JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Extraction of Protein Binding Pockets in Close Neighborhood of Bound Ligands Makes Comparisons Simple Due to Inherent Shape Similarity

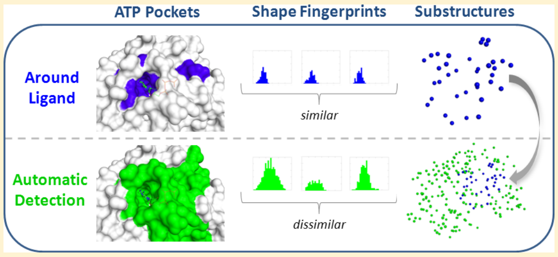
Timo Krotzky,[†] Thomas Rickmeyer,[†,‡] Thomas Fober,[§] and Gerhard Klebe*,[†,‡]

[†]Institute of Pharmaceutical Chemistry, University of Marburg, Marbacher Weg 6-10, 35032 Marburg, Germany

[‡]LOEWE-Zentrum für Synthetische Mikrobiologie (SYNMIKRO), Hans-Meerwein-Straße, 35043 Marburg, Germany

[§]Department of Mathematics and Computer Science, University of Marburg, Hans-Meerwein-Straße 6, 35032 Marburg, Germany

ABSTRACT: Methods for comparing protein binding sites are frequently validated on data sets of pockets that were obtained simply by extracting the protein area next to the bound ligands. With this strategy, any unoccupied pocket will remain unconsidered. Furthermore, a large amount of ligand-biased intrinsic shape information is predefined, inclining the subsequent comparisons as rather trivial even in data sets that hardly contain redundancies in sequence information. In this study, we present the results of a very simplistic and shape-biased comparison approach, which stress that unrestricted cavity extraction is essential to enable unexpected cross-reactivity predictions among proteins and function annotations of orphan proteins.



INTRODUCTION

Comparative analysis of protein binding pockets has emerged as an important field of interest in drug development. Sequence alignments or fold comparisons are often not appropriate to identify similar binding sites in unrelated proteins that originated from convergent evolution.[1,2] Nonetheless, they can still exhibit similar binding sites.[3] A comparison of pockets using geometric properties,[4,5] typed triangles, or physicochemical features in 3D space[6,7] has therefore become a popular strategy to unravel similarities of protein binding sites. A prominent alignment-free comparison method for protein−ligand binding sites is FuzCav.[8] Cavity fingerprints are defined for binding sites that store information about the presence of pharmacophoric feature triplets as lists of integers. These allow for an ultrafast comparison in the following step, attaining about 1000 calculations per second on a 3.4 GHz processor. Pocket-Surfer, which estimates global pocket similarity, and Patch-Surfer, which also detects local binding site similarities, have been introduced by Sael and co-workers.[9,10] The latter approach represents a pocket as a set of patches described by their shapes, electrostatic potential, degree of burial, and hydrophobicity. The comparisons are subsequently carried out by a bipartite matching procedure. Desaphy et al.[11] introduced the pocket description VolSite together with a tool for alignment and comparison called Shaper. The shape and physicochemical environment of a binding site are stored and then compared via Shaper, which aligns pockets by determination of an optimal surface overlap. In PocketAlign, shape descriptors from binding sites are derived that are enhanced by pharmacophoric features.[12] In the comparison step, matching pairings of the descriptors are combined into mappings that are subsequently evaluated using different metrics to achieve starting points for reasonable alignments. In order to accelerate binding-site comparisons, geometric hashing has become rather popular. Specific features of binding sites are transformed into a hash table, which is consulted in the following comparison step to obtain similarities to other cavities.[13−15] An extensive review of methods for the detection of similarity between protein binding sites can also be found in the contribution of Kellenberger et al.[16]

Usually the implementation of new algorithms for binding-site comparisons should accomplish one of the following three tasks. First, the prediction of putative off-target binding of drug molecules is highly desired, possibly providing an explanation for polypharmacology and adverse drug effects in early phases of a drug development project.[17] Second, the predictive functional annotation to orphan proteins is of high interest.[18] Third, the discovery of bioisosteric replacements[19,20] for specific ligand portions by retrieving similar binding sites that accommodate ligands with alternative scaffolds can support drug development. For the third application, it is undoubtedly reasonable to focus only on pockets (or subpockets) that have been extracted in the close neighborhood of a known ligand. Such pockets along with their bound ligands are successfully exploited, e.g., in KRIPO,[21] a method to identify valuable bioisosteric replacements of ligand portions recognized in specific subpockets. However, many binding-site comparison methods aimed at the other two goals have been developed, and they are subsequently validated by compiling test data sets

