Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein−Ligand Binding Sites

Nathanaël Weill and Didier Rognan*

Structural Chemogenomics, Laboratory of Therapeutic Innovation, UMR 7200 CNRS-UdS (Université de Strasbourg), F-67400 Illkirch, France

Received September 16, 2009

Inferring the biological function of a protein from its three-dimensional structure as well as explaining why a drug may bind to various targets is of crucial importance to modern drug discovery. Here we present a generic 4833-integer vector describing druggable protein−ligand binding sites that can be applied to any protein and any binding cavity. The fingerprint registers counts of pharmacophoric triplets from the Cα atomic coordinates of binding-site-lining residues. Starting from a customized data set of diverse protein−ligand binding site pairs, the most appropriate metric and a similarity threshold could be defined for similar binding sites. The method (FuzCav) has been used in various scenarios: (i) screening a collection of 6000 binding sites for similarity to different queries; (ii) classifying protein families (serine endopeptidases, protein kinases) by binding site diversity; (iii) discriminating adenine-binding cavities from decoys. The fingerprint generation and comparison supports ultra-high throughput (ca. 1000 measures/s), does not require prior alignment of protein binding sites, and is able to detect local similarity among subpockets. It is thus particularly well suited to the functional annotation of novel genomic structures with low sequence identity to known X-ray templates.

INTRODUCTION

Fast and exhaustive comparison of protein−ligand binding sites is of crucial importance to predict the biological function of a protein and design safer drugs able to modulate their activity.[1] Until the 1990s, the Protein Data Bank,[2] which stores all publicly available protein three-dimensional (3-D) structures, was relatively poor in terms of target diversity. Thanks to remarkable advances in molecular and structural biology, high-resolution structural information on druggable protein−ligand binding sites[3] has reached a level of maturation allowing general rules about protein−ligand interaction patterns to be derived[4] and paradigms in drug discovery to be changed. For example, the sc-PDB data set of druggable protein−ligand binding sites from the Protein Data Bank currently stores about 5900 binding sites from more than 2000 unique proteins and 3000 unique ligands.[3,5] Exhaustive comparisons of protein−ligand binding sites are notably believed to directly influence two recent research areas: structural genomics and chemogenomics.

Outstanding efforts of genomic consortia worldwide contribute to significantly improve the structural description of the proteome[6] until nearly full coverage of the current UniProt database,[7] which is anticipated in ca. 15 years.[8] However, designing low-molecular-weight ligands from a protein 3-D structure is not straightforward notably for ligand-free proteins (apo form). Fast comparison of novel druggable cavities to ligand-annotated binding sites can be used to identify putative targets of existing ligands.[9,10] Besides structural genomics, chemogenomics is a novel multidisplinary research area aimed at identifying all possible ligands of all possible targets.[11] Since the target−ligand interaction matrix is still very sparse and cannot be fully completed experimentally, bio- and chemoinformatics approaches have been developed to predict ligand binding to a wide array of protein cavities.[12] In both applications, there is a basic need to compare, at a high throughput, protein−ligand binding sites. Assuming that similar ligands bind to similar cavities, function and ligands for a novel protein may be inferred from structurally similar liganded cavities. Since binding site similarities may be quite difficult to detect from amino acid sequences, 3-D computational methods for quantifying global or local similarities between protein cavities have been developed in the past decade.[13] All described methods follow the same three-step flowchart. First, the structures of the two proteins to compare are parsed into meaningful 3-D coordinates to reduce the complexity of the pairwise comparison. Typically, only key residues/atoms are considered and described by a limited number of points, which are labeled according to pharmacophoric, geometric, and/or chemical properties of their neighborhood. Second, the two resulting patterns are structurally aligned using notably clique detection[14,15] and geometric hashing[16,17] methods to identify the maximum number of equivalent points. Last, a scoring function quantifies the number of aligned features. In the pairwise comparison, a critical step is the search for the best possible structural alignment. Threfore, an erroneous alignment will lead to an underestimated similarity score. Moreover, the alignment step can be computer-intensive and can prevent generation of all-against-all distance matrices for several thousand PDB entries. Alignment-free quantification of binding site similarities is therefore highly desirable, and only a few algorithms have tacked this problem up to now. Starting from GRID

* Corresponding author phone: +33 3 68 85 42 35; fax: +33 3 68 85 43 10; e-mail: rognan@unistra.fr.

10.1021/ci900349y

molecular interaction fields,[18] FLAP[19] converts energy minimum points into a pharmacophore fingerprint registering all possible quadruplets of features. FLAP has been shown to be quite useful in mapping ligand space to target space by matching their corresponding pharmacophoric fingerprints,[19] but its application to compare binding sites has been limited to a small number of protein kinases and is hampered by the complexity of the fingerprint as well as the computing time needed to generate molecular interaction fields for a set of nonredundant probe atoms. PocketMatch[20] describes a binding pocket as a set of 90 lists of sorted distances between three sets of critical atoms (Cα, Cβ, and the centroid of the side chain) of any cavity-lining residue classified into five groups according to their physicochemical properties. Similarity between two binding sites is scored as the net average of the number of matching distances in the 90 lists as a fraction of the total number of distance elements in the bigger set. The method is fast (250 comparisons/s on a single CPU) and was shown to be able to detect cavity similarities among unrelated proteins that state-of-the-art protein alignment algorithms could not find.[20] A putative drawback of the method is that residues are grouped according to standard amino acid similarity rules and that protein cavities accommodating the same ligand sometimes show no strong conservation of cavity-lining amino acids.[21] To account for the real shape of the cavity, Yin et al.[22] recently proposed a novel frame invariant descriptor of binding pockets in which surface patches are encoded by a geometric fingerprint of 60 bins describing a distance-dependent distribution of surface curvatures. A significant limit of the approach is its sensitivity to small variations of atomic coordinates for the same cavity (e.g., apoenzyme vs holoenzyme).

We therefore think that there is still room for improving currently existing alignment-free cavity comparison algorithms along three lines: (i) insensitivity to moderate variations (<3 Å) of atomic coordinates, (ii) mapping pharmacophoric properties to cavity-lining residues to accurately describe all possible interactions with putative ligands, (iii) suitability to ultra-high throughput. We herewith introduce a novel method (FuzCav) which allows a fast and systematic comparison of protein−ligand binding sites. It takes its foundation from previous studies to efficiently encode molecular properties of low-molecular-weight compounds into pharmacophore fingerprints.[23] IUPAC defines a pharmacophore as "an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response".[24] Notably, three-point (feature) pharmacophores have been extensively used to rationalize structure−activity relationships,[25] to control the design of compound libraries,[26] and to find novel ligands.[27] Since a single compound may be described by several three-point pharmacophores, automated description of all possible three-point pharmacophores can be encoded into a pharmacophore fingerprint in which every possible triplet is binned according to interfeature distances.[28] Determining how many bin counts are shared by two compounds is thus a direct measure of their similarity. Four-point pharmacophores were shown to be even more accuracte but at the cost of an increasing complexity of the fingerprint.[28] Describing protein−ligand binding sites by pharmacophore fingerprints has already been described.[28,29] In the later methods, pharmacophoric tuplets (triplets, quadruplets) are derived from computed coordinates of probe atoms ideally interacting with protein atoms. In FuzCav, tuplets are directly defined from true protein atoms. A single atom (Cα) was chosen to make the fingerprint insensitive to rotameric states of a single chain, and the fingerprint was made of triplets and not quadruplets to speed up calculation and pairwise comparison. The cavity descriptor is a vector of 4833 integers, applicable to any protein class, insensitive to variation of rotameric states for binding-site-lining amino acids. It enables an alignment-free exhaustive comparison of protein cavities at an ultrahigh throughput (ca. 1000 comparisons/s) which is compatible with the generation of full distance matrices and the systematic screening of thousands of protein−ligand binding sites for similarity to a query.

METHODS

Setting Up Data Sets of Protein Binding Sites. All protein−ligand binding sites have been retrieved from the 2008 release of the sc-PDB database,[3] which currently comprises 5952 entries. An sc-PDB binding site is defined by any monomer (amino acid, ion, cofactor, prosthetic group) presenting one heavy atom closer than 6.5 Å from any heavy atom of the pharmacological ligand. The binding site definition is in line with docking algorithms[30] using protein−ligand bound coordinates to fix binding site boundaries. It notably prevents a too strong dependency of the cavity definition on the size and pocket occupancy of the bound ligand.

Five different protein data sets have been used throughout this study.

Data set 1 is composed of 769 pairs of nonredundant similar binding sites seeded with 769 pairs of nonredundant dissimilar binding sites. The similar pairs have been selected as follows. First, all entries from the sc-PDB database have been clustered according to their UniProt name,[7] leading to 911 clusters and 1204 singletons. Second, an all-against-all comparison of all active sites within each cluster was realized with the 3-D alignment tool SiteAlign[31] to generate a distance matrix. Algorithmic details of SiteAlign have been described elsewhere.[31] Briefly, eight topological and physicochemical attributes are projected from the Cα atom of cavity-lining residues to an 80-triangle-discretized polyhedron placed at the center of the binding site, thus defining a cavity fingerprint of 640 integers. 3-D alignment is performed by moving the sphere within the target binding site while keeping the query sphere fixed. After each move, the distance of the newly described cavity descriptor is compared to that of the query, the best alignment being that minimizing the distance between both cavity fingerprints. Two distances are used in SiteAlign. The d1 distance is suited to measure global similarities and is a sum of normalized distances between the eight descriptors on all indexed triangles with non-null values for either the query or the target. Previous benchmarking studies suggest that a d1 distance of 0.60 is a good threshold for discriminating similar from dissimilar binding sites.[31] The d2 distance is suited to measure local similarities and is a sum of normalized distances between the eight descriptors on all indexed triangles with non-null values for both the query and the target. Previous benchmarking studies suggest that a d2 distance of 0.20 is a good threshold for

discriminating similar from dissimilar binding sites. In the current study, two entries were randomly selected from each of the 911 clusters if their binding sites were found similar (d1 ≤ 0.6 and d2 ≤ 0.2). To avoid further processing issues, only cofactor-free binding sites were selected and finally led to 769 pairs of similar binding sites. The same number of dissimilar binding sites was randomly selected from the initial set of 911 clusters. Binding site pairs with an Enzyme Commission (E.C.) annotation differing at the first level were retrieved until the final number of 769 pairs was reached.

