JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Pharmacophore Fingerprint-Based Approach to Binding Site Subpocket Similarity and Its Application to Bioisostere Replacement

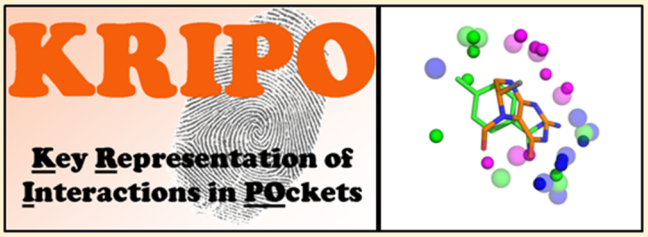
David J. Wood,[†,§] Jacob de Vlieg,[†,‡,⊥] Markus Wagener,[‡,∥] and Tina Ritschel*[,†]

[†]Computational Drug Discovery, CMBI 260, NCMLS, Radboud University Medical Centre, PO Box 9101, 6500 HB Nijmegen, The Netherlands

[‡]MSD, Department of Molecular Design and Informatics, Molenstraat 110, 5342 CC Oss, PO Box 20, 5340 BH Oss, The Netherlands

Supporting Information

ABSTRACT: Bioisosteres have been defined as structurally different molecules or substructures that can form comparable intermolecular interactions, and therefore, fragments that bind to similar protein structures exhibit a degree of bioisosterism. We present KRIPO (Key Representation of Interaction in POckets): a new method for quantifying the similarities of binding site subpockets based on pharmacophore fingerprints. The binding site fingerprints have been optimized to improve their performance for both intra- and interprotein family comparisons. A range of attributes of the fingerprints was considered in the optimization, including the placement of pharmacophore features, whether or not the fingerprints are fuzzified, and the resolution and complexity of the pharmacophore fingerprints (2-, 3-, and 4-point fingerprints). Fuzzy 3-point pharmacophore fingerprints were found to represent the optimal balance between computational resource requirements and the identification of potential replacements. The complete PDB was converted into a database comprising almost 300 000 optimized fingerprints of local binding sites together with their associated ligand fragments. The value of the approach is demonstrated by application to two crystal structures from the Protein Data Bank: (1) a MAP kinase P38 structure in complex with a pyridinylimidazole inhibitor (1A9U) and (2) a complex of thrombin with melagatran (1K22). Potentially valuable bioisosteric replacements for all subpockets of the two studied protein are identified.



INTRODUCTION

In the lead optimization stage of drug discovery, drug designers are often in the position where a lead candidate exhibiting a desired biological activity cannot be advanced further because of issues with its wider pharmacological profile. Common issues include insufficient oral bioavailability, problems with metabolic stability, or toxicity caused by nonselective binding or reactive functional groups. Bioisosterism represents a key strategy for improving the pharmacological profiles of lead candidates and involves replacing undesirable structural features with bioisosteric groups: molecular structures that are known to exhibit the same biological activities. To allow bioisosteric replacement strategies to be fully exploited, scientists require well-organized sources of information on possible replacements.

Traditional sources of bioisosteres include published reviews on bioisosterism that summarize the most versatile replacements from the literature.[1−8] The Bioster database collects these data in electronic form and currently contains over 24 000 pairs of related structures observed to exhibit similar biological properties.[9,10] More recently, computational methods for the identification of potential replacements have been proposed. IBIS is an approach that uses special 2D pharmacophore fingerprints that were optimized to identify potential bioisosteric replacements.[11] Kennewell et al. identified target-specific bioisosteres from the Protein Data Bank by aligning related protein structures and extracting ligand substructures occupying the same binding site regions as bioisostere.[12] A recent work of Moriaud et al. extended the work of Kennewell et al. to target specific and nontarget specific biosisostere replacement rules.[13]

In this paper, we present KRIPO (Key Representation of Interaction in POckets), a new method for mining crystal structure databases for bioisosteres. Binding sites with similar structures are identified with a novel, fast method for quantifying local binding site similarities that is independent of the proteins' sequence similarities or structural similarities. Comparing almost 300 000 entries in the KRIPO database to a query takes about 1 min on a Intel Core2 Duo CPU E6550, 4 GB RAM, using one thread only. Since only features of the proteins' 3D structures are used, the scope of the method is not restricted to intraprotein family comparisons but can also be applied to compare binding site structures across protein families, leading to a higher number of more diverse results.

The assessment of similarity between protein binding sites has been an area of major research interest over the past decade and a wide range of methods have been proposed. Some examples are as follows. CavBase was developed at the

Received: February 7, 2012
Published: July 25, 2012

   dx.doi.org/10.1021/ci3000776 | J. Chem. Inf. Model. 2012, 52, 2031−2043

University of Marburg and uses a clique detection approach on pharmacophoric representations of the binding sites to determine similarity.[14−17] Jackson et al. developed a binding site database called SitesBase, which identifies similar pairs of binding sites with a geometric hashing algorithm.[18] FLAP, developed at the University of Perugia in collaboration with Pfizer and Molecular Discovery Limited, uses pharmacophore fingerprints derived from GRID maps of the binding sites.[19] FLAP has been further developed to allow virtual high-throughput screening of proteins for predicting ligand activity across related targets and protein−protein interactions.[20] PocketMatch represents binding site structures with a list of 90 sorted distances that capture the sites' shapes and chemical natures.[21] Contributions from Hoffmann et al. represents each binding site as a cloud of atoms and assesses the similarity by an alignment of two atom clouds.[22] FuzCav stores binding site information as pharmacophore triplets from the Cα atomic coordinates of binding-site-lining residues, which are then used to perform similarity searches.[23] While methods for calculating binding site similarities are now well-established, application to local binding site regions is still a relatively new idea. Ramensky implemented a local similarity method that allows users to fill protein binding sites with atoms and small fragments that have been observed in similar protein structures.[24] More recently Moriaud et al. developed a local method for larger fragments with MED-SuMo technology;[25−27] Wallach et al. developed a method that characterizes protein-small-molecule binding patterns on a subcavity-level, enabling the discovery of similarities between structurally diverse proteins;[28] and Durrant et al. developed a program called CrystalDock that analyses binding site microenvironments defined by pocket-lining amino acid residues and places ligand fragments that are identified to occur in similar environments into the binding sites.[29]

In KRIPO, local protein binding sites are represented as protein structure-based pharmacophores that are converted into fingerprints, thus enabling extremely rapid comparisons to be made. The binding site fingerprint representation was optimized to identify similar binding sites both within and across protein families. Any fragments identified by KRIPO to bind to similar binding sites will form similar interactions with the protein and thus may represent possible bioisosteric replacements. In a validation experiment, the similarity distribution of approximately 5000 binding site pairs of known similarity were compared to the distribution of a decoy set of randomly selected binding sites pairs. The value of the approach is also demonstrated by identifying suitable replacements for three substructures of a pyridinylimidazole inhibitor of MAP kinase P38 and three substructures of the thrombin inhibitor melagatran.[30,31]

MATERIALS AND METHODS

Preparation of Protein Structures. Protein structures were prepared using the python module of the YASARA-Structure molecular modeling package.[32] Structures were retrieved from the Protein Data Bank (PDB, April 2011) and processed using the *Clean* protocol, which applies a number of corrections including addition of missing bonds and hydrogen atoms, reassignment of bond orders in protein and hetgroup structures, deletion of alternate atomic locations, and assignment of ionization states of amino acids and hetgroups. Next the protocol *Add Hydrogens and Optimize* was applied to adjust the positions of the hydrogen atoms and to optimize the hydrogen bonding network. This protocol additionally optimizes the tautomeric states of imidazole and indole substructures that are part of ligand structures or amino acid residues (histidine, tryptophan). Nucleic acids, metal ions, solvent molecules, and hetgroups with a molecular weight greater than 800 Da or less than 50 Da were removed. Finally, any remaining hetgroups were used to define the binding sites present in the protein structures. Cofactors or water molecules in contact with the hetgroup do not contribute to the binding site definition.

Generation of Ligand Fragments and Local Binding Sites. Ligands extracted from the PDB entries were fragmented using a set of SMIRKS transformations that cleave bonds connecting cyclic and acyclic parts of the structures.[33] All possible fragments were generated from each of the ligands by cleaving the ligand at one or more of these bond types. Each resulting fragment defined a local binding site region. In total 299 591 fragments were extracted from the PDB, and the fragments, together with representations of their local protein binding sites, form the KRIPO database.

Binding Site Pharmacophores. Local binding sites are defined as the amino acid residues in contact with a ligand or ligand fragment; more specifically, any amino acid with at least one atom within 6 Å of a ligand atom. Features representing intermolecular interactions are placed at geometric positions relative to the amino acids present in the binding site and include hydrogen bond acceptors and donors, aromatic T- and π- interactions, hydrophobic contacts, and positive and negative electronic charges. Once all the pharmacophore features have been placed in the binding site, all features with a distance of greater than 2.5 Å from any of the fragment atoms are removed. The remaining features make up the final protein structure-based pharmacophore representation of the local binding site. A complete list of the pharmacophore feature placements on all amino acids is provided in the Supporting Information (Table SI 1).

