JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Encoding Protein−Ligand Interaction Patterns in Fingerprints and Graphs

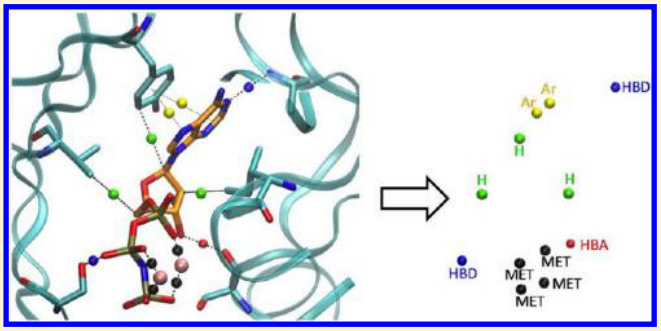
Jérémy Desaphy,[†] Eric Raimbaud,[‡] Pierre Ducrot,[‡] and Didier Rognan*[,†]

[†]Laboratory for Therapeutical Innovation, UMR 7200 Université de Strabsourg/CNRS, MEDALIS Drug Discovery Center, F-67400 Illkirch, France

[‡]Chemistry Partnership and Molecular Modeling, Institut de Recherches SERVIER, 125 chemin de Ronde, F-78290 Croissy sur Seine, France

Supporting Information

ABSTRACT: We herewith present a novel and universal method to convert protein−ligand coordinates into a simple fingerprint of 210 integers registering the corresponding molecular interaction pattern. Each interaction (hydrophobic, aromatic, hydrogen bond, ionic bond, metal complexation) is detected on the fly and physically described by a pseudoatom centered either on the interacting ligand atom, the interacting protein atom, or the geometric center of both interacting atoms. Counting all possible triplets of interaction pseudoatoms within six distance ranges, and pruning the full integer vector to keep the most frequent triplets enables the definition of a simple (210 integers) and coordinate frame-invariant interaction pattern descriptor (TIFP) that can be applied to compare any pair of protein−ligand complexes. TIFP fingerprints have been calculated for ca. 10 000 druggable protein−ligand complexes therefore enabling a wide comparison of relationships between interaction pattern similarity and ligand or binding site pairwise similarity. We notably show that interaction pattern similarity strongly depends on binding site similarity. In addition to the TIFP fingerprint which registers intermolecular interactions between a ligand and its target protein, we developed two tools (Ishape, Grim) to align protein−ligand complexes from their interaction patterns. Ishape is based on the overlap of interaction pseudoatoms using a smooth Gaussian function, whereas Grim utilizes a standard clique detection algorithm to match interaction pattern graphs. Both tools are complementary and enable protein−ligand complex alignments capitalizing on both global and local pattern similarities. The new fingerprint and companion alignment tools have been successfully used in three scenarios: (i) interaction-biased alignment of protein−ligand complexes, (ii) postprocessing docking poses according to known interaction patterns for a particular target, and (iii) virtual screening for bioisosteric scaffolds sharing similar interaction patterns.



INTRODUCTION

Three-dimensional (3D) structures of protein−ligand complexes provide crucial information to better understand molecular rules governing living cells and assist rational drug discovery. If analyzing a few structures at a graphic desktop is now common practice, mining and comparing a large array of protein−ligand complexes requires a simplification of the 3D information. Among the most useful simplification processes for analyzing protein−ligand interactions is the conversion of atomic coordinates into simpler 1D or 2D fingerprints.[1] Fingerprints are easy to generate, manipulate, compare and, therefore, enable a systematic analysis of large data sets. They are largely used to describe and compare molecular objects (small molecular weight ligands,[2] pharmacophores,[3] proteins,[4] and protein−ligand binding sites[5]) and represent descriptors utilized by computer-aided drug design programs, notably in silico screening tools. Computational chemists frequently manipulate these fingerprints independently in either ligand-based or structure-based approaches to drug design.[6] It would, however, be very interesting to combine both protein-based and ligand-based information in a single descriptor focusing on molecular interactions in order to answer the following questions: Do similar binding sites identically recognize similar ligands? Are protein−ligand interaction patterns conserved across target families? Which chemically different ligand structures or substructures share identical interaction patterns with a single target?

Two main approaches to merge ligand−target pairs in a single descriptor have been proposed up to now. The first one qualitatively describes the protein−ligand pair by annotating ligand descriptors with interaction features. For example, Bajorath et al. reported a way to augment classical ligand fingerprints with protein−ligand interaction-derived information.[1,7,8] The basic underlying idea is that all atoms of a bioactive ligand are not equally responsible for its biological activity.

Received: November 27, 2012
Published: February 22, 2013

Focusing a chemical fingerprint to protein-interacting ligand atoms (interacting fragment or IF) is likely to enhance the value of such a fingerprint by avoiding the possibility to recruit novel compounds by a pure ligand-based virtual screening approach for wrong reasons (atoms/groups not interacting with a protein pocket). Such IF-annotated fingerprints were shown to outperform conventional fingerprints in standard similarity searches to known ligands of diverse activity classes.[8] Standard descriptors for protein cavities and their cognate ligands can also be concatenated into a single fingerprint[9,10] and then used as input to train machine learning algorithms to discriminate true from false complexes.[11] In a prospective virtual screening study, this kind of fingerprint was found superior to conventional ligand-based descriptors (2D and 3D) in finding novel nonpeptide ligands for the oxytocin receptor.[12] However, the latter descriptors do not describe the physical intermolecular interactions (e.g., hydrogen bond, hydrophobic contact) between ligand and target.

The second possible approach to protein−ligand fingerprinting annotates protein descriptors (usually binding site-lining amino acids) with ligand-interaction features. The interaction fingerprint concept (SIFt: Structural Interaction Fingerprint) was pioneered by Biogen Idec.[13] and consists in converting a 3D protein−ligand complex into a 1D bit string registering intermolecular interactions (hydrophobic, hydrogen bonds, ionic interactions) between a ligand and a fixed set of active site residues. Interactions are computed on the fly using standard topological criteria between interacting atoms (distances, angles) and a bit is switched "on" or "off" as to whether the interaction occurs or not. The SIFt method was originally designed for analyzing ligand docking poses to protein kinases and shown several promising features: (i) enhancing the quality of pose prediction in docking experiments,[13] (ii) clustering protein−ligand interactions for a panel of related inhibitors according to the diversity of their interactions with a target subfamily,[14] and (iii) assisting target-biased library design.[15] The interaction fingerprint concept was further developed by other groups in order to define the directionality of the interactions (e.g., H-bonds donated by the ligand and by the active site are stored in distinct bits),[16] the strength of the interaction,[17] or assign a bit to every active site atom instead of every active site residue.[18] Interaction fingerprints (IFPs) are now part of many docking tools in order to postprocess docking poses according to known protein−ligand X-ray structures. Remarkably, IFPs show a great scaffold hopping potential in selecting virtual hits sharing the same interaction pattern than a reference ligand, but with different chemotypes.[19] Since a bit is defined for every active site atom/residue, the method is therefore limited to analyze interactions with highly homologous active sites sharing a fixed number of cavity-lining atoms/residues. To overcome this limitation, cavity-independent fingerprints (APIF)[20] do not consider the absolute but only the relative positions of pairs of protein−ligand interacting atoms and store information in a 294-bit fingerprint according to the interaction type and distance between interacting pairs. Like standard IFPs, APIF scoring by comparison to known references was shown to outperform conventional energy-based scoring functions in docking-based virtual screening of compound libraries. Unfortunately, obtained results are difficult to interpret since deconvoluting APIF into specific protein−ligand features or protein−ligand alignments is not possible. Databases of protein−ligand complexes (e.g., CREDO,[21] PROLIX[22]) focusing on observed interactions in X-ray structures have been described and use fingerprint representations of interaction patterns to retrieve PDB complexes fulfilling user queries (e.g., number and/or type of protein−ligand interactions, interaction to particular amino acids). However, neither a 3D alignment of protein−ligand complexes nor a generic similarity measure between the two complexes to evaluate, are proposed.

The novel protein−ligand descriptors (fingerprint, graph) and comparisons methods (Ishape, Grim) presented in this study were therefore designed to specifically enable the following features: (i) compare protein−ligand interactions whatever the size and sequence of the target binding sites (e.g., across target families), (ii) quantitatively describe molecular interactions with a specific frame-invariant descriptor, and (iii) provide an alternative 3D alignment of protein−ligand complexes to protein-based or ligand-based matches, by focusing on molecular interactions only.

It enables an exhaustive and easily interpretable pairwise comparison of all protein−ligand X-ray structures that may be used in several scenarios: postprocess protein−ligand docking poses, find off-targets sharing key interaction patterns to a known ligand, and identify bioisosteric fragments with a conserved interaction mode to a given target

METHODS

Data Sets of Protein−Ligand Complexes. All protein−ligand complexes were retrieved from the sc-PDB data set[23] which archives 9877 high resolution X-ray structures of druggable protein−ligand complexes. For each complex, protein and ligands were separately stored in TRIPOS mol2 file format.[24] Pairwise sc-PDB ligand similarity was expressed by the Tanimoto coefficient on either circular ECFP4 fingerprints[25] in PipelinePilot[26] or MACCS 166-bit structural keys[27] in MOE.[28]

Set 1: 900 Similar and 900 Dissimilar Protein−Ligand Complexes. Pairs of protein−ligand complexes were considered similar if (i) their pairwise binding site similarity (expressed by the Shaper similarity score[29]) was higher than 0.44 and (ii) their pairwise ligand similarity (expressed by a Tanimoto coefficient on ECFP4 fingerprints) was between 0.55 and 0.75. This selection protocol led to a set of 7426 pairs of similar complexes, out of which 900 pairs were finally selected (Supporting Information Table 1) by retrieving all possible nonredundant Uniprot names. The same number of pairs of dissimilar protein−ligand complexes was retrieved assuming that their pairwise binding site similarity and their pairwise ligand similarity were lower than 0.20. This cutoff was chosen arbitrarily to be sure that both active sites and bound-ligands were really dissimilar. This selection protocol led to 3 524 800 pairs of dissimilar complexes, out of which 900 pairs were randomly selected (Supporting Information Table 1) avoiding duplicates in protein names.

Set 2: sc-PDB Fragments. The 9877 ligands of the current sc-PDB release were submitted to a retrosynthetic fragmentation protocol using 11 RECAP[30] rules embedded in Pipeline Pilot.[26] About 78% of the ligands (7769) could be fragmented into 20 839 building blocks (15 828 cyclic, 5011 acyclic). Each fragment was annotated with descriptors from its parent ligand (HET identifier, fragment number) and PDB target (Uniprot target name, KEGG BRITE functional class[31]). PDB codes and HET codes of the fragments are given in Supporting Information Table 2.

Set 3: CCDC/Astex Subset of Protein−Ligand Complexes. The CCDC/Astex set of 95 clean high-resolution (<2.0 Å) protein−ligand X-ray structures[32] was downloaded from the Cambridge Crystallographic Data Centre.[33] Here, 45 entries,

dx.doi.org/10.1021/ci300566n | J. Chem. Inf. Model. 2013, 53, 623−637

already present in the sc-PDB were discarded, 14 additional entries were removed since the corresponding target (mainly antibodies) was completely absent in the sc-PDB, thus leading to a final set of 36 protein−ligand complexes (Supporting Information Table 3). Hydrogen atoms were added in SYBYL[24] and their atomic coordinates changed to manually optimize protein−ligand interactions first and intramolecular interactions in a second step. Ionization and tautomeric states of cavity-lining residues were accordingly updated whenever necessary. Proteins and ligands were separately stored in ready-to-dock file mol2 file format.