Data set 2 is composed of the entire sc-PDB archive[3] categorized into five classes. The first group is composed of 271 serine endopeptidase entries sharing a trypsin-like fold and a trypsin substrate cleavage specificity. The second group is composed of 17 other serine endopeptidase entries presenting a trypsin-like fold with a substrate cleavage specificity different from that of trypsin. The third group is composed of 11 serine endopeptidase entries with a subtilisin-like fold. The fourth group is composed of 13 entries with an α/β hydrolase fold. The last class is composed of the 5640 remaining sc-PDB entries. Fold and cleavage specificities were assigned from the CATH[32] and CUTDB[33] databases, respectively. The 1aq7 sc-PDB entry (trypsin in complex with the AEB ligand) was used as a reference to query this data set.

Data set 3 is composed of the entire sc-PDB categorized into four classes. The first class groups protein kinases (522 entries) according to the Enzyme Commission nomenclature (E.C. number 2.7.10.−, 2.7.11.−, 2.7.12.−, 2.7.13.−, or 2.7.99.−). The second class is made up of other kinases (181 entries with a UniProt name ending in "kinase"). The third class consists of 283 ATP- or ADP-binding sites which are not kinases (the ligand PDB HET code is "ATP" or "ADP"). The last class contains the remaining protein binding sites (4966 entries). Three ATP-binding sites of different sizes (34, 40, and 47 residues) from the human proto-oncogene serine/threonine-protein kinase Pim-1 (sc-PDB entries 1yi4, 3cy3, and 1yhs) have been used as references to query this data set.

Data set 4 was taken from a recent study from Aung et al.[34] It is made up of 126 nonredundant PDB entries with known adenine-binding motifs (34 entries, 18 folds) and 92 decoy proteins (21 different folds) of other functional types.

Data set 5 was retrieved from the previous work of Baroni et al.[19] on the FLAP methodology. It is composed of 23 ATP-binding sites of protein kinases from 4 different subfamilies (CDK2, GSK3β, LCK, P38).

FuzCav Fingerprint. The FuzCav fingerprint encodes a protein−ligand binding site by a vector of 4833 integers (Figure 1). Binding site residues are first listed and atomic coordinates of their Cα atoms annotated by six pharmacophoric properties of the corresponding amino acid (H-bond donor, H-bond acceptor, positive ionizable, negative ionizable, aromatic, aliphatic; Supporting Information Table S1). Each integer of the vector registers the count of unique pharmacophoric triplets (three properties and three related distances) occurring at binned interfeature distances. The distances between Cα atoms are currently discretized into five intervals (0−4.8, 4.8−7.2, 7.2−9.5, 9.5−11.9, and 11.9−14.3 Å). Please notice that the original first two intervals (0−2.4 and 2.4−4.8 Å) have been merged since the first one is too narrow to describe any distance between

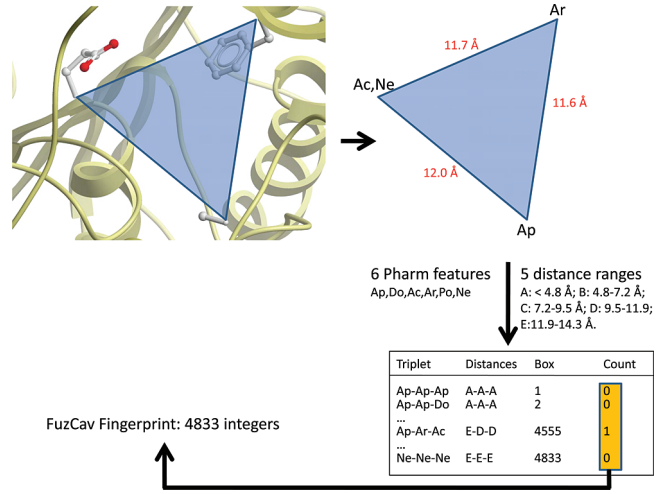

Figure 1. Flowchart for defining FuzCav cavity fingerprints. All possible pharmacophoric triplets (three points, three distances) are computed from Cα atomic coordinates of cavity-lining residues. Match counts for every possible case (triplet of pharmacophoric features separated by five possible distance ranges) are stored in a vector of 4833 integers.

two Cα atoms. Starting from the first interval, all triplet combinations are counted and stored, until the last interval is processed. To generate the shortest possible fingerprint, redundant triplets (property redundancy, isosceles and equilateral triangles) are removed. Last, the geometrical validity of the pharmacophoric triplet is checked by applying the triangle inequality rule stating that one distance cannot be longer than the sum of the two other distances.

Metrics. The similarity between two binding sites is computed as follows:

$$\mathrm{sim}(A, B) = \frac{a}{\min(\mathrm{nzA}, \mathrm{nzB})} \tag{1}$$

where a is the number of common non-null counts in both fingerprints and nzA and nzB are the numbers of non-null counts in fingerprints A and B, respectively. This similarity is symmetric and varies from 0 (not similar) to 1 (identical). To establish whether two binding sites are similar (classification), the best threshold for the similarity measure described above has been found using the maximum of the F-measure value:

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

with

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

and

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4}$$

where TP is the true positive rate, FN the false negative rate, and FP the false positive rate. The best threshold is found by the maximum of the F-measure curve when the threshold varies from 0 to 1 with an increment of 0.01.

The accuracy of classifications was assessed by computing the receiver-operating characteristic (ROC) curve[35] plotting

Table 1. Optimization of Interval Ranges and Numbers in the FuzCav Descriptor

unrestricted distance range		distance <14.3 Å		
intervals[a]	ROC[b]	number of intervals[c]	ROC	*F*-measure
1−7	0.95	2	0.95	0.89
1−6	0.96	3	0.97	0.92
1−5	0.96	4	0.96	0.91
1−4	0.97	5	0.98	0.94
1−3	0.96	5[d]	0.99	0.95
1−2	0.95			
1	0.50			

[a] Key: interval 1, 0−7.6 Å; interval 2, 7.6−10.1 Å; interval 3, 10.1−12.3 Å; interval 4, 12.3−14.3 Å; interval 5, 14.3−16.8 Å; interval 6, 16.8−20.0 Å; interval 7, >20.0 Å. [b] Area under the ROC curve[35] for classifying data set 1 entries. [c] Intervals are regularly distributed over a maximum range of 14.3 Å. [d] The first two intervals (0−2.4 and 2.4−4.8 Å) are merged.

the false positive rate versus the true positive rate. The 95% confidence intervals were calculated with the MedCalc software (MedCalc Software, 9030 Mariakerke, Belgium).

RESULTS AND DISCUSSION

Setting Up a Generic Cavity Fingerprint and a Similarity Threshold for Similar Binding Sites. The herein presented FuzCav cavity descriptor is inspired from a previous study on ligand pharmacophoric fingerprints which translates atomic coordinates of ligand atoms into all possible arrangements of three-point or four-point pharmacophores.[28] To introduce a certain level of fuzziness into the FuzCav descriptor, residue properties are matched to their corresponding Cα atoms and inter-Cα distances used for pharmacophore triplet definitions. Using Cα atoms to map binding site properties on either graphs[36] or vectors[31,37] presents the noticeable advantage to remove a strong dependency on atomic coordinates[31] (e.g., ligand-dependent rotameric state of a side chain, slight domain rearrangements up to 3 Å) without altering the quality of the 3-D alignment with respect to an all-atom match.[38] Having defined a representative data set of known similar protein−ligand binding site pairs and known dissimilar binding site pairs according to a previously reported distance metric specifically suited for druggable protein−ligand binding sites,[31] we computed all distances between Cα atoms of binding site residues and defined distance ranges for each interval to homogenize the distribution of inter-Cα distances over seven bins (Table 1). Using seven bins and three-point pharmacophores enables a clear distribution of similar and dissimilar binding sites in data set 1 with a high ROC value (Table 1). To optimize the number and range of distances used in the descriptor, we first reduced systematically the number of bins starting from the last one (interval 7, distance over 20.0 Å) until only the first distance range (interval 1, 0−7.6 Å) was encoded in the pharmacophoric triplet (Table 1). The best compromise was achieved with four bins coding for inter-Cα distances up to 14.3 Å (ROC value 0.97, Table 1). Having defined the maximum distance encoded in the fingerprint, we next varied the number of bins within this maximum distance by dividing the distance into two to five bins. The best compromise between ROC and *F*-measure values was obtained with five bins (Table 1, Figure 2A). It enables a nearly perfect separation of similar and dissimilar binding sites in data set 1 (area under the ROC plot of 0.99) and also allows precise determination of a similarity threshold for similar binding sites by systematically plotting statistical properties of classification models against increasing values of similarity thresholds (Figure 2B). The optimal classification was obtained with a similarity value of 0.16, which is specific for the FuzCav descriptor and the metric used (see the Methods) to measure distances between druggable binding sites. Since the current similarity threshold of 0.16 has been determined by comparing two sets of very similar and very dissimilar binding site pairs, it was next applied to various data sets of druggable protein−ligand binding sites to check its suitability for both screening and classification purposes.

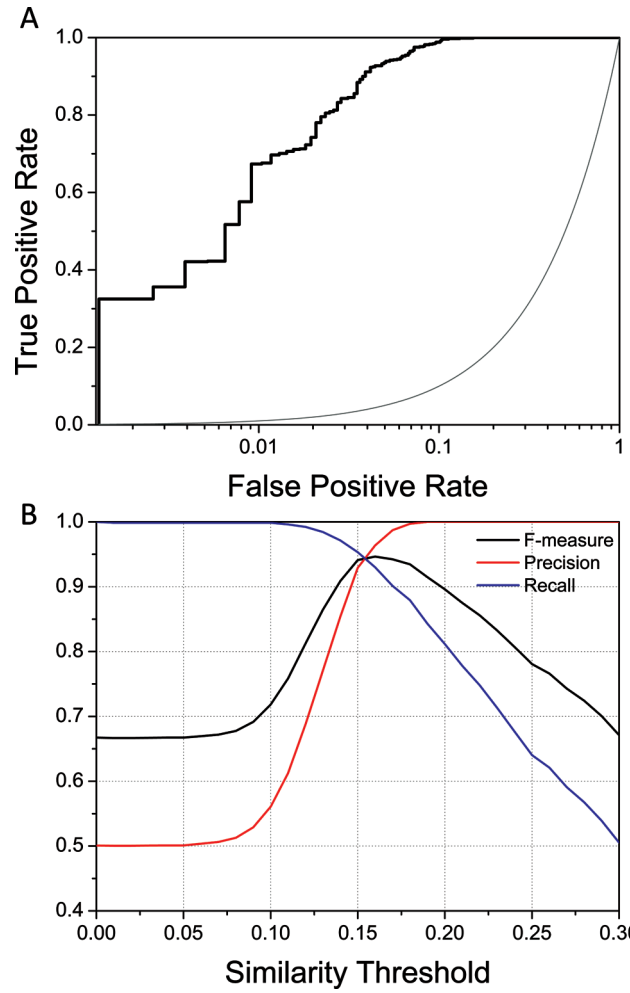

Figure 2. Distinction of similar from dissimilar protein−ligand binding site pairs (data set 1; see the Methods). (A) ROC plot (black line) obtained by sorting the pairwise similarity of 1538 binding sites (769 similar and 769 dissimilar pairs) by decreasing FuzCav similarity score. True positives are any of the similar pairs, whereas true negatives are any of the dissimilar pairs. The accuracy of random picking is represented by a gray line. (B) Variation of statistical parameters (*F*-measure, precision, recall) of a binary classification model (similar/dissimilar binding sites) for increasing similarity score thresholds.

