Published in final edited form as:

J Chem Inf Model. 2010 April 26; 50(4): 589–603. doi:10.1021/ci900397t.

Protein Pockets: Inventory, Shape, and Comparison

Ryan G. Coleman^{1,2,‡} and Kim A. Sharp^{1,2,*}

¹The Johnson Research Foundation and Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, PA 19104

²Genomics and Computational Biology Graduate Group, University of Pennsylvania, Philadelphia, PA 19104

Abstract

The shape of the protein surface dictates what interactions are possible with other macromolecules, but defining discrete pockets or possible interaction sites remains difficult. First, there is the problem of defining the extent of the pocket. Second, one has to characterize the shape of each pocket. Third, one needs to make quantitative comparisons between pockets on different proteins. An elegant solution to these problems is to sort all surface and solvent points by Travel Depth, and then collect a hierarchical tree of pockets. The connectivity of the tree is determined via the deepest saddle points between each pair of neighboring pockets. The resulting pocket surfaces tessellate the entire protein surface, producing a complete inventory of pockets. This method of identifying pockets also allows one to easily compute important shape metrics, including the problematic pocket volume, surface area, and mouth size. Pockets are also annotated with their lining residue lists, polarity, and other residue based properties. Using this tree and the various shape metrics pockets can be merged, grouped, or filtered for further analysis. Since this method includes the entire surface it guarantees that any pocket of interest will be found among the output pockets, unlike all previous methods of pocket identification. The resulting hierarchy of pockets is easy to visualize and aids users in higher level analysis. Comparison of pockets is done using the shape metrics, avoiding the complex shape alignment problem. Example applications show that the method facilitates pocket comparison along mutational or time-dependent series. Pockets from families of proteins can be examined using multiple pocket tree alignments to see how ligand binding sites or other pockets have changed with evolution. Our method is called CLIPPERS, for Complete Liberal Inventory of Protein Pockets Elucidating and Reporting on Shape.

Keywords

Pockets; depth; ligand binding; molecular surface; protein shape

Introduction

The shape and properties of the protein surface determine what interactions are possible with ligands and other macromolecules. Pockets are an important yet ambiguous feature of this

*Corresponding author: Kim A. Sharp, Dept. of Biochemistry and Biophysics, University of Pennsylvania, 37th and Hamilton Walk, Philadelphia, PA 19104-6059, Phone: 215-573-3506, Fax: 215-898-4217, sharpk@mail.med.upenn.edu.

‡Present Address: Department of Pharmaceutical Chemistry, University of California San Francisco, 1700 4th Street, San Francisco, CA 94158-2518, USA

Supporting Information: A movie depicting the transition of adenylate kinase is shown, pockets found along the pathway are shown colored by Travel Depth. This material is available free of charge via the Internet at <http://pubs.acs.org>.

surface. For example the first pass in screening for lead compounds and drug-like molecules is usually a filter based on the shape of the binding pocket¹, and shape plays a role in many computational pharmacological methods as reviewed by Kortagere et al². A study of drug-binding pockets found that most features important to predicting drug-binding were related to size and shape of the binding pocket, with the chemical properties of secondary importance³. The surface shape is also important for interactions between protein and water. This depends, for instance, on how wide or narrow the pocket, or how deep or shallow the pocket as reviewed by Levitt and Park⁴. However, defining discrete pockets or possible interaction sites remains difficult despite many studies, for example see the review of Campbell et al⁵. Compounding the problem is that the shape and location of nearby pockets can affect promiscuity and binding site diversity⁶. The primary difficulty is in defining the border of a pocket, as most pockets are open to solvent. Those closed to solvent we refer to as buried cavities. Buried cavities are more straightforward to locate as they have a well-defined extent, area and volume. In contrast, the border of an open pocket defines its mouth and it provides the cut-off for determination of the surface area and volume. The border definition problem for open pockets has been discussed before as a 'can-of-worms problem'⁷. Even defining the pocket as a set of residues does not define the volume or the mouth of the pocket.

Several very different solutions, and therefore pocket definitions, have been proposed. These include fattening the atoms to close off pockets⁷, defining pockets as clustered sets of spheres⁸⁻¹⁵, by using discrete flow analysis on alpha-shapes¹⁶, and by using a larger probe radius to construct a surface or alpha-shape that acts as the pocket mouth^{3, 17, 18}, by examining clusters of lines through solvent^{19, 20}, by defining pockets of interest to only fall in a narrow range of surface areas and shapes and then generating multiple overlapping pockets covering the protein surface for evaluation²¹. Other methods focus only indirectly on shape, for instance by examining pockets predicted by evolution²² or by protein motion changes upon binding²³. Various combinations of these methods are also employed^{24, 25}, including methods that find regions where certain combinations of features are clustered or combined within a statistical framework^{26, 27}.

A common problem with any specific definition of a pocket or any method for finding a small number of non-overlapping pockets on a protein is that they may miss the actual pocket of biological interest. For example defining pockets to be bottlenecks (a narrowed region of the pocket that defines the mouth) as several methods do will miss non-bottleneck pockets, such as clefts, entirely. Other methods and definitions can also miss certain types of pockets or need parameter adjustment to capture relevant pockets.

We present here an alternative definition of pockets, one general enough to create what we call a complete inventory of pockets: In this inventory the entire surface is tessellated into protein pocket regions, each pocket being organized into a hierarchical tree of sub-pockets. The basic idea is illustrated in Figure 1, and it described in detail in the methods section. If a protein had a molecular surface which was convex everywhere, this surface would be identical to what is known as its convex hull²⁸. Clearly such a protein would have no pockets, however relaxed the definition. However a real protein's molecular surface is not identical to its convex hull; it lies within the latter surface at many points (Figure 1a). Thus in seeking pockets our attention is directed to *both* the molecular surface that lies within the convex hull, *and* the solvent accessible volume that lies between the two surfaces (the intermediate volume). It is in this combined surface/volume region that every protein pocket must lie. The foundation for inventorying the pockets is Travel Depth²⁹. Travel Depth is an efficient way to determine the shortest distance, traveling only through solvent, from any point on the molecular surface point or in the intermediate volume to the convex hull. This distance provides the basis for the inventorying step.

In addition to presenting a new definition of pockets, a new way of comparing pockets is described. Most algorithms for comparing two binding sites assume the binding site is known or locate it solely based on proximity to a ligand in the co-crystal structure. After that most algorithms that use spatial information to come up with a motif of various chemical properties and their arrangements in space, and rely on some alignment or geometric hashing technique to compare binding sites based on these structural motifs^{25, 30-42}. Motif definitions can involve hydrogen bond donors or acceptors, residues, or atom types based on residues or can involve the complete set of docked substrates⁴³. Here we present a new method of comparison of pockets based solely on the shape features.

We first describe the use of Travel Depth to create a complete inventory of protein pockets, including construction of the complete tree of protein pockets, then we describe the computation of various pocket metrics and a way to quantitatively compare pockets. We then show various applications of the methods, including display of pockets and visualization of pocket properties, analysis of pockets along mutational and time series of structures, and the clustering of pockets from different members of evolutionarily related protein families.

Materials and Methods

Computation of Travel Depth

This work builds on the concept of Travel Depth, first used to analyze surfaces and ligand binding sites²⁹, with subsequent speed and algorithm improvements⁴⁴. The Travel Depth algorithm computes the shortest molecule interior-avoiding paths from all surface points to the convex hull of a given macromolecule. The algorithm also computes the Travel Depth of points in the intermediate volume between the molecular surface and the convex hull. Additionally the algorithm puts the surface points and volume grid points in a graph structure with the distances between each point as the edge lengths between adjacent nodes, which aids in later steps. The outline of the algorithm is as follows:

Starting with the atomic coordinates, the molecular surface⁴⁵ is generated using a 1.2Å solvent probe radius. The convex hull of this surface is generated using the Qhull algorithm²⁸. These surfaces are mapped onto an appropriately scaled cubic grid, and all grid points are assigned to either to the interior of the molecular surface, outside the convex hull, or between the two surfaces. The Travel Depth of all molecular surface points and intermediate volume grid points is computed as described previously using the multiple source shortest paths algorithm⁴⁶, avoiding the interior points.

We extend the original Travel Depth algorithm here to include a definition of Travel Depth for buried cavities. Previously these cavities were removed completely, which made analyzing ligands inside them impossible. The extension to buried cavities is done by adding one 'virtual' edge per cavity to connect it to the exterior molecular surface. This edge connects the closest cavity and exterior surface points. The length of this edge defines the Burial Depth of that cavity⁴⁷. After adding a virtual connecting edge to each buried cavity the Travel Depth algorithm is applied as described above. Due to these connecting edges, Travel Depth values are now propagated to all buried cavity surface points and their enclosed volume grid points.

