Protein Science (1998), 7:1884–1897. Cambridge University Press. Printed in the USA.
Copyright © 1998 The Protein Society

Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design

JIE LIANG,[1,3] HERBERT EDELSBRUNNER,[2] AND CLARE WOODWARD[1]

[1]Department of Biochemistry, University of Minnesota, 1479 Gortner Avenue, St. Paul, Minnesota 55108
[2]Department of Computer Science, University of Illinois, 1304 West Springfield Avenue, Urbana, Illinois 61801

(RECEIVED December 30, 1997; ACCEPTED April 14, 1998)

Abstract

Identification and size characterization of surface pockets and occluded cavities are initial steps in protein structure-based ligand design. A new program, CAST, for automatically locating and measuring protein pockets and cavities, is based on precise computational geometry methods, including alpha shape and discrete flow theory. CAST identifies and measures pockets and pocket mouth openings, as well as cavities. The program specifies the atoms lining pockets, pocket openings, and buried cavities; the volume and area of pockets and cavities; and the area and circumference of mouth openings. CAST analysis of over 100 proteins has been carried out; proteins examined include a set of 51 monomeric enzyme–ligand structures, several elastase–inhibitor complexes, the FK506 binding protein, 30 HIV-1 protease–inhibitor complexes, and a number of small and large protein inhibitors. Medium-sized globular proteins typically have 10–20 pockets/cavities. Most often, binding sites are pockets with 1–2 mouth openings; much less frequently they are cavities. Ligand binding pockets vary widely in size, most within the range 10^2–10^3 Å^3. Statistical analysis reveals that the number of pockets and cavities is correlated with protein size, but there is no correlation between the size of the protein and the size of binding sites. Most frequently, the largest pocket/cavity is the active site, but there are a number of instructive exceptions. Ligand volume and binding site volume are somewhat correlated when binding site volume is ≤700 Å^3, but the ligand seldom occupies the entire site. Auxiliary pockets near the active site have been suggested as additional binding surface for designed ligands (Mattos C et al., 1994, *Nat Struct Biol 1*:55–58). Analysis of elastase–inhibitor complexes suggests that CAST can identify ancillary pockets suitable for recruitment in ligand design strategies. Analysis of the FK506 binding protein, and of compounds developed in SAR by NMR (Shuker SB et al., 1996, *Science 274*:1531–1534), indicates that CAST pocket computation may provide a priori identification of target proteins for linked-fragment design. CAST analysis of 30 HIV-1 protease–inhibitor complexes shows that the flexible active site pocket can vary over a range of 853–1,566 Å^3, and that there are two pockets near or adjoining the active site that may be recruited for ligand design.

Keywords: alpha shape; delaunay triangulation; docking; molecular interactions; molecular recognition; protein binding site; protein ligand design; protein pockets and cavities; protein structure analysis; SAR by NMR

The complex shapes of protein surfaces are sculpted by numerous concavities and protrusions, which offer unique microenvironments for ligand binding and catalysis. In addition to pockets, internal cavities are often observed. Biologically functional ligands usually employ only one or a few pockets/cavities, those at the active site. A new program, CAST, has been developed for locating pockets and cavities in protein crystal structures and quantifying their size. The method is a computational geometry treatment of complex shapes, based on alpha shape and discrete flow theory, and a related suite of programs, of Edelsbrunner and colleagues (Edelsbrunner & Mucke, 1994; Edelsbrunner et al., 1995, 1996; Facello, 1995). CAST makes new extensions to current alpha-shape software, and provides a full description of protein pockets and cavities, including volume, surface area, protein atoms that line the concavity, and features of pocket mouth(s) including identification of mouth atoms as well as measurement of mouth area and circumference.

The computational methods employed by CAST, analysis of more than 100 protein structures, and web access to CAST are

Reprint requests to: Clare Woodward, Department of Biochemistry, University of Minnesota, 1479 Gortner Avenue, St. Paul, Minnesota 55108; e-mail: clare@biosci.cbs.umn.edu.

