JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

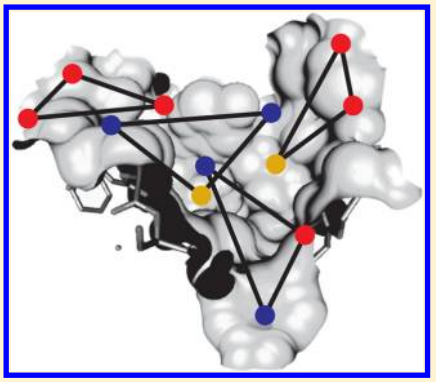
Fast Protein Binding Site Comparison via an Index-Based Screening Technology

Mathias M. von Behren, Andrea Volkamer, Angela M. Henzler, Karen T. Schomburg, Sascha Urbaczek, and Matthias Rarey*

Center for Bioinformatics, University of Hamburg, Bundesstrasse 43, 20146 Hamburg, Germany

Supporting Information

ABSTRACT: We present TrixP, a new index-based method for fast protein binding site comparison and function prediction. TrixP determines binding site similarities based on the comparison of descriptors that encode pharmacophoric and spatial features. Therefore, it adopts the efficient core components of TrixX, a structure-based virtual screening technology for large compound libraries. TrixP expands this technology by new components in order to allow a screening of protein libraries. TrixP accounts for the inherent flexibility of proteins employing a partial shape matching routine. After the identification of structures with matching pharmacophoric features and geometric shape, TrixP superimposes the binding sites and, finally, assesses their similarity according to the fit of pharmacophoric properties. TrixP is able to find analogies between closely and distantly related binding sites. Recovery rates of 81.8% for similar binding site pairs, assisted by rejecting rates of 99.5% for dissimilar pairs on a test data set containing 1331 pairs, confirm this ability. TrixP exclusively identifies members of the same protein family on top ranking positions out of a library consisting of 9802 binding sites. Furthermore, 30 predicted kinase binding sites can almost perfectly be classified into their known subfamilies.



INTRODUCTION

Due to large scale structural genomics projects, the amount of available protein structures in databases expands at an exponential rate.[1] Experimental methods for feature annotation are not able to keep up with this data quantity due to time and cost limitations. Thus, computational methods for automatic analysis, e.g., annotation of protein function or druggability, are of high practical relevance for pharmaceutical and biotechnological industry. Exploiting structures with annotated function, knowledge can be transferred to unknown proteins. Therefore, elucidating similarities between proteins or binding sites can help in many drug discovery contexts, e.g., to address drug promiscuity and polypharmacology. Success stories exist for predicting cross-reactivity,[2] adverse effects,[3] off-targets,[4] and multidrug resistance.[5] Furthermore, comparing active sites of enzymes can give hints about substrate specificity or potential mutation sites for enzyme optimization, and thus, assist in rational enzyme design for biotechnological research.

The number of computational methods for binding site comparisons is large[6,7] and notably either based on sequence or structural similarities. For a long period of time, sequence-based homology transfer has been the gold standard for protein annotation. Knowledge is transferred based on the similarity agreement of multiple sequence alignments of the complete protein sequence (BLAST[8]) or sequence motifs (PROSITE,[9] BLOCKS,[10] PRINT[11]). The fast progress in 3D protein structure elucidation, enhanced by the fact that structure was found to be more conserved than sequence,[12] recently promoted structure-based methods for protein comparison. Classically, structure-based comparison methods rely on multiple structural alignments of complete structures (FATCAT,[13] PAST,[14] VAST,[15] 3DCOMB[16]) or structural fragments (PROCAT[17]). Nevertheless, the amount of required overall sequence[18] or structure identity to reliably transfer function information is debatable. Examples showed that nonhomologous proteins—in terms of overall sequence or structure—also share functions, which shifted the focus toward binding site analysis. Specific interaction partners and their arrangement are responsible for the recognition and binding of small molecules, hence, determining the protein's function. Thus, most approaches follow the assumption that similar ligands bind to similar cavities.[19] As stated in several reviews,[6,7] the three main components of methods for binding-site comparison are molecular recognition feature encoding, similarity searching, and scoring. In the first step, the complexity of the comparison problem is reduced by using simplified representations of the binding site encoded in structural features. Second, similarities are identified for these representations, mostly by structural alignment overlap or fingerprint comparisons. Third, a scoring function is applied to quantify the similarity between two sites.

The number of approaches trying to solve the comparison problem is manifold and can be rudimentarily divided into alignment-based and alignment-free methods. Alignment-based algorithms rely on superimposing two structures. Strategies used for structural alignments are mostly geometric matching

Received: October 1, 2012
Published: February 7, 2013

   dx.doi.org/10.1021/ci300469h | J. Chem. Inf. Model. 2013, 53, 411−422

(ProSurfer,[20] SuMo,[21] SiteBase[22]), geometric hashing (SiteEngine[23]), or clique detection (CSC,[24] Cavbase,[25,26] eFsite,[27] eFseek,[28] IsoCleft,[29] ProBis[30]). Cavbase, e.g., uses a grid representation of the binding site, in which cavity-flanking residues are mapped to pseudocenters, representing the chemical properties of the binding site. Cavities are compared using a clique detection algorithm identifying three-dimensional (3D) pseudocenter arrangements that are common for two cavities. Since the alignment of structures is computationally expensive, effort is undertaken to develop alignment-free methods. A common approach is to convert cavity properties into simple 1D fingerprints, facilitating high-throughput comparisons.[31−38] Due to the speed of these methods, large screening scenarios with a few thousand up to a million binding site comparisons are feasible. One group of methods compares lists of sorted distances between, e.g., critical atoms (PocketMatch[31]), conserved atoms,[32] and surface curvature.[33] Other methods analyze distances between centroids of fragment pairs[34] or property-encoded shape distributions (PESD).[35] Using pharmacophore-based fingerprints is another prominent approach (FLAP,[36] SiteAlign,[37] FuzCav[38]). In SiteAlign,[37] a fingerprint overlap based on properties projected on an 80 triangle-discretized sphere is introduced. The mapping on the sphere comparison allows for easy alignments. A subsequent development, FuzCav[38] is completely alignment-free and performs comparisons based on a fingerprint of counts of pharmacophoric triplets. Moment-based methods use rotational invariant pocket representations by 3D mathematical functions that describe the protein surface space. Spherical-harmonics[19] or 3D Zernike descriptors[39] are employed to represent the structure as a vector of coefficients of the function series. Repurposing methods from other fields like image processing[40] or word processing[41] also proved useful. Merelli et al.,[40] e.g., used spin-images for surface matching. Pang et al.[41] calculate a "visual words" descriptor and uncover similarities between binding sites based on a fast algorithm from the information retrieval area. Ito et al.[42] achieved a good running time by representing the binding site as a bit string, combined with the application of ultra fast all pair similarity search methods.

Nevertheless, the speed of structural alignment free methods entails a lack of interpretability of the results. Besides the fingerprint-based similarity score, no information about the features responsible for the similarity is returned. Thus, the shortcomings of both—the slower character of alignment-based and the low interpretability of alignment-free methods—have recently been faced.[43,44] BSAlign,[43] e.g., finds the largest common subgraph based on clique detection and subgraph isomorphism with high-throughput. A sparse graph is built based on residues—together with geometric and physicochemical information—instead of point-based representations. An efficient algorithm has been invented to circumvent the computational expensive (NP-hard) problem of finding the maximum common subgraph.

