Journal of
CHEMICAL INFORMATION
AND MODELING

Article

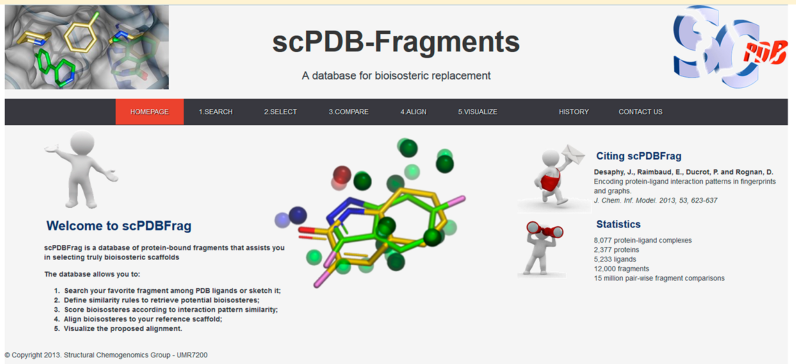
pubs.acs.org/jcim

sc-PDB-Frag: A Database of Protein–Ligand Interaction Patterns for Bioisosteric Replacements

Jérémy Desaphy and Didier Rognan*

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

S Supporting Information



scPDB-Fragments

A database for bioisosteric replacement

Welcome to scPDBFrag

scPDBFrag is a database of protein-bound fragments that assists you in selecting truly bioisosteric scaffolds

The database allows you to:

1. Search your favorite fragment among PDB ligands or sketch it;
2. Define similarity rules to retrieve potential bioisosteres;
3. Score bioisosteres according to interaction pattern similarity;
4. Align bioisosteres to your reference scaffold;
5. Visualize the proposed alignment.

Citing scPDBFrag

Desaphy, J., Raimbaud, E., Ducrot, P. and Rognan, D.
Encoding protein-ligand interaction patterns in fingerprints and graphs.
J. Chem. Inf. Model. 2013, 53, 623-637

Statistics

8,077 protein-ligand complexes
2,377 proteins
5,233 ligands
12,000 fragments
15 million pair-wise fragment comparisons

© Copyright 2013. Structural Chemogenomics Group - UMR7200

ABSTRACT: Bioisosteric replacement plays an important role in medicinal chemistry by keeping the biological activity of a molecule while changing either its core scaffold or substituents, thereby facilitating lead optimization and patenting. Bioisosteres are classically chosen in order to keep the main pharmacophoric moieties of the substructure to replace. However, notably when changing a scaffold, no attention is usually paid as whether all atoms of the reference scaffold are equally important for binding to the desired target. We herewith propose a novel database for bioisosteric replacement (scPDBFrag), capitalizing on our recently published structure-based approach to scaffold hopping, focusing on interaction pattern graphs. Protein-bound ligands are first fragmented and the interaction of the corresponding fragments with their protein environment computed-on-the-fly. Using an in-house developed graph alignment tool, interaction patterns graphs can be compared, aligned, and sorted by decreasing similarity to any reference. In the herein presented sc-PDB-Frag database (http://bioinfo-pharma.u-strasbg.fr/scPDBFrag), fragments, interaction patterns, alignments, and pairwise similarity scores have been extracted from the sc-PDB database of 8077 druggable protein–ligand complexes and further stored in a relational database. We herewith present the database, its Web implementation, and procedures for identifying true bioisosteric replacements based on conserved interaction patterns.

■ INTRODUCTION

Bioisosteric replacement is common practice in medicinal chemistry for changing the structure of an existing lead in order to improve both pharmacological (e.g., potency, selectivity), physicochemical (e.g., solubility), and ADMET (absorption, distribution, metabolism, excretion, toxicity) properties and eventually enabling patentability of the newly designed compounds.[1] From a semantic point of view, bioisosteric changes usually affect either a core heterocycle (scaffold) or its substituents. In the first case, the move, called a scaffold hop,[2] can be classified in four categories (heterocycle replacements, ring opening or closure, peptidomimetics, and topology/shape-based hopping) depending on the complexity of the change.[3] In the second case, smaller changes exemplified by substituent variations led to the concept of matched molecular pairs.[4] Experienced medicinal chemists usually known how to modify a lead while maintaining the parent biological activity. The proposed changes are however biased by their own history and projects to which they contribute. To guide and inspire chemists, databases storing bioisosteric moves gathered from literature appeared in the late 1990s.

The BIOSTER database,[5,6] which registers about 27000 bioisosteric transformations from 36000 compounds from the literature, has been used as a source to develop ligand-based similarity search methods using two-dimensional (2D) fingerprints,[7] molecular fields,[7] topological pharmacophore fingerprints,[8] reduced graphs,[9] or R-group descriptors.[10] Thanks to the rapid growth of public binding data, freely available bioactivity databases[11−13] can now be easily mined to extract allowed bioisosteric changes. For example, the SwissBioisostere database[14] reports 6 million matched molecular pairs from the ChEMBL database,[11] and the influence of user-defined changes on various properties (binding affinity, molecular weight, logP, polar surface area). For every possible move, probabilities are given that the change will positively or negatively affect the

Received: May 9, 2014
Published: July 3, 2014

corresponding in vitro binding affinity. In any case, there is no guarantee that the proposed change might be accommodated by the target protein of interest. For example, changing a 4-methylmorpholine into a 4-methylpyridine is neutral to the biological activity in ca. 50% of the cases, detrimental in 25% of cases, but beneficial in 25% of reported examples.[14] Taking the target into account requires (i) the knowledge of the target three-dimensional (3D) structure, preferably in complex with a low molecular weight compound, and (ii) a relatively small bioisosteric move. The VAMMPIRE database[15] is built on these principles and stores 16300 aligned matched molecular pairs (MMPs) gathered from ChEMBL.[11] A binding mode for a potential bioisostere is obtained by superposition to the parent compound cocrystallized with the same target and further refined by molecular mechanics energy minimization of the full complex. While this strategy is viable for small changes, larger variations might be problematic to predict since a single energy refinement is usually not sufficient to account for significant structural changes of the binding site.

A few methods focusing on existing protein−ligand 3D structures to retrieve potential bioisosteres have been reported but are restricted to different ligands in complex with the same target[16] or require the prior knowledge of similar binding sites.[17] Of interest is the recently described KRIPO method[18] which fragments existing protein-bound ligand X-ray structures into small fragments and further describe their binding subpockets by 3-point pharmacophore fingerprints. Bioisosteric replacements are considered for any fragment that shares a similar binding pocket with a reference substructure. Interestingly, the method enables the superposition of both bioisosteric groups in their respective protein-bound conformation. A small-sized fragment, as usually considered in fragment-based drug discovery,[19] may however bind to very different subsites.[20] The method we present herewith is, to the best of our knowledge, the first approach considering bioisosteric searches with no a priori on either ligand/fragment and/or protein/binding site similarity. Capitalizing on our previous work on converting protein−ligand interaction patterns in either one-dimensional (1D) fingerprints or 3D graphs,[21] we define here bioisosteres as any pair of ligand moieties sharing similar interaction patterns with their native target proteins. The selection directly operates on protein−ligand interaction pattern space and therefore does not require any pairwise similarity calculation on either a pair of compounds or binding sites.

■ METHODS

Database of Protein-Bound Fragments.

The sc-PDB data set (v. 2012) of 8077 druggable protein−ligand complexes[22] was used as a source of 3D information on protein-bound drug-like compounds. Starting from curated 3D mol2[23] files, the bound ligands were fragmented according to two protocols. The first procedure (Figure 1) from here on called “HOME protocol” is inspired by a method reported by Brenk et al.[24] aimed at detecting substituted ring cores. First, a ring perception algorithm[25] was used to automatically detect aromatic and aliphatic rings. Acyclic atoms are then parsed to assign either a linker or substituent label, as whether to the corresponding bonds are connecting two rings or not. Linker atoms are left unchanged. In case of substituent atoms, single bonds involving the closest apolar carbon (in terms of bond distance) to any ring are later cleaved at the condition that the cleaved bond is at least 3 bonds away from the cyclic root atom.