[3]Current address: Department of Cheminformatics, SmithKline Beecham Pharmaceuticals, 709 Swedeland Rd., PO Box 1539, UW2940, King of Prussia, Pennsylvania 19406-0939.

reported here. Statistical studies of total pockets and cavities, and of pockets and cavities known to be ligand binding sites, were carried out. Also, protein–ligand complexes of elastase, FK506 binding protein, and HIV-1 protease were examined in detail. Results indicate that CAST will be useful in important aspects of drug design, for example, (1) identification and measurement of the most probable active sites in a ligand-free enzyme or receptor structure, (2) identification and measurement of "unoccupied" space in the active site, and in nearby concavities, for incorporation into drug design strategies.

Methods

The alpha-shape and discrete-flow methods (Edelsbrunner & Mucke, 1994; Edelsbrunner, 1995; Edelsbrunner et al., 1995, 1996) are applied to protein binding site analysis with a significant new feature, the measurement of pocket size. A new program, CAST, (1) identifies the atoms forming pockets, (2) computes volume and area of pockets, (3) identifies atoms forming "rims" of the pocket mouth(s), (4) computes the number of mouth openings for each pocket, (5) computes the area and circumference of mouth openings, and (6) locates cavities and measures their size. Steps 1, 3, 4, and 6 employ previously reported methods (Edelsbrunner et al., 1995, 1996; Liang et al., 1998b); steps 2 and 5 are newly developed. Mathematical and technical details are described in references cited above. An introduction to alpha shape theory can be found on the web site http://alpha.ncsa.uiuc.edu/alpha. Here, we give a brief, qualitative explanation of the basic concepts used in CAST. Technical words are italicized when first mentioned.

A highly simplified model of a two-dimensional molecule formed by atom disks of uniform radius is shown in Figure 1A. If nails are figuratively hammered into the plane at each atom center, and a rubber band is stretched around the entire collection of nails, the band encloses a *convex hull* of the molecule, containing all atom centers within. The convex hull of the disk centers in Figure 1A is the shape enclosed by the outer boundary of the polygon in Figure 1B (shaded area). It can be triangulated, i.e., tessellated with triangles so that there is neither a missing piece, nor overlap, of the triangles. Triangulation of a convex hull is shown in Figure 1B, where triangles tile all of the shaded convex hull area. This particular triangulation, called the *Delaunay triangulation*, is especially useful because it is mathematically equivalent to another geometric construct, the Voronoi diagram (all dashed lines in Fig. 1A).

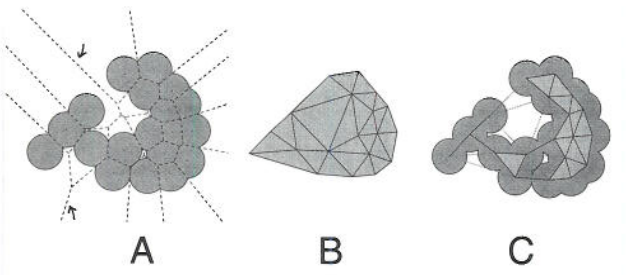

Fig. 1. Illustration of concepts in alpha shape theory. **A:** A two-dimensional molecule consisting of disks of uniform radii. The dashed lines show the Voronoi diagram of the molecule. Arrows indicate 2 of the 10 Voronoi edges that are completely outside the molecule. **B:** The convex hull of the atom centers in Figure 1A (all shaded area) with Delaunay triangulation (triangles defined by dark lines). **C:** The alpha shape of the molecule in Figure 1A. The alpha shape, or dual complex, consists of the light-shaded triangles, the dark line segments, and the atom centers. There are 10 shaded line segments corresponding to the 10 Voronoi edges that are completely outside the molecule. Any triangle with one or more shaded edges is an "empty triangle." A void formed by three empty triangles can be seen at the bottom center. It encloses a molecular cavity.