In this work, we introduce TrixP, a new method for index-based binding site comparison which falls in the latter category of alignment-based but efficient algorithms. TrixP allows for fast structure-based screening of a query binding site against a library of precalculated sites. The method exploits the main advantages of TrixX,[45] a method for structure-based high-throughput screening. Pharmacophoric features present in the binding site are identified and a triangle descriptor—together with an 80-ray bulk spanned from the triangle center—is used to represent physicochemical and spatial information of the binding site. The use of a bitmap index[46] and an efficient data partitioning scheme avoids the sequential evaluation of binding sites. Binding sites can either be identified by providing a reference ligand or automatically predicted by the built-in DoGSite[47] method. For a query protein, descriptors are calculated and binding sites with matching descriptors are returned from the bitmap index. The respective sites are superposed onto the query based on calculated clusters of matching descriptors. A scoring scheme, considering matching and mismatching pharmacophoric interaction sites, is introduced to rank the library binding sites by their similarity to the query. The method is evaluated on a set of 1331 pairs[38] and successfully retrieves 81.8% of the similar pairs while rejecting 99.5% of the dissimilar pairs. These results are in good agreement with the results achieved by FuzCav.[38] Furthermore, an index is built on 9802 structures from the sc-PDB[48] and screened against four different protein families. Querying the index with an estrogen receptor, e.g., delivers a ranked list of similar sites with 98.5% of the contained estrogen receptors among the top ranking positions. Next, the method is used to classify kinases into subfamilies, and achieves classifications similar to those of Cavbase and SCOP.[49] Furthermore, the quality and runtime of TrixP is compared to other recently published methods, on a data set containing eight protein pairs sharing only partial similarities which are hard-to-detect.[31] TrixP finds similarities for seven of the eight pairs in a few seconds per comparison. Thus, the running time is comparable to BSAlign, another alignment-based algorithm and faster than earlier alignment-based methods, which are in the order of minutes. Nevertheless, alignment-based methods are still slower than 1D fingerprint methods, which perform comparisons in the order of milliseconds. Finally, high-throughput screening studies are executed in parallel on the eight cores of an Intel(R) Xeon(R) E5630 @ 2.53 GHz machine with 32 GB RAM. Building an index takes 6.3 h for the sc-PDB data set with 9802 protein binding sites but has to be done only once. The estrogen receptor query on this library including query preprocessing, matching, and scoring takes 37.5 min.

The high recovery rate and the speed of the method show its importance in fields like protein function prediction, rational enzyme design, and polypharmacology.

DATA PREPARATION

Sc-PDB Data Sets for Method Evaluation. The sc-PDB[48] data set, released in 2011, containing 9877 entries, corresponding to 3034 different proteins and 5339 different ligands, has been downloaded and used throughout this work. Due to nonstandard PDB annotations or errors in the respective ligand mol2 files, 75 of the entries are discarded by NAOMI,[50] yielding a total of 9802 structures in our data set.

To determine a reliable score cutoff within TrixP, a similar and dissimilar pair sc-PDB subset is used. The pair sets have originally been setup by Weill and Rognan[38] to define a cutoff for FuzCav. Starting from the complete sc-PDB (Version 2008), Weill et al. clustered the entries according to their UniProt name. Subsequently, they randomly selected two entries (only cofactor-free ones) from each cluster based on the SiteAlign[37] distance value, yielding 769 pairs. The same number of dissimilar pairs has been selected from the clusters, with the requirement of having an EC number differing at the first level. Due to changes between the sc-PDB versions and some discards by NAOMI, only 683 similar and 648 dissimilar pairs

Table 1. Overview of All Protein Pairs of the Benchmark Data Set

first protein	protein family	second protein	protein family
1gjc	utpa[a]	1v2q	trypsin
1gjc		2ayw	trypsin
1gjc		1o3p	utpa[a]
1ecm	chorismate mutase	4csm	chorismate mutase
1m6z	cytochrome c4	1lga	peroxidase
1zid	enoyl-ACP reductase	2cig	dihydrofolate reductase
1v07	mini-hemoglobin	1hbi	hemoglobin
6cox	prostaglandin G/H synthase 2	1oq5	carbonic anhydrase 2

[a]urokinase type plasminogen activator.

could be recovered from the 2008 version (used in SiteAlign) and are used within the TrixP study.

Since the entries of the sc-PDB are always annotated with a drug-like cocrystallized ligand, we used those ligands to determine the binding sites of the proteins within the sc-PDB experiments. Therefore, we selected every amino acid within a radius of 6.5 Å as part of the binding site. To evaluate the performance of TrixP on this data set, different protein families with a sufficient large amount of representatives are chosen. Iteratively, one randomly chosen representative of each of these families is used to query the complete data set. The chosen families are estrogen receptors (PDB codes: 1qkt, 1l2j, 2ewp), proteases (2q54), reverse transcriptases (1klm), and carbonic anhydrases (3bet). The sc-PDB contains 34 estrogen receptors α, 23 estrogen receptors β, and 5 estrogen related receptors γ. Furthermore, the sc-PDB contains five estrogen receptors, which are not further specified. According to the information present in the PDB entries of those proteins, two of them can be counted as estrogen receptors α (3l03 and 3hlv). The remaining three estrogen receptors (2yat, 3os9, and 3osa) have to be labeled as "unknown form" during the result evaluation. The other three families consist of 174 proteases, 75 reverse transcriptases, and 105 carbonic anhydrases.

Kinase Data Set for Subfamily Based Classification. To show TrixP's sensitivity in protein family annotation, we perform a classification study on kinases, an enzyme class which is of special biochemical interest. In 2006,[26] Kuhn et al. collected a set of eukaryotic protein kinases to evaluate the performance of Cavbase. This data set consists of 30 binding sites of 28 kinases from five different kinase subfamilies. The challenge within this classification problem lies in the separation of closely related subfamilies, containing active and inactive conformations with significant differences in the local conformation of the ATP binding sites. In this experiment, binding sites are predicted using the DoGSite[47] method, likewise for holo- and apo-structures . To distinguish between the different families, an all vs. all comparison is performed. The resulting similarity matrix is used as input for a hierarchical clustering procedure.

Benchmark Data Set for Comparison Study. To directly compare TrixP with other recently published methods for binding site comparison, a data set originally introduced by Yeturu et al.[31] and extended by Weill et al.[38] is used. The data set contains eight binding site pairs: three from the same SCOP family and five belonging to different SCOP families. The ligands present in 1v2q, 2ayw, and 1o3p are all bound to the same binding site, but in different orientations and interacting with varying residues. The four remaining pairs introduced by Yeturu et al. contain different folds, and therefore even belong to different SCOP families, but show similar binding sites as reported in the literature. Furthermore, Weill et al. included an additional pair of two proteins showing a cross-reactivity, explained by the similarity of a small-sized subpocket within both binding sites. For this data set, we again used the respective ligands to determine the binding sites. The pdb codes of the eight pairs present in the data set can be seen in Table 1.

METHODS

TrixP is based substantially on the TrixX technology, a novel approach for structure-based virtual screening of large compound libraries. For a detailed description of the technology we refer to the original publications.[45,51] Here, we briefly overview the basic concepts of TrixX and focus on the explanation of the adaptations, necessary to employ this technology for pocket similarity prediction. TrixP compares pockets, i.e., it screens a library of protein binding sites and identifies matching binding site descriptors. Therefore, TrixP follows the general TrixX proceeding of a first library indexing step followed by a screening step, in order to avoid repetitive calculations and to perform efficient virtual screening runs. Similarly in both methods, triangle descriptors for a given library are calculated and stored in a bitmap index, during the indexing phase. This bitmap index is created once and is reusable in subsequent virtual screening runs. In the screening phase, descriptors are derived for the binding site of a query protein and only matching descriptors and their associated structures are extracted from the index. Those hits are then transformed and scored according to their agreement of pharmacophoric features with the binding site of the query protein. The TrixP method mainly differs in four points from TrixX: First, instead of small molecules TrixP stores descriptors derived from protein binding sites in the index. Second, in order to account for the inherent flexibility of proteins, TrixP employs an adapted descriptor matching method. Third, instead of placing small molecules into the binding site, it aligns binding sites. And finally, TrixP uses a new scoring scheme that assesses the hits according to their similarity with the query protein. We introduce the new concepts in the following sections in more detail.