Set 4: DUD-E Target and Ligand Sets. Active and decoy ligand sets for 10 targets covering 5 different protein families: G protein-coupled receptors (adenosine A2A receptor: AA2AR, adrenergic beta2 receptor: ADRB2), nuclear hormone receptors (androgen receptor: AND, glucocorticoid receptor: GCR); other enzymes (adenosine deaminase: ADA, prostaglandin G/H synthase 2: PGH2); proteases (angiotensin-converting enzyme: ACE, renin: reni); and protein kinases (fibroblast growth factor receptor 1: FGFR1, RAC-alpha serine/threonine-protein kinase: AKT1) were downloaded in 3D mol2 file format from the DUD-E[34] Web site (http:// http://dude.docking.org/). Since several pdb entries selected as DUD-E atomic coordinates for the host protein were already present in the sc-PDB and one of our rescoring procedures (Grim) relies on existing protein−ligand complexes, PDB entries not present in the sc-PDB and of the highest possible resolution (2pwh, 2am9, 1p93, 1a4l, 3zqz, 3sfc, 3tt0, 4ekl) were selected as host coordinates for docking. For two targets (ADRB2, PGH2), we decided to keep the original DUD-E PDB entry (3ny8, 3nt1) but removed it from the set of references for further scoring.

Detection of Protein−Ligand Interactions. Seven pharmacophoric properties (hydrophobic, aromatic, h-bond donor, h-bond acceptor, positive ionizable, negative ionizable, metal; Supporting Information Table 4) for protein and ligand atoms are assigned by parsing the atom and bond connectivity fields of the mol2 files. Protein−ligand interactions are then detected on the fly with respect to the above-defined pharmacophoric types and previously defined topological criteria[16] (Supporting Information Table 5). The protein−ligand interaction (Figure 1A) is characterized by the two interacting atoms and an interaction pseudoatom (IPA) located at three possible positions: (i) the geometric center of interacting atoms (*Centered mode*), (ii) the interacting protein atom (*InterProt mode*), and (iii) the interacting ligand atom (*InterLig mode*). IPAs can be computed using one of the three possible modes to enable a mapping of the interaction on either ligand or protein atoms (InterLig and InterProt modes, respectively) or more naturally at the mid-distance of interacting atoms (Centered mode).

Since hydrophobic atoms are by far the most frequent, two additional rules have been defined to limit hydrophobic IPAs (Figure 1B). First, only one IPA can be generated between a single ligand atom and a single protein amino acid, whatever the number of interacting atoms. In case a ligand atom verifies the condition of a hydrophobic interaction with two different atoms of the same protein residue, only the shortest distance is kept to assign the corresponding IPA. All hydrophobic IPAs are then iteratively clustered with a hierarchical agglomerative clustering using a 1.0 Å distance criterion and an average linkage method. Second, two aromatic rings not engaged in an aromatic interaction (edge-to-face or face-to-face) will define a hydrophobic interaction between their two closest atoms at the condition that the corresponding distance rule is verified (Figure 1C). When the aromatic interaction condition is verified, an aromatic IPA is set between centroids of the corresponding aromatic rings. In InterLig mode only, a last pruning step avoids property redundancy on ligand atoms by keeping a single occurrence of interaction type per ligand atom. For example, if a ligand atom makes one aromatic and two hydrophobic interactions, only one aromatic and one hydrophobic InterLig IPA are created. Finally, aromatic edge-to-face and face-to-face interactions are merged to represent only one aromatic IPA. When all interactions have been detected, IPAs are exported in a mol2 file format.

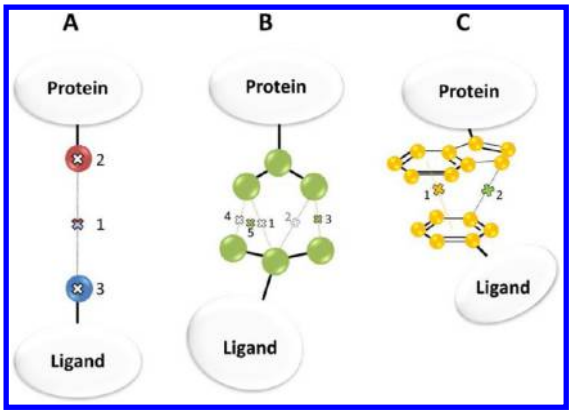


Figure 1. Definition of interaction pseudoatoms (IPAs). (A) Case of hydrogen bonding, ionic interaction, and metal complexation. Two atoms of complementary pharmacophoric properties (red and blue balls) fulfilling interaction rules describe an interaction (dotted line) at a pseudoatom (cross) located at three possible positions: (i) geometric center of interacting atoms (1), (ii) interacting protein atom (2), (iii) interacting ligand atom (3). (B) Case of hydrophobic interactions between protein and ligand hydrophobic atoms (green balls). Only the shortest interaction (number 1) between a single ligand atom and many protein atoms is kept (interaction 2 is not conserved). Remaining hydrophobic IPAs (1, 3, 4) are then clustered to yield IPAs 3 and 5. (C) Case of aromatic interactions between protein and ligand aromatic atoms (yellow balls). Protein and ligand aromatic atoms verifying the aromatic interaction rules define an aromatic interaction (yellow dotted line) with an aromatic IPA set at the mid-distance between both aromatic ring centroids (IPA 1). For aromatic interacting atoms, not respecting the aromatic interaction rule but fulfilling a hydrophobic interaction, a hydrophobic interaction (green dotted line) and IPA (IPA 2) is defined between the closest protein and ligand atoms.

Fingerprinting Triplets of Interaction Pseudoatoms (TIFP). Starting from a set of IPAs (whatever the mapping mode), the TIFP fingerprint encodes protein−ligand interactions by a vector of 210 integers (Figure 2). Each integer of the vector registers the count of unique IPA triplets (seven properties and three related distances) occurring at binned interfeature distances. Please note that we only consider seven properties since the two aromatic interactions (edge to face, face-to face) are merged. The distances between IPAs are currently discretized in six intervals (0−4, 4−6, 6−9, 9−13, 13−17, 17+ Å). Starting from the first interval, all triplet combinations are counted and stored, until the last interval is processed. Given seven pharmacophoric types and six distance ranges, the total number of triplets is thus equal to $7^3 \times 6^3 = 74\,088$. To generate the shortest possible fingerprint, redundant triplets (property redundancy, isoscele, and equilateral triangles) are removed. Last, the geometrical validity of the pharmacophoric triplet is checked by applying the triangle inequality rule stating that one

dx.doi.org/10.1021/ci300566n | J. Chem. Inf. Model. 2013, 53, 623−637

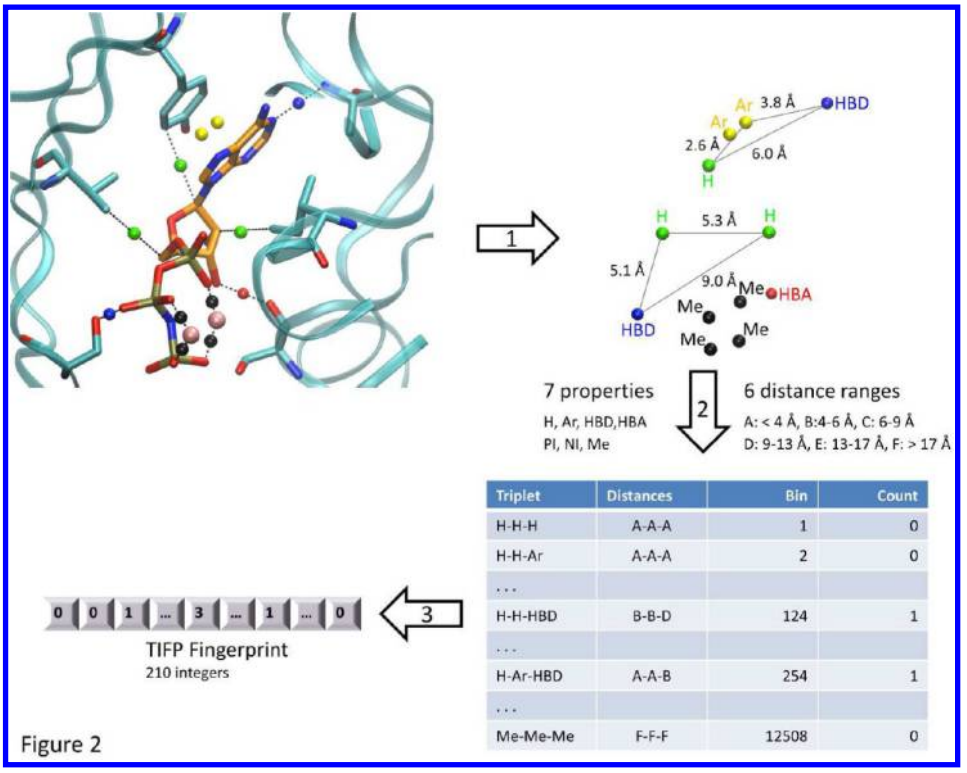


Figure 2. Generating a fingerprint of IPA triplets. From atomic coordinates of the protein−ligand complex (PDB code: 1j7u), interactions are detected on the fly and described by IPAs featuring seven interaction types (hydrophobic, H, green balls; aromatic, Ar, yellow balls; H-bond donor, HBD, blue balls; H-bond acceptor, HBA, red balls; positive ionizable, PI, not featured here; negative ionizable, NI, not featured here; metal complexation, Me, dark). All possible triplets (three properties, three distances) of IPAs are generated (step 1) and matched to a triplet list (step 2). The count of each triplet type is encoded through a fingerprint of 12 508 integers that is further pruned to 210 integers (step 3) to feature the most frequently occurring triplets.

distance cannot be longer than the sum of the two others. The full fingerprint accounts for 18 179 possible triplets stored in 12 508 bins. The higher number of triplets with regards to bins is simply due to the redundancy which is observed at the triplet (e.g., ABC and BAC triplets) and not at the bin level (both triplets could be assigned to a single bin).

To speed up fingerprint calculations and comparisons, a two-step compression was done as follows. Full length fingerprints were computed for the 9877 protein−ligand complexes of the sc-PDB data set (2011 release) and the count status of every triplet was computed. First, 17 612 weakly populated triplets corresponding to 12 119 bins (count <10) were removed. Second, 361 triplets (226 bins) with a frequency between 10 and 20 were merged into 185 triplets (46 bins) of the same composition but with no distance information. The remaining 206 triplets (164 bins), with a frequency higher than 20, were kept unchanged. The final size of the compressed TIFP fingerprint amounts to 391 triplets in 210 separate bins. The similarity between two TIFP fingerprints was expressed by a Tanimoto coefficient as follows:[35]

$$\mathrm{Tc} = \frac{\sum_{j=1}^{N} x_{jA}x_{jB}}{\sum_{j=1}^{N} (x_{jA})^2 + \sum_{j=1}^{N} (x_{jB})^2 - \sum_{j=1}^{N} x_{jA}x_{jB}}$$

where x_{jA} is the value of the bin j in the reference fingerprint A and x_{jB} the value of the bin j in the comparison fingerprint B.