Functional Annotation of a Protein Family (Serine Endopeptidases). Serine endopeptidases (data set 2) constitute a standard benchmarking data set for fold-independent protein alignment and comparison methods.[15,17,31,39−41] Although they all share a unique catalytic triad (Ser, Asp, His) with a single function, they exhibit different CATH

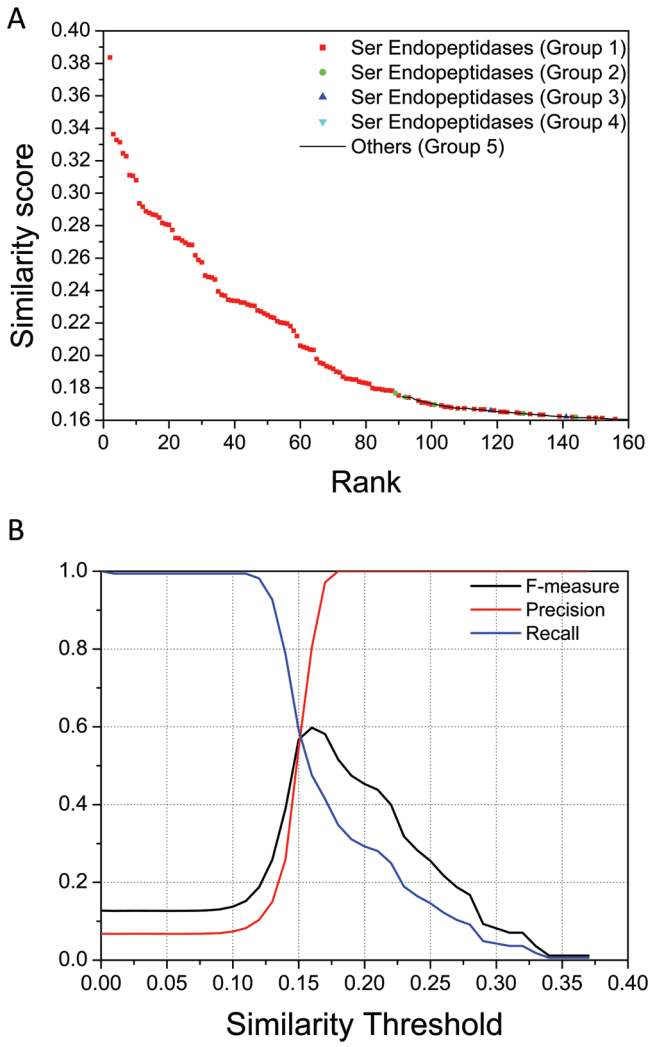

Figure 3. FuzCav similarity of 5951 sc-PDB binding sites to that of the 1aq7 entry (trypsin in complex with the ligand AEB). (A) Entries are sorted by decreasing similarity scores and displayed for scores above a similarity threshold of 0.16. (B) Variation of statistical parameters (*F*-measure, precision, recall) of a binary classification model (serine endopeptidases/other binding sites) for increasing similarity score thresholds.

folds[32] and substrate cleavage specificities.[33] In the current validation, all 5952 sc-PDB binding sites were classified into five groups (see the Methods) and fingerprinted as described above and their distances to the 1aq7 binding site (trypsin in complex with ligand AEB) evaluated. Figure 3 plots the distribution of these five group members by decreasing similarity to the 1aq7 binding site. The previously defined similarity threshold of 0.16 was used to retrieve putative "hits" (similar binding sites). A total of 157 out of 5951 entries pass the 0.16 similarity threshold. 75% of these entries are serine endopeptidases of group 1 (trypsin fold, trypsin substrate cleavage specificity) like the 1aq7 reference. The top-ranked 66 sites are trypsin−inhibitor binding sites with similarity scores to 1aq7 above 0.20. Five entries above the 0.16 similarity cutoff belong to group 2 (trypsin fold, other substrate cleavage specificity), and two entries belong to group 3 (subtilisin fold). Only 32 binding sites out of 5640 (0.56%) are false positive with similarity scores just above the threshold (in the 0.16−0.17 range) and functionally annotated as similar to the 1aq7 trypsin binding site (Figure 3A). False negatives are easy to explain from the nature of

Table 2. ROC Plot Values for Classifying sc-PDB Entries in Data Set 2

			FuzCav		BSAlign[a]	
group	fold	cleavage specificity	AUC[b]	95% CI[c]	AUC	95% CI
1	trypsin	trypsin	0.90	0.89−0.91	0.91	0.90−0.92
2	trypsin	other	0.78	0.77−0.79	0.67	0.66−0.68
3	subtilisin		0.65	0.64−0.66	0.67	0.66−0.68
4	α/β hydrolase		0.64	0.63−0.65	0.57	0.55−0.58
5	others		0.12	0.11−0.13	0.12	0.11−0.13

[a] Using default settings.[34] [b] AUC = area under the ROC curve.[35] [c] CI = confidence interval.

protein−ligand binding sites in serine endopeptidases in which several subpockets accommodate side chains of endogenous peptides to be cleaved. Depending on the subpocket occupancy by various inhibitors, the corresponding binding sites may only have a few residues in common (e.g., amino acids lining the oxyanion hole) with that of the 1aq7 reference.

Considering the entire data set, the classification according to the FuzCav fingerprint is excellent for serine endopeptidases of group 1 (ROC score of 0.90, Table 2) and good for entries of group 2 (ROC value of 0.78). Binding sites from subtilisin-like (group 3) and α/β hydrolases (group 4) exhibit FuzCav fingerprints different from that of trypsin. Other binding sites of the sc-PDB are found quite different from that of trypsin with a very weak area under the ROC curve (0.12). Interestingly, plotting *F*-measure values of a binary classification model (serine proteases vs others) against increasing similarity thresholds indicates a clear peak at 0.16 (Figure 3B), thus validating the previously chosen cutoff. The virtual screen was repeated with the recently described BSAlign algorithm.[34] BSAlign uses a graph representation of proteins with Cα atoms defining property-annotated vertices (solvent accessibility, physicochemical type, secondary structure type) and corresponding edges characterized by a distance (between Cα atoms) and an angle (between Cα−Cβ vectors of the corresponding residue-defining vertices). A subgraph isomorphism algorithm is used to detect the maximum common subgraph between two binding sites, and a similarity score is outputted depending on the number of aligned residues and the corresponding rms deviations. BSAlign was shown to be as efficient as SiteEngine, a state-of-the art 3-D binding site alignment tool,[17] with the advantage of being 14 times faster (ca. 7 s/comparison). When applied to classify the serine peptidase data set, BSAlign was found as good as FuzCav in detecting serine peptidases of group 1 (ROC score of 0.91, Table 2), but was clearly less accurate for group 2 members (trypsin fold and cleavage specificity different from that of trypsin). For the remaining three subgroups, very similar classification accuracies were obtained (Table 2).

Comparing ATP-Binding Sites across the sc-PDB Data Set. We next look at whether our generic cavity descriptor is able to distinguish binding sites for a very permissive ligand (ATP) known to bind to active sites exhibiting quite different shapes.[42] All sc-PDB binding sites were screened and ranked by decreasing similarity to the ATP-binding site of the protein kinase Pim-1 (PDB entry 3cy3). As to be expected, ATP-binding sites of protein kinases are statistically similar to each other but resemble neither ATP-binding sites of other kinases nor other ATP-

Table 3. ROC Plot Values for Classifying sc-PDB Entries in Data Set 3

	FuzCav		BSAlign	
group	AUC[a]	95% CI[b]	AUC[a]	95% CI[b]
protein kinases	0.89	0.88−0.90	0.53	0.52−0.54
other kinases	0.59	0.58−0.60	0.52	0.51−0.53
other ATP/ADP-binding sites	0.56	0.55−0.57	0.52	0.51−0.53
others	0.23	0.22−0.24	0.52	0.51−0.53

[a] AUC= area under the ROC curve.[35] [b] CI= confidence interval.

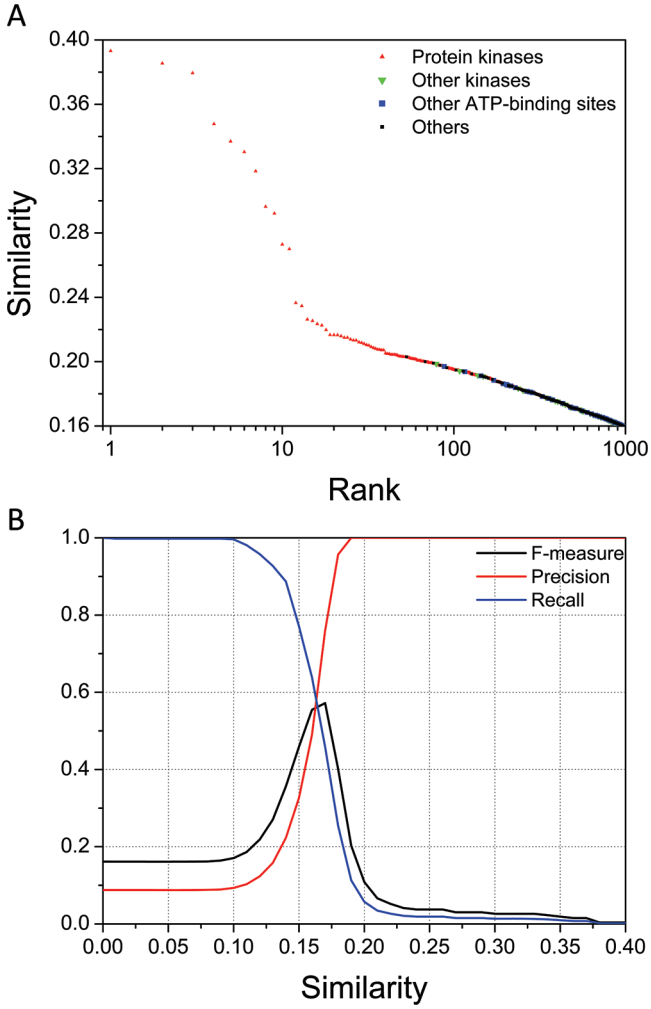

Figure 4. FuzCav similarity of 5951 sc-PDB binding sites to that of the 3cy3 entry (Pim-1 in complex with the ligand JN5). (A) Entries are sorted by decreasing similarity scores and displayed for scores above a similarity threshold of 0.16. (B) Variation of statistical parameters (*F*-measure, precision, recall) of a binary classification model (protein kinases/other binding sites) for increasing similarity score thresholds.