The rationale for defining the burial depth of a cavity by the shortest distance to the main surface is that this route would require the least amount of protein motion to open the cavity to bulk solvent. Of course the protein may open by a different route, and if experimental or simulation data were available, a more accurate burial depth estimate could be made.

Nevertheless, the closest distance connection is a useful device to seamlessly include cavities in the analysis of pockets.

Pocket Inventory

The goal of this step of the algorithm is to enumerate all pockets by analyzing all regions of the molecular surface that lie below the convex hull. By enumerating all pockets over the entire protein surface we produce an unbiased collection, rather than focusing *a priori* on a subset of possible pockets.

The inventory algorithm has two phases. In the first phase, all surface and intermediate volume grid points with a defined Travel Depth are put into a list and that list is sorted so the deepest points are first. Ties are broken randomly, but the sorted order is kept fixed throughout the algorithm. To keep track of pockets, a union-find data structure P , is initialized^{48 49}. P is essentially a list of lists, each sub-list containing the surface and volume points belonging to a single pocket, P_j . Also a tree data structure T , whose nodes will be pockets, is initialized.

In the second phase of the algorithm, each point in the sorted list is examined in turn, starting with the point with the greatest Travel Depth. For each point, i , there are three possible cases:

- i. The point i has no neighbors already in P . In this case, a new pocket P_j is added to P , the point i is added to P_j 's list of points, and a new leaf node P_j added to the tree T . The depth of point i will be the maximum depth of the new pocket.
- ii. The point i has neighbor(s) in only one pocket of P , P_k . The point is added to P_k 's list of points.
- iii. The point i has neighbors in two or more pockets in P , say pockets $P_j \dots P_k$. The point i and the point lists of all sub-pockets $P_j \dots P_k$ are added into the point list of a new pocket P_l . The pocket P_l is added as a new node in T , and the existing sub-pockets nodes $P_j \dots P_k$ are indexed as descendents of P_l . The depth of point i will be simultaneously the minimum depth of all the pockets $P_j \dots P_k$ and the height of the deepest saddle point connecting these sub-pockets.

In summary, in this phase of the algorithm there are three possible operations: i) finding a new pocket, ii) adding to an existing pocket, iii) merging pockets. An example of each kind of point is shown in Figure 1 as stars on points representing each case by color: i) yellow ii) cyan iii) magenta.

Once all points have been examined, the points in all the top level pockets of T are unioned into a final mother of all pockets which forms the root of T . This pocket contains all parts of the molecular surface that lie within the convex hull, and the entire intermediate volume.

The result of the algorithm is therefore a complete tree of pockets, T . Each node of T is a pocket, and each pocket contains all the volume and surface points of each of its descendent pockets, plus points specific to itself, i.e. the smaller pockets are nested inside the larger pockets. Every molecular surface point and intermediate volume point has been assigned to a pocket and hence to all antecedents of that pocket. Each saddle point has been assigned to two or more pockets, and the resulting merged pocket. Each leaf node of this tree represents a pocket containing a single local maximum in Travel Depth, i.e. a simple pocket. As we ascend the tree, the pockets become increasingly larger and more complex, with multiple local maxima in depth (sub-pockets), i.e. they are compound pockets. The mouth or mouths of a given pocket are defined as the union of surface and volume points belonging to that pocket which are on its boundary. i.e. that have at least one neighbor that is *not* in that

pocket. Each pocket has other associated shape, physical and protein related properties as described in the next section.

Pocket Collation

To facilitate collation, filtering, comparison and clustering of pockets, various features or metrics of each pocket are computed.

First are the global geometric features: volume, surface area, and principal axis dimensions. Second are the mouth geometric features: number of mouths, mouth area(s), and largest mouth linear dimension(s). Third are residue based properties: Lists of residues lining the entire pocket and lining the mouth. Fourth are physico-chemical properties: including surface area of positively charged, negatively charged, or neutral (apolar) atoms. Fifth are secondary surface properties: mean curvature and mean absolute curvature (roughness). The sixth set of properties, unique to this work, are Travel Depth related: height (maximum Travel Depth – minimum Travel Depth), mean height (mean Travel Depth – minimum Travel Depth), absolute maximum Travel Depth.

Curvatures are computed by analyzing the angle between adjacent triangles of the surface, and these are mapped from edges to points by weighting according to the length of the edge. This gives local curvatures, not regional curvatures as computed by other methods⁵⁰. The mouth linear dimension and pocket dimensions are computed by finding the principal components⁵¹ of the mouth or pocket points and then measuring the distance along each dimension. The pocket principal dimensions could be considered similar to finding the global fit of an ellipsoid through all pocket surface points to judge how open the pocket is, similar to previous work that fit spheres^{50, 52}. Partial charges are assigned using the PARSE parameter set⁵³, using a cutoff of -0.45 and 0.45 to determine polarity of lining atoms.

These pocket properties are principally designed for quantitative comparison of pockets, as described in the next section. We note that these features could also be used to automate the qualitative classification into pocket types, i.e. bottlenecks, clefts, tunnels, etc based on ratios of appropriate metrics, although we don't pursue that application here.

Another use for these metrics is to identify biological activity associated with various pockets. This would include assessing the likelihood the pocket is an active site, or if the pocket is druggable. This application will be pursued in future work.

Pocket Comparison

To compare the shape of two pockets using either the actual surfaces or lining residue positions requires first, that the surface points or residue atoms of the two pockets be put into a 1-1 correspondence (aligned). The two objects are then overlaid using rigid body superposition, to yield the minimum root mean square deviation (rmsd) for that set of pair alignments. Since it may not be *a priori* evident which parts of each pocket correspond with the other, especially in pure shape matching, many alternate alignments may have to be considered until the global minimum rmsd is found. An alternative is to examine motifs of lining atoms or residues, which may generate thousands of descriptors which have to be matched. Thus pocket shape comparison using positional alignment or indirect lining residue information is fraught with difficulty. In this work each pocket is described by a modest number of shape descriptors, and our goal is to use these descriptors to quantitatively compare pockets avoiding the aforementioned alignment problem.

Since the numerical range and units of each descriptor differ widely, we first express them in dimensionless, normalized units using the information contained in the pocket tree(s), as follows. For the protein or set of proteins of interest, and their resulting pocket trees we first

select all the relevant shape descriptors for the particular application. The mean and standard deviation of each descriptor is calculated over all these trees. Each descriptor for the two pockets to be compared is turned into a Z-score by subtracting the mean (for that descriptor) and dividing by the standard deviation (again for that descriptor). Each pocket now has an n-dimensional vector of Z-scores where n is the number of descriptors. The rectilinear, or 'Manhattan' distance in shape space between two pockets P_i and P_j is defined as

$$D_{ij} = \sum_{m=1}^n |Z_i^m - Z_j^m| \quad (1)$$

where Z_i^m is the Z-score of the m'th descriptor of the pocket i .

The default set of descriptors used for shape comparison in this work are: volume, surface area, height, mean height, mean curvature, principal dimensions, number of mouths, mean mouth area and mouth longest dimension.

Use of Z-scores removes differences in numerical range and units for each descriptor and gives each descriptor equal weight in the final analysis. So for example a difference in surface area equal to one standard deviation over the set of all pockets is the same as a difference in one standard deviation in volume. This method of pocket shape comparison requires no alignment, and hence is extremely rapid. It does however use the descriptors as a proxy for full shape comparison. False negative type errors are demonstrably small: If two pockets are significantly different in a single descriptor, say volume or height, then they really must be different. Conversely, if two pockets are similar in all descriptors, and the descriptors are well chosen to represent non-redundant aspects of shape, it is highly likely that they truly are similar in shape and size. However, it does not preclude the possibility that the pockets differ in some aspect of shape that is not measured by the descriptors, so false positive type errors are possible. Using visual examination of many dozens of pairs of matched pockets we found no egregious examples of this error, so we judge it uncommon enough to consider this method of shape comparison robust.

To estimate the descriptor means and standard deviations to compute Z-scores we use the population of pockets for the protein or protein trees under comparison. An alternative approach to this internal standard would be means and standard deviations calculated from a suitable 'standard set' of protein structures. This choice of reference will likely have little effect as the means and standard deviations of the many shape descriptors across several of our data sets were found to be very similar.

Selecting Unique Pockets

For various applications, it is useful to have a measure of pocket uniqueness. This was calculated by comparing each pocket in a given tree to all other pockets in that tree that did not have any lining atoms in common. The distances between the pocket of interest P_j and all m non-overlapping pockets P_i are computed, and the uniqueness score of P_j is defined as

$$R_j = 1 / \sum_{i=1}^m 1 / D_{ij} \quad (2)$$

the mean of the reciprocal distances. Unique pockets will have a low value of redundancy, R , since there will be no pockets close in shape space. Conversely, pocket types seen frequently (like small dimples occurring between two or three neighboring non-bonded atoms) will have a high value of R . The uniqueness score allows one to filter out 'uninteresting pockets' to focus on ones that have a unique shape and that are therefore more likely to support specific ligand binding.

