Published in final edited form as:
Proteins. 2010 March ; 78(4): 825–842. doi:10.1002/prot.22608.

A Novel and Efficient Tool for Locating and Characterizing Protein Cavities and Binding Sites

Ashutosh Tripathi and **Glen E. Kellogg**
Department of Medicinal Chemistry & Institute for Structural Biology and Drug Discovery, Virginia Commonwealth University, Richmond, Virginia 23298-0540 USA

Abstract

Systematic investigation of a protein and its binding site characteristics are crucial for designing small molecules that modulate protein functions. However, fundamental uncertainties in binding site interactions and insufficient knowledge of the properties of even well-defined binding pockets can make it difficult to design optimal drugs. Herein, we report the development and implementation of a cavity detection algorithm built with HINT toolkit functions that we are naming VICE (Vectorial Identification of Cavity Extents). This very efficient algorithm is based on geometric criteria applied to simple integer grid maps. In testing, we carried out a systematic investigation on a very diverse data set of proteins and protein-protein/protein-polynucleotide complexes for locating and characterizing the indentations, cavities, pockets, grooves, channels and surface regions. Additionally, we evaluated a curated data set of unbound proteins for which a ligand-bound protein structures are also known; here the VICE algorithm located the actual ligand in the largest cavity in 83% of the cases and in one of the three largest in 90% of the cases. An interactive front-end provides a quick and simple procedure for locating, displaying and manipulating cavities in these structures. Information describing the cavity, including its volume and surface area metrics, and lists of atoms, residues and/or chains lining the binding pocket, can be easily obtained and analyzed. For example, the relative cross-sectional surface area (to total surface area) of cavity openings in well-enclosed cavities is 0.06 ± 0.04 and in surface clefts or crevices is 0.25 ± 0.09.

Keywords

active site; cavity detection; binding pocket; surface area; buried volume; protein structure; molecular modeling; computer-aided drug design

Introduction

Modulation of the dynamics of a target protein binding site to elicit a pharmacological response is the major therapeutic approach for the treatment of a plethora of diseases. This is usually accomplished by developing small molecules that occupy a ligand recognition site. Drug development is a challenging process, owing to fundamental uncertainties in structural determination and associated issues such as structural and physicochemical characterization of the binding pockets, even under relatively static conditions such as in crystals subjected to x-ray analyses. Reliable, rational and efficient approaches to locating and characterizing the binding sites of protein and other bioactive molecules should be valuable in the design of new drugs.[1] In recent years there has been a surge in the number of crystal structures deposited in Protein Data Bank [2]. Concomitantly, NMR and X-ray crystallography have played an increasingly crucial role in drug discovery through structure based methods and virtual screening of extensive libraries of compounds. Facilitating this has been the design and development of many computational tools with a large range of functions. In particular,

a number of programs have been developed to *de novo* locate the binding pockets in proteins [1,3]. Such tools have provided valuable information for better understanding protein binding site architecture. However, the accurate identification and quantitation of binding pockets is not an entirely straightforward process, and the existing computational tools have numerous strengths and weaknesses.

Proteins have "pockets" for molecules to bind; however, these pockets may not be observed from an initial inspection. Protein surfaces are formed by numerous cavities and protrusions that are interlinked through small narrow channels and are often are interspersed with numerous holes or voids. The size and shape of these protein cavities dictates the three-dimensional geometry of ligands that will bind within, and guides the important intermolecular contacts that mediate this binding. Binding sites that are formed by several neighboring pockets/cavities and auxiliary pockets near the active site are often suggested as providing additional ligand binding surface, which adds further to the complexity. Efficient analysis of the shape and size of protein pockets and cavities thus becomes important as structural changes due to side-chain rotations and backbone movements, loop motion and/or ligand-induced conformational changes may significantly alter the topography of the active site. A thorough structural analysis of the target binding site is critical to propel a drug discovery project forward. There has been significant progress in this endeavor in recent years [1,3,4].

Theoretical approaches for locating binding sites on proteins

Identification and characterization of active sites is key in studying protein structure, particularly when designing molecules that interfere with function and modulate activity. There are a number of ways in which binding sites or cavities in proteins can be located and characterized, e.g., with several existing programs such as VOIDOO [5], LIGSITE [6], POCKET [7], POCKET-FINDER [8], CAST [9], PASS [10], APROPOS [11], SURFNET [12], Q-SITEFINDER [13], POCKETPICKER [14] and others. These programs can be generally classified into categories according to the approach they take to locate and characterize the cavity: i) evolutionary methods (structure/sequence alignments); ii) probe/energy based methods; and iii) geometric approaches.

Evolutionary methods use a heuristic approach to predict cavities in unknown proteins from known protein structures based on family and/or functional criteria. With the abundance of structural-and sequence-related data for many protein families, this approach has found increased application in finding and characterizing protein target binding sites [15,16]. Structural similarity and three-dimensional templates are used to compare and classify putative binding sites in uncharacterized protein structures with unknown functions, e.g., with similarity searches over functional site databases like LigBase [17] and INTERPRO [18] that detect functional similarity when homology is minimal. The approach by Bickel *et al.* [19] uses statistical methods to identify active sites by residue identity within and outside functional subfamilies. Programs like ConSurf [20] identify functional regions of proteins by surface mapping of phylogenetic information, while Rate4Site [21] applies evolutionary determinants in mapping the functional regions on a protein surface. These methods are likely to continually evolve with the increasing availability of structural and sequence data from structural genomics projects.

The idea of *in silico* mapping of protein surfaces was first conceptualized by Lee & Richards (1971) [22] based on the idea of an "accessible" surface area. Connolly (1982, 1983) [23] suggested the concept of "solvent excluded surface" and developed the eponymous algorithm for calculating molecular surfaces with a rolling spherical probe. Later, Kuntz *et al.* developed an algorithm that fills all pockets and grooves on the surface of receptor molecule with a set of balls [24]. While the probe sphere radius is generally 1.4 Å to

approximate a water molecule surveying the solvent accessible surface of the protein, this sphere radius can be varied to map other representations such as the van der Waals surface. Kuntz *et al*. used this approach to define the binding site in the first implementations of the DOCK automated docking program [25–26]. Another novel approach of using spherical probes on a regular Cartesian point grid was implemented by Peter Goodford in GRID [27] and by Martin Karplus in MCSS (multiple copy simultaneous searches) [28]. In GRID, a binding region on a protein is mapped by calculating interaction energies between a (functional) probe group placed at each grid point and the atoms of the protein. In MCSS, about 1000 to 5000 small functional groups (probes) are interacted with the protein surface simultaneously and energy minima are calculated to define favorable interaction sites. The generated functional maps of the binding site indicate the most favorable regions for placing ligand groups with properties similar to the probes. A number of cavity detection algorithms based on this approach have been reported: Voorintholt *et al*. adopted an approach where grids are used to store the distance to the nearest atom [29]; a similar approach was taken by Del Carpio *et al*. [30] in searching for pocket regions in a protein; the POCKET program by Levitt and Banaszak [7] uses a 3D Cartesian grid and spherical probes to map protein surfaces and pockets using a modification of the marching cubes algorithm; and the CHANNEL algorithm [31] uses a sphere of radius R to probe a node space that fills the unit cell of a crystal lattice.

Some probe/energy based approaches to detect cavities overlap with geometric approaches in that a probe of a specified volume is only used to exclude van der Waals overlap as the protein surface is surveyed. The VOIDOO program reported by Kleywegt and Jones [5] uses atom fattening or a flood fill algorithm on a regular 3D grid to locate and delineate cavities. Another method totally relying on geometric criteria is the PASS algorithm developed by Brady and Stouten [10] where the cavities in a protein are filled with a set of spheres. Cavity detection based on alpha shape theory [32–33] incorporates a different, perhaps purely algorithmic, approach. The Automatic PROtein Pocket Search (APROPOS) method developed by Peters, Fauck and Frömmel [11] is based on purely geometric criteria for finding binding sites using atomic coordinates. Atoms are represented as a set of points in 3D Euclidean space and the envelope or surface is derived by Delaunay triangulation [34] of these points. The alpha shape algorithm describes these surfaces as lists of adjacent triangles and, depending on the value of alpha, delineates the cavity shape. The program CAST developed by Liang and Woodward [9] also applies alpha shape principles along with discrete flow theory to determine the shape of the binding pocket as a negative image of cavity derived from Delaunay tetrahedrons [34]. Alpha shapes and Delaunay triangulations are rich in geometric information from which area and volumes of inaccessible cavities can be calculated.

Another such widely used algorithm for cavity detection is LIGSITE developed by Hendlich, Rippmann and Barnickel [6]. This algorithm is similar to POCKET, but circumvents many of its drawbacks: first, grid points within a protein atom's van der Waals sphere are discarded; next, the remaining lattice points are scored according to their degree of burial by scanning grid points along the three Cartesian axes and the four cubic diagonals; and finally, the area delineating these grid points is clustered to describe contiguous cavities. Similarly, Stahl *et al.* [35] described an algorithm based on "degree of buriedness" (accessibility). The accessibility of a grid point is found with a set of 45 points distributed evenly about a sphere centered on each grid point. Vectors projected from each grid point through all points on its sphere determine the point's accessibility depending on how many vectors pass through the van der Waals radii of a protein atom in 15 Å or less. Points whose accessibility value is below a threshold are clustered. Also similar is the approach of Schneider *et al.* (the PocketPicker algorithm [14]) that has a somewhat different algorithmic definition of buriedness and additionally creates a shape descriptor that enables comparison

of pocket shapes. Most of these algorithms can fairly easily locate moderately to well-defined binding pockets and can be used in combination with other drug design tools to provide valuable information for structure based drug design projects.