binding sites (see the ROC values, Table 3). About 17% of all sc-PDB binding sites (ca. 1000 entries) pass the 0.16 similarity threshold although the top 100 entries are almost exclusively ATP-binding sites of protein kinases (Figure 4A). The large hit list reflects both the number of protein kinases in the sc-PDB data set (522 entries) and the abundance of nucleotide-binding sites in the sc-PDB. False negatives at a threshold of 0.16 are either peptide-binding sites in protein kinases or larger binding sites for ATP-competitive inhibitors (e.g., a typical Cam kinase II inhibitor binding site has 25 residues more than the Pim-1 inhibitor site) exhibiting a more extended conformation than that of the reference-bound

Table 4. ROC[a] Plot Values for Classifying PDB Entries in Data Set 4

	ROC	
method	AUC[a]	95% CI[b]
BSAlign[34]	0.57	0.48−0.66
SiteAlign[31]	0.77	0.69−0.85
PocketMatch[20]	0.85	0.77−0.91
FuzCav	0.84	0.76−0.90

[a] AUC = area under the ROC curve.[35] [b] CI = confidence interval.

ligand. Repeating the same screening with the BSAlign algorithm did not allow separation of protein kinases from other classes (Table 3). A reasonable explanation for the observed BSAlign failure resides in the cavity composition (15 residues within 5 Å of the JN5 ligand) being different from that used in FuzCav (40 residues within 6.5 Å from the ligand). Since BSAlign scores the alignment on the number of matched residues, having a small-sized cavity as the reference may have penalized this method.

Since binding sites are defined from the bound ligand, the conformation of the ligand (folded, extended) significantly influences the cavity definition and thus the FuzCav fingerprint. To ascertain whether the virtual screen is relatively independent of the reference cavity, the same virtual screen was repeated by selecting other Pim-1 protein kinase entries as a reference with a smaller (PDB entry 1yi4, 34 residues) or larger (PDB entry 1yhs, 47 residues) active site. Fortunately, it did not change ROC statistics for the four subgroups of data set 3 (Supporting Information Table S2). A variation of up to ca. 20 residues for the same binding site will not dramatically influence the FuzCav fingerprint as long as the cavity contains more than 10−15 amino acids. Last, it should be noticed that "DFG-in" and "DFG-out" binding site conformations[43] could not be distinguished by the current cavity descriptor since it focuses on Cα atoms only. As for the serine peptidase data set, the selected threshold (0.16) also corresponds to the best possible *F*-measure of a binary classification model (protein kinase versus other targets, Figure 4B), therefore validating the usage of a universal cutoff for druggable protein−ligand binding sites.

Comparing Substructure-Binding Motifs. The fourth data set was taken from previous studies[17,34] and focuses on the difficult problem of finding local similarity among binding sites for the adenine substructural motif (in ligands such as ATP, ANP, FAD, NAD, etc.). In the data set, 34 adenine-binding proteins are seeded with 92 other proteins and their binding sites compared to the ATP-binding site in the 1atp PDB entry (cAMP-dependent protein kinase in complex with ATP). The main criterion of evaluation in previous studies was the number of adenine-binding pockets among the 15 top-ranked binding sites.[17,34] Whereas other algorithms (SiteEngine,[16] SiteAlign,[19] PocketMatch,[20] and BSAlign[22]) retrieve 9−11 true hits, our alignment-free descriptor is more accurate and selects 13 true adenine-binding proteins among the 15 top-ranked entries. The ability to distinguish, in a binary classification model, adenine-binding pockets from other cavities is significantly higher using the FuzCav descriptor than two 3-D alignment tools (BSAlign,[34] SiteAlign[31]) on the same data set (Table 4, Figure 5A). This observation suggests that a main reason

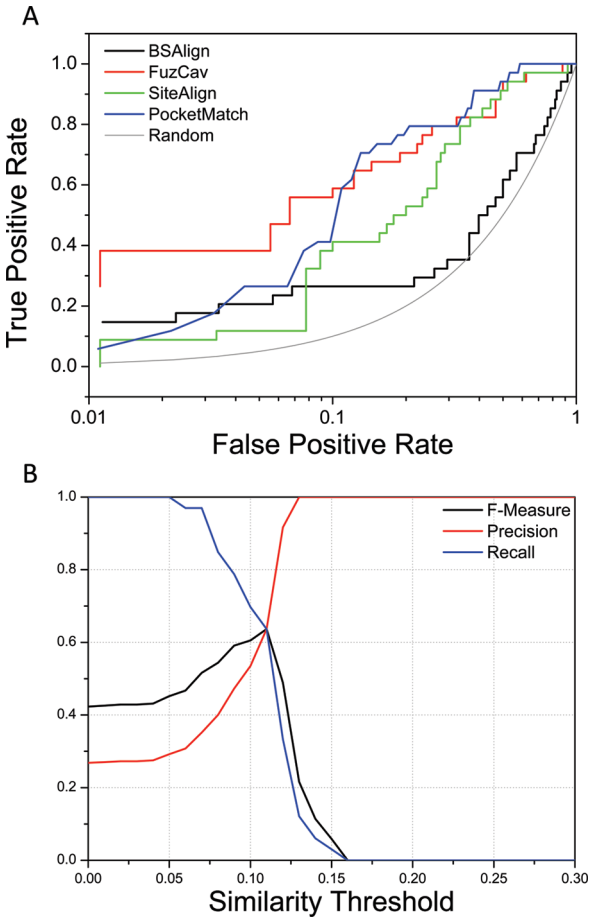

Figure 5. Distinction of adenine-binding pockets from other binding sites (data set 4; see the Methods). (A) ROC plot obtained by sorting the similarity of 126 binding pockets (34 adenine-binding pockets and 92 other binding sites) to the 1atp binding pocket by decreasing similarity score for three 3-D comparison tools (BSAlign,[34] SiteAlign,[31] PocketMatch[20]) and the herein presented FuzCav method. Random picking is represented by a gray line. True positives are any of the adenine-binding pockets, whereas true negatives are any of the other 92 cavities. (B) Variation of statistical FuzCav parameters (*F*-measure, precision, recall) for recovering adenine-binding pockets for increasing similarity score thresholds.

for the lower accuracy of alignment-dependent tools is the underestimation of similarity scores due to misalignment of unrelated proteins. We therefore applied a recently described alignment-free binding site comparison tool (PocketMatch)[20] to the same data set. PocketMatch uses a frame-invariant representation of binding sites as 90 lists of sorted distances capturing both the shape and the physicochemical properties of the cavity.[20] As to be expected for this peculiar data set, PocketMatch performs equally to FuzCav in terms of ROC score for classifying adenine-binding pockets from decoys (area under the ROC curve of 0.85, Table 4). PocketMatch is however still inferior to our method when enrichment in true positives among the top-scoring entries is considered (Figure 5A). This difference may be explained by the different physicochemical properties used to annotate protein Cα atoms. PocketMatch uses a standard amino acid similarity matrix to classify amino acids into five groups, whereas FuzCav uses a finer representation based on pharmacophoric properties. For example, Asp and Asn residues will be differently annotated in FuzCav (Asp, H-bond acceptor and negatively charged; Asn, H-bond acceptor and H-bond donor) than in PocketMatch (both residues being classified in the same group, i.e., group 2[20]).

Since many of the adenine-binding pockets in the current data set do not fall into the category of druggable protein−ligand binding sites,[3] it was interesting to check whether the optimal similarity threshold value of 0.16, previously derived from the analysis of similar druggable pockets, could be transferred to the present data set. Plotting statistical parameters of the classification (*F*-measure, recall, precision) as a function of the similarity threshold clearly shows an optimal value at 0.11 (Figure 5B), far below that observed for druggable cavities. Although the cavity descriptor has captured the relevant structural information, this benchmark clearly demonstrates that the similarity threshold score for classifying binding sites is not invariant and depends on the physicochemical properties of cavities to compare. Our experience with druggable protein−ligand cavities (molecular weight of the ligand in the 250−700 range, 30−60 active-site-lining residues, protein-bound ligand buried area above 70%) suggests that a similarity cutoff of 0.16 is robust enough to be used in various scenarios (classification, virtual screening).