Recognition Feature Encoding. TrixP identifies the similarity between proteins by comparing pharmacophoric binding site features presumably responsible for the recognition of ligands. The starting point for the calculation of descriptors encoding these features is a binding site of a protein which can be determined using different strategies. The most straightforward way is the use of a reference ligand to identify surrounding residues or atoms. In this case, every amino acid within a distance threshold of 6.5 Å has been selected. Alternatively,

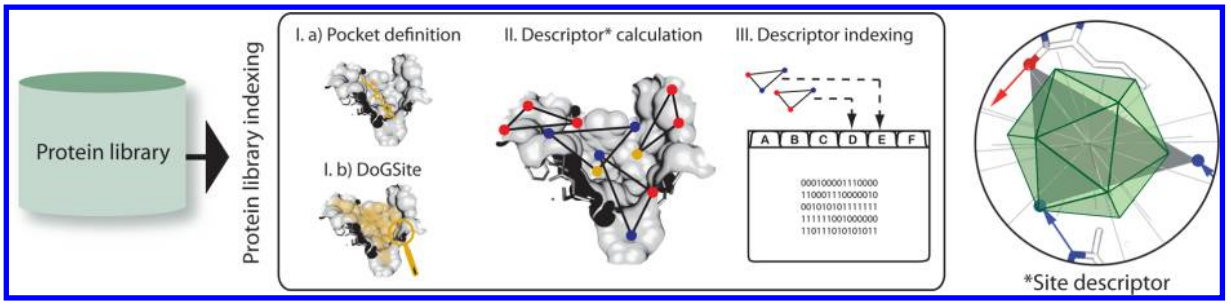


Figure 1. (left) Recognition feature encoding. For a given protein library, binding sites are detected and attributes of site descriptors are converted and stored in a bitmap index. (right) The Binding site descriptor encodes three types of pharmacophoric features in its triangle corners, three main interaction directions if the corners are of hydrophilic type, three triangle side lengths that describe the relative arrangement of the pharmacophores, and a set of 80 bulk rays through the 20 triangle faces of an icosahedron (four rays per triangle face) that locally describe the interior volume of a pocket.

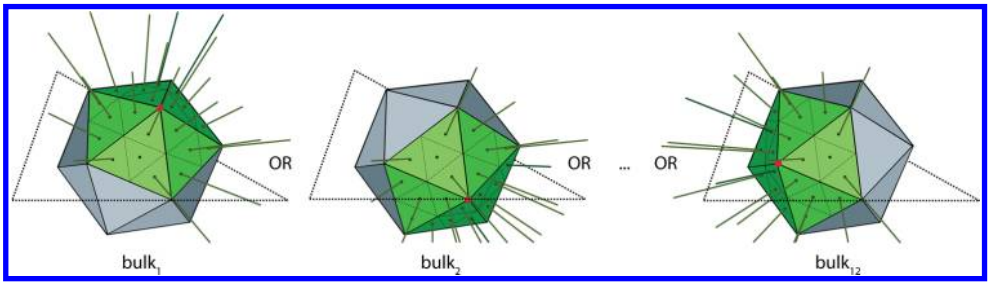


Figure 2. Partial bulk implementation: An icosahedron is orientated relative to a pharmacophoric triangle. For each vertex of the icosahedron, all rays going through a surrounding triangle are used to define a partial bulk. Combining them with a logical OR during descriptor matching indirectly introduces protein flexibility as only 25% of the shape of the compared binding sites have to match.

potential binding sites can be detected ligand-independently using the built-in DoGSite[47] method.

For each detected pocket, TrixX triangle descriptors are calculated based on present pharmacophoric features. The used triangle descriptors, hereby, resemble three-point pharmacophores, and the triangle corners have one of the three types: donor, acceptor, or hydrophobic. Hydrophilic features are generated from hydrogen-bond donor and acceptor atoms and possess potential main interaction directions. These directions indicate the locations of hydrogens or lone pairs, respectively. Hydrophobic features are derived from a grid placed in the binding site and are undirected. Grid points with a sufficient number of surrounding hydrophobic atoms represent hydrophobic regions and therefore become hydrophobic features. Since hydrophilic features are far more specific than hydrophobic features, triangle descriptors with only hydrophobic corners are not allowed. Nevertheless, hydrophobic features are of great importance since they increase the number of possible active site superpositions. Additionally, the descriptor is equipped with 80 steric bulk rays, aligned in an icosahedron, radiating from the center of the triangle. These rays represent the shape of the pocket relative to the triangle, since the length of each ray is the distance of the triangle center to the surface of the binding site. Due to the large number of possible triangle descriptors (on average 6090) per binding site, the derived data requires an efficient space management. For descriptors derived from the protein library, this is realized by binning the features of the descriptors and converting them into bitmaps. Thereby, the descriptors are separated and stored according to their descriptor attributes (types of corners, directions, lengths of triangle sides and of bulk rays) in the triangle descriptor index. Figure 1 summarizes the workflow of recognition feature encoding and depicts a descriptor of a binding site.

Similarity Searching. A comparison of query and library descriptors can reveal similarities between associated proteins. Therefore, descriptors are generated from the query protein and used to formulate logical expressions that directly access and extract only similar descriptors, and thus, similar pockets from the index. A query descriptor matches if the types of the corners, the directions, the lengths of the triangle sides, and the lengths of each of the 80 bulk rays are in accordance with a descriptor of the index. In order to allow a certain amount of structural flexibility during the matching procedure, tolerances are added to the lengths of the triangle sites and bulk rays. Then, the associated structures are identified as potential similar proteins. Due to data partitioning by type, the index structure avoids the evaluation of dissimilar descriptors and supports a rapid data querying. However, since even closely related binding sites exhibit differences in their overall shape, the bulk descriptor matching in its original form turned out to be a too rigorous matching criterion. Furthermore, the TrixX bulk descriptor ensures that the ligand completely fits into the binding site avoiding steric overlap. While steric overlap is forbidden in general, shape similarity considered in TrixP can also occur partially. Therefore, the shape-descriptor matching procedure is adapted to allow matches with only partial shape agreement. As depicted in Figure 2, this partial bulk implementation uses multiple subsets of rays as matching criteria for querying. A shape requirement of 25% is realized using only those rays going through triangles surrounding a particular icosahedron vertex. Since a vertex is enclosed by five triangles, only 20 out of 80 rays are selected as matching constraints at a time. Each vertex of the icosahedron defines a subset of rays leading to 12 possible sets of rays for a single query descriptor. These subsets are logically ORed during the evaluation of a query descriptor, i.e., it is sufficient if only 25%

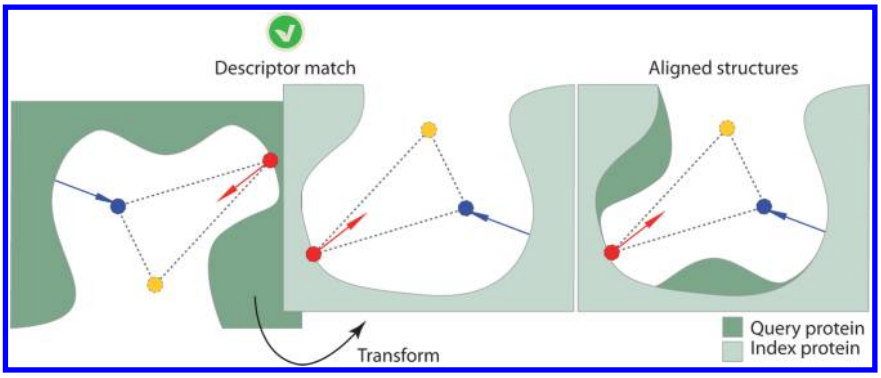


Figure 3. Structure alignment: Schematic superposition of binding sites based on a descriptor match. The colors of the triangle corners indicate their respective type: hydrophobic (yellow), hydrogen-bond donor (blue), or hydrogen-bond acceptor (red).

of the binding site shape of a library protein matches to indicate similarity with the query protein. Note that this matching 25% has to occur in one connected region of the active site.

Structure Alignment. The search procedure of TrixP results in a list of matching query and library descriptors. Each triangle descriptor match holds the information to superpose a pair of binding sites. The transformation of the query triangle onto the matching index triangle is calculated and applied to the coordinates of the query binding site. In order to reduce the number, as well as to improve the quality, of the transformations, matching descriptor pairs are clustered.[52] The aim of the clustering is to identify groups of descriptor pairs whose transformation results in a very similar superimposition of the binding sites. A complete linkage clustering algorithm compares the descriptor matches on the basis of their transformation result. To evaluate the distance between any two descriptor matches, both query descriptors are transformed once with each of the two corresponding transformations. The RMSD between the resulting triangle descriptor corner coordinates of both transformations is used as distance measure for the clustering algorithm. In the end, only one combined transformation is calculated for each cluster by optimizing the simultaneous overlay of all included query pharmacophore points on their respective matching points in the index structure. This transformation is applied to superimpose the binding sites for subsequent scoring. Figure 3 illustrates the structure alignment resulting from a transformation based on a matching triangle descriptor.