Received: July 10, 2014

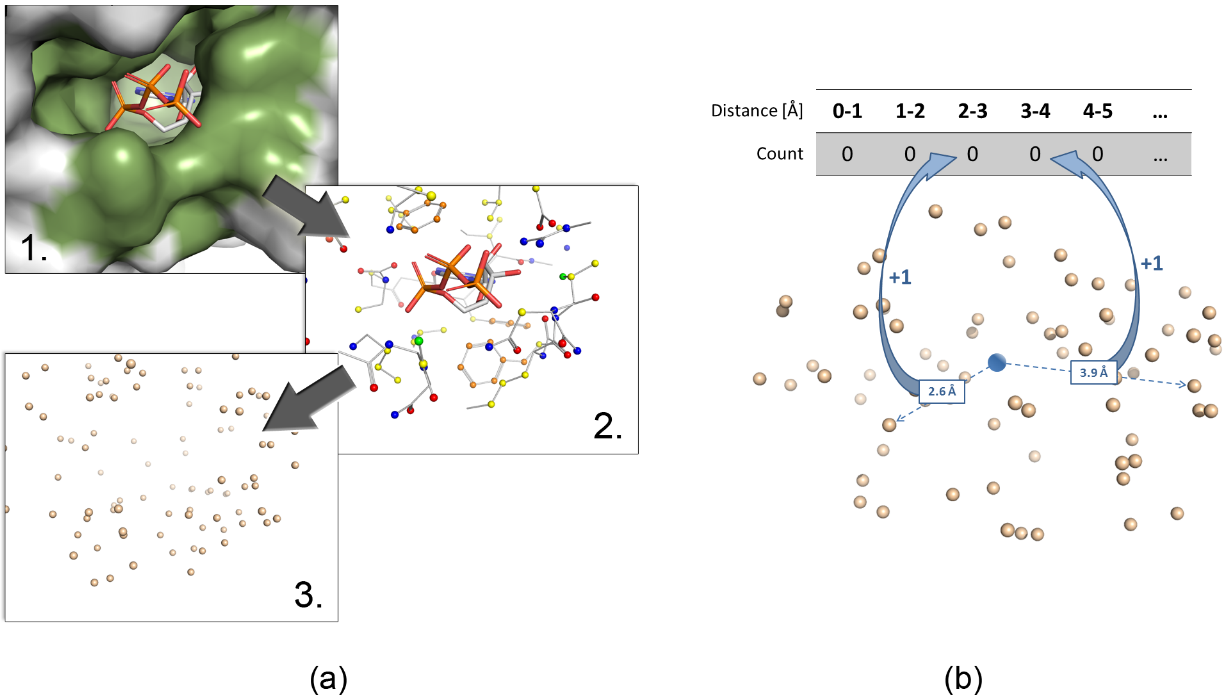


Figure 1. (a) Workflow resulting in a binding-site representation considering only shape information. First, all protein atoms approaching any ligand atom to ≤6 Å are defined as the binding site (green area). Next, all atoms are considered that agree with one of the following physicochemical properties: H-bond donor (blue), H-bond acceptor (red), H-bond doneptor (green), aromatic (orange), or hydrophobic (yellow) character (as classified by the program fconv). In the final step, any differentiation of physicochemical properties is discarded, revealing a pure pocket shape description. (b) Fingerprint generation to capture the spatial distribution of distances of interaction points with respect to their common centroid. All bin counts of the assigned fingerprint are initially set to zero. Next, distances determined between all interaction points (beige) and the centroid (blue) are assigned to the corresponding 1 Å sized bins. Any match to a bin augments the corresponding fingerprint element by one. In a very similar way, the spatial atom distribution of bound ligands is also analyzed. Here the atomic coordinates of the ligands are directly used as input.

of binding sites extracted as regions adjacent to bound ligands within a 4−6.5 Å sphere.[7,8,12,22,23] As a matter of fact, such data sets will lack binding sites originating from uncomplexed structures. As a consequence, any putative binding site that is previously unknown cannot be detected as a potential off-target for the drug molecule of interest unless the site of interest was incidentally occupied in the same or highly overlapping region by another ligand during crystallization. Furthermore, it is rather likely that such extracted pockets resemble inflated representations of the ligand shape, as only the region close to the accommodated ligand is considered. Thus, with respect to the prediction of drug side effects or functional annotations of orphan proteins, it may be beneficial to apply an automated cavity detection method that is independent of the presence or absence of a bound ligand. This will be of utmost importance when ligands that address different subpockets of proteins with large binding sites are studied. Several methods have been developed and successfully tested on putative binding cavities,[24−27] extracted independently of the presence or absence of a bound ligand. Hence, they will incorporate pockets of uncomplexed proteins. Nonetheless, any pocket data set extracted in the sole neighborhood of bound ligands will be biased toward intrinsic ligand shape information. This can strongly incline the obtained results, as an exaggerated weight is assigned to the ligand-based pocket shape rather than to the exposure of physicochemical properties available to recognize an arbitrary ligand.

Binkowski and Joachimiak[28] alluded to this fact that shape alone cannot be expected to serve as a comprehensive binding-site descriptor, a statement that matches with our assumption. In another study, Kahraman et al.[29] used spherical harmonics to describe binding-site shapes. Although they found that the success of retrieving similar pockets depends on the ligand shape, particularly if rigid host molecules are considered, the success rate declines once increasingly flexible ligands such as ATP, NAD, and FAD are subjected to the analysis. The latter ligands involve a large number of rotatable bonds that allow them to adopt multiple conformations of deviating shape (even when bound to members of the same superfamily).[30] This fact stimulated us to use these cofactor ligands in our evaluation. Moreover, it is suggested that the shapes of the hosting pockets vary more strongly than the accommodated ligands. The more it appears important to assess the extent to which predefined ligand shape affects the pocket representation and distorts subsequent comparisons if the pockets are extracted as close environments around bound ligands.

In the present study, we want to face the results of a pocket comparison using pockets extracted immediately around the ligands and pockets that result from an unbiased analysis of surface-exposed depressions on proteins. In the latter case, we use physicochemical properties to describe the pockets. Furthermore, we examine whether the geometries of ligands and pockets extracted around the bound ligands show high shape-based similarity. To perform these comparisons, we use a very simple geometric approach and describe the extracted binding sites (or bound ligands) in terms of spatial distance distributions of pocket-attributed interaction points (or ligand atoms). With this approach we do not intend to develop a new comparison algorithm but seek a fast method to compute similarity. A related method was suggested by Binkowski and

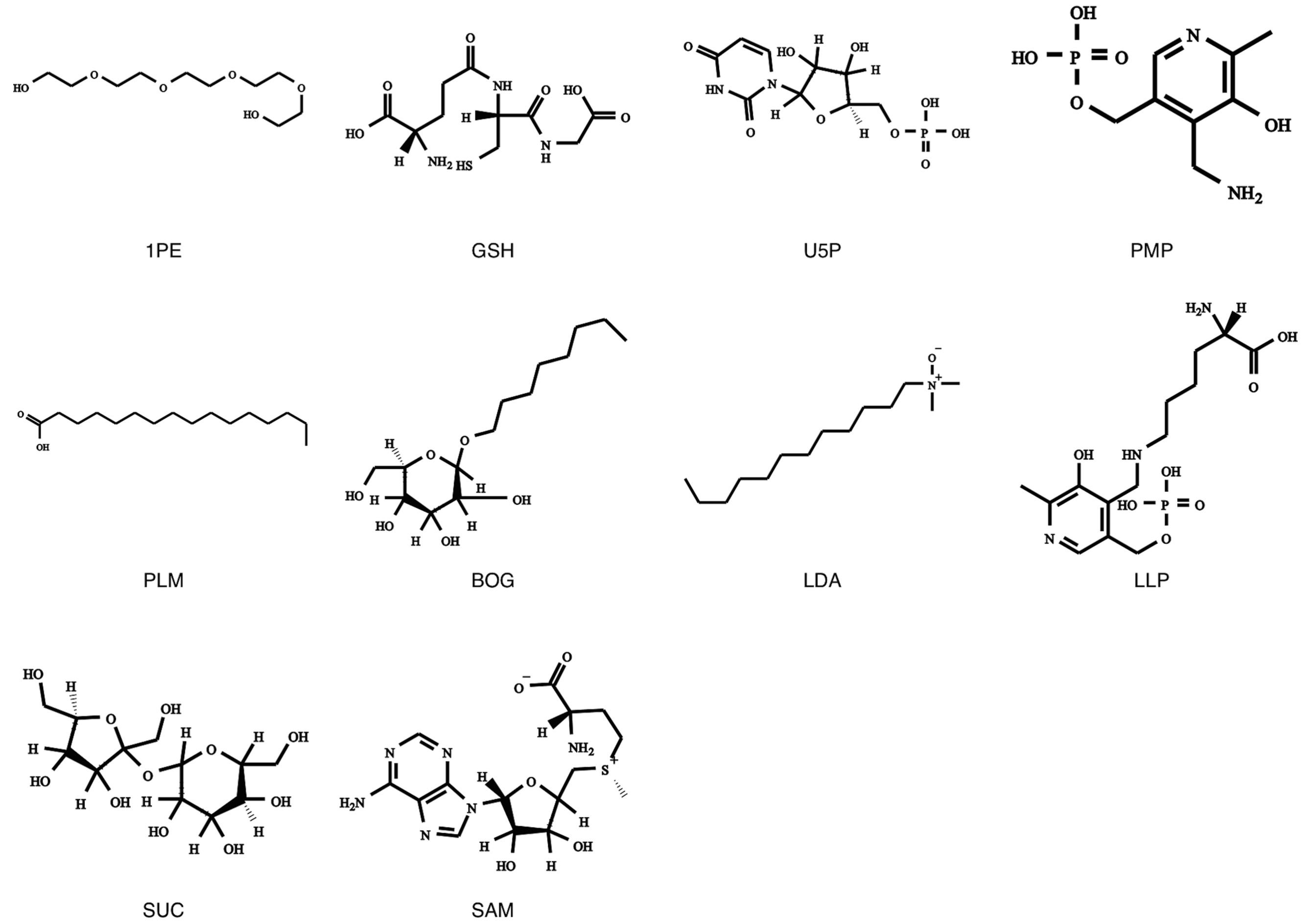


Figure 2. Ten ligands that were regarded in the comparative study of Hoffmann et al.[6] All of the structures are labeled with the respective PDB ligand identifier.