The uniqueness score is most useful for pockets lower on the pocket tree, where there are many non-overlapping pockets to compare. Pockets very high up on the pocket tree contain large amounts of surface, and there will be few, perhaps no pockets without any atom overlap. These would correctly get low uniqueness scores, but only because the sample size is small. For this reason, in most applications one would only use a uniqueness score combined with some suitable upper volume bound.

The uniqueness filter step in our algorithm takes the place of filtering strategies or parameter variation employed by other methods to generate only the most interesting pockets or those likely to be active sites. The difference is that here all pockets of interest are already contained in the complete pocket tree, so if a particular filtering step does not pick out the required pockets, one can re-examine the complete list.

Clustering and Ordering Pockets

With a well-defined pocket-pocket distance in shape space it is straightforward to cluster trees of pockets using standard clustering algorithms. To get useful clustering, however, we add the uniqueness score R as a penalty into the distance formula. This penalizes common uninteresting pockets such as dimples, which would otherwise dominate the clustering. The term in the penalty function used for clustering, due to a pocket pair A-B is

$$\frac{1}{D_{AB}} - (R_A + R_B) \quad (3)$$

where D_{AB} is the rectilinear distance in shape space between pockets A and B, and R_A and R_B are the two pockets' uniqueness scores. In clustering whole trees we also exclude pockets with volume less than $25\text{\AA}^3$. These are too small to be of any relevance. Pockets larger than $2000\text{\AA}^3$ are also not clustered explicitly, since they are invariably compound pockets that consist of multiple sub-pockets which are already included individually in the clustering operation. An upper cutoff is used primarily to avoid clustering the largest pockets that are near the root of the tree and represent most of the intermediate volume. For the protein families examined $2000\text{\AA}^3$ seemed to be a reasonable upper limit, though this parameter can be increased if the ligand is known to be very large.

For applications involving transitions along a single dimension (like a transition pathway or molecular dynamics run), we found it useful to create minimum spanning "lines". These are similar to minimum spanning trees^{49, 54} except the maximum degree of any node is 2 so when the minimum spanning line is fully constructed it gives a connected series from one end to another, each end being defined as having degree one. This is an approximation to the Traveling Salesman Problem⁴⁹, where the best solution is one that minimizes the total pocket-pocket distance while visiting each pocket exactly once.

Output files are created that can be used to visualize these clusters or minimum spanning trees in the graph drawing software packages GraphViz⁵⁵, 56 and aiSee⁵⁷. The aiSee version is annotated with snippets of code that can be used to quickly display the pockets of

interest in PyMOL⁵⁸, a common operation. Nodes can be colored according to which tree they belong to, or by the amount of residue overlap (ignoring ordering) of each pocket to all adjoining pockets.

Additionally, heatmaps of the pocket-pocket distance matrix can be created, which are useful for looking at the variation between sets of pockets of interest or among pockets from a single tree.

Pocket Selection

Once an entire tree of pockets has been collated, a common task will be to examine a pocket of interest. This can be done interactively with PyMOL⁵⁸ using our customized scripts. The tree can be followed up or down the branches to look at progressively larger or smaller pockets.

Another common task is to select a pocket or pockets based on a set of residues of interest. This is done most simply by computing a Tanimoto type overlap score: the size of the intersection of the list of residues of interest with the list of pocket lining residues, divided by the size of the union of the same two lists. Perfect overlap gives a score of one, no overlap gives zero. The pocket that maximizes the Tanimoto overlap score, T , is then picked. This part of the procedure is automated. The user can then use this pocket as a good starting point for an interactive search of related pockets up and down the tree using PyMol to refine the pocket selection for a specific application.

A more advanced pocket selection routine for a series of closely related pocket trees involves the following procedure. One initial pocket is selected from each tree based on residue overlap using the Tanimoto type score. All pocket-pocket distances for this pocket set are computed. The pocket with the greatest mean distance to all other pockets is removed, and all other pockets from the same tree with at least 0.5 in overlap to the removed pocket are examined to see which has the lowest mean distance to the other pockets remaining in the set. The one with the lowest mean distance is added to the set so there remains one pocket from each tree. This swapping operation is done iteratively until the pockets remain the same even after examining all pockets in descending order of mean pocket distance. The swapping optimization potentially involves a large number of steps so as a failsafe the procedure is terminated if a large cutoff number of swaps is reached, though this cutoff was not reached in our experiments. The swapping optimization leads to a consistent set of pockets along a transition pathway or a mutational series so the differences can be analyzed with minimal bias from the initial residue overlap selection step. We refer to this method of pocket selection as refined-Tanimoto.

The overall workflow used in a CLIPPERS analysis of one of more proteins is as follows

- Generation of pocket trees
 - Preparation of PDB file, removal of solvent entries
 - Generation of triangulated molecular surface (MS) using specified probe radius
 - Calculated of Travel Depth for each point on MS
 - Construction of pocket tree using travel depth
 - Calculation of pocket metrics, and annotation with lining atoms
- Display in PyMol
 - Display of depth coded MS

- Visual navigation up or down tree using Python interface module
- Clustering of pocket trees using AiSee
 - Generation of cluster diagram and display of tree in GraphViz
 - Selection of nodes from GraphViz to display in PyMol

Code for the pocket finding program, named CLIPPERS, is available at <http://crystal.med.upenn.edu/>.

Results and Discussion

We now present various application of the CLIPPERS program for finding and analyzing pockets. As part of this we include several important objective tests of CLIPPERS. First, we claim to generate pockets for every portion of the surface and therefore at least one pocket for any given bound ligand should exist. This is tested on a diverse set of structures with bound ligands, where the resulting pocket trees are searched for pockets that have a high Tanimoto score between the residues lining the pocket and the residues near each ligand.

Second, given a series of structural snapshots of a protein undergoing a transition between two very different conformations, one should be able to follow an evolving pocket through this transition pathway. More specifically, if the pocket shape distance measure is robust, distances between pockets in structures that are neighbors should be smaller than between non-neighbors. In other words, a complete reconstruction of the pocket ordering through the transition pathway should be possible from just the pocket-pocket distance matrix. This is tested in the section of the paper on adenylate kinase.

Finally, the ability to distinguish between pockets associated with less dramatic conformational change, such as those in protein tyrosine phosphatase 1b (ptp1b), can be tested by comparing the pocket-pocket distances between and within different conformations.

Comparison of binding site location in SURFNET, CAST and CLIPPERS

As a comparison to two other widely used approaches to finding pockets, we analyze a data set of 67 monomeric proteins with diverse enzymatic activity, originally compiled and analyzed using SURFNET¹⁰. SURFNET identifies all active sites at least partially, but we note that the algorithm has several parameters that were adjusted to get this recognition. This same data set was also used to test against CAST, though only 51 of the structures were used¹⁶. 14 structures were excluded since CAST could not analyze the known binding site since the discrete flow method could not find the pocket. Two other structures were eliminated in the original CAST work since they had been superseded in the PDB. We use the newer versions of these two structures here.

These 67 monomers were downloaded from the PDB⁵⁹. Waters were removed and ligands were separated for later analysis. Some complexes contained multiple ligands bound in spatially separated sites. These were split by clustering using a 5Å cutoff, resulting in 92 individual binding sites in these 67 structures. To correctly identify the ligand, special attention was paid to including non-standard residues with the protein and to identify peptide ligands correctly. Radii were assigned to the atoms using the radius set of Bondi⁶⁰, which is a standard set in the area of macromolecular analysis. SURFNET does not use an explicit probe sphere to construct the surface it uses. However CAST does use a solvent probe sphere, of radius 1.4Å. For CLIPPERS, we used a probe radius of 1.2Å as previously described⁴⁴. The choice of 1.2Å instead of the more common 1.4Å is for two reasons. First, using this surface construction algorithm and set of radii with the water sized probe radius of

1.4Å we find heavy atoms from ligands quite often penetrate the surface. When a radius of 1.2Å is used these heavy atoms are correctly placed outside the surface. Second, using a probe radius of 1.2Å allows the algorithm to define pockets for the smaller ions such as Mg^{2+} , which would otherwise be missed. Since the three methods have different methods of surface generation, and different radii sets, the surfaces will differ somewhat leading to minor differences in volumes and surface areas. This may contribute to differences in results, although the major effect is the method of pocket finding. In collecting pockets, a lower bound volume cutoff of $25\text{\AA}^3$ was used in CLIPPERS since this represents the volume of a typical heavy atom. This is the smallest pocket that could be considered relevant to molecular recognition, as one ion, water, or other heavy atom could fit into a dimple of that size. Since some structures had ligands in buried cavities, we included these cavities while computing the pockets, as described in the methods section. We note that several of these 67 structures have ligands binding in the non-physiological active site, and some of the active site ligands are much smaller than the actual substrate, as in PDB code 1PII⁶¹ which contains phosphates and not the entire substrate and 1ONC⁶² which contains a sulfate in the active site of an RNase, so while these are valid ligands for the test, they do not reflect accurately the physiological ligand.