All-against-All Comparison of sc-PDB Binding Sites. To identify similar binding sites in the absence of amino acid sequence similarities, an all-against-all comparison of 5952 sc-PDB binding sites was realized. Out of the 17 710 176 possible pairs, 55 100 (0.31%) have a similarity score above 0.16 and are thus potentially interesting. To avoid trivial matches resulting from inaccurate biological annotations, we only selected binding sites from proteins differing at the first level of the E.C. classification and retrieved the top 32 pairs with a similarity score higher than 0.225 (Table 5). Out of these 32 pairs, 6 share very similar ligands as indicated by their pairwise chemical similarity (Table 5), which thus provides an immediate validation for the comparisons. In most of the cases, no similarity between the corresponding protein amino acid sequences could be inferred by either a BLASTp[44] alignment or comparison of SCOP[45] families (Table 5). Noteworthy, some similarity between adenosine-binding patches could be observed in the 1x14−1jqh pair although the corresponding ligands (NAD in 1x14, ANP in 1jqj) occupy pockets of different shapes and volumes. Due to the discretization of distances between pharmacophoric features in five intervals, local similarity between patches of binding sites is better treated in FuzCav than in standard 3-D alignment tools which are more suited to detect global similarity patterns. Since FuzCav scores the sum of similar counts in all possible pharmacophoric triplet keys, an absence of match for a few residues is not detrimental to the similarity score. A total of 13 out of the 33 pairs exhibit a pairwise similarity of their ligands higher than 0.5, which reflects the detection of local similar patches in their cognate binding sites. Whether the latter may cross-react with similar ligands is difficult to predict and would necessitate experimental validation. The ultrahigh-throughput comparison of cavities (ca. 1000 pairwise measures/s on an Intel Pentium 3.4 GHz processor) opens the door to the systematic comparison of all druggable cavities in the PDB whether they are cocrystallized with a ligand or not. The possibility to find remote local similarities in binding site pairs should be useful for the functional annotation of genomic structures in the absence of known templates with globally similar cavities, a situation

Table 5. Top 32 Ranked sc-PDB Binding Site Pairs Differing at the First Level of the E.C. Annotation[a]

						BLASPp	
PDB1[b]	PDB2[c]	Name1[d] (SCOP[e] fold)	Name2[f] (SCOP[e] fold)	sim[g]	Tc[h]	*E*-value[i]	positives[j]
1oyj	*2pbj*	*glutathione S-transferase (52832)*	*prostaglandin E synthase 2 (52832)*	*0.25*	*1.00*	*0.007*	*37/75*
1xov	2b35	Ply protein (53162)	enoyl-[acyl-carrier-protein] reductase (51734)	0.24	0.01	0.20	23/47
1tvp	2hbl	endoglucanase 5A (51350)	exosome complex exonuclease (53066)	0.24	0.43	9.0	7/10
1od6	1hk8	phosphopantetheine adenylyltransferase (52373)	ribonucleoside-triphosphate reductase (51977)	0.24	0.50	0.14	23/54
1pzp	*1w2d*	*metallo β-lactamase (56600)*	*inositol-trisphosphate 3-kinase A (56103)*	*0.23*	*0.38*	*25.0*	*9/21*
2cun	1fgx	phosphoglycerate kinase (53747)	α-lactalbumin (53447)	0.23	0.41	7.5	6/13
1dia	2veg	methylenetetrahydrofolate dehydrogenase (53222)	dihydropteroate synthase (51350)	0.23	0.62	0.18	12/18
1fds	1j4h	estradiol 17β-dehydrogenase (51734)	peptidyl-prolyl cis−trans isomerase A (54533)	0.23	0.22	1.4	8/13
2j4h	2ecp	deoxycytidine triphosphate deaminase (51268)	glycogen phosphorylase (53755)	0.23	0.46	1.1	7/8
1ndc	1dia	nucleoside diphosphate kinase (54861)	methylenetetrahydrofolate dehydrogenase (53222)	0.23	0.49	1.2	20/47
1hiy	2j4h	nucleoside diphosphate kinase (54861)	deoxycytidine triphosphate deaminase (51268)	0.23	0.74	0.93	7/11
1pzp	*2jav*	*metallo β-lactamase (56600)*	*casein kinase I homologue 1 (56111)*	*0.23*	*0.41*	*8.1*	*11/26*
1qyx	1tsl	estradiol 17β-dehydrogenase (51734)	thymidylate synthase (55830)	0.23	0.30	0.44	31/83
2pvr	1onz	casein kinase I homologue 1 (56111)	tyrosine-protein phosphatase type 1 (52798)	0.23	0.39	4.6	10/22
2cun	1pgp	phosphoglycerate kinase (53747)	6-phosphogluconate dehydrogenase (51734)	0.23	0.85	0.50	12/17
1x3n	1pl6	propionate kinase (53066)	sorbitol dehydrogenase (51734)	0.23	0.49	1.4	8/14
1d7o	1hiy	enoyl-[acyl-carrier-protein] reductase (51734)	nucleoside diphosphate kinase (54861)	0.23	0.19	0.74	11/16
1tu7	1goy	glutathione S-transferase (52832)	ribonuclease (53932)	0.22	0.32	0.31	12/18
2b35	1uoo	enoyl-[acyl-carrier-protein] reductase (51734)	prolyl endopeptidase (54473)	0.22	0.11	20.0	6/8
2gss	*1nwl*	*glutathione S-transferase (52832)*	*tyrosine-protein phosphatase type 1 (52798)*	*0.22*	*0.27*	*1.3*	*6/8*
1x14	1jqh	NAD(P) transhydrogenase (52171)	basic fibroblast growth factor receptor 1 (56111)	0.22	0.86	1.1	9/14
1hk8	1fkg	ribonucleoside-triphosphate reductase (51997)	peptidyl-prolyl cis−trans isomerase A (54533)	0.22	0.40	0.017	13/21
1e8m	1p9p	prolyl endopeptidase (54473)	tRNA (guanine-N(1)-)-methyltransferase (75216)	0.22	0.53	7.8	8/14
1uoq	2iaj	prolyl endopeptidase (54473)	reverse transcriptase/ribonuclease H (53066)	0.22	0.55	1.8	22/45
1uoo	1kf0	prolyl endopeptidase (54473)	phosphoglycerate kinase (53747)	0.22	0.52	5.4	30/74
1e8n	1g99	prolyl endopeptidase (54473)	acetate kinase (53066)	0.22	0.56	1.1	36/86
2vqm	2iaj	histone deacetylase 8 (52767)	reverse transcriptase/ribonuclease H (53066)	0.22	0.51	2.8	11/19
2jav	1dia	casein kinase I homologue 1 (56111)	methylenetetrahydrofolate dehydrogenase (53222)	0.22	0.44	0.12	14/29
3eng	1hrk	endoglucanase 5A (51350)	ferrochelatase (53799)	0.22	0.38	0.66	41/104
1fki	2b35	peptidyl-prolyl cis−trans isomerase A (54533)	enoyl-[acyl-carrier-protein] reductase (51734)	0.22	0.18	0.82	17/42
2g2f	1j4h	proto-oncogene tyrosine kinase Src (56111)	peptidyl-prolyl cis−trans isomerase A (54533)	0.22	0.45	1.4	7/9
1x14	*2olk*	*NAD(P) transhydrogenase (52171)*	*amino acid ABC transporter (52539)*	*0.22*	*0.85*	*7.6*	*6/9*

[a] Pairs sharing a similar 3-D fold are displayed in italics. [b] PDB identifier of the first protein in the pair. [c] PDB identifier of the second protein in the pair. [d] Name of the first protein. [e] SCOP[45] class number. [f] Name of the second protein. [g] FuzCav similarity of the ligand-binding sites. [h] Chemical similarity (Tanimoto coefficient) of the cocrystallized sc-PDB ligand, measured from MACCS public keys in Pipeline Pilot.[50] [i] BLASTp[44] expectation value for sequence stretches producing a significant alignment. [j] Number of similar residues/total length of the alignment.

in which most 3-D alignment tools fail to find any local similarity.[46]

Alignment-Free versus Alignment-Dependent Binding Site Comparisons. A major issue with most 3-D binding site comparison tools is that they require first alignment of structures before computation of a similarity score. An incorrect alignment of two sites will dramatically underestimate the similarity score, as suggested by results previously obtained when benchmarking different tools using data set 4 of adenine-binding pockets (Table 4, Figure 5).

To estimate the frequency of such an event, we systematically compared our alignment-free descriptor with that of an alignment-dependent 3-D method (SiteAlign).[31] Like the FuzCav fingerprint, SiteAlign also maps pharmacophoric properties to Cα atom coordinates; therefore, the influence of the prior alignment on the similarity score can be directly estimated. The previous 522 ATP-binding sites of protein kinases from data set 3 were compared to the ATP-binding site in Pim-1 kinase (PDB entry 3cy3) in both FuzCav and SiteAlign. A total of 75% of the ATP-binding sites were found similar to that of Pim-1 according to FuzCav (score ≥ 0.16), whereas SiteAlign only finds some similarity ($d2 \leq 0.20$)[31] in 48.5% of the comparisons. Plotting FuzCav versus SiteAlign similarity scores permit four situations to be distinguished (Figure 6A). In 42% of the comparisons, both tools agree that the two sites are similar (lower right quarter of the plot). In 18.4% of the cases, the two algorithms agree that both sites are dissimilar (upper left quarter of the plot) notably for screened binding sites of large dimension (over 60 residues, Figure 6A). In 33.1% of the comparisons, SiteAlign fails to find any similarity whereas FuzCav does (upper right quarter of the plot). These cases correspond to SiteAlign misalignments underestimating the corresponding similarity score. A typical example is found with the comparison between 3cy3 (Pim-1 kinase bound to inhibitor JN5) and 1uv5 (Gsk-3β bound to inhibitor BRW) binding sites. The SiteAlign fit is not optimal with respect to the sequence-based match of Cα atoms (Figure 6B) and thus underestimates the SiteAlign similarity score. The opposite situation (good SiteAlign score and bad FuzCav score) occurs much less frequently (6.4% of the test cases, Figure 7) and almost exclusively when one of the two binding site pairs exhibits many more charged residues than the other (e.g., 3cys vs 2jdr; Figure 6C). Since charged features are less frequent in FuzCav, significant variations in their number may lead to quite different pharmacophoric triplet counts. Conversely, the SiteAlign d2 score registers similarity for pairs of matched binding site residues and is thus insensitive to extra residues missing a counterpart in one of the two cavities to compare.