Hydrophobic features are placed on hydrophobic amino acid side chains along vectors defined by C−H and terminal C−C bonds, as shown in Figure 1a, and along vectors perpendicular

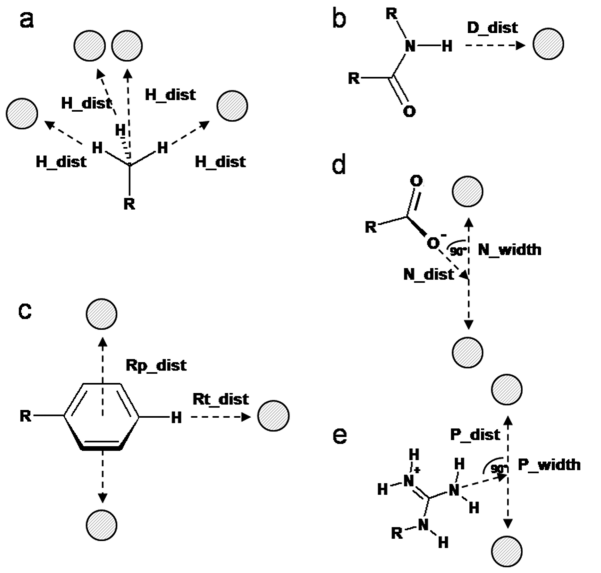


Figure 1. Pharmacophore feature placement: (a) hydrophobic features, (b) donor features, (c) aromatic features, (d) negatively charged features, and (e) positively charged features.

to the center of hydrophobic rings. The parameter *H_dist* specifies the distance from the relevant hydrogen atoms, terminal carbon atoms, or ring centers

Pharmacophore features representing hydrogen bond donors are placed at positions along vectors defined by the N—H and O—H bonds at a distance of *D_dist* from the hydrogen atoms (cf. Figure 1b). For the hydrogen bond acceptors, features are placed along vectors defined by carbonyl C═O bonds and at positions corresponding to the two lone electron pairs on hydroxyl oxygen atoms at a distance of *A_dist* from the oxygen atoms. If a donor and an acceptor pharmacophore feature are placed within 1.2 Å of each other, they are assumed to be interacting and are removed. Placing the hydrogen bond acceptor features along the vector of the carbonyl bond, rather than on each of the lone pairs, was found to result in improved recognition of interacting hydrogen bonding pairs, particularly for β-sheet protein structures.

Both π-stacking and T-stacking aromatic interactions were considered for the binding site pharmacophores, and the placement of the aromatic features is governed by the parameters *Rp_dist* and *Rt_dist*, as shown in Figure 1c. π-Stacking interactions are represented by aromatic features placed at a distance of *Rp_dist* along vectors perpendicular to the ring centers of all aromatic amino acid side chains. T-stacking interactions are represented by aromatic features placed at a distance of *Rt_dist* from the substituent atoms along vectors defined by R—H bonds in tyrosine and phenylalanine amino acids, where R is an aromatic ring atom.

Histidine residues occur in both positively charged and neutral states at physiological pH, and when a ligand forms an interaction, the equilibrium is likely to be shifted to one side or the other. If the ionization state of a histidine residue is assessed by YASARA-Structure to be neutral, a hydrogen bond acceptor feature is assigned to the unprotonated ring nitrogen, and a hydrogen bond donor feature is assigned to the protonated nitrogen. If the histidine side chain is determined by YASARA-Structure to be protonated, hydrogen donor features are placed on each of the two N—H hydrogens. The delocalized positive charge on the protonated ring is represented by a set of 8 positive charged pharmacophore features. The features are placed at positions that correspond to distances of *P_dist* from the ring atoms in the plane of the ring, and distances of *P_width* in the directions perpendicular to the ring systems.

The two parameters *N_width* and *N_dist* control placement of features representing negative charged electrostatic interactions, as shown in Figure 1d. In total, five negative charged pharmacophore features are placed on each side-chain acid group. Four of the features are placed at positions that correspond to distances of *N_dist* from the oxygen atoms along vectors defined by the C—O bond vectors, and distances of *N_width* in the perpendicular direction to the acid groups. The fifth feature is placed at a distance *N_dist* along a vector defined by the C—C bond so that it lies between the oxygen atoms.

The placement of pharmacophore features representing positively charged electrostatic interactions is governed by two parameters: *P_dist* and *P_width*. For lysine side chains, positively charged features are placed along vectors defined by the N—H and C—N bonds at distances of *P_dist* from the hydrogen or terminal nitrogen atoms. For arginine side chains, six positively charged features are placed at positions that correspond to distances of *P_dist* from the nitrogen atoms along vectors defined by the CZ—N bonds and distances of *P_width* in the perpendicular direction to the guanidinium group (Figure 1e).

In Figure 2, the pharmacophore features of the binding site of MAP kinase P38a cocrystallized with a pyridinylimidazole inhibitor (PDB entry 1A9U) are given as an example.[30]

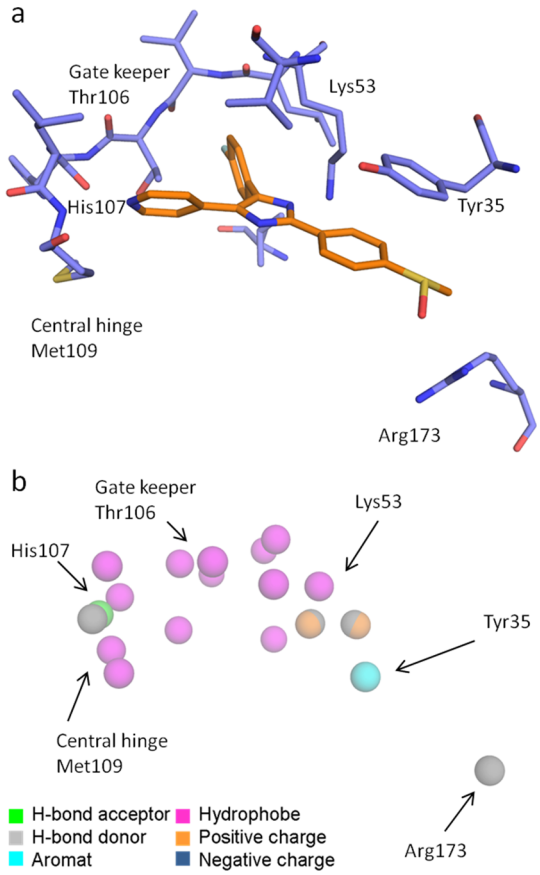


Figure 2. (a) Binding site and (b) resulting binding site pharmacophore of a map kinase P38a structure in complex with a pyridinylimidazole inhibitor (PDB code: 1A9U).

The automatic pharmacophore generation method has identified many of the key structure aspects of the binding site that are known to be important for the design of MAP kinase P38a inhibitors (Figure 2b), including the hydrogen bond donor and acceptor pair in the hinge region of the binding pocket (residues 108 and 109), the hydrophobic pocket surrounding the 3-fluoro benzene group, the positively charged lysine 53 donating a hydrogen bond, an aromatic π-stacking interaction with the tyrosine 35, and a hydrogen bond donor of arginine 173 at the end of the triphosphate binding site.[34]

Pharmacophore Fingerprint Encoding. The process of converting pharmacophores into fingerprints requires that the distances between the features in the pharmacophore are binned into discrete distance ranges. A set of potential binning schemes was generated for each fingerprint from two parameters: the width of the initial bin and the bin width multiplier (the factor by which each successive bin increases in size relative to the previous). An initial bin width *i* of 0.5 and a multiplier *m* of 1.0 would result in a linear binning scheme of 0—0.5—1.0—1.5, ..., whereas an initial width *i* of 1.0 and a multiplier *m* of 1.5 would result in a nonlinear binning scheme

dx.doi.org/10.1021/ci3000776 | J. Chem. Inf. Model. 2012, 52, 2031—2043

of 0−1.0−2.5−4.75−8.125, ... The start of the nth bin can therefore be calculated with the equation below. The final bin is the last bin with a start value below 20 Å and has no upper bound.

$$\text{start}(1) = 0 \qquad \text{for } n = 1$$

$$\text{start}(n) = \sum_{k=0}^{n-2} l \times m^k \qquad \text{for } n > 1$$

Three levels of pharmacophore fingerprint complexity were assessed: 2-, 3-, and 4-point pharmacophore fingerprints with a distance, a triangle, or a tetrahedral of pharmacophore features corresponding to one bit in the fingerprint, respectively. In 2-point representation, each binding site is described by all pairs of features, together with the distances between them, that are present in the full binding site pharmacophore. Feature pairs are represented by three characters in the format {f1}{f2}-{d12}, where the first two characters indicate the types of the two features and the third character indicates the binned distance between them. The features are prioritized alphabetically to ensure that any feature pair has a single unique code.