Shape Matching of IPAs (IShape). IShape uses an algorithm very similar to that recently described in Shaper, a tool to align protein−ligand binding sites.[29] It relies on OEChem and OEShape toolkits[36] which present the advantage to describe molecular shapes by a smooth Gaussian function and to align two molecular objects (IPAs) by optimizing the overlap of their corresponding volumes.[37−39] During the alignment, a reference IPA set (Centered mode only) is kept rigid while the set of IPAs to fit (fit object) undergoes rigid body rotations and translations. To speed-up calculations, the "Grid" volume overlap method was chosen to represent the volume of the target IPAs and all atom radii were set to that of carbon (1.7 Å). Once the best shape alignment has been achieved, it is scored by a "Color Force Field" (a color being a pharmacophoric feature) similar to that used by the ligand matching tool ROCS[36] to account for pharmacophoric properties matching. In other words, the alignment proposed by the simple shape matching is scored to account for pharmacophoric feature superposition.

The force field (Supporting Information Table 6) consists in SMARTS patterns for seven pharmacophoric properties (H, hydrophobic; Ar, aromatic; HBA, H-bond acceptor; HBD, H-bond donor; A−, negative ionizable; D+, positive ionizable; Me, metal complexation) and seven pattern matching rules (H to H, Ar to Ar, HBA to HBA, HBD to HBD, A− to A−, D+ to D+, Me to Me) to score the shape-based alignment by pharmacophoric similarity. Color matches were considered for interaction points up to 1.5 Å apart with a single weight for all matching rules.

The similarity $\mathrm{Sim}_{A,B}$ between IPAs A (reference) and B (fit) was calculated by a Tversky index as follows:

$$\mathrm{Sim}_{A,B} = \frac{O_{A,B}}{0.95I_A + 0.05I_B + O_{A,B}}$$

dx.doi.org/10.1021/ci300566n | J. Chem. Inf. Model. 2013, 53, 623−637

Table 1. Pharmacophoric Property Distribution of Interactions in the sc-PDB Data Set (n = 9877)

position	Tot[a]	Hyd[b]	Ar[c]	HBA[d]	NI[e]	HBD[f]	PI[g]	Me[h]
centered	284 186	72%	1%	6%	1%	16%	3%	1%
ligand	182 262	66%	1%	8%	1%	18%	4%	2%
protein	284 186	72%	1%	6%	1%	16%	3%	1%

[a]Total number of protein−ligand interacting pseudoatoms (IPAs). [b]IPA with hydrophobic property. [c]IPA with aromatic property. [d]IPA with H-bond acceptor property (from protein side). [e]IPA with negative ionizable property (from protein side). [f]IPA with H-bond donor property (from protein side). [g]IPA with positive ionizable property (from protein side). [h]IPA with metal complexation property.

where $O_{A,B}$ is the overlap between colors of IPAs A and B, and I is nonoverlapped colors of each entity A and B. Classification models based on pairwise similarity values were assessed by computing the area under the receiver operating characteristic (ROC) curve,[40] and the F-measure as follows:

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

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

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

where TP is true positives, FN is false negatives, FP is false positives, and FN is false negatives. The best similarity threshold is found by the maximum of the F-measure curve when the threshold was varied from 0 to 1 with an increment of 0.01. ROC and Boltzmann enhanced discrimination of ROC (BEDROC) curves were computed using the CROC program.[41] For computing BEDROC values, the alpha-parameter was set to the default value of 20.

Graph Matching of IPAs (Grim). At the noticeable difference of IPA fingerprinting, Grim matching simultaneously considers the three IPA definition modes (InterLig, InterProt, Centered) in order to take into account the ligands, the proteins and the protein−ligand interactions at the same time. A set of IPAs (either reference or target) can be represented as a graph where vertices are IPAs and edges represents all possible links (distances) between two IPAs. Therefore, all IPAs are linked to each other in the graph. In order to compare two sets of IPAs, we rely on the conventional graph theory and product graphs. Each vertex of a product graph describes a possible match between a reference IPA and a target one, only if both IPAs have the same label (e.g., h-bond acceptor) and the same "point of view" (Centered, InterLig, or InterProt). For example, centered IPAs of reference and target will be compared but centered IPAs of reference and InterLig IPAs of the target will not. Vertices of this product graph therefore lists all possible matches between two sets of IPAs. Moreover, since hydrophobic interactions are more frequent than other interaction types, one needs to take into account the relative ferquency of each interaction type. Therefore, a weight is added to each vertex that is inversely proportional to the observed frequency among the 284 186 IPAs generated from the 9877 sc-PDB protein−ligand complexes. Assigned weights were as follows: hydrophobic IPA (0.299), aromatic IPA (0.990), h-bond acceptor (0.930), h-bond donor (0.834), negative ionizable (0.993), positive ionizable (0.966), and metal complexation (0.985). It should be recalled that the different weights assigned to hydrogen bonds (acceptor vs donor) comes from a previously observed bias in the sc-PDB in which donors occur more frequently from protein than from ligand atoms.[42]

Since a graph is not a 3D object, we need to take into account the relative spatial orientation of IPAs by looking at distances connecting them. Assuming a pair of IPAs for the reference (IPA_{R1}, IPA_{R2}) and the target (IPA_{T1}, IPA_{T2}), the two vertices Vg_1 and Vg_2 in the product graph are defined by

$$\text{Vg}_1 = (\text{IPA}_{R1}, \text{IPA}_{T1})$$

$$\text{Vg}_2 = (\text{IPA}_{R2}, \text{IPA}_{T2})$$

An edge is created between Vg_1 and Vg_2 if

$$|d(\text{IPA}_{R1}, \text{IPA}_{R2}) - d(\text{IPA}_{T1}, \text{IPA}_{T2})| < d_{\text{Thres}}$$

The acceptable distance threshold d_{Thres} depends on the IPA definition (InterLig, InterProt, Cenetred) with corresponding values listed in Supporting Information Table 7.

This product graph gives the possibility to find same interaction types with the same spatial environment. The largest cliques, which corresponds to the maximal set of possible IPAs matches with the same spatial environment, are detected using the Bron−Kerbosch algorithm[43] with pivoting and pruning improvements.[44] Each IPA of the target is then matched with the corresponding reference IPA using a quaternion-based characteristic polynomial.[45] It returns both the translation vector and the rotation matrix to match target and reference graphs as well as a Graph-alignment score (Grscore). Cliques are then scored by decreasing Grscore, a score which was empirically determined by fitting six Grim parameters to the previously described IShape similarity score on the set of 1800 protein−ligand complexes (900 similar and 900 dissimilar) as follows:

- $\text{Grscore} = 0.5006 + 0.0151N_{\text{Lig}} + 0.0039N_{\text{Center}} + 0.0143N_{\text{Prot}} + 0.2098\text{Sum}_{\text{Cl}} - 0.0720\text{RMSD} - 0.0003\text{DiffI}$
- N_{Lig}: number of matched InterLig IPAs
- N_{Center}: number of matched Centered IPAs

$$\text{Sum}_{\text{CI}}: \frac{\sum \text{pair weights in clique}}{\sum \text{all possible pair weights}}$$

- N_{Prot}: number of matched InterProt IPAs
- RMSD: root-mean square deviation of the matched clique
- DiffI: absolute value of the difference in the number of IPAs between reference and query.

Docking. All ligands from sets 2−4 were docked into their original X-ray structure with the Surflex-Dock (v.2601) software.[46] Protomols were first generated from the list of cavity-lining residues defined as any amino acid within a 6.5 Å-radius sphere centered on the bound-ligand center of mass. Compounds were then docked with default settings (excepted for the "−pgeom" option) of the docking engine keeping the best 20 poses according to the native Surflex-Dock scoring function. All poses were finally reranked by decreasing Tanimoto similarity value of the TIFP and decreasing Grscore to either the native X-ray pose (self-docking) or any sc-PDB pose with the same sc-PDB target name (target-directed docking).

dx.doi.org/10.1021/ci300566n | J. Chem. Inf. Model. 2013, 53, 623−637

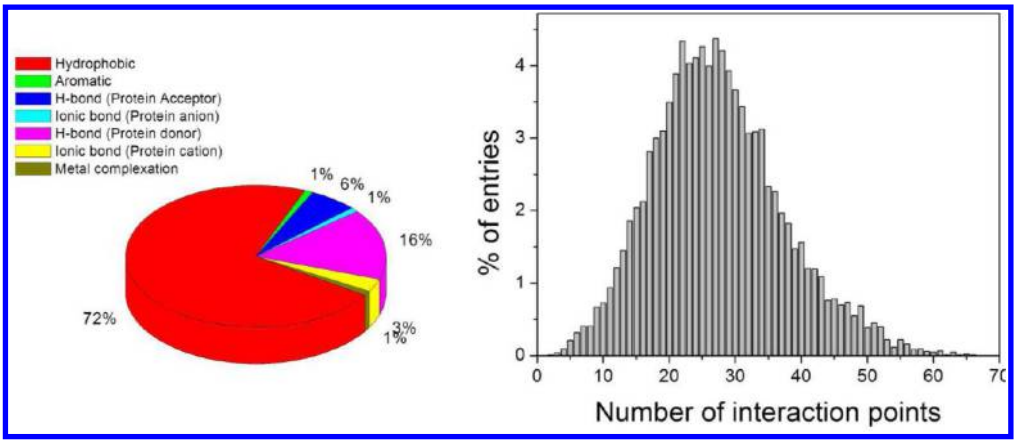


Figure 3. Analysis of IPAs from 9877 protein complexes of the sc-PDB data set: (A) distribution of interaction types, (B) distribution of interaction points.

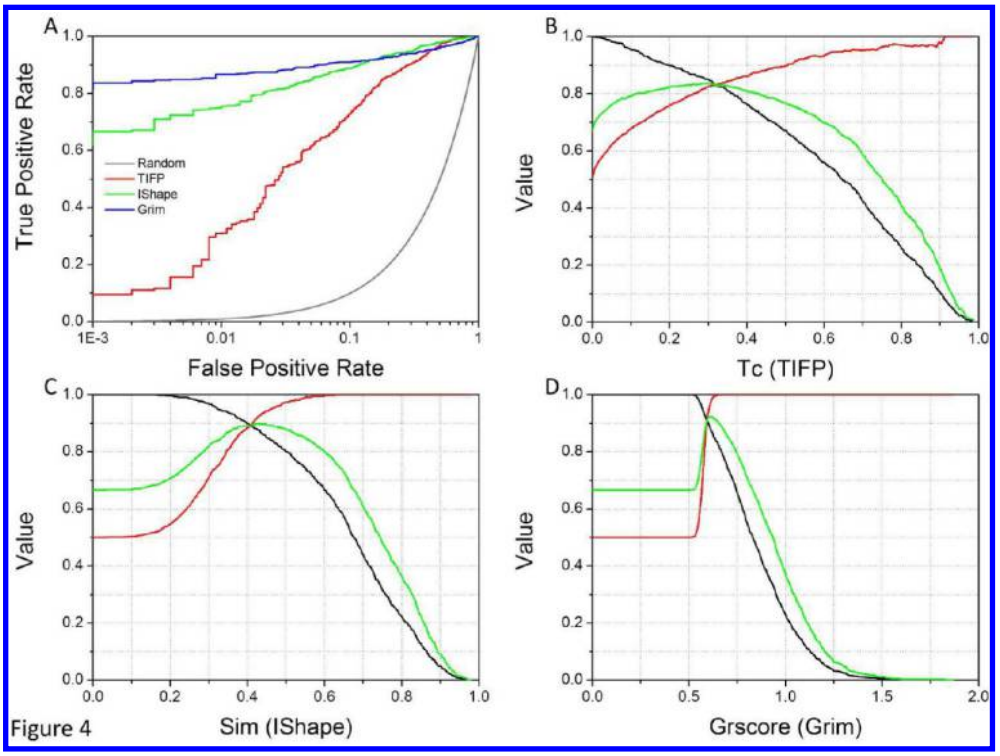


Figure 4. Statistical evaluation of protein−ligand complex similarity scores. (A) ROC plot obtained by sorting 1800 sc-PDB pairs of complexes by decreasing similarity values (red, TIFP Tc score; green, Ishape Sim score; blue, Grim Grscore). True positives are pairs of similar protein−ligand complexes predicted similar whereas false positives are pairs of dissimilar complexes predicted similar. Accuracy of random picking is represented by a gray line. (B−D) Variation of statistical parameters (recall, dark line; precision, red line; *F*-measure, green line) for a binary classification model (similar/dissimilar) of all 1800 pairs, according to the TIFP similarity threshold value (panel B), IShape similarity score (panel C), and Grim Grscore (panel D).