Sensitivity of the FuzCav Fingerprint to the Size of the Binding Site. The FuzCav descriptor registers pharmacophoric triplets and is therefore dependent on the way binding site residues are selected, notably the maximal distance between a ligand and the corresponding cavity-lining residues. In the current study, binding site residues are selected according to a maximal distance threshold of 6.5 Å

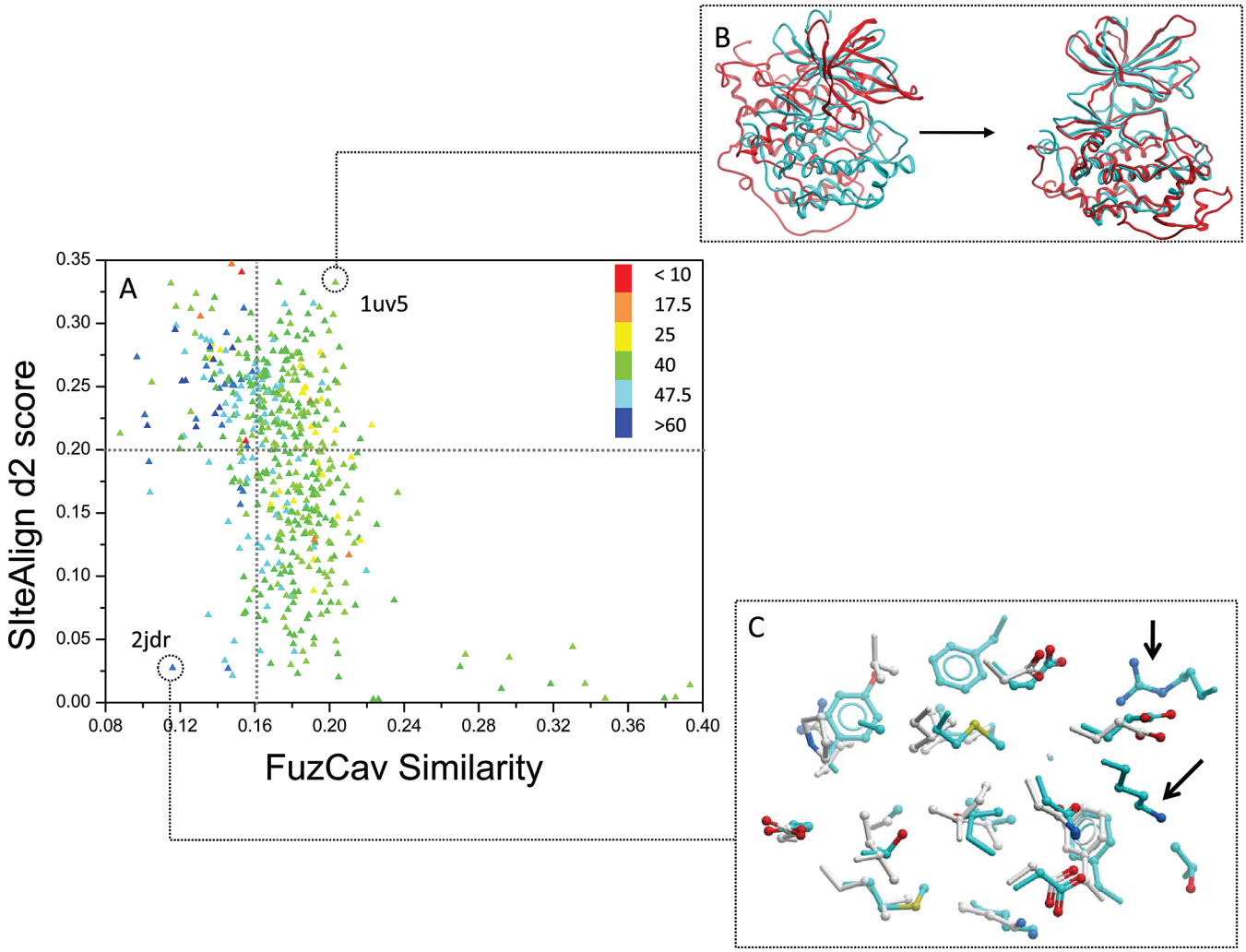

Figure 6. (A) FuzCav versus SiteAlign[31] similarity values of 521 ATP-binding sites of protein kinases to the ATP-binding site of human Pim-1 kinase (3cy3 PDB entry). Similarity thresholds for similar binding sites are indicated for FuzCav and SiteAlign by dotted lines. Data are colored according to the size of the binding site (number of residues, color ramp in the upper right part). (B) SiteAlign failure in measuring the similarity between 3cy3 (cyan) and 1uv5 (red) binding sites. The SiteAlign alignment is proposed in the left panel. The optimal sequence-based alignment is proposed in the right panel. (C) FuzCav failure in measuring the similarity between 3cy3 (white sticks) and 2jdr (cyan sticks) binding sites. Black arrows indicate two positively charged residues not present in the 3cy3 reference.

to any ligand atom. To investigate the influence of the size of the binding site (number of amino acids) on the variability of the FuzCav descriptor, the biggest sc-PDB binding site (2gmj entry) was iteratively trimmed residue by residue and a FuzCav similarity matrix between all possible binding sites was computed (Figure 7). The corresponding heat plot clearly shows that the FuzCav similarity rapidly decreases for small-sized binding sites (less than 10 residues). For a binding site of 13 residues, similarity to much bigger sites (30 residues more) can still be detected. For a binding site lined by at least 20 amino acids (98.7% of all sc-PDB protein−ligand binding sites),[3] the descriptor is fuzzy enough to recover similarity with binding sites having up to 70 residues (Figure 8). The FuzCav descriptor is thus insensitive to the binding site definition for a vast majority of druggable protein−ligand binding sites.

Sensitivity of the FuzCav Fingerprint to Variations of Atomic Coordinates. Existing 3-D alignment programs, notably those using property-mapped molecular-surface-based descriptors are notoriously sensitive to variations of atomic coordinates of binding site residues.[21] Since active site comparison algorithms may be used to predict the function of novel protein structures solved within structural genomic consortium initiatives, it is important that comparison algorithms are still suited to infer a putative function from a ligand-free protein structure (apo structure). To verify this property with the herein presented tool, five different proteins of known X-ray structure were selected in both ligand-bound and ligand-free states. The proteins were selected to mimick different conformational changes upon ligand binding: (i) very modest conformational rearrangement of the binding cavity upon ligand binding (estrogen-related receptor γ), (ii) disclosure of a subpocket (aldose reductase), (iii) side chain motions at the target−inhibitor interface (uracil DNA-glycosylase inhibitor), significant to large motions (HIV-1 protease, cell division protein kinase 2, glucokinase). Measuring the binding site similarity of apo and holo forms of the six targets shows that FuzCav is quite robust in detecting true similarity. In none of the cases does the computed similarity fall below the 0.16 similarity threshold score, despite quite significant motions in some binding sites (Figure 8A). An illustrative example of FuzCav tolerance to active site motions is given by human glucokinase, which upon binding of its substrate (α-D-glucose)

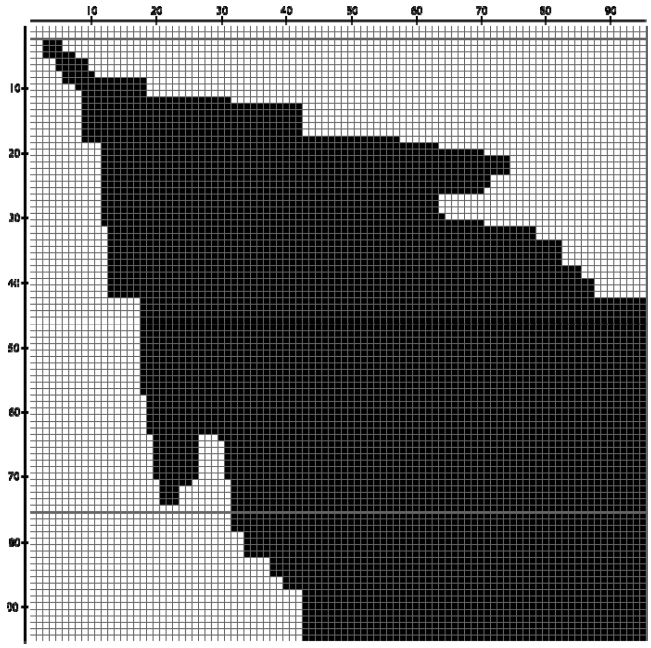

Figure 7. Similarity plot between truncated binding sites from the 2gmj entry. The native 2gmj binding site (92 residues) has been iteratively trimmed residue by residue to generate a set of 91 truncated sites whose pairwise similarity was measured according to the FuzCav descriptor. Similarities higher than 0.16 are colored in black, and similarities lower than 0.16 are colored in white. The size of the binding sites (number of residues) is indicated as upper and lower left labels.

undergoes substantial conformational rearrangements of its binding site (18 residues in total) which particularly affect a stretch of four amino acids (Phe150−Pro153) that drift more than 10 Å away from their ligand-free coordinates to allow substrate enclosure.[47] Nevertheless, FuzCav still detects strong similarities between ligand-free and ligand-bound active sites (similarity of 0.287) because the 14 remaining residues share similar coordinates (rmsd of 1.1 Å). To confirm this observed tolerance on six test cases, 2000 molecular dynamics snapshots of a protein−ligand binding site were compared to the starting structure (Figure 8B,C). Athough heavy atoms of binding site residues exhibit rms deviations up to 3.5 Å from the starting structure, the corresponding FuzCav similarity score decreases to a value (0.27) much above the acceptable threshold (0.16) for similar sites.

Comparison of FuzCav with Other Binding Site Comparison Methods. Numerous site-matching methods have been reported in the literature (for an exhaustive review see ref 13). Comparing them to the herein presented method is quite difficult for numerous reasons, among the most important are (i) unavailability of the program, (ii) different assumptions for defininining a ligand binding site (e.g., automated vs ligand-based detection), (iii) use of very different benchmarking data sets, (iv) prohibitive computation costs (notably for high-throughput virtual screening of site collections).

The suitability of FuzCav in a true virtual screening exercise (querying a data set of ca. 6000 sites for similarity to a single entry) could only be compared to that of BSAlign (Tables 2 and 3). In both virtual functional annotations, FuzCav performed equally to (serine endopeptidases test case) or much better than (protein kinases test case) BSAlign.

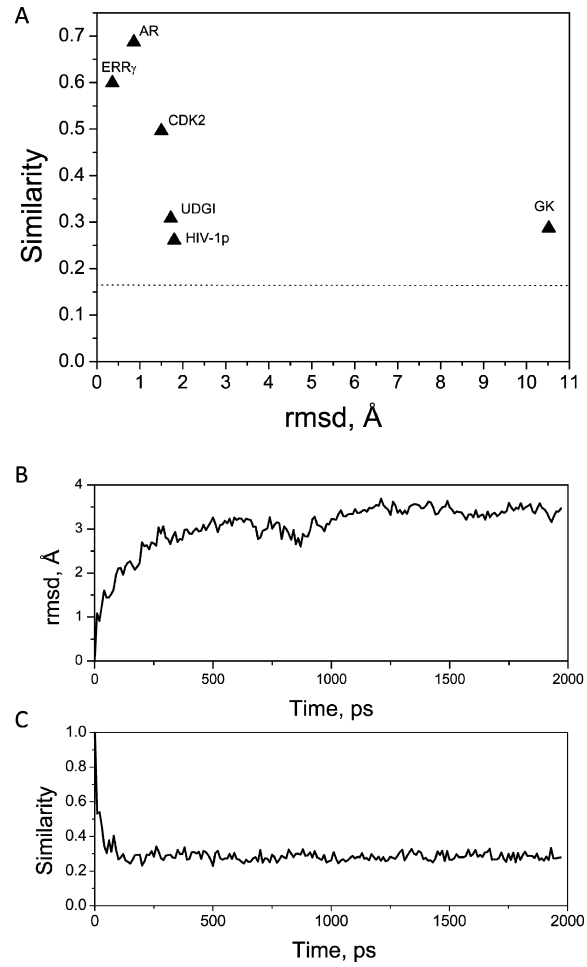

Figure 8. Sensitivity of FuzCav similarity scores to variations of binding site coordinates. (A) FuzCav similarity versus the rms deviations of the holo from the apo structure (active site only) of six targets: uracil DNA-glycosylase inhibitor (UDGI; 36 residues, pdb identifier 1udi vs 1ugi), cell division protein kinase 2 (CDK2; 36 residues, 1dm2 vs 2jgz), HIV-1 protease (HIV-1p; 52 residues, 1qbs vs 1hhp), estrogen-related receptor γ (ERRγ; 27 residues, 2zkc vs 2zbs), aldose reductase (AR; 24 residues, 1ads vs 2nvd), and glycokinase (GK; 18 residues, 1v4t vs 1v4s). The horizontal dotted line represents the similarity threshold (0.16) used throughout this study to discriminate similar from dissimilar protein−ligand binding sites. (B) Rms deviations (Å) from the input conformer of 2000 snapshots of active site heavy atoms generated by the molecular dynamics (MD) simulation of the water-solvated glucagon type-1 receptor structure (homology model, unpublished data). (C) FuzCav similarity score of the 2000 MD snapshots (active site residues only) from the input structure.