Considering first the success rate in finding ligand binding pockets the mean number of pockets per protein generated by CLIPPERS for this dataset of 67 proteins is 431 ± 161 . Thus a large number of possible pockets are found covering the entire surface. To score these pockets, the set of residues within a cutoff of 5Å from any ligand atom is generated, and then the Tanimoto overlap score of this residue set with the lining residues of all the CLIPPERS pockets is computed. For all 92 ligands, at least one pocket is generated with a significant Tanimoto overlap, indicating 100% success in generating the binding pocket. Selecting the most overlapped pocket for each ligand, the mean Tanimoto overlap score over the 92 sites was 0.5 ± 0.2 , even though the set contained very exposed sites or sites that bound very small ligands like sulfate or phosphate. In other words using CLIPPERS there are enough pocket candidates generated that one finds on average a pocket that overlaps at least 50%, as identified by proximity to the ligand. This is in contrast to CAST, which fails completely in 14 cases to define the ligand binding pocket, since the discrete flow method cannot find pockets without bottleneck mouths. In examining all 92 pockets found for these ligands, we note that most cases of a low Tanimoto overlap are with ligands that are bound to a very shallow pocket near the convex hull of the protein. The pockets near such ligands tend to be less 'pocket' like. The Tanimoto overlap score can be less than 1 if either the pocket is too small or too large. One example is shown in Figure 2. The middle panel on the bottom row has a pocket far larger than one would expect, with a Tanimoto score of 0.25. Despite this poor overlap, CLIPPERS outperforms CAST which cannot find this ligand at all. To stress the point, this analysis was on an independent dataset first used in the SURFNET analysis and the methodology was not trained on this data at all, and given the caveats about some of the ligands in certain structures mentioned earlier, a mean Tanimoto score of 0.5 is good.

CAST also fails on the ligand in the upper right panel of Figure 2, which CLIPPERS finds easily. Interesting cases where $T\ll 1$ because the pockets are too large are shown in the upper middle and the lower right panels of Figure 2. Low Tanimoto scores for this reason are not necessarily bad: These pockets contain additional volume that could guide the design by medicinal chemists of more specific or higher affinity ligands by indicating areas where functional groups can be added. More generally, once having identified a ligand binding pocket, nearby pockets may be a good target for fragment based drug design⁶³⁻⁶⁹ or interaction sites for added groups. Since CLIPPERS inventories all the pockets and places them in a tree, it facilitates such an approach. For example one may easily search for 'siblings' pockets in the tree: Ones which are joined by the lowest barriers forming natural

routes across which the fragments would be joined. While CAST and SURFNET can sometimes identify these nearby pockets, only CLIPPERS identifies all such pockets and the saddle points joining them.

Comparing now the number, shape and size of pockets generated by the different methods, CAST typically generate tens of pockets per protein, SURFNET generates more, typically a hundred or so. CLIPPERS generates considerably more candidate pockets, usually several hundred per protein, and due to the hierarchical and inclusive way they are generated, smaller pockets are nested inside larger pockets, all the way down to the smallest dimple. Neither CAST nor SURFNET generates overlapping or nested pockets. Both methods also prune the number of possible pockets to focus on ones that hopefully include the site of interest. In SURFNET this is done by adjusting the parameters use in the sphere clustering method. In CAST this is done using the discrete flow technique to join the tetrahedra and decide where pocket mouths lie. However, in each method the number of pockets is well correlated with the protein volume, shown for CLIPPERS in Figure 3a. In the SURFNET study, only the volume of the biggest and second biggest clefts were compared the protein volume. As another comparison to CAST, we show that the pocket areas and volumes correlate linearly with total protein area and volume, respectively as shown in Figures 3b and 3c.

Analyzing the 92 ligand binding pockets further, we find, as does CAST, that there is no correlation of protein size with binding site pocket size, as measured either by volume or surface area (Figures 4a and 4b). The mean of various statistics of these 92 pockets is as follows: Volume: $530 \text{ \AA}^3$, Surface Area: $319 \text{ \AA}^2$, mean Travel Depth: $12.8 \text{ \AA}$, maximum Travel Depth: $17.2 \text{ \AA}$, height: $7.2 \text{ \AA}$, mean height: $2.8 \text{ \AA}$, mean curvature: 5 degrees, principal dimensions: $16.8 \text{ \AA}$, $11.6 \text{ \AA}$, $7.1 \text{ \AA}$, fraction apolar surface area, 0.31: fraction negative surface area: 0.25, fraction positive surface area: 0.44.

Analyzing the mouth statistics in CLIPPERS, there is only one cavity in the set of 92, 83 pockets have single mouths, 5 have 2 mouths, 1 has 3 and 2 have 4. The mean mouth area is $147.5 \text{ \AA}^2$ and the mean mouth longest dimension is $14.5 \text{ \AA}$. The relationship between mouth number and pocket volume is shown in Figure 4c, as in CAST there is a slight correlation with mouth number and volume. The relationship between mouth diameter and mouth area is shown in Figure 4d, a line representing a perfect circle is shown for reference. Most mouths show some deviation from this, many mouths tend to be longer in one dimension than would be expected of perfectly circular mouths, since our mouths are not constrained to be bottlenecks as in CAST. This makes sense as mouths of grooves or clefts would by nature be very elongated. The mouth diameter is measured from point to point and not necessarily along the Travel Depth isosurface representing the mouth. This explains the few mouths with diameters smaller than possible for two dimensional circles.

As an additional analysis, a subset of the ligands were taken in the molecular mass range of drug-like molecules (150 to 500 daltons)⁷⁰, 71, 43 of the 92 ligands met this criteria. These are shown in each panel of Figure 4 as a subset of the complete set of pockets, there are very few additional trends. The binding site pockets for drug-like ligands appear slightly smaller than for the general ligands and they almost always have 1 mouth (1 ligand was in a cavity, 1 was in a pocket with 2 mouths).

A major feature of the CLIPPERS program is improved visualization of pockets with PyMOL⁵⁸ using customized python scripts. Once pockets have been inventoried and the resulting pocket data file loaded, each pocket surface can be displayed and colored individually. Several examples are shown in a montage in Figure 2. The default coloring is by Travel Depth, but other coloring schemes include pocket size, curvature, electrostatic

potential and polarity. Another feature of CLIPPERS is that the lining atoms can be easily highlighted.

Adenylate Kinase Transition Pathway

Adenylate kinase undergoes a significant conformational transition between the open inactive form PDB code 4AKE⁷² and the closed active form PDB code 1AKE⁷³. This transition has been modeled by examining various crystal structures at the endpoints and in the middle of the transition pathway⁷⁴⁻⁷⁶, or by more extensive experiments^{77, 78}. We generated a full transition from the closed to open forms using Climber, a morphing method that takes into account the energy of each structure when determining the step size to the next structure⁷⁹. 82 structures were generated along the pathway and analyzed. The purpose of generating this transition pathway was twofold. First, to show how CLIPPERS can be used to track and examine pocket shape changes due to conformational changes. Second, to test the objectivity of the pocket-pocket distance function. Adjacent pockets in the pathway should have smaller separations in shape space than pockets further apart in the transition pathway. We do not claim that the intermediate conformations that are generated follow the actual transition pathway, nor that the intermediate pocket shapes accurately represent the active site pockets during the transition. We only required a smoothly varying, ordered set of conformations to check the pocket-pocket distance metric.

To select the initial CLIPPER pocket series a set of 41 lining residues around the active site pocket was chosen, based on the bound ATP-mimetic inhibitor, and the iterative Tanimoto overlap/swapping procedure described in Methods was used to pick a single pocket from each of the 82 structures. The resulting pairwise distance matrix was computed for this set of 82 pockets. We then used just this distance matrix, without reference to the known conformational sequence, to construct the minimum spanning lines of these pockets, i.e. the pocket sequence that minimized the total neighbor neighbor distance. We then compared this reconstructed sequence with the actual sequence through the transition pathway.

We varied the descriptors included in the distance function and the distance metric (Manhattan or Euclidean) to determine which gave us the best reconstructed pathway. The Manhattan metric provided the best results, along with the following 11 descriptors: 1) surface area 2) volume 3) height 4) mean curvature 5) mouths 6) longest dimension 7) middle dimension 8) short dimension 9) area of biggest mouth 10) diameter of biggest mouth 11) mean height. The reconstructed ordering of the minimum spanning line had a Spearman Rank Correlation Coefficient of 0.999 with the actual ordering, indicating almost perfect ordering. This is excellent considering the degree of similarity of many pockets to each other in the open form. These 11 descriptors and the Manhattan metric were used for all further pocket-pocket distance comparisons.