Scoring. For each returned alignment of the query to a binding site of the library, named target in the following, a similarity score is calculated based on the compliance of pharmacophoric features. Therefore, let Q be the set of features q of a query protein and T be the set of target features t. Furthermore, let $A(Q, T)$ be the set of alignments of Q onto T gained by the clustering method. The similarity $S(Q, T)$ between Q and T maximizes the similarity scores $s_a(Q,T)$ of all structural alignments of $A(Q, T)$. $s_a(Q, T)$ is determined by scanning the environments of the query features for matching features in the target.

$$S(Q, T) = \max_{a \in A(Q,T)}\{s_a(Q, T)\} \tag{1}$$

$$s_a(Q, T) = \sum_{i=1}^{|Q|} (s_a(q_i, T_{\text{sphere}}(q_i))) \tag{2}$$

The function mtype(q, t) discriminates between three different matching scenarios:

$$\text{mtype}(q, t) = \begin{cases} \text{dir:} & \text{if } q, t \text{ have the same directed type (directed match)} \\ \text{undir:} & \text{if } q, t \text{ have the same undirected type (undirected match)} \\ \text{mis:} & \text{if } q, t \text{ have different types (mismatch)} \end{cases} \tag{3}$$

Furthermore, we define $T_{\text{sphere}}(q_i)$ to be the set of target features with a maximum distance $d_{\max}$ = 1.5 Å from q_i, i.e., $T_{\text{sphere}}(q_i) = \{t \in T | d(q_i,t) \leq d_{\max}\}$. For each query feature q_i of an alignment a, $s_a(q_i, T_{\text{sphere}}(q_i))$ honors matching and penalizes mismatching features and, thus, reflects the similarity of feature q_i to its close environment:

$$s_a(q_i, T_{\text{sphere}}(q_i)) = \begin{cases} \max\limits_{t_j \in T_{\text{sphere}}(q_i)} \{s_{\text{undir}}(q_i, t_j)\}: & \text{if only hydrophobic matches} \\ \frac{1}{n}\left(\sum\limits_{j=1}^{n} (s_{\text{mtype}(q_i,t_j)}(q_i, t_j)\right): & \text{otherwise} \end{cases} \tag{4}$$

Where n is the number of target features within $T_{\text{sphere}}(q_i)$. The individual similarity scores $s_{\text{dir}}(q_i, t_j)$ of directed hydrophilic matches, $s_{\text{undir}}(q_i, t_j)$ of hydrophobic matches, and $s_{\text{mis}}(q_i, t_j)$ of mismatches of the query feature q_i and the target feature(s) are defined as follows:

$$s_{\text{dir}}(q_i, t_j) = s_{\max}\left[\left(1 - \frac{d(q_i, t_j)}{d_{\max}}\right) + w\left(1 - \frac{\alpha(q_i, t_j)}{\alpha_{\max}}\right)\right] \tag{5}$$

$$s_{\text{undir}}(q_i, t_j) = s_{\max}\left(1 - \frac{d(q_i, t_j)}{d_{\max}}\right) \tag{6}$$

$$s_{\text{mis}}(q_i, t_j) = p_{\max}\left(1 - \frac{d(q_i, t_j)}{d_{\max}}\right) \tag{7}$$

Figure 4 illustrates possible matching cases that might occur during the scanning and scoring of local query feature environments.

(a) The similarity $s_{\text{dir}}(q_i, t_j)$ between directed hydrophilic features is determined linearly based on the distance $d(q_i, t_j)$ and the angle difference $\alpha(q_i, t_j)$ between the main interaction directions of the features. Therefore, we define the maximal score $s_{\max}$ = 10 and the angle weight parameter w = 0.8, resulting in an absolute score of 18 for

dx.doi.org/10.1021/ci300469h | J. Chem. Inf. Model. 2013, 53, 411−422

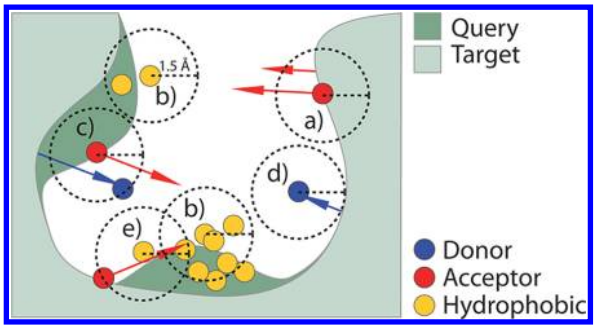


Figure 4. Schematic depiction of the scoring procedure: Around a query feature a sphere with 1.5 Å radius is placed and scanned for matching target features. (a) Match of hydrophilic features. (b) Match of hydrophobic features. (c) Mismatch of features. (d) No match within the 1.5 Å around a donor feature. (e) Mismatch between a hydrophobic and an acceptor feature, and a simultaneous match with a hydrophobic feature.

a perfect overlay of a query and a target feature. The hydrophilic score drops to 0 if d_{max} = 1.5 Å and α_{max} = 65° is reached.

(b) As hydrophobic features are undirected, the score $s_{undir}(q_i, T_{sphere}(q_i))$ only depends on the distance of the features. However, since hydrophobic features often appear in clusters of discrete features describing hydrophobic binding site volumes, a special case is introduced for the match of a hydrophobic feature with multiple other hydrophobic features. In this case, the score $s_{undir}(q_i, t_i)$ is maximized over all hydrophobic target features in $T_{sphere}(q_i)$.

(c) If the types of features differ but are located close to each other, this mismatch of query and target features is penalized by $p_{max} = -2$.

(d) Generally, if no match or mismatch is identified in $T_{sphere}(q_i)$, there is no contribution to the score since there is no evidence for similarity or an explicit mismatch in such a case.

(e) Simultaneous matches or mismatches in the query feature sphere are seldom. However, in these rare cases their contributions are averaged in order to account for the heterogeneity of matched regions.

Finally, in order to grant comparability, the similarity score $s_a(Q, T)$ is normalized with respect to the query protein to reflect a value between 0 and 1. All chosen parameters within the equations, e.g. maximal score, distance, and angle, as well as angle weight and maximal penalty, have been optimized on a small training set (see Supporting Information Material A) and proved to produce a reliable score for the overall similarity of two binding sites.

RESULTS AND DISCUSSION

In the following, different aspects of the presented binding site comparison tool are analyzed. First, TrixP is evaluated in terms of its ability to find similar sites while discarding dissimilar ones based on ligand-defined binding sites, with respect to studies from FuzCav.[38] Second, the methods capability to distinguish between subfamilies based on predicted binding sites is investigated and the results are compared to Cavbase.[53] Finally, several benchmark studies are executed comparing TrixP to other recently published efficient algorithms[38] and showing its potential as a high-throughput method.

Separating Similar from Dissimilar Protein Pairs. In a first experiment, the pair data set introduced by Weill et al.[38] is used to determine a reliable cutoff value for the TrixP similarity score. Two indices are built containing all similar and dissimilar pairs, respectively. For each pair A, B, protein A is used to query the corresponding index.

First, the TrixP similarity score between each pair is investigated. The average score of all similar pairs is 0.46, while the average score of all dissimilar pairs is 0.17. A histogram of the achieved scores for similar and dissimilar pairs (see Figure 5) shows that a TrixP score of 0.3 is well suited to distinguish between similar and dissimilar binding sites. With this cutoff value, 81.8% of all similar pairs can be retrieved, while 99.5% of all dissimilar pairs are discarded.

