. 5). ROC plots unambiguously discriminated protein kinases from decoys as true staurosporine-binding targets with ROC scores of about 0.8 whatever the reference binding site (Table X). Other kinases ($n = 329$) present significantly different ligand-binding sites, as acknowledged by ROC score values close to random picking for any of the six *in silico* screens (Table X). The ATP-binding sites of the six investigated protein kinases were also different from a set of 166 generic ATP/ADP binding sites or from the 5683 remaining decoys.

*A search for staurosporine targets in the PDSP database (http://pdsp.med.unc.edu/pdsp.php) returned 119 targets, all of them being protein kinases (accessed in July 2007).

Table VII
Similarity of 6415 Ligand Binding Sites, Annotated by Protein Name, to an Estrogen Receptor β Ligand-Binding Site (pdb Entry: 1x7b, Ligand HET code: O41)

Protein	Rank	ROC score	n	D_2^{mean}	D_2^{min}	D_2^{max}	D_2^{sd}
Estrogen receptor β	1	0.999	7	0.02	0.00	0.05	0.01
Estrogen receptor α	2	0.967	13	0.16	0.04	0.24	0.08
HIV-2 protease	3	0.884	10	0.24	0.18	0.28	0.02
Chorismate-pyruvate lyase	4	0.871	6	0.34	0.17	1.00	0.29
HIV-1 protease	5	0.848	140	0.27	0.18	1.00	0.14
p38 MAP kinase 14	6	0.839	26	0.35	0.16	1.00	0.27
Protein kinase Pim-1	7	0.802	9	0.32	0.18	1.00	0.24
Tryptophan synthase	8	0.798	23	0.37	0.17	1.00	0.28

D_2 score is used to rank entries by decreasing distance to the 1x7b reference. If the D_1 score of the alignment is higher than 0.6, the D_2 distance was arbitrary set to 1, and corresponding sc-PDB entries ranked by decreasing original D_2 score and if necessary, by decreasing D_1 score. A ROC score is computed, within an in-house Pipeline Pilot workflow,[21] for each occurrence of a protein name represented by at least five entries in the sc-PDB dataset. Proteins were filtered according to the minimal D_2 distance observed for all entries of that protein ($D_2^{min} \leq 0.20$) in the sc-PDB dataset and ranked by decreasing ROC score. Only proteins with ROC scores higher than 0.75 were considered as sufficiently close to the reference. N, D_2^{mean}, D_2^{min}, D_2^{max}, and D_2^{sd} are defined as in Table IV.

Last, we investigated the case of a very permissive small molecular weight ligand, namely adenosine diphosphate (ADP). There are 136 true ADP-binding sites in the sc-PDB dataset (for which the bound ligand is ADP itself) covering a total of 84 proteins and 44 different E.C. numbers. ADP is known to bind to very heterogeneous binding sites with multiple binding modes[18] such that low similarity scores should be expected among these binding sites. All sc-PDB active sites were thus compared to an ADP-binding site of a nucleoside diphosphate kinase. Only 43 ligand-binding sites were found similar enough to the reference with the top 14 ranked entries describing exactly nucleoside diphosphate kinase bound to ADP and/or ADP analogues. Among known ADP-binding proteins, only three others sites from a pancreatic endoribonuclease could be retrieved (Fig. 6).

Systematic calculation of ROC scores for all protein names in the sc-PDB dataset suggests only three putative targets for ADP (Table XI) out of which the top ranked entry is the query protein itself. True ADP from other binding sites could not be discriminated by ROC scores which were identical to what should be expected by random picking (ROC scores of 0.50 for both categories). Furthermore, no ligand-binding sites from less populated proteins ($n < 5$) were found similar to the ATP site of the 1nlk entry.

Clustering ligand-binding sites from homology models of a target gene family

A possible application of the herein presented methodology is the clustering of ligand-binding sites from the same protein family. Because SiteAlign was designed to be fuzzy enough to be applicable to homology models, we decided to evaluate the sensitivity of the alignment and comparison procedures on human G protein-coupled receptors (GPCRs). Two randomly-chosen receptors for each of the previously-defined 22 GPCR receptor clusters[33] were thus selected for defining a full D_2 distance matrix calculated from the corresponding 44 homology models[34] and using our in-house definition of a generic antagonist-binding site[33] applicable to any nonolfactive GPCR. The obtained distance matrix

Table VIII
Similarity of 6415 Ligand Binding Sites, Annotated by Protein Name, to an Estrogen-Related Receptor γ-Binding site (pdb Entry 1s9q, ligand HET Code: OHT)

Protein	Rank	ROC score	n	D_2^{mean}	D_2^{min}	D_2^{max}	D_2^{sd}
Estrogen receptor β	1	0.998	7	0.09	0.07	0.11	0.01
Estrogen receptor α	2	0.982	13	0.12	0.07	0.25	0.06
Heat shock protein HSP 90-α	3	0.875	18	0.31	0.18	1.00	0.24
Nicotinamide mononucleotide adenylyltransferase 3	4	0.840	5	0.39	0.18	1.00	0.30
(+)-Acetone-cyanohydrin lyase	5	0.835	6	0.47	0.18	1.00	0.37
HIV-1 protease	6	0.780	140	0.37	0.15	1.00	0.27
p38 MAP kinase 14	7	0.766	26	0.47	0.16	1.00	0.35
Guanidinoacetate *N*-methyltransferase	8	0.765	5	0.38	0.18	1.00	0.30

D_2 distance is used to rank entries by decreasing distance to the 1s9q reference. If the D_1 score of the alignment is higher than 0.6, the D_2 distance was arbitrary set to 1 and corresponding sc-PDB entries ranked by decreasing original D_2 distance and if necessary, by decreasing D_1 score. A ROC score is computed, within an in-house Pipeline Pilot workflow,[21] for each occurrence of a protein name represented by at least five entries in the sc-PDB dataset. Proteins were filtered according to the minimal D_2 distance observed for all entries of that protein ($D_2^{min} \leq 0.20$) in the sc-PDB dataset and ranked by decreasing ROC score. Only proteins with ROC scores higher than 0.75 were considered as sufficiently close to the reference. n is the number of sc-PDB entries for a particular protein. N, D_2^{mean}, D_2^{min}, D_2^{max}, and D_2^{sd} are defined as in Table IV.

Table IX
Rare Proteins ($n < 5$) Found Similar to the 4-Hydroxy-tamoxifen Binding Sites in Estrogen Receptors α, β, and the Estrogen-Related Receptor γ

Ref	Protein	Enrichment	n	PDB	HET	D_2^{min}
3ert	Retinoic acid receptor RXR-α	100	2	1mzn	BM6	0.15
	Catechol 12-dioxygenase	100	2	1dlq	LIO	0.16
	Seed lipoxygenase-3	50	4	1no3	4NC	0.16
1x7b	Estrogen-related receptor γ	100	2	1s9p	DES	0.10
	Major urinary protein 1	100	2	1qy1	PRZ	0.17
	Major urinary protein 6	66	3	1mup	TZL	0.17
1s9q	Estrogen-related receptor γ	100	2	1s9q	OHT	0.00
	Retinoic acid receptor RXR-α	100	2	1mvc	BM6	0.17

D_2 distance is used to rank sc-PDB entries by decreasing distance to each of the three references 3ert, 1x7b, and 1s9q). If the D_1 score of the alignment is higher than 0.6, the D_2 distance was arbitrary set to 1 and corresponding sc-PDB entries ranked by decreasing original D_2 distance and if necessary, by decreasing D_1 score. Proteins were filtered according to the enrichment (>50%) in ligand-binding sites found similar ($D_2 \leq 0.2$) to that of the reference. For each reference and each protein, the entry is referenced by the PDB identifier (PDB) and ligand HET code (HET) exhibiting the smallest D_2 score (D_2^{min}).

(Fig. 7) enables the unambiguous clustering of all 22 GPCR pairs in separate clusters (see the distance scores along the crossdiagonal of the matrix). It should be noticed that receptor clusters, for which no small molecular-weight ligands (e.g., both receptors from the Frizzled cluster) or at least very few ligands (e.g., MAS cluster, metabotropic glutamate receptors) exist, are predicted to have unique ligand-binding sites not resembling that of their congeners. Conversely, several receptor subfamilies (e.g., amines, peptides, opiates) seem to present more permissive ligand-binding sites. There is however no strict relationships between binding site permissivity and number of known ligands within a cluster. For example, the orphan subfamily of super conserved receptors expressed in the brain (SREBs) is nevertheless found similar, regarding its ligand-binding site, to several liganded receptors (Fig. 7).

DISCUSSION

We herewith present a novel structural alignment method (SiteAlign) aimed at finding the best possible match between ligand-binding sites and quantifying the distance/similarity among them. SiteAlign is basically different from existing methods[5,6,8,10,11,37] able to compare binding sites in many features which have been specifically addressed in the design of the algorithm.

First, instead of describing ligand–binding sites by a formal atom/pseudo-atom definition with attached physicochemical properties (e.g., pharmacophoric features), we use a generic 80 triangle-discretised sphere placed at the centre of the cavity onto which topological and physicochemical descriptors are projected and define a fixed length cavity fingerprint. Projecting atomic properties (unit vectors connecting consecutive Cα atoms) onto a sphere has already been reported for computing unit vector root-mean square deviations (uRMSD) between two protein chains[38] and for estimating dissimilarity of two protein pockets by computing orientational rmsd (oRMSD) between cavity residues projected onto the unit sphere.[39] Our method differs from that previously reported by projecting much more properties (Table I) but paying less attention to atomic coordinates (single distance descriptor) to achieve the desired balance between accuracy and fuzziness.

Second, only druggable ligand-binding sites[15,17] are compared which means that cavities are carefully selected according to several properties (e.g., drug-like ligand occupancy, volume and buried surface area ranges) to avoid the comparison of very different and/or unsuitable cavities.

Third, the use of 80 triangles to discretise the local sphere ensures an adequate description of all druggable cavities which generally contains around 20–30 residues (see a distribution of binding site sizes in supplementary Fig. 1).