RESULTS AND DISCUSSION

Fingerprinting Interaction Patterns in sc-PDB Complexes. All 9877 protein−ligand complexes of the sc-PDB were parsed to detect protein−ligand interactions and defined 284 186 interaction pseudoatoms in total (Table 1). As to be expected, about 70% of these interactions represent apolar hydrophobic contacts (Figure 3). Due to the prevalence of anionic compounds (mostly nucleotides) among sc-PDB ligands, H-bonds donated by protein atoms are significantly more abundant than H-bonds donated by ligands (16% vs 6%, respectively). The same statistical discrepancy also occurs when comparing salt bridges (3% with a protein cation, 1% with a protein anion; Figure 3). For a protein−ligand complex, the number of interactions goes from 2 (*E. coli* dihydroorotate dehydrogenase in complex with orotic acid, PDB code: 1f76) to 78 (orange carotenoid protein R155L mutant in complex with beta-caroten-4-one; PDB code: 3mg3). A protein makes in average 28 ± 10 interactions with its ligand. However, when positioning the pseudoatoms on the ligand (InterLig mode), fewer interactions are output (182 262 in total, Table 1) due to the additional filtering process of hydrophobic IPAs (see Methods). Using the InterLig mode, a sc-PDB complex has on average 18 ± 6 IPAs, with a minimum of 1 and a maximum of 44.

Reliable Similarity Metric to Compare Protein−Ligand Interaction Patterns. A data set of 900 pairs of similar protein−ligand complexes and 900 pairs of dissimilar complexes

dx.doi.org/10.1021/ci300566n | *J. Chem. Inf. Model.* 2013, 53, 623−637

(see Methods) was setup to investigate the possibility to quantitatively assess the similarity of protein−ligand complexes from either TIFP fingerprints (alignment-free comparison) or the set of matched IPAs (Ishape: shape-based alignment, Grim: graph-based alignment). First, a binary classification of all 1800 pairs using a ROC curve analysis of three possible lists, ordered by decreasing Tc (TIFP), Sim (IShape), and Grscore (Grim) values, clearly indicates that all three metrics discriminate similar from dissimilar complexes (Figure 4A, Table 2). Almost perfect

Table 2. Quantitative Estimation of Pairwise Similarity for 1800 pairs (900 similar, 900 dissimilar) of Protein−Ligand Complexes

statistics	TIFP[a]	IShape[a]	Grim[a]
AUROC[b]	0.908	0.954	0.959
BEDROC[c]	0.973	0.999	0.999
Best threshold[d]	0.318	0.407	0.594
F-measure[e]	0.830	0.892	0.909
precision100%[f]	0.911	0.627	0.659

[a]Similarity measured by the Tanimoto coefficient from TIFP fingerprints (TIFP), the IShape Sim similarity index (IShape), and the Grscore (Grim). [b]Area under the ROC curve for a binary classification (similar, dissimilar). [c]Boltzmann enhanced discrimination of ROC. [d]Similarity score enabling the best possible classification. [e]*F*-measure of the best classification model. [f]Similarity score enabling a perfect classification (precision of 100%).

area under the ROC curves are obtained with either Grim or IShape comparisons (0.96 and 0.95, respectively). The corresponding value using the simple TIFP fingerprint similarity is still very high but significantly lower (0.90) and therefore indicates some limited but true noise in the fingerprints that does not exist in IPAs themselves. Plotting the performance of a binary classification model (complexes are either predicted similar or dissimilar) as a function of the similarity threshold used for deciding upon similarity, gives an explanation for the higher permissivity behavior of the TIFP score. Hence, the corresponding classification models exhibit recall values slowly decaying and precisions slowly increasing when the similarity Tc value increases (Figure 4**B**). The gap between the similarity value at the highest *F*-measure (Tc = 0.318) and that at the maximal precision (Tc = 0.911) is very large. Conversely, using the Grscore as a metric of complex similarity produces classification models whose performance (recall, precision, *F*-measure) are optimal in a very narrow similarity threshold window (0.59 < Grscore < 0.65, Figure 4D) and therefore more robust. In between, the IShape similarity score varies between 0.41 at the *F*-measure optimum and 0.63 at the precision optimum (Figure 4C).

The greater noise in the TIFP fingerprint comparison with respect to either shape or graph matching arises for two main reasons: (i) the information loss upon converting 3D information into 1D data and (ii) the relative importance of hydrophobic triplets in TIFP fingerprints (with either 2 or 3 features) which is minored in the graph alignment-based scores. Hence, our clique detection method uses a weight on pharmacophoric features which is inversely proportional to their abundance in the sc-PDB (hydrophobes are less important than polar features in the clique ranking).

Current benchmarks on a Intel Core2 Duo E8500 processor (3.16 GHz, 6 M cache) indicate that all three comparison methods are fast enough (10, 20, and 35 ms/comparison for TIFP, Ishape, and Grim, respectively) to be applied to large scale comparisons.

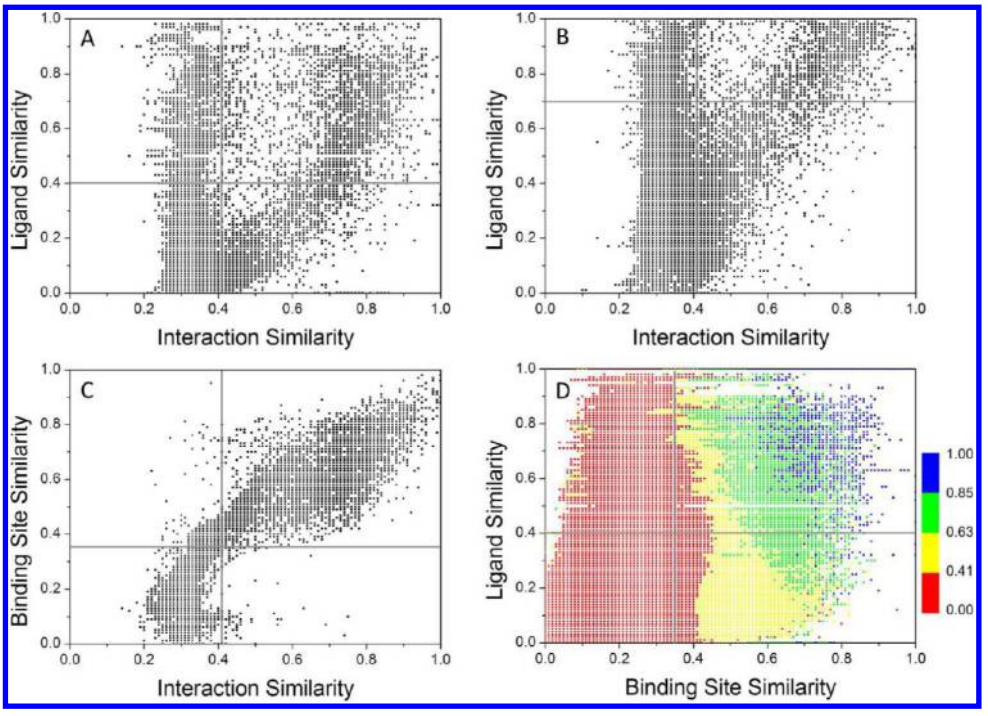


Figure 5. Relationships between ligand similarity, binding site similarity, and interaction pattern similarity for 9877 sc-PDB entries. (A) Ligand similarity (Tanimoto coefficient on ECFP4 fingerprints) vs interaction pattern similarity (IShape similarity score). (B) Ligand similarity (Tanimoto coefficient on MACCS public keys) vs interaction pattern similarity (IShape similarity score). (C) Binding site similarity (Shaper[29] similarity score) vs interaction pattern similarity (Ishape similarity score). (D) Ligand similarity (Tanimoto coefficient on ECFP4 fingerprints) vs binding site similarity (Shaper[29] similarity score). Data are colored according to the interaction pattern similarity score (IShape similarity).

dx.doi.org/10.1021/ci300566n | *J. Chem. Inf. Model.* 2013, 53, 623−637

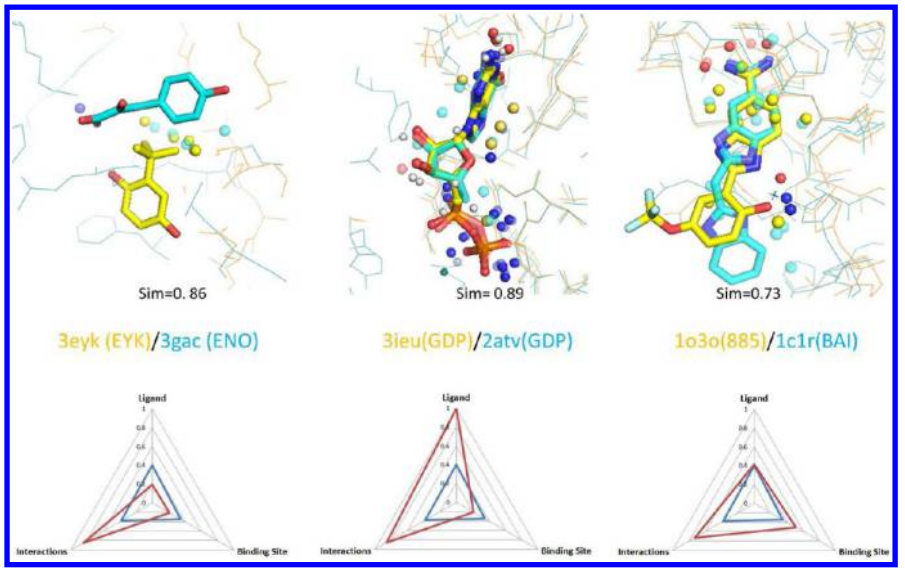


Figure 6. Ishape alignment of protein−ligand complexes. Interaction points are labeled for the two complexes, according to their pharmacophoric properties (hydrophobic, yellow or cyan; H-bond donor, blue; H-bond acceptor, red, negative ionizable, blue; positive ionizable, red; metal chelation, white). Bound ligands heavy atoms are displayed by cpk-colored sticks. The radar plot on lower panel indicates the pairwise similarity of the corresponding ligands (Tanimoto coefficient on ECFP4 fingerprints), binding sites (Shaper similarity), and interaction patterns (IShape similarity). The blue line indicates the similarity threshold for the three metrics, the red line indicates the similarity values for the current protein−ligand complex. (A) IShape alignment of an influenza hemagglutinin-inhibitor complex (PDB code: 3eyk, HET code: EYK, yellow) and of a macrophage inhibitory factor-substrate complex (PDB code: 3gac, HET code: ENO, cyan). (B) IShape alignment of protein ERA-GDP complex (PDB code: 3ieu, HET code: GDP, yellow) and of a RAS-like estrogen-regulated growth inhibitor RERG-GDP complex (PDB code: 2atv, HET code: GDP, cyan). (C) IShape alignment of two trypsin-inhibitor complexes (PDB code: 1o3o, HET code: 885, yellow; PDB code: 1c1r, HET code: BAI, cyan).