Vectorial Identification of Cavity Extents (VICE)

The present paper is in a series of articles describing our work in developing computational tools for drug design [36–37]. The development of the VICE cavity detection algorithm was initially motivated by our need for a tool that could be tightly integrated with other algorithms in the HINT toolkit suite [38]. While implementing VICE, we realized that, although there are quite a number of available cavity detection algorithms, most, if not all, of these programs have minor or major flaws. In particular: 1) many are not flexible enough to locate the wide variety of cavity and pocket shapes and sizes in which ligands can bind or with which proteins associate; 2) most do not have an adjustable and user-interpretable parameter for defining the cavity opening(s); 3) many programs fail to characterize unusual cavities like those in multi-domain channel and pore proteins; and 4) to our knowledge none of the programs provide what we consider to be a complete set of quantitative data describing the cavity.

In this paper we describe the new VICE computer algorithm for finding and delineating the active site in proteins or other biomacromolecules based on geometric criteria applied to simple integer grid maps using very minimal floating point mathematics. Like many of the algorithms described above, VICE uses a grid-based approach and immediately discards grid points which fall within the van der Waals radii of protein atoms. The remaining grid points are scored according to a metric roughly similar to degree of burial and VICE is thus similar to some of the methods listed in the previous section [6,14,35]. Our objective in this report is to find pockets and shallow binding regions that have the characteristics of receptor sites, identify the amino acid residues surrounding them, and calculate descriptive metrics regarding the sites. The algorithm was applied to a diverse set of over 60 proteins in order to locate, investigate and characterize their various kinds of cavities on proteins. This is a starting point towards comprehensive analysis of protein topography with respect to its function and an efficient and robust method for finding active sites that would be compatible with other tools and protocols we have developed based on our HINT empirical force field model [39–41].

Methods

The dataset of protein complexes in this study consisted of examples from the literature having binding pockets of diverse shapes, sizes and types. Table 1 lists the proteins evaluated by their PDB code and the associated cavity type for which the binding sites were calculated. All protein structure coordinates, in PDB format, were retrieved from the RCSB (Brookhaven) Protein Data Bank [2]. Molecular modeling was performed using the Sybyl 7.3 program suite (www.tripos.com) on Irix and Linux workstations. The protein structures were prepared for this study by removing all the water molecules, ions, and any cofactors associated with the structure. Hydrogen atoms were added to the structures using the “Add Hydrogens” tool within the Sybyl Biopolymer module before further analysis.

The cavity detection and analysis programs were constructed using subroutines from the HINT toolkit [38]. Several new subroutines were composed for 3D map manipulation and analysis. Of particular value were an enhanced suite of functions for Boolean maps (where each grid value can only be zero or one) that forms the basis of the search algorithm as described in the Results and Discussion section. The algorithm provides several user-adjustable options to optimize the cavity calculation. With these parameters it is possible to change the focus from the entire protein to a small region for a detailed investigation. For

the initial surveys in this study, the grid boxes were defined as the molecular extents of each biomacromolecule with a grid resolution of 1 Å and margin of 3–5 Å. Most importantly, the cavity definition ("cavityness") was set at 0.5, which is the fraction of vectors reaching a protein "wall" instead of the box edge (see Figure 1). The maximum unrestrained path-length (vector length) was set to 20 Å by default but was increased to 40–60 Å to explore very large cavities or channels. The minimum closed contour volume was set to 100 $Å^2$ to eliminate small clusters or irrelevant voids. The shaping factor was usually set to be 0.50, but was varied from 0.35 to 0.6 to interactively smooth some pockets that presented small and inaccessible sub-pocket regions. In the figures shown in this work, the surface of the pocket was displayed by contouring the cavity map at a value of 0.5, i.e., matching the cavity definition.

For the reevaluation of the bound/unbound data set of Huang *et al.* [42], a somewhat different set of parameters was used as we intended this investigation to proceed without parameter tinkering. Thus, the cavityness definition was set to 0.55, the maximum unrestrained path-length was set to 10 Å, the minimum closed contour volume was set to 150 $Å^2$ and the shaping factor was set to 0.60. All maps were created with 1.0 Å resolution with margins (exceeding the molecular extents) of 2.0 Å. As defined by Huang *et al*. [42], the cavity search is successful if any atom of the ligand is found within 4.0 Å of the cavity center; this is evaluated for the largest cavity (most stringent) and for any of the three largest cavities.

Results and Discussion

Protein binding regions provide a microenvironment for substrates, inhibitors, other proteins or biomacromolecules to interact and modulate the protein's activity. This paper describes a computational tool for locating and investigating the binding regions of protein from a standard PDB file. This section describes and illustrates the algorithm, outlines the quantitative cavity metrics that can be derived through this algorithm, and highlights in some detail several of the more than sixty cases we have used to validate the methodology for this work. The rather remarkable variation that is observed in shapes and sizes of binding cavities is evident even from this small number of examples.

The VICE Algorithm

The VICE (Vectorial Identification of Cavity Extents) algorithm is schematically illustrated in Figure 1. After the region of interest, which can be the entire target protein or portions thereof, is defined, a grid cage with user selectable resolution is created. While 1 Å resolution is typical, larger or smaller values may be appropriate depending on computational requirements. These requirements may include a very high resolution over a restricted spatial region for defining channels or a low resolution when surveying the entire extents of a very large protein. This degree of fine-tuning capability is an advantage over probe-based methods. The key advantage of this algorithm is that many of the calculations are performed on integers and on integer (Boolean) grid maps so that the method is very efficient. In the first step grid points occupied by atoms in the target molecule are set to zero, while those unoccupied are set to one. These latter points are potentially in the cavity; each will be examined by the algorithm. The search tools are sets of vectors whose directions are determined by the grid nodes (see Figure 1a). In the first shell the set of 2D vectors are {(1,0);(1,1);(0,1);(−1,1);(−1,0);(−1,−1);(0,−1);(1,−1)}, while in the second shell set the unique 2D vectors are {(2,1);(1,2);(−1,2);(−2,1);(−2,−1);(−1,−2);(1,−2);(2,−1)}. Each vector is projected until it reaches an edge of the grid box (Figure 1b) and the nodes that the vector passes through constitute a path list. This is a major difference between VICE and other methods using maps and vectors [14,35]: VICE deploys test vectors that are keyed by

grid box paths, not by compass directions; thus performing this critical part of the cavity search completely with faster integer (rather than floating point) arithmetic.

Each vector is classified through analysis of its path list (Figure 1c) as having: a clear path to edge, i.e., it does not pass through an occupied node; a blocked path; or is “stalled”, i.e., it has neither reached the box edge nor has it passed through an occupied node. These latter vectors are treated as having clear path; their purpose is to ameliorate the possibility that a very long vector may inadvertently pass through occupied nodes belonging to another biomacromolecular subunit or because of a slightly curved pocket entrance. The stalled vector length is a parameter that may be adjusted depending on the anticipated dimensions of the cavity. The fraction of vectors classified as blocked is evaluated for each grid point. Thus, each grid point is classified as “inside” or “outside” the putative cavity based on a parameter with nominal cutoff value of 0.5 (Figure 1d). A few grid points, mostly at the cavity mouth, are ambiguous (e.g., 0.5 ± 0.05); these are recalculated with additional shells of vectors and tightening criteria until a final disposition can be determined. This intuitive fraction is the defining parameter for the cavity entrance. With relatively small adjustments, the entrances to deeply buried pockets and shallow grooves can be detected. However, as illustrated in Figure 1d, openings to the cavity are not necessarily at “sea level” but are generally a more natural description of these openings that reference the cavity’s shape.

Two steps are applied to refine the cavity definition. First, narrow pseudo-channels, i.e., one grid node in width, and tendrils are eliminated by forcing a requirement that each “inside” point have a minimum of “inside” neighbors (Figure 1e). This can be applied recursively to “shape” the cavity. Lastly, to eliminate stray irrelevant pockets, each enclosed surface must have a minimum volume. While these steps can be performed automatically without user input, the algorithm is designed to display the intermediate raw maps and allow interactive application of the refinement.

Overview of Protein Structure Studies

We carried out a systematic investigation of VICE on a diverse set of proteins to locate and investigate cavities of different shapes and sizes on these proteins. The dataset consisted of examples of proteins from the literature having binding pockets of diverse shapes and sizes. All protein structure coordinates, retrieved from the RCSB (www.rcsb.org) [2], were prepared as described in the Methods section. Our test set included: 16 cases where the binding pocket is a well-defined, well-enclosed, deeply buried pocket; 9 cases where the cavity or groove is on the protein’s surface; 10 cases where the cavity is created by a protein–protein interface (more challenging since protein–protein dimers do not often show deep well-defined cavities that are putative binding sites for small molecules); 10 cases of cavities at DNA- or RNA-protein interfaces; 5 cases of protein structure pairs with very flexible binding pockets due to movements of flexible loops resulting in both open and closed cavities; 5 cases of proteins with channels or tunnels, i.e., ion channels, porins, and ligand gated channels; and lastly, 4 cases of proteins with multiple and/or allosteric sites including some with adjacent auxillary sub-pocket sites that may have additional biochemical roles. To our knowledge this is the most structurally challenging data set used to validate cavity detection software; it includes several proteins that have never been subjected to this type of analysis as well as a number that have been studied more than once.

