JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

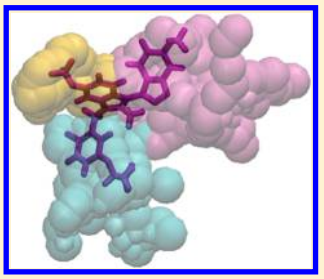
Graph-Based Clustering of Predicted Ligand-Binding Pockets on Protein Surfaces

Jennifer Degac, Uwe Winter, and Volkhard Helms*

Center for Bioinformatics, Saarland University, 66041 Saarbruecken, Germany

Supporting Information

ABSTRACT: Detecting appropriate ligand binding pockets on protein surfaces has several important applications in the drug discovery process. In pocket sets identified by two software packages, PASS and Fpocket, we found a sizable number of protein−ligand complexes where more than one pocket overlaps with the ligand. In such cases, it would be desirable if a merged set of contacting pockets would represent the small molecule. Thus, we tested three clustering approaches to merge the given pockets, a classical clustering method and two methods based on algorithms from graph theory. We found that hierarchical clustering, as well as an approach based on the concept of maximum flow, could be favorably used for clustering pockets predicted either by PASS or by Fpocket.



INTRODUCTION

Proteins play essential roles in all cellular processes. Their biological function is typically defined by their three-dimensional structure and by their interactions with other molecules, such as small molecules, nucleic acids, and other proteins. The reversible, noncovalent binding of small molecules to proteins is critical for many important cellular activities, such as signal transmission, regulatory processes, and immune response. In addition, the majority of approved drugs are small molecules that activate or inhibit the function of a target protein. Hence, the analysis and understanding of protein−ligand interactions are also of high interest for the drug design process. In this context, the identification of potential ligand binding pockets and cavities on the protein surface is of fundamental importance for a range of applications such as molecular docking, structure-based drug design, and high-throughput virtual screening. We note that identification of druggable binding pockets requires, beyond purely geometric parameters, also an analysis of additional features such as physicochemical properties and so on. Our study aims at providing a better defined starting position for such analysis.

Various computational approaches have been proposed in the recent decades for detecting pockets and cavities on protein surfaces. Most of these algorithms can be classified into either geometry-based or energy-based methods. The geometry-based methods are based on geometrical features of the protein surface and identify more or less deep cavities. They can be further subdivided into grid-based, sphere-based, and alpha-shape-based methods. Grid-based approaches place the protein structure onto a three-dimensional grid. Then each grid point located outside the protein is analyzed to decide whether the point belongs to a surface cavity or not. Examples for this type of software tools are POCKET,[1] LIGSITE,[2] PocketPicker,[3] and GHECOM.[4] Sphere-based methods, such as SURFNET,[5] PASS,[6] and PHECOM,[7] add small spheres with different radii on the protein surface and then define accumulations of spheres showing specified properties as putative pockets. The third class of geometry-based methods is based on the alpha-shape theory or the related Voronoi diagrams. Examples are the algorithms CAST,[8,9] Fpocket,[10] and APROPOS.[11] In contrast to the purely geometry-based methods, energy-based methods such as Q-SiteFinder[12] and SITEHOUND[13] estimate the interaction energy between the protein and a putative small molecule from the pairwise interactions between the protein atoms and ligand probe spheres. These types of algorithms are computationally more expensive and also require atom typing and the definition of a force field. In the last years, methods using a combination of several such approaches were developed to improve the prediction accuracy.[14]

In this study, we applied the two popular pocket detection methods—PASS[6] and Fpocket[10]—to a set of 195 protein−ligand complexes retrieved from the PDBbind database.[16] Beside ligands that were not totally enclosed by pockets and pockets that were too large, we found many cases where the small molecule is not covered by only one but by two or even more predicted pockets. Figure 1 shows as an example an X-ray structure,[17] where a potent nonamidine inhibitor is bound to the human factor Xa. In this case, both methods found two overlapping pockets each covering about half of the inhibitor. For cases such as this, it would be preferable if a single pocket resulting from the fusion of multiple contacting pockets could be identified that represents the entire ligand in order to simplify and possibly improve further processing steps. A higher ligand coverage could be desirable, for example, for the comparison of binding sites, for finding small molecule fragments mirroring the protein binding site, for a more accurate druggability prediction, or for virtual screening projects. Therefore, we introduced and tested here three

Received: January 26, 2015
Published: September 1, 2015

DOI: 10.1021/acs.jcim.5b00045
J. Chem. Inf. Model. 2015, 55, 1944−1952

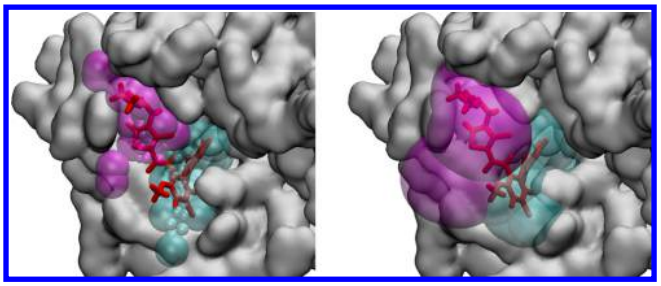


Figure 1. Pockets detected by PASS (left) and Fpocket (right) that overlap with a nonamidine inhibitor bound to human factor Xa (PDB ID: 1mq6). The two PASS pockets have 9 probe sphere contacts whereas the two Fpocket pockets have 1028 alpha-sphere contacts. Figures were generated using Visual Molecular Dynamics (VMD).[15]

clustering methods to merge the pockets generated by PASS and Fpocket.

METHODS

Dataset. To establish the pocket-clustering method, we used the core set of 195 protein−ligand complexes from the PDBbind database version 2014.[16] Protein−ligand complexes in PDBbind are noncovalent assemblies of one protein and a single ligand with a mass of, at most, 1000 Da for which a crystal structure of the bound complex exists with an overall resolution equal to or better than 2.5 Å, and their K_d or K_i value has been experimentally measured. Furthermore, the ligand molecule must contain only common organic elements and the protein molecule has no nonstandard amino acid residues as part of its binding pocket. In order to reduce the redundancy, the eligible complexes were clustered by the PDBbind team on the basis of protein sequence similarity using BLAST and a cutoff of 90%. From each cluster with at least five members, the one with the strongest binding affinity, the one with the weakest binding affinity, and the one with a binding affinity closest to the mean value were selected as representatives. This yielded the set of 195 protein−ligand complexes derived from 65 clusters.