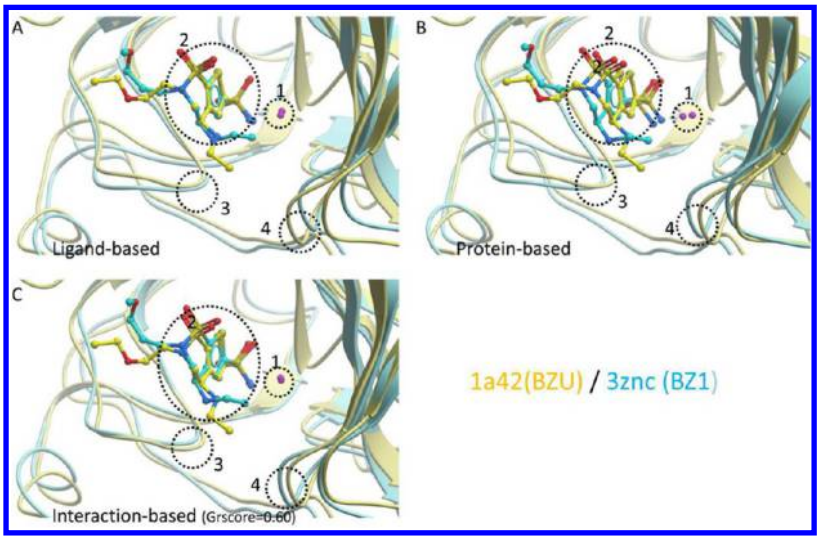


Figure 7. Ligand-based (panel A), protein-based (panel B), and interaction-based (panel C) alignments of two complexes of brinzolamide bound to human carbonic anhydrase II (yellow ribbons, PDB id: 1a42) and murine carbonic anhydrase IV (cyan ribbons, PDB id: 3znc). The bound inhibitor is displayed by cpk-colored ball and sticks (yellow carbon atom, 1a42-bound inhibitor; cyan carbon atom, 3znc-bound inhibitor). The catalytic zinc ion is displayed by a magenta ball. Four areas of interest are circled: (1) catalytic ion; (2) scaffold of the bound inhibitor; (3 and 4) site-enclosing loops.

Interaction Pattern Similarity Depends Tightly on Binding Site Similarity. The above-described similarity offers us the opportunity to check across all sc-PDB complexes for a possible dependence between the similarity of protein−ligand interactions and the corresponding ligand and/or protein binding site similarities. The similarity of two protein−ligand complexes was then computed with three different metrics: the pairwise similarity of their ligands (Tanimoto coefficient on ECFP4 and MACCS fingerprints), the pairwise similarity of their binding sites (Shaper similarity), and the pairwise similarity of their interaction patterns (IShape similarity). Ligands were considered similar if their Tanimoto coefficient was above 0.4 for circular ECFP4 fingerprints or above 0.7 for MACCS keys. Binding sites were considered similar if their pairwise Shaper similarity was above 0.35.[29] Last, interaction patterns were considered similar if their IShape similarity was above 0.41, as previously suggested in Table 2. Plotting these three possible values against each other clearly shows that interaction pattern similarity is not related to ligand similarity whatever the descriptor used (Figure 5A and B) but strongly correlates with binding site similarity (r = 0.876, sd = 0.11; Figure 5C). As to be expected, there are very few cases of similar interaction patterns

dx.doi.org/10.1021/ci300566n | J. Chem. Inf. Model. 2013, 53, 623−637

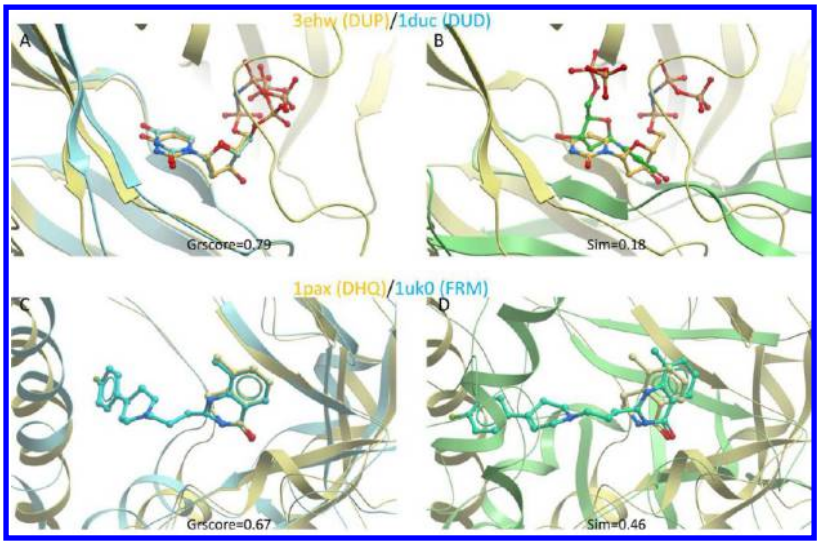


Figure 8. IShape versus Grim alignments for difficult cases. (A) Grim alignment of two related ligands (DUP: α,β-imido-dUTP, DUD: dUDP) cocrystallized with dUTPase as a trimer (PDB code: 3ehw, yellow ribbons) or monomer (PDB code: 1duc, cyan ribbons). Bound ligands are represented by cpk-colored ball and sticks (carbon in 3ehw, yellow; carbon in 1duc, cyan; oxygen, red; nitrogen, blue; phosphorus, orange. The Grim Grscore is indicated below the alignment. (B) Ishape alignment of two related ligands (DUP: α,β-imido-dUTP, DUD: dUDP) cocrystallized with dUTPase as a trimer (PDB code: 3ehw, yellow ribbons) or monomer (PDB code: 1duc, green ribbons). Bound ligands are represented by cpk-colored ball and sticks (carbon in 3ehw, yellow; carbon in 1duc, green; oxygen, red; nitrogen, blue; phosphorus, orange. The Ishape similarity score is indicated below the alignment. (C) Grim alignment of two inhibitors (DHQ, FRM) cocrystallized with poly(ADP-ribose) polymerase (PDB code: 1pax, yellow ribbons; PDB code: 1uk0, cyan ribbons). Bound ligands are represented by cpk-colored ball and sticks (carbon in 1pax, yellow; carbon in 1uk0, cyan; oxygen, red; nitrogen, blue; fluorine, light green). The Grim Grscore is indicated below the alignment. (D) Ishape alignment of two inhibitors (DHQ, FRM) cocrystallized with poly(ADP-ribose) polymerase (PDB code: 1pax, yellow ribbons; PDB code: 1uk0, cyan ribbons). Bound ligands are represented by cpk-colored ball and sticks (carbon in 1pax, yellow; carbon in 1uk0, cyan; oxygen, red; nitrogen, blue; fluorine, light green). The Ishape similarity score is indicated below the alignment.

between dissimilar ligands and dissimilar binding sites (lower left quadrant, Figure 5D). In most of the cases, this relates to small molecular weight ligands exhibiting a simple and promiscuous hydrophobic interaction pattern (see a prototypical example Figure 6A). Cases for which similar ligands exhibit similar interaction patterns to dissimilar binding sites are still rare (upper left quadrant, Figure 5D). This situation mainly occurs either when one of the two binding sites undergoes a significant change (mutation, monomer vs dimer-lining interface) without altering ligand recognition, or for primary metabolite-binding sites (e.g., GDP-binding sites, Figure 6B) which have evolved to share conserved features even in absence of sequence and fold conservation. Interestingly, as far as binding sites are similar, interaction patterns are conserved irrespectively of the corresponding ligand similarity (in 93 and 88% of cases for similar and dissimilar ligands, respectively; Figure 5D). Despite a bias due to the still limited ligand diversity among sc-PDB ligands, this observation suggests that a single interaction mode to a single druggable cavity remains the rule because a few key interactions to a few key residues need to be fulfilled to achieve significant binding. Careful inspection of ligand structures revealed that dissimilar ligands sharing both a conserved cavity and interaction pattern are usually sharing a common substructure which is the main anchoring moiety to their target (e.g., trypsin inhibitor binding to the catalytic site, Figure 6C).

Some Applications of Interaction Pattern Fingerprints and Graphs. *Interaction-Based Alignment of Protein–Ligand Complexes.* When aligning protein–ligand complexes, modelers usually have the choice between two scenarios: (i) align protein-bound ligands hoping that protein atoms will overlap adequately and (ii) align protein atoms (main chain or all heavy atoms) hoping that ligand atoms will match. IShape and Grim enables to merge both options by focusing on interaction patterns, thereby optimizing both target and ligand alignment in a single step. A prototypical example of the advantage in aligning interaction patterns is provided Figure 7 in the alignment of two complexes between the inhibitor brinzolamide and two carbonic anhydrase isoforms (carbonic anhydrase II, PDB code 1a42; carbonic anhydrase IV, PDB code: 3znc). A ligand-based alignment is not satisfactory because of the flexibility of the two alkyl side chains that induces a significant shift of the ligand and the proteins (Figure 7A). A sequence-based structural alignment of both proteins (using the SYBYL "Align Structure by Homology" method) better matches the two proteins structures with however some mismatches in loops enclosing the binding site and in the bicyclic scaffold of the inhibitor (Figure 7B). The best compromise is obtained by the interaction pattern alignment generated by Grim (Figure 7C) which optimally matches all partners (inhibitor, protein, zinc) at the same time since it simultaneously takes the three kinds of IPAs (Centered, InterLig, InterProt) into consideration during the graph alignment procedure.