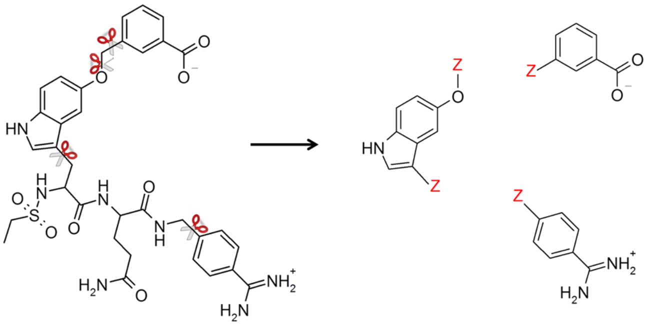


Figure 1. Example of ligand fragmentation (PDB id 2zzu, HET code 359). Cleavable bonds are marked by scissors. A Z pseudoatom indicates the cleavage site for each retained cyclic core.

An anchoring atom (Z label) is then added to each of the remaining fragments to indicate the cleavage site (Figure 1). When applied to the ensemble of 8077 sc-PDB ligands, 7150 entries (88%) could be fragmented into 16224 fragments using the HOME fragmentation protocol. In a second and independent procedure, the same set of sc-PDB ligands were submitted to a retrosynthetic fragmentation using 11 RECAP rules[26] that were directly applied to protein-bound sc-PDB 3D structures using an in-house tool developed around the IChem toolkit.[21] 5930 ligands (73%) could be fragmented into 17295 fragments (13100 cyclic, 4195 acyclic).

Whatever the fragmentation method (HOME or RECAP), each fragment was annotated with descriptors from its parent ligand (HET identifier, fragment number) and PDB target (Uniprot[27] target name, KEGG BRITE[28] functional class). Of course, annotated fragments generated by the first method were not mixed with the set of fragments identified by the second one.

Interaction Patterns.

Protein-fragment interaction patterns were computed with the in-house IChem toolkit.[21] This toolkit computes molecular interactions on-the-fly from the 3D structure of the protein-fragment 3D structure (MOL2 file format) using standard geometrical rules[21] to detect 5 possible interaction types (apolar, aromatic, h-bond, ionic bond, metal chelation). A protein−ligand interaction (Figure 2) is represented by three interaction pseudoatoms (IPA) located at (i) the ligand atom in interaction (*"InterLig"* mode), (ii) the protein atom in interaction (*"InterProt"* mode), (iii) the geometric center of the latter two atoms (*"Centered"* mode). First, every protein-fragment complex was converted into a TIFP fingerprint[21] storing the occurrence of all possible IPA triplets (*"Centered"* mode only) in 210 bins. Then, an interaction pattern graph in which nodes are IPAs (all modes)[21] was created for every protein-fragment pair. Only fragments exhibiting at least 4 distinct interactions (among which at least one is either an aromatic interaction, a hydrogen-bond or an ionic bond) with the local protein environment were kept, therefore pruning the initial list of protein-bound fragments to a total of 11358 molecular entities for the HOME fragmentation protocol and 12343 fragments for the RECAP procedure.

Interaction Pattern and Structural Similarity Searches.

A TIFP similarity value is computed for every pair of TIFP fingerprints using the Tanimoto coefficient as follows

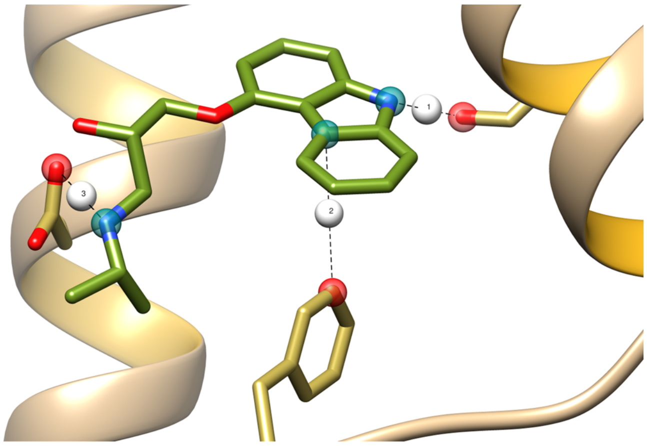


Figure 2. Description of protein−ligand interactions by pseudoatoms. Interactions between the ligand (olive sticks) and the protein (tan sticks) is represented by a dashed line. Each interaction is encoded by pseudoatoms (transparent spheres) located at the ligand-interacting atom (blue sphere, InterLig mode), the protein-interacting atom (red sphere, InterProt mode), and the geometric center of the two heavy atoms in interaction (gray sphere, Centered mode). Only the 'Centered' pseudoatoms (1,2,3) are taken into account to register the corresponding triplet in the TIFP fingerprint, while all interactions pseudoatoms are selected for graph-matching of two interaction patterns. Every pseudoatom has a pharmacophoric property depending on the described interaction (apolar contact, aromatic, H-bond with ligand as donor, H-bond with ligand as acceptor, positive ionizable, negative ionizable, metal coordination).

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

where x_{jA} is the value of bin j in the vector A (reference), and x_{jB} is the value of bin j in the vector B (comparison).

The second computed similarity values applies to interaction pattern graphs (nodes being interaction peudoatoms and edges the distance between them) calculated by the Grim method.[21] Grim uses a clique detection algorithm to find the best possible alignments of two interaction pattern graphs and scores the alignment using an empirical scoring function (Grim score).[21]

Last, third and fourth similarity values were computed, in PipelinePilot,[29] between the molecular graphs of the two fragments under consideration, from their 166 public MDL keys (MACCS) and ECFP4 extended connectivity fingerprints,[30] respectively.

sc-PDB-Frag Database Architecture. In its current version, the interface to the scPDBFrag database is running under the Apache License version 2[31] using Tomcat 6[31] and MySQL 5.0[32] as relational database management system. The Web interface is realized according to a model-view-controller template using vTemplate 1.3.3,[33] is written in PHP 5.3/HTML, and uses the JQuery user interface[34] on top of the jQuery JavaScript Library.[35] MarvinSketch 6.0.4[36] and Open-AstexViewer 3.0[37] modules are utilized to draw fragment structures and visualize their alignments, respectively. The DataTables plug-in[38] is used to organize, filter, and export tabulated results. Export in Microsoft Excel format is realized thanks to the PHPExcel collaborative library.[39]

Querying the Web interface is possible thanks to a wizard (Figure 3) driving the user throughout the 5-step procedure involving the following:

- the definition of the reference fragment (known PDB identifier, user-defined structure sketch);
- the selection of the fragment (identical, most similar) in our fragment database;
- the definition of various filters based on the four computed similarity values, to retrieve potential bioisosteres;
- the choice of the potential bioisosteres to further consider;
- the visualization of the interaction pattern-based alignment between the reference and selected bioisosteres within their respective protein environment.

RESULTS AND DISCUSSION

Ligand Fragmentation. From the starting 8077 sc-PDB ligands, the HOME and RECAP fragmentation protocols generate a comparable number of fragments (16224 and 17295, respectively). Whereas the HOME protocol defines only cyclic moieties, the RECAP procedure yields to 4195 acyclic fragments. Our pruning methodology retains only fragments capable of a sufficient number of protein−ligand interactions (at least 4) out of which one must be directional or polar (aromatic interaction, h-bond, ionic bond). This filter was added to prevent the selection of promiscuous bioisosteres from purely hydrophobic interaction pseudoatoms. Hence, we have noticed that such hits are of low interest, since corresponding protein-fragment interactions have no directionality at all. Using a single alkyl chain as a query for example will therefore return no hits.