This dataset was preprocessed by us using two different molecular modeling tools. In the first approach, we used the Biochemical Algorithms Library (BALL) version 1.4[18] to add missing hydrogens and to optimize the positions of all hydrogens using the Amber force field. In the second approach, we prepared the coordinate files with the tool PDB2PQR with the options <nodebump>, <noopt>, <chain>, and the Charmm force field.[19] For all types of analyses, the results obtained for the datasets prepared either with BALL or with PDB2PQR were highly similar. Therefore, in the following, we will present and discuss only results that were obtained for the dataset prepared with PDB2PQR. Parts of the results obtained for the complexes prepared with BALL are shown in the Supporting Information.

For the subsequent pocket detection step, all ligands and water molecules should be removed from the protein. This was already the case for all complexes in the core set of the PDBbind database.

Pocket Detection. Pockets on the protein surfaces that are suitable to bind small molecule ligands were detected with two different algorithms, namely, PASS[6] and Fpocket.[10] Both methods use geometric criteria to detect buried cavities on the protein surface and inside the protein. In both cases, we applied the standard parameters suggested by the authors.

The PASS (Putative Active Sites with Spheres) algorithm characterizes cavities on the protein surface by iteratively attaching small nonoverlapping probe spheres to the three-dimensional (3D) molecular structure using three-point Connolly-like sphere geometry.[6] To coat the surface of the protein, an initial layer of spherical probes is generated by placing probes in tangential orientation of the protein atoms. Additional layers of probes then are accreted onto the previously identified probe spheres also utilizing three-point geometry. After each layer, a filtering step removes those probes that clash with any protein atom and retains only those probes with low solvent exposure. The process stops when a new accretion layer produces no new buried probe spheres. In the end, the obtained probes are clustered together, so that each cluster represents a putative binding pocket. In each cluster, a so-called active site point is identified that most likely represents the center of the potential binding pocket. Here, we applied a reimplementation of the PASS algorithm using the BALL library,[18] as previously described.[20]

Fpocket is an open source pocket detection package based on Voronoi tessellation and alpha-spheres.[10] In this scope, an alpha-sphere is a sphere that contacts four protein atoms on its boundary but does not overlap with any protein atom. The center of a thus-defined alpha-sphere corresponds to a Voronoi vertex, which is a point at which Voronoi regions intersect. Hence, alpha-spheres are determined by executing a Voronoi tessellation. The determined alpha-spheres are then filtered according to a minimum size and a maximum size. Alpha-spheres with a large radius have a tendency to be mostly exposed to the solvent, whereas alpha-spheres with a small radius are very often solvent-inaccessible. Pockets are identified by clustering nearby alpha-spheres of proper radius. Alpha-sphere clusters of low interest are removed. Note that, in contrast to PASS probes, the alpha-spheres in a Fpocket pocket can overlap. At the end, the clustered pockets are ranked according to their potential ability to bind small molecules.

We applied both software packages to the "bound" protein structures after stripping away the small molecule ligands. We left out those complexes from the analysis for which no pocket was found that overlaps with the small molecule ligand in the X-ray structure of the complex. For PASS, only one complex (3ov1) was affected, whereas for Fpocket, seven complexes (2jdm, 2jdy, 2r23, 2xy9, 3kwa, 3ov1, 4gqq) had to be removed.

Pocket Clustering. In this study, we merged pockets only on the pairwise relations, namely, the degree of contact between each pair of pockets; hence, this is a case of relational clustering. For clustering the predicted binding pockets, we used one classical clustering method and two popular methods from graph theory. As representative of classical clustering approaches, we chose the commonly used hierarchical clustering.[21] As the name suggests, this method creates a multilevel hierarchy of clusters, in which clusters at any level are composed of two clusters at the next lower level. The highest level consists of a single cluster containing all objects whereas at the lowest level each cluster contains a single object. Hierarchical clustering procedures can be agglomerative or divisive. Agglomerative strategies start at the bottom level and join, at each step, the two clusters that are most similar. In contrast, divisive strategies start at the top level and split, in each step, one of the existing clusters into two new clusters. Here, we used the agglomerative technique implemented in the hclust function of R.[22]

DOI: 10.1021/acs.jcim.5b00045
J. Chem. Inf. Model. 2015, 55, 1944−1952

In order to decide which clusters should be merged, a measure of dissimilarity between two groups of observations is required. For this, one first must specify a distance function that defines the pairwise dissimilarities of single observations. In our case, where predicted pockets should be combined to larger pockets, we took the negated number of sphere contacts (see next paragraph) between two pockets as the dissimilarity between them. To compute the dissimilarity between two groups of observations, we chose the average linkage method. This means the dissimilarity between two clusters is defined as the average of distances between all pairs of objects in the two considered clusters. The resulting cluster hierarchy can be illustrated as a hierarchical tree diagram called a dendrogram. The height of a cluster in this dendrogram corresponds to the dissimilarity between its two daughter clusters. Therefore, cutting the tree at a particular height partitions the dataset into disjoint clusters that would be produced by terminating the clustering procedure when the optimal intergroup dissimilarity exceeds that threshold value.

For the two graph-based methods, the identified pockets were interpreted as the vertices of an undirected graph. Two vertices in the graph were then connected by an edge if the two respective pockets are in physical contact. In both detection methods, each pocket consists of a set of spheres with defined radii. In the case of PASS, they are called probe spheres; in the case of Fpocket, they are termed alpha-spheres. Thus, we defined two pockets to be in contact if at least one sphere of a pocket touches a sphere of the other pocket. With this construction, we can then assign weights to the edges (see below) and use graph-based clustering approaches to merge tightly connected pockets.