In most cases, the alignments produced by IShape (shape-based alignment) and Grim (graph-based alignment) are similar. We however recommend the usage of Grim which is insensitive to the difference in the number of IPAs between the reference and the fit object. Significant variations in size of either the target or the ligand will produce two interaction patterns, one being a subset of the other. A local alignment (Grim) will therefore usually outperforms a global match (IShape) in these conditions, as shown in the following two examples. In the first one, the same target (dUTPase) is complexed to two related ligands (α,β-imido-dUTP in 3ehw, dUDP in 1duc) but in different oligomeric states (one molecule at the interface of a trimer in 3ehw,

dx.doi.org/10.1021/ci300566n | J. Chem. Inf. Model. 2013, 53, 623–637

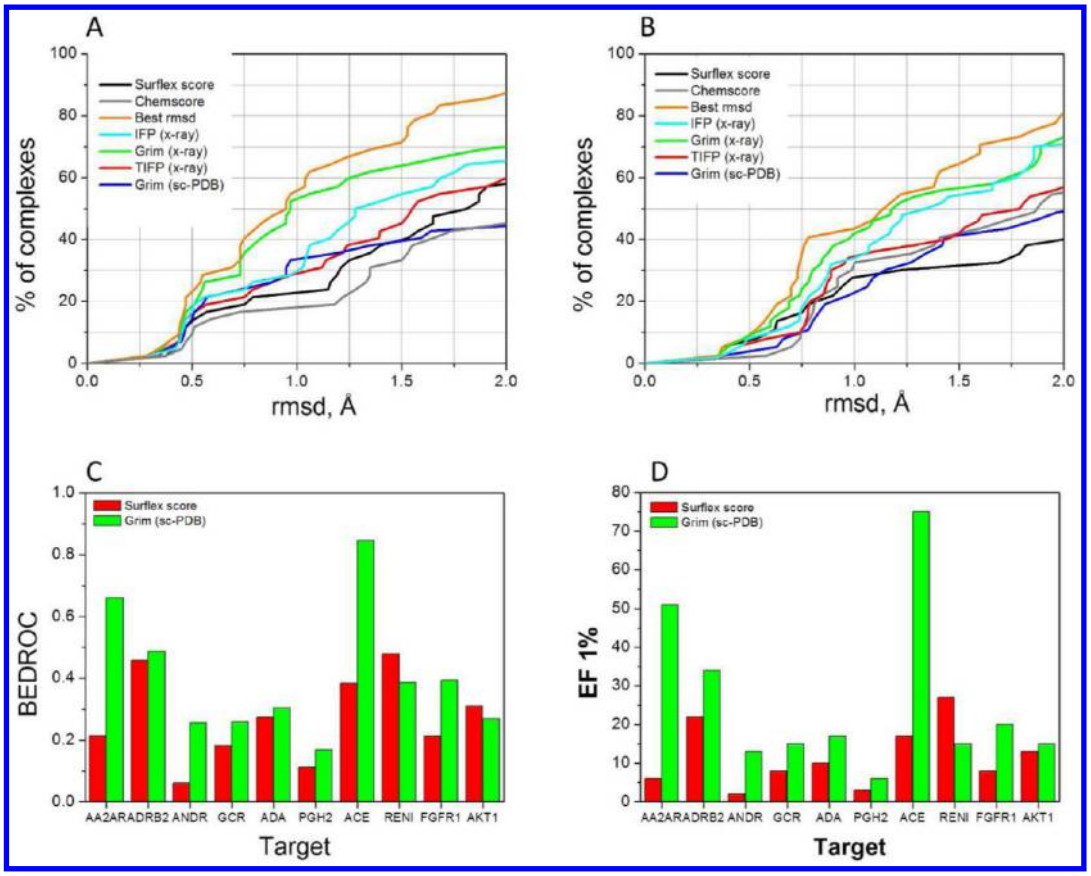


Figure 9. Postprocesssing docking poses by similarity of interaction fingerprints (IFP) and interaction patterns (TIFP, Grim). (A) Posing accuracy or 42 low molecular weight ligands[16] (set 2) obtained upon Surflex-Dock docking. For each complex, top ranked poses are stored according to the native Surflex-Dock score (black line), the Tanimoto similarity of standard interaction fingerprints to that of the native X-ray pose (IFP-xray, cyan line), the Tanimoto similarity of interaction fingerprint triplets to that of the X-ray pose (TIFP-xray, red line), the similarity of interaction pattern graphs to that of the X-ray pose (Grim-xray, green line), or the similarity of interaction pattern graphs to that of X-ray poses of all sc-PDB ligands sharing the same target (Grim-scPDB, blue line). (B) Posing accuracy of 36 ligands from the CCDC/Astex subset. For each complex, top ranked poses are stored according to the native Surflex-Dock score (black line) or the similarity of interaction pattern graphs to that of X-ray poses of all sc-PDB ligands sharing the same target (blue line). (C) Surflex-Dock vs Grim scoring of 10 poses for a set of DUD-E[34] actives and decoys and 10 representative targets: adenosine A2A receptor (AA2AR), beta2 adrenergic receptor (ADRB2), androgen receptor (ANDR), glucocortocoid receptor (GCR), adenosine deaminase (ADA), prostaglandin G/H synthase 2 (PGH2), angiotensin-converting enzyme (ACE), renin (RENI), fibroblast growth factor receptor 1 (FGFR1), RAC-alpha protein kinase (AKT1). The top-ranked pose according to the Surflex-Dock score was retained. For Grim rescoring, scores were fused by ligand and protein−ligand sc-PDB reference complexes of the same target and the best Grscore retained. The discrimination of actives from inactives is measured by the area under the BEDROC[49] curve. (D) Enrichment in true actives at a constant 1% false positive rate upon scoring docking poses by either the native Surflex-Dock score or the Grim Grscore. Targets and ligand sets are identical to that indicated in panel C.

monomer in 1duc). As a consequence, the 3ehw ligand exhibits many more interactions (46 interactions, 36 IPAs) than the 1duc ligand (14 interactions, 11 IPAs). Whereas Grim accommodates fairly well this discrepancy by finding the largest common subgraph and perfectly aligning both complexes (Figure 8A), IShape fails to align the two protein−ligand complexes by proposing a global shape matching that does not coincide with a proper alignment (Figure 8B). In the second example, variations occur at the ligand level with one ligand (PDB code: 1pax) being a substructure of the second one (PDB code: 1uko), both compounds being cocrystallized with chicken and human poly(ADP-ribose) polymerase. IPA numbers being quite different in both cases (17 in 3ehw, 37 in 1uk0), the same erroneous alignment is produced by IShape whereas Grim perfectly overlaps both protein−ligand structures (Figure 8C and D).

Visual inspection of overlaid protein−ligand complexes is a common task for modelers involved in structure-based lead optimization programs. Aligning these molecular objects by focusing on interactions and not structures permits to compare ligands (from a single series) in complex with a single protein, but also multiple ligands complexed to homologous targets. One of the two alignment tools proposed here (Grim) is particularly interesting since it is insensitive to large variations in one of the two partners and should therefore be of interest for target family based ligand optimization as well as for structure-based fragment growing for example.

Postprocessing Docking Poses. The very first motivation in designing the TIFP fingerprint was to remove the dependency of standard IFPs to the active site definition. To check the comparative performance of the conventional IFP and the newly defined TIFP in rescoring docking poses, a data set of 42 protein−ligand complexes[16] (set 2) in which ligands have been chosen to cover fragment-like space was retained. We previously reported for this data set that IFP rescoring (selecting the pose that has the highest IFP similarity to the X-ray solution) was superior to conventional docking scores in self-docking experiments.[16] The 42 fragments were docked again to their cognate protein X-ray structure with Surflex-Dock,[46] and 20 poses were

dx.doi.org/10.1021/ci300566n | J. Chem. Inf. Model. 2013, 53, 623−637

Table 3. Area under the ROC Plot of a Binary Classification (Active, Inactive) of Docked Poses to the X-ray Structure of 10 Representative Targets

		DUD-E		ROC[a]		BEDROC[b]		EF1[c]		sc-PDB
	PDB code	actives	decoys	SF-Dock	Grim	SF-Dock	Grim	SF-Dock	Grim	refs
G protein-coupled receptors										
adenosine A2A receptor (AA2AR)	3pwh	482	31500	0.736	0.911	0.214	0.660	6	51	4
beta2 adrenergic receptor (ADRB2)	3ny8	231	15000	0.854	0.846	0.457	0.487	22	34	3
nuclear hormone receptors										
androgen receptor (ANDR)	2am9	269	14350	0.470	0.730	0.060	0.256	2	13	29
glucocortocoid receptor (GCR)	1p93	258	15000	0.557	0.742	0.183	0.259	8	15	8
other enzymes										
adenosine deaminase (ADA)	1a4l	93	5450	0.828	0.749	0.274	0.304	10	17	20
prostaglandin G/H synthase 2 (PGH2)	3nt1	435	23150	0.620	0.626	0.113	0.169	3	6	7
proteases										
angiotensin-converting enzyme (ACE)	3zqz	282	16900	0.840	0.952	0.383	0.845	17	75	14
renin (RENI)	3sfc	104	6958	0.878	0.850	0.478	0.386	27	15	29
protein kinases										
fibroblast growth factor receptor 1 (FGFR1)	3tt0	139	8700	0.721	0.836	0.213	0.394	8	20	7
RAC-alpha protein kinase (AKT1)	4ekl	293	16450	0.759	0.709	0.310	0.270	13	15	10

[a]Area under the ROC curve for a binary classification of ligands (actives, decoys) from their docked poses scored by either the Surflex-Dock score (SF-Dock) or the Grscore of the interaction pattern graphs (Grim). [b]Area under the Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) curve. [c]Percent of actives found when 1% of decoys have been retrieved.

generated and stored for every ligand. Poses were scored according to six scoring functions: (i) the native docking score (Surflex score), (ii) the Chemscore empirical scoring function,[47] (iii) IFP similarity to the native X-ray pose (IFP-xray score), (iv) interaction pattern graph similarity to the X-ray pose (Grim-xray), (v) TIFP similarity to the X-ray pose (TIFP-xray score); and (vi) interaction pattern graph similarity to that of any known sc-PDB ligand of the same target (Grim-scPDB). The top scored pose for every scoring scheme was retained and its root-mean square (rmsd) deviation to the X-ray pose computed.

Considering the docking successful if the top-ranked pose deviates less than 2.0 Å from the X-ray solution, we confirmed that the quality of poses generated by Surflex-Dock is quite remarkable in ca. 90% of the cases. We also confirm that rescoring according to IFP similarity statistically enhances the quality of the top ranked posed (65% of success vs 58% for the Surflex score and 44% for Chemscore rescoring; Figure 9A). It is noteworthy that transforming interaction fingerprint (IFP) into interaction fingerprint patterns (TIFP) slightly alters the quality of the rescoring with only 60% of success, thereby not providing any significant advantage with respect to energy scoring at least for this data set and the default Surflex scoring function. We suspect, as previously reported before, that some noise has been introduced in TIFP representation because of the large proportion of hydrophobic interaction-containing triplets which dominates the similarity calculation. To avoid this flaw, poses were rescored by interaction pattern graph similarity with Grim. Grim rescoring significantly improves the performance of the rescoring (70% of success, Figure 9A) notably for high precision poses (<1 Å rmsd). Graph-based rescoring is very efficient provided that a good pose is available within the pool of possible solutions and does not suffer from the previously reported TIFP drawback since graph nodes are weighted according to the scarcity of the encoded interaction. If we do not consider the best rmsd score which indeed is the best possible score according to the quality of generated poses, Grim rescoring with respect to the X-ray pose (Grim_xray) is the only method that is statistically superior to the native Surflex-Dock scoring (Kolmogorov–Smirnov test, $D = 0.3156$; $p = 0.016$)

To remove all possible bias, we discarded the true X-ray pose as a reference for interaction pattern similarity and used instead all existing sc-PDB ligands cocrystallized with the target under investigation (Grim-scPDB rescoring). This scoring procedure is only possible with either interaction pattern fingerprints or graphs and not doable with conventional IFPs since defining a unique binding site of constant composition is almost impossible for most targets. Reference sc-PDB entries were selected according to the recommended Uniprot name of the target of interest and provided for each ligand a variable number of references (from none to 115). If the target was absent in the sc-PDB (10 out of 42 cases), the closest sc-PDB target was manually selected as a surrogate (see full list of references in the Supporting Information Table 2). In four cases (PDB entries 1pu7, 1pu8, 1qpr, 1qy2), no surrogate could be found and no rescoring solution could be proposed. As to be expected, rescoring by similarity of interaction pattern to target-specific sc-PDB ligand sets (Grim_scPDB score) is inferior to rescoring with respect to the true X-ray solution (Figure 9A, Kolmogorov–Smirnov test $D = 0.2821$, $P = 0.0073$). It however presents the considerable advantage of a target-specific rescoring taking into account all available structural information and does not necessitate the selection of a particular reference for comparing target-ligand interactions. For the data set under investigation, we should however acknowledge that this advantage with respect to conventional energy-based scoring (Surflex, Chemscore) vanishes for pose prediction with rmsd higher than 1.5 Å.