Joachimiak[28] as the first coarse-filtering method in a two-step comparison procedure. They determined the distances between all pairs of atoms defining the pocket surface to generate a probability distribution. In our comparison, we try an even simpler and thus faster approach by considering all distances to the pocket-describing points with respect to one common center point.

■ METHODS

Shape-Based Comparative Analysis. For the considered data sets, the pocket-describing points (or ligand atoms) were obtained following the protocol illustrated in Figure 1. Any protein atom approaching an atom of the bound ligand (to ≤6 Å) was supposed to be part of the binding pocket. In case of the ligands, we simply considered the composing atoms. Next, we used the program fconv[31] to perform an atom-type assignment to all of the thus-defined binding-site atoms. Subsequently, they were filtered in terms of represented physicochemical properties: only those atoms were considered that could be attributed to groups showing either H-bond donor, H-bond acceptor, H-bond "doneptor" (either donor or acceptor), aromatic, or hydrophobic character. In the following, this physicochemical information was neglected, and solely the spatial location of the retrieved interaction points was used to describe the pocket. Thus, this procedure provided binding-site representations solely reflecting shape and no physicochemical information. The comparison of two pockets was then accomplished by the following two-step procedure. First, a fingerprint was calculated for each pocket that captures the distances of all interaction points with respect to their geometric center (centroid). The obtained distances were represented histographically in bins of 1 Å size, and the occurrence frequencies of the found distance ranges were compiled (see Figure 1b). After fingerprints were assigned to all pockets of the data set, the comparative distance between two pockets was calculated by using the Jensen–Shannon divergence. When two fingerprints varied in length, the shorter one was extended by adding unoccupied bins. For the evaluation of the ligands we proceeded similarly, only taking the composing atoms directly.

Data Sets. A data set of cofactor complexes that had already been compiled in an earlier study[32] was evaluated. It comprised 420 ATP, 380 NAD, and 432 FAD pockets, considering each PDB entry only once. A second data set, suggested by Hoffmann et al.,[6] that comprised pockets accommodating ligands of similar size was assembled. This set considered 100 nonredundant proteins with pockets hosting one of the 10 ligands of approximately equal size shown in Figure 2. A third data set was extracted from the PDB (83 000 entries in the used release) by using LIGSITE to find putative binding pockets. A database of 451 100 pockets was compiled. All of the pocket atoms were annotated according to fconv atom types.

RESULTS AND DISCUSSION

Cofactor Binding Pockets. As a first example, we evaluated the set of cofactor binding pockets accommodating either ATP, FAD, or NAD (NADH and NAD$^+$) present with deviating conformations. In this experiment, we used the k-nearest-neighbor (kNN) method for classification, which has also been applied previously with success.[32−34] We calculated an all-against-all scoring matrix that was subsequently used as input for a 10-fold cross-validation with a kNN classifier using k = 1. Unexpectedly, the two-class classification experiment of the ATP and NAD data sets revealed an excellent rating (96.4 ± 2.0%). Since we considered only unlabeled surface points as pocket descriptors and the analyzed cofactors are of rather different size, it might well be that simply the number of interaction points defining the binding pocket is already responsible for the impressive discrimination. However, using only the number of points for the comparisons led to a poor rate of only 63.7 ± 5.3%. Also, normalization of the fingerprint (FP) distributions (calibrating the area under each curve to 1) to exclude any influence of the total number of points per pocket led to hardly any change in the above-mentioned classification rate (95.5 ± 1.7%).

As next experiment, we incorporated FAD pockets. FAD and NAD evidently vary less in size than ATP and NAD. Although the resulting heat map of the distance scores (Figure 3) suggests that FAD and NAD pockets are less well discriminated, we still obtained a convincing classification of 94.3 ± 1.5% in this three-class experiment. Furthermore, a success rate of 96.2 ± 2.1% was achieved when only the FAD and NAD pockets were considered in the classification. To estimate the robustness of the results obtained in this three-class experiment, we also evaluated the scoring matrix by another method that is closely related to the 10-fold cross-validation. We applied the k-leave-one-out cross-validation and varied the number of nearest neighbors k in the range from 1 to 15. As shown in Table 1, the obtained rates did not deteriorate rapidly and all exceeded 90%, which indicates either robustness of our method or simplicity of the data set used.

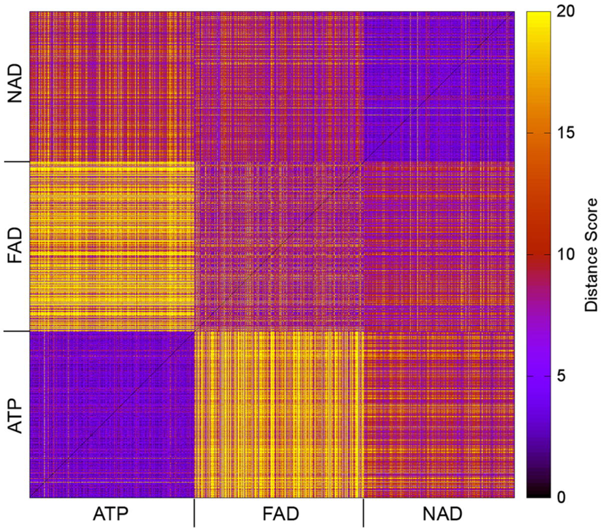


Figure 3. Heat map of the scoring matrix resulting from the classification experiment using ATP, FAD, and NAD pockets. The areas of correctly assigned ATP (lower left corner) and NAD pockets (upper right corner) generally display rather low distance scores among each other, as indicated by the bluish coloring. They are obviously well-separated from the other pockets. The FAD pockets (center) appear to be more similar to the NAD pockets; however, they can also be correctly classified with a success rate of over 96% in a two-class experiment that regards FAD and NAD pockets only. The black main diagonal from bottom left to top right indicates the distance values of zero in the cases of a self-comparison.