Fourth, structural alignment is operated by simply rotating/translating one sphere in its active site and generating at each move a new set of different cavity descriptors, while keeping the reference sphere rigid. Scoring the alignment is done by finding which move of the mobile sphere minimises the difference between the two cavity fingerprints. Because a normalised distance between both fingerprints is calculated, the similarity/distance score is easy to interpret, does not rely on a number of aligned features (e.g., atom triplets, pharmacophores) or the root-mean square deviation between them, but simply varies from 0 to 1. Two distances are outputted, the first one (D_1) is a global measure of binding site similarity, the second one (D_2) is more suited to detect local similarity among unrelated binding sites.

Fifth, avoiding a strict dependency to atomic coordinates is reached by projecting cavity descriptors to the sphere centre from a single representative atom (Cβ) of each residue whose coordinates are less dependent of a peculiar rotameric state. To quantify the degree of fuzziness of our method, molecular dynamics (MD) snapshots of a test protein were recorded and compared to the starting set of coordinates [Fig. 8(A)]. Plotting the

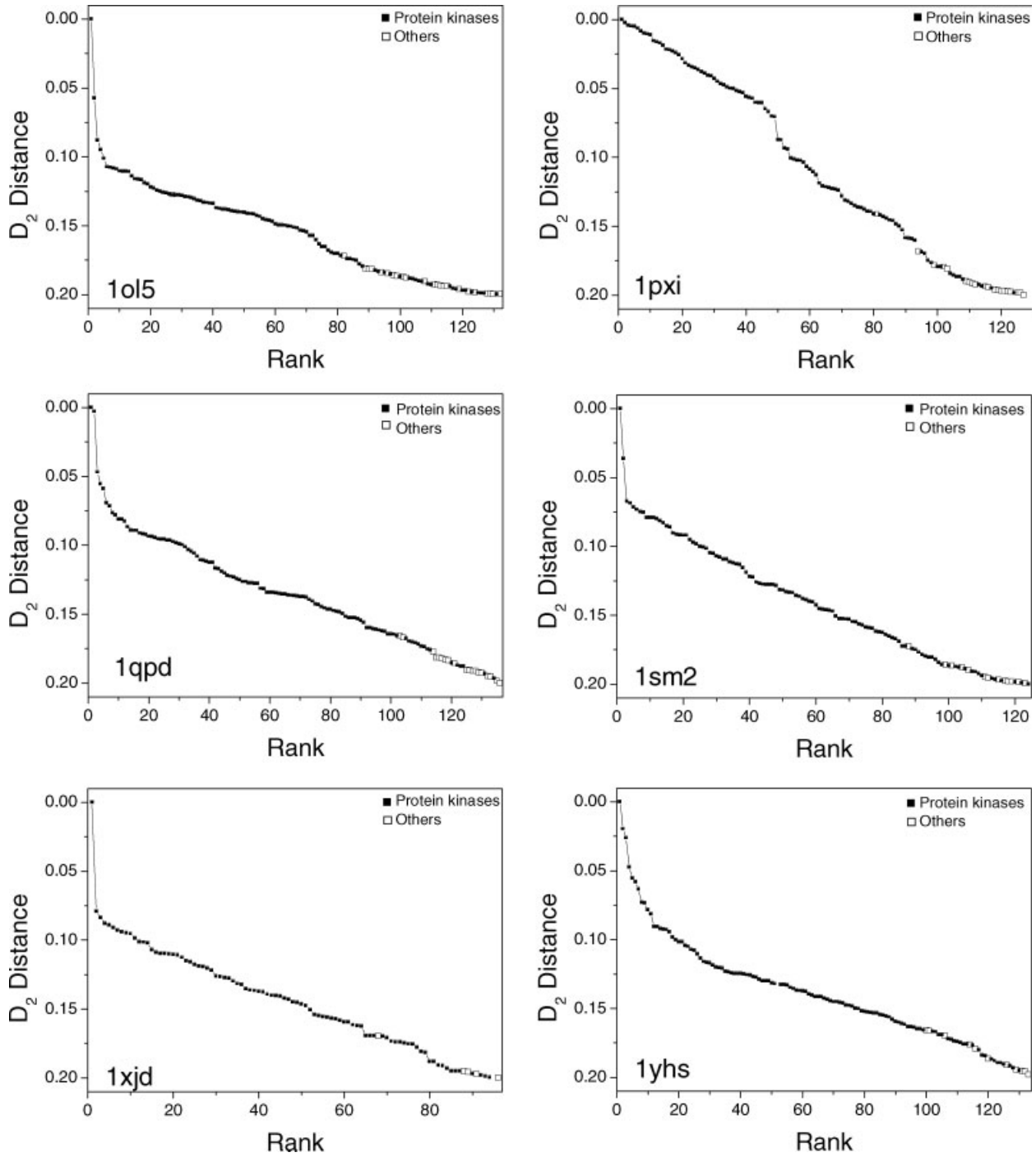


Figure 5

D_2 distance between the staurosporine binding site of six protein kinases (pdb entries 1ol5, 1pxi, 1yhs, 1qpd, 1sm2, 1xjd) and 6415 sc-PDB entries. Protein kinases encompass here any protein with an E.C. number equal to 2.7.10, 2.7.11, or 2.7.12. Alignments have been generated using slightly modified settings (Rn = 15, Tn = 3) of SiteAlign (See Methods). The D_2 distance was set to 1 for any binding site whose D_1 distance was higher than 0.6.

variation of D_1, D_2 scores as well as N_1, N_2 number of matched triangles against the observed atomic rmsd calculated from Cβ-atoms of cavity-lining residues clearly shows that our method tolerates large variation in atomic coordinates (up to 3 Å) without affecting much the D_2 score but maintaining significantly high N_2 number of matched residues [Fig. 8(B)]. Forcing the protein to begin unfolding by raising the temperature to 600 K

Table X
Similarity of 6415 Ligand Binding Sites, Annotated by Protein Family, to Various Staurosporine-Binding Sites

Targets	Reference	ROC	D_2^{mean}	D_2^{min}	D_2^{max}	D_2^{sd}
Protein kinases (n = 237)	1ol5	0.736	0.48	0.00	1	0.39
	1pxi	0.760	0.51	0.00	1	0.44
	1qpd	0.770	0.50	0.00	1	0.42
	1sm2	0.771	0.49	0.00	1	0.41
	1xjd	0.744	0.52	0.00	1	0.40
	1yhs	0.773	0.48	0.00	1	0.41
Other kinases (n = 329)	1ol5	0.475	0.79	0.12	1	0.32
	1pxi	0.489	0.94	0.21	1	0.18
	1qpd	0.497	0.92	0.20	1	0.21
	1sm2	0.476	0.92	0.19	1	0.21
	1xjd	0.486	0.88	0.19	1	0.26
	1yhs	0.488	0.90	0.19	1	0.24
Non-kinase	1ol5	0.580	0.63	0.17	1	0.36
ATP and ADP sites (n = 166)	1pxi	0.503	0.92	0.12	1	0.21
	1qpd	0.508	0.91	0.14	1	0.24
	1sm2	0.520	0.86	0.13	1	0.28
	1xjd	0.518	0.83	0.10	1	0.30
	1yhs	0.551	0.81	0.15	1	0.31
Others (n = 5683)	1ol5	0.405	0.77	0.12	1	0.33
	1pxi	0.411	0.94	0.14	1	0.18
	1qpd	0.402	0.94	0.09	1	0.20
	1sm2	0.409	0.91	0.15	1	0.23
	1xjd	0.415	0.88	0.16	1	0.26
	1yhs	0.395	0.90	0.13	1	0.24

The D_2 distance is used to rank entries by decreasing distance any of the eight references. If the D_1 score of the alignment is higher than 0.6, the D_2 distance was arbitrary set to 1 and corresponding sc-PDB entries ranked by decreasing original D_2 distance and if necessary, by decreasing D_1 score. A ROC score is computed for, within an in-house Pipeline Pilot workflow,[21] each occurrence of the above-mentioned protein families. Protein kinases have been defined by E.C. number (2.7.10.−, 2.7.11.−, 2.7.12.−), according to Manning *et al.*[32] Other kinases encompass all kinases (E.C. number beginning with 2.7) not selected by the above definition. ADP/ATP-binding sites refer to any non-kinase entry cocrystallized with a ligand of a HET code equal to ADP or ATP. Others represent all other entries not selected by any of the three preceding filters. N, D_2^{mean}, D_2^{min}, D_2^{max}, and D_2^{sd} are defined as in Table IV.

effectively results in a concomittent increase of the rmsd value and of both D_1 and D_2 scores which then pass the maximum distance cut-off (0.6 and 0.2 , respectively). In addition, the number of matched triangles (N_1, N_2) decreases significantly [Fig. 8(B)].

It is important to recall that the D_2 score does not mirror only sequence identity of matched residues but also side chain orientation, and the distance to the sphere centre of matched residues. Therefore, D_2 scores do not follow a theoretical model of extreme value distributions but a simpler normal distribution. For each possible match of 7–30 residues, the probability (P-value) to obtain by chance a D_2 score of 0.2 was estimated by comparing three randomly-generated cavity fingerprints (assuming of course only possible values for each of the eight reals defining the full vector) to 100,000 randomly-defined cavity fingerprints. In the worst possible scenario (match of only seven residues and D_2 score of 0.2), the P-value is 2.3×10^{-3}. In a typical scenario (alignment of 20 residue pairs and a D_2 score of 0.2), this P-value drops to 5.3×10^{-6}. Plotting the P-value against the N_2 number of matched triangles is given in the supplementary Figure 2.