The final amount of fragments stored in the database is therefore lower (11358 and 12343 for the two protocols, respectively) out of which about one-third are unique (3978 for the HOME method and 4556 for the RECAP method). In both procedures, the most frequent fragments logically originate from promiscuous cofactors (ATP, NAD, FAD; Supplementary Figure S1). Analysis of standard molecular properties (molecular weight, AlogP, H-bond donor and acceptor counts,

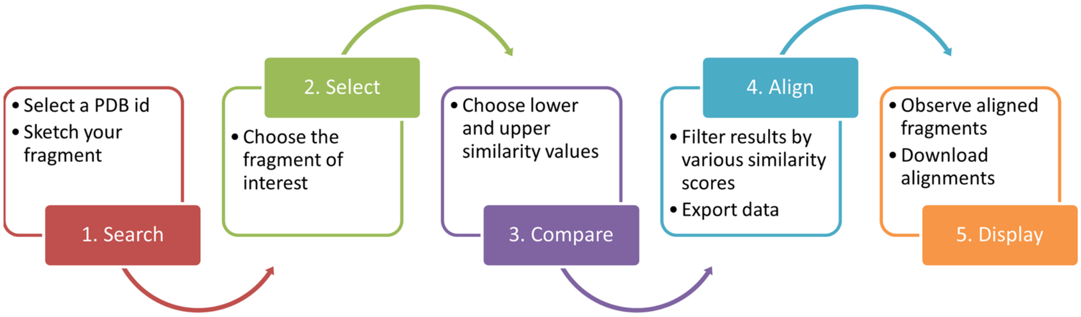

• Select a PDB id
• Sketch your fragment
1. Search
2. Select
• Choose the fragment of interest
• Choose lower and upper similarity values
3. Compare
4. Align
• Filter results by various similarity scores
• Export data
• Observe aligned fragments
• Download alignments
5. Display

Figure 3. Five-step protocol for searching potential bioisosteres in the sc-PDB-Frag database.

rotatable bonds) does not reveal major differences among the two sets of fragments (Table 1, Supplementary Figure S2). Due to the pruning scheme of the fragmentation protocol, HOME fragments are slightly smaller and more rigid. 46 and 48% of the HOME and RECAP fragments, respectively, comply to the Astex rule of 3.[40] On average, each fragment makes 16 interactions with its protein environment with a strong dominance of hydrophobic contacts (ca. 78%) over hydrogen bonds (ca. 18%). Ionic bonds, metal chelation, and aromatic interaction are much less frequent and equally contribute to the remaining interactions (Supplementary Figure S2). Hence, we use relatively stringent rules[21] notably with respect to the distance between the 2 aromatic ring centers ($d \leq 4.0$ Å), which explains the observed low frequency of aromatic interactions (face-to-face and edge-to face interactions are here merged into the same interaction type).

Table 1. Summary of Fragments Properties[a]

	fragmentation method	
property	HOME	RECAP
total number	11358	12343
unique	3978	4556
acyclic	0	435
molecular weight	232.7 ± 101.6	225.6 ± 100.2
AlogP	0.5 ± 2.1	0.6 ± 2.3
H-bond donors	1.7 ± 1.4	1.7 ± 1.3
H-bond acceptors	3.7 ± 2.5	3.6 ± 2.7
rotatable bonds	2.5 ± 2.1	3.3 ± 2.6
total number of interactions	16.1 ± 8.2	16.2 ± 8.0
% Ro-3 compliant[b]	46.3	48.2

[a] Fragments properties were computed by Pipeline Pilot v.9.1.[29] [b] Molecular weight <300, H-bond donors ≤3, H-bond acceptors ≤3, ALogP ≤3.

Pairwise Similarity of Fragments. In the herein developed approach, a bioisostere is defined as any molecular fragment sharing with the query the two following properties: (i) a low chemical similarity (assessed by Tanimoto coefficients from MDL public keys and ECFP4 fingerprints) and (ii) a high interaction pattern similarity (assessed by TIFP fingerprint similarity and the graph alignment Grim score). Based on our experience, the following thresholds have been selected:

Tc-ECFP4 < 0.30
Tc-MDL keys < 0.50
Tc-TIFP > 0.50
Grim score > 0.65

Out of the ca. 15 million possible HOME fragment pairs, only 9556 (0.06%) correspond to a true bioisosteric move. Interestingly, selected bioisosteres correspond to fragments bound to the same target (identical target name) in only 24% of the cases. In 76% of bioisosteric pairs, the change is proposed among a set of fragments derived from ligands bound to related but not identical targets. Last and very interestingly, 26% of proposed changes correspond to completely different protein−ligand environments in which the potential bioisosteric fragment is bound to a totally different target. The percentages given here are just indicative of major trends are should not be strictly considered. The 4 thresholds which are used in the database search enable the selection of a reasonably low number of hits (<20−25) in most cases. Obviously, the Grim score threshold is the parameter that will influence the most the number of hits, provided that the two ligand-based similarity scores (ECFP4 and MDL keys-based Tanimoto coefficient) are meaningful (0.30 and 0.50, respectively). The TIFP similarity score has been chosen to a minimal value of 0.50 but may be varied by 10−15% without major impact on the results.

As an example of a fully conservative change (fragments bound to the same target), a 4-substituted phenol (PDB ID 1X76) can be replaced by a 2-substituted-1,3-benzoxazol-6-ol fragment (PDB ID 1u3r), both originating from ligands cocrystallized with the same target (estrogen receptor beta) and reproducing very similar protein-fragment interactions (Figure 4A). Likewise, selected fragment pairs often come from ligands bound to highly related targets (Figure 4B), like 4-substituted-1,2-dihydrophthalazin-1-one (PDB ID 3uy9, target: tankyrase-2) and 2-substituted-1,3-benzodiazole-4-carboxamide (PDB ID 2RCW, target: poly [ADP-ribose] polymerase 1). Last, there are fewer cases of bioisosteric moves from fragments bound to unrelated targets. An example is given here between 6-hydroxy-1H-indazole-5-carboxamide (PDB ID 4EFU) and 5-chloro-2,4-dihydroxybenzoate (PDB ID 2K8I) which both bind to different targets (Heat shock protein HSP 90-alpha and Pyruvate dehydrogenase lipoamide kinase isozyme 3, respectively) (Figure 4C).

At this point, it must be stated that the proposed bioisosteric change does not implicitly take into consideration the position of exit vectors (or Z groups, displayed in magenta Figure 4B) to reconstruct a novel and full ligand. Once a putative bioisostere has been selected, the linker(s) need therefore to be carefully optimized.

The pairwise chemical similarity among selected bioisosteres does not vary much depending on whether the fragments originate from the same target or just targets from the same class (Figure 5A, B). It should be recalled that the fine specificity of sc-PDB fragments toward related targets is not known but is likely to be very low, therefore explaining the above-described observation. As to be expected, the chemical similarity of the bioisosteric fragments decreases when picked from a ligand bound to a different target class (Figure 5A, B).

Interestingly, the interaction pattern similarity, measured from TIFP fingerprints, is independent of the nature of the targets binding the corresponding bioisosteric pairs (Figure 5C), confirming our previous observation that TIFP similarity is strongly biased by ligand shape conservation and hydrophobic contacts.[21] It should be considered as a simple filter to eliminate fragments pairs not sharing a similar shape. The GRIM graph-based alignment score is much more sensitive to conservation of polar contacts (notably key hydrogen bonds)[21] and is therefore much more dependent on the nature of the bound targets supporting the bioisosteric move (Figure 5D). As seen previously, it does not preclude finding bioisosteres sharing similar interaction patterns with unrelated targets, but the corresponding Grim score is distributed in a very narrow window (0.66−0.68) just above the acceptable similarity threshold of 0.65.

Representative Examples of Bioisosteric Searches. In the following section, we will illustrate two standard ways of querying the database depending on whether the starting fragment is already present or not in or database.