Table 1. Classification Results of the Shape FP When the Scoring Matrix Was Evaluated Using a k-Leave-One-Out Cross-Validation with the Number of Nearest Neighbors k in the Range 1 to 15

k	correct classifications [%]
1	94.3
3	93.8
5	93.5
7	92.9
9	92.7
11	91.6
13	90.5
15	90.1

Data Set of Equal-Sized Ligands. Hoffmann et al.[6] suggested as a real challenge for a binding-site comparison approach to discriminate pockets accommodating ligands of similar size. To evaluate their comparison method, they compiled a benchmark data set of 100 nonredundant proteins with pockets hosting one of the 10 ligands of approximately equal size shown in Figure 2.

For each ligand, 10 pockets were extracted by defining the protein atoms within a distance of up to 5.3 Å. The authors hence ended up with a 10-class data set in which each class consisted of 10 pockets, which they called a *homogeneous data set*. In their study, a total of nine pocket comparison methods were tested using this validation set, and the classification rates were analyzed using receiver operating characteristic (ROC) curves. For each method, 100 ROC curves were calculated by performing comparisons of each single pocket against the 99 remaining structures. Finally, the average areas under the curve (AUCs) of all ROC curves that correspond to a single method were calculated. An AUC of 0.5 denotes a method that detects hits (pockets of the same class) equally as well as a random assignment. On the contrary, a value of 1.0 would be obtained for a method that assigns the highest similarity scores to the nine remaining pockets of its class and achieves perfect classification. The results reported in the above-mentioned study revealed average AUCs between 0.58 and 0.77. We performed a similar analysis of this data sample using our shape-based fingerprint descriptors. Our approach performed surprisingly well, reaching an average AUC of 0.66 (Figure 4).

Comparison with Unbiased Surface-Exposed Pockets. All of the reported examples demonstrate convincingly well that the success of a comparative binding-site analysis is intrinsically given if only shape complementarity next to the bound ligand is used to describe the considered pockets. We therefore applied a ligand-unbiased cavity detection algorithm to extract pockets from protein structures in order to analyze again our ATP, NAD, and FAD data sets. A variety of structure-based methods has emerged to accomplish the task of finding putative binding pockets on the protein surface. They can be divided into either geometry-based methods, such as PASS,[36] SURFNET,[37] CAST,[38] APROPOS,[39] SiteFinder,[40] fpocket,[26] and PocketPicker,[41] or energy-based approaches, such as PocketFinder[42]

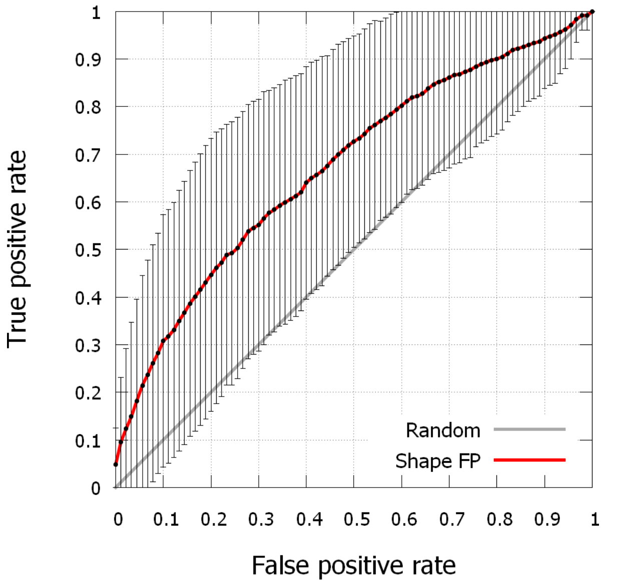


Figure 4. Average ROC curve of the shape FP (red) when applied on the homogeneous data set of Hoffmann et al.[6] Random performance is indicated by the gray diagonal from bottom left to top right. The plot represents the average of all 100 curves that were obtained and exhibits an average AUC of 0.66 ± 0.16. In addition, we display the standard deviations for the individual data points, shown as black error bars.

and SuperStar.[43] Comprehensive overviews of the current binding-site prediction methods are provided, e.g., by Perót et al.[44] and Leis et al.[45]

Here we applied LIGSITE,[46] a grid-based (and thus also geometry-based) method to detect depressions on protein surfaces. They optionally comprise hosted ligands but usually they extend beyond the actual contact area of the ligand with the protein. This strategy appears to be a less biased protocol to define a binding pocket. To apply LIGSITE, the protein is placed onto a regular grid with a spacing of 0.5 Å. Each grid intersection point is evaluated with respect to its degree of burial. A cluster of at least 320 adjacent buried grid points is then defined as a putative binding site. A detailed description can be found in the original publication.[46] All atoms flanking the thus-detected cavities are potentially capable of binding a ligand. The extracted pockets were likewise classified in terms of atom types using fconv and subsequently used to construct three new data sets: $ATP_{LIGSITE}$, $NAD_{LIGSITE}$, and $FAD_{LIGSITE}$. The LIGSITE pockets are different in shape and generally larger than the ligand-based pockets (on average, the number of extracted interaction points is increased by a factor of 2.5), which indicates additional areas competent to recognize a ligand beyond the area actually addressed by the regarded cofactors (Figure 5).

In contrast to the results obtained with the ligand-shape-based pockets of ATP, NAD, and FAD, the success rates of classifying by use of the shape-based fingerprints decreased substantially from 94.3% to 61.8%. Regarding the actual atom-type assignment for the pocket representation (see the workflow in Figure 1a) enabled us to apply a previously presented approach for the comparison of protein binding sites, the so-called labeled point cloud superposition method (LPCS).[34] The approach suggested by Fober et al. was applied using the parameter setting recommended by the authors. Applying LPCS in the current case led to an accuracy of 97.7 ± 1.3% when it was submitted to the ligand-based pockets. Thus, the success rates agreed well with the results obtained with our simple fingerprint approach. However, in contrast to the fingerprint approach, LPCS was still able to achieve a comparable accuracy of 93.1 ± 2.8% when it was applied to the larger pockets extracted by LIGSITE (Figure 6a).

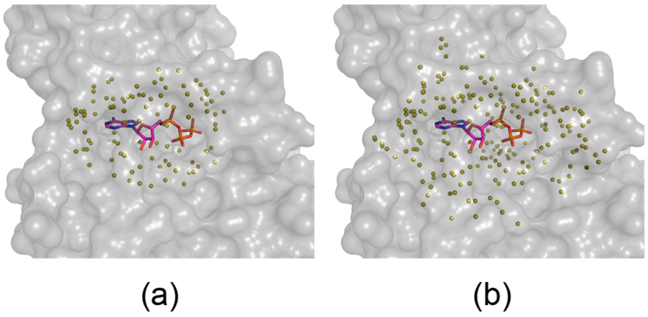


Figure 5. Example of an ATP binding pocket (PDB code 1B38) that is defined by either (a) extracting the area of 6 Å around the ligand or (b) the automated cavity detection procedure LIGSITE. The automatically detected cavities are in general much larger than the pockets solely defined by considering the bound ligand.