Second, aside from the pairwise comparisons of proteins, the screening procedure allows to rank the respective partner relative to all other 1366 proteins in the index of the pair screening run. The first finding is that the method recovers the protein itself, which is also contained in the index, as top ranking hit (self-match) in all cases. This self-match is excluded from the following analysis. Figure 5 shows the distribution of the position at which the respective similar pair occurs within the result list. TrixP retrieves 69.4% of all similar pairs at the top ranking position. Furthermore, only 18.7% of the pairs are found on a rank below four. In total, 209 respective pairs do not occur at the first rank. In 54% of the cases, the best ranking hit as well as the query have an annotated EC number, which can be used to assess the quality of those matches. In the majority

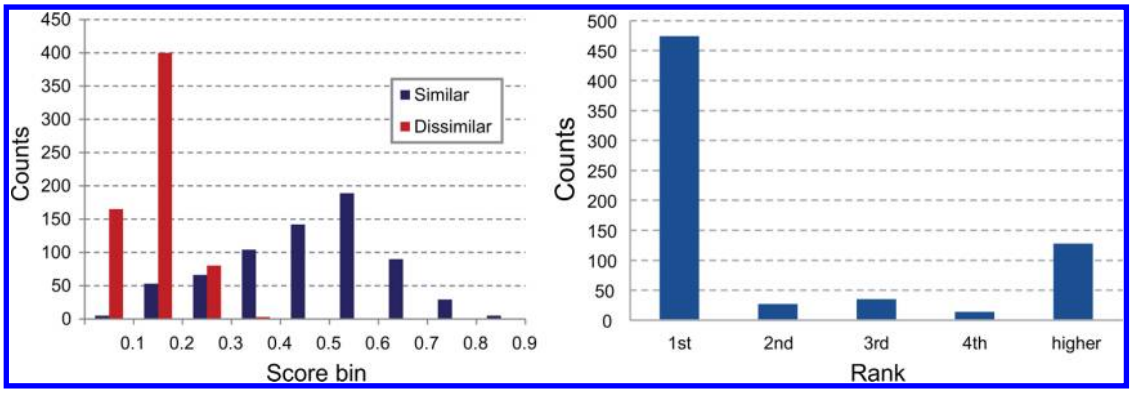


Figure 5. Results of the sc-PDB pair data set screening. (left) Distribution of the achieved scores for similar (blue) and dissimilar (red) pairs, displayed is the respective number of pairs within a certain score range. (right) Number of respective pairs found on a certain rank within the screening results sorted by similarity score.

dx.doi.org/10.1021/ci300469h | J. Chem. Inf. Model. 2013, 53, 411−422

of those cases, the top ranking match had at least the first three EC digits in common with the query. Only for as few as 0.7%, the method is unable to recover the respective pair at all. A further examination of those cases showed problems during the correct binding site determination, leading to disproportionally small binding sites.

In order to further demonstrate the discriminative power of TrixP, the scores of all similar (true positives) and dissimilar pairs (false positives) are sorted in descending order resulting in a ROC curve (see Supporting Information Material B). Generally, in the first 41.5% of the ranked data points true positives are exclusively found. An AUC of 0.96 is achieved for the performance of TrixP. Although the pair data set slightly differs (see above), these results are in good agreement with the data published by Weill et al.[38]

sc-PDB Screening. Multiple screening runs are performed on an index containing all 9802 binding sites of the sc-PDB data set, measuring the potential of TrixP to select binding sites similar to a query site. Six proteins from four different protein families, i.e., carbonic anhydrase 2 (CA2), protease (PR), reverse transcriptase (RT), and estrogen receptor (ER), are chosen as examples to query the index. For all queries, in total only 1−9% of the binding sites in the data set are returned as matches with a TrixP score above 0.3 (Table 2). For each target, the number of family members present in the index is assigned beforehand and the recovery rate per target is analyzed. Between 84% and 100% of the contained family members can be recovered for the respective queries. Furthermore, the 50 top ranking positions are occupied by members of the same family in 88% up to 100% of the cases.

Similar to an experiment performed in the evaluation of SiteAlign,[37] the ERα, ERβ, and ERγ queries are further investigated. The query with an ERα receptor (1qkt) retrieved in total 98.5% of the ERs present in the library, more precisely all ERα, all ERβ, all ERγ structures, and two out of the three nonspecified estrogen receptors. The missed nonspecified estrogen receptor (1qkn) achieved a score of 0.20. In contrast to the query, which had been crystallized in complex with the estrogen estradiol, the antagonist raloxifene is bound to 1qkn. These two ligands differ significantly, especially concerning their size. Since the bound ligands have been used to determine the binding sites of the proteins, the significant differences of the bound ligands might be the reason for the low similarity score in this case. Using an ERβ structure (1l2j) as query retrieved all 23 ERβ structures, and additionally all 67 other present estrogen receptors. Finally, for an ERγ (2ewp) structure as query, four of five ERγ, 34 of 36 ERα, all ERβ structures, and two out of three nonspecified estrogen receptors are recovered. The results for ERα are further analyzed, with respect of high-scoring family and nonfamily members. Two out of the four missed ERs during this screening run, ERγ 2gpp and nonspecified ER 3os9, still achieved a score higher than 0.29. Even if the threshold of 0.3 had not been exceeded in these two cases, both receptors still show a relatively high similarity to the query. Note, that similarity is always rated with respect to the query protein. In the case of the two missed ERα proteins, their ligands differ from the ligand bound to 2ewp and might cause the low similarity. Figure 6 shows the top ranking binding sites up to rank 150. The 16 top ranking positions are exclusively occupied by ERαs, which are still dominant on ranks up to 40.

Table 2. sc-PDB Screening Results Using Different Queries

protein family	PDB code	family hits	present in sc-PDB	members within top 50	general hits with score > 0.3
CA2	3bet	100	105	50	385
PR	2q54	147	174	50	151
RT	1klm	63	75	48	162
ERα	1qkt	36(66)[a]	36(67)	31(49)	843
ERβ	1l2j	23(67)	23(67)	14(44)	467
ERγ	2ewp	4(63)	5(67)	4(45)	245

[a]Number in parenthesis represents the combined number of all estrogen receptors.

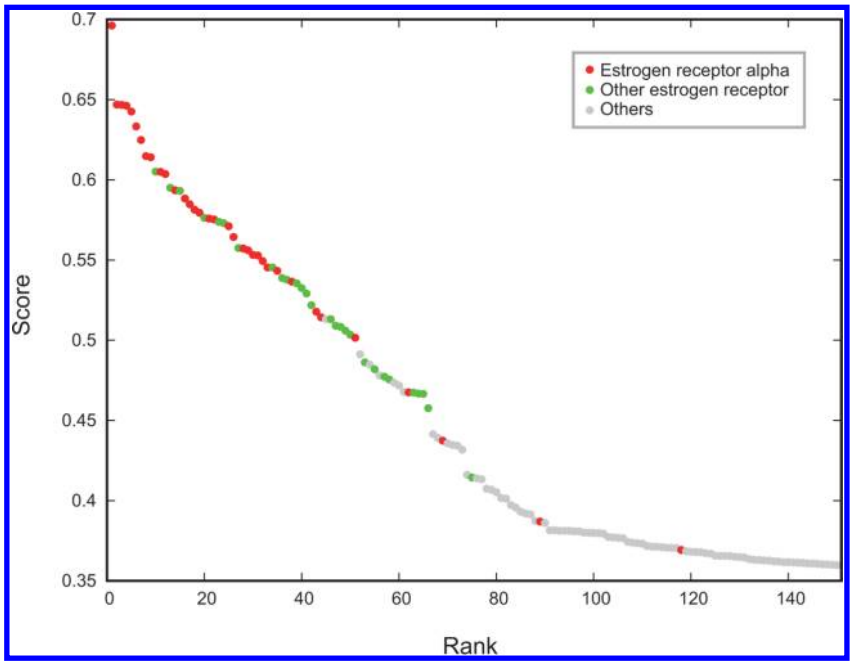


Figure 6. Results for the query with an ERα (PDB code: 1qkt) ranked by TrixP score. ERα and other estrogen receptors are colored in red and green, respectively. Nonestrogen receptor family matches are colored in gray.