A variety of metrics can be obtained or calculated for protein cavities. Of the most potential interest is the cavity volume that can be reported in terms of both its ligand-occupied and unoccupied fractions. Figure 2 illustrates how these metrics are calculated through manipulation of integer grid maps. We have also derived an automated algorithmic method (Figure 3) to estimate the cavity cross-sectional entrance areas. These volume and area metrics for the 64 biomacromolecules, some with multiple pockets or symmetry-related

sites, in this study are set out in Table 1. Lastly, identification of protein residues and/or atoms lining the cavity may also be useful information for drug design and/or site-directed mutagenesis studies. These data are indicated below for a few cases, but are readily available from the analysis module in the algorithm. In the following paragraphs we focus on several examples, and present, somewhat qualitatively, the level of success the VICE algorithm has obtained in describing these cavities for a broad range of variations in the architecture of binding pocket viz. deeply buried binding pockets, cavities at protein-protein dimer, and with DNA/RNA interface. The program also addresses the problem of defining metrics that indicate quantitatively and qualitatively the limits of a cavity, especially its boundary with free space, i.e., at the entrance (vide infra).

Well-enclosed cavities/deeply buried pockets

In the initial examples, we characterized deeply buried binding pockets that are, in other terms, well-enclosed cavities. These cases also may be thought of as essentially closed continuous volumes in the interior of protein molecules. While these binding pockets, which might bind small molecules, are sometimes not obvious from initial inspection, most available cavity detection software can effectively detect them. Although there may be a number of these voids inside a protein, it has been observed that the active site is usually the largest cavity in a protein [8,13] because a large pocket provides increased surface area and hence increased opportunity for small molecule binding. Thus, one of the problems faced by these algorithms is identifying the primary binding pocket amongst (often) numerous small clefts and voids. In addition, the boundary of the active site is often not well demarcated and numerous snake-like tendrils can project from the binding envelope. An important success factor of a cavity detection algorithm is in presenting a single, clean well-bounded cavity.

Prostaglandin H_2 synthase (PDB 1eqg) is an example of this class of cavity. A detailed structural analysis of NSAID binding with prostaglandin H_2 synthase is discussed by Selinsky et *al.* [43]. Figure 4 illustrates this protein and its detected cavity. The inset at the upper left shows the relatively small opening (calculated as 22 $Å^2$ by our algorithm) while the inset at the lower left extracts the cavity, ligand and surrounding residues (Pro86, Ile89, His90, Leu93, Met113, Val116, Arg120, Phe205, Val344, Ile345, Tyr348, Val349, Leu352, Ser353, Tyr355, Leu357, Leu359, Phe381, Leu384, Tyr385, Trp387, His513, Phe518, Glu520, Met522, Ile523, Glu524, Gly526, Ala527, Ser530, Leu531 and Leu 534). The cavity volume is estimated at 814 $Å^3$ of which only 214 $Å^3$ are occupied by ligand. We have not included any volume contribution from water in calculated volume estimates as the number of water molecules detected by x-ray crystallography varies greatly with crystallographic resolution [44].

Similarly, the anti-malarial compound fosmidomycin binds to IspC (PDB 1onp) [45] and the detected cavity is well-defined (Figure 5), surrounded by residues Ser151, Glu152, Gly185, Ser186, Gly187, Gly188, Trp212, Ser213, Ile218, Ser222, Asn227, Lys228, Glu231, Ser254, Met276 and a Mn ion. Here, the binding site is deeply buried with a volume of 342 $Å^3$, while the volume of fosmidomycin is 136 $Å^3$ of which 127 $Å^3$ occupies the active site. Most cavities in this class have opening surface areas that are about 10% or less of the total cavity surface area and have occupancy factors of around 35–50% (See Table 1).

Groove/cleft on the surface of a protein

The more shallow cavities and surface grooves are also potential sites for binding of drugs, ligands, proteins and other biomacromolecules. Identification and size characterization of surface pockets and occluded cavities are often the initial steps in protein structure-based drug design. The most important of these binding pockets are generally found to be particularly large and deep clefts. While internal cavities are fairly easy to define as they

generally correspond to well-enclosed regions completely bounded by surrounding atoms, in many cases interactions between protein and small molecule tend to involve what can appear to be a nearly planar surface on the surface of the protein. However, on the nano-scale protein surfaces are irregular with many clefts and grooves of varying shapes and sizes, and it is often difficult to define the boundaries of these shallow pockets. In particular the "open" boundary at the mouth is ambiguously defined even in the best of circumstances with this class of protein cavity. Our algorithm, as described in Figure 1, defines this boundary in terms of a user-adjustable parameter that represents the ratio of vectors finding the cavity wall over all vectors for each grid point. For this work we used the default value of 0.5 for this parameter, but it should be reiterated that this simple to comprehend parameter is user-adjustable and a crucial factor in the success of the VICE algorithm. In summary, most shallow cavities can be characterized by one key metric: they generally have opening cross-sectional areas (Table 1) of about 30% of the total cavity surface area.

One example of a shallow cavity on the surface of protein is illustrated with cytokine interleukin-2 (1m48) [46] in Figure 6. Here, the binding site is mapped to a shallow groove on the surface of protein. This particular protein is a symmetric homodimer so that there are two essentially identical binding sites. Cytokine interleukin-2 has been implicated as one of the principal mediators in proliferation and differentiation of activated cells in an immune response. It attaches through its surface to the trimeric IL-2R receptor, thereby triggering an immune response. Although the binding pocket is actually present as a surface cleft divided by a ridge, the cavity detection algorithm was able to capture both sides of the pocket. Interestingly, while a large portion of the ligand hangs out of the pocket, the two terminal ends are buried within the pocket.

In another example, as illustrated in Figure 7, a cavity was identified on the surface of the BCL-X_L protein (1bxl, 2yxj) [47,48], a pro-survival protein whose function is regulated by the binding of anti- or pro-apoptotic factors. Several anti-apoptotic proteins can bind to the BH3 domain of BCL-X_L in tumor cells where it is overexpressed. These interactions increase the survival rate of the cancer cell and may contribute to drug resistance. In contrast, pro-apoptotic proteins such as BAK can induce apoptosis by their binding to the BH3 domain; thus, the BH3 domain on BCL-X_L could be exploited as an attractive drug target in cancer chemotherapy. The BH3 domain has a largely hydrophobic surface with an estimated volume of 1300 $Å^3$. The lower left inset of Figure 7 shows BAK bound to the BH3 domain of BCL-X_L (1bxl). The associated cavity is indicated in yellow. However, a smaller sub-pocket (indicated in orange) can also be identified on the BH3 domain that binds small molecule modulators such as ABT-737 (2xyj) as shown in the upper right inset of Figure 7. The overlap of these two sites is shown in the central portion of Figure 7, and suggests that the bound ABT-737 ligand would block the binding of BAK. Exploitation of such cavities and sub-pockets at the interface between proteins could have important implications in drug discovery as more is learned about the role of protein-protein interactions in biological processes.

Cavity formed at a protein-protein interface

Next, we consider examples of cavities at protein–protein interfaces. These interactions have an important role in many biological processes and cavities at the interface of protein-protein dimers offer particularly attractive, but as yet largely unrealized, opportunities for therapeutic intervention. However, uncertainties owing to the structural changes due to domain movement upon binding and the often insufficient knowledge of well-defined binding pockets, coupled with the irregular shape and size of typical protein–protein interfaces, have made it difficult to design inhibitory ligands that can modulate protein-protein interactions. Although a large surface area is usually buried on each side of the

actual interface, there is often only a relatively small cavity or groove where a small molecule can fit and thus inhibit the protein-protein interaction.

However, in some cases, cavities at protein-protein interfaces can be observed, either at the joint between two subunits of the same protein or for a protein-protein complex. In one example, for αβ-tubulin (1z2b) (Figure 8) [49], our cavity detection algorithm defined the binding envelope at the wide interface between protein-protein units. Tubulin is the basic building block of microtubules, critical for mitosis and cell division, and an important target for anti-cancer drugs. Tubulin exists as a heterodimer and joins end-to-end to form a protofilament with alternating α and β subunits. The staggered assembly of 13 protofilaments forms hollow, cylindrical microtubule filaments. Three distinct binding sites have been identified on tubulin heterodimers for the taxol, colchicinoids and vinca classes of drugs. Although Taxol binds wholly on the β subunit, the colchicine binding site lies at the intradimeric interface of α and β subunits of tubulin and the vinblastine binding site is located at the interdimeric interface of αβ-subunits. The colchicine and vinblastine binding sites have been difficult to map as these binding pockets are poorly demarcated between the big subunit interfaces and the crystallographic resolution is rather poor at 3.58 Å. However, our algorithm was able to clearly find and delineate binding envelopes at these subunit interfaces: the colchicine binding site (Figure 8, left inset) has a volume of 842 $Å^3$ with an opening directly at the α-β interface with a estimated opening area of 28 $Å^2$; and the vinblastine site cavity has an estimated volume of 1457 $Å^3$ and an opening of 381 $Å^2$.