The 3- and 4-point pharmacophore fingerprint representations list all the pharmacophore feature triplets or quadruplets present in the full pharmacophore. In the 3-point format, the feature opposite the shortest distance is assigned the highest priority, and if two or more of the binned distances are equal, the features are prioritized alphabetically. Each pharmacophore triplet is represented by 6 characters in the format {f1}{f2}-{f3}{d23}{d13}{d12}, where fx is the feature with priority x and dij is the binned distance between the features with priorities i and j.

In 4-point fingerprints, the features are prioritized according to the sum of the squared binned distances from the other features and the feature type. Pharmacophore quadruplets are assigned 10 characters in the format {f1}{f2}{f3}{f4}{d12}-{d13}{d14}{d23}{d24}{d34}. For the range of feature distances considered in the fingerprints, this approach was sufficient to ensure that all 4-point pharmacophores have single unique representation.

The binning of distances between features into discrete distance ranges can cause highly similar pharmacophores to be considered different if one or more of the distances fall either side of a binning threshold, and this effect has been cited as the reason for the relatively poor neighborhood behavior demonstrated by 3- and 4-point, ligand-based pharmacophore fingerprints.[35] This limitation has been addressed previously and a similar approach has been taken here to fuzzify the 2- and 3-point fingerprints: given two features (A, D) separated by a distance that falls into bin c, three bits are set in the fingerprint corresponding to ADb, ADc, and ADd.[11] Identical feature pairs in different pharmacophores will result in three matching bits in their fingerprint representations, whereas a near miss will contribute just a single match. For pharmacophore triplets, three discrete distances are assigned to each of the three feature distances in the triplet, which results in the triplet being represented by up to 27 bits. With two identical pharmacophore triplets, all the bits that represent the triplet in the fingerprint will match and the triplet will make a full contribution to the similarity calculation. Pharmacophore triplets that differ slightly will match a fraction of the bits and will therefore make a partial contribution to the similarity calculation. Fuzzified 4-point pharmacophore fingerprints resulted in unmanageably large fingerprints and were not considered for the fragment database.

Quantifying Similarity Between Fingerprints. The problem of the exponential increase in the density of the fingerprints with increasing numbers of features in the pharmacophore is a key issue that affects the choice of similarity coefficient, as coefficients for calculating similarity between binary fingerprints are known to be affected by size biases.[36] For example, the Tanimoto coefficient, which bases the similarity calculation on the number of common attributes in the two binary fingerprints, is more likely to consider densely populated fingerprints similar, whereas the Hamming distance, which judges similarity according to the number of differences between the attributes of the fingerprints, is more likely to judge sparsely populated fingerprints as similar. Fligner and Verducci proposed a modification to the Tanimoto coefficient that aims to remove the size bias inherent in the coefficient.[37] Given two fingerprints of length n with a and b bits set in each fingerprint, respectively, and c bits set in both fingerprints, selected from a data set of fingerprints with a mean bit density of ρ_0, the modified Tanimoto similarity S_{MT} is calculated as

$$S_{MT} = \left(\frac{2 - \rho_o}{3}\right) S_T + \left(\frac{1 + \rho_o}{3}\right) S_{T0}$$

where S_T is the standard Tanimoto coefficient

$$S_T = \frac{c}{a + b - c}$$

and S_{r0} is the inverted Tanimoto coefficient

$$S_{T0} = \frac{n - a - b + c}{n - c}$$

The length of the fingerprint n is calculated as the total number of feature pairs, triplets, or quadruplets, given a binning scheme. This modified version of the Tanimoto coefficient has been used as the metric for the comparison of binding site fingerprints.

Optimization of the Binding Site Fingerprints. Preliminary experiments showed that binding site fingerprints can efficiently be used to identify similar binding sites originating from proteins with an overall high similarity. However, it proved to be much more difficult to identify similar binding sites from proteins with lower sequence homologies. Therefore, a procedure was set up to optimize the parameters of the binding site fingerprints so that the fingerprints' abilities to identify cross binding site similarities is improved while the good performance within a single family is still maintained. The parameters that have been optimized include the complexity of fingerprints (2-, 3-, or 4-point pharmacophore fingerprints), the width of the initial distance bin, the bin width multiplier, whether or not the fingerprints are fuzzified, and the parameters governing the placement of each of the pharmacophore features.

Fragment Validation Set. Validation of the binding site fingerprints with examples of known bioisosteres was not viable for several reasons: first, there are not enough examples of well-known bioisosteric replacements available in the PDB to achieve statistical significance in a validation experiment, and second, accepting additional, less obvious replacements to circumvent the first problem requires specifying varying degrees of bioisosterism, which is difficult to do in an unbiased manner. Therefore, the optimization of the fingerprints has been based

on pairs of identical fragments bound to proteins in different PDB entries with the assumption that a fragment compared to itself is a perfect example of a bioisostere. A distinction has been made between pairs of fragments bound to proteins of the same family (intraprotein fragment pairs) and pairs of fragments bound to proteins of different families (interprotein fragment pairs). Protein families have been classified using version 1.7.4 of the Structural Classification of Proteins (SCOP) database, with the SCOP "family" classifications determining to which of the two sets a pair of fragments belongs. The separation of intrafamily protein pairs from the decoys indicates the degree to which the fingerprints are able to recognize specific protein structures, whereas the separation of interprotein family fragment pairs indicates the degree to which the fingerprints capture the aspects of the binding site that are responsible for fragment binding and recognition. In our view the latter of these two characteristics is of primary importance in the optimization of the fingerprints. While it is true that a specific fragment may bind to a wide range of different binding site structural motifs, on average the range of binding site structures that the fragment can bind to will have a greater degree of structural similarity than randomly selected binding site pairs, and it is these aspects of similarity that the fingerprint optimization was intended to develop.

The validation set of molecular fragments was taken from ligands extracted from the PDB. Only fragments that were represented at least 20 times in the PDB were selected for the set. In total, 398 unique fragments were used, of which 144 were R-groups (fragments with a single attachment point), 180 were linkers (two attachment points), and 74 were cores (three or more attachment points). Additionally, decoy sets with 10 000 random fragment pairs were generated for each fragment size as measured by number of atoms.

Quantifying the Performance of the Fingerprints. Up to 50 examples were randomly selected from the PDB for each of the 398 fragments and all pairwise similarities were calculated from the fingerprint representations of their local protein environments. To assess the statistical relevance of the similarities determined for validation pairs, the quantified similarity values were compared to those of a set of 10 000 decoy fragment pairs with the same number of atoms as the fragment under consideration. It was necessary to compare the similarities of true pairs to pairs with the same number of atoms to ensure that the optimization experiment was not optimizing any size biases inherent in the fingerprints. The decoy fragment pairs were taken from preselected sets of 200 diverse PDB ligand fragments for each fragment size. These diverse fragment sets were generated from the fragment database by first grouping the database fragments according to number of atoms and then applying Pipeline Pilot's *Select Diverse Set* component to each of the fragment sets.[38]

The separation of the true fragment pairs from the decoy pairs was measured by both the area under the receiver operating characteristic curves (AUC) and the area under logarithm biased receiver operating characteristic curves proposed by Clark and Webster-Clark (LAUC).[39] The LAUC places a greater emphasis on the highest ranked pairs and was therefore considered to be a more appropriate measure of performance; given that the database contains around 300 000 fragments, only the first few hits obtained from a search will be considered by the users. The expected values for random and perfect separations for the LAUC metric differ from the AUC: the expected AUC value for a random distribution is 0.5, and the equivalent LAUC value is 0.434; the AUC value for a perfect separation corresponds to 1.0, whereas the LAUC has no upper bound and the maximum LAUC value depends on the numbers of negative and positive instances in the data set.

Superposition of Local Binding Sites. A KRIPO fingerprint similarity search returns only a list of binding sites that are similar to the query; however, these results can be visualized in 3D by superimposition of the query and hit fragments. The superimposition process involves first identifying the maximum common subpharmacophore between the query and hit pharmacophores with clique detection. For the clique detection, we used the program Cliquer with a distance tolerance of 1.2 Å[40] The protein and ligand structures from the hit complex can then be transferred onto the query structures by applying the transformation matrix obtained from the superposition of the common subpharmacophores.