Its performance was notably insensitive to the size of the reference binding site. The processing of the smaller data set 4 (126 entries) was possible for BSAlign, SiteAlign, and PocketMatch thanks to its Web Interface (Table 4, Figure 5). PocketMatch and FuzCav were clearly the best two methods for distinguishing 34 adenine-binding pockets from 92 decoys, as measured by the area under the ROC plot (Table 4). A closer look at the early enrichment in true positives (Figure 5) indicates a better capability of FuzCav for placing true adenine-binding pockets among the best scored cavities and thus to lower the false positive rate.

FuzCav was next compared to the full-atom-based FLAP method,[19] which relies on molecular interaction fields. On a data set of 23 ATP-binding sites of protein kinases spanning 4 different subfamilies, a perfect discrimination of the binding

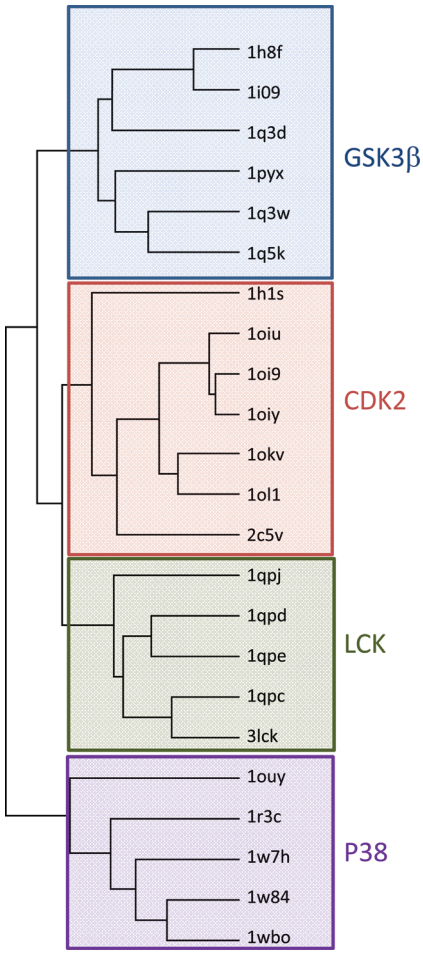

Figure 9. Hierarchical clustering of 23 ATP-binding sites of 4 protein kinase subfamilies[19] according to the FuzCav similarity matrix. FuzCav distances were computed as defined in the Methods. A hierarchical clustering of the matrix was realized with the Cluster3.0 program (Human Genome Center, Institute of Medical Science, University of Tokyo) using an uncentered correlation as the similarity metric and the complete linkage clustering method. The tree is rendered with JavaTreeView (http://sourceforge.net/projects/jtreeview/).

sites is observed, as previously reported for the FLAP method, but with a much simpler and faster approach (Figure 9).

Last, a data set of seven binding site pairs (three sharing the same SCOP family, four belonging to different SCOP families) was retrieved from the previous work of Yeturu et al.,[20] to which we added the difficult case of celecoxib-binding pockets in cyclooxygenase-2 (COX-2) and carbonic anhydrase II (CA-II). Cross-reactivity of celecoxib derivatives with the latter two enzymes was reported by Weber et al.[9] on the basis of shared chemical features of known COX-2 and CA-II inhibitors and could only be explained a posteriori by the similarity of small-sized subpockets for a sulfonamide and a trifluoromethyl group. It is thus an interesting test case for checking the suitability of site comparison methods to detect local similarities. Thanks to Web interfaces to several programs (ProFunc,[48] SitesBase,[16] SuMo,[39] SiteEngine,[17] PocketMatch[20]) and the availability of some executables (BSAlign,[34] SiteAlign[31]), seven methods could be compared to FuzCav in detecting site similarities for these eight pairs (Table 6). These methods could be roughly classified into four categories of increasing complexity. ProFunc determines the best possible match between ligand-binding templates consisting of amino acid triads. PocketMatch and FuzCav converts the binding site coordinates into a frame-invariant fingerprint as lists of either distances or pharmacophoric triplets between Cα and/or Cβ atoms. Pairwise similarity is computed by counting a normalized sum of equivalent elements. BSAlign and SiteAlign represent protein−ligand binding sites by property-annotated Cα atoms and use either a subgraph isomorphism algorithm or a systematic iterative search for finding the largest common subgraph (BSAlign) or the most similar fingerprints (SiteAlign). Last, SitesBases, SuMo, and SiteEngine rely on a full atomic representation of binding site residues by true atoms or representative points. The best alignment is inferred by matching triplets of atoms/points using either clique detection or geometric hashing algorithms. The match is scored according to the number of equivalent atoms/representative points. All methods usually find some similarity among pairs of binding sites from

Table 6. Comparison of FuzCav to Other Binding Site Comparison Tools in Detecting Similarity among Difficult Binding Site Pairs

PDB1−Lig1[a]	PDB2−Lig2[b]	ProFunc[c]	SitesBase[d]	SuMo[e]	SiteEngine[f]	PocketMatch[g]	BSAlign[h]	SiteAlign[i]	FuzCav[j]
speed order (1 measure)		min	s−min	min	min	ms	s	min	ms
				Pairs of Proteins Belonging to the Same SCOP Family					
1gjc−**130**	1v2q−**ANH**	229	29.38	89	32.56	50.17	31.77	0.03	0.19
	2yaw−**ONO**	217	NA	40	x	52.29	31.51	0.02	0.18
	1o3p−**655**	x[k]	54.80	NA[l]	38.48	88.01	42.26	0.01	0.18
				Pairs of Proteins Belonging to Different SCOP Families					
1ecm−**TSA**	4csm−**TSA**	x	54.65	x	x	55.56	x	x	0.18
1m6z−**HEC**	1lga−**HEM**	x	x	x	x	63.85	x	x	x
1zid−**ZID**	2cig−**IDG**	x	NA	NA	x	56.01	x	x	x
1v07−**HEM**	1hbi−**HEM**	x	46.81	x	x	61.42	x	0.20	0.18
6cox−**S58**	1oq5−**CEL**	x	x	x	33.14	x	x	x	0.16

[a] PDB1 = PDB identifier of protein 1, and Lig1 = chemical identifier of ligand 1. [b] PDB2 = PDB identifier of protein 2, and Lig2 = chemical identifier of ligand 2. [c] ProFunc Site Steer score assessed online from the Web server http://www.ebi.ac.uk/thornton-srv/databases/profunc/index.html. [d] SitesBase score computed online from the Web server http://www.modelling.leeds.ac.uk/sb/. The reported score is the ratio between the Score and the Max score expressed in percentage.[16] [e] SuMo score computed online from the server http://www.sumo-pbil.ibcp.fr/ [f] SiteEngine MatchScore(M)[17] computed online from the Web server http://bioinfo3d.cs.tau.ac.il/SiteEngine/. [g] PocketMatch PMScore[20] computed online from the Web server http://proline.physics.iisc.ernet.in/pocketmatch/. [h] BSAlign alignment score.[34] [i] SiteAlign d2 score.[31] [j] FuzCav similarity score. [k] No detectable similarity at a defined threshold (Profunc, *E*-value <1.00 × 10^{-6}, SitesBase, ratio ≥10; SuMo, SiteEngine, Matchscore(M) ≥ 40; PocketMatch, PMScore ≥40; BSAlign, score ≥10; SiteAlign, d2 ≤ 0.20; FuzCav, score ≥0.16). [l] PDB entry not available on the Web server.

proteins sharing the same fold (Table 6). For pairs of sites binding the same ligand in the absence of any fold similarity (second section of Table 6), the lowest (ProFunc) as well as the highest (SuMo, SiteEngine, SitesBase) resolution methods meet significant difficulties to recover pairwise similarities. Lower resolution methods focusing on protein Cα atoms only (BSAlign, SIteAlign) to align protein coordinates also fail in almost all examples. Interestingly, both frame-invariant comparison methods (PocketMatch, FuzCav) present the best compromise between the complexity of the binding site descriptor and the success rate. PocketMatch outperforms FuzCav for large binding site pairs for which the occurrence of charged residues varies (e.g., 1m6z vs 1lga, 1zid vs 2cig), thus affecting the pharmacophoric triplet distribution. Importantly, only FuzCav and SiteEngine manage to find some similarity between 6cox and 1oq5 protein−ligand binding sites, thus suggesting that they may be used to detect subpocket similarities, a very important feature for annotating proteins with novel folds or no representative templates.[46] Additional benchmarks on common data sets are clearly needed to unambiguously compare site-matching programs. However, the current analysis shows that the FuzCav method is robust, relatively insensitive to the binding site definition, and fuzzy enough to be applied to ligand-bound as well as ligand-free protein structures.

A clear drawback of the current method is the lack of interpretability of outputted results. It is currently not possible to explain why two cavity fingerprints are similar and which pairs of residues account for the observed similarity. If the user is interested in aligning two sites, FuzCav may be first used as a first filter to check a putative similarity and then coupled to an independent 3-D alignment tool if the score is above any user-defined similarity threshold. A second limit lies in the different distribution of charged residues observed at the rim of some large but similar binding pockets which directly affect the distribution of pharmacophoric triplets. Since charged residues are less frequent than neutral ones, the resulting fingerprints will be significantly different and the corresponding binding site pairs more difficult to recover.