Scaffold Hopping. In this first scenario, we seek to replace the scaffold of a known PDB ligand by that of known inhibitors from the same target or target class. In the present case, we replace the 3-substituted 1,2-dihydroquinoline-2-one scaffold of a checkpoint kinase-1 inhibitor (PDB id 2hxq) by a bioisosteric group while keeping its key interactions with the kinase hinge

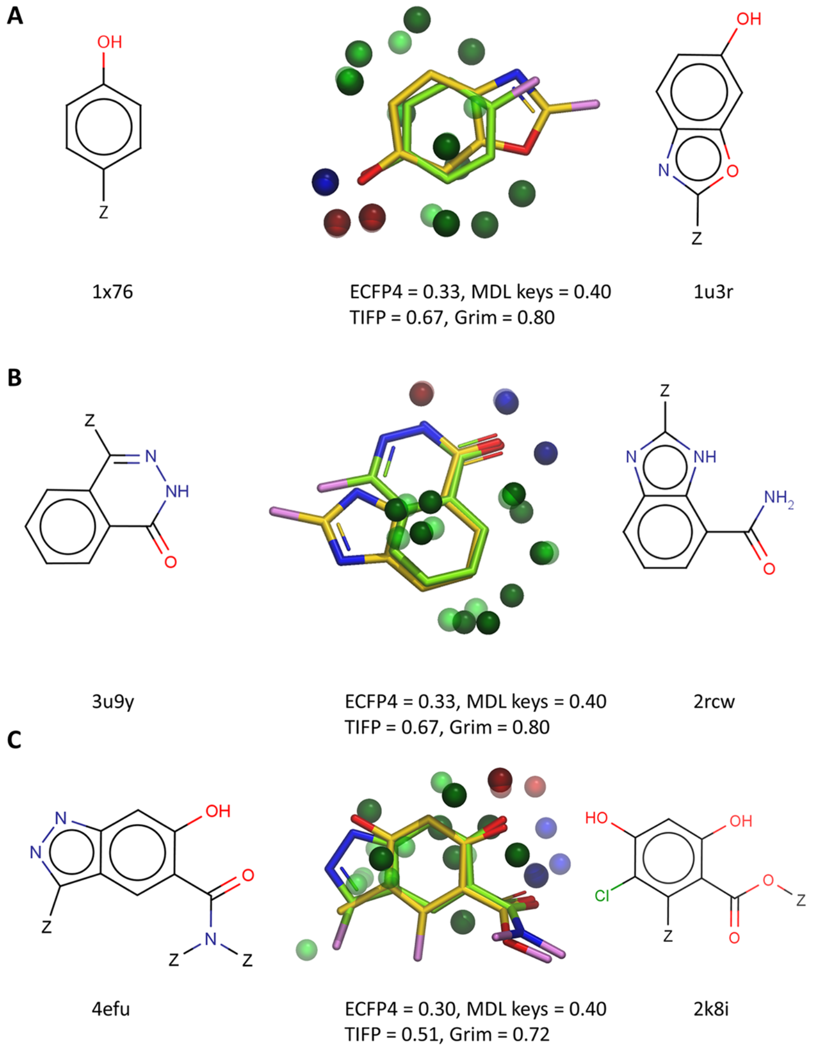


Figure 4. Representative bioisosteric changes. A) Conservative change: fragments are derived from ligands bound to the same target, B) Semiconservative change (ligands bound to related but not identical targets), C) Nonconservative change (ligands bound to unrelated targets). The reference fragment is indicated in the right panel and the bioisostere in the left panel. PDB identifiers of the corresponding protein−ligand complex are indicated below the structures. In the middle panel are presented the Grim-based alignments of the interaction patterns with the corresponding similarity scores. Carbons atoms of the reference and bioisisoteric fragment are colored in green and yellow, respectively. Nitrogen and oxygen atoms are colored in blue and red, respectively. Interaction pseudoatoms are represented by transparent (reference) and solid (bioisostere) spheres with the following color coding: red, hydrogen-bond (ligand donor) blue, hydrogen-bond (ligand acceptor), green (hydrophobic contact). Target relationships were inferred from their respective KEGG-BRITE functional annotation (second level for nonenzymes) and Enzyme Commission numbers (second level) for enzymes.

region. In the starting panel of our Web application (1. SEARCH; Figure 6A) the user has just to give the correct PDB identifier (2hxq), choose the fragments set (HOME), and click the upper RUN tab and the wizard will proceed to the second phase (2. SELECT) in order to select the appropriate fragment (fragment id 3599, Figure 6B). We can see that this fragment is one of the two which are registered for the parent compound (HET code 373). Alternatively, the user could have directly sketched the 3-substituted dihydroquinolinone structure in the Marvin sketcher and selected the proper identifier among the 15 different proposals. The EXPORT and FILTER upper tabs allows the user to either save the results in various formats or to filter the answers according to any value of any column. Selecting the proper fragment (ID: 3599) is done by clicking the corresponding 'Compare' tab and the wizard brings the user to the next stage (3. COMPARE). The selected fragment structure and its main properties are summarized on the top section of the new window (Figure 6C). In the lower section, 5 sliding bars permit the selection of potential bioisosteres according to the above-described 4 similarity scores as well as molecular weight. The similarity scores correspond to two structural similarity assessments (chemical similarity from ECFP4 fingerprints and public MACCS keys) and two interaction pattern similarities (interaction pattern similarity: TIFP score, graph interaction matching: Grim score). Lower and upper thresholds have been defined according to our previous experience with interaction pattern fingerprints and graphs[21] but can be easily changed by sliding the corresponding rolling bars to the desired values. Once thresholds have been set, clicking the upper RUN tab bring the user to the fourth stage (4. ALIGN) of the wizard in which all potential bioisosteres are listed by decreasing graph interaction matching (GRIM) score (Figure 6D) along with other properties (structure, identifier, PDB id, HET code, bound-target name,