dx.doi.org/10.1021/ci300469h | J. Chem. Inf. Model. 2013, 53, 411−422

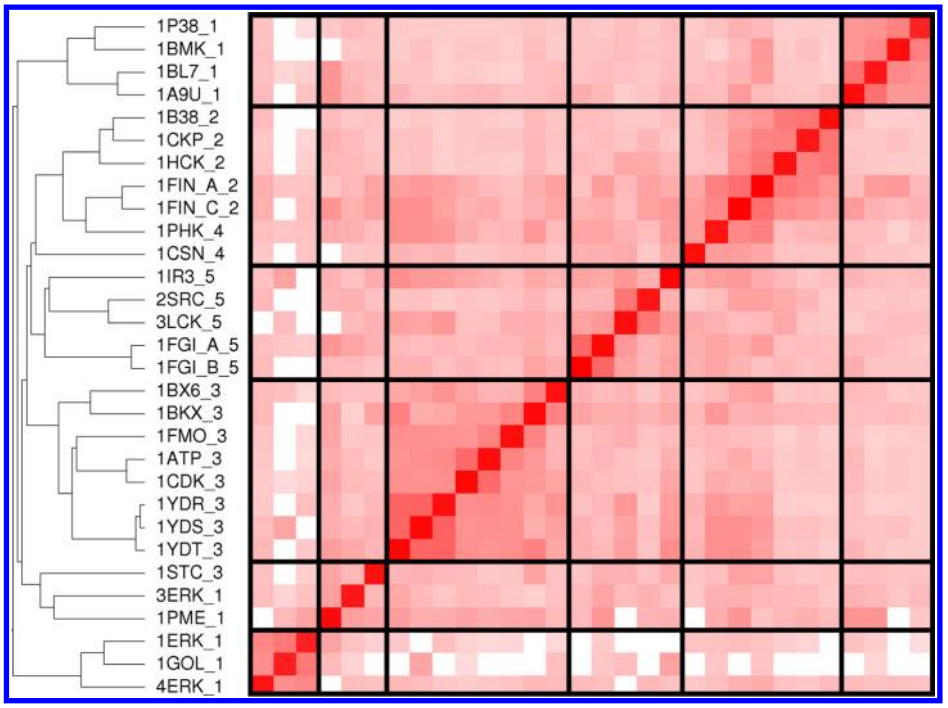


Figure 7. Agglomerative clustering of 30 kinase pockets by TrixP similarity score. SCOP annotation of the structures is indicated as a number: MAP kinases (1), CDK2 (2), PKA (3), Ser/Thr kinases (4), and Tyr kinases (5).

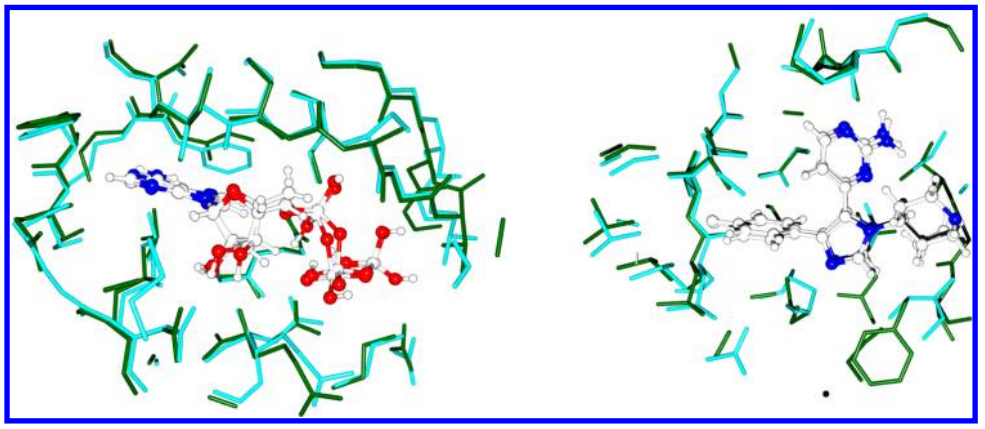


Figure 8. Superimposition of the two cycline-dependent kinases 2 (CDK2) 1b38 (cyan) and 1hck (dark green) on the left and of the two mitogen-activated protein (MAP) kinases 1bmk (cyan) and 1bl7 (dark green) on the right.

ERβs and ERγs capture most of the following ranks up to 70. Furthermore, the eight non-ER proteins found within these ranks belong to other human nuclear receptors also binding steroid hormones like progesterone (four), aldosterone (one), glucocorticoid (one), and mineralocorticoid (two).

Kinase Subfamily Detection. In a third experiment, the kinase data set introduced by Kuhn et al.[53] is used to evaluate the ability of TrixP to distinguish between closely related binding sites. Binding sites are predicted with the built-in DoGSite method,[47] and the resulting 30 kinase pockets are stored in the index. The index is queried with all pockets and a hierarchical clustering is performed on the resulting similarity matrix. As previously described in the scoring section, the score of TrixP is calculated with respect to the query binding site and, therefore, not symmetric by design. In order to account for the fuzziness in the definition of the pocket boundary, the maximum of the two respective scores for comparing A vs B and B vs A is used. Applying an agglomerative clustering results in six clusters (Figure 7). SCOP annotations are indicated by number: MAP kinases (1), CDK2 (2), PKA (3), Ser/Thr kinases (4), and Tyr kinases (5). Clearly, the clustering based on TrixP similarity is in good agreement with the SCOP classification. All Tyr kinases (5) and PKA (3) structures aggregate within one cluster, respectively. The Tyr kinase cluster hereby contains two active (1ir3 and 3lck) as well as three inactive structures (1fgi_a, 1fgi_b, and 2src), which were nevertheless correctly classified as members of the same family. The only exception, hereby, is the PKA structure 1stc, which ended up in a cluster mostly occupied by MAP kinase structures. The binding of a large rigid inhibitor (STU) to 1stc may have introduced a change in its binding site conformation, causing the missclassification. Similar to the findings by Cavbase[53] on this data set, TrixP is able to distinguish between the different activation states of CDK2s. The CDK2 (2) main cluster contains two subclusters: one is occupied by inactive CDK2s (1b38, 1ckp, 1hck), and the other one contains active

dx.doi.org/10.1021/ci300469h | J. Chem. Inf. Model. 2013, 53, 411−422

CDK2s (1fin chain A and C). Furthermore, the two Ser/Thr kinases (4) are assigned to the CDK2 cluster. Although the ligand is not considered within this experiment, all structures (except 1ckp) of this cluster contain a bound ATP, and thus, the method detects the common interaction points within similar distances present in these structures. The MAP kinases (1) span over three clusters. One cluster exclusively contains all structures from the p38α subfamily (1bmk, 1p38, 1bl7, and 1a9u). A second MAP kinase cluster holds only structures from the Erk2 subfamily (1gol, 1erk, and 4erk). The third cluster contains the remaining two Erk2 structures (1pme and 3erk), paired with the only miss-annotated PKA kinase 1stc. The correct classification of active as well as inactive structures within certain families, like the Tyr kinases and CDK2 kinases, proved the flexibility of TrixP regarding local changes within overall similar binding sites. Figure 8 shows the superimpositions of the CDK2 structures 1b38 and 1hck and of the MAP kinases 1bmk and 1bl7, as examples.

Qualitative and Quantitative Comparison to Other Methods. To compare the performance of TrixP with other recent methods, a small set of eight difficult targets is investigated. Yeturu et al.[31] and Weill et al.[38] evaluated the performance of multiple recent methods on this data set, concerning their ability to identify similarities as well as their run time requirements. As shown in Table 3, TrixP was able to assign a similarity score to seven out of those eight difficult pairs. Regarding the three pairs of proteins belonging to the same SCOP families, TrixP like most other methods detects similarities between the sites and exhibits a similar score trend as BSAlign[43] and PocketMatch,[31] by assigning a higher score to the pair of urokinase type plasminogen activators (1gjc and 1o3p). For the five pairs of proteins belonging to different SCOP families, TrixP and PocketMatch are the only methods able to derive a score for four out of the five present pairs, while FuzCav[38] is the only method assigning a score to the new pair of a prostaglandin G/H synthase 2 (6cox−1oq5). Furthermore, the TrixP and PocketMatch comparably assign higher scores to the pairs of a cytochrome c4 with a peroxidase (1m6z−1lga) and of a mini-hemoglobin with a hemoglobin (1v07−1hbi). Using the determined threshold of 0.3 for the TrixP similarity score would only yield two cases of possible cross-reactions or related function among the eight pairs present in this data set. Nevertheless, the results of TrixP show a certain degree of similarity for five of the six remaining pairs and therefore confirm the possibility of the observed cross-reactions. Furthermore, TrixP is able to reproduce the same score trends among the different pairs as BSAlign. Figure 9 shows the superimposition of mini-hemoglobin 1v07 and hemoglobin 1hbi as calculated by TrixP. The calculated score of 0.43 indicates high similarity between the two binding sites even if they belong to different SCOP families and therefore have different folds. The figure shows an almost perfect superimposition of the two heme groups with an RMSD of 0.93 and reasonable alignments of some residues common in both sites. In terms of run-time requirements, the pairwise comparison of the eight protein pairs, using TrixP, takes on average 19 s, including index querying and scoring. Thus, TrixP performs in the same speed order (seconds) as BSAlign, another method designed for efficient alignment-based comparison. Furthermore, both methods are faster than general alignment-based methods executed on this data set (ProFunc,[54] SitesBase,[55] SuMo,[21] SiteEngine[56]), as can be seen in the extended table within the publication of Weill et al. But clearly, the performance of alignment based-methods is still slower than the millisecond range of efficient fingerprint-based methods, such as PocketMatch and FuzCav, which on the other hand often produce results with a lack of interpretability.