The edge betweenness centrality clustering[23] is a popular method to determine the centrality of edges in a given network. It was designed to rank edges in a graph according to the number of shortest paths between all vertices of the graph that use a specific edge. Edge betweenness (or the related Girvan−Newman algorithm) has already been used before in the area of computational chemistry, e.g., to identify protein regions showing correlated dynamics.[24,25]

Given are an undirected graph $G = (V,E)$ and a weight function w: $E \to \mathbb{R}^+$ defined on the edges of the graph, where $w(e) > 0$ for all edges $e \in E$. In our case, the weight is set as the inverse of the number of sphere contacts between the two pockets linked by the edge. Let a path P from $s \in V$ to $t \in V$ be an alternating sequence of vertices and connecting edges that starts at s and ends at t. We then define the weight of this path as the sum of the weights of the traversed edges. A shortest path between s and t is a path P with $w(P) \leq w(F)$ for any other path F connecting vertices s and t. Let us now denote the number of shortest paths connecting vertices s and t by $\sigma_{st} = \sigma_{ts}$ and the number of shortest paths between s and t that contain the edge e by $\sigma_{st}(e)$. The edge betweenness centrality is then defined as follows:

$$C(e) = \sum_{\substack{s,t \in V \\ s \neq t}} \frac{\sigma_{st}(e)}{\sigma_{st}}$$

The actual clustering algorithm calculates the betweenness centrality for all edges in the graph and then removes iteratively the edge with the highest betweenness until the highest betweenness centrality is lower than a given threshold. After each removal, the betweenness centrality for all edges affected by the removal is recalculated. In our study, we used the betweenness centrality method presented in ref 26, as implemented in the Boost Graph Library.

The second method for graph clustering used in this study is based on the concept of maximum flow. This type of clustering was originally introduced for image segmentation.[27] The clusters are either grown using a maximum flow, or they are carved out using a minimum cut, which is equivalent. To understand how this approach works, we must introduce the definition of a flow and the related maximum flow. Given are again the undirected pocket graph $G = (V,E)$ and a capacity function c: $E \to \mathbb{R}^+$ defined on the edges of the graph. In the clustering case, the capacity of an edge reflects the similarity between the two linked vertices. Hence, we used the number of sphere contacts between the linked pockets as edge capacities. Defining a source $s \in V$ and a sink $t \in V$ with $t \neq s$, a non-negative function f: $E \to \mathbb{R}^+$ is called a flow if it fulfills both

(1) the capacity constraint: $\forall(u,v) \in E: f((u,v)) \leq c((u,v))$ and
(2) the flow conservation: $\forall v \in V\backslash\{s,t\}: \sum_{(u,v)\in E} f((u,v)) = \sum_{(v,u)\in E} f((v,u))$.

The value of the flow is then defined as the total sum of outgoing flow from the source vertex s:

$$|f| = \sum_{(s,v)\in E} f((s,v))$$

Given this specification, we can define a maximum flow as a flow of maximum value. It does not need to be unique.

Clustering via maximum flow is straightforward. The idea is to pick a vertex randomly and assign all vertices that have a maximum flow higher than a specified threshold to this vertex to the same cluster. This is repeated until all vertices are assigned to a cluster. For this type of clustering approach, we applied a modified method of Edmonds and Karp[28] as implemented in the Boost Graph Library.

In order to get reference points for the clustering results produced by the three methods described above, we additionally applied a rather simple clustering algorithm. In this approach, two predicted pockets are merged into one single pocket if they both contain either probe spheres (in the case of PASS) or alpha-spheres (in the case of Fpocket) that are near the same protein atom. In this study, we defined a maximal distance of 5 Å for such neighbors.

Pocket Parameters. To compare the results obtained from different clustering approaches, we defined four pocket parameters (PPL, OLV, UPV, and PV). The first parameter is the "pockets per ligand" (PPL) value, which specifies the average number of pockets on a protein surface detected by PASS or Fpocket that overlap with the small molecule ligand of the related protein. The next three parameters are calculated only for those pockets that have the largest overlap with the ligands. This means that, for each complex, a single pocket is considered. OLV is an abbreviation for "overlapping ligand volume" and measures the fraction of the volume of a small molecule in the given dataset that is covered by the respective pocket. The pockets themselves are described by two parameters: "uncovered pocket volume" (UPV) and "pocket volume" (PV). As the name denotes, PV is the average volume of the relevant pockets and UPV is the fraction of the pocket volume that is not occupied by the according ligand. During the volume calculations for the Fpocket pockets, we corrected the radii of the alpha-spheres by a factor of −1.6 Å, as is done in the

DOI: 10.1021/acs.jcim.5b00045
J. Chem. Inf. Model. 2015, 55, 1944−1952

source code of Fpocket when computing pocket volumes. In the following text, these four pocket parameters are usually given as the average over the full dataset or over a specified subset.

Pocket Clustering Validation. To assess the clustering results and determine default parameters, we used a criterion that has the objective of optimizing both OLV and UPV at the same time. This means that the fraction of the volume of the ligand that is covered by the resulting pocket (OLV) should be maximal and the fraction of the pocket volume that is not occupied by the ligand (UPV) should be minimal. We cubed the OLV to place emphasis on the ligand coverage. Precisely, we determined that threshold to be the default parameter that gave a minimal value of UPV/OLV^3 over the full dataset. We computed the average ratio value 10 times for different subsets where we systematically omitted 10% of the protein complexes and obtained almost identical optimal parameters. Moreover, it turned out that the default parameters determined on 90% of the data led, in almost all cases, to a lower average UPV/OLV^3 ratio for the 10% omitted data than for the "training set" of 90%. This validates the robustness of this approach. In the end, we selected the median of the 10 runs as the default parameter.

RESULTS AND DISCUSSION

In this study, we are interested in pockets on protein surfaces predicted by the two algorithms PASS and Fpocket which are overlapping with the small molecule ligands in our dataset that bind specifically to these proteins. We observed that, in many cases, the ligand is not covered by only one pocket but rather by two or even more pockets. As shown in Table 1, this is the case for 47% of the complexes in our dataset when using PASS and for 25% of the complexes when using Fpocket. Thus, we defined the measure pockets per ligand (PPL) as the number of detected pockets that are overlapping with the ligand. PASS detected for the dataset, on average 1.67 PPL, whereas Fpocket gave a lower average of 1.3 PPL. This fact matches with the observation that Fpocket generated larger pockets with an average volume (PV) of 2774.2 $Å^3$, whereas PASS produced pockets with an average pocket volume of 1067.8 $Å^3$. Note that the average volume of the ligands is 316.2 $Å^3$ for the PASS dataset and 316.8 $Å^3$ for the Fpocket dataset. Thus, one would assume that one large pocket of Fpocket replaces two or more smaller pockets of PASS. However, PASS pockets have an average overlapping ligand volume (OLV) of 69%, resulting in a higher coverage of the ligand than Fpocket pockets that have an average overlapping ligand volume of 60%. This fits the observation made on individual complexes where a set of PASS pockets almost covered the entire ligand, whereas a single large Fpocket pocket only covered a part of the ligand (see Figure 2).