To set up a more challenging task with respect to conformational and structural diversity, we culled the three-class data set to remove sequential redundancies. Therefore, the protein sequence culling server PISCES of the Dunbrack lab[35] was employed, which kept only PDB structures that agreed to the following conditions: the sequence identity does not exceed 25%, the method of structure determination is X-ray crystallography with an *R* factor of ≤0.3, and the resolution is 3 Å or better. As a result, 268 elements (135 ATP pockets, 73 NAD pockets, and 60 FAD pockets) remained in the data set. Table 2 provides an overview of the contained structures.

When the culled data set was used, the actual problem became even more apparent. In the case of the ligand-based pockets extracted 6 Å around the bound molecules, both LPCS and the shape FP still achieved satisfactory results (Figure 6b). Although the success rate of FP was worse than that of LPCS, the difference was not significant (80.4 ± 4.6% vs 88.5 ± 7.0%). When the automatically detected LIGSITE pockets were used, however, the resulting rates decreased substantially, by 34% in the case of FP to reveal 44.6 ± 8.9% correct classifications, which hardly deviated from a random assignment (37.8% in the present example considering the unequal population of the subsets). The LPCS approach was still clearly better, attaining a correct classification rate of 66.5 ± 7.2% even though it became obvious that culling the data set increased the complexity of the problem of classifying the automatically detected cavities. Thus, this experiment shows once again that the degree of complexity is highly diminished when ligand-based pockets are used for the comparisons instead of automatically detected ones.

Ligand Atom Distributions versus Ligand-Shape-Based Pockets. The minor loss in accuracy indicates that LPCS is obviously quite robust and independent of the actual size and shape of the pocket whereas the fingerprint approach is strongly affected. Supposedly, the consideration of pockets extracted in the close neighborhood of the ligands provides a remarkable advantage in cavity comparison. In order to examine the extent to which the latter pockets resemble just an inflated representation of the ligand shape, we performed a comparison of the data sets in which the ligands were used instead of the pockets. Therefore, the ligand atoms were processed in the same way as the pocket atoms beforehand.

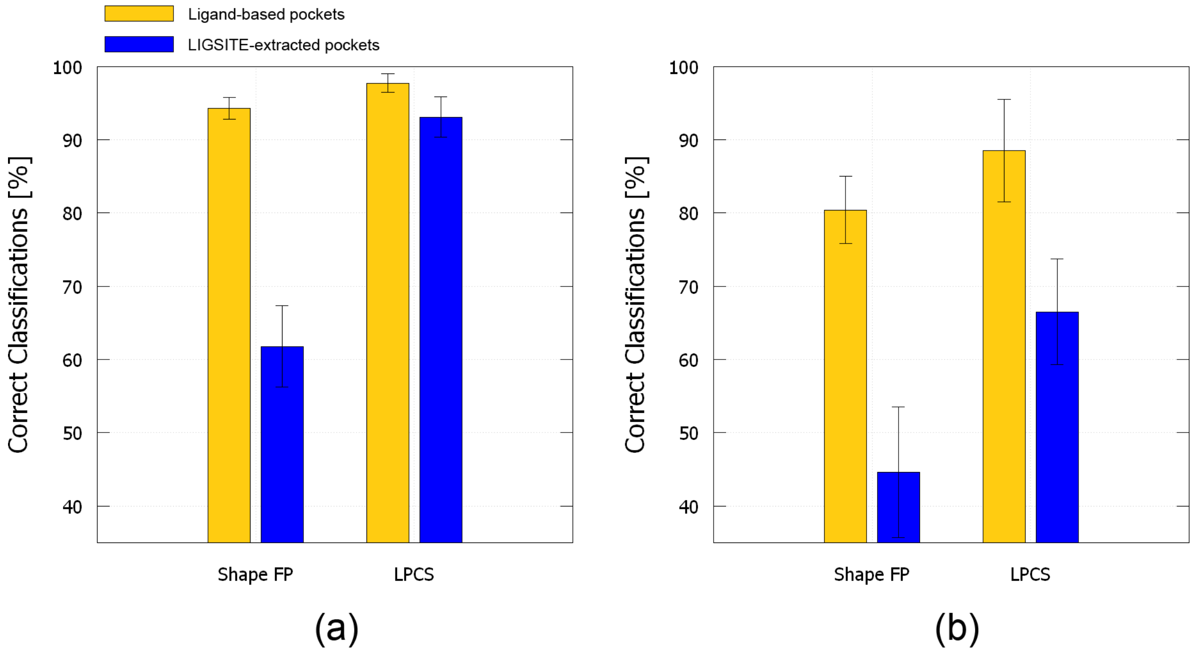


Figure 6. Comparison of the ligand-shape-based fingerprint and the LPCS approach when subjected to a classification experiment using binding pockets of ATP-, NAD-, and FAD-binding proteins. The binding pockets were defined by either extracting protein atoms in the close neighborhood (≤6 Å) about the bound ligand (yellow) or applying the LIGSITE algorithm to detect cavities in an unbiased way as depressions on the protein surface (blue). The complete data set of 1232 pockets was used to obtain the results in (a). Here the accuracy of the fingerprint approach (left), which is heavily biased by the actual shape information complementarity to the bound ligand, drops by more than 30% when applied to the LIGSITE-extracted pockets. In addition, the standard deviation increases strongly by a factor of almost 4. However, the LPCS results (right) exhibit a deterioration of only 4.6% and the standard deviation increases minimally by a factor of 2 among the data sets of differently extracted pockets. In (b) the culled data set containing 268 pockets was used. In this case, both approaches still show satisfactory success rates when ligand-based pockets are used (yellow). However, the classification rates decrease dramatically when the LIGSITE pockets are used (blue), especially if the shape FP is applied. It is no longer able to reach a success rate better than a random classification.

Table 2. Overview of the Culled Data Set of ATP, NAD, and FAD Pockets, Giving the Considered PDB Entries by Their Reference Codes