Table 3. Comparison of TrixP to an Extraction of Other Recently Published Binding Site Comparison Tools[a]

		efficient alignments		fingerprints	
PDB1−lig1	PDB2−lig2	TrixP[b]	BSAlign[c]	PocketMatch[d]	FuzCav[e]
pairs of proteins belonging to the same SCOP family					
1gjc−130	1v2q−ANH	0.18	31.77	50.17	0.19
	2ayw−ONO	0.27	31.51	52.29	0.18
	1o3p−655	0.65	42.26	88.01	0.18
pairs of proteins belonging to different SCOP families					
1ecm−TSA	4csm−TSA	0.16	×	55.56	0.18
1m6z−HEC	1lga−HEM	0.24	×	63.85	×
1zid−ZID	2cig−IDG	0.19	×	56.01	×
1v07−HEM	1hbi−HEM	0.43	×	61.42	0.18
6cox−S58	1oq5−CEL	×	×	×	0.16
speed order		s	s	ms	ms

[a]The full list can be found in the publication of Weill et al.[38] [b]TrixP similarity score. [c]BSAlign alignment score.[43] [d]PocketMatch PMScore.[31] [e]FuzCav similarity score.[38]

Pocket-Based High-Throughput Screening. The TrixP high-throughput screening process can be parallelized by splitting up the data into subindices, simultaneously screening each on one CPU core. As a test scenario, the index in this study is split into eight equal parts, and TrixP is run on the eight cores of an Intel(R) Xeon(R) E5630 @ 2.53 GHz with 32 GB RAM.

The most time-consuming task within TrixP is the initial creation of indices. Calculating descriptors for one binding site and writing them into the bitmap index takes on average 14.25 s. Thus, building the sc-PDB index containing 9802 structures, when equally split onto eight cores for parallel screening, takes 6.3 h, but has to be done only once. The time for screening an index with a protein query depends on two components: First, the number of structures in the index and second, the size of the query's binding site. The first part within TrixP is the matching phase. The average time needed to evaluate the sc-PDB index is 1.76 s per query descriptor. The efficiency within TrixP arises from the usage of the index technology. First, the sequential screening scheme is overcome by efficient horizontal data partitioning based on the descriptor's triangle corner types. Second, the number of binding sites to be scored is greatly reduced to the number of matches returned by the initial index query. The second part of TrixP captures postprocessing—from reinitialization to superposition and scoring. Hence, postprocessing can be done on average in 1.18 s per matching binding site returned by the index query. For ERα (1qkt), 5179 descriptors are calculated, a number close to the average number of 6090 descriptors per binding site for the sc-PDB data set. Using this ERα to query the sc-PDB index, the parallel screening finishes after 37.5 min. Note that the scoring time needed to screen each of the subindices could be further reduced by splitting the data either more reasonable or onto more cores.

dx.doi.org/10.1021/ci300469h | J. Chem. Inf. Model. 2013, 53, 411−422

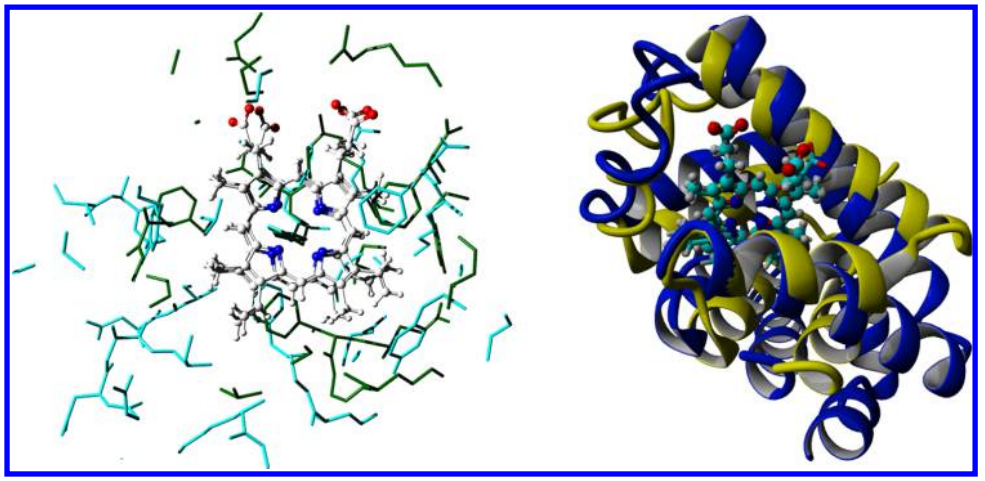


Figure 9. (left) Superimposition of the binding sites of mini-hemoglobin 1v07 (cyan) and the hemoglobin 1hbi (dark green). The bound hemoglobin of each structure is shown in ball-and-stick modus, color-coded in white (except oxygens (red) and nitrogens (blue)). (right) Three-dimensional depiction of the superposed secondary structures of 1v07 (yellow) and 1hbi (blue). The bound hemoglobins are color-coded in cyan (except oxygens (red), nitrogens (blue), and hydrogens (white)).

CONCLUSION

Due to the growing amount of available protein structures, computer methods are required to efficiently tackle the annotation problem. In this study, we introduced TrixP, a new method for fast binding site comparison and function prediction based on structural alignments of the binding sites. The main invention is hereby the representation of binding sites by chemical and structural triangle descriptors, stored in a bitmap index technology, allowing for time-efficient screening.

In multiple experiments, the ability of TrixP to efficiently produce reliable results, comparable or partially superior to other state of the art methods, is shown. Screening two data sets containing known similar and dissimilar binding sites, a reliable cutoff value for the TrixP similarity score is determined. With this cutoff value, 81.8% of all similar pairs can be recovered with TrixP, while rejecting 99.5% of all dissimilar pairs. Furthermore, 69.4% of all similar pairs have been ranked at position one of 1331 screened binding sites. Large scale screening experiments using four different protein families as a query against the sc-PDB index containing 9802 structures are performed. TrixP is capable of identifying similar binding sites to the respective query, to assign an appropriate score to them, and thus, rank related above unrelated binding sites. For each tested protein family, TrixP recovers at least 84% of all family members present in the library. Another experiment on a small data set containing representatives of five kinase subfamilies proved TrixP's ability to distinguish between closely related binding sites.

Besides the quality assessment of TrixP, the efficiency of the method is investigated on a prereleased comparison study on eight binding site pairs. The experiments showed that TrixP is able to perform pairwise comparisons in a few seconds while recovering similarities between so-classified difficult binding sites. Parallel screening, using eight cores, allows TrixP to build the index for the whole sc-PDB database within 6.3 h and afterward to screen it within only 37.5 min.

The application scenarios show the assistance of binding site comparison tools like TrixP to solve important and challenging tasks of today's biochemical research. Nevertheless, as some studies indicate, geometric rearrangements of some amino acid side chains result in different similarity scores. As demonstrated with the kinase data set, TrixP already is able to take into account a certain amount of protein flexibility by its representation of rotatable hydrophilic interactions as well as by using tolerance values for the matching of the lengths of triangle sides and bulk rays. However, there is still room for further improvement. Especially, large changes of the structure like different possible folds could not be handled by the recent version of TrixP. Another improvement of TrixP might be to also value the shape similarity of two binding sites during the scoring procedure.

ASSOCIATED CONTENT

Supporting Information

Data set used for parameter training (A) and the ROC curve for the similar and dissimilar pair data set (B). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The first two authors, Mathias von Behren and Andrea Volkamer, contributed equally to this work. We thank Christin Schärfer who originally developed the method for partial matching of bulk rays during her diploma thesis in 2008. Furthermore, we thank Didier Rognan for providing us with the FuzCav pair data and Daniel Kuhn for help with the Cavbase kinase cluster experiment. Components of the presented work emerged from the COMPASITES project from the Biokatalyse2021 cluster and were funded by the BMBF under grant 0315292A.

REFERENCES

(1) Sleator, R.; Walsh, P. An overview of in silico protein function prediction. *Arch. Microbiol.* **2010**, *192*, 151−155.