Table 1. Number of Structures in the Dataset of 195 Protein−Ligand Complexes Possessing the Specified Number of Pockets Overlapping with the Ligand before the Clustering[a]

	Number of Structures				
	1 PPL[b]	2 PPL[b]	3 PPL[b]	4 PPL[b]	5 PPL[b]
PASS	103	61	22	7	1
Fpocket	141	40	5	2	0

[a]The pockets were detected using either PASS or Fpocket. [b]PPL = pockets per ligand.

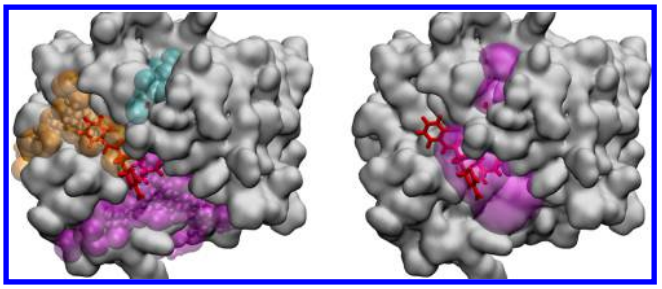


Figure 2. Pockets detected by PASS (left) and Fpocket (right) that overlap with an inhibitor bound to the catalytic domain of human stromelysin-1 (PDB ID: 1sln). Figures were generated using VMD.[15]

For further processing of the complexes and the according pockets, and possibly for applications in drug design projects, we suggest that it would be preferable if a single pocket or a merged set of multiple contacting pockets could be identified that represents the entire ligand. For this, we introduced three clustering methods (see the Methods section) to merge the given pockets based on the magnitude of contact between them. Ideally, the fused pockets should perfectly overlap with the ligand and show a value near 1 for the PPL parameter. However, this number alone is not a good measure for the quality of the clustering. In the extreme case, all detected pockets could be fused into one very large pocket spreading over the entire protein surface. All small molecules then would indeed most likely be covered by single pockets; however, this, of course, is not desirable. Thus, we introduced another criterion that evaluates the percentage of the pocket volume that is not occupied by the ligand. In the following, we call this the uncovered pocket volume (UPV). To compute the UPV, as well as PV and OLV already mentioned above, we used the pocket that has the largest overlap with the ligand. Therefore, in an ideal case, where a single pocket completely covers the entire ligand but occupies no additional space, we would get a PPL value of 1, an UPV value of 0%, and an OLV value of 100%. To be able to better judge the UPV values of the clustering, we first must look at the values obtained for the unclustered case. For PASS, an UPV value of 46% was measured, and for Fpocket, a value of 71% was measured. This means that PASS pockets fit more snugly around the ligand.

In a first step, we applied a simple agglomerative clustering algorithm to the dataset. In this method, two pockets are combined to a single pocket if any of their sphere centers are near to the same protein atom (see the Methods section). This approach indeed gave an average of 1.0 PPL for both pocket detection methods but also an average PV of 15 764.9 $Å^3$ for PASS and 15 520.0 $Å^3$ for Fpocket. Thus, using this method probably leads, in most cases, to the extreme case described above, where most of the initial pockets are merged into one large pocket. For our dataset, this results in an improvement of the OLV value from 69% to 75% for PASS and from 60% to 64% for Fpocket. In exchange, the UPV increases from 46% to 94% for PASS and from 71% to 87% for Fpocket, which is clearly undesirable.

In order to examine the influence of the tunable clustering parameter (edge betweenness, maximum flow, and cut threshold), we tested a range of appropriate thresholds for all three clustering methods. Figure 3 shows the results for the hierarchical clustering (gray lines) and the clustering via maximum flow (black lines), either for the full dataset or for the subset of complexes that possess a PPL value of >1 in the