ATP-binding pockets	1A0I, 1KVK, 1YFR, 2E89, 2QUI, 3A8T, 3GAH, 3LL3, 3RGL, 1BCP, 1MAU, 1YP3, 2FXU, 2R7L, 3AMT, 3GNI, 3LL5, 3RK1, 1CSN, 1MB9, 1Z0S, 2GBL, 2R9V, 3ATT, 3GQK, 3LMI, 3RTG, 1DY3, 1NSF, 1ZAO, 2HIX, 2VHQ, 3C16, 3H39, 3LRR, 3SEZ, 1E2Q, 1OBD, 1ZFN, 2HMU, 2VT3, 3C9R, 3H8V, 3LSS, 3SL2, 1E8X, 1PK8, 2A5Y, 2HS0, 2XCW, 3CQD, 3HAV, 3MHY, 3T54, 1EE1, 1QHX, 2A84, 2IVP, 2Y27, 3CRC, 3HY2, 3NA3, 3TLX, 1GN8, 1R0X, 2AQX, 2IYW, 2YCH, 3DKC, 3I7V, 3NEM, 3TUT, 1GTR, 1RDQ, 2ARU, 2O0H, 2YJ4, 3DNT, 3IBQ, 3OS3, 3V2U, 1GZ4, 1S9J, 2BEK, 2OGX, 2YW2, 3E7E, 3IE7, 3OVB, 3VH4, 1J09, 1SVM, 2BUP, 2OLR, 2YWW, 3EHG, 3IKH, 3Q60, 3ZS7, 1J7K, 1UF9, 2C8V, 2Q0D, 2Z08, 3EPS, 3INN, 3QUO, 4A2A, 1KJ9, 1VJD, 2C96, 2Q7G, 2Z1U, 3ETH, 3IQ0, 3QXC, 4AFF, 1KMN, 1WKL, 2DTO, 2QKM, 2ZAN, 3FKQ, 3LEV, 3R1R, 4DW1, 1KP2, 1XDN, 2E5Y, 2QRD, 2ZSF, 3G59, 3LFZ, 3R5X, 4ED4
NAD-binding pockets	1M8G, 1SBY, 1WPQ, 2BJK, 2IXA, 3AJR, 3I9K, 3P2O, 3VDQ, 1MEW, 1SG6, 1X15, 2D37, 2IZZ, 3B6J, 3JSA, 3PJF, 1NVM, 1T2D, 1X7D, 2D4V, 2JHF, 3BTS, 3JYO, 3Q3C, 1OG3, 1UP7, 1Z0Z, 2DT5, 2NSD, 3C7D, 3LN3, 3Q9O, 1PJS, 1UWK, 1Z45, 2DVM, 2O2S, 3CEA, 3M2T, 3QVX, 1PL8, 1UXG, 1ZJZ, 2EKP, 2O2Z, 3CIN, 3NRC, 3RF7, 1RKX, 1V9L, 2A5F, 2G76, 2QG4, 3F3S, 3NT2, 3RIY, 1RLZ, 1VBI, 2A9K, 2G82, 2VUT, 3GGG, 3ORF, 3SYT, 1S20, 1VM6, 2B69, 2I2F, 2YVF, 3H9U, 3OX4, 3UQ8
FAD-binding pockets	1C0P, 1RM6, 2C12, 2GPJ, 2QA1, 3AH5, 3G5S, 3NLC, 3TEM, 1EP2, 1RYI, 2CUL, 2GQT, 2QCU, 3AXB, 3G6K, 3NYC, 3ZXS, 1GVH, 1TEZ, 2CZ8, 2GQW, 2QDX, 3D1C, 3GWL, 3P0K, 4DNA, 1JR8, 1U8V, 2DJI, 2HQ9, 2UXW, 3DJL, 3GWN, 3PND, 4FEH, 1JU2, 1UMK, 2ED4, 2IJG, 2V5Z, 3E2Q, 3LLI, 3QJ4, 1N4W, 2AQJ, 2FG9, 2OLN, 2XRY, 3F8D, 3LO8, 3QVP, 1R2J, 2B9W, 2GJ3, 2PGN, 2YYJ, 3FST, 3M31, 3RP8

They were typed using fconv, and subsequently the fingerprints were calculated with the help of the centroid to facilitate a comparison (cf. Figure 1b). In this case, a correct classification rate of 98.6 ± 1.0% was obtained. This is not surprising, as the spatial arrangement of ligand atoms is in general less complex than the arrangement of binding-site atoms.[28] However, generation of the scoring matrix enabled us to compare this matrix to the scoring matrices obtained for the pocket comparisons based on either ligand-shape-based pockets or surface-exposed pockets (LIGSITE). We calculated the correlation between the ligand scoring matrix and the ligand-shape-based pocket scoring matrix. Alternatively, we faced the ligand scoring matrix to the LIGSITE pocket scoring matrix. To calculate a correlation of two matrices, the Spearman's rank correlation coefficients of all matching pairs of rows were determined and, finally, normalized by the total number of row pairs. A high positive correlation of 0.68 between the ligand matrix and the matrix of ligand-shape-based pockets was obtained, which underscores the general similarity of ligand shape and pocket shape in this case. On the contrary, there was hardly any correlation between the ligand matrix and the matrix of LIGSITE pockets (correlation coefficient of 0.09), which demonstrates the minor relationship of ligand shapes and the shapes of automatically detected surface-exposed pockets.

Ligand-Shape-Based Pockets versus Unbiased Surface-Exposed Pockets. The above-described examples show that LPCS is obviously still able to extract the relevant information required to match common substructures competent to bind the same ligand when unbiased surface-exposed depressions on proteins are considered in the analysis. As mentioned, the latter approach usually extracts larger pockets as additional areas in the environment, not addressed by the bound ligand, can still provide binding epitopes capable of recognizing another ligand. This fact may be responsible for undesired cross-reactivity. To assess whether LPCS outperforms the ligand-shape-based pocket fingerprints, we accomplished another experiment. The PDB contains a significant number of crystal structures determined with the same protein where the bound ligands do not bind to overlapping binding epitopes. This situation can increasingly be expected for fragment binding. Figure 7 displays the crystal structures of thrombin with benzamidine as an S_1-accommodated ligand and a second fragment exclusively binding to the S_2−S_4 pocket.[47] Both ligands address hardly any shared binding region, and accordingly, an approach that extracts binding pockets solely in the close neighborhood of bound ligands will likely fail to provide a similarity signature for the two thrombin pockets.

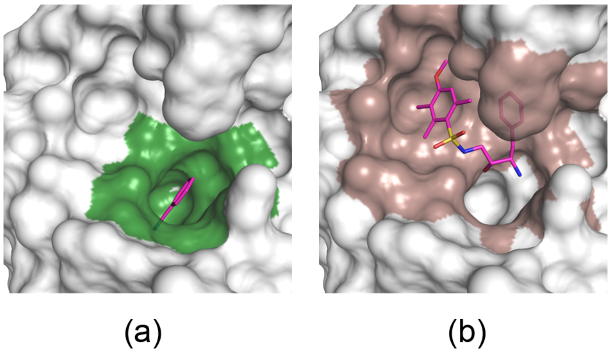


Figure 7. (a) Example of a pocket extracted 6 Å around benzamidine (PDB code 1DWB) in the green S_1 pocket. (b) Thrombin shown from the same angle of view with another ligand that accommodates a distinct region of the binding pocket (PDB code 2C93, ligand identifier C4M), in the pale S_2−S_4 pocket. Hardly any overlap of the two pockets is given if the pockets are extracted next to the bound ligands.