(2) Stauch, B.; Hofmann, H.; Perkovic, M.; Weisel, M.; Kopietz, F.; Cichutek, K.; Munk, C.; Schneider, G. Model structure of APOBEC3C reveals a binding pocket modulating ribonucleic acid interaction

required for encapsidation. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 12079−12084.
(3) Xie, L.; Wang, J.; Bourne, P. In silico elucidation of the molecular mechanism defining the adverse effect of selective estrogen receptor modulators. *PLoS Comput. Biol.* **2007**, *3*, e217.
(4) Xie, L.; Bourne, P. Structure-based systems biology for analyzing off-target binding. *Curr. Opin. Struct. Biol.* **2011**, *21*, 189−199.
(5) Kinnings, S.; Liu, N.; Buchmeier, N.; Tonge, P.; Xie, L.; Bourne, P. Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput. Biol.* **2009**, *5*, e1000423.
(6) Nisius, B.; Sha, F.; Gohlke, H. Structure-based computational analysis of protein binding sites for function and druggability prediction. *J. Biotechnol.* **2012**, *159*, 123−134.
(7) Kellenberger, E.; Schalon, C.; Rognan, D. How to measure the similarity between protein ligand-binding sites? *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 209−220.
(8) Altschul, S.; Madden, T.; Schaffer, A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(9) Henikoff, J.; Greene, E.; Pietrokovski, S.; Henikoff, S. Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.* **2000**, *28*, 228−230.
(10) Attwood, T. The PRINTS database: a resource for identification of protein families. *Briefings Bioinf.* **2002**, *3*, 252−263.
(11) Sigrist, C.; Cerutti, L.; Hulo, N.; Gattiker, A.; Falquet, L.; Pagni, M.; Bairoch, A.; Bucher, P. PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings Bioinf.* **2002**, *3*, 265−274.
(12) Illergard, K.; Ardell, D.; Elofsson, A. Structure is three to ten times more conserved than sequence−a study of structural response in protein cores. *Proteins* **2009**, *77*, 499−508.
(13) Ye, Y.; Godzik, A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* **2003**, *19* (Suppl 2), ii246−255.
(14) Taubig, H.; Buchner, A.; Griebsch, J. PAST: fast structure-based searching in the PDB. *Nucleic Acids Res.* **2006**, *34*, W20−3.
(15) Gibrat, J.; Madej, T.; Bryant, S. Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **1996**, *6*, 377−385.
(16) Wang, S.; Peng, J.; Xu, J. Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling. *Bioinformatics* **2011**, *27*, 2537−2545.
(17) Wallace, A.; Laskowski, R.; Thornton, J. Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.: Publication Protein Soc.* **1996**, *5*, 1001−1013.
(18) Rost, B. Enzyme function less conserved than anticipated. *J. Mol. Biol.* **2002**, *318*, 595−608.
(19) Morris, R.; Najmanovich, R.; Kahraman, A.; Thornton, J. Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics* **2005**, *21*, 2347−2355.
(20) Minai, R.; Matsuo, Y.; Onuki, H.; Hirota, H. Method for comparing the structures of protein ligand-binding sites and application for predicting protein-drug interactions. *Proteins* **2008**, *72*, 367−381.
(21) 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.
(22) Brakoulias, A.; Jackson, R. Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching. *Proteins* **2004**, *56*, 250−260.
(23) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. Recognition of functional sites in protein structures. *J. Mol. Biol.* **2004**, *339*, 607−633.
(24) Milik, M.; Szalma, S.; Olszewski, K. Common Structural Cliques: a tool for protein structure and function analysis. *Protein Eng.* **2003**, *16*, 543−552.
(25) 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.
(26) Kuhn, D.; Weskamp, N.; Schmitt, S.; Huellermeier, 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.
(27) Kinoshita, K.; Furui, J.; Nakamura, H. Identification of protein functions from a molecular surface database, eF-site. *J. Struct. Funct. Genomics* **2002**, *2*, 9−22.
(28) Kinoshita, K.; Murakami, Y.; Nakamura, H. eF-seek: prediction of the functional sites of proteins by searching for similar electrostatic potential and molecular surface shape. *Nucleic Acids Res.* **2007**, *35*, W398−402.
(29) Najmanovich, R.; Kurbatova, N.; Thornton, J. Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics* **2008**, *24*, i105−11.
(30) Konc, J.; Janezic, D. ProBiS-2012: web server and web services for detection of structurally similar binding sites in proteins. *Nucleic Acids Res.* **2012**, *40*, W214−221.
(31) Yeturu, K.; Chandra, N. PocketMatch: a new algorithm to compare binding sites in protein structures. *BMC Bioinf.* **2008**, *9*, 543.
(32) Binkowski, T.; Joachimiak, A. Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites. *BMC Struct. Biol.* **2008**, *8*, 45.
(33) Yin, S.; Proctor, E.; Lugovskoy, A.; Dokholyan, N. Fast screening of protein surfaces using geometric invariant fingerprints. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 16622−16626.
(34) Xiong, B.; Wu, J.; Burk, D.; Xue, M.; Jiang, H.; Shen, J. BSSF: a fingerprint based ultrafast binding site similarity search and function analysis server. *BMC Bioinf.* **2010**, *11*, 47.
(35) Das, S.; Kokardekar, A.; Breneman, C. Rapid comparison of protein binding site surfaces with property encoded shape distributions. *J. Chem. Inf. Model.* **2009**, *49*, 2863−2872.
(36) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. 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.
(37) Schalon, C.; Surgand, J.-S.; Kellenberger, E.; Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins* **2008**, *71*, 1755−1778.
(38) Weill, N.; Rognan, D. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *J. Chem. Inf. Model.* **2010**, *50*, 123−135.
(39) Sael, L.; Chitale, M.; Kihara, D. Structure- and sequence-based function prediction for non-homologous proteins. *J. Struct. Funct. Genomics* **2012**, (epub).
(40) Merelli, I.; Cozzi, P.; D'Agostino, D.; Clematis, A.; Milanesi, L. Image-based surface matching algorithm oriented to structural biology. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2011**, *8*, 1004−1016.
(41) Pang, B.; Zhao, N.; Korkin, D.; Shyu, C.-R. Fast protein binding site comparisons using visual words representation. *Bioinformatics* **2012**, *28*, 1345−1352.
(42) Ito, J.-I.; Tabei, Y.; Shimizu, K.; Tomii, K.; Tsuda, K. PDB-scale analysis of known and putative ligand-binding sites with structural sketches. *Proteins* **2012**, *80*, 747−763.
(43) Aung, Z.; Tong, J. BSAlign: a rapid graph-based algorithm for detecting ligand-binding sites in protein structures. *International Conference on Genome Informatics*, Gold Coast, Australia, Dec 1-3, 2008; pp 65−76.
(44) Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes. *J. Chem. Inf. Model.* **2012**, *52*, 2287−2299.
(45) Schellhammer, I.; Rarey, M. TrixX: structure-based molecule indexing for large-scale virtual screening in sublinear time. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 223−238.
(46) Wu, K. FastBit: an efficient indexing technology for accelerating data-intensive science. *J. Phys.: Conf. Ser.* **2005**, *16*, 556−560.

(47) Volkamer, A.; Griewel, A.; Grombacher, T.; Rarey, M. Analyzing the topology of active sites: on the prediction of pockets and subpockets. *J. Chem. Inf. Model.* **2010**, *50*, 2041−2052.
(48) 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.
(49) Hubbard, T.; Murzin, A.; Brenner, S.; Chothia, C. SCOP: a structural classification of proteins database. *Nucleic Acids Res.* **1997**, *25*, 236−239.
(50) Urbaczek, S.; Kolodzik, A.; Fischer, J.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI: On the Almost Trivial Task of Reading Molecules from Different File formats. *J. Chem. Inf. Model.* **2011**, *51*, 3199−3207.
(51) Schlosser, J.; Rarey, M. Beyond the Virtual Screening Paradigm: Structure-Based Searching for New Lead Compounds. *J. Chem. Inf. Model.* **2009**, *49*, 800−809.
(52) Rarey, M.; Wefing, S.; Lengauer, T. Placement of medium-sized molecular fragments into active sites of proteins. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 41−54.
(53) 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.
(54) Laskowski, R.; Watson, J.; Thornton, J. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.* **2005**, *33*, W89−93.
(55) Gold, N.; Jackson, R. Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. *J. Mol. Biol.* **2006**, *355*, 1112−1124.
(56) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. Recognition of functional sites in protein structures. *J. Mol. Biol.* **2004**, *339*, 607−633.