We expect our method to be as accurate as existing cavity alignment programs but less prone to generate false positives (notably in comparing binding sites of

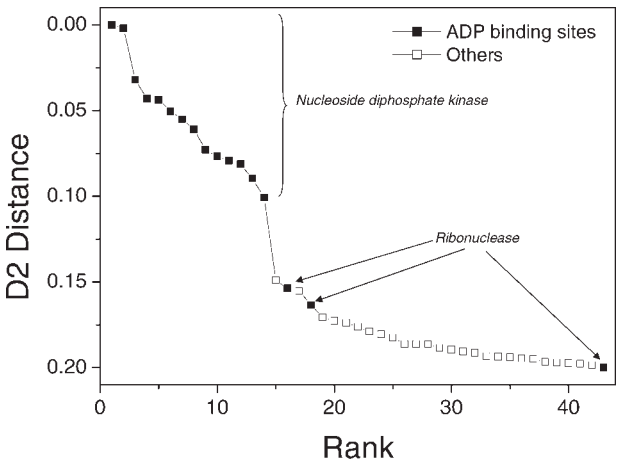


Figure 6
D_2 distance between the ADP binding site of a nucleoside diphosphate kinase (pdb entry 1nlk, ligand HET code: ADP) and 6415 sc-PDB entries. Binding sites are ranked by decreasing D_2 distance and classified in two groups: ADP-binding sites (filled squares), others (empty squares). Alignments have been generated using default settings of SiteAlign (See Methods). The D_2 distance was set to 1 for any binding site whose D_1 distance was higher than 0.6.

Table XI
Similarity of 6415 Ligand Binding Sites, Annotated by Protein Name, to a Nucleoside Diphosphate Kinase ADP-Binding Site (pdb Entry 1nlk, Ligand HET Code: ADP)

Name	Rank	ROC	n	D_2^{mean}	D_2^{min}	D_2^{max}	D_2^{sd}
Nucleoside diphosphate kinase	1	0.999	14	0.05	0.00	0.10	0.02
FK506-binding protein	2	0.820	13	0.52	0.17	1.00	0.37
Aldose 1-epimerase	3	0.753	19	0.64	0.14	1.00	0.37

D_2 distance is used to rank entries by decreasing distance to the 1nlk reference. If the D_1 score of the alignment is higher than 0.6, the D_2 distance was arbitrary set to 1 and corresponding sc-PDB entries ranked by decreasing original D_2 distance and if necessary, by decreasing D_1 score. A ROC score is computed, within an in-house Pipeline Pilot workflow,[21] for each occurrence of a protein name represented by at least five entries in the sc-PDB dataset. Proteins were filtered according to the minimal D_2 distance observed for all entries of that protein ($D_2^{min} \leq 0.20$) in the sc-PDB dataset and ranked by decreasing ROC score. Only proteins with ROC scores higher than 0.75 were considered as sufficiently close to the reference. n is the number of sc-PDB entries for a particular protein. N, D_2^{mean}, D_2^{min}, D_2^{max}, and D_2^{sd} are defined as in Table IV.

different sizes with a very low number of matched residues) and suitable for comparing homology models, a feature which has not been demonstrated up to now by any other method. In the current SiteAlign release, neither metals nor covalently-bound ligands are treated. Explicit treatment of metal ions is not necessary as far as all coordinating residues are present in the binding site definition (usually metal ions are characterized by specific arrangement of neighbouring residues). It could be a problem if the metal-binding site only partially overlaps the ligand-binding site. However, preliminary attempts addressing this issue on phosphodiesterases for example showed that it did not alter the quality of the structural alignment. Because we focused our work on druggable

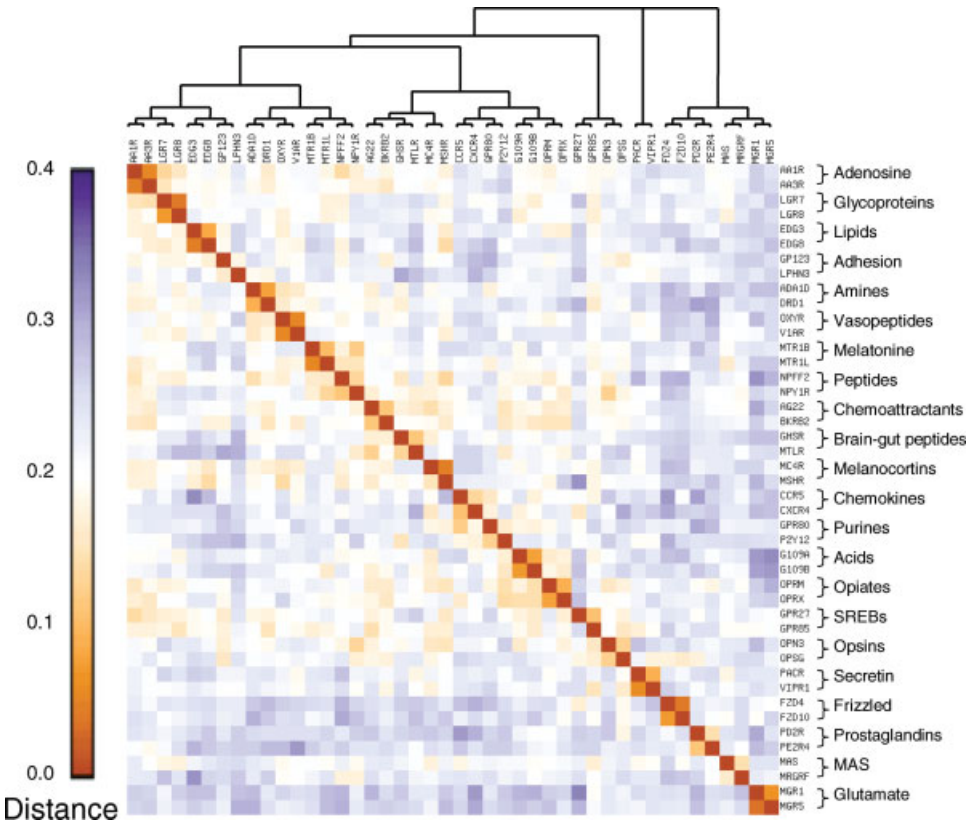


Figure 7
D_2 Distance matrix between transmembrane-binding sites of 44 human G Protein-coupled receptors (GPCR) originating from previously defined 22 clusters.[33] GPCR 3D models were built using the GPCRmod program[34] and ligand-binding sites defined from a list of 30 consensus transmembrane residues known to line the antagonist-binding site of most nonpeptide GPCR ligands.[33] The phylogenetic tree (top axis) was obtained from the D_2 distance matrix using the Neighbour-Joining method as implemented in Mega3.[35] GPCRs are labelled according to their SwissProt entry name and cluster names with their receptor representatives (two entries per cluster) indicated on the right axis. The figure was rendered with the matrix2png program.[36] [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

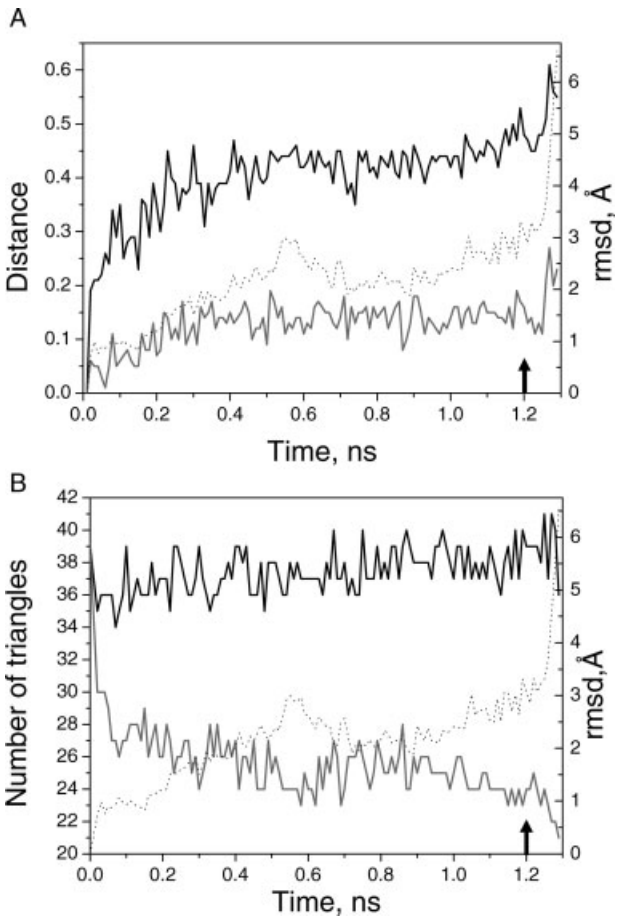


Figure 8

*Plotting the variation of SiteAlign descriptors in MD snapshots of a the smNACE enzyme[27] against the rmsd of C-β coordinates (cavity-lining residues only) to the starting coordinates. The starting protein coordinates[27] were embedded in a box of 12,410 TIP3P water molecules and simulated in AMBER8.0.[28] After 100 ps equilibration at constant pressure (1 atm), a 1-ns production trajectory was recorded at 300 K using standard AMBER8.0 parameters. After 1.2 ns (time indicated by a thick arrow), the temperature was gradually increased to reach 600 K after 1.3 ns. The ligand-binding site is composed of 39 residues located less than 6.5 Å away from the ligand (cyclic ADP-ribose) in the corresponding complex.[27] Snapshots were aligned to the first set of coordinates using standard parameters (see Methods) of SiteAlign. (**A**) Variation of D_1 (dark solid lines) and D_2 (gray solid line) distance scores [Eq. (3)] against variations of the rmsd (dotted line) on active site Cβ-atoms to the starting coordinates, (**B**) Variation of N_1 (dark solid lines) and N_2 (gray solid line) number of matched triangles [Eq. (3)] against variation of the rmsd (dotted line) on Cβ-atoms to the starting coordinates.*

protein–ligand binding sites, we made the initial choice to discard active sites with covalently-bound ligands.

How similar are similar ligand-binding sites?