Using the refined Tanimoto pocket selection criteria that involve iterative swapping of pockets that have good residue overlap led to a transition pathway of pockets that was visually smooth and plausible (See the movie in supplementary material). With this sequence many useful pocket properties can now be tracked smoothly throughout the entire transition pathway, as shown in Figure 5.

The heatmap of the matrix of pocket-pocket distances across the entire transition pathway was computed (Figure 6). This representation confirms that adjacent pockets (near the diagonal) have low distances and pockets far away in the pathway have high distances. One interesting observation is that the open pockets are more similar to each other than the closed and intermediate pockets, indicating that for a given overall structural change, the pocket shape change goes through an obvious transition where most of the shape measures are changing rapidly. This transition of the pocket happens quickly upon moving from the

closed towards the open form. The detection and identification of this transition of the active site pocket as opposed to the entire structure could not be done with previous methods. Results obtained with another transition pathway generation method based on normal modes were similar⁸⁰.

Enzyme Pocket Shape

To test the ability of our shape comparison to discern differences between active site pockets of proteins we used a benchmark dataset used to train a geometric hashing comparison algorithm based on atoms in the active site³⁵. This data set contains 79 proteins from 13 diverse protein families. Although the activity and enzyme classification of these proteins within a family are identical, it is not necessarily true that binding pocket shapes within one class will be similar. In the original study describing this dataset, it was possible to cluster these binding sites into the correct classifications, but knowledge of the binding location was used. In contrast, in the test here of CLIPPERS, no prior information about the binding site was used in the clustering. Instead, each family was examined in turn to see if pocket shapes cluster together. Then these clusters were examined to see if they corresponded to the ligand binding sites.

First protein structures were downloaded, waters and ligands removed, and nonstandard amino acids preserved. Then CLIPPERS was run on each protein to inventory the pockets. To compute shape descriptor Z scores the means and standard deviations of the descriptors here taken from the 67 proteins in the SURFNET/CAST dataset. For clustering, the penalty score given by Eqn. 3 was used. Again no sequence or structural alignment of pockets was necessary. Clustering of the complete pocket trees of two or more proteins in a family proceeds by comparing pair wise pocket distances and connecting pocket pairs with an edge if their similarity rises above some threshold. As each edge is added to the clustering, the residue overlap (ignoring residue order) of each distinct cluster was computed, along with how many structures have a pocket in that cluster. We search this output for clusters that have pockets representing all structures and that have the highest residue overlap score otherwise. Each pocket in the cluster is then examined to see if it corresponds to the ligand binding site. The results are summarized in Table 1. The total number of connections used up to the point where that cluster is created is also reported.

Analysis of eight of the thirteen families are complete successes; the cluster with at least one pocket from each structure with the highest residue overlap contains pockets representing each individual binding site. Though the residue overlap scores may not seem high, considering that mutations and size variation among pockets will affect this score, they are reasonable in the successful cases. In these cases we presume the shape and enzymatic activity are linked and note that the relatively simple scoring system of finding the cluster with at least one pocket from each structure with the highest residue overlap is sufficient to identify the binding sites for all such structures. An example is shown for the phosphoglycerate mutase family, the clusters formed at the threshold indicated in Table 1 are shown in Figure 7. The cluster with one pocket from each of the 4 structures is in the middle top and is colored red since it is conserved. The other red cluster to the right is a slightly larger site that is not represented in all structures at this clustering threshold, the one to the left is a conserved site between domains only present in 2 of the 4 structures.

The cases where this simple scoring scheme fails to identify a cluster of active site pockets were examined further. In the set of ten serine/threonine kinases, no cluster with a nonzero overlap score containing pockets from all ten structures existed, and the highest scoring cluster with nine structures represented was not the binding site. The highest scoring cluster with four structures represented does indeed contain the binding sites from those structures, and has a residue overlap of 0.487. Further examination of the structures shows that the ones

found represent an open conformation of the binding site, leading to a wide-mouthed but similarly shaped pocket in those four structures. Other structures in the set of ten for this enzyme class have a more closed conformation of the binding site, leading to a very different pocket with a much smaller mouth. So these would not show up in a clustering scheme based on shape, even though they do cluster together when residue type, position and known binding site location are the basis of the clustering³⁵.

The class of tyrosine kinases is represented by only two structures in this dataset. The highest overlap cluster found only contains one binding site, clustered with a similarly shaped cleft in the other structure. As there are only two structures in this cluster, these small clusters of just two pockets are very common and drown out possible clusters where many similar pockets all cluster together that have lower residue overlap due to size differences. Also, the binding site shapes of these two structures are somewhat different, while both bind in a cleft with an open mouth, one structure has two very deep lobes that extend beyond the volume taken up by the ligand, the other structure has much less volume below the ligand and no lobes. So, while CLIPPERS and the simple scoring scheme fail to identify the binding site here, this is not unreasonable as the binding site is not similarly shaped.

Methionine gamma-lyase has only two structures in this dataset, a cluster is found containing perfect residue overlap between 8 residues, however this is not the binding site cluster. D-xylose isomerase has many more structures but again suffers a similar result, the highest residue overlap cluster is not the binding site. Both these classes share some features: they have multiple binding sites per structure, their ligands are small, their binding sites are very deep and limited to just the volume near the ligand. In the methionine gamma-lyase structures, the probe radius of 1.2Å used appears to be a bit small and some of the ligand is inside the surface, this is possible where the surrounding protein is packed very closely to many parts of the ligand. Having multiple binding sites per structure, and having those binding sites be very small bottleneck pockets means they will be penalized highly by our redundancy clustering scheme.

The class of D-Glutamate ligase MurD contains five structures and the cluster with the most overlap is a conserved shallow dimple on the surface. The cluster ranked 3rd by residue overlap contains the five correct binding site pockets in a cluster of size 54 but with a low residue overlap of 0.211. Since there are 54 pockets in this cluster, and they are of variable size, the union of their residue counts is from the largest pocket while the intersection of the residue counts is from the smallest pocket, accounting for the very low overlap score.

Overall, while other methods can completely cluster these classes correctly³⁵, they have prior knowledge of the binding site location. Without binding site location, CLIPPERS correctly clusters the shapes of about two thirds of the classes. The other classes present a challenge for any shape based comparison. While binding and functional site location is not the major motivation for developing CLIPPERS, we note that the successes here show promise that additional methods or a better clustering and scoring system could prove useful. Regardless, similarly shaped binding sites can be identified using the clustering and scoring system, a useful test of the pocket similarity and redundancy formulas.

PTP-1b

As a closer examination of the ability of CLIPPERS to discern differences among related conformations and bound states, four structures of PTP-1b were examined. These four structures all have bound inhibitors that exploit both the main site and the nearby secondary pocket, though the structures are in different conformations. Two structures have a closed active site: PDB codes 1PTY⁸¹ and 1Q1M⁸². Two structures have an open active site: PDB

codes 1NNY⁸³ and 1ONZ⁸⁴. The four pockets are shown in Figure 8. The open ones have higher volumes, surface areas, bigger mouths and longer first principal dimensions than the closed sites, the other shape descriptors do not vary much. 31 residues were chosen that lie near any of the ligands in the structures and used to choose pockets for comparison. The table of distances is shown in Table 2 for both the standard method of picking based on residue Tanimoto overlap and the refined Tanimoto method. When the chosen pockets are compared, the open and closed states are discriminated according to their pocket-pocket distances, i.e. the two distances between closed pockets, or between open pockets are less than any of the four open-to-closed pocket distances. When the refined Tanimoto method is used, slightly larger pockets are chosen for all four structures, indicating that the smaller pockets are less alike than the larger pockets that contain them. The fraction of apolar surface area of these pockets is between 0.3 and 0.4, confirming analysis that the PTP-1b site is not very druggable¹⁴ and that the site is hard to search for using a formula based on finding hydrophobic concave regions²¹.

MDM2-p53

To additionally highlight a challenging, very open protein-peptide pocket, the MDM2-p53 interaction⁸⁵ from PDB code 1YCR was examined. By treating the p53 peptide (which consists of a single alpha helix) as a ligand and with the same procedure of finding nearby residues and using residue overlap to find pockets in the MDM2 protein, a representative pocket was found with 21 of the 23 nearby residues (and 3 residues not found nearby). This pocket is consistent with the original analysis⁸⁵, it has a large surface area and volume ($454\text{\AA}^2$ and $682\text{\AA}^3$ respectively), has a large mouth (diameter of $23\text{\AA}$), and is relatively shallow and flat (mean height of $2.3\text{\AA}$, height of $7.9\text{\AA}$). This pocket would be impossible to find using methods that look for bottleneck mouths.