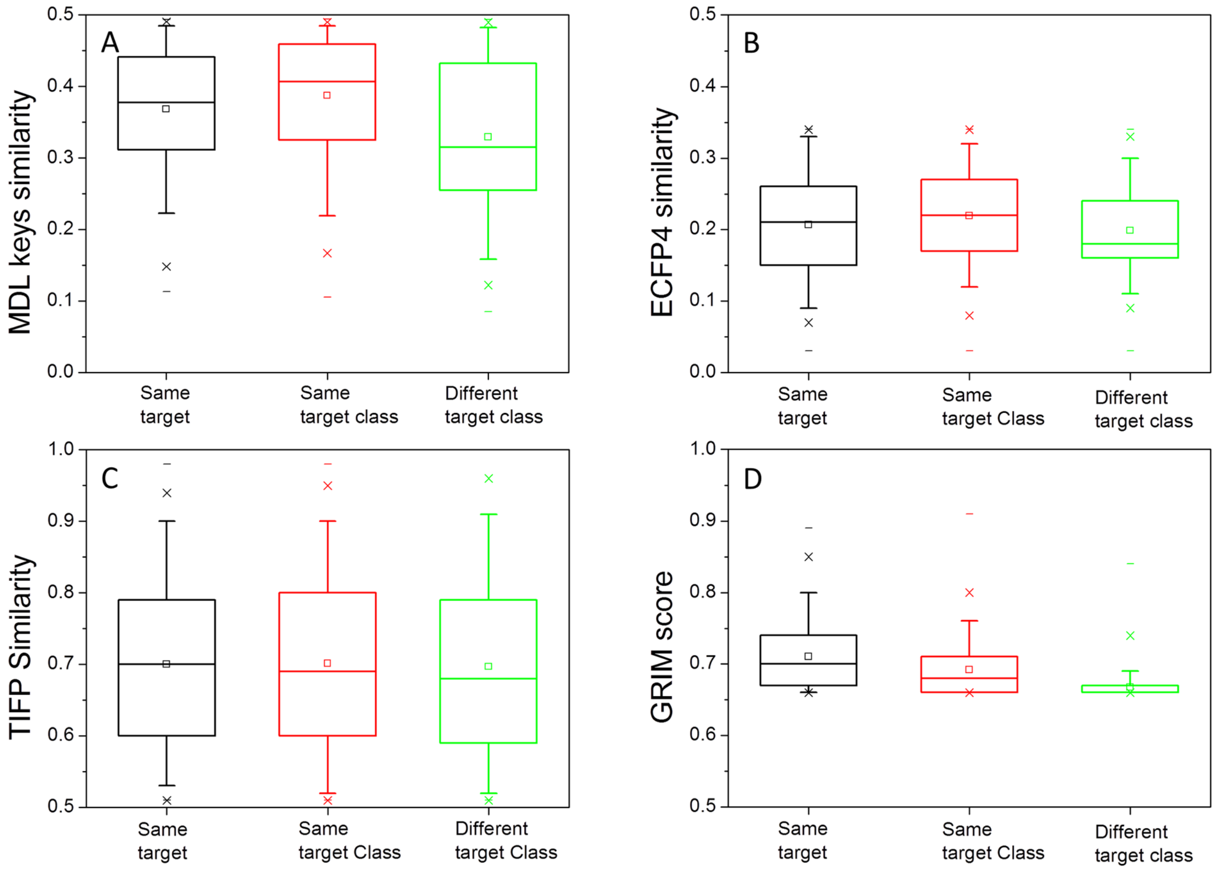
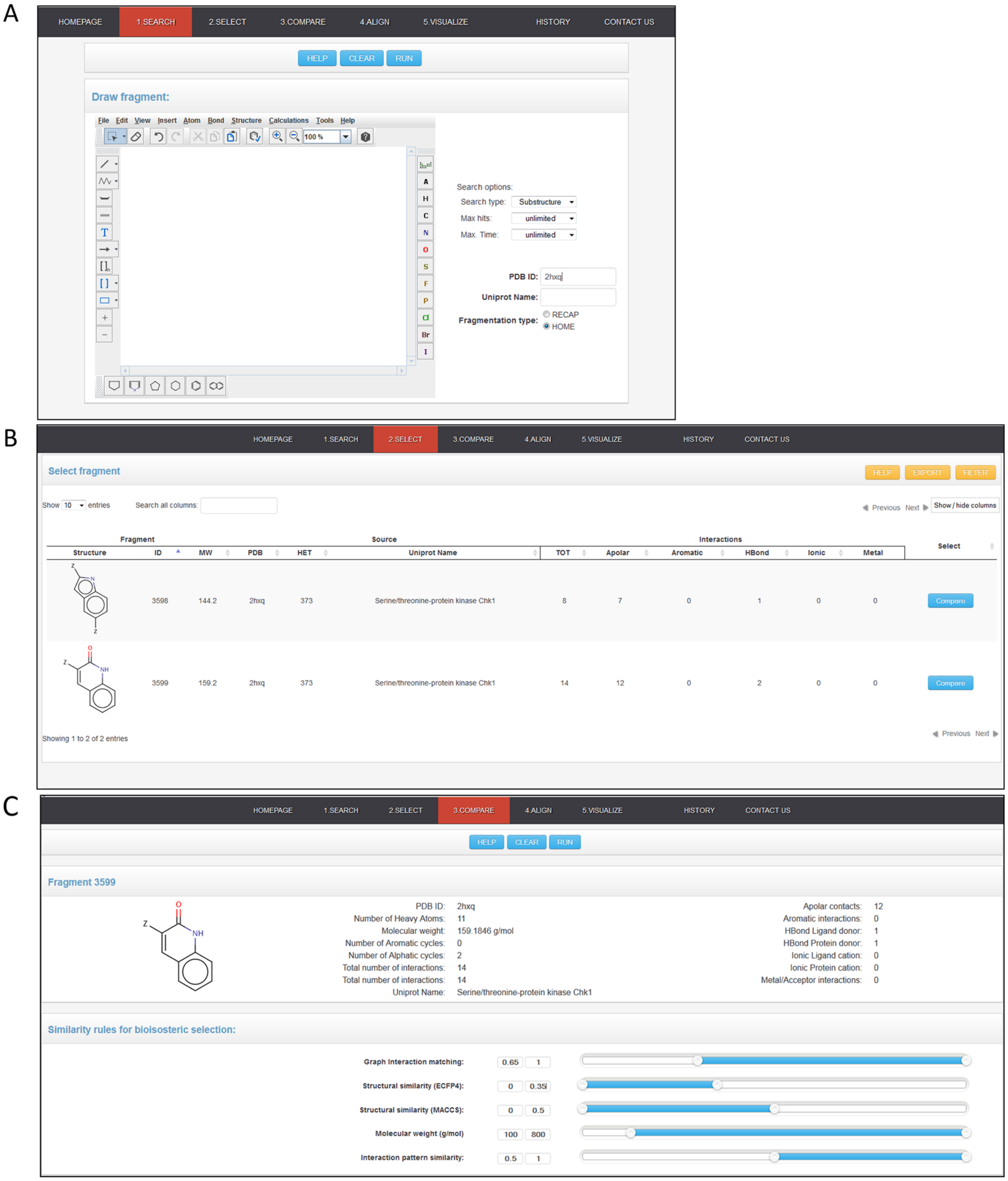


Figure 5. Box-and-whisker plot of the pairwise similarity distribution of HOME-fragmented bioisosteres. A) MDL keys similarity, B) ECFP4 fingerprint similarity, C) TIFP fingerprint similarity, D) Grim graph-based alignment score. The box delimits the 25th and 75th percentiles, the whiskers delimit the 5th and 95th percentiles. The median and mean values are indicated by a horizontal line and an empty square in the box. Crosses delimit the 1% and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash.

molecular weight, GRIM score, MACCS and ECFP4 similarity, TIFP similarity). Among the 27 proposed bioisosteres, 15 originate from known protein kinase inhibitors out of which the top 3 ranked proposals are fragments from inhibitors of the same target (checkpoint kinase-1) as the query fragment (Supplementary Table S1). Analysis can be focused by typing any character in the 'search all columns' tab. In case the proposed bioisostere and the reference fragment share the same target, its name is colored in green. Last, to check the alignment of the proposed scaffold (e.g., the fifth ranked indazole fragment, ID = 3595) to the reference fragment, a graph-based alignment using the in-house Grim algorithm[21] is proposed by clicking the 'Align' tab. The user is brought to the last page of the wizard (5. VISUALIZE section, Figure 6E) in which reference and the selected bioisostere are aligned in 3D together with their color-coded IPAs (*"Centered"* mode only). In the present case, we can clearly see a very good overlap of H-bond acceptor and donor points (to the undisplayed kinase hinge) as well as a good match of hydrophobic interaction points. The OpenAstexViewer browser enables various display scenarios and permits to check whether the exit Z atoms attached to each scaffold are also well matched. In the lower right section (Z information), passing the mouse over each line highlights the corresponding Z groups in the aligned structures and displays the distance between the two exit Z atoms in both fragments as well as the angle between the two exit vectors (C-Z bonds). The case presented here is a true bioisosteric replacement since linking the proposed bioisostere to the remaining fragment of the query is a 12 nM checkpoint kinase-1 inhibitor.[42] The distance (in Å) between the two Z groups and the angle (in degrees) between the two exit vectors are small. Aligned fragments, interaction pseudoatoms, and binding sites (in MOL2 file formats) can be downloaded to enable visualization in any third party software. At any time, the user can go back to any of the preceding steps by selecting the corresponding top menu and modify the course of the wizard by changing prior selections.

Rigidification of a Flexible Fragment. Since HOME fragments are acyclic, looking for rigid bioisosteres of a flexible acyclic group requires searching among RECAP fragments. Drawing for example a valine in the sketcher (Figure 7A) enables the selection of the closest RECAP fragments (178 in total) in the database, sorted by decreasing ECFP4 fingerprint similarity. Selecting the *N*-methyl C-terminal analogue from an HIV-protease inhibitor (ID = 441, Figure 7B) as a reference (by just typing 441 in the 'search all columns' tab) and default similarity thresholds, 7 bioisosteres are proposed (Figure 7C), all being derived from HIV-1 protease inhibitors and constituting rigid valine analogs. Please, note that same fragment can be retrieved several times (e.g., IDs 3137, 327, 2231, 1123) but in a different context (e.g., bound to different X-ray structures of the same protein), thereby leading to different interaction pattern similarity values (GRIM and TIFP similarity scores). Alignment of the query to one hit (fragment 3136) illustrates the perfect match of the corresponding interaction pseudoatoms (Figure 7D).