The very first task in our study has been to define a distance threshold under which two sites can be estimated to be similar. For this purpose, a database of 376 related ligand-binding site pairs was specifically designed from the sc-PDB collection of druggable binding sites.[15] To simplify the biological annotation of cavities, only ligand-binding sites from enzymes with a well-defined E.C. number[22] have been selected. A pair of binding sites was extracted for each nonredundant E.C. number and chosen according to the molecular weight of the bound ligand. The two cavities cocrystallized with the lowest and highest molecular weight ligands were thus finally selected for each new E.C. number which ensures a collection of diverse active site pairs. A systematic comparison of the 376 pairs enables the visualization of distance score distributions (Fig. 2). Two distance scores (D_1 and D_2) are computed by SiteAlign. The D_1 score is computed from triangles receiving a cavity projection for at least one of the two sites to compare. It is therefore well suited to guide the structural alignment but also to measure the global similarity of the two binding sites under investigation. The D_2 score is computed only from triangles receiving a cavity projection for both sites and more suited to measure local similarities irrespectively of the cavity dimensions. D_1 and D_2 score distributions among the 376 aligned pairs are both bimodal. The major distribution (75% of the data) describes binding sites of comparable dimensions (ca. less than 15 residues difference) whereas the minor distribution (25% of the data) describes cavities of very different dimensions. Although the site pairs have been chosen to originate from catalytically equivalent enzymes, our procedure which selects binding sites from extreme molecular weight ligands still retrieves cavities of very different dimensions having very few residues in common (ca. 20–25% of the dataset, see Table III). We did not want to remove these pairs in our comparison since they allow a better definition of the applicability range of the alignment method.

Eighty percentage of the pairs present a D_1 score ≤ 0.6 or a D_2 score ≤ 0.2. Applying both thresholds in consensus afforded to predict 75% of the starting pairs as truly similar. However, does a good score really correspond to a good alignment? To answer this question, we compared the SiteAlign superimposition to a full protein sequence-based match ("align structures" Biopolymer routine of the SYBYL7.3 package[25]) which unambiguously aligns the two entry pairs bearing the same E.C. number and enables the identification of equivalent residues (conserved or very homologous amino acids sharing the same set of main chain coordinates). After a systematic survey of the 376 pairs, the SiteAlign fit could generally been tagged as "good" (similar to that produced by SYBYL) when the percentage of volume overlap (VO) of both protein-bound ligands, computed by an in-house SYBYL programming language script,[25] was higher than 30. Using this rough analysis, alignments could be theoretically described as either true positive (good score and good alignment), false positive (good score and bad alignment), true negative (bad score and bad alignment) or false negative (bad score and good alignment). As a matter of fact, only the first three cases could be

detected. 75% of the alignments correspond to true positives, with a good score ($D_1 \leq 0.6$ and $D_2 \leq 0.2$) and a good alignment (Table III). This corresponds to binding sites having at least 50% of their cavity-lining residues in common for which a very good fit can be obtained within 30–45 s. Applying the consensus score threshold, only 1.6% of false positive alignments (good score but bad alignment) could be detected. This situation corresponds to similar binding sites cocrystallized with ligands of very different sizes although some residues are in common. Last, the remaining 25% of the cases correspond to true negative alignments, meaning two ligand-binding sites having very few residues in common (usually less than five) and accommodating very different ligands and different binding modes. This situation is typically observed when comparing for example the inhibitor and the cofactor binding sites from the same protein.

To assess the sensitivity of our structural alignment method to the binding site definition, we generated ligand-binding sites of increasing sizes by varying the maximal distance between any ligand heavy atom and any binding-site heavy atom (4.5, 5.0, 6.5 Å). On a subset of 12 randomly-chosen active site pairs, D_1 and D_2 scores were remarkably constant (mean standard deviation of 0.056 and 0.019 for D_1 and D_2 distances, respectively) assuming a similar binding site definition for both entries to compare.

Binding site similarity across a gene family

Having quantified similarity thresholds, we next looked at the similarity of ligand-binding sites from proteins of the same family. For that purpose, serine proteases were an ideal choice for many reasons: (i) numerous protein-inhibitor X-ray structures are available in the Protein Data Bank[40]; (ii) many competitive serine protease inhibitors exhibit broad specificity for different members of the serine protease protein family[41] although they are characterized by different folds[42] (e.g., trypsin, subtilisin, α/β hydrolase); (iii) the substrate cleavage specificity can vary for a given fold illustrating subtle but functionally important differences in the catalytic site.[43]

A binding site for a bovine trypsin inhibitor (pdb entry 1aq7) was chosen as reference to measure its similarity to 6415 ligand-binding sites of the sc-PDB dataset, out of which 357 originate from serine proteases according to their E.C. annotation (3.4.21−).[22] A vast majority (92%) of binding sites passing similarity thresholds ($D_1 \leq 0.6$ and $D_2 \leq 0.2$) come from serine endopeptidases (Fig. 3). Interestingly, although proteases with the same fold and substrate cleavage specificity were retrieved first, proteins with other folds and cleavage preferences were also found similar enough to the reference as exemplified by the systematic calculation of ROC scores according to two protein annotations (Tables IV and V). The physicochemical significance of the match and related D_1, D_2 scores can be easily assessed by looking at the matched residue pairs and the physicochemical properties at the corresponding matched triangle (Table XII). Highly similar active sites present not only low D_1 and D_2 scores but also a match involving a higher proportion of very similar active site residues. A good indicator is to compute the percentage of matched residues presenting a local distance $D_{t,ij}$ [$D_{t,ij} = 1 - S_{t,ij}$; Eq. (2)] below the 0.2 threshold. In the case of a very similar binding site in another trypsin entry (e.g., 1cw5 entry, Table XII), 22 out of 26 residues contribute to the match and all 22 residues present a pair-wise distance to their matched counterpart in 1aq7 below the 0.2 distance threshold. For a more divergent binding site of still another serine protease (thrombin, pdb entry 1t4u), D_1 and D_2 scores are higher but a lower proportion of residues (21 out of 32) contribute to the match and an even lower proportion (10/21) present a pairwise distance to their counterparts in 1aq7 below 0.2.

Only serine proteases with a α/β hydrolase fold are significantly different from our reference bovine trypsin ligand-binding site. For example, a peptide binding site in proteinase K (pdb entry 1p7v) was found similar ($D_1 = 0.54$, $D_2 = 0.15$) to that of bovine trypsin (pdb entry 1aq7) despite very different folds and ligands. This is a typical case illustrating significantly different binding sites with a local similarity at residues defining the catalytic triad, and suggests that our alignment and comparison method may be used for the biological annotation of genomic structures. Among the 245 similar binding sites, eight describe proteins not annotated as serine endopeptidases. Of particular interest is the malonamidase A2 (pdb entries 1obl and 1ocl) which is a peptide amidase bearing a novel Ser-cisSer-Lys catalytic triad supposed to present catalytic mechanism reminiscent from that of serine proteases.[44] A ligand-binding site in a completely different enzyme (purine nucleoside phosphorylase, pdb entry 1lvu) was also found to possess a local similarity to that of bovine trypsin. The latter binding sites are globally different as acknowledged by their different shapes (Fig. 9) and a global distance ($D_1 = 0.59$) just above the upper limit, but they show a local strong similarity at seven conserved residues which are nicely matched by SiteAlign. This observation does not necessarily means that the corresponding ligands crossreact with both enzymes but more likely that this local subpocket similarity would enable the binding of common low molecular-weight fragments[45] to both targets.

Predicting off-targets by comparing binding sites

A basic assumption in chemogenomics is that similar ligands bind to similar active sites.[46] Predicting secondary targets for a given ligand may thus be achieved by finding cavities which are similar enough to the binding

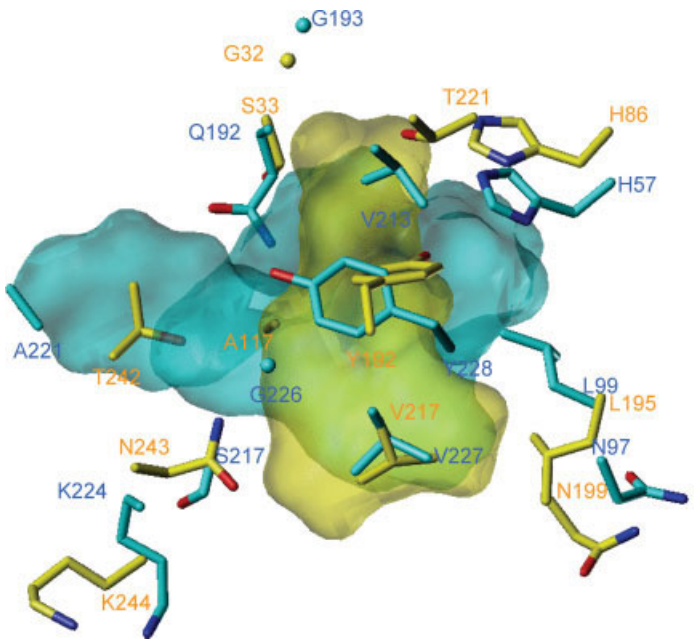
Table XII
Fine Details of SiteAlign Fit to 1aq7 Ligand-Binding Site (Size: 35 Residues)

Entry[a]	Protein	Size[b]	Rank[c]	D_1	D_2	N_1	N_2	Match[d]	$D_{t,ij}$[e]
1c5w	UtPA[f]	26	85	0.39	0.06	34	22	H57:H57[g]	0.000
								Y94:H99	0.116
								G148:S146	0.266
								Y172:Y172	0.004
								A183:A183	0.125
								D189:D189	0.004
								S190:S190	0.000
								C191:C191	0.000
								Q192:Q192	0.004
								G193:G193	0.004
								D194:D194	0.000
								G196:S195	0.129
								V213:V213	0.000
								S214:S214	0.187
								W215:W215	0.187
								G216:G216	0.004
								G219:G219	0.000
								C220:C220	0.066
								A221:L222	0.066
								G226:G226	0.125
								V227:V227	0.004
								Y228:Y228	0.000
1t4u	Thrombin	32	371	0.57	0.23	38	21	Y94:H79	0.129
								L99:Y126	0.462
								S96:Y83	0.412
								N97:W128	0.270
								N143:E232	0.308
								K145:C261	0.295
								Y172:I209	0.420
								A183:Y267	0.487
								D189:D262	0.020
								S190:A230	0.241
								C191:C231	0.016
								G193:G233	0.004
								D194:D234	0.008
								S195:S235[g]	0.141
								V213:V255	0.008
								S214:D135	0.191
								W215:E130	0.483
								K224:E259	0.550
								G226:G258	0.129
								V227:N131	0.337
								Y228:Y270	0.012

[a]PDB code.
[b]Number of cavity-lining residues.
[c]Rank, assessed by the D_2 distance score to 1aq7, among 6415 scPDB entries.
[d]SiteAlign matched residues (reference:target).
[e]Local distance [$1 - S_{t,ij}$; see Eq. (2)] of the macthed triangles.
[f]Urokinase-type plasminogen activator.
[g]Residue of the catalytic triad.

site of the primary target, as recently exemplified by the binding of cyclooxygenase-2 inhibitors to carbonic anhydrase.[47] We therefore chose three ligands (4-OHT, STAU, ADP) of increasing target promiscuity to check whether our comparison method would be able to recover their known targets by a simple comparison of ligand-binding sites.