Future Work

We present here three basic pocket finding/analysis operations now possible with CLIPPERS: identifying all pockets, calculating pocket shape properties, and comparing pockets by shape. These three basic operations, along with the hierarchical pocket tree data structure, can be used as elements of higher level more applied protein shape analyses. For instance pocket shape analysis can aid in functional site location and prediction⁵, in finding druggable binding sites^{14, 21, 52} or especially druggable binding spots in protein-protein interfaces⁸⁶⁻⁹⁰, in finding sites amenable to fragment based drug design⁶³⁻⁶⁹ and in identifying transient pockets as proteins undergo motions⁹¹.

The influence of pocket shape on chemical shape space and ligand shape space is obviously important as well. Complete classification of protein pocket shapes can assist or provide guidance in these areas⁹²⁻⁹⁸. Allosteric site discovery^{21, 99, 100} is another very important application for pocket analysis.

The approach used for characterizing pockets could also be easily integrated with tools like multiple sequence alignment, so residue overlaps could be scored based on alignment profiles rather than residue identity scores with a single sequence. This would be expected to improve the pocket classification of multi-protein families.

Conclusion

CLIPPERS is a new computational technique capable of cataloging all the potential pockets on a protein surface, and this cataloging is done without any tunable parameters or user intervention. CLIPPERS passes three objective tests, first, it always finds a pocket with a good residue Tanimoto score to known bound ligands in a diverse test set of proteins, giving

a mean score of $T=0.5$. Second, it can reconstruct the ordering of pockets formed along a transition pathway purely from the pocket-pocket distances, as shown for the adenylate kinase transition pathway. Finally it gives lower pocket-pocket distances within groups of similar structures than between them, as is shown with PTP1B.

CLIPPERS provides excellent visualization and characterization of pocket shape through customized PyMOL scripts⁵⁸ and output of many shape features, including the difficult volume and mouth descriptors. CLIPPERS computes pocket-pocket distances without doing computationally expensive shape based alignments. It can cluster pockets according to shape and uniqueness to visualize the possible interacting pockets on a set of protein surfaces. It also facilitates the discovery of neighboring pockets. Since CLIPPERS provides a complete inventory of pockets, it is especially suited to feed downstream applications that need a list of pockets as their starting point, such as docking.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

RGC and KAS thank Dr. Lyle Ungar for useful discussions about comparing trees. RGC received partial funding for this project from the NIH Structural Biology Training Grant GM008275 and the Computational Genomics Training Grant T32HG000046 from the National Human Genome Research Institute. Funding for KAS and RGC was provided by NIH GM48130.

References

1. Alvarez, J.; Shoichet, BK. Virtual Screening in Drug Discovery. Taylor & Francis; New York: 2005.
2. Kortagere S, Krasowski MD, Ekins S. The importance of discerning shape in molecular pharmacology. Trends in Pharma Sci 2009;30:138–147.
3. Nayal M, Honig B. On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites. Proteins: Struct Funct Bioinf 2006;63:892–906.
4. Levitt M, Park BH. Water: now you see it, now you don't. Structure 1993;1:223–226. [PubMed: 8081736]
5. Campbell SJ, Gold ND, Jackson RM, Westhead DR. Ligand binding: functional site location, similarity and docking. Curr Opin Struct Biol 2003;13:389–395. [PubMed: 12831892]
6. Nobeli I, Favia AD, Thornton JM. Protein promiscuity and its implications for biotechnology. Nat Biotech 2009;27:157–167.
7. Kleywegt GJ, Jones TA. Detection, Delineation, Measurement and Display of Cavities in Macromolecular Structures. Acta Cryst D 1994;50:178–185. [PubMed: 15299456]
8. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule-ligand interactions. J Mol Biol 1982;161:269–288. [PubMed: 7154081]
9. Laskowski RA. SURFNET: A program for visualizing molecular surfaces, cavities and intermolecular interactions. J Mol Graph 1995;13:323–330. [PubMed: 8603061]
10. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. Protein clefts in molecular recognition and function. Prot Sci 1996;5:2438–2452.
11. Brady GP Jr, Stouten PFW. Fast prediction and visualization of protein binding pockets with PASS. J Comput-Aided Mol Des 2000;14:383–401. [PubMed: 10815774]
12. Zhong S, MacKerell AD Jr. Binding Response: A Descriptor for Selecting Ligand Binding Site on Protein Surfaces. J Chem Inf Model 2007;47:2303–2315. [PubMed: 17900106]
13. Harris R, Olson AJ, Goodsell DS. Automated prediction of ligand-binding sites in proteins. Proteins: Struct Funct Bioinf 2008;70:1506–1517.

14. Cheng AC, Coleman RG, Smyth KT, Cao Q, Soulard P, Caffrey DR, Salzberg AC, Huang ES. Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotech* 2007;25:71–75.
15. Weisel M, Proschak E, Schneider G. PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem Central J* 2007;1:7.
16. Liang J, Edelsbrunner H, Woodward C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Prot Sci* 1998;7:1884–1897.
17. Peters KP, Fauck J, Frommel 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. [PubMed: 8609611]
18. Xie L, Bourne PE. A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. *BMC Bioinformatics* 2007;8:S9. [PubMed: 17570152]
19. Hendlich M, Rippman F, Barnickel G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graphics Modell* 1997;15:359–363.
20. Kalidas Y, Chandra N. PocketDepth: A new depth based algorithm for identification of ligand binding sites in proteins. *J Struct Biol* 2008;161:31–42. [PubMed: 17949996]
21. Coleman RG, Salzberg AC, Cheng AC. Structure-Based Identification of Small Molecule Binding Sites Using a Free Energy Model. *J Chem Inf Model* 2006;46:2631–2637. [PubMed: 17125203]
22. Lichtarge O, Sowa ME. Evolutionary predictions of binding surfaces and interactions. *Curr Opin Struct Biol* 2002;12:21–27. [PubMed: 11839485]
23. Ming D, Cohn JD, Wall ME. Fast dynamics perturbation analysis for prediction of protein functional sites. *BMC Struct Biol* 2008;8:5. [PubMed: 18234095]
24. Glaser F, Morris RJ, Najmanovich RJ, Laskowski RA, Thornton JM. A Method for Localizing Ligand Binding Pockets in Protein Structures. *Proteins: Struct Funct Bioinf* 2006;62:479–288.
25. Chen BY, Bryant DH, Fofanov VY, Kristensen DM, Cruess AE, Kimmel M, Lichtarge O, Kavraki LE. Cavity Scaling: Automated Refinement of Cavity-Aware Motifs in Protein Function Prediction. *J Bioinf Comp Biol* 2007;5:353–382.
26. Joughin BA, Tidor B, Yaffe MB. A computational method for the analysis and prediction of protein:phosphopeptide-binding sites. *Prot Sci* 2005;14:131–139.
27. Pettit FK, Bare E, Tsai A, Bowie JU. HotPatch: A Statistical Approach to Finding Biologically Relevant Features on Protein Surfaces. *J Mol Biol* 2007;369:863–879. [PubMed: 17451744]
28. Barber, C.; Dobkin, D.; Huhdanpaa, H. The Quickhull Algorithm for Convex Hull. *Geometry Center Technical Report GCG53*; Univ. of Minnesota, MN: 1993.
29. Coleman RG, Sharp KA. Travel Depth, a New Shape Descriptor for Macromolecules: Application to Ligand Binding. *J Mol Biol* 2006;362:441–458. [PubMed: 16934837]
30. 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. [PubMed: 12381328]
31. Kuhn D, Weskamp N, Schmitt S, Hüllermeier E, Klebe G. From the Similarity Analysis of Protein Cavities to the Functional Classification of Protein Families Using Cavbase. *J Mol Biol* 2006;359:1023–1044. [PubMed: 16697007]
32. Weskamp N, Hüllemmeier E, Klebe G. Merging chemical and biological space: Structural mapping of enzyme binding pocket space. *Proteins: Struct Funct Bioinf* 2009;76:317–330.
33. Powers R, Copeland JR, Germer K, Mercier KA, Ramanathan V, Revenz P. Comparison of Protein Active Site Structures for Functional Annotation of Proteins and Drug Design. *Proteins: Struct Funct Bioinf* 2006;65:124–135.
34. Petsalaki E, Stark A, García-Urdiales E, Russell RB. Accurate Prediction of Peptide Binding Sites on Protein Surfaces. *PLoS Comp Biol* 2009;5:1–10.
35. Kinnings SL, Jackson RM. Binding Site Similarity Analysis for the Functional Classification of the Protein Kinase Family. *J Chem Inf Model* 2009;49:318–329. [PubMed: 19434833]
36. Caffrey DR, Lunney EA, Moshinsky DJ. Prediction of specificity-determining residues for small-molecule kinase inhibitors. *BMC Bioinformatics* 2008;9