The Voronoi diagram is formed by the collection of Voronoi cells. For the hypothetical molecule in Figure 1A, Voronoi cells include the convex polygons bounded all around by dashed lines, as well as the polygons with edges defined by dashed lines, but extending to infinity. Each cell contains one atom, and those extending to infinity contain boundary atoms of the convex hull. A Voronoi cell consists of the space around one atom so that the distance of every spatial point in the cell to its atom is less than or equal to the distance to any other atom of the molecule. The Voronoi diagram has many applications in chemistry and biology (Finney, 1975; Richards, 1977; David & David, 1982; Gellatly & Finney, 1982; Richards, 1985; Procacci & Scateni, 1992; Gerstein et al., 1995). The Delaunay triangulation can be mapped directly from the Voronoi diagram. Across every Voronoi edge separating two neighboring Voronoi cells, a line segment connecting the corresponding two atom centers is placed. For every Voronoi vertex where three Voronoi cells intersect, a triangle whose vertices are the three atom centers is placed. In this way, the full Delaunay triangulation is obtained by mapping from the Voronoi diagram. That is, both the Delaunay triangulation and the Voronoi diagram contain equivalent information.

To obtain the *alpha shape*, or *dual complex*, the mapping process is repeated, except that the Voronoi edges and vertices completely outside the molecule are omitted (two such edges are indicated by arrows in Fig. 1A). Figure 1C shows the dual complex for the two-dimensional molecule in Figure 1A. The omitted edges of the Delaunay triangulation are the dotted edges in Figure 1C; a triangle with one or more dotted edges is designated an "empty" triangle (although not all empty triangles have dotted edges). The dual complex and the Delaunay triangulation are two key constructs that are rich in geometric information; from them the area and volume of the molecule, and of the interior inaccessible cavities, is measured. As an example, a void at the bottom center in the dual complex (Fig. 1C) is easily identified as a collection of empty triangles (three in this case) for which the enclosing polygon has solid edges. There is a one-to-one correspondence between such a void in the dual complex, and an inaccessible cavity in the molecule. The actual size of the molecular cavity is obtained by subtracting from the sum of the areas of the triangles, the fractions of the atom disks contained within the triangle. Details for computing cavity area and volume are in Edelsbrunner et al. (1995) and Liang et al. (1998b).

For identifying and measuring pockets, the *discrete-flow* method is employed. For the two-dimensional model, discrete flow is defined only for empty triangles, that is, those Delaunay triangles that are *not* part of the dual complex. An obtuse empty triangle "flows" to its neighboring triangle, whereas an acute empty triangle is a *sink* that collects flow from neighboring empty triangles. Figure 2A shows a pocket formed by five empty Delaunay triangles. Obtuse triangles 1, 4, and 5 flow to the sink, triangle 2. Triangle 3 is also obtuse; it flows to triangle 4, and continues to flow to triangle 2. All flows are stored, and empty triangles are later merged when they share dotted edges (dual, non-complex edges). Ultimately, the pocket is delineated as a collection of empty triangles. The actual size of the molecular pocket is computed by

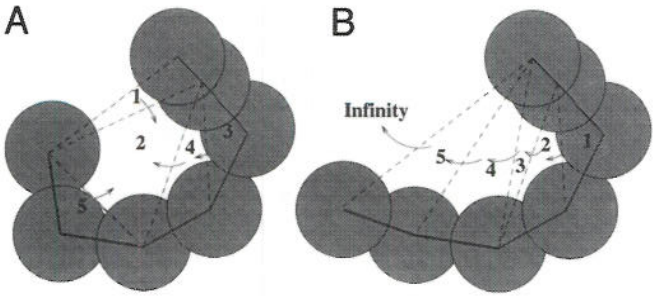

Fig. 2. Illustration of discrete flow for two-dimensional pockets. **A:** Discrete flow of a pocket. Obtuse empty triangles (1, 3, 4, and 5) flow to the acute triangle (2). Collectively, they form a pocket of the dual complex, which can be mapped to the molecular pocket. **B:** A depression for which obtuse triangles sequentially flow to the outside (to infinity). Depressions of this type are not identified as pockets in this paper.

subtracting the fractions of atom disks contained within each empty triangle. The two-dimensional *mouth* is the dotted edge on the boundary of the pocket (upper edge of triangle 1, in this case). The size of the mouth of the pocket is the length of the outside boundary edge of triangle 1, minus the two radii of the atoms connected by the edge. The type of surface depression not identified as a pocket is illustrated in Figure 2B; it is one formed by five obtuse triangles that flow sequentially from 1 to 5 to the outside, or infinity.