4-OHT is known to bind to the family of estrogen receptors (estrogen subtypes α and β) as well as to the more recently identified estrogen-related receptor γ, for which high-resolution X-ray structure of the corresponding complexes are available.[48,49] Three 4-OHT binding sites from each of the three major targets were thus iteratively taken as reference for a systematic comparison to all sc-PDB binding sites. For all three *in silico* screens, the main targets were retrieved indeed among the top-ranked targets with even a sharp drop in similarity scores when shifting to a priori unknown off-targets (Fig. 4). The ERRγ site was found to resemble more the ERβ than the ERα ligand-binding site. As expected from the known major conformational changes occurring upon



1aq7 : 1lvu	$D_{t,ij}$
H57 : H86	**0.008**
S96 : P198	0.250
N97 : N199	**0.008**
L99 : L195	**0.141**
N143 : L35	0.233
Q192 : S33	**0.191**
G193 : G32	**0.054**
D194 : T114	0.500
G196 : S220	0.250
V213 : T221	**0.120**
G216 : V243	0.308
S217 : N243	**0.187**
A221 : T242	**0.137**
K224 : K244	**0.145**
G226 : A117	**0.016**
V227 : V217	**0.008**
Y228 : Y1912	**0.125**

Figure 9

SiteAlign superposition of the ligand binding site in bovin trypsin (1aq7 entry, ligand HET code: AEB, size: 35 residues) with the ligand binding site in calf spleen purine nucleoside phosphorylase (1lvu entry, ligand HET code: 9PP, size: 33 residues). Carbon atoms of trypsin and purine nucleoside phosphorylase are coloured in cyan and yellow, respectively. Nitrogen atoms are coloured in blue and oxygen atoms in red. The bound ligand envelopes computed by MOLCAD,[27] although not used in the alignment, are here presented as transparent surfaces to delineate the shape of both ligand-biding sites (1aq7 in cyan, 1lvu in yellow). The matched residues and their local pairwise distance $D_{t,ij}$ [$1 - S_{t,ij}$; see Eq. (2)] are tabulated on the right panel of the figure. Pairs with a local distance above 0.2 are indicated in bold and displayed for sake of clarity. The alignment ($D_1 = 0.59$, $N_1 = 39$; $D_2 = 0.15$, $N_2 = 19$) has been computed using default settings of SiteAlign (see Methods). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

activation at helix 12 of nuclear hormone receptors,[50] the three antagonist-binding sites were more similar to other antagonist-binding sites than to agonist-binding sites. Nevertheless, the alignment method is fuzzy enough to select most agonist-binding sites of the estrogen receptor α and β. The best compromise was achieved using the ERRγ site as reference with 19 out of the true 22 copies of all known estrogen receptors present in the selection, while retrieving 10/10 antagonist-binding sites and 9/12 agonist-binding sites. A rigorous ROC statistical analysis ranks both estrogen receptors at the top two positions of all three screens (Tables VI–VIII). Interestingly, several off-targets have been predicted for 4-OHT in all screens. Among those found at least twice are HIV proteases (type 1 and type 2) and the p38 MAP kinase 14. Local similarity between estrogen receptor and HIV protease is found by matching hydrophobic residues only [Fig. 10(A)]. Interestingly, one may notice some chemical similarity in the scaffold of known ligands of these two targets [Fig. 10(B)] which may be a consequence of the herein proposed similarity of both binding sites.

Another interesting target worth investigating is the p38 MAP kinase 14. 4-OHT and antiestrogens in general are known to activate p38 in a nongenomic mechanism that can be selectively blocked by p38 MAPK inhibitors.[30] The p38 MAP kinase 14 is indeed found among the top scoring proteins according to our ROC scoring in the three *in silico* screens (Tables VI–VIII). Inspecting the matched active sites indicate a good fit at several hydrophobic residues delineating common hydrophobic patches in both cavities [Fig. 11(A)].

We could not find direct evidence of 4-OHT binding to any of the other similar targets with high ROC score values. It is likely that the hydrophobic nature of the 4-OHT binding site in estrogen receptors explains this hypothesized promiscuity. In most if not all of these cases, the active site match was performed on hydrophobic residues exclusively, as previously shown for the p38 MAP kinase.

Among less populated proteins, a few ligand-binding sites were also found similar to each of the three references (Table IX). Interestingly, another nuclear receptor (retinoic acid RXR-α) was retrieved twice, likely because of the hydrophobic predominance of both binding sites. However, similarity to a ligand-binding site in a totally unrelated enzyme (catechol-1,2 dioxygenase, Table IX) involves polar residues and notably the important Asp351 which forms a salt-bridge with the ERα-bound 4-OHT [Fig. 11(B)], suggesting that both active sites may readily share similar ligands, although the envelopes of co-

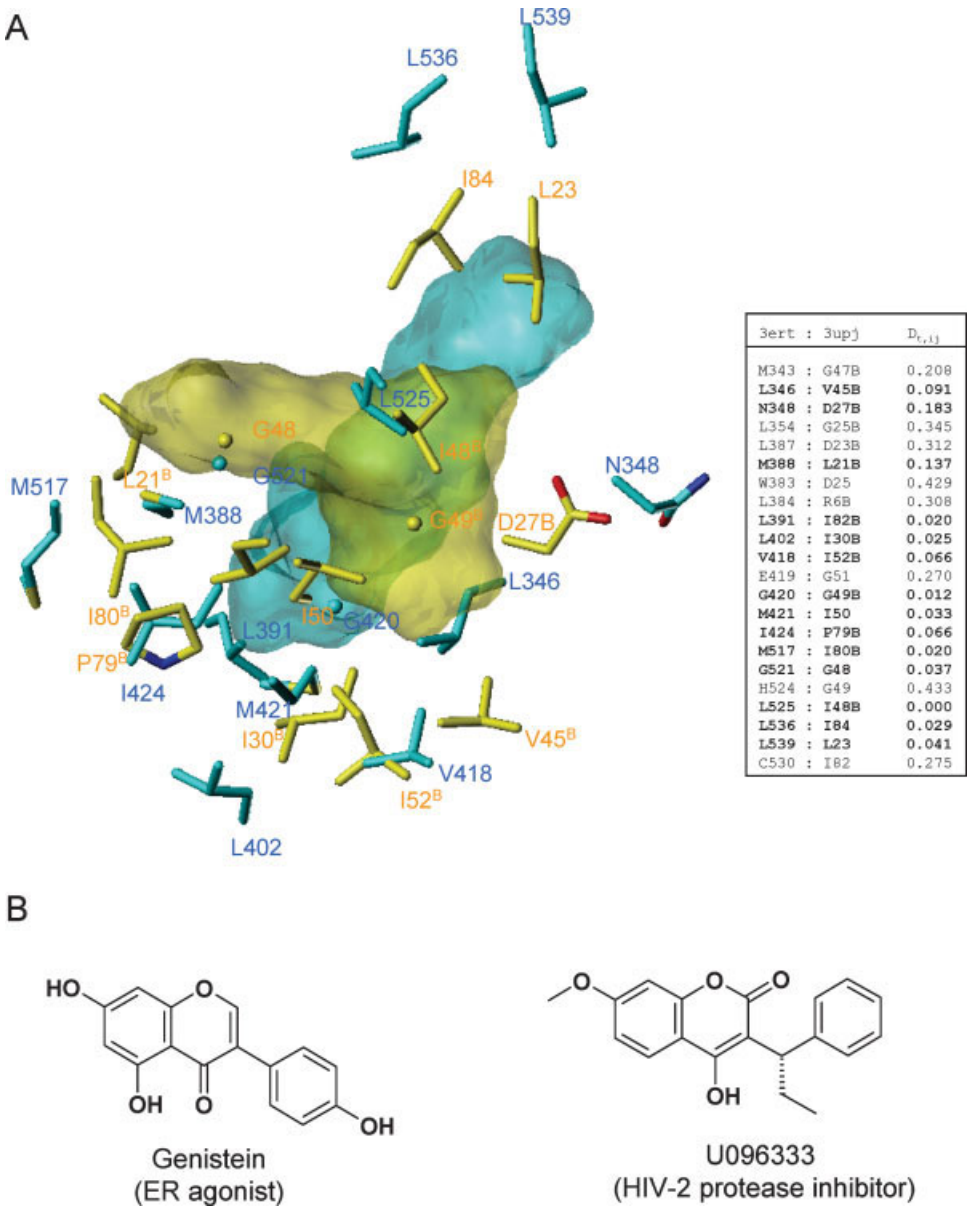
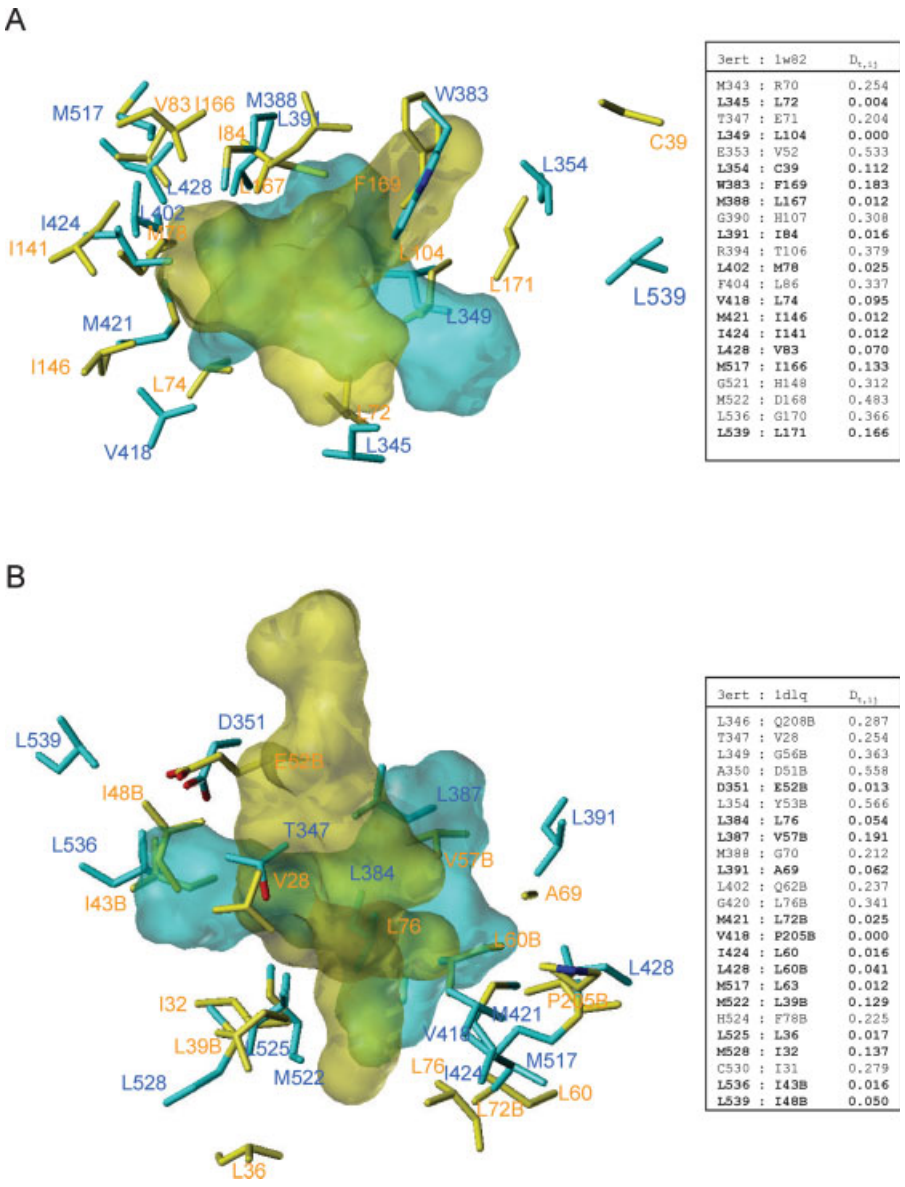