DOI: 10.1021/acs.jcim.5b00045
J. Chem. Inf. Model. 2015, 55, 1944−1952

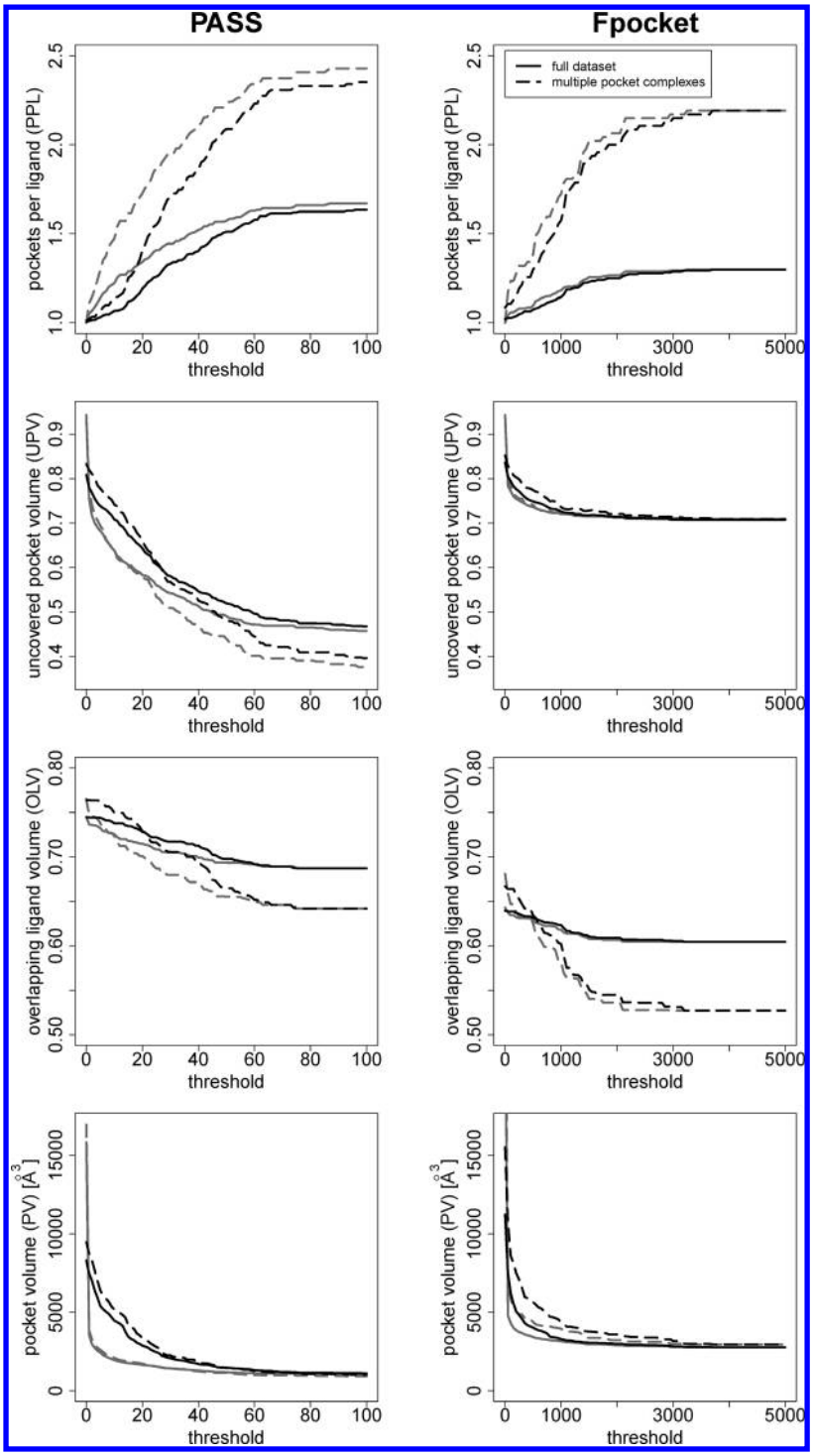


Figure 3. Pocket parameters when pockets are clustered with maximum flow (black lines) or hierarchical approach (gray lines) for different clustering thresholds. Solid lines are for the full dataset and dashed lines are for mp complexes.

unclustered pocket set. This latter subset is meaningful for judging the performance of the three pocket clustering methods. In the following, these complexes are called multiple pocket complexes, abbreviated as mp complexes. The findings from the hierarchical clustering and the maximum flow clustering were highly similar for the pocket sets predicted by Fpocket and similar for the pocket sets predicted by PASS.

Using either of the two approaches, all four parameters (PPL, UPV, OLV, and PV) smoothly converge with increasing threshold toward the limiting values for the unclustered case with no obvious transition point. For small thresholds, almost all pockets are clustered together into one pocket, because of already a small flow and a low similarity, meaning a loose contact between the pockets, suffices to be merged into one cluster. Using a threshold of zero for the hierarchical clustering causes the fusion of all pockets into a single pocket, whereas for the clustering via maximum flow, only those pockets that actually touch are merged into one pocket. This circumstance is reflected on the left side of the UPV and PV graphs, where the curve for the hierarchical clustering abruptly rises more strongly

DOI: 10.1021/acs.jcim.5b00045
J. Chem. Inf. Model. 2015, 55, 1944−1952

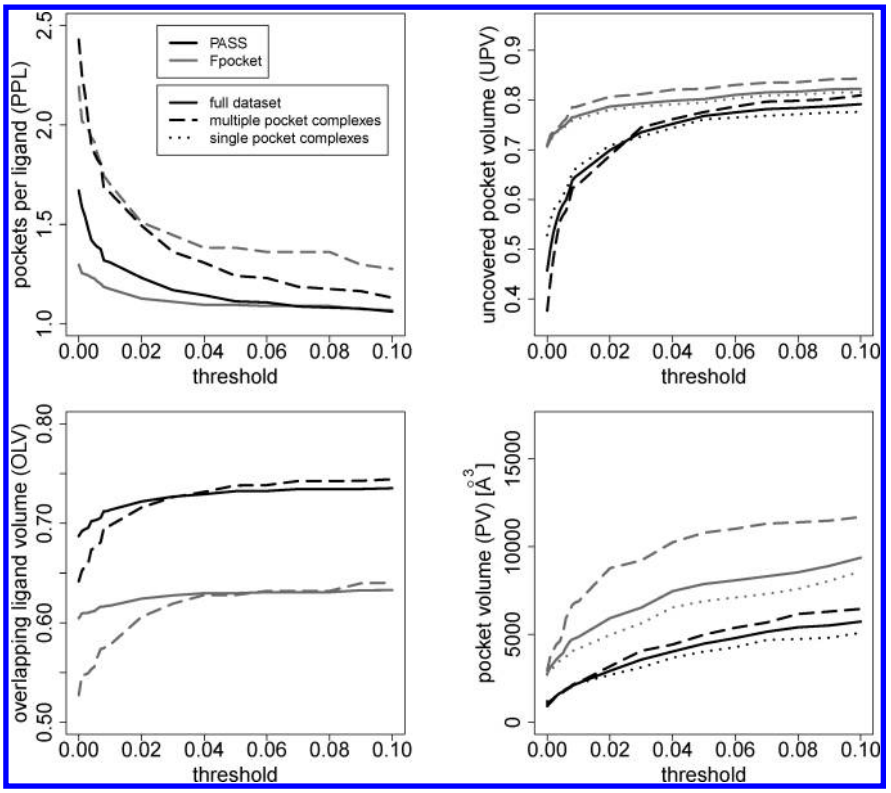


Figure 4. Pocket parameters using clustering via edge betweenness centrality as a function of the betweenness threshold used for deleting edges. Solid lines represent the full dataset, dashed lines represent mp complexes, and dotted lines represent sp complexes. Black lines belong to the PASS results, and the gray lines correspond to the Fpocket results.

than the curve for the maximum flow clustering. Unfortunately, for intermediate thresholds, the UPV increases as the PPL decreases. Thus, the selection of the optimal threshold is a balancing act between too much empty space in the pockets and pockets that are too small to cover the entire ligand. Interestingly, although the OLV is not changing strongly neither for the full dataset nor for the mp complexes, the change for the mp complexes is larger. This observation indicates that the clustering is advantageous for the mp complexes, because it leads to better coverage of the ligand. With respect to PV, no large difference is found between the full dataset and the mp complexes, i.e., the volume of the pockets formed by the clustering is not dependent on the PPL value in the beginning.