All the features of the two-dimensional description have more complex three-dimensional counterparts. The convex hull in three dimensions is a convex polytope instead of a polygon, and its Delaunay triangulation is a tessellation of the polytope with three-dimensional tetrahedra. When atoms have different radii, the *weighted Delaunay triangulation* is required, and the corresponding weighted Voronoi cells are also different (Edelsbrunner, 1995; Liang et al., 1998a). Figures 3A and 3B show the pockets and cavities of staphylococcal nuclease (1snc) identified by CAST. The empty Delaunay tetrahedra, the three-dimensional counterparts of empty triangles in Figure 2C, are visualized in Figure 3A. The tetrahedra represent negative images of pockets and cavities, and the wire frame represents the alpha shape of the molecule. Figure 3A contains precise information for size calculation; purple surfaces are pocket mouths open to bulk solvent, while gold surfaces line the interior of the pocket or cavity. Pockets/cavities in Figure 3B correspond to the empty tetrahedra in Figure 3A. Atoms lining the pockets and cavities are shown as spheres, each pocket/cavity in a different color. Among the proteins surveyed, staphylococcal nuclease has the least number (11) of cavities and pockets; all are displayed in Figure 3B. A binding-site pocket or cavity is defined as any pocket or cavity in which at least one atom is in contact with ligand. CAST identifies one large pocket (green) in the staphylococcal nuclease active site; mouth rim atoms are darker green. Note that mouth area is not the surface area of mouth rim atoms; rather it is the area of the mouth opening (computed from the purple Delaunay triangles in Fig. 3A, by subtracting the fractions of the triangles occupied by rim atoms).

In the current implementation of CAST, if two neighboring empty tetrahedra share a common triangle, then they belong to the same pocket. If two empty tetrahedra do not share an interfacial triangle, but share a common edge, then they belong to different pockets, if each flows to a different sink. An example is Figure 5C, which shows the active site of porin composed of three pockets. In the computation of CAST parameters, each atom shared by two pockets is cut by a bisector plane, and the shared edge lies in the bisector plane. Each half belongs to a different neighboring pocket, and should be colored accordingly. However, current software does not accommodate a user-defined bisector plane as a color divider within a single atom.