Search options:
Search type: Substructure
Max hits: unlimited
Max. Time: unlimited

PDB ID: 2hxq
Uniprot Name:
Fragmentation type: ○ RECAP ● HOME

Show 10 entries Search all columns:

Fragment					Source	Interactions						Select
Structure	ID	MW	PDB	HET	Uniprot Name	TOT	Apolar	Aromatic	HBond	Ionic	Metal	
	3598	144.2	2hxq	373	Serine/threonine-protein kinase Chk1	8	7	0	1	0	0	Compare
	3599	159.2	2hxq	373	Serine/threonine-protein kinase Chk1	14	12	0	2	0	0	Compare

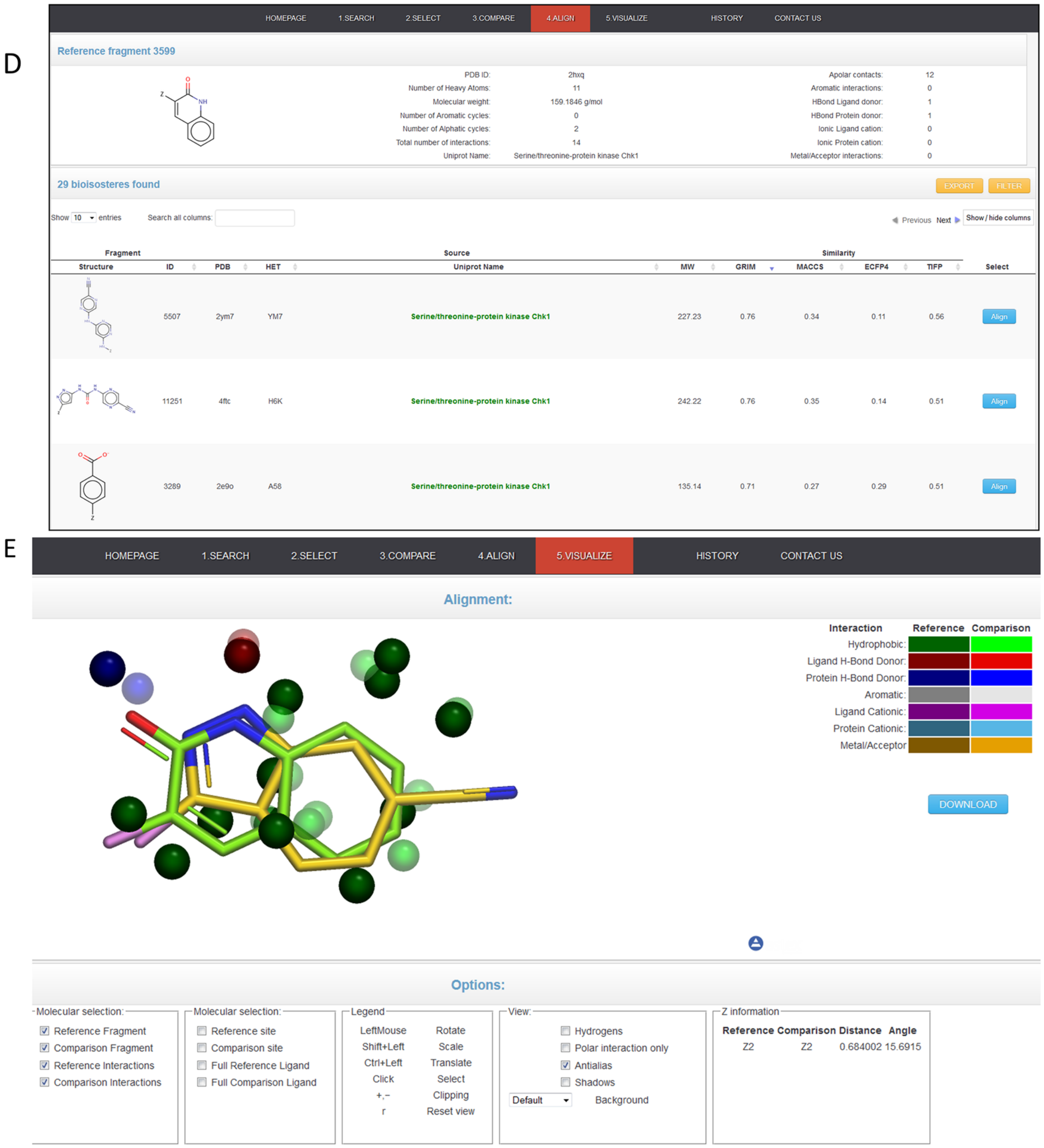
Showing 1 to 2 of 2 entries

PDB ID: 2hxq
Number of Heavy Atoms: 11
Molecular weight: 159.1846 g/mol
Number of Aromatic cycles: 0
Number of Alphatic cycles: 2
Total number of interactions: 14
Total number of interactions: 14
Uniprot Name: Serine/threonine-protein kinase Chk1

Apolar contacts: 12
Aromatic interactions: 0
HBond Ligand donor: 1
HBond Protein donor: 1
Ionic Ligand cation: 0
Ionic Protein cation: 0
Metal/Acceptor interactions: 0

Graph Interaction matching: 0.65 1
Structural similarity (ECFP4): 0 0.35
Structural similarity (MACCS): 0 0.5
Molecular weight (g/mol) 100 800
Interaction pattern similarity: 0.5 1

Figure 6. continued



Reference fragment 3599

PDB ID:	2hxq	Apolar contacts:	12
Number of Heavy Atoms:	11	Aromatic interactions:	0
Molecular weight:	159.1846 g/mol	HBond Ligand donor:	1
Number of Aromatic cycles:	0	HBond Protein donor:	1
Number of Alphatic cycles:	2	Ionic Ligand cation:	0
Total number of interactions:	14	Ionic Protein cation:	0
Uniprot Name:	Serine/threonine-protein kinase Chk1	Metal/Acceptor interactions:	0

29 bioisosteres found

Fragment				Source		Similarity				
Structure	ID	PDB	HET	Uniprot Name	MW	GRIM	MACCS	ECFP4	TIFP	Select
	5507	2ym7	YM7	Serine/threonine-protein kinase Chk1	227.23	0.76	0.34	0.11	0.56	Align
	11251	4ftc	H6K	Serine/threonine-protein kinase Chk1	242.22	0.76	0.35	0.14	0.51	Align
	3289	2e9o	A58	Serine/threonine-protein kinase Chk1	135.14	0.71	0.27	0.29	0.51	Align

Interaction	Reference	Comparison
Hydrophobic:	🟩	🟩
Ligand H-Bond Donor:	🟥	🟥
Protein H-Bond Donor:	🟦	🟦
Aromatic:	⬛	⬜
Ligand Cationic:	🟪	🟪
Protein Cationic:	🟦	🟦
Metal/Acceptor:	🟫	🟧

Molecular selection:
☑ Reference Fragment
☑ Comparison Fragment
☑ Reference Interactions
☑ Comparison Interactions

Molecular selection:
☐ Reference site
☐ Comparison site
☐ Full Reference Ligand
☐ Full Comparison Ligand

Legend

LeftMouse	Rotate
Shift+Left	Scale
Ctrl+Left	Translate
Click	Select
+,−	Clipping
r	Reset view

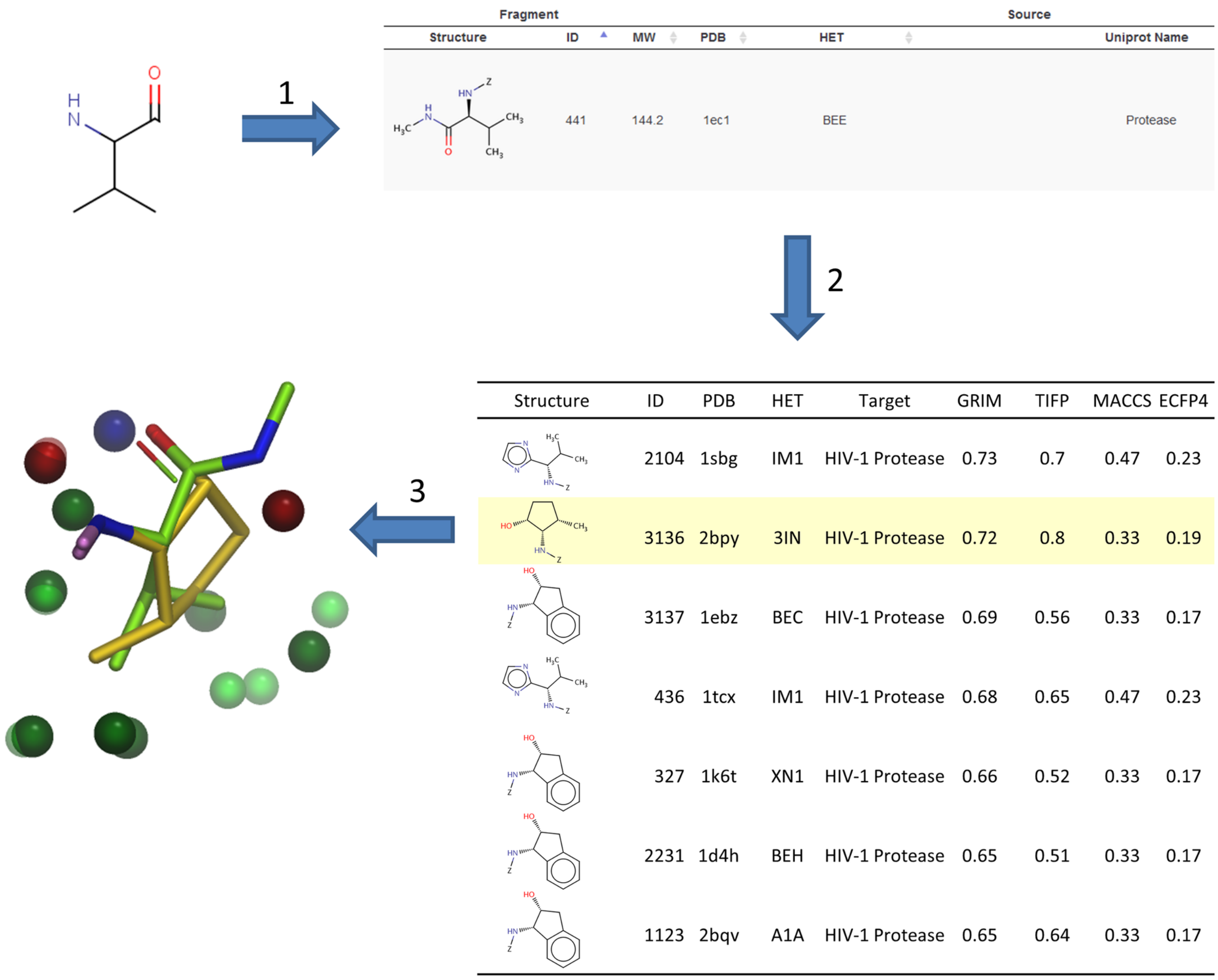
View:
☐ Hydrogens
☐ Polar interaction only
☑ Antialias
☐ Shadows
Default Background