We applied LIGSITE to extract putative binding pockets from the PDB and compiled a database of 450 100 pockets. Next, three thrombin query pockets were defined and subjected to this pocket database. First, the structure 3UWJ was used, and all of the pocket atoms falling within 6 Å next to the accommodated ligand, *N*-(benzylsulfonyl)-D-leucyl-*N*-(4-carbamimidoylbenzyl)-L-prolinamide (ligand identifier TIF), were extracted. As this ligand fills the pocket quite extensively, the entire thrombin active site was captured. Second, only the S_1 subpocket of 1DWB was retrieved by extracting all of the atoms in a range of 6 Å around benzamidine. Third, the S_2−S_4 pockets of PDB entry 2C93 were extracted using the bound fragment, *N*-[(2*R*,3*S*)-3-amino-2-hydroxy-4-phenylbutyl]-4-methoxy-2,3,6-trimethylbenzenesulfonamide (ligand identifier C4M) (see Figure 7). Subsequently, we performed retrieval experiments based on these three query pockets in order to detect other thrombin cavities in the database. To detect the total number of thrombin entries in our database, we searched for a match with the EC number 3.4.21.5 (thrombin) and the presence of Asp189, a key residue in S_1 for substrate recognition, to guarantee that only the catalytic pockets were captured. In total, we detected 430 thrombin pockets, which were used as a reference to evaluate our subsequent retrieval results. Figure 8 displays the resulting ROC curves obtained using the three query pockets and either the LPCS or ligand-shape-based fingerprint approach.

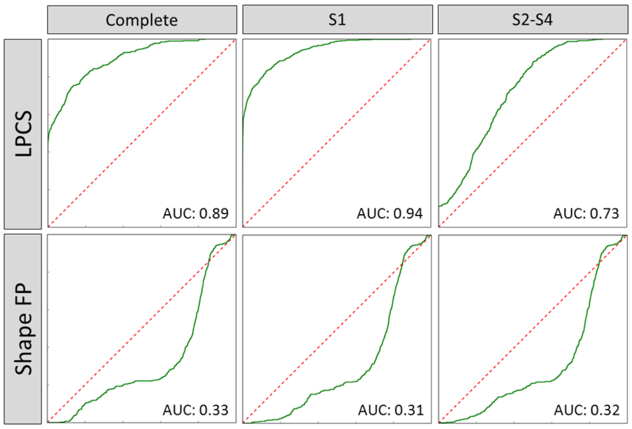


Figure 8. (a) ROC curves illustrating the retrieval rates of database screenings based on the LPCS (first row) or shape fingerprint (second row) approach, respectively. The first column depicts the ROC curves using the complete binding pocket, the second the S_1 subpocket, and the third the S_2−S_4 subpocket as a query. The dashed red lines indicate the random retrieval rate (AUC = 0.5).

As mentioned above, ROC curves are widely used to validate retrieval and enrichment results. True-positive retrieval rates (*y* axis) are plotted against false-positive ones (*x* axis), and the AUC indicates the success of the method. As shown in Figure 8, LPCS achieved very convincing retrieval success (AUCs of 0.94 and 0.89, respectively) and remarkable early enrichment slopes when the complete and S_1 pockets were subjected as queries. Using the S_2−S_4 pocket as a query resulted in a somewhat worse ROC curve, though still much better than random (AUC = 0.73). The plots based on the ligand-shape-based pocket fingerprint analysis show the unsatisfactory performance of this method when applied to compare any of the query pockets against the database of surface-exposed cavities. All of the ROC curves exhibit AUCs worse than random retrieval. This result demonstrates that the latter approach is much less robust than the LPCS approach with respect to substructure detection.

CONCLUSION

The presented study has uncovered the inherent and highly biased shape information on binding sites when they are extracted in the close neighborhood of the bound ligands. Simply considering the coarse distribution of potential interaction points in such a ligand-shape-based pocket reveals retrieval success rates of more than 95% in our classification experiments even when ligands of deviating size and conformations are analyzed. Even when high redundancies in protein sequence are eliminated from the data set (so-called "culling"), this simple comparison method still achieves high success rates of around 80%. Any information about the distribution of physicochemical properties across the pockets was neglected, and a simple ligand shape-determined fingerprint, assigned to each pocket, was sufficient to accomplish a successful comparison with minimal computational effort (more than 500 000 comparisons per second on a customary computer). We could show that the sole pocket size expressed by the number of interaction points is not discriminative. Thus, the information that enables classifications is stored in the spatial distribution pattern of the interaction points next to the ligands. This pattern is likewise determined as a kind of inflated ligand shape, as the spatial positions of bound ligands were used to extract a binding pocket. The fact that these pockets can be regarded as size-inflated ligands is demonstrated by a significant correlation of the distance distributions derived from the ligand atoms and the pocket interaction points defined in the close neighborhood of the bound ligands.

An unbiased approach seeking for a cavity comparison of surface-exposed depressions on proteins does not make use of ligand information. Thus, pockets found in uncomplexed proteins will also be extracted and analyzed. The same holds for pockets extracted from the same reference protein that are accommodated by ligands addressing nonoverlapping epitopes of the binding pocket. This strongly argues that one should only analyze and compare automatically detected surface-exposed cavities that are extracted unbiased from any ligand information. Only then can surprising results with respect to putative cross-reactivity and functional annotation of orphan proteins be expected. Most likely, comparative methods seeking similarities between automatically extracted cavities will require more computational effort, since the pockets will be larger and similarity may be detected in terms of subpockets (detecting a subset in pocket A that is also present in pocket B). As a major advantage, putative binding sites of uncomplexed or spatially differently accommodated proteins can also be studied, which considerably expands the pocket space. This is of utmost importance in predicting unexpected cross-reactivity of newly developed drugs and will only be of relevance if the evaluation algorithm still detects similarity in subpockets. As convincingly shown in this study, these criteria are matched by the LPCS approach, whereas the ligand-shape-based fingerprint fails at this challenge.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: klebe@staff.uni-marburg.de.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors gratefully acknowledge financial support by the Center for Synthetic Microbiology, Marburg (SYNMIKRO).

■ ABBREVIATIONS

PDB, Protein Data Bank; ATP, adenosine triphosphate; NAD, nicotinamide adenine dinucleotide; kNN, *k*-nearest-neighbor; FAD, flavin adenine dinucleotide; LPCS, labeled point cloud superposition; ROC, receiver operator characteristic; AUC, area under the curve.

■ REFERENCES

(1) Lee, D.; Redfern, O.; Orengo, C. Predicting Protein Function from Sequence and Structure. *Nat. Rev. Mol. Cell Biol.* **2007**, *12*, 995−1005.

(2) Rost, B. Twilight zone of protein sequence alignments. *Protein Eng.* **1999**, *2*, 85−94.

(3) Chalk, A. J.; Worth, C. L.; Overington, J. P.; Chan, A. W. E. PDBLIG: Classification of Small Molecular Protein Binding in the Protein Data Bank. *J. Med. Chem.* **2004**, *47*, 3807−3816.

(4) Russell, R. B. Detection of protein three-dimensional side-chain patterns: New examples of convergent evolution. *J. Mol. Biol.* **1998**, *279*, 1211−1227.

(5) Das, S.; Kokardekar, A.; Breneman, C. M. Rapid Comparison of Protein Binding Site Surfaces with Property Encoded Shape Distributions. *J. Chem. Inf. Model.* **2009**, *49*, 2863−2872.

(6) 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*, No. 99.

(7) 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.

(8) Weill, N.; Rognan, D. Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein−Ligand Binding Sites. *J. Chem. Inf. Model.* **2010**, *50*, 123−135.