CAST computation of pockets and cavities involves no human interactions. The only parameters required are the atomic van der Waals radii, and the radius of the probe sphere (usually 1.4 Å to approximate water). The probe sphere is used only to define molecular surface. All water molecules are first removed. The analytical area of each atom of the molecule, before and after removal of the ligand, is computed by an alpha-shape–based program VOLBL (Edelsbrunner et al., 1995; Liang et al., 1998a). Atoms showing a difference in area are designated as "in contact" with ligand. Pockets and voids are then computed for the ligand-removed protein, and those containing any atoms "in contact" with the ligand are selected and defined as (part of) the ligand binding pocket. The initial stage of the computational procedure involves three steps: (a) the atomic radius for each atom of the PDB file is assigned using the utility program PDB2ALF; (b) the program DELCX is used to compute the three-dimensional weighted Delaunay triangulation, which takes into account the nonuniform nature of the atom radii; (c) the alpha shape or dual complex is computed using the program MKALF. The outputs of these three programs may then be utilized in three additional programs: (1) For measurements of pockets and cavities, CAST accesses information computed by MKALF and computes the volume and area of pockets/cavities, as well as mouth area and circumference of pockets. Molecular visualizations of pockets and cavities are generated using RASMOL (Sayle & Milner-White, 1995) with pocket atoms specified by CAST. (2) VOLBL can be used to compute the area and volume of each atom and of the whole molecule, and to identify and measure inaccessible cavities in the molecule. (3) ALVIS may be used for visualization of cavity/pockets in the form of Delaunay tetrahedra (e.g., Fig. 3B). The outputs of all computations give two sets of parameters, one based on molecular surface (MS) (Connolly, 1983) and one based on solvent accessible surface (Lee & Richards, 1971). Only MS measurements are reported in this paper. For the protein acetylcholinesterase (1ack, 531 residues, 4,099 protein atoms), on an SGI indigo 2 (195 MHz R10000 processor with 128 Mb memory), run times for PDB2ALF, DELCX, MKALF, and CAST are, respectively, 1, 29, 61, and 8 s (total about 1 min 40 s). For the 24mer of ferrittin (1rcd), the largest protein we have run (33,552 atoms), CAST takes about 1 min 53 s, and the full computation takes about 14 min. The visualization step for this protein takes about 8 min for RASMOL to load the space filling atoms on the screen. All of the above programs, except CAST, can be downloaded from the National Center for Supercomputing and Applications at http://alpha.ncsa.uiuc.edu/alpha. A CAST web server is provided at http://sunrise.cbs.umn.edu/cast. Users may obtain pocket and cavity parameters for any structure in the Brookhaven protein data bank, or for molecular coordinates uploaded in the PDB file format. Results of the CAST calculations will be sent to the user by e-mail, together with a RASMOL script for visualization. In this paper, color figures are generated by VMD (Humphrey et al., 1996) and rendered through Raster3D (Merritt & Bacon, 1997); the surface in Figure 8a is generated by SURF (Varshney et al., 1994) as bundled in VMD.