Similar Binding Site Pairs Extracted from the PDB. The effectiveness of the KRIPO fingerprints in identifying similar binding sites was assessed by comparing the fingerprint similarities of a set of highly similar binding site pairs extracted from the PDB to a set of dissimilar decoy binding site pairs. Pairs of PDB entries that contain similar protein sequences were identified with an all-against-all BLAST search on the PDB.[41] Any pairs of PDB entries containing sequences with homologies of greater than 0.9 were recorded. For each recorded pair, the query and the reference (hit) structures were read into YASARA and the *Clean* protocol was applied (see Preparation of Protein Structures for full details).[32] Any solvent molecules, metal ions, and other unwanted molecules were removed (including lipids, sugars, and small molecules from the crystallization buffer). Alternate atom locations were removed, and only the first structures were kept from NMR bundles. Finally any hetgroup structures with sizes of greater than 800 or less than 50 Da or hetgroups that were not in contact with the protein (5 Å) were removed. The reference structure was then aligned to the query structure with the YASARA MOTIF protocol, which produces a 3D alignment of the two protein sequences using substructures of high sequence homology.[42] Each of the remaining eligible hetgroups (ligands) in the reference structure was then taken in turn. The reference binding site was defined as all reference amino acids with atoms within 5 Å of the reference ligand currently under consideration, and any nonbinding site amino acids in the reference structure were deleted. Following this, all query amino acids that did not have an atom within 5 Å of a query hetgroup atom were removed from the query structure, and the alignment of the query structure amino acids with the reference binding site was refined by applying the MustangPP protocol.[43] In this alignment, only the residues that can be considered to be structurally aligned are realigned. This process should result in a clear alignment between the reference binding site and the equivalent query binding site. All other query amino acids and hetgroups were removed, leaving a pair of aligned binding sites and an overlaid pair of ligand structures. The binding sites were considered to be similar, and the ligands considered bioisosteric, if the following criteria were met: identity of the amino acid pairs in the two binding sites was greater than 0.9, with amino acids pairs considered to match if the BLOSUM score for the residue pair was 0 or greater; the binding sites had at least 3 amino acids in common; the α-carbon rmsd of the amino acids used for the MustangPP alignment was less than 0.8; none of the pairs of matching amino acids have an all-atom rmsd of greater than 2.0.[44]

dx.doi.org/10.1021/ci3000776 | J. Chem. Inf. Model. 2012, 52, 2031−2043

Bioisosteric substructure pairs were identified from the overlaid ligands by application of the method outlined by Kennewell et al.[12] This method involves breaking the query ligand into overlapping fragments and reference ligand into nonoverlapping fragments. Ligands were fragmented at the bonds between ring atoms and nonring atoms using the same process described in the section Generation of Ligand Fragments and Local Binding Sites above. Each query fragment is taken in turn, and any reference fragments that overlap with the query fragment to a sufficient degree are recorded. The reference fragments are combined where possible, and the largest combined reference substructure is identified as a bioisostere to the query fragment. The degree to which pairs of fragments overlap is determined with a scoring function based on the pairwise atomic distances of the two fragments. The same overlap score cutoff values were used as described in the paper of Kennewell, which are 0.5 for the uncombined reference fragments with the query fragment and 0.7 for the final combined reference fragment with the query fragment.

In total, 5262 similar fragment−protein complex pairs were obtained from a total of 1992 unique PDB entries with 910 unique ligands. An equal number of dissimilar binding site pairs were generated by randomly selecting complexes from the extracted set. KRIPO fingerprints were generated from each of the complexes, and the pairs were used to assess the ability of the binding site fingerprints to separate similar binding site pairs from dissimilar binding site pairs.

RESULTS AND DISCUSSION

The optimization of KRIPO fingerprint progressed in three stages. In the first stage, the three levels of fingerprint complexity (2-, 3-, or 4-point) were compared. Following this, fuzzy and nonfuzzy versions of the fingerprints were assessed, and finally, the placement of features in the binding site pharmacophores was optimized. The fingerprint and feature placement parameters could have been further optimized with a second round of calculations; however, these calculations were computationally intensive and we did not expect further optimizations to result in a significant improvement to the performance of the fingerprints. The fingerprints were therefore optimized with a single, linear pass through the set of parameters.

Effect of Pharmacophore Fingerprint Complexity on Bioisostere Retrieval. In the first stage, combinations of the initial bin width and the bin width multiplier were tested for each of the nonfuzzy fingerprints. Initial bins widths of {0.25, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0, 5.0} and multipliers of {1.0, 1.1, ..., 1.8} were tested for the 2- and 3-point fingerprints, and initial bin widths of {1.5, 2.0, 2.5, 3.0, 4.0, 5.0} and multipliers of {1.0, 1.1, ..., 1.5} were tested for the 4-point fingerprints.

Figures 3a and 3b show the mean LAUC and AUC results respectively for all the validation fragments using each of the fingerprint parameter combinations. Looking first at the intraprotein family data, the 3-point fingerprints tend to have a higher true-pair recall rate than the 2-point and 4-point fingerprints, with the highest mean AUC determined to be 0.89 for the 3-point fingerprints compared to 0.88 for the 2- and 4-point fingerprints, and the highest mean LAUC determined to be 1.37 for the 3-point fingerprints compared to 1.36 and 1.32 for the 2- and 4-point fingerprints. The best 3- and 4-point fingerprint encryptions are comparable, as measured by the interprotein family fragment pairs, with mean AUCs of around 0.65 and mean LAUCs of around 0.66.

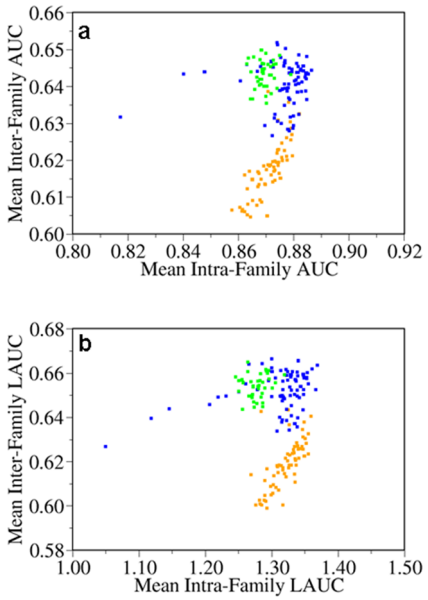


Figure 3. Comparison of fingerprint implementations. Comparison of 2- (orange), 3- (blue), and 4-point (green) fingerprints with the (a) AUC and (b) LAUC.

4-Point pharmacophore fingerprints are recognized to have some advantages over the lower-order representations: they provide implicit information on the volume of the binding sites, and although our implementation was achiral, they represent the minimum level of complexity required to describe chirality.[45,46] However, the number of pharmacophore bitcodes represented in the 4-point fingerprints increases exponentially with the number of features in the binding site pharmacophore and at a much greater rate than the 2- and 3-point fingerprints. The 4-point fingerprints therefore require considerably more disk storage space and result in longer preparation and search times. With a database of 300 000 fragments, this represented a significant problem for our implementation. As they did not result in significantly improved separation of true binding site pairs from the decoy pairs, chiral versions were not investigated, and the 4-point fingerprints were rejected from further consideration. The 2-point fingerprints were less effective at identifying interprotein family pairs and were also rejected from further consideration. 3-Point pharmacophore fingerprints seemed to represent the right balance between computational requirements and separation of true pairs from decoys and were selected to represent local binding sites in the fragment database. The optimal values for the initial bin width and the multiplier for the 3-point fingerprints were 3.0 and 1.2, respectively. For all three of the fingerprint complexity levels, the intraprotein family fragment pairs could be separated better from decoy pairs than interprotein family pairs. This result is unsurprising and can be explained by the fact that the binding site structures in interprotein family fragment pairs will not always exhibit a high degree of structural similarity.

Fuzzification of the Fingerprints. The 3-point pharmacophore fingerprints were tested in both the fuzzified and nonfuzzified forms with all combinations of the initial bin widths {0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0, 5.0} and bin width multipliers {1.0, 1.1, ..., 1.8}. Figures 4a and 4b show the mean LAUCs and mean AUCs of the various parameter combinations. The results of this test show that fuzzification of the 3-point fingerprints tends to improve the separation of true

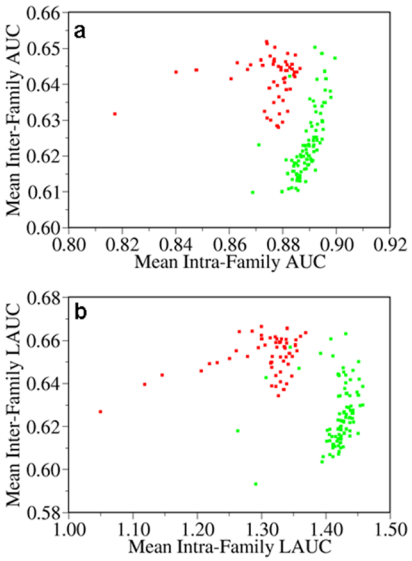


Figure 4. Comparison of fuzzy (green) and nonfuzzy (red) 3-point fingerprints with the (a) AUC and (b) LAUC.

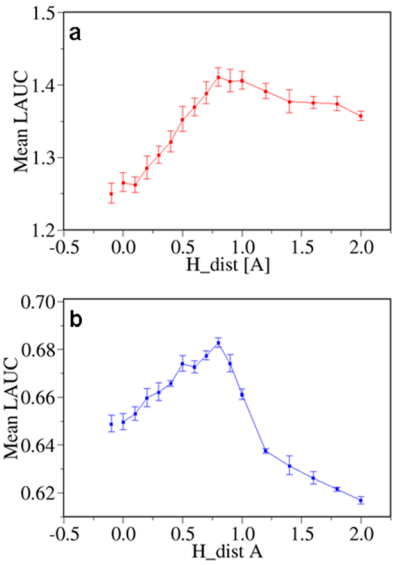


Figure 5. Optimization curves for the placement of hydrophobic features (a) intraprotein family pairs and (b) interprotein family pairs.

intraprotein family fragment pairs from the decoys but results in comparable separations of the interprotein family fragment pairs. We therefore selected the fuzzified 3-point fingerprints with an initial bin width of 0.8 and a bin width multiplier of 1.0 to represent the local binding sites in the fragment database.