On the second data set of 36 CCDC/Astex complexes, rather similar trends were observed (Figure 9B). Surflex provided adequate poses for 80% of the ligands. Rescoring using the true X-ray pose evidently provides a significant advantage with a better performance of graph rescoring (Grim) and conventional IFP scoring (70% of good poses) with respect to TIFP scoring (57% of success only) or energy-based scoring (40% and 55% of success for Surflex and Chemscore, respectively). Grim-scPDB rescoring also produces better poses than the conventional Surflex score but not Chemscore rescoring for this data set (Figure 9B). We clearly acknowledge that a much larger docking

dx.doi.org/10.1021/ci300566n | J. Chem. Inf. Model. 2013, 53, 623–637

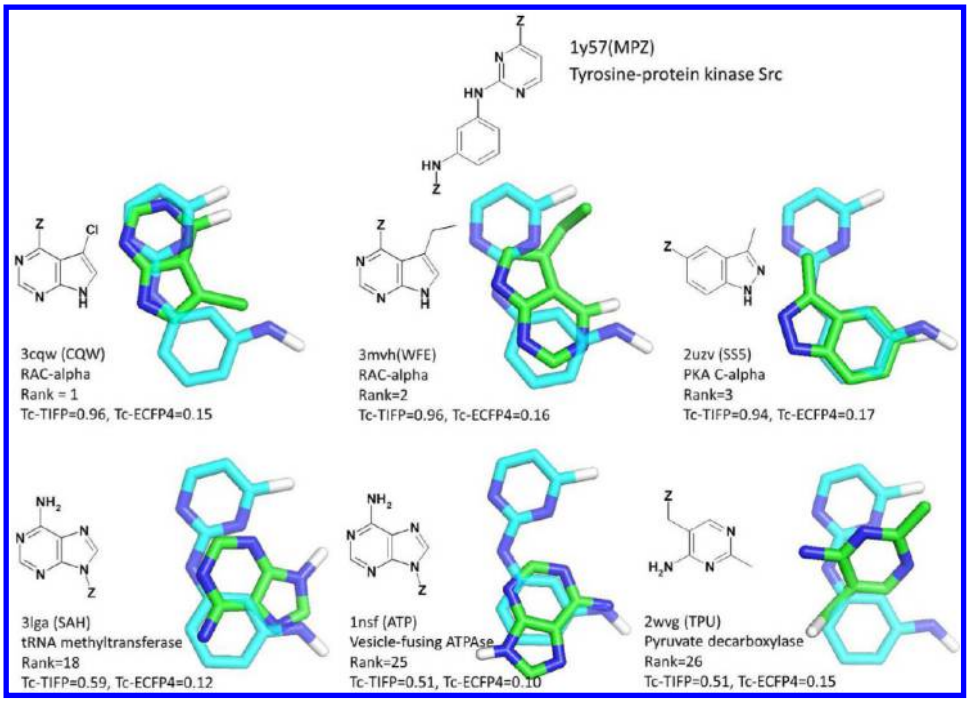


Figure 10. Search for bioisosteric scaffolds to the 1-*N*-(pyrimidin-2-yl)benzene-1,3-diamine fragment bound to tyrosine-protein kinase Src (PDB code: 1y57, HET code: MPZ). In the upper panel are displayed the top three scored scaffolds bound to a protein kinase (according to TIFP fingerprint similarity) and aligned with Grim to the reference (reference carbon atoms in cyan, aligned carbon atoms of the selected scaffold in green, oxygen and nitrogen atoms in blue and red, respectively). PDB and HET codes of the aligned ligands, common name of the aligned scaffold-bound protein, rank of the selected scaffold (scored by decreasing TIFP similarity), and pairwise protein-scaffold similarities (TIFP similarity, ECFP4 similarity) are indicated for every hit. In the lower panel are displayed the top three scaffolds bound to a nonprotein kinase target. The Z atom indicates the branching points generated by RECAP fragmentation.

set may be necessary to really appreciate the benefit of Grim rescoring on the quality of docking poses for known actives.

We next ask the question whether a real advantage is also found in virtual screening experiments for which Grim-scPDB rescoring may be particularly adapted to distinguish true actives from decoys. For that purpose, ten targets of pharmaceutical interest (Table 3) covering 5 major target families (proteases, G protein-coupled receptors, other enzymes, protein kinases, nuclear hormone receptors), along with a set of prepared actives and decoy ligands, were chosen from the DUD-E data set.[34] To avoid any possible bias in Grim-scPDB rescoring, caution was given to select for each target an X-ray structure absent from the sc-PDB data set, or to remove it from the pool of references. Starting from the same set of docking poses generated for each target by Surflex-Dock, the respective ability of the native Surflex-Dock score and of the Grscore to discriminate actives from chemically similar decoys was further inspected. In 5 out of 10 cases (AA2AR, ANDR, GCR, ACE, FGFR1), the area under the ROC curve was significantly improved (more than 0.1 unit) upon Grim rescoring (Table 3). In three cases (ADRB2, PGH2, RENI) both scoring methods could be considered as equally potent in segregating actives from decoys. A slight advantage of the Surflex-Dock native scoring function could only be found in the remaining two cases (ADA, AKT1). In virtual screening, it is however of utmost importance to enrich the list of top scorers (to be experimentally confirmed) in true actives. Following accepted recommendations,[48] we therefore focused the analysis in early enrichment in true actives by computing two important statistical parameters: the Boltzmann enhanced discrimination of the ROC (BEDROC) curve[49] as well as the enrichment in true actives at the low false positive rate of 1% (EF1). Both metrics demonstrates an enhanced advantage in rescoring docking poses with Grim in 8 out of 10 cases when considering the BEDROC metric, and in 9 out of 10 cases when considering the EF1 value (Table 3, Figure 9). From 1.5−4 times more true actives at a constant 1% false positive rate are found with the Grim rescoring method, therefore demonstrating its usefulness in virtual screening scenarios. For 7 out of the 10 targets, the benefit in Grim rescoring was related to the number of protein−ligand X-ray reference structures. However, it is currently impossible to draw general conclusions since such retrospective virtual screening studies are known to be very dependent on the chosen ligand set (actives and decoys) and protein coordinates (e.g., active vs inactive state of a receptor). The herein proposed Grim rescoring mode is however very interesting since it enables a user-independent scoring strategy capitalizing on existing knowledge about known ligand binding modes to the target of interest. Since sc-PDB protein−ligand interaction patterns only have to be computed once and are further stored in a look-up table, postprocessing is relatively straightforward and only necessitates (in addition to the set of docked poses) the name of the target of interest.

Scaffold Hopping with Interaction Pattern Conservation. Since the TIFP fingerprint generically describe spatial interactions between a ligand and its target, it can be applied to search for truly bioisosteric scaffolds to any reference of known binding mode. All 9877 sc-PDB ligands were thus fragmented according to retrosynthetic RECAP rules, therefore generating a set of 20 839 fragments whose TIFP fingerprint was deduced, after pairwise atom matching, from that of the full fingerprint from the parent ligand. To establish the proof-of-concept, we selected as reference the 1-*N*-(pyrimidin-2-yl)benzene-1,3-diamine scaffold

dx.doi.org/10.1021/ci300566n | *J. Chem. Inf. Model.* 2013, 53, 623−637

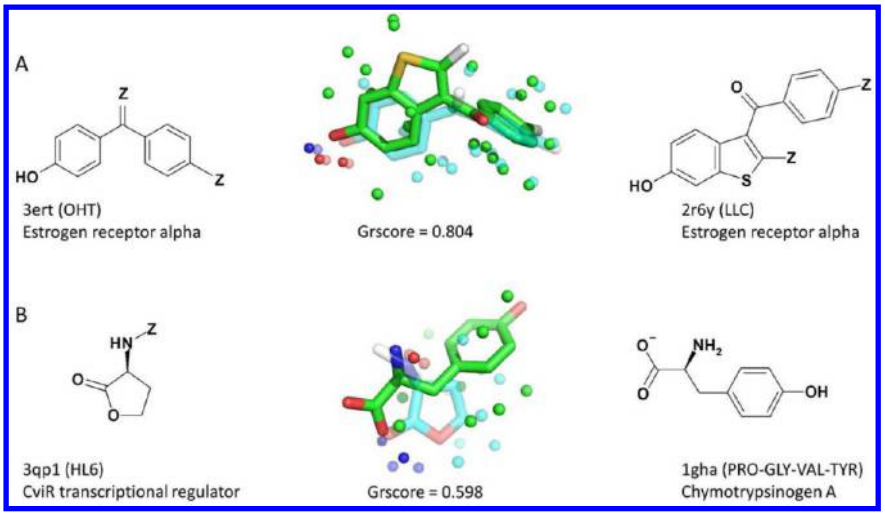


Figure 11. Grim alignment of the top-ranked bioisosteric scaffold to two references, one fragment from the estrogen receptor alpha ligand OHT (panel A, PDB id: 3ert), one from the CviR transcriptional protein ligand HL6 (panel B, PDB id: 3qp1). The structure of the most bioisosteric scaffold (HET code, PDB id) and its target protein are indicated on the right-hand side. Query (cyan carbon atoms) and top-ranked scaffolds (green carbon atoms) are aligned according to Grim along with the fitted IPAs (green, hydrophobic interactions of the query; cyan, hydrophobic interactions of the hit; red, hydrogen bond (protein acceptor); blue, hydrogen bond (protein donor)). IPAs of the query are displayed by transparent spheres, IPAs of the hits are displayed by solid spheres. The similarity scores of the interaction pattern graphs (Grscore) are indicated below the alignments.

(Figure 10) from a tyrosine-protein kinase inhibitor (HET code: MPZ, PDB id: 1y57). This scaffold presents the prototypical interaction observed between most ATP-competitive inhibitors and the hinge region of protein kinases. A bioisostere to this scaffold was here defined as any sc-PDB fragment fulfilling a TIFP similarity higher than 0.50 and a chemical similarity (expressed by a Tanimoto coefficient on ECFP4 fingerprint) below 0.25. Out of the 27 selected fragments, 23 originate from protein kinase inhibitors (Supplementary Table 8) and also interact with the above-described kinase hinge region. Aligning the corresponding fragments from their interaction patterns (IPAs in Centered mode) confirm that all hits are really bioisosteric to the reference, the h-bond acceptor and donor atoms (to the hinge region) being well matched (Figure 10). Interestingly, some hits could be retrieved from ligands interacting with unrelated proteins (methyltransferase, ATPase, pyruvate decarboxylase, sugar epimerase) albeit with scaffolds (adenine, aminopyrimidine) frequently observed in protein kinase inhibitors. Of course, the likelihood of finding bioisosteric scaffolds is higher among ligands sharing the same target or target class (e.g., estrogen receptor alpha-binding scaffolds, Figure 11A). However, fragment interaction patterns may be conserved among completely unrelated proteins exhibiting only subpocket similarities (e.g., transcriptional regulator-bound 3-aminooxolan-2-one and chymotrypsinogen-bound tyrosine, Figure 11B).