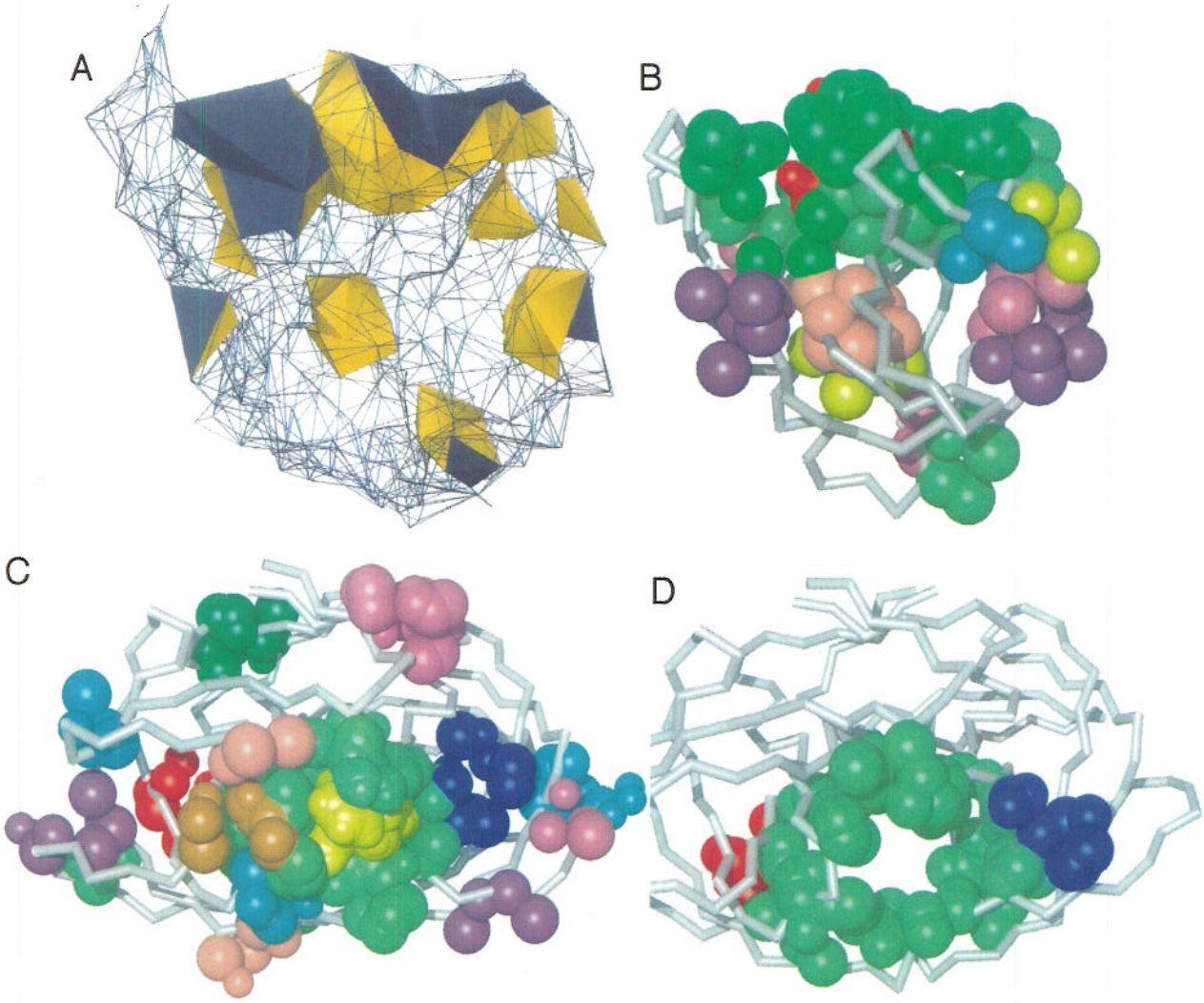

Fig. 3. Pockets, cavities, and inhibitor binding sites of staphylococcal nuclease and HIV-1 protease. **A:** The empty Delaunay tetrahedra for pockets and cavities in staphylococcal nuclease. The wire frame represents the alpha shape of the molecule, and tetrahedra represent negative images of pockets and cavities. The purple-colored surfaces are pocket mouths at the bulk solvent interface. Gold surfaces line the interior of pockets/cavities. The orientation is the same as in **B**. **B:** Pockets and cavities of staphylococcal nuclease (1snc) corresponding to **A**. Atoms lining the concavities are spheres, with each pocket/cavity in a different color. The largest pocket (green) binds the ligand (red). Atoms that rim the mouth of the active site pocket are darker green. **C:** The pockets of HIV-1 protease complexed with an inhibitor (1hvi). Each of the 18 pockets is shown in a different color. The inhibitor binding site (green), shown with the inhibitor (yellow), is 1,566 $Å^3$ and is the largest among the complexes analyzed. Two symmetrical pockets not occupied by ligand, one in each subunit, are near the mouths of the inhibitor binding site (red, rear left) and (dark blue, front right). **D:** The active site pocket (green) in ligand-free HIV-1 protease (3hvp, dimer generated by symmetry). Two nearby auxiliary pockets (red and blue) are present, essentially the same as the red and blue pockets in **C**.

Results and discussion

Pocket and cavity analysis

Shape and size parameters of protein pockets and cavities are important for active site analysis and structure-based ligand design. Research toward this end includes analytical area and volume calculation (Connolly, 1983; Richmond, 1984; Gibson & Scheraga, 1987); cavity identification and measurement (Rashin et al., 1986; Voorintholt et al., 1989; Ho & Marshall, 1990; Alard & Wodak, 1991; Nicholls et al., 1993; Smart et al., 1993; Hubbard et al., 1994; Kleywegt & Jones, 1994; Williams et al., 1994); pocket or cleft computation (Kuntz et al., 1982; Delaney, 1992; Levitt & Banaszak, 1992; Laskowski, 1995; Peters et al., 1996); and molecular shape representation (Lin et al., 1994). Although useful, their application to pocket calculations is limited by lack of fully automatic computations, lack of analytical measurements of area and volume with real physical meaning, and/or use of arbitrarily adjusted parameters.