Regarding the results obtained for PASS pockets, the UPV for clustering via maximum flow is larger for all thresholds (except zero), compared to the UPV for hierarchical clustering. The OLV is equal for both clustering methods for large clustering parameters and starts to differ at a threshold of ∼60. For smaller thresholds, the OLV for hierarchical clustering is again lower than that for the maximum flow clustering. Thus, it seems that, for larger thresholds, the hierarchical clustering is slightly better suited for combining PASS pockets than the maximum flow clustering.

For the hierarchical clustering and the clustering via maximum flow, a totally different range of values appears to be optimal for the two pocket detection methods. This method dependency was not so pronounced for the clustering via edge betweenness centrality (see Figure 4). Also here, the results for the edge betweenness centrality clustering showed smooth progress with no clear transition. One must remember that higher thresholds lead to the tendency to cluster all pockets together, cf. the right side of the graphs in Figure 4. The unclustered situation is found on the left side with a threshold of zero. Since the same threshold range was suitable for this type of clustering, the results attained for Fpocket (gray lines) and PASS (black lines) are plotted into joint diagrams. But also with the edge betweenness centrality, the threshold for merging Fpocket pockets only converges at larger values to a PPL value close to 1 for mp complexes, what means that as many pockets as possible are clustered into a single cluster. Thus, again, the choice of threshold is dependent on the pocket identification algorithm and therefore must be optimized for each algorithm separately.

In the UPV and PV diagrams, additional dotted lines represent the results for single pocket (sp) complexes. In contrast to mp complexes, these sp complexes have a PPL value of 1 in the unclustered pocket set. In these cases, only one pocket overlaps with the ligand. Yet, also these pockets can merge with neighboring pockets. It is striking that, for Fpocket, the PV parameter is dependent on the number of pockets in the unclustered case, as reflected by the gap between the curves representing mp and sp complexes. As the threshold increases, the PV value rises faster for the mp complexes than for the sp complexes as more pockets are merged into a larger pocket. Actually, it appears as an advantage for Fpocket that, for sp complexes, where the clustering is useless, the overlapping pocket is not unnecessarily expanded. However, when looking at single examples, we observed that, for sp complexes, the single overlapping pocket is often isolated from other detected pockets and, therefore, there are no neighboring pockets with which to merge.

As already mentioned above, the OLV for the mp complexes shows a more favorable trend than the OLV for the full dataset

DOI: 10.1021/acs.jcim.5b00045
J. Chem. Inf. Model. 2015, 55, 1944−1952

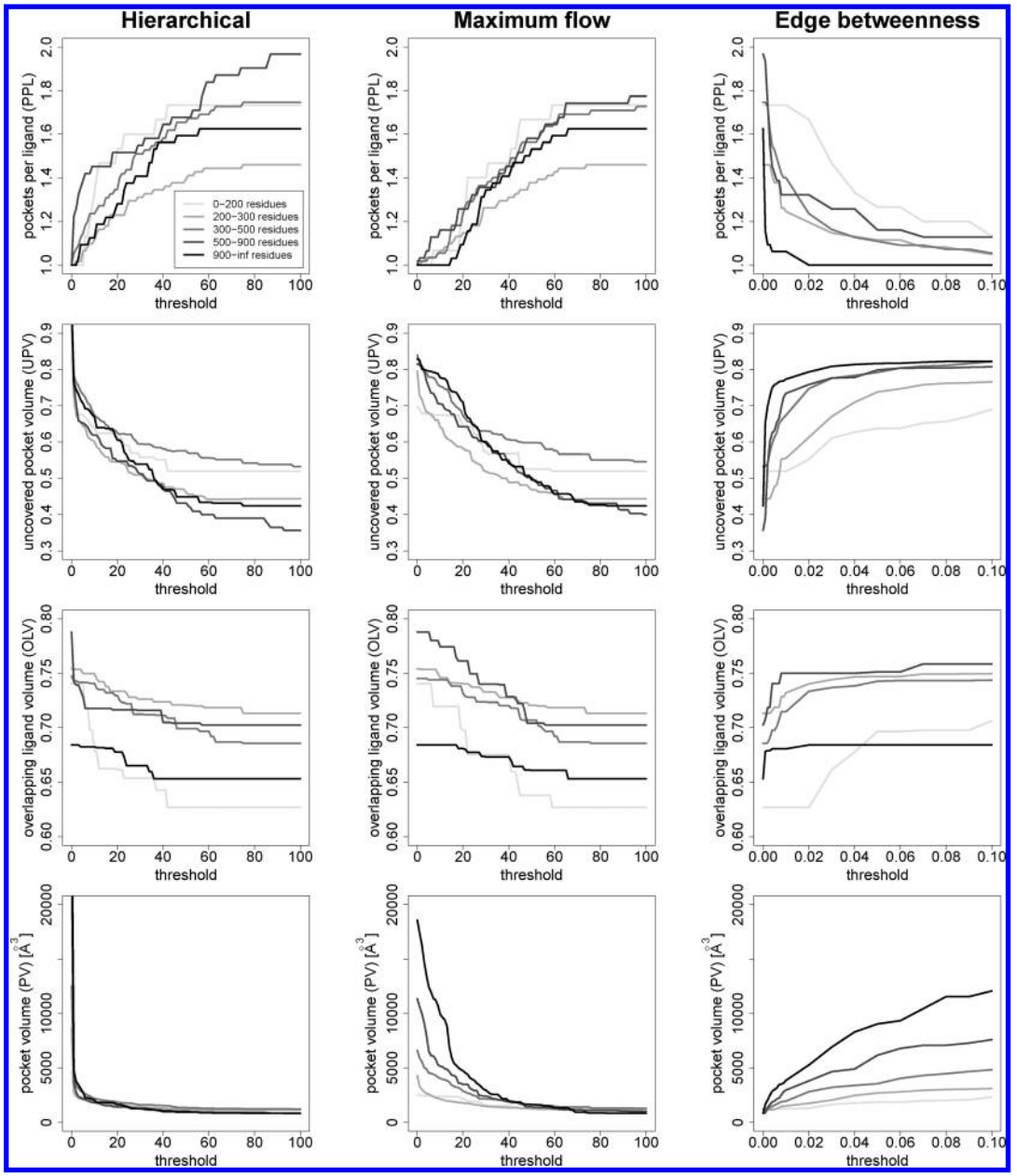


Figure 5. Pocket features for complexes grouped according to the protein size. The initial pockets were detected by PASS. Lines colored in light gray to black are the results for the complexes with 0−200 residues (light gray), 200−300 residues, 300−500 residues, 500−900 residues, or 900−∞ residues (black).