Traditional sources for finding bioisosteric groups rely on existing structure−activity (SAR) knowledge.[50,51] Computational approaches to find potential replacements may be derived from pairwise 2D and 3D similarity searches.[52,53] A few methods focusing on existing protein−ligand 3D structures have been reported but are restricted to different ligands complexed with the same target,[54] or require prior knowledge of similar binding sites.[55,56] To the best of our knowledge, we report here the first method considering biosiosteric searches from a set of existing protein−ligand interactions in the PDB with no a priori on either ligand and/or binding site similarity.

CONCLUSION

We herewith propose a generic fingerprint (TIFP) of protein−ligand interaction patterns as well as two computational methods (IShape, Grim) to efficiently compare and align protein−ligand complexes. The TIFP fingerprint currently describes standard intermolecular interactions but could be easily extended to less frequent, weaker, but sometimes important interactions like weak hydrogen bonds (C−H···O), halogen bonds, or cation-(donor)-π interactions. It enables an ultrafast comparison of protein−ligand complexes but suffers from the predominance of hydrophobic contacts among most PDB-ligand X-ray structures. We therefore prefer to directly manipulate the interaction pattern as a list of pseudoatoms describing the interactions engaged between the ligand and the target. Interaction patterns were computed for 10 000 protein−ligand complexes of the sc-PDB data set, therefore providing a framework for two important applications: (i) postprocessing docking poses while taking into account all known interactions with the target of interest and (ii) search for truly bioisosteric scaffolds to a reference by prioritizing substructures exhibiting conserved interaction patterns. Last, the proposed alignment tools (notably the Grim method) enables to directly fit protein−ligand complexes from the corresponding interaction patterns, without having to choose between a ligand-based or a target-based point of view. With the rapid growth of structural information in the Protein Data Bank, such methods are believed to play an important role in assisting molecular modelers to visualize common features or differences among protein−ligand complexes of biological interest.

ASSOCIATED CONTENT

Supporting Information

Set 1 of 900 pairs of similar and 900 pairs of dissimilar protein−ligand complexes from the sc-PDB database; set 2 of 42 fragments selected for Surflex-Dock docking; set 3 of 36 high-resolution protein−ligand structures from the clean CCDC/Astex Set; pharmacophoric properties of protein and ligand atoms; rules for detecting protein−ligand interactions; color force-field to postprocess shape matching in IShape; distance

dx.doi.org/10.1021/ci300566n | J. Chem. Inf. Model. 2013, 53, 623−637

thresholds (Å) between IPAs for generating a graph node; sc-PDB fragments bioisosteric to 1-*N*-(pyrimidin-2-yl)benzene-1,3-diamine (HET code: MPZ, PDB id: 1y57). This material is available free of charge via the Internet at http://pubs.acs.org

AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The Calculation Center of the IN2P3 (CNRS, Villeurbanne, France) and the GENCI (Project x2010075024) are acknowledged for allocation of computing time. The Institut de Recherches Servier (Croissy/Seine, France) is warmly acknowledged for a doctoral grant to J.D. and for useful discussions.

REFERENCES

(1) Tan, L.; Batista, J.; Bajorath, J. Computational methodologies for compound database searching that utilize experimental protein-ligand interaction information. *Chem. Biol. Drug. Des.* **2010**, *76*, 191−200.

(2) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046−1053.

(3) 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.

(4) Leslie, C.; Eskin, E.; Noble, W. S. The spectrum kernel: a string kernel for SVM protein classification. *Pac. Symp. Biocomput.* **2002**, 564−575.

(5) Rognan, D. Structure-Based Approaches to Target Fishing and Ligand Profiling. *Mol. Inf.* **2010**, *29*, 176−187.

(6) Schneider, G. Virtual screening: an endless staircase? *Nat. Rev. Drug Discovery* **2010**, *9*, 273−276.

(7) Crisman, T. J.; Sisay, M. T.; Bajorath, J. Ligand-target interaction-based weighting of substructures for virtual screening. *J. Chem. Inf. Model.* **2008**, *48*, 1955−1964.

(8) Tan, L.; Lounkine, E.; Bajorath, J. Similarity searching using fingerprints of molecular fragments involved in protein-ligand interactions. *J. Chem. Inf. Model.* **2008**, *48*, 2308−2312.

(9) Bock, J. R.; Gough, D. A. Virtual screen for ligands of orphan G protein-coupled receptors. *J. Chem. Inf. Model.* **2005**, *45*, 1402−1414.

(10) 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.

(11) Wang, F.; Liu, D.; Wang, H.; Luo, C.; Zheng, M.; Liu, H.; Zhu, W.; Luo, X.; Zhang, J.; Jiang, H. Computational screening for active compounds targeting protein sequences: methodology and experimental validation. *J. Chem. Inf. Model.* **2011**, *51*, 2821−2828.

(12) Weill, N.; Valencia, C.; Gioria, S.; Villa, P.; Hibert, M.; Rognan, D. Identification of Nonpeptide Oxytocin Receptor Ligands by Receptor-Ligand Fingerprint Similarity Search. *Mol. Inf.* **2011**, *30*, 521−526.

(13) Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **2004**, *47*, 337−344.

(14) Chuaqui, C.; Deng, Z.; Singh, J. Interaction profiles of protein kinase-inhibitor complexes and their application to virtual screening. *J. Med. Chem.* **2005**, *48*, 121−133.

(15) Deng, Z.; Chuaqui, C.; Singh, J. Knowledge-based design of target-focused libraries using protein-ligand interaction constraints. *J. Med. Chem.* **2006**, *49*, 490−500.

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

(17) Kelly, M. D.; Mancera, R. L. Expanded interaction fingerprint method for analyzing ligand binding modes in docking and structure-based drug design. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1942−1951.

(18) Mpamhanga, C. P.; Chen, B.; McLay, I. M.; Willett, P. Knowledge-based interaction fingerprint scoring: a simple method for improving the effectiveness of fast scoring functions. *J. Chem. Inf. Model.* **2006**, *46*, 686−698.

(19) Venhorst, J.; Nunez, S.; Terpstra, J. W.; Kruse, C. G. Assessment of scaffold hopping efficiency by use of molecular interaction fingerprints. *J. Med. Chem.* **2008**, *51*, 3222−3229.

(20) Perez-Nueno, V. I.; Rabal, O.; Borrell, J. I.; Teixido, J. APIF: a new interaction fingerprint based on atom pairs and its application to virtual screening. *J. Chem. Inf. Model.* **2009**, *49*, 1245−1260.

(21) Schreyer, A.; Blundell, T. CREDO: a protein-ligand interaction database for drug discovery. *Chem. Biol. Drug. Des.* **2009**, *73*, 157−167.

(22) Weisel, M.; Bitter, H. M.; Diederich, F.; So, W. V.; Kondru, R. PROLIX: Rapid Mining of Protein-Ligand Interactions in Large Crystal Structure Databases. *J. Chem. Inf. Model.* **2012**, *52*, 1450−1461.

(23) Meslamani, J.; Rognan, D.; Kellenberger, E. sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins. *Bioinformatics* **2011**, *27*, 1324−1326.

(24) *SYBYL*, version X2.0; Certara: St-Louis, MO, 2012.

(25) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(26) *PipelinePilot*, version 8.5; Accelrys Software Inc.: San Diego, CA, 2012.

(27) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.

(28) *MOE*, version 2011.10; Chemical Computing Group: Montreal, Quebec, Canada, 2011.

(29) Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes. *J. Chem. Inf. Model.* **2012**, *52*, 2287−2299.

(30) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP–retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511−522.

(31) Tanabe, M.; Kanehisa, M. Using the KEGG Database Resource. In *Current Protocols in Bioinformatics*; Wiley: New York, 2012; Chapter 1, Unit1 12.

(32) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726−741.

(33) CCDC/Astex validation set for docking software, http://www.ccdc.cam.ac.uk/products/life_sciences/gold/validation/astex/pdb_entries/ (accessed Jan 2013).

(34) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* **2012**, *55*, 6582−6594.

(35) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(36) *ROCS*, version 3.1.2; OpenEye Scientific Software: Santa Fe, NM, 2012.

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

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

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

(40) 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.
(41) Swamidass, S. J.; Azencott, C. A.; Daily, K.; Baldi, P. A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. *Bioinformatics* **2010**, *26*, 1348−1356.
(42) Barillari, C.; Marcou, G.; Rognan, D. Hot-spots-guided receptor-based pharmacophores (HS-Pharm): a knowledge-based approach to identify ligand-anchoring atoms in protein cavities and prioritize structure-based pharmacophores. *J. Chem. Inf. Model.* **2008**, *48*, 1396−1410.
(43) Bron, C.; Kerbosch, J. Algorithm 457: Finding all cliques of an undirected graph. *Commun. ACM* **1973**, *16*, 575−577.
(44) Johnston, H. C. Cliques of a graph—variations on the Bron−Kerbosch algorithm. *Int. J. Parallel. Prog.* **1976**, *5*, 209−238.
(45) Theobald, D. L. Rapid calculation of RMSDs using a quaternion-based characteristic polynomial. *Acta Crystallogr. A* **2005**, *61*, 478−480.
(46) Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499−511.
(47) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.
(48) Jain, A. N.; Nicholls, A. Recommendations for evaluation of computational methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133−139.
(49) Truchon, J. F.; Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* **2007**, *47*, 488−508.
(50) Wirth, M.; Zoete, V.; Michielin, O.; Sauer, W. H. SwissBioisostere: a database of molecular replacements for ligand design. *Nucleic Acids Res.* **2013**, *41*, D137−D1143.
(51) Ujvary, I. BIOSTER - a database of structurally analogous compounds. *Pestic. Sci.* **1997**, *51*, 92−95.
(52) Schuffenhauer, A.; Gillet, V. J.; Willett, P. Similarity searching in files of three-dimensional chemical structures: analysis of the BIOSTER database using two-dimensional fingerprints and molecular field descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 295−307.
(53) Wagener, M.; Lommerse, J. P. The quest for bioisosteric replacements. *J. Chem. Inf. Model.* **2006**, *46*, 677−685.
(54) Kennewell, E. A.; Willett, P.; Ducrot, P.; Luttmann, C. Identification of target-specific bioisosteric fragments from ligand-protein crystallographic data. *J. Comput.- Aided Mol. Des.* **2006**, *20*, 385−394.
(55) Wood, D. J.; de Vlieg, J.; Wagener, M.; Ritschel, T. Pharmacophore fingerprint-based approach to binding site subpocket similarity and its application to bioisostere replacement. *J. Chem. Inf. Model.* **2012**, *52*, 2031−2043.
(56) Moriaud, F.; Doppelt-Azeroual, O.; Martin, L.; Oguievetskaia, K.; Koch, K.; Vorotyntsev, A.; Adcock, S. A.; Delfaud, F. Computational fragment-based approach at PDB scale by protein local similarity. *J. Chem. Inf. Model .* **2009**, *49*, 280−294.