Z information

Reference	Comparison	Distance	Angle
Z2	Z2	0.684002	15.6915

Figure 6. Example of a focused search. A) Selection of the reference fragment by typing a PDB identifier of the parent compound (2hxq). The corresponding fragment and associated Z group(s) could be sketched as well in the Marvin sketcher. B) Selection of 1,2-dihydroquinoline-2-one (ID:3599) among the two proposed fragments for the parent ligand. C) Defining the conditions for retrieving bioisosteres. The structure and main properties of the reference fragment are summarized in the upper panel. In the lower panel, the user defined lower and upper thresholds for 5 properties: graph interaction matching (Grim score: 0.65-1), ECFP4 structural similarity to the reference fragment (0–0.30), MDL keys (MACCS) structural similarity to the reference fragment (0–0.5), molecular weight (100–800), interaction pattern similarity (TIFP score: 0.5–1). D) List of retrieved bioisosteres ranked by decreasing Grim score. For every hit, structures are presented along with PDB annotations (PDB identifier, HET code of the parent compound, target name), molecular weight (MW), and 4 similarity scores (Grim, MACCS, ECFP4, TIFP). E) Graph-based alignment of interaction patterns from the reference (ID: 3599) and the 5th ranked hit (ID: 3595). The structures of the two fragments are displayed as sticks (reference, green carbon; hit, yellow carbon) along with the corresponding interaction pseudoatoms (reference, solid sphere; hit: transparent sphere). Interactions pseudoatoms are color-coded according to the scheme displayed in the upper right panel. In the lower panel, the user can modify the view (e.g., enabling/disabling structures, interaction pseudoatoms and binding sites, hiding hydrogen atoms and apolar interaction pseudoatoms). By default, the full ligand structure from which the fragment originate is undisplayed but can be shown by selecting the corresponding item in the Astex viewer. Z-information box presents distances and angles between exit Z atoms and C-Z exit vectors, respectively.

Comparison to Other Approaches. In order to compare our approach with state-of-the-methods to identify bioisosteres, we repeated the query previously illustrated in Figure 6 (search for bioisosteres to 1-dihydroquinoline-2-one) using a knowledge-based molecular replacement method (SwissBioisostere[14]) and a structure-based scaffold hopping algorithm



Fragment				Source	
Structure	ID	MW	PDB	HET	Uniprot Name
	441	144.2	1ec1	BEE	Protease

Structure	ID	PDB	HET	Target	GRIM	TIFP	MACCS	ECFP4
	2104	1sbg	IM1	HIV-1 Protease	0.73	0.7	0.47	0.23
	3136	2bpy	3IN	HIV-1 Protease	0.72	0.8	0.33	0.19
	3137	1ebz	BEC	HIV-1 Protease	0.69	0.56	0.33	0.17
	436	1tcx	IM1	HIV-1 Protease	0.68	0.65	0.47	0.23
	327	1k6t	XN1	HIV-1 Protease	0.66	0.52	0.33	0.17
	2231	1d4h	BEH	HIV-1 Protease	0.65	0.51	0.33	0.17
	1123	2bqv	A1A	HIV-1 Protease	0.65	0.64	0.33	0.17

Figure 7. Search for valine biosisosteres. The 2D structure is sketched in the wizard and leads (step 1) to chemically close RECAP fragments sorted by decreasing ECFP4 fingerprint similarity. Using default similarity thresholds, the selection of fragment 441 as a reference yields (step 2) to a list of 7 potential biosisoteres, which all are rigid valine analogs (C). Alignment of the reference to one hit (fragment 3136, step 3) confirms a perfect matching of their interaction patterns with their respective target protein. The structures of the two fragments are displayed as sticks (reference, green carbon; hit, yellow carbon) along with the corresponding interaction pseudoatoms (reference, solid sphere; hit: transparent sphere; same color coding than Figure 6E).

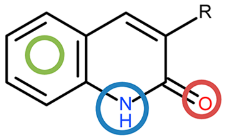
(BROOD).[41] The SwissBioisostere database (http://www.swissbioisostere.ch) contains information on 5 million molecular replacements and their performance in 350 000 biochemical assays gathered from ChEMBL.[11] BROOD generates analogs of a lead by replacing selected fragments (scaffolds) in the molecule with fragments that have similar shape and electrostatics.

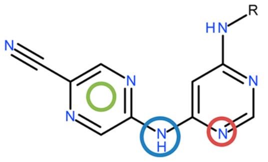
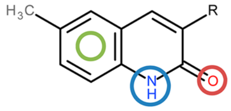
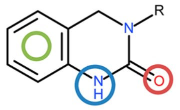
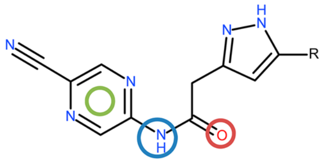
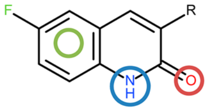
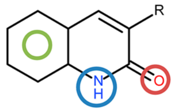
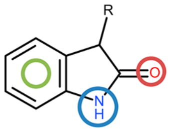
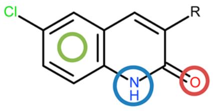
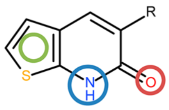
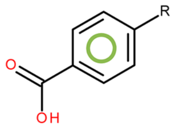
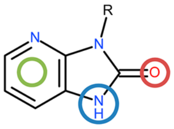
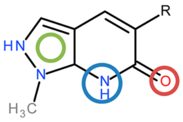
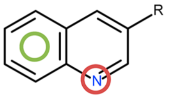
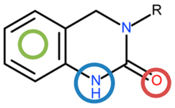
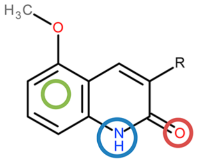
Since we scan a very small chemical space (ca. 12 000 PDB fragments), the herein presented method proposes less hits than SwissBioisostere and BROOD, which browses 5 and 6 million structures, respectively. Top-ranked hits (Table 2) are however much more diverse and chemically dissimilar to the query than proposals of the other software that tend to propose highly similar structures among top-ranked hits (Table 2) with very few overlap with our proposals. We must acknowledge a tendency of our method to generated false positives (e.g., rank 4 and 5, Grim method, Table 2) when the interaction pattern alignments are biased by dominating hydrophobic interaction points and thefore omit a key polar interaction. It is therefore of utmost importance to visualize aligned bioisosteres along with their interaction pseudoatoms to remove this proposals from a wish list. This brief comparison suggests that searching bioisoteric moves from protein−ligand interaction pattern is orthogonal to ligand-based scaffold hopping and molecular replacement methods, as far as only top-ranked solutions (<100) are examined. On the one hand, the former protocol generates ideas and structures that are chemically different from the query but may be difficult to reconcile with respect to full ligand reconstruction, since linking to remaining moieties is not explicitly taken into account. On the other hand, the latter methods suggest moves that can be easily incorporated into new ligands at the cost of a lower chemical diversity. Both strategies are however quite complementary and should be used synergistically.