Cavity formed at a protein-polynucleotide interface

Protein-DNA/RNA interactions primarily are related to regulation of gene expression and are thus associated with important functions. Cavities or pockets formed by proteins at protein-nucleic acid interfaces are designed to mediate interactions and allow sequence-specific recognition of a gene. Each nucleic acid binding motif on a protein consists of a specific binding pocket that recognizes and stabilizes the DNA/RNA. To bind in this fashion a protein must make contact with the nucleic acid in such a way that the nucleotide sequence can be recognized. Ligands that can interfere with this recognition, either by occupying the putative nucleic acid binding site and blocking DNA/RNA binding, or by exploiting cavities formed in the protein-polynucleotide complex, may be therapeutically significant. As an example of the latter strategy, Figure 9 shows binding pockets detected on the 30S ribosomal subunit (1fjg) [50]. Three well-defined major cavities are detected indicating the binding sites for the antibiotics spectinomycin, paromomycin and streptomycin. The binding pocket for spectinomycin, which inhibits elongation factor G catalyzed translocation of the peptidyl-tRNA from the A-site to the P-site, has a volume of 633 $Å^3$ with spectinomycin completely enclosed within the cavity. The majority of interactions are with RNA bases C1063, G1064, C1066, G1068, C1069, A1191, C1192, G1193, U1194, G1386, G1387, with protein residues Ala121 & Gly 122 lining the cavity envelope. Paromomycin, an aminogycoside, binds in the major groove at the decoding center on H44 and induces errors in translation by increasing the affinity and stability of tRNA for the A-site. The volume of this cavity is 1605 $Å^3$ and it is lined by bases C1404, G1405, U1406, C1407, A1408, C1409, G1410, G1488, G1489, C1490, G1491, A1492, A1493, G1494, U1495, C1496, G1497 and protein residue Lys47. Adjacent to this binding pocket is a third cavity which binds streptomycin, a drug that inhibits protein synthesis by interfering with the initial selection and proofreading of tRNA. The volume of the predicted binding pocket is 988 $Å^3$ with numerous nucleotides from 16S RNA and residues from the S12 protein lining the binding envelope. While a limited numbers of base pairs are involved in recognition and stabilization, designing an inhibitor that binds at an interface must involve sufficient nucleic acid and protein contact so that the ligand fits snugly.

Flexible cavities with loop or domain movements

All proteins have an intrinsic flexibility that is required for a wide range of biochemical processes in catalysis, regulation, and protein assembly. However, in some cases experimental evidence has indicated that the shape and size of the ligand binding envelope may change due to domain movements; e.g., molecular recognition and ligand binding is induced by large loop movements where flexibility in the protein main chain influences the ligand binding [51]. Ligand binding may involve a wide range of structural changes in the receptor protein, from hinge movement of entire domains to small side-chain rearrangements in the binding pocket residues. Many protein functions in fact involve conformational transitions that involve opening and closing of relatively rigid parts of that protein about flexible joints. The analysis of side chain flexibility gives insight valuable for improving docking algorithms and for ligand design when domain movement and/or loop flexibility opens and closes the binding pocket. Instead of well-defined binding pockets, most proteins that have 'induced' domain movement lack deep clefts or clearly shaped binding pockets. Thus, this is an interesting case study for cavity detection – where the change in the size and shape of binding pocket due to domain movement is calculated by comparison between pairs of *holo* and *apo* proteins. Figure 10 shows the example of citrate synthase, 5cts [52] and 5csc [53], which are the *apo* (unliganded) and *holo* (ligand-bound) forms with cavity volumes of 439 $Å^3$ and 967 $Å^3$, respectively. The bound ligand, oxaloacetate, which has a volume of 704 $Å^3$, appears to induce this large domain movement in the enzyme and causes binding pocket residues to undergo side-chain conformational changes as well as changes in overall shape. Residues His238, Asn242, Leu273, His274, Val314, Val315, Gly317, Tyr318, Gly319, His320, Ala321, Arg329, Gln364, Ala367, Ala368, Asn373, Asp375, Phe397 surround the binding pocket in the closed structure, while only residues His238, His274, His320, Arg329, Asp375, and Phe397 are lining the unliganded pocket.

Multi-domain proteins with channels or tunnels

Understanding the structure and function of channels and pores within biomolecules is important, e.g., to a large number of critical disease states and in compensating for drug resistance due to efflux. Channels and pores and other passages across cell membranes facilitate the movement of small molecules and ions. These transmembrane proteins, such as ion channels, transporters and G-protein coupled receptors, are exceptionally significant drug targets. Apart from this, channels and tunnels also facilitate the access and exit for substrates/products in some catalytic processes. Channels/pores are often dynamic in nature and can be relatively flexible in size and shape and access through them is often regulated by small molecules binding to an active site. Thus, while many of the available algorithms and associated programs developed to detect and characterize binding pockets are successful with well-enclosed pockets and surface grooves, for the most part these procedures fail to detect long, twisted tunnels connecting the interior of a binding pocket to the exterior environment. In fact, it is surprisingly difficult mathematically to differentiate between true channels and tunnels and random voids if the tunnel has a narrow diameter or constriction point(s).

With the recent availability of crystal structures for large membrane-bound proteins, detection and mapping of the interior of these channels can give insight into the binding process for design and development of more selective drugs. Our cavity detection algorithm provides sufficient flexibility and interactivity to map binding sites as well as the channels and tunnels through a protein. In the example illustrated in Figure 11, the KcsA potassium channel (1j95) [54], our cavity detection algorithm was able to locate the binding pocket along with a part of channel, which is occupied by tetrabutylammonium in this structure. This was easily detected with the program's default parameters, e.g., a grid resolution of 1.0

Å. However, to visualize the channel exclusively grid resolution was decreased to 0.3 Å, and molecular extents were redefined with a margin of 2.0 Å around the channel. The program successfully delineated a long, narrow porous channel traversing the entire length of the protein's transmembrane axis. It should be noted that this latter calculation was resource intensive due to the very large number of surveyed grid points, but this level of computation was necessary in order to adequately sample the protein structure. The total volume of channel was calculated to be 1342 $Å^3$ with the binding cavity of 615 $Å^3$, while the tetrabutylammonium occupies 168 $Å^3$ inside the binding cavity of the channel and is well-enclosed by hydrophobic residues.

Multiple cavities and allosteric binding pockets

The detection of auxiliary binding sites is becoming increasingly crucial as many proteins have more than one biochemical role and are likely to employ separate binding sites in performing these distinct biochemical tasks. Allosteric binding pockets may offer additional recognition sites that modulate the catalytic function of a protein. These auxiliary binding pockets may be located far away from the catalytic site, as in case of glycogen phosphorylase, or may overlap with the active site. Traditionally, allosteric sites were considered to be distal binding sites for molecules that may modulate the function of a protein by a feedback mechanism. While the mechanisms of allosteric modulation of proteins have been extensively studied, discovery efforts to efficiently find and characterize these binding sites continue as exploiting them may lead to development of entirely new classes of drugs. However, it can be a non-trivial matter to find and characterize allosteric binding sites when these sites are present as auxiliary pockets overlapping with the main active site. Figure 12 illustrates an example of an allosteric site on glycogen phosphorylase b (1c50) [55]. The crystal structure shows an allosteric binding site for the co-crystallized molecule CP320626. Our program identified this binding site with a volume of 431 $Å^3$ close to the AMP binding site with a volume of 728 $Å^3$. The main PLP catalytic site, with a volume of 849 $Å^3$, is about 30 Å distant from the allosteric site.

Summary and outlook

The location, delineation and visualization of protein active sites is a critical facet of drug design. These site topographies play crucial roles in molecular recognition. Proteins may have many pockets and cavities of various sizes, some of which whose function, e.g., protein-protein interaction, is unknown; it is possible that some may be usefully exploited with selective molecules that bind and modulate that protein's function. Thus it is important to be able to characterize these binding pockets, quantitatively and qualitatively. This algorithm and program provides a simple and interactive tool for locating and delineating all different kinds of pockets on a protein whose structure is known. In most cases the default parameters produce good results very rapidly because the majority of the calculations are performed with integer arithmetic. For example, full protein scans at 1.0 Å resolution for 2yxj (21.6 kDa), 1onp (43.7 kDa) and 1z2b (220 kDa), required 19 s, 63 s and 2695 s, respectively, on a 1.3 gHz AMD64 processor to calculate the raw cavity maps. Also, cavity volumes calculated by VICE are generally independent of Cartesian axes orientation; in 30 random orientations of 121p, the calculated cavity volumes were within ± 11% and in all cases these cavities enclosed the ligand. Thus, we believe that this tool could be a useful starting point for virtual screening by automatically and reproducibly locating potential binding sites in a first pass.

A few recent publications have explored success rates in locating binding pockets using various algorithms. This was first described by Huang et al. [42] in 2006, and revisited by Weisel et al. in 2007. Success is recorded when the actual ligand is located in the largest cavity (or one of the three largest cavities) found by an algorithm. Table 2 presents a

summary of this metric for VICE, Fpocket [56], PocketPicker [14], LIGSITE^{cs} [42], CAST [9], PASS [10] and SURFNET [12] on a data set of 48 unbound and proteins. The VICE success rate is more than 10% higher in locating the ligand in the largest pocket than any of the other methods except the very recently published Fpocket with which it compares very favorably (83% vs. 69%) for the most difficult and relevant problem of locating the binding pocket in the structures on unbound proteins.