(9) Chikhi, R.; Sael, L.; Kihara, D. Real-Time Ligand Binding Pocket Database Search Using Local Surface Descriptors. *Proteins* **2010**, *78*, 2007−2028.

(10) Sael, L.; Kihara, D. Detecting Local Ligand-Binding Site Similarity in Nonhomologous Proteins by Surface Patch Comparison. *Proteins* **2012**, *80*, 1177−1195.

(11) 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.

(12) Yeturu, K.; Chandra, N. PocketMatch: A New Algorithm To Compare Binding Sites in Protein Structures. *BMC Bioinf.* **2008**, *9*, No. 543.

(13) Brakoulias, A.; Jackson, R. M. 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.

(14) Bachar, O.; Fischer, D.; Nussinov, R.; Wolfson, H. A Computer Vision Based Technique for 3-D Sequence-Independent Structural Comparison of Proteins. *Protein Eng.* **1993**, *6*, 279−287.

(15) Pennec, X.; Ayache, N. A Geometric Algorithm To Find Small but Highly Similar 3D Substructures in Proteins. *Bioinformatics* **1998**, *14*, 516−522.

(16) 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.

(17) Jalencas, X.; Mestres, J. Identification of Similar Binding Sites To Detect Distant Polypharmacology. *Mol. Inf.* **2013**, *32*, 976−990.

(18) 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.

(19) Wagener, M.; Lommerse, J. P. The Quest for Bioisosteric Replacements. *J. Chem. Inf. Model.* **2006**, *46*, 677−685.

(20) 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.

(21) Wood, D. J.; de Vlieg, J.; Wagener, M.; Ritschel, T. Pharmacophore Fingerprint-Based Approach to Binding Site Subpocket Similarity and its Application to Bioisostere Replacement. *J. Chem. Inf. Model.* **2012**, *52*, 2031−2043.

(22) Feldman, H. J.; Labute, P. Pocket Similarity: Are α Carbons Enough? *J. Chem. Inf. Model.* **2010**, *50*, 1466−1475.

(23) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. SiteEngines: Recognition and Comparison of Binding Sites and Protein−Protein Interfaces. *Nucleic Acids Res.* **2005**, *33*, W337−W341.

(24) Volkamer, A.; Kuhn, D.; Grombacher, T.; Rippmann, F.; Rarey, M. Combining Global and Local Measures for Structure-Based Druggability Predictions. *J. Chem. Inf. Model.* **2012**, *52*, 360−372.

(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) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinf.* **2009**, *10*, No. 168.

(27) Xiong, B.; Wu, J.; Burk, D. L.; Xue, M.; Jiang, H.; Shen, J. BSSF: A Fingerprint Based Ultrafast Binding Site Similarity Search and Function Analysis Server. *BMC Bioinf.* **2010**, *11*, No. 47.

(28) Binkowski, T. A.; Joachimiak, A. Protein Functional Surfaces: Global Shape Matching and Local Spatial Alignments of Ligand Binding Sites. *BMC Struct. Biol.* **2008**, *8*, No. 45.

(29) Kahraman, A.; Morris, R. J.; Laskowski, R. A.; Thornton, J. M. Shape Variation in Protein Binding Pockets and Their Ligands. *J. Mol. Biol.* **2007**, *368*, 283−301.

(30) Stockwell, G. R.; Thornton, J. M. Conformational Diversity of Ligands Bound to Proteins. *J. Mol. Biol.* **2006**, *356*, 928−944.

(31) Neudert, G.; Klebe, G. fconv: Format Conversion, Manipulation and Feature Computation of Molecular Data. *Bioinformatics* **2011**, *27*, 1021−1022.

(32) Krotzky, T.; Fober, T.; Hüllermeier, E.; Klebe, G. Extended Graph-Based Models for Enhanced Similarity Search in Cavbase. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2014**, *11*, 878−890.

(33) Fober, T.; Mernberger, M.; Klebe, G.; Hüllermeier, E. Evolutionary Construction of Multiple Graph Alignments for the Structural Analysis of Biomolecules. *Oxford Bioinf.* **2009**, *25*, 2110−2117.

(34) Fober, T.; Glinca, S.; Klebe, G.; Hüllermeier, E. Superposition and Alignment of Labeled Point Clouds. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2011**, *8*, 1653−1666.

(35) Wang, G.; Dunbrack, R. L., Jr. PISCES: A Protein Sequence Culling Server. *Bioinformatics* **2003**, *19*, 1589−1591.

(36) Brady, G. P.; Stouten, P. F. W. Fast Prediction and Visualization of Protein Binding Pockets with PASS. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 383−401.

(37) Laskowski, R. A. SURFNET: A Program for Visualizing Molecular Surfaces, Cavities, and Intermolecular Interactions. *J. Mol. Graphics* **1995**, *13*, 323−330.

(38) Binkowski, T. A.; Adamian, L.; Liang, J. Inferring Functional Relationships of Proteins from Local Sequence and Spatial Surface Patterns. *J. Mol. Biol.* **2003**, *332*, 505−526.

(39) Peters, K. P.; Fauck, J.; Frömmel, C. The Automatic Search for Ligand Binding Sites in Proteins of Known Three-Dimensional Structure Using Only Geometric Criteria. *J. Mol. Biol.* **1996**, *256*, 201−213.

(40) Labute, P.; Santavy, M. Locating Binding Sites in Protein Structures. Chemical Computing Group, 2007; https://www.chemcomp.com/journal/sitefind.htm (accessed Oct 8, 2014).

(41) Weisel, M.; Proschak, E.; Schneider, G. PocketPicker: Analysis of Ligand Binding-Sites with Shape Descriptors. *Chem. Cent. J.* **2007**, *1*, No. 7.

(42) An, J.; Totrov, M.; Abagyan, R. Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelopes. *Mol. Cell Proteomics* **2005**, *4*, 752−761.

(43) Verdonk, M. L.; Cole, J. C.; Taylor, R. SuperStar: A Knowledge-Based Approach for Identifying Interaction Sites in Proteins. *J. Mol. Biol.* **1999**, *289*, 1093−1108.

(44) Pérot, S.; Sperandio, O.; Miteva, M. A.; Camproux, A.-C.; Villoutreix, B. O. Druggable Pockets and Binding Site Centric Chemical Space: A Paradigm Shift in Drug Discovery. *Drug Discovery Today* **2010**, *15*, 656−667.

(45) Leis, S.; Schneider, S.; Zacharias, M. In silico prediction of binding sites on proteins. *Curr. Med. Chem.* **2010**, *17*, 1550−1562.

(46) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: Automatic and Efficient Detection of Potential Small Molecule-Binding Sites in Proteins. *J. Mol. Graphics Modell.* **1997**, *15*, 359−363.

(47) Howard, N.; Abell, C.; Blakemore, W.; Chessari, G.; Congreve, M.; Howard, S.; Jhoti, H.; Murray, C. W.; Seavers, L. C. A.; van Montfort, R. L. M. Application for Fragment Screening and Fragment Linking to the Discovery of Novel Thrombin Inhibitors. *J. Med. Chem.* **2006**, *49*, 1346−1355.