Optimization of the Pharmacophore Representations of the Binding Sites. The pharmacophore representations of the binding sites were optimized in the following order: *H_dist*, *D_dist*, *A_dist*, *Rp_dist*, *Rt_dist*, *P_dist*, *P_width*, *N_dist*, *N_width*. Some preliminary experiments had already been run that had provided a rough guide to the feature placement. The initial values for the parameters *H_Dist*, *A_dist*, *P_dist*, *P_width*, *N_Dist*, and *N_width* were set to 1.0; the initial values for *D_dist* and *Rt_dist* were set to 0.0; and the initial value for *Rt_dist* was set to 3.0. Each parameter was optimized in turn by assessing the intra- and interprotein family pair separation over a range of values. During parameter optimization, all other parameters were set to either their initial or optimized values, depending on whether or not they had already been optimized. The optimization curves are provided in the Supporting Information and show how the separation of the fragment pairs from the decoys, as measured by the LAUC, changes over the range of possible values for the different parameters. Curves were computed for both the intra- and interprotein family pairs, and the value −0.1 was used to represent the removal of the feature from the pharmacophore. Because each experiment involved the random selection of 50 examples of the fragment from the PDB, the resulting LAUC values were subject to some variance, and each calculation was repeated five times to establish confidence intervals and to smooth out the optimization curves. The two curves for the *H_dist* parameter are shown in Figure 5 as an example and reveal a clear peak in the separation at 0.8 Å. The optimal values for all other feature placement parameters are shown in Table 1.

The reasons for the optimal parameter values are not always clear. For hydrophobic feature placement, if the value of the parameter *H_dist* is too low, amino acids that provide a hydrophobic character on the binding site will not be represented in the binding site pharmacophores. High values

Table 1. Parameters for Pharmacophore Feature Placement

feature placement parameter	pharmacophore feature type	optimized value
H_dist	hydrophobic	0.8
A_dist	acceptor	0.8
D_dist	donor	0.2
Rp_dist	aromatic/π	4.0
Rt_dist	aromatic/t	excluded
P_dist	positive charge	0.3
P_width	positive charge	0.0
N_dist	negative charge	1.4
N_width	negative charge	1.0

of *H_dist* would often result in hydrophobic features distributed throughout the entire binding sites, with no clear distinction between the hydrophobic and nonhydrophobic regions. A value of *H_dist* of 0.8 appears to represent the optimal balance between representation, and over-representation of hydrophobic functional groups in the binding sites.

Values of 0.8 and 0.2 for the acceptor and donor parameters, *A_dist* and *D_dist*, maximized the separation of true interprotein family pairs from the decoy pairs. Higher values for both these parameters did result in a slight but significant increase in the separation of intraprotein family pairs; however, as the higher values also caused the interprotein fragment pair separation to decrease, this gain appears to be because of the inclusion of structural characteristics of the binding site that help to identify the protein family but that are not directly responsible for the fragment binding and recognition.

The aromatic parameter *Rp_dist*—which governs the placement of pharmacophore features that represent π-stacking—had an optimal value in the range 3.2−4.0 Å for the intraprotein family pairs, but adjusting the parameter had virtually no effect on the separation of interprotein family pairs. Aromatic rings in a π-stacking arrangement lie in parallel planes approximately 3.6 Å apart. As the rings are slightly offset, feature placement in the range 3.2−4.0 Å provides the greatest chance that a genuine π-stacking interaction will be incorporated into the binding site pharmacophore. Inclusion of the T-stacking features, governed by the parameter *Rt_dist*,

dx.doi.org/10.1021/ci3000776 | J. Chem. Inf. Model. 2012, 52, 2031−2043

improved the recognition of intraprotein family fragment pairs but reduced separation of interprotein fragment pairs, and the features representing these interactions were therefore not included in the KRIPO binding site pharmacophores.

The positively charged electrostatic features added very little to the discriminating ability of the binding site pharmacophores, which may be because the positively charge amino acids side chains were already represented by distinct formations of donor and aromatic pharmacophore features. The inclusion of negatively charged pharmacophore features significantly improved the separation of both intra- and interprotein family pairs, and the optimal parameter values of *P_dist* = 1.4 and *P_width* = 1.0 suggest broad and delocalized interactions.

KRIPO Fingerprints Assessed with Similar Binding Site Pairs. KRIPO similarities were calculated for the similar binding site pairs and the dissimilar, decoy binding site pairs generated from the PDB (PDB June 2009; for details see Materials and Methods). The separation between the similar pairs and the decoy pairs was assessed by calculation of the receiver operating characteristic curve. The area under the curve (AUC) was calculated to be 0.83 which indicates that the method successfully separated similar from nonsimilar binding sites. A clear separation between the two sets of binding site pairs can be seen in the distributions shown in Figure 6. 82% of

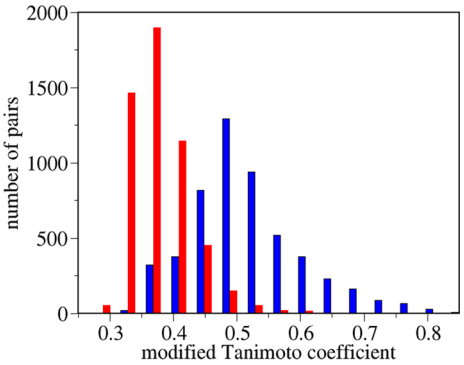


Figure 6. Distribution of the modified Tanimoto similarities for the true binding site pairs (blue) and the randomly generated decoy pairs (red).

the true binding site pairs have a modified Tanimoto coefficient of greater than 0.45, whereas only 27% of the decoy pairs fall into that range. A visual inspection of the true pairs with low binding site similarity showed that the low scores are caused by mutations in the binding site, ligand induced loop dis-/relocations, side chain rotations, and high solvent accessibility.

CASE STUDIES

We used the KRIPO system to search for bioisosteric replacements for an inhibitor of MAP kinase (1A9U) and an inhibitor of thrombin (1K22). Three fragment queries were identified from each inhibitor and applied as search queries to the KRIPO fragment database. We used a minimum pharmacophore clique of 5 features to filter out false positive hits from the hit list, and the analysis in the previous section indicated that a similarity value of 0.45 (using the modified Tanimoto coefficient) provides a reasonable discrimination between similar and dissimilar binding sites. Using these thresholds, hit lists of up to 100 fragments bound to similar binding sites were obtained with each of the queries. In these case studies, we have focused on the first intrafamily hit and the top three interfamily hits from different protein families. We consider bioisosteric replacements obtained from interprotein family hits to be of particular interest because they are more likely to be previously unknown binders for the subpocket. For each of the searches, a complete hit list of the search results can be found in the Supporting Information.

1A9U: A Pyridinylimidazole Inhibitor Bound to p38. The pyridinylimidazole inhibitor bound to MAP kinase p38 in PDB entry 1A9U was divided into three fragments as shown in Figure 7. The hinge fragment forms a key hydrogen bond with

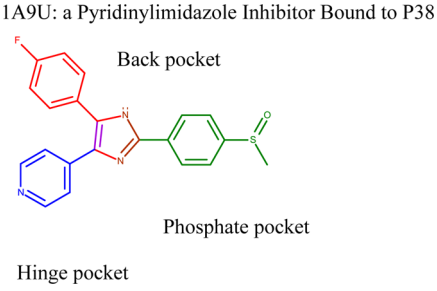


Figure 7. Structure of MAP kinase inhibitor SB6. The ligand is divided into three fragments that define different query binding sites. Hinge pocket (blue, purple), back pocket (red, purple, brown), and phosphate pocket (green, brown).

the backbone amine of the central hinge residue Met109 (Figure 2a). The hinge pocket query pharmacophore also included a hydrogen bond acceptor feature from His107, a positive charge and hydrogen bond donor feature from Lys53, and several hydrophobic interaction features. The highest ranked hit from this search query was a pyrimidine group that can form the same interaction with the hinge backbone as seen in the query (Figure 8a). The first interfamily hit (Figure 8b, rank 5) is an adenine nucleoside fragment in complex with the cofactor binding site of acetyl-CoA carboxylase that can also form the key hydrogen bond interactions with the hinge backbone (His107 and Met109). This fragment is the hinge-binding fragment of adenine triphosphate (ATP), the natural substrate of protein kinases, and is therefore a clear example of a bioisostere to the query fragment. The second interfamily hit (Figure 8c, rank 12) is from the binding site of camphor 5-monooxygenase and, although the binding site pharmacophore has comparable hydrophobic and hydrogen bond donor features to the query, the fragment cannot form the key interaction with the hinge backbone and we therefore do not consider it to be a viable replacement. The third interfamily hit (Figure 8d, rank 21) is a fragment bound to the subpocket of alcohol dehydrogenase. Again, this fragment does not form the required interactions with the hinge backbone, and we do not consider it to be a viable replacement, although the two hydroxyl groups can form interactions with the carbonyl group of the hinge (His107) and the side chain of Lys53.