Large proteins can yield many cavities that may require iterative refinement/visualization cycles. Of course, more reliable and biologically meaningful results will be obtained if the user can focus on particular regions or features by selecting one or more of the pockets and investigating them in detail by adjusting a relatively small number of calculational parameters and by restricting the scan to the region of interest. The fairly common presence of a co-crystallized ligand in the structure yields a particularly simple means of focusing on the pocket of interest, but can bias the user into assuming that only one pocket exists.

A second major advantage of this program is that it calculates a fairly extensive set of metrics describing a binding pocket and its occupants. Cavity volumes, cavity surface areas, and lists of atoms, residues and/or chains lining the binding pocket can be retrieved. The estimated cross-sectional surface area of cavity openings is particularly interesting as it may suggest methods to describe the maximum size of ligands to enter a site, although significant flexibility in this regard is expected. It is surprising that these types of quantitative metrics are reported by few of the other available cavity detection programs. This makes comparison between methods difficult and ultimately only qualitative in nature.

With our rapid and robust VICE cavity algorithm in place, we are exploring virtual screening and docking algorithms that use property-encoded cavities, e.g., the HINT complement map, as first stage targets. Such cavity maps would be inherently suited for grid-based pose generation and scoring.

Acknowledgments

We gratefully acknowledge the helpful advice and suggestions of Profs. J. Neel Scarsdale, Philip D. Mosier, John C. Hackett, and Jason P. Rife (VCU). Mr. Hardik Parikh and Mr. Chenxiao Da (VCU) provided technical assistance. This work was partially supported by the U.S. National Institutes of Health grant GM071894 to G.E.K.

References

1. Sotriffer C, Klebe G. Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. Il Farmaco. 2002; 57:243–251. [PubMed: 11989803]
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucl Acids Res. 2000; 28:235–242. [PubMed: 10592235]
3. 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]
4. Pazos F, Sternberg MJE. Automated prediction of protein function and detection of functional sites from structure. Proc Natl Acad Sci. 2004; 101:14754–14759. [PubMed: 15456910]
5. Kleywegt GJ. Detection, delineation, measurement and display of cavities in macromolecular structures. Acta Crystallogr D. 1994; 50:178–185. [PubMed: 15299456]
6. Hendlich M, Rippmann F, Barnickel G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model. 1997; 15:359–363. [PubMed: 9704298]
7. Levitt DG, Banaszak LJ. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J Mol Graph. 1992; 10:229–234. [PubMed: 1476996]

8. An J, Totrov M, Abagyan R. Pocketome via comprehensive identification and classification of ligand binding envelopes. Mol Cell Proteomics. 2005; 4:752–761. [PubMed: 15757999]
9. Liang J, Edelsbrunner H, Woodward C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci. 1998; 7:1884–1897. [PubMed: 9761470]
10. Brady GP Jr, Stouten PF. Fast prediction and visualization of protein binding pockets with PASS. J Comput Aided Mol Des. 2000; 14:383–401. [PubMed: 10815774]
11. 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]
12. Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph. 1995; 13:323–330. [PubMed: 8603061]
13. Laurie ATR, Jackson RM. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics. 2005; 21:1908–1916. [PubMed: 15701681]
14. Weisel M, Proschak E, Schneider G. PocketPicker: analysis of ligand binding-sites with shape descriptors. Chemistry Central Journal. 2007; 1:7. [PubMed: 17880740]
15. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol. 1996; 257:342–358. [PubMed: 8609628]
16. Aloy P, Querol E, Aviles FX, Sternberg MJ. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol. 2001; 311:395–408. [PubMed: 11478868]
17. Stuart AC, Ilyin VA, Sali A. LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. Bioinformatics. 2002; 18:200–201. [PubMed: 11836232]
18. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJ, Zdobnov EM. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res. 2001; 29:37–40. [PubMed: 11125043]
19. Bickel PJ, Kechris KJ, Spector PC, Wedemayer GJ, Glazer AN. Inaugural article: finding important sites in protein sequences. Proc Natl Acad Sci. 2002; 99:14764–14771. [PubMed: 12417758]
20. Armon A, Graur D, Ben-Tal N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol. 2001; 307:447–463. [PubMed: 11243830]
21. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics. 2002; 18:S71–S77. [PubMed: 12169533]
22. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol. 1971; 55:379–400. [PubMed: 5551392]
23. Connolly ML. Analytical molecular surface calculation. J Appl Cryst. 1983; 16:548–558.
24. 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]
25. DesJarlais RL, Sheridan RP, Seibel GL, Dixon JS, Kuntz ID, Venkataraghavan R. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. J Med Chem. 1988; 31:722–729. [PubMed: 3127588]
26. Meng E, Shoichet B, Kuntz ID. Automated docking with grid based energy evaluation. J Comput Chem. 1992; 13:505–524.
27. Goodford PJ. A computational Procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem. 1985; 28:849–857. [PubMed: 3892003]
28. Miranker A, Karplus M. Functionality maps of binding sites: a multiple copy simultaneous search method. Proteins: Structure, Function, Genetics. 1991; 11:29–34.

29. Voorintholt R, Kosters MT, Vegter G. A very fast program for visualizing protein surfaces, channels and cavities. J Mol Graphics. 1989; 7:243–245.
30. Del Carpio CA, Takahashi Y, Sasaki SI. A new approach to the automatic identification of candidates for ligand receptor sites in proteins: (I) search for pocket regions. J Mol Graphics. 1993; 11:23–29.
31. Kisljuk OS, Kachlova GS, Lanina NP. An algorithm to find channels and cavities within protein crystals. J Mol Graphics. 1994; 12:305–307.
32. Edelsbrunner H, Mücke E. Three-dimensional alpha shapes. ACM Trans Graph. 1994; 13:43–72.
33. Edelsbrunner, H.; Facello, M.; Fu, P.; Liang, J. Measuring proteins and voids in proteins. Proceedings of the Twenty-Eighth Hawaii International Conference on System Sciences; 3–6 January 1995; Wailea. 1995. p. 256-264.
34. Lee DT, Schachter BJ. Two Algorithms for constructing a Delaunay triangulation. Int J Comput Inf Sci. 1980; 9:219–242.
35. Stahl M, Bur D, Schneider G. Mapping of proteinase active sites by projection of surface-derived correlation vectors. J Comput Chem. 1999; 20:336–347.
36. Kellogg GE, Fornabaio M, Chen DL, Abraham DJ, Spyrakis F, Cozzini P, Mozzarelli A. Tools for building a comprehensive modeling system for virtual screening under real biological conditions: The Computational Titration algorithm. J Mol Graph Model. 2006; 24:434–439. [PubMed: 16236534]
37. Chen DL, Kellogg GE. A computational tool to optimize ligand selectivity between two similar biomacromolecular targets. J Comput Aided Mol Des. 2005; 19:69–82. [PubMed: 16075302]
38. Kellogg GE, Fornabaio M, Chen DL, Abraham DJ. New application design for a 3D hydropathic map-based search for potential water molecules bridging between protein and ligand. Internet Electron J Mol Des. 2005; 4:194–209.
39. Kellogg GE, Abraham DJ. Hydrophobicity: is LogP(o/w) more than the sum of its parts? Eur J Med Chem. 2000; 35:651–661. [PubMed: 10960181]
40. Cozzini P, Fornabaio M, Marabotti A, Abraham DJ, Kellogg GE, Mozzarelli A. Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 1. Models without explicit constrained water. J Med Chem. 2002; 45:2469–2483. [PubMed: 12036355]
41. Spyrakis F, Amadasi A, Fornabaio M, Abraham DJ, Mozzarelli A, Kellogg GE, Cozzini P. The consequences of scoring docked ligand conformations using free energy correlations. Eur J Med Chem. 2007; 42:921–933. [PubMed: 17346861]
42. Huang B, Schroeder M. LIGSITE^{csc}: preedecting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct. Biol. 2006; 6:19.
43. Selinsky BS, Gupta K, Sharkey CT, Loll PJ. Structural analysis of NSAID binding by prostaglandin H2 synthase: time-dependent and time-independent inhibitors elicit identical enzyme conformations. Biochemistry. 2001; 40:5172–5180. [PubMed: 11318639]
44. Carugo O, Bordo D. How many water molecules can be detected by protein crystallography? Acta Crystallogr D Biol Crystallogr. 1999; 55:479–483. [PubMed: 10089359]
45. Steinbacher S, Kaiser J, Eisenreich W, Huber R, Bacher A, Rohdich F. Structural basis of fosmidomycin action revealed by the complex with 2-C-methyl-D-erythritol 4-phosphate synthase (IspC). Implications for the catalytic mechanism and anti-malaria drug development. J Biol Chem. 2003; 278:18401–18407. [PubMed: 12621040]
46. Arkin MA, Randal M, DeLano WL, Hyde J, Luong TN, Oslob JD, Raphael DR, Taylor L, Wang J, McDowell RS, Wells JA, Braisted AC. Binding of small molecules to an adaptive protein-protein interface. Proc Natl Acad Sci USA. 2003; 100:1603–1608. [PubMed: 12582206]
47. Sattler M, Liang H, Nettesheim D, Meadows RP, Harlan JE, Eberstadt M, Yoon HS, Shuker SB, Chang BS, Minn AJ, Thompson CB, Fesik SW. Structure of Bcl-xL-Bak peptide complex: recognition between regulators of apoptosis. Science. 1997; 275:983–986. [PubMed: 9020082]
48. Lee EF, Czabotar PE, Smith BJ, Deshayes K, Zobel K, Colman PM, Fairlie WD. Crystal structure of ABT-737 complexed with Bcl-xL: implications for selectivity of antagonists of the Bcl-2 family. Cell Death Differ. 2007; 14:1711–1713. [PubMed: 17572662]