CONCLUSION

We herewith present a pure structure-based approach to the search for bioisosteric analogues of small fragments whose protein-bound X-ray structure are known. All druggable PDB ligands have been fragmented and the interactions to their protein environment encoded a specific fingerprint. Searching

Table 2. Top 5-Ranked Bioisosteres to 1,2-Dihydroquinolin-2-one[d]



Rank	Grim[a]	SwissBioisostere[b]	Brood[c]
1	 0.76	 0.85	 1.99
2	 0.76	 0.77	 1.99
3	 0.72	 0.77	 1.98
4	 0.71	 0.73	 1.93
5	 0.69	 0.73	 1.74

[a]Scored by decreasing Grimscore.[21] [b]Scored by decreasing replacement score.[14] [c]Scored by decreasing Comboscore (shape and electrostatics) using standard Brood parameters[41] and the protein (PDB 2hxq) as constraint. [d]Green, red, and blue circles represent atoms onto which are mapped the main pharmacophoric features of the query.

for bioisosteres is therefore as simple as identifying fragments which are chemically dissimilar but share similar protein-fragment interaction patterns. By opposition to many approaches, the search for bioisosteres needs no prerequisite on either ligand similarity or binding site similarity since it directly operates at the protein–ligand interaction level. All fragments and interaction patterns have been included in a relational database that can be easily queried to assist the medicinal chemist in proposing, within a few mouse clicks, bioisosteric modifications supported by existing crystallographic data. We acknowledge that biososteric fragments selected by our approach are just novel idea sources for a medicinal chemist and that many more steps have to be processed to ensure that the proposed move is feasible: correct linking of the new bioisosteric group to remaining susbstructures in order to complete ligand definition, possible constrained docking of the full ligand to the cognate binding site, predicting the synthetic accessibility of the full ligand. With contrast to other approaches, our method presents the advantage to be founded on existing protein–ligand X-ray structures and is therefore complementary to pure ligand-based approaches to bioisosteric search.

■ ASSOCIATED CONTENT

S Supporting Information

List of bioisosteres to 1,2-dihydroquinoline-2-one (Table S1), most frequent HOME and RECAP fragments (Figure S1), chemical properties of HOME and RECAP fragments (Figure S2). 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 High-performance Computing Center at the University of Strasbourg (Illkirch, France) and the Calculation Center of the IN2P3 (CNRS, Villeurbanne, France) are acknowledged for allocation of computing time and excellent support. We thank The Institut de Recherches Servier (Croissy/Seine, France) for allocating a Ph.D. grant to J.D.

■ REFERENCES

(1) Meanwell, N. A. Synopsis of some recent tactical application of bioisosteres in drug design. *J. Med. Chem.* **2011**, *54*, 2529−2591.

(2) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 2894−2896.

(3) Sun, H.; Tawa, G.; Wallqvist, A. Classification of scaffold-hopping approaches. *Drug Discovery Today* **2012**, *17*, 310−324.

(4) Leach, A. G.; Jones, H. D.; Cosgrove, D. A.; Kenny, P. W.; Ruston, L.; MacFaul, P.; Wood, J. M.; Colclough, N.; Law, B. Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *J. Med. Chem.* **2006**, *49*, 6672−6682.

(5) IUjváry, I. BIOSTER-a database of structurally analogous compounds. *Pest. Sci.* **1997**, *51*, 92−95.

(6) http://www.digitalchemistry.co.uk/prod_bioster.html (accessed Nov. 2013).

(7) 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.

(8) Wagener, M.; Lommerse, J. P. The quest for bioisosteric replacements. *J. Chem. Inf. Model.* **2006**, *46*, 677−685.

(9) Birchall, K.; Gillet, V. J.; Willett, P.; Ducrot, P.; Luttmann, C. Use of reduced graphs to encode bioisosterism for similarity-based virtual screening. *J. Chem. Inf. Model.* **2009**, *49*, 1330−1346.

(10) Holliday, J. D.; Jelfs, S. P.; Willett, P.; Gedeck, P. Calculation of intersubstituent similarity using R-group descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 406−411.

(11) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2011**, *40*, D1100−D1107.

(12) Irwin, J. J.; Shoichet, B. K. ZINC-a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.

(13) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34*, D668−D672.

(14) Wirth, M.; Zoete, V.; Michielin, O.; Sauer, W. H. SwissBioisostere: a database of molecular replacements for ligand design. *Nucleic Acids Res.* **2012**, *41*, D1137−D1143.

(15) Weber, J.; Achenbach, J.; Moser, D.; Proschak, E. VAMMPIRE: a matched molecular pairs database for structure-based drug design and optimization. *J. Med. Chem.* **2013**, *56*, 5203−5207.

(16) 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.

(17) 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.

(18) 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.

(19) Murray, C. W.; Rees, D. C. The rise of fragment-based drug discovery. *Nat. Chem.* **2009**, *1*, 187−192.

(20) Barelier, S.; Pons, J.; Gehring, K.; Lancelin, J. M.; Krimm, I. Ligand specificity in fragment-based drug design. *J. Med. Chem.* **2010**, *53*, 5256−5266.

(21) Desaphy, J.; Raimbaud, E.; Ducrot, P.; Rognan, D. Encoding Protein-Ligand Interaction Patterns in Fingerprints and Graphs. *J. Chem. Inf. Model.* **2013**, *53*, 623−637.

(22) 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.

(23) Certara USA, Inc., St. Louis, MO 63132, USA.

(24) Brenk, R.; Schipani, A.; James, D.; Krasowski, A.; Gilbert, I. H.; Frearson, J.; Wyatt, P. G. Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem.* **2008**, *3*, 435−44.

(25) Hanser, T.; Jauffret, P.; Kaufmann, G. A. New Algorithm for Exhaustive Ring Perception in a Molecular Graph. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1146−1152.

(26) 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.

(27) UniProt Consortium. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.* **2013**, *41*, D43−D7.

(28) Tanabe, M.; Kanehisa, M. Using the KEGG Database Resource. *Curr. Protoc. Bioinformatics* **2012**, Chapter 1, Unit1 12.

(29) *Pipeline Pilot*, version 9.1; Accelrys, Inc.: San Diego, CA 92121.

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

(31) http://www.apache.org/ (accessed Sep. 2013).

(32) http://dev.mysql.com/ (accessed Sep. 2013).

(33) http://vtemplate.sourceforge.net/ (accessed Sep. 2013).

(34) http://jqueryui.com/ (accessed Sep. 2013).

(35) http://jquery.com/ (accessed Sep. 2013).

(36) ChemAxon Kft: 1031 Budapest, Hungary.

(37) http://openastexviewer.net/web/ (accessed Sep. 2013).

(38) http://www.datatables.net/ (accessed Sep. 2013).

(39) http://phpexcel.codeplex.com/ (accessed Sep. 2013).

(40) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'rule of three' for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8*, 876−877.

(41) *Brood, version 2.0.0*; OpenEye Scientific Software: Santa Fe, U.S.A.

(42) Huang, S.; Garbaccio, R. M.; Fraley, M. E.; Steen, J.; Kreatsoulas, C.; Hartman, G.; Stirdivant, S.; Drakas, B.; Rickert, K.; Walsh, E.; Hamilton, K.; Buser, C. A.; Hardwick, J.; Mao, X.; Abrams, M.; Beck, S.; Tao, W.; Lobell, R.; Sepp-Lorenzino, L.; Yan, Y.; Ikuta, M.; Murphy, J. Z.; Sardana, V.; Munshi, S.; Kuo, L.; Reilly, M.; Mahan, E. Development of 6-substituted indolylquinolinones as potent Chek1 kinase inhibitors. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 5907−5912.