The back pocket fragment consists of an imidazole core and a fluorophenyl ring that occupies the hydrophobic cleft between gate keeper residue Thr106 and cationic residue Lys53. The corresponding binding site pharmacophore includes features that represent the positive charge and hydrogen bond donors of Lys53, an aromatic feature from Tyr35, and multiple hydrophobic features surrounding the hydrophobic cleft. The top hits are superposed onto the query in Figures 8e–h. The highest ranked hit is from the same kinase family as the query and the fragment binding is analogous (Figure 8e). The first

dx.doi.org/10.1021/ci3000776 | J. Chem. Inf. Model. 2012, 52, 2031–2043

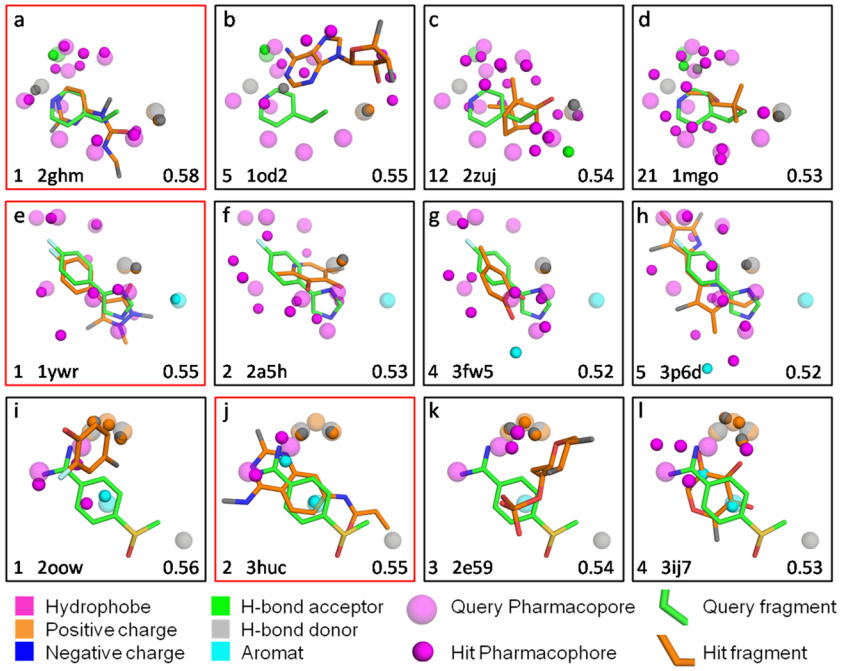


Figure 8. Superposition of the queries (C green; N blue; F light blue; O red; S yellow; R gray) and the corresponding top hits [(bottom left) rank; (bottom right) modified Tanimoto coefficient; C orange; N blue; F light blue; O red; S yellow; R gray] based on the local binding site pharmacophores. The pharmacophores are shown as spheres; the pharmacophore features of the query are transparent, and the pharmacophore features of the hits are solid. Hits with a red outline are from kinase structures.

two interfamily hits are from lysine-2,3-aminomutase (Figure 8f, rank 2) and neutrophil gelatinase-associated lipocalin (Figure 8g, rank 4). Both of the binding site pharmacophores from these hits include one positive charge feature, one hydrogen-bond donor feature, and several hydrophobic features. The fragments in these binding sites are methylated and hydroxylated aromats and are therefore potentially viable replacements for the back pocket fragment. The third interfamily hit is fragment bound to phycoerythrobilin synthase (Figure 8h, rank 5) and may offer an alternative scaffold that binds into the backpocket and links up to the hinge and phosphate pocket fragments.

The search with the phosphate pocket query was less successful than the other searches for this inhibitor. This cavity is highly solvent exposed and the query pharmacophore represents only three highly flexible amino acids: Arg137, Tyr35, and Lys53. The side chain of Arg137 is often dislocated, and the side chain of Tyr35 is frequently orientated away from the binding site. The Lys53 forms an ionic interaction to an aspartate side chain of Cα-helix that fixes the conformation of the side chain. In some kinase structures, this interaction is not formed and the aspartate adopts multiple conformations or is dislocated. The flexibility of the phosphate subpocket leads to different pharmacophore representations of this binding site among kinase structures, and it is therefore unsurprising that the first intrafamily hit was ranked second (Figure 8j). The highest ranked hit is a fragment bound to macrophage migration inhibitory factor (Figure 8i); the third highest hit is from Lymphocyte antigen 96 (Figure 8k, rank 3); and the fourth is from pancreatic alpha-amylase (Figure 8l, rank 4). In all three of the interfamily hits, the fragments are hydroxylated ring systems that could interact with Lys53 in a similar way to the natural substrate, ATP.

1K22: Melagatran, an Approved Drug for Thrombin Inhibition. In the second case study, we applied the KRIPO system to the inhibitor melagatran in complex with a thrombin structure (PDB: 1K22). The example was taken from a study by Gerlach et al., in which the Cavbase approach was used to assemble a focused combinatorial library for thrombin.[47,14] Thrombin belongs to the serine protease family, which has peptides as natural substrates. It has three well-defined subpockets that allow a selective cleavage of the substrate. Three fragments of melagatran defined the queries for the bioisostere search, as shown in Figure 9.

1K22: Melagatran, an approved drug for thrombin inhibition

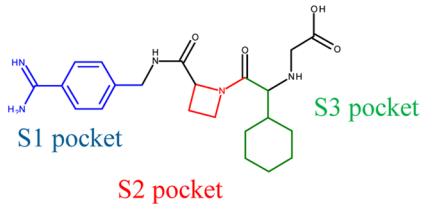


Figure 9. Structure of thrombin inhibitor melagatran. The ligand is divided into fragments, which define the query binding sites. S1 pocket (blue), S2 pocket (red), and S3 pocket (green).

The benzamidine fragment of melagatran in the S1 subpocket forms two hydrogen bonds with Asp189 at the bottom of the subpocket. It is well-known that the interaction of this substructure to Asp189 is responsible for high affinity binding of the inhibitors. The carboxyl group of Asp189 is represented by negative charged features and a hydrogen bond acceptor feature. In addition, hydrophobic features are placed near the aromatic ring of melagatran, and negative charge and hydrogen bond acceptor features are placed at the entry of the S1 pocket. The most similar intrafamily binding site is shown superposed onto the query in Figure 10a and is the same

dx.doi.org/10.1021/ci3000776 | J. Chem. Inf. Model. 2012, 52, 2031−2043

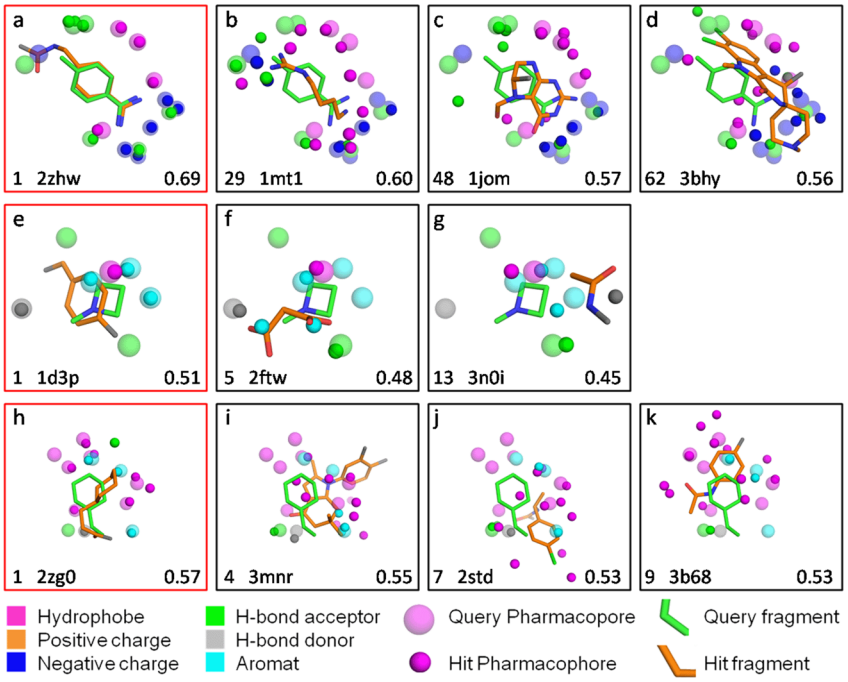


Figure 10. Superposition of the queries (C green; N blue; F light blue; O red; S yellow; R gray) and the corresponding top hits [(bottom left) rank; (bottom right) modified Tanimoto coefficient; C orange; N blue; F light blue; O red; S yellow; R gray] based on the local binding site pharmacophores. The pharmacophores are shown as spheres, whereas the pharmacophor features of the query are transparent and the pharmacophore features of the hits are solid. Hits with a red outline are from thrombin structures.

benzamidine fragment bound to the S1 pocket of thrombin. The first interprotein family hit (Figure 10b, rank 29) is a fragment with an aliphatic primary amine bound to pyruvoyl-dependent arginine decarboxylase. There are several examples of thrombin inhibitors with aliphatic primary or secondary amine groups that bind to the S1 pocket in the literature, and this provides strong support to the viability of this replacement.[48−51] The second highest ranked interfamily fragment (Figure 10c, rank 48) is a 2-amino-tetrahydro-pteridinone fragment bound to a subpocket of dihydrofolate reductase. Support for this replacement is provided by Lam et al., who identified several guanidine/benzamidine mimics exhibiting thrombin inhibition with functional groups binding to the S1 pocket that are very similar to the dihydrofolate reductase hit fragment.[50] The third interfamily hit is a carboline fragment from DAP kinase 3 (Figure 10d, rank 62). An ionic interaction to Asp186 is possible with the fragments from all three interfamily hits, which is important for high affinity binding to the S1 pocket.