Recent developments in the field of computational geometry provide the means of overcoming these limitations. The methods are based on fast and precise algorithms for weighted Delaunay triangulation and alpha shape (Edelsbrunner & Mucke, 1994; Facello, 1995; Edelsbrunner & Shah, 1996). Atoms are treated as discrete objects and numerical approximations involving, for ex-

ample, grids or dots are not employed (Edelsbrunner, 1995). The alpha-shape theory is used for analytical computation of protein volume and area, and for cavity identification and measurement (Edelsbrunner et al., 1995, Liang et al., 1998a, 1998b). It is also applied to binding site identification (Peters et al., 1996), molecular mesh generation (Akkiraju & Edelsbrunner, 1996), and electrostatics calculations based on boundary elements (Liang & Subramaniam, 1997). Other computational geometry methods applied to biological questions include unweighted Delaunay tetrahedra for analysis of protein structures (Singh et al., 1996) and for recognition of protein folds in sequence threading (Munson & Singh, 1997). A recent advance in characterization of protein pockets is the discrete flow method used in conjunction with alpha-shape theory (Edelsbrunner et al., 1996); it has been applied to estimation of hydration changes due to osmotic stress of antithrombin (Liang & McGee, 1998), identification of potential binding sites for calcium ions in the Gla domain of prothrombin (McGee et al., 1998), and identification of the D-site binding pocket thought to contain the proton acceptor of tyrosine radical in photosystem II (Kim et al., 1997).

An extension of discrete flow methods to measurement of pocket and pocket mouth size is implemented in the program CAST, which computes volume, surface area, and protein atoms lining the concavity of pockets and cavities, along with rim atoms, area, and circumference of pocket mouths. CAST operations are completely automatic, and calculated quantities have real physical meaning. New features computed by CAST, and not by other programs, are mouth parameters. Pocket mouths are analytically quantified in terms of the number of mouths, the area of the mouth opening(s), the identification of "rim" atoms, and the circumference of the mouth(s). Mouth parameters may be useful in analyses of molecular recognition and binding. Even though mouth openings are no doubt flexible, quantification of binding pocket mouth number and size in static crystal structures provide an indication of ligand accessibility to the pocket interior. One type of surface concavity not identified as a pocket by the present implementation of CAST is the case in which discrete flow goes to infinity (Fig. 2B). This occurs for shallow depressions with mouth openings wider than any cross section of the depression interior, analogous to a soup plate. We have experimented with an alternative implementation where discrete flows are constructed and maintained for shallow depressions. In the results, numerous tiny pieces of shallow depressions are identified. In general, these cluttered small defects distract from the main features of the pockets of interest. A mathematically and algorithmically satisfactory solution for the shallow depressions is a research subject we are exploring further.