(see Figure 4). Generally, PASS pockets have a larger (i.e., more desirable) OLV and, in addition, a lower (i.e., more desirable) UPV ratio than Fpocket pockets. These observations hold for the unclustered pocket set as well as for any other situation. When considering that Fpocket assigns only 25% of the complexes in the dataset as mp complexes and taking into account the stronger increase of the PV for the mp complexes during pocket clustering, we suggest that the clustering approach is slightly better suited for pocket sets identified by PASS than those identified by Fpocket. However, also for Fpocket pockets, the clustering is advantageous, as is reflected by the decrease of the PPL value and the increase in the OLV value.

The comparison between the results obtained for the clustering by maximum flow/hierarchical approach and by edge betweenness centrality is remarkable. For PASS as well as Fpocket, the UPV and PV values increase faster for the edge betweenness than for the maximum flow, whereas the OVL shows the same trend for all three clustering approaches. At equal coverage of the ligand and equal PPL value, this means that the edge betweenness centrality clustering makes the pockets unnecessarily large. Thus, the maximum flow clustering and the hierarchical clustering seem to match our purposes better than the edge betweenness centrality clustering. The hierarchical clustering has the additional benefit that it is a conceptually simpler algorithm that can be straightforwardly implemented, and its results are easy to interpret.

Finally, we grouped the dataset into five groups of proteins of different sizes and analyzed how this affected the clustering results (see Figure 5). The size of a protein was defined by the number of residues. When using the edge betweenness for clustering, the curves show a staggered behavior, indicating a

DOI: 10.1021/acs.jcim.5b00045
J. Chem. Inf. Model. 2015, 55, 1944−1952

dependency of the threshold on the protein size. This can be explained by the definition of edge betweenness centrality as the number of shortest paths using this edge. Because of this, the clustering is highly dependent on the structure and size of the graph and less on the weights of the edges. In this context, the graph size can be related to the protein size, since, for a large protein, typically a larger number of pockets is detected, which leads to a larger graph. In contrast, the hierarchical clustering as well as the clustering via maximum flow show little dependency on the protein size. Here, the clustering is mostly influenced by the degree of contact between two neighboring pockets, represented by the edge weights used for clustering via maximum flow and by the distance function of the hierarchical clustering. Thus, again, the clustering by maximum flow and the hierarchical clustering are better suited for our purpose than the clustering by edge betweenness.

For our two favored clustering methods, maximum flow and hierarchical clustering, we determined a default parametrization by minimizing the average of UPV/OLV[3] (see the Methods section). When examining examples of single mp complexes, this criterion appeared to identify an optimal threshold with maximum ligand coverage, avoiding the addition of unnecessary pockets to the ligand overlapping pocket. In contrast to PASS probes that are not allowed to overlap each other, Fpocket spheres are highly overlapping. In addition, the radii of the spheres identified by Fpocket and PASS differ. These two facts influence the number of sphere contacts between neighboring pockets and thus the clustering process. Therefore, we had to choose different default parameters for the two pocket detection methods. For clustering pockets predicted by Fpocket, a default parameter of 450 was determined for both clustering algorithms. As already stated above, when applied to PASS pockets, the two clustering methods gave slightly diverging results. Therefore, the default parameters also differ. For clustering via maximum flow, a threshold of 40 was obtained as the default; for hierarchical clustering, a threshold of 45 was obtained as the default.

So far, we have presented results for a dataset of 195 bound protein structures. To validate if the introduced clustering approach is also beneficial for unbound protein structures, we also applied it to a dataset of 48 unbound proteins and their corresponding protein−ligand complexes compiled by Huang and Schroeder.[29] Pockets were predicted for the unbound structure using Fpocket and PASS. The pocket parameters were then computed using the ligand of the related complex. For three proteins (1krn, 2rta, 3app) no overlapping PASS pockets were found; for five proteins (1hel, 1pdy, 2rta, 3app, 6ins), no overlapping Fpocket pockets were found. However, for one protein (2fbp), two different ligands were used. For comparison, we also determined pockets and pocket parameters for the associated complex structures. Those results are presented in Table 2 (in parentheses).

Table 2 shows that there are indeed proteins for which more than one pocket was found that overlap with the ligand of the related complex. Hence, we applied the hierarchical clustering using default thresholds to this dataset. The clustering procedure caused an improvement of the PPL value from 1.57 (1.48) to 1.48 (1.37) for PASS and from 1.36 (1.34) to 1.11 (1.16) for Fpocket. With regard to the OLV, that means an increase from 69% (68%) to 70% (70%) for PASS and an increase from 58% (65%) to 60% (66%) for Fpocket; with regard to the UPV, a slight increase is observed, from 56% (53%) to 57% (55%) for PASS and from 69% (68%) to 71% (70%) for Fpocket. Given these observations, we suggest that the clustering routine is useful when starting pocket detection from apo structures, as is often the case in the *de novo* design of ligands.

Table 2. Number of Structures in the Dataset of 48 Unbound Proteins Possessing the Specified Number of Pockets Overlapping with the Ligand of the Related Bound Complex before and after Hierarchical Clustering Using Default Thresholds[a]

	Number of Structures			
	1 PPL	2 PPL	3 PPL	4 PPL
PASS				
unclustered	27 (32)	13 (7)	5 (6)	1 (1)
clustered	28 (35)	14 (5)	4 (6)	0 (0)
Fpocket				
unclustered	31 (29)	10 (15)	3 (0)	0 (0)
clustered	39 (37)	5 (7)	0 (0)	0 (0)

[a]Values given in parentheses represent the corresponding numbers for the bound complexes. The pockets were detected using either PASS or Fpocket.