The S2 pocket of thrombin is covered by the S60 loop, which creates a smaller S2 pocket compared to other serine proteases.[47] When thrombin is in complex with its natural substrate, the narrow cleft of the subpocket is occupied by a valine side chain, whereas when thrombin is in complex with the inhibitor melagatran, this cleft is occupied by the azetine moiety. The pharmacophoric representation of the subpocket consists of aromatic features from Tyr60A, Trp60D, and His57, hydrophobic features from Leu99 and Trp60D, hydrogen bond acceptor features from Tyr60A and Ser214, and a hydrogen bond donor feature of Gly 216. These searches typically yield more than 100 hits, but with the S2 query, only 14 hits were retrieved of which 12 were from thrombin structures. The most similar intrafamily hit was the benzyl ring fragment (Figure 10e). Only two interfamily hits were found: a malonic acid fragment bound to dihydropyrimidinase (Figure 10f, rank 5) and a carboxamide fragment bound to Ad37 fiber knob (Figure 10g, rank 13).

The final search query for melagatran was the ethyl-cyclohexyl ring system bound in the hydrophobic S3 pocket. The query pharmacophore for this pocket consisted of hydrophobic features representing the aromatic ring system of Trp60D and the hydrophobic residues Ile174, Leu99, and Try60D, and a single hydrogen bond acceptor from the side chain of Ser97. The highest ranked hit yielded by the search was a cyclohexyl ring bound to a another thrombin structure (Figure 10h); however, the linker to the core of the inhibitor is longer than the equivalent linker in melagatran, which will lead to a different binding mode for the aliphatic ring system in the S3 pocket. The first interfamily hit was ranked fourth and is an indole fragment bound to HSP 90-alpha (Figure 10i, rank 4). The bicyclic ring system of this fragment is a putative bioisostere. The second and third interfamily hits were ranked seventh and ninth on the hit list and are a chlorobenzene fragment bound to scytalone dehydratase (Figure 10j, rank 7) and a phenylamide fragment bound to the dihydrotestosterone receptor (Figure 10k, rank 9). The binding mode of the phenylamide fragment is similar to the binding mode of the melagatran query fragment.

SUMMARY

These case studies each provided lists of up to 100 fragments for each query. Although the lists included some fragments from the same protein family that were similar, if not identical to, the query fragments, in general, the fragments that were returned were quite distinct from fragments retrieved with simple 2D similarity searches. For comparison, we searched the fragment database with 2D IBIS fingerprints and compared the top 1000 similar fragments obtained with the two approaches

for each query (see the Supporting Information section).[11] Only a few hits are common to both methods, which emphasizes that the KRIPO system represents an alternative and complementary approach to the identification of bioisosteric replacements.

The ability of the fingerprints to identify replacements from different protein families was highly dependent on the query. Query fragments from highly conserved binding sites of proteins families that are abundant in the PDB, e.g. the hinge region of kinases and the S1 pocket of thrombin, are likely to result in many top scoring hits found in the equivalent region of proteins belonging to the same family. In such cases, interfamily hits are ranked later in the hit list. The search results are also dependent on the query fragment used, and we have found that it is sometimes necessary to define more than one query to retrieve all valuable hits. This kind of on-the-fly searching of the database is possible because searches typically take no longer than five minutes.

We made a decision to focus on the first three interfamily hits to ensure that the results were presented in a consistent way and to avoid introducing bias by cherry picking the best examples. For some searches, viable replacements are not identified within the top three interfamily hits; for example, the first interfamily hit for the kinase hinge-binding fragment is from ATP, the natural substrate of protein kinases, and the second and third interfamily hits are not viable replacements. Digging further down into the hit list can sometimes yield better replacements (the smiles code for all top 100 hits are provided in the Supporting Information section), and so users of KRIPO should be prepared to manually screen the top ranked hits to find ideas for replacements. In some cases, it is not possible to link the hit fragments to the core molecule. We decided to present all hits to the users regardless of whether they represent viable replacements as incompatible hit fragments may provide inspiration for alternative substructural replacements to the query molecule.

The case studies demonstrated that the similarity between pockets in the PDB is not equally distributed. For most of the queries, we find both intra- and interfamily hits, which indicates that the different protein pockets have a significant similarity to subpockets of other proteins. Our query for the thrombin S2 pocket yielded only 2 interfamily hits with a very low similarity, and all the intrafamily hits were from other thrombin structures, which is in agreement with the role of the S2 pocket as the substrate selectivity determining pocket. The small number of interfamily hits found by our method suggests that the S2 pocket, as it is defined in this paper, is quite unique. Gerlach et al. came to the same conclusion following their analysis of thrombin.[47] This finding suggests an alternative application of KRIPO, which is to identify possible off target activities and to identify the structural features of the target protein that should be targeted to reduce unwanted side effects of the drug candidates.

CONCLUSIONS

We have presented a searchable fragment database derived from the PDB that can be used to identify possible bioisosteric replacements for ligand substructures. Central to the database is KRIPO: a novel method for calculating localized binding site similarities that is based upon pharmacophore fingerprints of the local binding sites. A series of optimization steps were performed to tune the various attributes of the fingerprints and the binding site pharmacophores, including the positioning of the pharmacophore features relative to the amino acids of the binding sites, and the method by which the pharmacophores are converted into a fingerprint representation. The best results could be achieved with a fuzzified 3-point pharmacophore fingerprint representation.

The ability of the binding site pharmacophore fingerprints to distinguish similar protein binding sites from unrelated protein binding sites was demonstrated by calculating the similarities for a set of pairs of known similar binding sites and a set of randomly generated decoy pairs. The area under the receiver operating curve for the distributions of similarity values of these two sets was calculated to be 0.83, which indicates that fingerprints can effectively identify similar binding sites.

Finally, the value of KRIPO for the identification of bioisosteric replacements for ligand substructures from the PDB was demonstrated by its application to an inhibitor bound to a MAP kinase (1A9U) and the inhibitor melagatran bound to thrombin (1K22). Potential replacements were identified for all the substructures of the inhibitors, including some replacements from protein families that were unrelated to the query structures. The hits from interfamily proteins increase the diversity of bioisosteres and facilitate the replacement with novel fragments.

The S2 pocket of thrombin as defined in this paper appears to be rather unique in the PDB. An analysis of all subpockets stored in the KRIPO database against each other could help to discover more unique subpockets in the future. In addition, the general architecture of protein binding sites can be studied independently from protein sequence or protein folding using the KRIPO structure-based pharmacophores.

ASSOCIATED CONTENT

Supporting Information

Table of pharmacophore feature assignment to each amino acid; figure showing the optimization curves for the placement of the different features; tables with case study hits. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: t.ritschel@cmbi.ru.nl. Phone: +31 24 36 19674. Fax: +31 24 36 19395.

Present Addresses

[§]AstraZeneca, Mereside, Alderley Park, Macclesfield SK10 4TG, UK.

[⊥]The Netherlands eScience Center, Science Park 140, 1098 XG Amsterdam, The Netherlands.

[∥]Grünenthal GmbH, Global Drug Discovery, Discovery Informatics, 52099 Aachen, Germany.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to thank NBIC and DFG for funding. T.R. is recipient of a personal grant Ri 2087/1-1 from the DFG.

REFERENCES

(1) Lima, L. M.; Barreiro, E. J. Bioisosterism: a useful strategy for molecular modification and drug design. *Curr. Med. Chem.* **2005**, *12*, 23−49.