Figure 10

*(**A**) SiteAlign superposition of the 4-OHT binding site in human estrogen receptor α (3ert entry, ligand HET code: OHT, size: 35 residues) with the ligand binding site in human HIV-2 protease (3upj entry, ligand HET code: U03, size: 29 residues). Carbon atoms of estrogen receptor α and HIV-2 protease are coloured in cyan and yellow, respectively. Nitrogen atoms are coloured in blue and oxygen atoms in red. The bound ligand envelopes computed by MOLCAD,[32] although not used in the alignment, are here presented as transparent surfaces to delineate the shape of both ligand-biding sites (3ert in cyan, 3upj in yellow). The matched residues and their local pairwise distance $D_{t,ij}$ [$1 - S_{t,ij}$; see Eq. (2)] are tabulated on the right panel of the figure. Pairs with a local distance above 0.2 are indicated in bold and displayed for sake of clarity. The alignment ($D_1 = 0.56$, $N_1 = 42$; $D_2 = 0.15$, $N_2 = 22$) has been computed using default settings of SiteAlign (see Methods). (**B**) Structures of genistein (ER receptor agonist) and of U096333 (HIV-2 protease inhibitor). The Tanimoto similarity coefficient between both compounds, computed in Pipeline Pilot[21] according to MDL public keys, is 0.75. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]*

crystallized ligands are pretty different. Many secondary targets have been suggested for 4-OHT[51] but only a few of them (protein kinase C, calmodulin, COX-1, COX-2, quinone oxidoreductase) have been confirmed by *in vitro* binding assays. Out of the true validated targets for 4-OHT, none of their ligand-binding sites were selected in the three screens. Quite often, several ligand-binding sites from those targets were just above the D_2 distance upper threshold (e.g., 0.21 for the KAR-2 binding site in calmodulin pdb entry 1xa5). The ligand-binding site selection procedure may also explain these discrepancies since different binding sites (inhibitor and cofactor-bind-



3ert : 1w82	$D_{t,ij}$
M343 : R70	0.254
L345 : L72	**0.004**
T347 : E71	0.204
L349 : L104	**0.000**
E353 : V52	0.533
L354 : C39	**0.112**
W383 : F169	**0.183**
M388 : L167	**0.012**
G390 : H107	0.308
L391 : I84	**0.016**
R394 : T106	0.379
L402 : M78	**0.025**
F404 : L86	0.337
V418 : L74	**0.095**
M421 : I146	**0.012**
I424 : I141	**0.012**
L428 : V83	**0.070**
M517 : I166	**0.133**
G521 : H148	0.312
M522 : D168	0.483
L536 : G170	0.366
L539 : L171	**0.166**

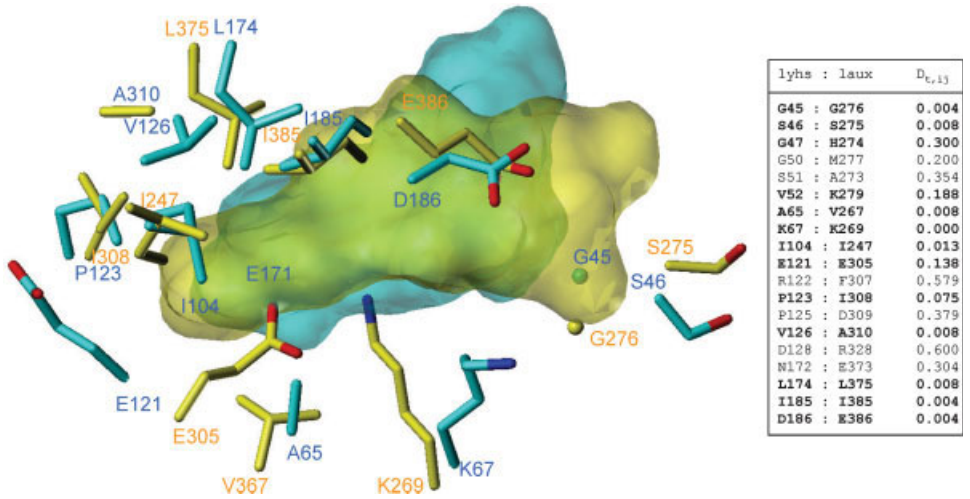
3ert : 1dlq	$D_{t,ij}$
L346 : Q208B	0.287
T347 : V28	0.254
L349 : G56B	0.363
A350 : D51B	0.558
D351 : E52B	**0.013**
L354 : Y53B	0.566
L384 : L76	**0.054**
L387 : V57B	**0.191**
M388 : G70	0.212
L391 : A69	**0.062**
L402 : Q62B	0.237
G420 : L76B	0.341
M421 : L72B	**0.025**
V418 : P205B	**0.000**
I424 : L60	**0.016**
L428 : L60B	**0.041**
M517 : L63	**0.012**
M522 : L39B	**0.129**
H524 : F78B	0.225
L525 : L36	**0.017**
M528 : I32	**0.137**
C530 : I31	0.279
L536 : I43B	**0.016**
L539 : I48B	**0.050**

Figure 11

*(**A**) SiteAlign superposition of the 4-OHT binding site in human estrogen receptor α (3ert entry, ligand HET code: OHT, size: 35 residues) with the ligand binding site in human p38 MAP kinase 14 (1w82 entry, ligand HET code: L10, size: 31 residues). Carbon atoms of estrogen receptor α and of p38 MAP kinase are coloured in cyan and yellow, respectively. Nitrogen atoms are coloured in blue and oxygen atoms in red. The bound ligand envelopes computed by MOLCAD,[32] although not used in the alignment, are here presented as transparent surfaces to delineate the shape of both ligand-biding sites (3ert in cyan, 1w82 in yellow). The matched residues and their local pairwise distance $D_{t,ij}$ [$1 - S_{t,ij}$; see Eq. (2)] are tabulated on the right panel of the figure. Pairs with a local distance above 0.2 are indicated in bold and displayed for sake of clarity. The alignment ($D_1 = 0.53$, $N_1 = 40$; $D_2 = 0.18$, $N_2 = 23$) has been computed using modified settings of SiteAlign (see Methods). (**B**) SiteAlign superposition of the 4-OHT binding site in human estrogen receptor α (3ert entry, ligand HET code: OHT, size: 35 residues) with the ligand binding site in catechol 1,2-dioxygenase (1dlq entry, ligand HET code: LIO, size: 38 residues). Carbon atoms of estrogen receptor α and catechol 1,2-dioxygenase are coloured in cyan and yellow, respectively. Nitrogen atoms are coloured in blue and oxygen atoms in red. The bound ligand envelopes computed by MOLCAD,[32] although not used in the alignment, are here presented as transparent surfaces to delineate the shape of both ligand-biding sites (3ert in cyan, 1dlq in yellow). The matched residues and their local pairwise distance $D_{t,ij}$ [$1 - S_{t,ij}$; see Eq. (2)] are tabulated on the right panel of the figure. Pairs with a local distance above 0.2 are indicated in bold and displayed for sake of clarity. The alignment ($D_1 = 0.55$, $N_1 = 44$; $D_2 = 0.17$, $N_2 = 24$) has been computed using default settings of SiteAlign (see Methods). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]*

ing sites) may be stored for the same protein. Last, 4-OHT may bind to still undiscovered allosteric binding sites for which no templates are available within our collection.