49. Gigant B, Wang C, Ravelli RB, Roussi F, Steinmetz MO, Curmi PA, Sobel A, Knossow M. Structural basis for the regulation of tubulin by vinblastine. Nature. 2005; 435:519–522. [PubMed: 15917812]
50. Carter AP, Clemons WM Jr, Brodersen DE, Morgan-Warren RJ, Wimberly BT, Ramakrishnan V. Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. Nature. 2000; 407:340–348. [PubMed: 11014183]
51. Hayward S. Identification of specific interactions that drive ligand-induced closure in five enzymes with classic domain movements. J Mol Biol. 2004; 339:1001–1021. [PubMed: 15165865]
52. Karpusas M, Branchaud B, Remington SJ. Proposed mechanism for the condensation reaction of citrate synthase: 1.9-Å structure of the ternary complex with oxaloacetate and carboxymethyl coenzyme A. Biochemistry. 1990; 29:2213–2219. [PubMed: 2337600]
53. Liao D-I, Karpusas M, Remington SJ. Crystal structure of an open conformation of citrate synthase from chicken heart at 2.8-Å resolution. Biochemistry. 1991; 30:6031–6036. [PubMed: 2043641]
54. Zhou M, Morais-Cabral JH, Mann S, MacKinnon R. Potassium channel receptor site for the inactivation gate and quaternary amine inhibitors. Nature. 2001; 411:657–661. [PubMed: 11395760]
55. Oikonomakos NG, Skamnaki VT, Tsitsanou KE, Gavalas NG, Johnson LN. A new allosteric site in glycogen phosphorylase b as a target for drug interactions. Structure Fold Des. 2000; 8:575–584. [PubMed: 10873856]
56. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics. 2009; 10:168. [PubMed: 19486540]

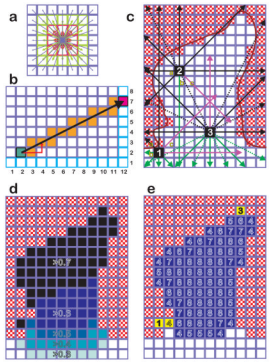

Figure 1. VICE Algorithm

a) Vector representations of direction: red = shell 1, green = shell 2, blue = shell 3; b) Vector (starting in green) continues until reaching grid box edge (red) and all nodes in path (orange shading) are tested; c) Each grid point is surveyed with set of vectors that: are blocked by molecule (black), have clear path to box edge (green), or are stalled (pink) because with their finite length they do not reach box edge and thus are considered as having a clear path. Node 1 is clearly outside the cavity (more clear than blocked paths), node 2 is clearly inside (more blocked than clear), while node 3 is ambiguous requiring further examination with shell 2 vectors; d) The fraction of blocked vectors is represented as a contourable scalar quantity that most impacts the definition of “cavityness” at the mouth; and e) Tendrils, very narrow channels and other vague regions are tested with neighbor count that requires each node to have a minimum number of neighbors defined to be inside the cavity. The nodes indicated in yellow are subject to this filter, which may be applied recursively. Not shown: each closed solid contour must have a minimum volume or it will be deleted.

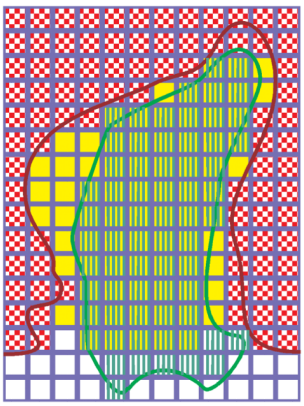

Figure 2. Cavity Volume Metrics

The volume of the cavity (V_C) is indicated by yellow shading, the volume of the ligand (V_L) is indicated by vertical green bars, the volume of the ligand occupying the cavity (V_O) is the intersection of V_C and V_L, i.e., yellow shading + green bars. The unoccupied cavity volume is $V_C - V_O$, and the volume of the ligand outside the cavity is $V_L - V_O$.

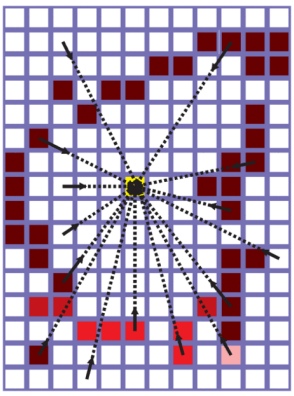

Figure 3. Cavity Entrance Calculation

The cavity entrance is calculated from the derivative of the map illustrated in Figure 1d. Vectors are projected from each grid node toward the center-of-gravity of the cavity (dashed lines); the path (as in Figure 1b) is determined and the absolute value of the difference between the starting grid point and the first node on that path is calculated as the derivative. Paths completely inside or outside will have close to zero slope (white), paths clearly crossing from outside to inside will have slope values close to one (dark red), while the ambiguous cavity mouth grid points will have intermediate slope values.

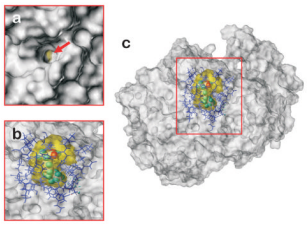

Figure 4. Well-enclosed Cavity

Prostaglandin H_2 synthase (1eqg) examined with the VICE algorithm and displayed with MOLCAD and Sybyl. a) The protein Connolly surface is displayed with opaque rendering. The small opening to the cavity is indicated by the red arrow; b) the ligand, ibuprofen rendered in CPK (space-filled), and the residues lining the cavity are shown. The yellow translucent surface illustrates the extents of the cavity. The protein is rendered with a translucent Connolly surface; c) shown as in b) but displaying the entire protein.

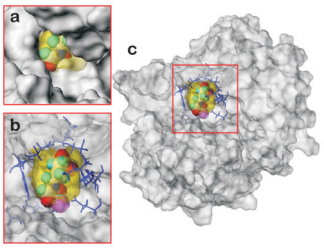

Figure 5. Well-enclosed Cavity

IspC (1onp) examined with the VICE algorithm and displayed with MOLCAD and Sybyl. a) The protein Connolly surface is displayed with opaque rendering. The relatively small opening to the cavity can be seen; b) the ligand, the anti-malarial compound fosmidomycin rendered in CPK, and the residues lining the cavity are shown. The yellow translucent surface illustrates the extents of the cavity. The protein is rendered with a translucent Connolly surface and the space-filling magenta sphere is the manganese ion; c) shown as in b) but displaying the entire protein.

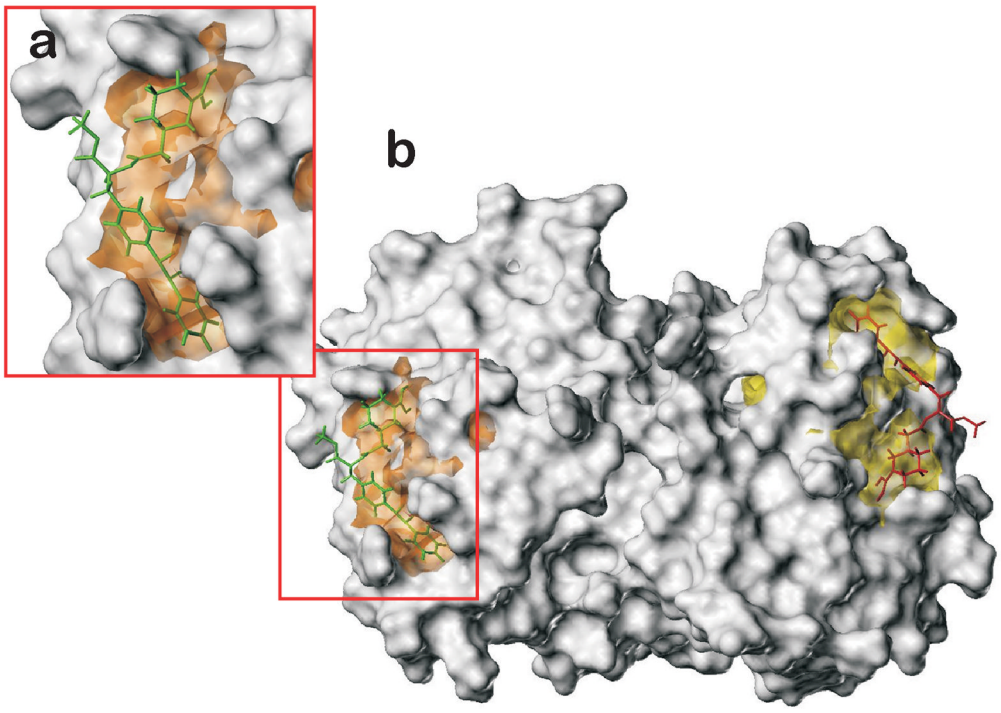

Figure 6. Shallow Cavity on Protein Surface

The cytokine interleukin-2 dimer (1m48) has one essentially identical shallow cavity binding site on each of the two chains. a) The inhibitor Ro26-4550 is bound in the cavity of chain A: the cavity extents are displayed as the orange contour volume. Both ends are well-bound but much of the middle of the ligand is external to the cavity; b) both sites are displayed in this view of the entire protein.