(2) Olesen, P. H. The use of bioisosteric groups in lead optimization. *Curr. Opin. Drug Discovery Devel.* **2001**, *4*, 471−478.
(3) Martin, Y. C. A practitioner's perspective of the role of quantitative structure-activity analysis in medicinal chemistry. *J. Med. Chem.* **1981**, *24*, 229−237.
(4) Friedman, H. L. *Influence of Isosteric Replacements upon Biological Activity*; NAS-NRS Publication No. 206; NAS-NRS: Washington, DC, 1951; Vol. *206*, pp 295−358.
(5) Thornber, C. W. Isosterism and molecular modification in drug design. *Chem. Soc. Rev.* **1979**, *8*, 563−580.
(6) Lipinski, C. A. Bioisosterism in Drug Design. *Annu. Rep. Med. Chem.* **1986**, *21*, 283−291.
(7) Burger, A. Isosterism and bioisosterism in drug design. *Prog. Drug Res.* **1991**, *37*, 287−371.
(8) Patani, G. A.; LaVoie, E. J. Bioisosterism: A Rational Approach in Drug Design. *Chem. Rev.* **1996**, *96*, 3147−3176.
(9) Ujváry, I. BIOSTER—a database of structurally analogous compounds. *Pestic. Sci.* **1997**, *51*, 92−95.
(10) *Bioster*, version 12.1, Digitalchemistry: Sheffield, United Kingdom, 2011.
(11) Wagener, M.; Lommerse, J. P. The quest for bioisosteric replacements. *J. Chem. Inf. Model.* **2006**, *46*, 677−685.
(12) 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.
(13) Moriaud, F.; Adcock, S. A.; Vorotyntsev, A.; Doppelt-Azeroual, O.; Richard, S. B.; Delfaud, F. A Computational Fragment Approach by Mining the Protein Data Bank: Library Design and Bioisosterism. In *Library Design, Search Methods, and Applications of Fragment-Based Drug Design*; Bienstock, R. J., Ed.; ACS Symposium Series; American Chemical Society: Washington, DC, 2011; Vol. *1076*, pp 71−88.
(14) Schmitt, S.; Hendlich, M.; Klebe, G. From Structure to Function: A New Approach to Detect Functional Similarity among Proteins Independent from Sequence and Fold Homology. *Angew. Chem.* **2001**, *40*, 3141−3144.
(15) Schmitt, S.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **2002**, *323*, 387−406.
(16) Kuhn, D.; Weskamp, N.; Schmitt, S.; Hullermeier, E.; Klebe, G. From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. *J. Mol. Biol.* **2006**, *359*, 1023−1044.
(17) Kuhn, D.; Weskamp, N.; Hullermeier, E.; Klebe, G. Functional classification of protein kinase binding sites using Cavbase. *ChemMedChem* **2007**, *2*, 1432−1447.
(18) Gold, N. D.; Jackson, R. M. SitesBase: a database for structure-based protein−ligand binding site comparisons. *Nucleic Acids Res.* **2010**, *34*, 231−234.
(19) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application. *J. Chem. Inf. Model.* **2007**, *47*, 279−294.
(20) Sciabola, S.; Stanton, R. V.; Mills, J. E.; Flocco, M. M.; Baroni, M.; Cruciani, G.; Perruccio, F.; Mason, J. S. High-throughput virtual screening of proteins using GRID molecular interaction fields. *J. Chem. Inf. Model.* **2010**, *50*, 155−169.
(21) Yeturu, K.; Chandra, N. PocketMatch: a new algorithm to compare binding sites in protein structures. *BMC Bioinf.* **2008**, *9*, 543.
(22) Hoffmann, B.; Zaslavskiy, M.; Vert, J. P.; Stoven, V. A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. *BMC Bioinf.* **2010**, *11*, 99.
(23) Weill, N.; Rognan, D. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *J. Chem. Inf. Model.* **2010**, *50*, 123−135.
(24) Ramensky, V.; Sobol, A.; Zaitseva, N.; Rubinov, A.; Zosimov, V. A novel approach to local similarity of protein binding sites substantially improves computational drug design results. *Proteins* **2007**, *69*, 349−357.
(25) Jambon, M.; Andrieu, O.; Combet, C.; Deleage, G.; Delfaud, F.; Geourjon, C. The SuMo server: 3D search for protein functional sites. *Bioinformatics* **2005**, *21*, 3929−3930.
(26) Jambon, M.; Imberty, A.; Deleage, G.; Geourjon, C. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins* **2003**, *52*, 137−145.
(27) 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.
(28) Wallach, I.; Lilien, R. H. Prediction of sub-cavity binding preferences using an adaptive physicochemical structure representation. *Bioinformatics* **2009**, *25*, 296−304.
(29) Durrant, J. D.; Friedman, A. J.; McCammon, J. A. CrystalDock: a novel approach to fragment-based drug design. *J. Chem. Inf. Model.* **2011**, *51*, 2573−2580.
(30) Wang, Z.; Canagarajah, B. J.; Boehm, J. C.; Kassisa, S.; Cobb, M. H.; Young, P. R.; Abdel-Meguid, S.; Adams, J. L.; Goldsmith, E. J. Structural basis of inhibitor selectivity in MAP kinases. *Structure* **1998**, *6*, 1117−1128.
(31) Dullweber, F.; Stubbs, M. T.; Musil, D.; Sturzebecher, J.; Klebe, G. Factorising ligand affinity: a combined thermodynamic and crystallographic study of trypsin and thrombin inhibition. *J. Mol. Biol.* **2001**, *313*, 593−614.
(32) Krieger, E.; Koraimann, G.; Vriend, G. Increasing the precision of comparative models with YASARA NOVA–a self-parameterizing force field. *Proteins* **2002**, *47*, 393−402.
(33) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(34) Aronov, A. M.; Murcko, M. A. Toward a pharmacophore for kinase frequent hitters. *J. Med. Chem.* **2004**, *47*, 5616−5619.
(35) Horvath, D.; Jeandenans, C. Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces–a benchmark for neighborhood behavior assessment of different in silico similarity metrics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 691−698.
(36) Holliday, J. D.; Salim, N.; Whittle, M.; Willett, P. Analysis and display of the size dependence of chemical similarity coefficients. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 819−828.
(37) Fligner, A. M.; Verducci, J. S.; Blower, P. E. A Modification of the Jaccard−Tanimoto Similarity Index for Diverse Selection of Chemical Compounds Using Binary Strings. *Technometrics* **2002**, *44*, 110−119.
(38) *Pipeline Pilot*, version 7.7; Accelrys: San Diego, CA, 2011.
(39) Clark, R. D.; Webster-Clark, D. J. Managing bias in ROC curves. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 141−146.
(40) Niskanen, S.; Östergård, P. R. J. *Cliquer*, version 1.0; available at http://users.tkk.fi/~pat/cliquer.html.
(41) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403−410.
(42) Vriend, G.; Sander, C. Detection of common three-dimensional substructures in proteins. *Proteins* **1991**, *11*, 52−58.
(43) Konagurthu, A. S.; Whisstock, J. C.; Stuckey, P. J.; Lesk, A. M. MUSTANG: A multiple structural alignment algorithm. *Proteins* **2006**, *64*, 559−574.
(44) Henikoff, S.; Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **1992**, *89*, 10915−10919.
(45) 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.
(46) Mason, J. S.; Cheney, D. L. Ligand-receptor 3-D similarity studies using multiple 4-point pharmacophores. In *Pacific Symposium on Biocomputing*, River Edge, NJ; World Scientific: Singapore, 1999; Vol *4*, pp 456−467.

(47) Gerlach, C.; Munzel, M.; Baum, B.; Gerber, H. D.; Craan, T.; Diederich, W. E.; Klebe, G. KNOBLE: a knowledge-based approach for the design and synthesis of readily accessible small-molecule chemical probes to test protein binding. *Angew. Chem.* **2007**, *46*, 9105−9109.
(48) Rittle, K. E.; Barrow, J. C.; Cutrona, K. J.; Glass, K. L.; Krueger, J. A.; Kuo, L. C.; Lewis, S. D.; Lucas, B. J.; McMasters, D. R.; Morrissette, M. M.; Nantermet, P. G.; Newton, C. L.; Sanders, W. M.; Yan, Y.; Vacca, J. P.; Selnick, H. G. Unexpected enhancement of thrombin inhibitor potency with o-aminoalkylbenzylamides in the P1 position. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 3477−3482.
(49) Linusson, A.; Gottfries, J.; Olsson, T.; Ornskov, E.; Folestad, S.; Norden, B.; Wold, S. Statistical molecular design, parallel synthesis, and biological evaluation of a library of thrombin inhibitors. *J. Med. Chem.* **2001**, *44*, 3424−3439.
(50) Lam, P. Y.; Clark, C. G.; Li, R.; Pinto, D. J.; Orwat, M. J.; Galemmo, R. A.; Fevig, J. M.; Teleha, C. A.; Alexander, R. S.; Smallwood, A. M.; Rossi, K. A.; Wright, M. R.; Bai, S. A.; He, K.; Luettgen, J. M.; Wong, P. C.; Knabb, R. M.; Wexler, R. R. Structure-based design of novel guanidine/benzamidine mimics: potent and orally bioavailable factor Xa inhibitors as novel anticoagulants. *J. Med. Chem.* **2003**, *46*, 4405−4418.
(51) Pierce, A. C.; Jorgensen, W. L. Estimation of binding affinities for selective thrombin inhibitors via Monte Carlo simulations. *J. Med. Chem.* **2001**, *44*, 1043−1050.