In a next series of validation screens, we took a more permissive ligand (staurosporine) known to bind in the ATP-binding site of most protein kinases.[31] Since many ATP binding sites not belonging to protein kinases are



1yhs : 1aux	$D_{t,ij}$
G45 : G276	**0.004**
S46 : S275	**0.008**
G47 : H274	**0.300**
G50 : M277	0.200
S51 : A273	0.354
V52 : K279	**0.188**
A65 : V267	**0.008**
K67 : K269	**0.000**
I104 : I247	**0.013**
E121 : E305	**0.138**
R122 : F307	0.579
P123 : I308	**0.075**
P125 : D309	0.379
V126 : A310	**0.008**
D128 : R328	0.600
N172 : E373	0.304
L174 : L375	**0.008**
I185 : I385	**0.004**
D186 : E386	**0.004**

Figure 12

SiteAlign superposition of the staurosporin-binding site in human Pim-1 kinase (1yhs entry, ligand HET code: STO, size: 32 residues) with the ligand binding site of rat synapsin-I (1aux entry, ligand HET code: SAP, size: 24 residues). Carbon atoms of Pim-1 and synapsin-I are coloured in cyan and yellow, respectively. Nitrogen atoms are coloured in blue and oxygen atoms in red. The bound ligand envelopes computed by MOLCAD,[32] although not used in the alignment, are here presented as transparent surfaces to delineate the shape of both ligand-biding sites (1yhs in cyan, 1aux in yellow). The matched residues and their local pairwise distance $D_{t,ij}$ [1 − $S_{t,ij}$; see Eq. (2)] are tabulated on the right panel of the figure. Pairs with a local distance above 0.2 are indicated in bold and displayed for sake of clarity. The alignment (D_1 = 0.56, N_1 = 36; D_2 = 0.17, N_2 = 19) has been computed using modified settings (Rn = 15, Tn = 3) of SiteAlign (see Methods). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

stored in the sc-PDB dataset, it was challenging to check whether our comparison method could discriminate ATP binding sites across different protein families. Six ATP competitive inhibitor-binding sites were selected from different protein kinases (cdk2, Aurora-1, Pim-1, Lck, IL-2 tyrosine kinase, protein kinase C) and iteratively compared to all sc-PDB binding sites. To precisely study the influence of the binding site selection, four entries (1yhs, 1qpd, 1sm2, 1xjd) were cocrystallized with staurosporine itself whereas the latter two entries had been cocrystallized with either another ATP-competitive inhibitor (1pxi) or ADP (1ol5). In all six screens, top-ranked binding sites were almost originating from protein kinases with only 15–20 exceptions out of about 120 selected entries (Fig. 5). About half of protein kinases stored in the sc-PDB dataset was missed by our selection protocol for two main reasons: (i) entries were ranked just above the minimum $D_1 - D_2$ score thresholds (ca. 30% of failure), (ii) entries were cocrystallized with peptide inhibitors binding to another site than the ATP-binding site (ca. 50% of failures).

Interestingly, three out of the six unexpected proteins which were selected in at least 50% of all screens were in fact protein kinases (Rio2, TAO2, Aurora-B) which were not biologically annotated at the time of the screening, thus demonstrating that functional annotation of genomic structures can be accomplished with SiteAlign. Among the three other unrelated proteins whose active sites are predicted to be close from the ATP-binding site of protein kinases, are the HIV-1 protease and the estrogen receptor α, which present both hydrophobic features. Avoiding this artefact may be obtained by computing a third similarity score accounting for polar residues, only. The last protein in the target list is synapsin-I, a synaptic vesical protein regulating neurotransmitter release and the organization of cytoskeletal architecture in the presynaptic terminal.[52] Synapsin-I exhibits a calcium-dependant ATP binding site,[53] which presents striking similar features to ATP-binding sites in protein kinases, especially at polar and charged residues (Fig. 12) which raises the possibility that protein kinase inhibitors may also bind to synapsin-I. Interestingly, a clear distinction of protein kinases from other kinases is achieved in all *in silico* screens, whatever the binding site reference (Table X). Other ADP/ATP ligand-binding sites are also statistically different from ATP-binding sites in protein kinases (Table X), confirming that ATP is quite permissive in recognizing protein binding sites of very different shapes.[18]

To confirm this observation, we last took ADP as a reference ligand. An ADP-binding site in a randomly-selected ADP binding protein (nucleoside diphosphate kinase) was chosen as reference for computing distance scores to all sc-PDB entries. As expected, very few binding sites were found similar to that of the reference (pdb entry 1nlk, HET code ADP), out of which the top-ranked entries were all nucleoside diphosphate kinases (Fig. 6). ADP-binding sites do not resemble each other since only

one ADP-binding protein (ribonuclease) is retrieved by computing binding site similarities (Fig. 6). Two features of ADP-binding sites may explain this observation: (i) ADP-binding sites present by analogy to AMP and ATP binding sites with very different shapes[18] and binding modes to their cognate ligands (in the 1nlk entry, the adenine ring is surprisingly stacking with Ile111 side chain); (ii) ADP frequently develops H-bonds to main chain atoms which are not taken into account in SiteAlign. The latter observation is also true in protein kinase inhibitors but does not prevent to clearly discriminate ATP-binding sites in protein kinases form other kinases or ATP/ADP binding sites (see above). A thorough statistical analysis computing ROC scores for every protein of our dataset only selected two relevant proteins beside the reference itself (Table XI). FK506-binding protein was notably retrieved because of the presence of an aromatic patch in its ligand-binding site which is common to that observed in the ADP-binding site of nucleoside diphosphate kinase.

Comparing ligand-binding sites in homology models

Because of their high dependency to 3D coordinates, most if not all ligand-binding site comparisons have focussed on high-resolution X-ray structures. However, comparing binding sites for proteins of unknown 3D structures would be highly desirable in many situations: (i) comparing ligand-binding properties of pharmacologically-important proteins not registered in the Protein Data Bank, (ii) predicting putative ligands of new genomic entries for which an homology model might be obtained.[54]

The first question was addressed by comparing the canonical antagonist-binding site[33] of human GPCRs, from previously published high-throughput homology models.[34] To avoid discussing a full and complex binding-site biased phylogenetic tree, we randomly selected two entries for each of the 22 clusters recently proposed in our group[33] to account for GPCR binding site diversity, and computed a 44 × 44 distance matrix according to the SiteAlign D_2 distance, while refining the 3D alignment previously obtained by our comparative homology modelling tool.[34] Since all 3D models have been generated from a restricted set of templates, D_2 distances are smaller than that previously observed in screening the sc-PDB dataset (Fig. 7). However, the method is suitable to unambiguously cluster each pair in its subfamily. Some clusters (e.g., Adhesion, Frizzled, Glutamate, MAS, Prostaglandins, Secretin) describe unique ligand-binding sites, corresponding to either allosteric regulatory sites[55] (e.g., Adhesion, Secretin, Frizzled, Glutamate) or very peculiar binding regions (Prostaglandin cluster).[56] Conversely, some other receptor clusters (e.g., Opiates, Chemoattractants) exhibit more resemblance in their transmembrane

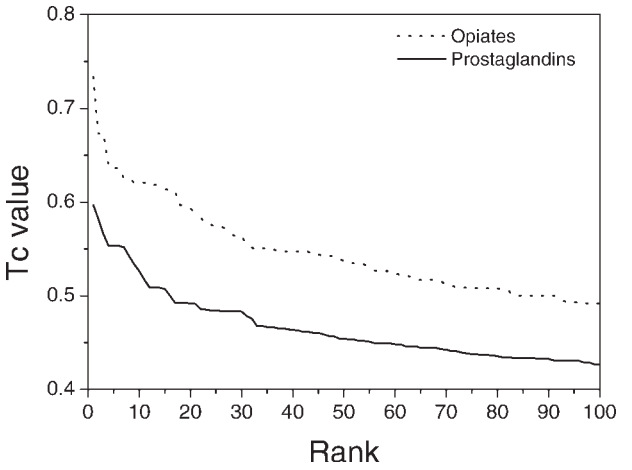


Figure 13

Chemical similarity of 912 known prostaglandin receptor ligands (solid line) and of 1285 known opiate receptor ligands (dotted line) to a panel of 17,906 other GPCR receptor ligands (with no binding data for either prostaglandin or opiate receptors) from the MDL Drug Data Report database.[57] Similarity is expressed as the Tanimito coefficient computed from SciTegic ECFP4 fingerprints.[21] For both activity classes, the closest other GPCR ligand to any of the reference ligand is selected and Tanimoto coefficients fused and ranked by decreasing values. Only the top 100-ranked other GPCR ligands are plotted for sake of clarity.

cavity to other GPCRs (Fig. 7). Although it is almost impossible to quantitatively relate binding site similarity to ligand similarity, we notice that ligands for specific binding sites (e.g., Prostaglandin receptors) usually have less similar compounds among known GPCR ligands than molecules binding to more permissive binding sites (e.g., opiate receptor ligands; Fig. 13). Interestingly, a few clusters of orphan or quasi-orphan receptors[33] (e.g., Acids, MAS, SREBs) show binding site similarity to some liganded GPCR clusters suggesting novel directions to identify low molecular-weight ligands for these new receptors.

CONCLUSIONS

We herewith present a novel method for comparing druggable ligand-binding sites which significantly differs from existing approaches in both fingerprinting cavity properties and measuring distances between two binding sites. Using a dataset of similar binding site pairs, we could define generic distance thresholds to estimate whether or not two binding sites may be considered similar. The two distance scores D_1 and D_2 are quantitative and qualitative measurements of the similarity, respectively and are therefore indicative of the global and local similarity between two cavities. Importantly, the method is insensitive to the definition of the size of the ligand-

binding site as far as the latter is structurally druggable, and is fuzzy enough to tolerate variations in atomic coordinates up to 3 Å. It should be stressed that detection of local similarity (low D_2 distance) between two cavities does not ensure crossreaction with the same ligand, a feature that probably requires a concomitant low D_1 score, a conservation of key polar interactions, and a flexible ligand. We are currently pursuing a systematic pair-wise comparison of 1600 nonredundant sc-PDB active sites to determine which percentage of binding sites from unrelated proteins verify these conditions. Two main applications of our comparison method in structural biology can be foreseen: guiding the functional annotation of new protein structures solved in structural genomic programs,[58] predicting off-targets of drug candidates by detecting similarity with a known ligand-binding site.[47] Since cavity descriptors are independent of rotameric states of cavity-lining side chains, the method is fuzzy enough to be applied to rough 3D homology models, thus extending the applicability range of binding site comparisons to a much wider biological space[54] than that spanned by current PDB structures.