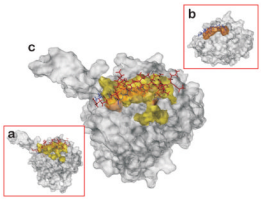

Figure 7. Shallow Cavity on Protein Surface

Two structures of the BCL-X_L protein with BAK protein and inhibitor ABT-737 bound within its binding cavity. a) BCL-X_L protein (1bxl) with sixteen residue BAK protein (red capped stick representation) bound within the surface cavity (yellow translucent envelope); b) BCL-X_L protein (2yxj) with ABT-737 inhibitor (blue capped sticks) bound in a relatively smaller sub-pocket (orange translucent surface); c) overlap superposition of 1bxl and 2yxj structures showing the correspondence of the two pockets. Cavity extents illustrated with yellow and orange translucent envelopes.

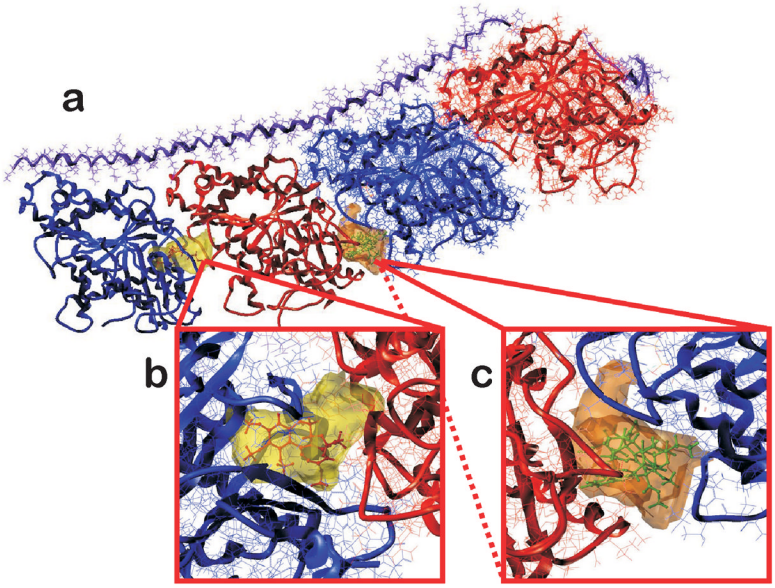

Figure 8. Cavity at Protein-Protein Interface

a) The tubulin protein (1z2b) with colchicine and vinblastine binding sites at interfaces between the α and β subunits. The tubulin polymer is rendered in ribbon and tube with the α subunits shown in red and β subunits shown in blue; b) inset shows the colchicine binding pocket (yellow contour) at the intra-dimeric interface of the αβ-subunit; c) inset shows the vinblastine binding site (orange contour) at the inter-dimeric interface between αβ-subunits.

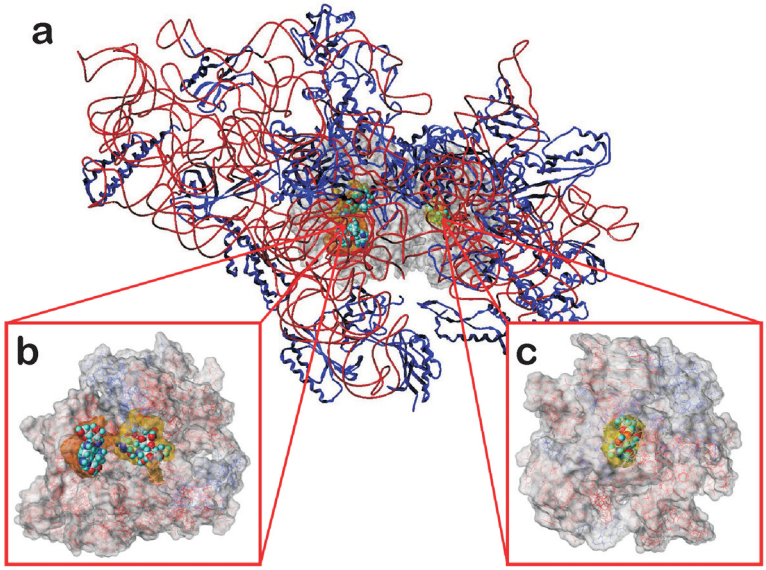

Figure 9. Cavity at Protein/Polynucleotide Interface

a) The 30S ribosomal subunit (1fjg) is rendered as ribbon and tube, except within 20 Å of binding region where a MOLCAD surface display is shown to highlight the binding pockets for the antibiotics spectinomycin, paromomycin and streptomycin; b) the binding site for paromomycin (orange envelope) and streptomycin (yellow envelope) are illustrated. The antibiotic drugs are rendered in spacefill; c) the binding pocket for spectinomycin (yellow envelope) is illustrated.

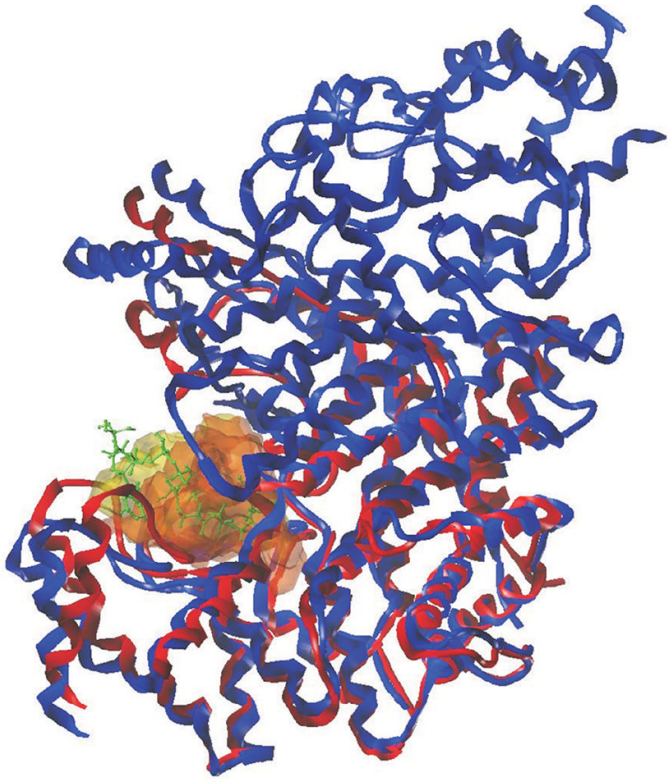

Figure 10. Flexible Cavity with Loop or Domain Movement

The citrate synthase protein, 5cts (red) and 5csc (blue), the *apo* (unliganded) and *holo* (ligand-bound) forms, respectively, is illustrated. A relatively smaller binding pocket is detected in 5cts (orange envelope); however, the native ligand oxaloacetate (green capped sticks) induces a domain movement that significantly alters the shape and size of the binding pocket (yellow envelope) in 5csc.

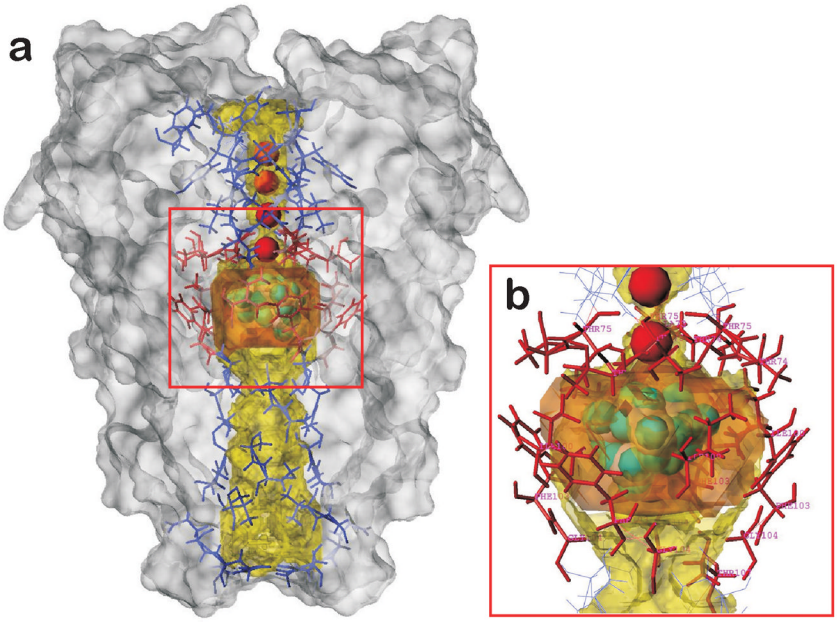

Figure 11. Channels and Tunnels

a) The KscA K^+ ion channel (1j95) plotted with translucent MOLCAD surface. The binding pocket at the center of the channel is illustrated with an orange contour map; its tetrabutylammonium inhibitor is rendered in CPK (space-fill). The channel, traversing the entire length of the protein, is highlighted with the yellow contour map. Detection of the channel required calculations with a very large number of grid map points and high resolution. The potassium ions are rendered as the red spheres; b) expanded view of the inhibitor binding cavity.