37. Gold ND, Jackson RM. SitesBase: a database for structure-based protein– ligand binding site comparisons. *Nuc Acid Res* 2006;34:231–234.
38. Kinjo AR, Nakamura H. Comprehensive Structural Classification of Ligand-Binding Motifs in Proteins. *Structure* 2009;17:234–246. [PubMed: 19217394]
39. Rosen M, Lin SL, Wolfson H, Nussinov R. Molecular shape comparisons in searches for active sites and functional similarity. *Prot Eng Des Sel* 1998;11:263–277.
40. Xie L, Wang J, Bourne PE. In Silico Elucidation of the Molecular Mechanism Defining the Adverse Effect of Selective Estrogen Receptor Modulators. *PLoS Comp Biol* 2007;3:e217.
41. Xie L, Bourne PE. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile–profile alignments. *Proc Natl Acad Sci USA* 2008;105:5441–5446. [PubMed: 18385384]
42. Halgren TA. Identifying and Characterizing Binding Sites and Assessing Druggability. *J Chem Inf Model* 2009;49:377–389. [PubMed: 19434839]
43. Hermann JC, Marti-Arbona R, Fedorov AA, Fedorov E, Almo SC, Shoichet BK, Raushel FM. Structure-based activity prediction for an enzyme of unknown function. *Nature* 2007;448:775–779. [PubMed: 17603473]
44. Coleman RG, Sharp KA. Finding and Characterizing Tunnels in Macromolecules with Application to Ion Channels and Pores. *Biophys J* 2009;96:632–645. [PubMed: 18849407]
45. Connolly ML. Solvent-accessible surfaces of proteins and nucleic acids. *Science* 1983;221:709–713. [PubMed: 6879170]
46. Dijkstra EW. A note on two problems in connexion with graphs. *Numerische Mathematik* 1959;1:269–271.
47. Coleman RG, Sharp KA. Shape and Evolution of Thermostable Protein Structure. *Proteins: Struct Funct Bioinf* 2010;78:420–433.
48. Tarjan RE. Efficiency of a Good But Not Linear Set Union Algorithm. *J Assoc Comput Mach* 1975;22:215–225.
49. Cormen, TH.; Leiserson, CE.; Rivest, RL.; Stein, C. *Introduction to Algorithms*. 2nd. McGraw-Hill Higher Education; 2001.
50. Coleman RG, Burr MA, Souvaine DL, Cheng AC. An intuitive approach to measuring protein surface curvature. *Proteins: Struct Funct Bioinf* 2005;61:1068–1074.
51. Pearson K. On lines and planes of closest fit to systems of points in space. *Philos Mag* 1901;2:559–572.
52. Cheng AC. Predicting Selectivity and Druggability in Drug Discovery. *Ann Rep Comp Chem* 2008;4:23–37.
53. Sitkoff D, Sharp K, Honig B. Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. *J Phys Chem* 1994;98:1978–1988.
54. Graham RL, Hell P. On the History of the Minimum Spanning Tree Problem. *Ann History Computing* 1985;7:43–57.
55. Gansner ER, Koutsofios E, North SC, Vo K. A technique for drawing directed graphs. *IEEE Trans Software Eng* 1993;19:214–230.
56. Ellson, J.; Gansner, ER.; Koutsofios, E.; North, SC.; Woodhull, G. *Graphviz and Dynagraph -- Static and Dynamic Graph Drawing Tools*. In: Junger, M.; Mutzel, P., editors. *Graph Drawing Software*. Springer-Verlag; 2004.
57. aiSee, 3.0.5. AbsInt; Saarbruecken, Germany: 2009.
58. DeLano, WL. *The PyMOL Molecular Graphics System*. DeLano Scientific; San Carlos, California: 2002.
59. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nuc Acid Res* 2000;28:235–242.
60. Bondi A. van der Waals Volumes and Radii. *J Phys Chem* 1964;68:441–451.
61. Wilmanns M, Priestle JP, Neirmann T, Jansonius JN. Three-dimensional Structure of the Bifunctional Enzyme Phosphoribosylanthranilate Isomerase : Indoleglycerolphosphate Synthase from *Escherichia coli* Refined at 2.0 Å Resolution. *J Mol Biol* 1992;223:477–507. [PubMed: 1738159]

62. Mosimann SC, Ardelt W, James MNG. Refined 1.7 Å X-ray crystallographic structure of P-30 protein, an amphibian ribonuclease with anti-tumor activity. *J Mol Biol* 1994;236:1141–1153. [PubMed: 8120892]
63. Allen KN, Bellamacina CR, Ding X, Jeffery CJ, Mattos C, Petsko GA, Ringe D. An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins. *J Phys Chem* 1996;100:2605–2611.
64. Teotico DG, Babaoglu K, Rocklin GJ, Ferreira R, Giannetti AM, Shoichet BK. Docking for fragment inhibitors of AmpC β -lactamase. *Proc Natl Acad Sci USA* 2009;106:7455–7460. [PubMed: 19416920]
65. Hubbard RE, Chen I, Davis B. Informatics and modeling challenges in fragment-based drug discovery. *Curr Opin Drug Discovery Dev* 2007;10:289–297.
66. Chen Y, Shoichet BK. Molecular docking and ligand specificity in fragment-based inhibitor discovery. *Nat Chem Biol* 2009;5:358–364. [PubMed: 19305397]
67. Verlinde CLMJ, Rudenko G, Hol WGJ. In search of new lead compounds for trypanosomiasis drug design: A protein structure-based linked-fragment approach. *J Comput-Aided Mol Des* 1992;6:131–147. [PubMed: 1624956]
68. Hubbard RE, Davis B, Chen I, Drysdale MJ. The SeeDs Approach: Integrating Fragments into Drug Discovery. *Curr Topics Med Chem* 2007;7:1568–1581.
69. Verdonk ML, Hartshort MJ. Structure-guided fragment screening for lead discovery. *Curr Opin Drug Discov Devel* 2004;7:404–410.
70. Irwin JJ, Shoichet BK. ZINC - A free database of commercially available compounds for virtual screening. *J Chem Inf Model* 2005;45:177–182. [PubMed: 15667143]
71. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 1997;23:3–25.
72. Müller CW, Schlauderer GJ, Reinstein J, Schulz GE. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure* 1996;4:147–156. [PubMed: 8805521]
73. Müller CW, Schulz GE. Structure of the complex between adenylate kinase from *Escherichia coli* and the inhibitor Ap5A refined at 1.9 Å resolution: A model for a catalytic transition state. *J Mol Biol* 1992;224:159–177. [PubMed: 1548697]
74. Beckstein O, Denning EJ, Perilla JR, Woolf TB. Zipping and Unzipping of Adenylate Kinase: Atomistic Insights into the Ensemble of Open $\leftrightarrow$ Closed Transitions. *J Mol Biol* 2009;394:160–176. [PubMed: 19751742]
75. Vonrhein C, Schlauderer GJ, Schulz GE. Movie of the structural changes during a catalytic cycle of nucleoside monophosphate kinases. *Structure* 1995;3:483–490. [PubMed: 7663945]
76. Beckstein O, Denning EJ, Woolf TB. The Closed $\leftrightarrow$ Open Transition of Adenylate Kinase From Crystal Structures and Computer Simulations [abstract]. *Biophys J* 2009;96:70a–71a.
77. Henzler-Wildman KA, Thai V, Lei M, Ott M, Wolf-Watz M, Fenn T, Pozharski E, Wilson MA, Petsko GA, Karplus M, Hübner CG, Kern D. Intrinsic motions along an enzymatic reaction trajectory. *Nature* 2007;450:838–844. [PubMed: 18026086]
78. Henzler-Wildman KA, Lei M, Thai V, Kerns SJ, Karplus M, Kern D. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature* 2007;450:913–916. [PubMed: 18026087]
79. Weiss DR, Levitt M. Can Morphing Methods Predict Intermediate Structures? *J Mol Biol* 2009;385:665–674. [PubMed: 18996395]
80. Yang Q, Sharp KA. Building alternate protein structures using the elastic network model. *Proteins: Struct Funct Bioinf* 2009;74:682–700.
81. Puius YA, Zhao Y, Sullivan M, Lawrence DS, Almo SC, Zhang ZY. Identification of a second aryl phosphate-binding site in protein-tyrosine phosphatase 1B: a paradigm for inhibitor design. *Proc Natl Acad Sci USA* 1997;94:13420–13425. [PubMed: 9391040]
82. Liu G, Xin Z, Hajduk PJ, Abad-Zapatero C, Hutchins CW, Zhao H, Lubben TH, Ballaron SJ, Haasch DL, Kaszubska W, Rondinone CM, Trevillyan JM, Jirousek MR. Fragment screening and assembly: a highly efficient approach to a selective and cell active protein tyrosine phosphatase 1B inhibitor. *J Med Chem* 2003;46:4232–4235. [PubMed: 13678400]