CONCLUSION

Here, we applied the two pocket identification methods PASS and Fpocket to the core set of 195 protein−ligand complexes retrieved from the PDBbind database. For 47% of the complexes using PASS and 25% of the complexes using Fpocket, we found that more than a single pocket on the protein surface overlapped with the ligand bound to the related protein. Since it would be desirable if only a single pocket or a merged set of neighboring pockets represented a small molecule, we tested three different methods to cluster the identified pockets for a protein by using either maximum flow, edge betweenness centrality, or hierarchical clustering. We noticed that the selection of an optimal clustering threshold is a balancing act between pockets that are too small to cover the entire ligand and pockets that contain too much empty space. For all three clustering approaches, the threshold showed a dependency on the method used for pocket identification. The edge betweenness clustering was found to be additionally dependent on the protein size. Our analysis revealed that both the hierarchical and the maximum flow approach are better suited for the clustering of pockets predicted by PASS and Fpocket than the edge betweenness centrality approach. We determined default parameters for those two clustering algorithms. For both pocket detection methods, the procedure of merging their pockets was beneficial, not only when starting from bound protein structures, but also when starting from unbound protein structures. It is suggested that the merging of pockets will also be applicable to post-process the output of other pocket detection tools (although this was not tested here).

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00045.

Pockets per ligand (PPL) distribution for the datasets prepared with the Biochemical Algorithms Library (BALL), using the different clustering approaches (PDF)

DOI: 10.1021/acs.jcim.5b00045
J. Chem. Inf. Model. 2015, 55, 1944−1952

AUTHOR INFORMATION

Corresponding Author

*Tel.: +49 (0)681 302 70701. Fax: +49 (0)681 302 70702. E-mail: volkhard.helms@bioinformatik.uni-saarland.de.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Levitt, D. G.; Banaszak, L. J. POCKET: A Computer Graphics Method for Identifying and Displaying Protein Cavities and their Surrounding Amino Acids. *J. Mol. Graphics* **1992**, *10*, 229−234.
(2) 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.
(3) Weisel, M.; Proschak, E.; Schneider, G. PocketPicker: Analysis of Ligand Binding-Sites with Shape Descriptors. *Chem. Cent. J.* **2007**, *1*, 7.
(4) Kawabata, T. Detection of Multiscale Pockets on Protein Surfaces Using Mathematical Morphology. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 1195−1211.
(5) Laskowski, R. A. SURFNET: A Program for Visualizing Molecular Surfaces, Cavities, and Intermolecular Interactions. *J. Mol. Graphics* **1995**, *13*, 323−330.
(6) Brady, G. P., Jr.; Stouten, P. F. Fast Prediction and Visualization of Protein Binding Pockets with PASS. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 383−401.
(7) Kawabata, T.; Go, N. Detection of Pockets on Protein Surfaces Using Small and Large Probe Spheres to Find Putative Ligand Binding Sites. *Proteins: Struct., Funct., Genet.* **2007**, *68*, 516−529.
(8) Liang, J.; Woodward, C.; Edelsbrunner, H. Anatomy of Protein Pockets and Cavities: Measurement of Binding Site Geometry and Implications for Ligand Design. *Protein Sci.* **1998**, *7*, 1884−1897.
(9) Dundas, J.; Ouyang, Z.; Tseng, J.; Binkowski, A.; Turpaz, Y.; Liang, J. CASTp: Computed Atlas of Surface Topography of Proteins with Structural and Topographical Mapping of Functionally Annotated Residues. *Nucleic Acids Res.* **2006**, *34*, W116−W118.
(10) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinf.* **2009**, *10*, 168.
(11) 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.
(12) Laurie, A. T. R.; Jackson, R. M. Q-SiteFinder: An Energy-Based Method for the Prediction of Protein-Ligand Binding Sites. *Bioinformatics* **2005**, *21*, 1908−1916.
(13) Ghersi, D.; Sanchez, R. EasyMIFs and SiteHound: A Toolkit for the Identification of Ligand-Binding Sites in Protein Structures. *Bioinformatics* **2009**, *25*, 3185−3186.
(14) Zhang, Z.; Li, Y.; Lin, B.; Schroeder, M.; Huang, B. Identification of Cavities on Protein Surface Using Multiple Computational Approaches for Drug Binding Site Prediction. *Bioinformatics* **2011**, *27*, 2083−2088.
(15) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.
(16) Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.
(17) Adler, M.; Kochanny, M. J.; Ye, B.; Rumennik, G.; Light, D. R.; Biancalana, S.; Whitlow, M. Crystal Structures of Two Potent Nonamidine Inhibitors Bound to Factor Xa. *Biochemistry* **2002**, *41*, 15514−15523.
(18) Hildebrandt, A.; Dehof, A.; Rurainski, A.; Bertsch, A.; Schumann, M.; Toussaint, N.; Moll, A.; Stockel, D.; Nickels, S.; Mueller, S.; Lenhof, H. P.; Kohlbacher, O. BALL—Biochemical Algorithms Library 1.3. *BMC Bioinf.* **2010**, *11*, 531.
(19) Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A. PDB2PQR: Expanding and Upgrading Automated Preparation of Biomolecular Structures for Molecular Simulations. *Nucleic Acids Res.* **2007**, *35*, W522−W525.
(20) Eyrisch, S.; Helms, V. Transient Pockets on Protein Surfaces Involved in Protein−Protein Interaction. *J. Med. Chem.* **2007**, *50*, 3457−3464.
(21) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edition; Springer Series in Statistics; Springer Science+ Business Media: New York, 2009; pp 520−528.
(22) R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2014.
(23) Girvan, M.; Newman, M. E. J. Community Structure in Social and Biological Networks. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 7821−7826.
(24) McClendon, C. L.; Kornev, A. P.; Gilson, M. K.; Taylor, S. S. Dynamic Architecture of a Protein Kinase. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, E4623−E4631.
(25) Sethi, A.; Eargle, J.; Black, A. A.; Luthey-Schulten, Z. Dynamical Networks in tRNA Protein Complexes. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 6620−6625.
(26) Brandes, U. A. Faster Algorithm for Betweenness Centrality. *J. Math. Sociol.* **2001**, *25*, 163−177.
(27) Wu, Z.; Leahy, R. An Optimal Graph Theoretic Approach to Data Clustering: Theory and Its Application to Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **1993**, *15*, 1101−1113.
(28) Edmonds, J.; Karp, R. M. Theoretical Improvements in Algorithmic Efficiency for Network Flow Problems. *J. Assoc. Comput. Mach.* **1972**, *19*, 248−264.
(29) Huang, B.; Schroeder, M. LIGSITEcsc: Predicting Ligand Binding Sites Using the Connolly Surface and Degree of Conservation. *BMC Struct. Biol.* **2006**, *6*, 19.

DOI: 10.1021/acs.jcim.5b00045
J. Chem. Inf. Model. 2015, 55, 1944−1952