ACKNOWLEDGMENT

We sincerely thank the Calculation Centre of the IN2P3/CNRS (Villeurbanne, France) for the allocation of computing time on the PC/Linux farm.

REFERENCES

1. Kubinyi H. Success stories of computer-aided design. In: Ekins S, editor. Computer applications in pharmaceutical research and development. New York: Wiley-Interscience; 2006. pp 377–424.
2. Harris CJ, Stevens AP. Chemogenomics: structuring the drug discovery process to gene families. Drug Discov Today 2006;11:880–888.
3. Grabowski M, Joachimiak A, Otwinowski Z, Minor W. Structural genomics: keeping up with expanding knowledge of the protein universe. Curr Opin Struct Biol 2007;17:347–353.
4. Kinoshita K, Furui J, Nakamura H. Identification of protein functions from a molecular surface database, eF-site. J Struct Funct Genomics 2002;2:9–22.
5. Schmitt S, Kuhn D, Klebe G. A new method to detect related function among proteins independent of sequence and fold homology. J Mol Biol 2002;323:387–406.
6. Jambon M, Imberty A, Deleage G, Geourjon C. A new bioinformatic approach to detect common 3D sites in protein structures. Proteins 2003;52:137–145.
7. Brakoulias A, Jackson RM. Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching. Proteins 2004;56:250–260.
8. Shulman-Peleg A, Nussinov R, Wolfson HJ. Recognition of functional sites in protein structures. J Mol Biol 2004;339:607–633.
9. Ferre F, Ausiello G, Zanzoni A, Helmer-Citterich M. Functional annotation by identification of local surface similarities: a novel tool for structural genomics. BMC Bioinformatics 2005;6:194.
10. Powers R, Copeland JC, Germer K, Mercier KA, Ramanathan V, Revesz P. Comparison of protein active site structures for functional annotation of proteins and drug design. Proteins 2006;65:124–135.
11. Gold ND, Jackson RM. Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. J Mol Biol 2006;355:1112–1124.
12. Kinoshita K, Nakamura H. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. Protein Sci 2003;12:1589–1595.
13. Kuhn D, Weskamp N, Schmitt S, Hullermeier E, Klebe G. From the Similarity Analysis of Protein Cavities to the Functional Classification of Protein Families Using Cavbase. J Mol Biol 2006;359:1023–1044.
14. Willett P. Similarity-based virtual screening using 2D fingerprints. Drug Discov Today 2006;11:1046–1053.
15. 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.
16. 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.
17. Hajduk PJ, Huth JR, Fesik SW. Druggability indices for protein targets derived from NMR-based screening data. J Med Chem 2005; 48:2518–2525.
18. Kahraman A, Morris RJ, Laskowski RA, Thornton JM. Shape variation in protein binding pockets and their ligands. J Mol Biol 2007; 368:283–301.
19. Liang J, Edelsbrunner H, Woodward C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci 1998;7:1884–1897.
20. Teague SJ. Implications of protein flexibility for drug discovery. Nat Rev Drug Discov 2003;2:527–541.
21. Pipeline Pilot release 6.0, SciTegic, San Diego, CA; 2007.
22. Bairoch A. The ENZYME database in 2000. Nucleic Acids Res 2000;28:304–305.
23. Wisconsin Package Version 10.2. Genetics Computer Group (GCG), Madison, WI; 2001.
24. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
25. Sybil release 7.3, TRIPOS, St. Louis, MO; 2007.
26. Triballeau N, Acher F, Brabet I, Pin JP, Bertrand HO. 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.
27. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
28. Igarashi Y, Eroshkin A, Gramatikova S, Gramatikoff K, Zhang Y, Smith JW, Osterman AL, Godzik A. CutDB: a proteolytic event database. Nucleic Acids Res 2007;35:D546–549.
29. Coward P, Lee D, Hull MV, Lehmann JM. 4-Hydroxytamoxifen binds to and deactivates the estrogen-related receptor gamma. Proc Natl Acad Sci USA 2001;98:8880–8884.
30. Seval Y, Cakmak H, Kayisli UA, Arici A. Estrogen-mediated regulation of p38 mitogen-activated protein kinase in human endometrium. J Clin Endocrinol Metab 2006;91:2349–2357.
31. Fabian MA, Biggs WH 3rd, Treiber DK, Atteridge CE, Azimioara MD, Benedetti MG, Carter TA, Ciceri P, Edeen PT, Floyd M, Ford JM, Galvin M, Gerlach JL, Grotzfeld RM, Herrgard S, Insko DE, Insko MA, Lai AG, Lelias JM, Mehta SA, Milanov ZV, Velasco AM, Wodicka LM, Patel HK, Zarrinkar PP, Lockhart DJ. A small molecule-kinase interaction map for clinical kinase inhibitors. Nat Biotechnol 2005;23:329–336.
32. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science 2002;298:1912–1934.

33. Surgand JS, Rodrigo J, Kellenberger E, Rognan D. A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors. Proteins 2006;62:509–538.
34. Bissantz C, Logean A, Rognan D. High-throughput modeling of human G-protein coupled receptors: amino acid sequence alignment, three-dimensional model building, and receptor library screening. J Chem Inf Comput Sci 2004;44:1162–1176.
35. Kumar S, Tamura K, Nei M. MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 2004;5:150–163.
36. Pavlidis P, Noble WS. Matrix2png: a utility for visualizing matrix data. Bioinformatics 2003;19:295–296.
37. Zhang Z, Grigorov MG. Similarity networks of protein binding sites. Proteins 2006;62:470–478.
38. Kedem K, Chew LP, Elber R. Unit-vector RMS (URMS) as a tool to analyze molecular dynamics trajectories. Proteins 1999;37:554–564.
39. Binkowski TA, Adamian L, Liang J. Inferring functional relationships of proteins from local sequence and spatial surface patterns. J Mol Biol 2003;332:505–526.
40. Kouranov A, Xie L, de la Cruz J, Chen L, Westbrook J, Bourne PE, Berman HM. The RCSB PDB information portal for structural genomics. Nucleic Acids Res 2006;34:D302–D305.
41. Nar H, Bauer M, Schmid A, Stassen JM, Wienen W, Priepke HW, Kauffmann IK, Ries UJ, Hauel NH. Structural basis for inhibition promiscuity of dual specific thrombin and factor Xa blood coagulation inhibitors. Structure 2001;9:29–37.
42. Rawlings ND, Morton FR, Barrett AJ. MEROPS: the peptidase database. Nucleic Acids Res 2006;34:D270–D272.
43. Bartoli L, Calabrese R, Fariselli P, Mita DG, Casadio R. A computational approach for detecting peptidases and their specific inhibitors at the genome level. BMC Bioinformatics 2007;8(Suppl 1):S3.
44. Koo HM, Choi SO, Kim HM, Kim YS. Identification of active-site residues in *Bradyrhizobium japonicum* malonamidase E2. Biochem J 2000;349:501–507.
45. Erlanson DA. Fragment-based lead discovery: a chemical update. Curr Opin Biotechnol 2006;17:643–652.
46. Rognan D. Chemogenomic approaches to rational drug design. Br J Pharmacol 2007;152:38–52.
47. Weber A, Casini A, Heine A, Kuhn D, Supuran CT, Scozzafava A, Klebe G. Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. J Med Chem 2004;47:550–557.
48. Shiau AK, Barstad D, Loria PM, Cheng L, Kushner PJ, Agard DA, Greene GL. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. Cell 1998;95:927–937.
49. Greschik H, Flaig R, Renaud JP, Moras D. Structural basis for the deactivation of the estrogen-related receptor gamma by diethylstilbestrol or 4-hydroxytamoxifen and determinants of selectivity. J Biol Chem 2004;279:33639–33646.
50. Moras D, Gronemeyer H. The nuclear receptor ligand-binding domain: structure and function. Curr Opin Cell Biol 1998;10:384–391.
51. Chen YZ, Zhi DG. Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. Proteins 2001;43:217–226.
52. De Camilli P, Cameron R, Greengard P. Synapsin I (protein I), a nerve terminal-specific phosphoprotein. I. Its general distribution in synapses of the central and peripheral nervous system demonstrated by immunofluorescence in frozen and plastic sections. J Cell Biol 1983;96:1337–1354.
53. Brautigam CA, Chelliah Y, Deisenhofer J. Tetramerization and ATP binding by a protein comprising the A, B, and C domains of rat synapsin I. J Biol Chem 2004;279:11948–11956.
54. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D, Shen MY, Kelly L, Melo F, Sali A. MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 2006;34:D291–D295.
55. Raddatz R, Schaffhauser H, Marino MJ. Allosteric approaches to the targeting of G-protein-coupled receptors for novel drug discovery: A critical assessment. Biochem Pharmacol 2007;74:383–391.
56. Stitham J, Stojanovic A, Merenick BL, O'Hara KA, Hwa J. The unique ligand-binding pocket for the human prostacyclin receptor. Site-directed mutagenesis and molecular modeling. J Biol Chem 2003;278:4250–4257.
57. MDDR version 2007-2, Elsevier MDL, San Leandro, CA; 2007.
58. Marti-Renom MA, Rossi A, Al-Shahrour F, Davis FP, Pieper U, Dopazo J, Sali A. The AnnoLite and AnnoLyze programs for comparative annotation of protein structures. BMC Bioinformatics 2007;8(Suppl 4):S4.