83. Szczepankiewicz BG, Liu G, Hajduk PJ, Abad-Zapatero C, Pei Z, Xin Z, Lubben TH, Trevillyan JM, Stashko MA, Ballaron SJ, Liang H, Huang F, Hutchins CW, Fesik SW, Jirousek MR. Discovery of a potent, selective protein tyrosine phosphatase 1B inhibitor using a linked-fragment strategy. *J Am Chem Soc* 2003;125:4087–4096. [PubMed: 12670229]
84. Liu G, Szczepankiewicz BG, Pei Z, Janowick DA, Xin Z, Hajduk PJ, Abad-Zapatero C, Liang H, Hutchins CW, Fesik SW, Ballaron SJ, Stashko MA, Lubben T, Mika AK, Zinker BA, Trevillyan JM, Jirousek MR. Discovery and structure-activity relationship of oxalylarylaminobenzoic acids as inhibitors of protein tyrosine phosphatase 1B. *J Med Chem* 2003;46:2093–2103. [PubMed: 12747781]
85. Kussie PH, Gorina S, Marechal V, Eilenbaas B, Moreau J, Levine AJ, Pavletich NP. Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science* 1996;274:948–953. [PubMed: 8875929]
86. Wells JA, McClendon CL. Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature* 2007;450:1001–1009. [PubMed: 18075579]
87. Zhong S, Macias AT, MacKerell AD Jr. Computational Identification of Inhibitors of Protein-Protein Interactions. *Curr Topics Med Chem* 2007;7:63–82.
88. Fuller JC, Burgoyne NJ, Jackson RM. Predicting druggable binding sites at the protein-protein interface. *Drug Discovery Today* 2009;14:155–161. [PubMed: 19041415]
89. Chen T, Kablaoui N, Little J, Timofeevski S, Tschantz WR, Chen P, Feng J, Charlton M, Stanton R, Bauer P. Identification of small-molecule inhibitors of the JIP–JNK interaction. *Biochem J* 2009;420:283–294. [PubMed: 19243309]
90. Dömling A. Small molecular weight protein–protein interaction antagonists—an insurmountable challenge? *Curr Opin Chem Biol* 2008;12:281–291. [PubMed: 18501203]
91. Eyrisch S, Helms V. Transient Pockets on Protein Surfaces Involved in Protein-Protein Interaction. *J Med Chem* 2007;50:3457–3464. [PubMed: 17602601]
92. Kahraman A, Morris RJ, Laskowski RA, Thornton JM. Shape Variation in Protein Binding Pockets and their Ligands. *J Mol Biol* 2007;368:283–301. [PubMed: 17337005]
93. Hert J, Keiser MJ, Irwin JJ, Oprea TI, Shoichet BK. Quantifying the Relationships among Drug Classes. *J Chem Inf Model* 2008;48:755–765. [PubMed: 18335977]
94. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. *Nat Biotech* 2007;25:197–206.
95. Rush TS III, Grant JA, Mosyak L, Nicholls A. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein-Protein Interaction. *J Med Chem* 2005;48:1489–1495. [PubMed: 15743191]
96. Stockwell GR, Thornton JM. Conformational Diversity of Ligands Bound to Proteins. *J Mol Biol* 2005;356:928–944. [PubMed: 16405908]
97. Koch MA, Schuffenhauer A, Scheck M, Wetzel S, Casaulta M, Odermatt A, Ertl P, Waldmann H. Charting biologically relevant chemical space: A structural classification of natural products (SCONP). *Proc Natl Acad Sci USA* 2005;102:17272–17277. [PubMed: 16301544]
98. Favia AD, Nobeli I, Glaser F, Thornton JM. Molecular Docking for Substrate Identification: The Short-Chain Dehydrogenases/Reductases. *J Mol Biol* 2007;375:855–874. [PubMed: 18036612]
99. Hardy JA, Lam J, Nguyen JT, O'Brien T, Wells JA. Discovery of an allosteric site in the caspases. *Proc Natl Acad Sci USA* 2004;101:12461–12466. [PubMed: 15314233]
100. Hardy JA, Wells JA. Searching for new allosteric sites in enzymes. *Curr Opin Struct Biol* 2004;14:706–715. [PubMed: 15582395]

Figure 1.

Top panel: Schematic protein, molecular volume shown in black, the convex hull shown as red lines. Pockets are labeled and the split line where two sub-pockets are joined is shown in green. See the methods section for a description of the algorithm and explanation of the cyan, magenta and yellow stars. Bottom panel: Corresponding pocket tree.

Figure 2.

9 example pockets found using CLIPPERS having the greatest Tanimoto score to ligand-neighboring residues. From left to right, top to bottom, the structures are PDB codes 1ADS, 1BYH, 1FUT, 1GPB, 1PDA, 1PPL, 1SMR, 1THG, 2CND. The protein is shown as grey lines, the ligand is shown in red sticks, the pocket is colored according to Travel Depth, figure created using PyMOL⁵⁸.

Figure 3. 67 protein structures⁹ analyzed using CLIPPERS. a) The protein volume compared with the total number of pockets with volume greater than 25 Å³. b) Protein volume compared to mean pocket volume for pockets with volume greater than 25 Å³. c) Protein surface area compared to mean pocket surface area for pockets with volume greater than 25 Å³.

Figure 4.

92 binding sites in 67 protein structures⁹ analyzed using CLIPPERS show as gray filled in squares. Outline black squares indicate 43 ligands with molecular mass in the drug-like range of 150 to 500 daltons^{70, 71}. a) Protein volume compared to binding site volume (on log scale). b) Protein surface area compared to binding site surface area (on log scale). c) Number of mouths compared to the binding site volume (on log scale). d) Mouth area compared to mouth diameter (on log-log scale). The line corresponds to perfect circles.

Figure 5.

Properties of the binding pocket tracked over the transition pathway between the closed and open adenylate kinase structures. a) Volume, Surface Area, and Area of Biggest Mouth. Additionally the RMSD of carbon alphas to either endpoint structure are shown, scaled to fit the graph. b) Height, Mean height, Diameter of Biggest Mouth, principal dimensions, all in Å, and mean Curvature in degrees.

Figure 6.

The differences between all 82 structures along the transition pathway. Upper left half: root pocket-pocket distance. Lower right half: binding site pocket-pocket distance. Note that the scale for the two comparisons is different.

Figure 7.

Clustering of 4 phosphoglycerate mutase pocket trees. Node boundary colors indicate trees. Node interiors colored by lining residue conservation (increasing from blue to white to red). Grey arrows indicate within tree connections. Black lines indicate between tree connections (shape similarity). Only the top 5000 connections are shown for clarity. Spatial layout of nodes is designed to place similar pockets close together. Visualization produced with aiSee⁵⁷.

Figure 8.

Pockets from the PTP-1b set of structures colored by Travel Depth, with the ligands shown in gray and the proteins shown in red. Top two panels: Open form. Bottom two panels: Closed form.

Table 1

Enzyme Shape Clustering

Enzyme Name	Total Connections	Cluster Size	Overlap	Structures in Cluster	Found Ligands	Total Structures
Aldose reductase	1823	130	0.368	8	8	8
Isocitrate dehydrogenase	3441	16	0.5	7	7	7
p-Hydroxybenzoate hydroxylase	640	15	0.73	7	7	7
Kinases (serine/threonine)	2859	51	0.019	9	0	10
Kinases (tyrosine)	166	2	0.5	2	1	2
Thymidylate kinase	6572	114	0.13	11	11	11
Subtilisin	243	2	0.763	2	2	2
Acid protease	694	26	0.725	7	7	7
Carbonic anhydrase	3140	16	0.444	6	6	6
Methionine gamma-lyase	6	2	1	2	0	2
D-Xylose isomerase	3801	32	0.115	8	0	8
Phosphoglycerate mutase	4990	31	0.507	4	4	4
D-Glutamate ligase MurD	166	11	0.917	5	0	5

Table 2PTP-1b Pocket-Pocket Distances^a

	1NNY	1ONZ	1PTY	1Q1M
1NNY		0.609	1.234	1.241
1ONZ	0.361		1.104	1.046
1PTY	0.679	0.785		0.414
1Q1M	0.689	0.856	0.281	

^aUpper right distances are for pockets selected by the direct Tanimoto method, lower left distances are for pockets selected by the refined Tanimoto method. Bold indicates distances within the set of closed or or the set of open conformations.