CONCLUSIONS

We herewith present a generic cavity fingerprint (FuzCav) to compare protein−ligand binding sites. In contrast to most existing tools, the present method does not require a prior 3-D structural alignment of proteins to compare and thus achieves an incomparable pace to quantify pairwise similarities. It is applicable to any druggable cavity from any protein class. The fingerprint was designed to incorporate a certain level of fuzziness by assigning pharmacophoric properties to Cα atoms of cavity-lining residues. It is insensitive to the binding site definition. Moderate ligand-induced fit notably by variation of side chain rotameric states is therefore easily handled in the present descriptor. By discretizing distances between pharmacophoric features into five bins and using a sum of identical counts in the pharmacophore key, the FuzCav descriptor is suited to detect either local or global similarities between protein cavities. It is robust enough to define a single similarity threshold above which two druggable binding sites can be considered as similar. Adaptation to nondruggable cavities is still possible at the condition that the similarity threshold is customized from a suitable data set.

The present descriptor could be used easily to screen a database of binding sites for similarity to any druggable cavity, notably those arising from novel genomic structures, for guiding their functional annotation and identifying their first ligands. Since the cavity descriptor is generic, it can also be associated with ligand descriptors in fixed-sized protein−ligand fingerprints[49] to mine a broad and complex chemogenomic space.

ACKNOWLEDGMENT

We thank the French Ministry of Research and Technology for a Ph.D. grant to N.W. We also acknowledge calculation centers at the CINES (Montpellier, France) and IN2P3 (Villeurbanne, France) for allocation of computing time (Research Project No. x20080725024). Dr. Z. Aung and Dr. J. C. Tong (Institute for Infocomm Research, A*STAR, Singapore) are acknowledged for providing us with the BSAlign executable. The FuzCav program is available for nonprofit research upon request from D.R.

Supporting Information Available: Pharmacophoric properties of standard residues and ions and impact of the active site size of the reference on the ROC values in classifying four subgroups of data set 3. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Xie, L.; Li, J.; Bourne, P. E. Drug discovery using chemical systems biology: Identification of the protein−ligand binding network to explain the side effects of CETP inhibitors. *PLoS Comput. Biol.* **2009**, *5*, e1000387.
(2) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(3) 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.
(4) Ji, H. F.; Kong, D. X.; Shen, L.; Chen, L. L.; Ma, B. G.; Zhang, H. Y. Distribution patterns of small-molecule ligands in the protein universe and implications for origin of life and drug discovery. *Genome Biol.* **2007**, *8*, R176.
(5) sc-PDB: An Annotated Database of Druggable Binding Sites from the Protein Data Bank. http://bioinfo-pharma.u-strasbg.fr/scPDB (accessed Dec 2, 2009).
(6) Dessailly, B. H.; Nair, R.; Jaroszewski, L.; Fajardo, J. E.; Kouranov, A.; Lee, D.; Fiser, A.; Godzik, A.; Rost, B.; Orengo, C. PSI-2: Structural genomics to cover protein domain family space. *Structure* **2009**, *17*, 869–881.
(7) Schneider, M.; Lane, L.; Boutet, E.; Lieberherr, D.; Tognolli, M.; Bougueleret, L.; Bairoch, A. The UniProtKB/Swiss-Prot knowledgebase and its Plant Proteome Annotation Program. *J. Proteomics* **2009**, *72*, 567–573.
(8) Nair, R.; Liu, J.; Soong, T. T.; Acton, T. B.; Everett, J. K.; Kouranov, A.; Fiser, A.; Godzik, A.; Jaroszewski, L.; Orengo, C.; Montelione, G. T.; Rost, B. Structural genomics is the largest contributor of novel structural leverage. *J. Struct. Funct. Genomics* **2009**, *10*, 181–91.
(9) Weber, A.; Casini, A.; Heine, A.; Kuhn, D.; Supuran, C. T.; 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.
(10) Kinnings, S. L.; Liu, N.; Buchmeier, N.; Tonge, P. J.; Xie, L.; Bourne, P. E. Drug discovery using chemical systems biology: Repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput. Biol.* **2009**, *5*, e1000423.
(11) Caron, P. R.; Mullican, M. D.; Mashal, R. D.; Wilson, K. P.; Su, M. S.; Murcko, M. A. Chemogenomic approaches to drug discovery. *Curr. Opin. Chem. Biol.* **2001**, *5*, 464–470.
(12) Rognan, D. Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* **2007**, *152*, 38–52.
(13) Kellenberger, E.; Schalon, C.; Rognan, D. How to measure the similarity between protein ligand-binding sites. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 209–220.

(14) 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.
(15) 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.
(16) Gold, N. D.; Jackson, R. M. Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. *J. Mol. Biol.* **2006**, *355*, 1112–1124.
(17) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of functional sites in protein structures. *J. Mol. Biol.* **2004**, *339*, 607–633.
(18) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.
(19) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): Theory and application. *J. Chem. Inf. Model.* **2007**, *47*, 279–294.
(20) Yeturu, K.; Chandra, N. PocketMatch: A new algorithm to compare binding sites in protein structures. *BMC Bioinf.* **2008**, *9*, 543.
(21) Das, S.; Kokardekar, A.; Breneman, C. M. Rapid comparison of protein binding site surfaces with property encoded shape distributions. *J. Chem. Inf. Model.* **2009**, *49*, 2863–2872.
(22) Yin, S.; Proctor, E. A.; Lugovskoy, A. A.; Dokholyan, N. V. Fast screening of protein surfaces using geometric invariant fingerprints. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 16622–16626.
(23) Mason, J. S.; Good, A. C.; Martin, E. J. 3-D pharmacophores in drug discovery. *Curr. Pharm. Des.* **2001**, *7*, 567–597.
(24) Wermuth, C. G.; Ganellin, C. R.; Lindberg, P.; Mitscher, L. A. Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998). *Pure Appl. Chem.* **1998**, *70*, 1129–1143.
(25) Langer, T.; Hoffmannn, R. D. *Pharmacophores and Pharmacophore Searches*; Wiley-VCH: Weinheim, Germany, 2006.
(26) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity profiling and design using 3D pharmacophores: Pharmacophore-derived queries (PDQ). *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214–1223.
(27) Flohr, S.; Kurz, M.; Kostenis, E.; Brkovich, A.; Fournier, A.; Klabunde, T. Identification of nonpeptidic urotensin II receptor antagonists by virtual screening based on a pharmacophore model derived from structure−activity relationships and nuclear magnetic resonance studies on urotensin II. *J. Med. Chem.* **2002**, *45*, 1799–1805.
(28) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: Overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251–3264.
(29) Karnachi, P.; Kulkarni, A., Application of pharmacophore fingerprints to structure-based design and data mining. In *Pharmacophores and Pharmacophore Searches*; Langer, T., Hoffmannn, R. D., Eds.; Wiley-VCH: Weinheim, Germany, 2006; pp 193−206.
(30) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470–489.
(31) Schalon, C.; Surgand, J. S.; Kellenberger, E.; Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins* **2008**, *71*, 1755–1778.
(32) Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. CATH−A hierarchic classification of protein domain structures. *Structure* **1997**, *5*, 1093–1108.
(33) Igarashi, Y.; Eroshkin, A.; Gramatikova, S.; Gramatikoff, K.; Zhang, Y.; Smith, J. W.; Osterman, A. L.; Godzik, A. CutDB: A proteolytic event database. *Nucleic Acids Res.* **2007**, *35*, D546–549.
(34) Aung, Z.; Tong, J. C. BSAlign: A rapid graph-based algorithm for detecting ligand-binding sites in protein structures. *Genome Inf.* **2008**, *21*, 65–76.
(35) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005**, *48*, 2534–2547.
(36) Park, K.; Kim, D. Binding similarity network of ligand. *Proteins* **2008**, *71*, 960–971.
(37) Powers, R.; Copeland, J. C.; Germer, K.; Mercier, K. A.; Ramanathan, V.; Revesz, P. Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins* **2006**, *65*, 124–135.
(38) Sheridan, R. P.; Holloway, M. K.; McGaughey, G.; Mosley, R. T.; Singh, S. B. A simple method for visualizing the differences between related receptor sites. *J. Mol. Graphics Modell.* **2002**, *21*, 71–79.
(39) 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.
(40) Rosen, M.; Lin, S. L.; Wolfson, H.; Nussinov, R. Molecular shape comparisons in searches for active sites and functional similarity. *Protein Eng.* **1998**, *11*, 263–277.
(41) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. Derivation of 3D coordinate templates for searching structural databases: Application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.* **1996**, *5*, 1001–1013.
(42) Kahraman, A.; Morris, R. J.; Laskowski, R. A.; Thornton, J. M. Shape variation in protein binding pockets and their ligands. *J. Mol. Biol.* **2007**, *368*, 283–301.
(43) Hubbard, S. R.; Wei, L.; Ellis, L.; Hendrickson, W. A. Crystal structure of the tyrosine kinase domain of the human insulin receptor. *Nature* **1994**, *372*, 746–754.
(44) Altschul, S. F.; Wootton, J. C.; Gertz, E. M.; Agarwala, R.; Morgulis, A.; Schaffer, A. A.; Yu, Y. K. Protein database searches using compositionally adjusted substitution matrices. *FEBS J.* **2005**, *272*, 5101–51109.
(45) Andreeva, A.; Howorth, D.; Chandonia, J. M.; Brenner, S. E.; Hubbard, T. J.; Chothia, C.; Murzin, A. G. Data growth and its impact on the SCOP database: New developments. *Nucleic Acids Res.* **2008**, *36*, D419–425.
(46) Wallach, I.; Lilien, R. H. Prediction of sub-cavity binding preferences using an adaptive physicochemical structure representation. *Bioinformatics* **2009**, *25*, i296–304.
(47) Kamata, K.; Mitsuya, M.; Nishimura, T.; Eiki, J.; Nagata, Y. Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase. *Structure* **2004**, *12*, 429–438.
(48) Laskowski, R. A.; Watson, J. D.; Thornton, J. M. ProFunc: A server for predicting protein function from 3D structure. *Nucleic Acids Res.* **2005**, *33*, W89–93.
(49) Weill, N.; Rognan, D. Development and validation of a novel protein−ligand fingerprint to mine chemogenomic space: Application to G protein-coupled receptors and their ligands. *J. Chem. Inf. Model.* **2009**, *49*, 1049–1062.
(50) *Pipeline Pilot*, version 7.5; SciTegic Inc.: San Diego, CA, 2009.

CI900349Y