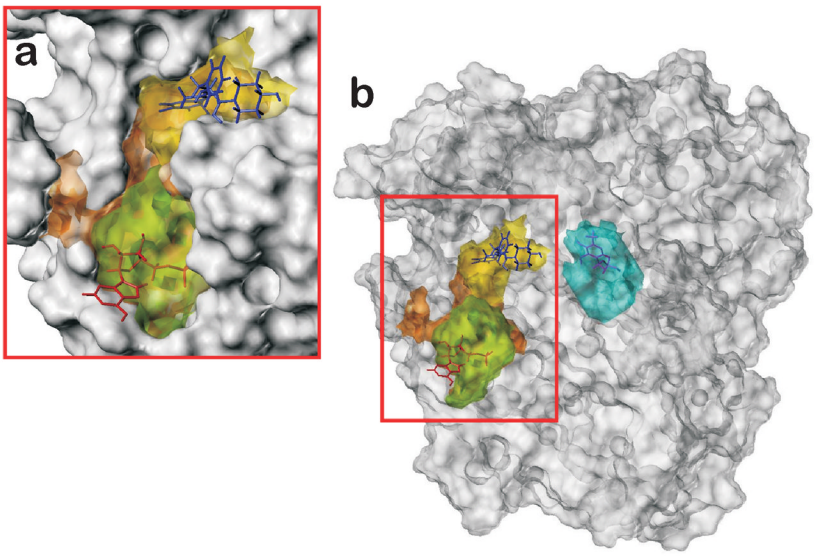

Figure 12. Auxiliary and allosteric sites

The glycogen phosphorylase b (1c50) with multiple binding pockets. a) A close-up view of the allosteric/auxiliary site. The AMP cofactor (red sticks, green cavity contour) and allosteric site (yellow contour) with inhibitor CP320626 (blue sticks) are in separate subsites of the overall surface groove (orange contour); b) the main catalytic site (cyan contour) is bound with PLP and is quite deeply buried in the protein.

Table 1

Protein Cavity Data

PDB Code	Cavity Type	Subcavity ID	Cavity vol. (Å^3)	Ligand volume (Å^3)			Occupied fraction	Cavity surface area (Å^2)		Opening fraction
				Total	Inside	Outside		Total	Opening(s)	
3cox	Deeply buried pockets		1451	575	565	10	0.39	1834	42	0.022
1onp			342	136	127	9	0.37	405	49	0.121
1aes[a]			179	61	61	0	0.34	209	0	0.000
1mbd			903	521	418	103	0.46	1400	154	0.110
1piv			564	330	314	16	0.59	823	35	0.043
1gf1			426	321	250	71	0.59	458	2	0.004
1tys			1499	576	549	27	0.37	2079	194	0.093
1asc			380	198	182	16	0.48	477	33	0.069
1eqg			814	215	214	1	0.26	1230	22	0.018
1m6y			596	295	269	26	0.45	695	29	0.042
1ju3			203	113	101	12	0.50	208	10	0.048
1ydd			273	147	121	26	0.54	308	11	0.036
1hor			402	183	176	7	0.44	510	11	0.022
121p			924	344	333	11	0.36	1116	107	0.096
2acs			947	656	547	109	0.58	1147	97	0.085
1njs			578	431	305	126	0.53	791	71	0.090
1dr1	Cavity from surface depression		2010	686	673	13	0.33	2728	348	0.128
1m48			400	418	141	277	0.35	558	190	0.341
1fmp			396	238	173	65	0.44	510	120	0.236
1k2v			395	426	247	179	0.63	604	203	0.336
1b12			293	244	174	70	0.59	309	62	0.201
2ngr			856	282	264	18	0.31	964	164	0.170
1e3e			2352	474	464	10	0.20	3261	734	0.225
1bxl			1498	1476	518	980	0.35	2065	791	0.383
1adg			1288	457	422	35	0.33	1597	435	0.272

PDB Code	Cavity Type	Subcavity ID	Cavity vol. (Å^3)	Ligand volume (Å^3)			Occupied fraction	Cavity surface area (Å^2)		Opening fraction
				Total	Inside	Outside		Total	Opening(s)	
1cqp	Cavity formed at protein-protein interface	Chain A	982	396	380	16	0.38	1200	191	0.159
		Chain B	941	389	359	30	0.36	1175	198	0.168
1dn2		Chain E	620	1253	231	1022	0.14	1086	245	0.225
		Chain F	578	1264	198	1066	0.12	863	280	0.324
1i8l		Chain C	923	10475	79	10396	0.007	1378	190	0.137
		Chain D	807	10302	63	10239	0.005	1395	221	0.158
1jhl			527	11955	95	11860	0.18	507	196	0.387
1mlc			423	11927	167	11760	0.39	483	138	0.286
1rv1			692	516	328	188	0.47	881	218	0.247
1t4f			834	1011	402	609	0.48	208	676	0.308
1vfb			413	20770	61	20710	0.15	389	130	0.334
3ink			215	12306	18	12228	0.08	217	67	0.309
1z2b		Colchicine	842	374	343	31	0.39	1015	28	0.027
		Vincristine	1457	705	615	90	0.39	1956	381	0.194
1aud	a protein-polynucleotide		881	6104	270	5834	0.31	1056	285	0.270
1esg			2848	3131	100	3031	0.04	4337	2804	0.647
1euq			924	343	324	19	0.34	1097	154	0.140
1×9n			12267	7761	2491	5270	0.20	18884	6207	0.329
2c62			1791	3002	486	2516	0.27	2985	1017	0.341
2cdm			4197	4797	1388	3409	0.33	8018	2467	0.308
2euv			1662	4958	487	4471	0.29	1742	414	0.238
2hhs			4160	5171	705	4466	0.17	6377	3025	0.474
2hw8			1435	7379	466	6913	0.32	2159	1391	0.644
2i0q			2672	2214	914	1300	0.34	3862	1263	0.343
1ama 9aat	Flexible cavities with loop or domain movements	*Holo*	330	290	222	68	0.67	425	42	0.099
		Apo	300	290	211	79	0.70	402	64	0.159
1ank 4ake		*Holo*	950	572	438	134	0.46	1488	21	0.014
		Apo	358	572	94	478	0.26	472	171	0.362

PDB Code	Cavity Type	Subcavity ID	Cavity vol. (Å^3)	Ligand volume (Å^3)			Occupied fraction	Cavity surface area (Å^2)		Opening fraction
				Total	Inside	Outside		Total	Opening(s)	
1atp 1ctp		*Holo*	707	332	316	16	0.45	914	9	0.010
		Apo	531	332	188	144	0.35	669	129	0.193
2ohx 8adh		*Holo*	722	464	390	74	0.54	910	125	0.137
		Apo	590	464	278	186	0.47	700	182	0.260
5cts 5csc		*Holo*	967	704	507	197	0.52	1181	167	0.141
		Apo	439	704	167	537	0.38	662	328	0.495
1av2[b]	Proteins with channels or tunnels		305	-	-	-	-	509	92	0.180
1j95[c]		channel	1575	319	260	58	0.15	995	137	0.137
		TBA ligand	658	168	165	3	0.24	850	2	0.002
1okc			6192	608	592	16	0.09	7592	608	0.080
2al5			3266	518	501	17	0.15	4526	271	0.063
2byq[b]			37512	-	-	-	-	51168	2816	0.055
1c50	Multiple cavities and allosteric binding pockets	PLP	849	184	169	15	0.19	1114	9	0.008
		AMP	728	237	142	95	0.17	945	256	0.270
		CHI	431	371	112	259	0.16	681	164	0.240
1kv2		allosteric	1293	484	458	26	0.34	1676	261	0.155
		main	639	327	178	149	0.22	787	35	0.044
1q0b		allosteric	515	256	221	35	0.40	667	85	0.127
		main	558	282	219	63	0.35	915	281	0.307
1v4s		allosteric	722	264	261	3.0	0.36	960	62	0.064
		main	282	138	133	5.0	0.46	330	4	0.012

[a] no apparent opening to cavity

[b] no ligand in model

[c] channel calculated with 0.3 Å resolution; ligand site with 1.0 Å resolution.

Table 2

Comparison of success (percent) for 48 complexed and 48 unbound protein structures.[a]

Method	Top 1		Top 3	
	Unbound	Bound	Unbound	Bound
VICE	83	85	90	94
Fpocket	69	83	94	92
PocketPicker	69	72	85	85
LIGSITE^{cs}	60	69	77	87
CAST	58	67	75	83
PASS	60	63	71	81
SURFNET	52	54	75	78

[a]Data represents the rate of success in finding the actual ligand pocket in cavities calculated by the various algorithms. If the largest cavity contains the ligand it is “Top 1”; if one of the three largest cavities contains the ligand it is “Top 3”. The data for VICE is from this work; the data for Fpocket were first reported by Le Guilloux et al. [56]; the data for PocketPicker were first reported by Weisel *et al.* [14]; the data for LIGSITE^{cs}, CAST, PASS and SURFNET were first reported by Huang *et al.* [42].