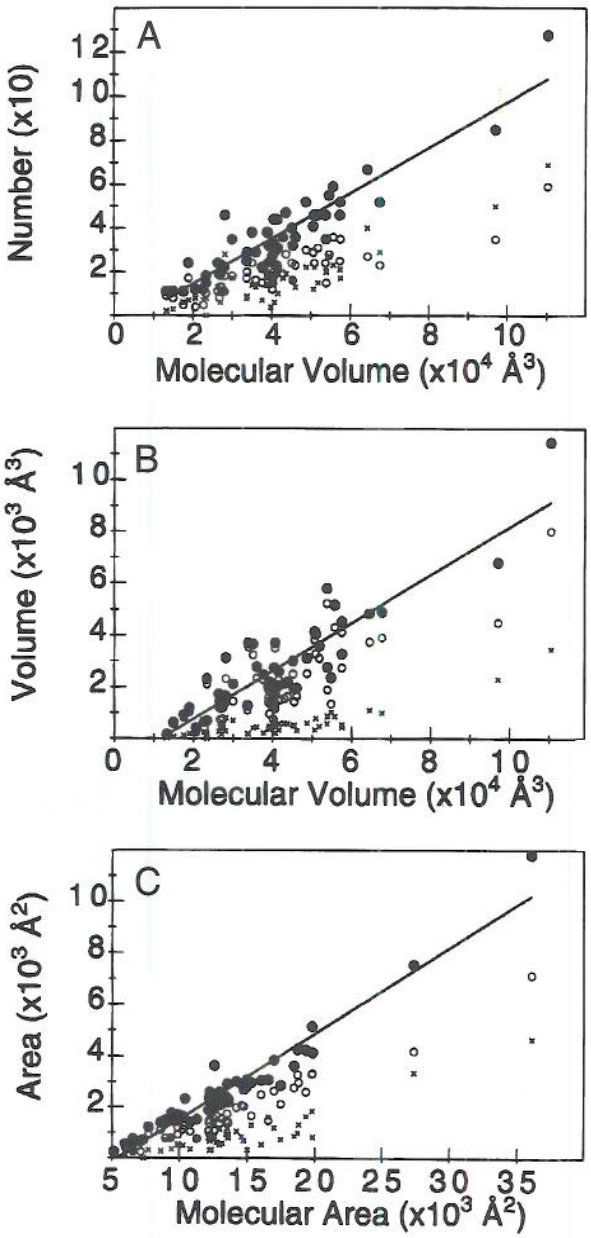

Fig. 4. Pockets and cavities statistics of a set of 51 monomeric enzymes. **A:** The total number of all pockets plus cavities (filled circles) of a protein is linearly correlated with the protein molecular surface volume (MS volume, or Connolly volume). The numbers of pockets (open circles) and cavities (crosses) when plotted separately are also linearly correlated with the protein MS volume. **B:** The total MS volume of pockets plus cavities (filled circles) is linearly correlated with the protein MS volume. When plotted separately, the volume of pockets, but not cavities, is also linearly correlated with the protein MS volume. **C:** The total MS area of pockets plus cavities (filled circles) is linearly correlated with the protein MS area.

Pocket and cavity statistics

The collective properties of protein pockets and cavities for 51 monomeric enzymes were analyzed. Number and size statistics of all pockets and cavities identified by CAST were determined for the proteins listed in Electronic Supplementary Material. Although CAST computes parameters for both solvent accessible surface (Lee & Richards, 1971) and molecular surface (MS) (Connolly, 1983), only MS measurement are reported here. The number of pockets and cavities for proteins in the data set is linearly correlated with protein MS volume (Fig. 4A, filled circles). The MS volume of a protein is the volume enclosed by the envelope inscribed by the protein molecular surface, minus the volume of the interior cavities. Roughly, an increase of 1,000 Å^3 of protein volume supports about one additional pocket or cavity. When pockets and cavities are distinguished, the numbers for both are still linearly correlated with protein volume (Fig. 4A, open circles and crosses). Total volume of pockets plus cavities is also correlated with protein volume, but when plotted separately, the volume of pockets, but not cavities, shows a correlation with protein volume (Fig. 4B). As expected from the volume correlation, total protein MS area is also correlated with MS area of pockets plus cavities

(Fig. 4C). The MS area of the protein is the area of the envelope of the protein molecular surface, plus the area of the interior inaccessible cavities.

In pockets, hydrogen-bond donor or acceptor groups lining the surface are presumably hydrogen bonded to water, which are commonly (but not always) crystallographically observable. Cavities of small size may contain water(s), while those of larger size may accommodate non-native ligands. For example, the noble gas xenon diffuses into the myoglobin interior; its four crystallographically observed xenon-binding sites were identified in the native form (no xenon) as cavities and measured using the alpha shape method (Liang et al., 1998b).

Pockets and cavities with known ligands

Of the many surface pockets and buried cavities in a typical protein, only one, or a few, bind biological ligands. Less frequently, a binding site is formed by several neighboring pockets/cavities. The remarkable variation in known ligand binding pockets and cavities is illustrated in Figure 5. A binding site may be a single spherical cavity (endonuclease, Fig. 5A), a deep pocket of simple geometry (acetylcholinesterase, Fig. 5B), a curved groove composed of several connected pockets (porin, Fig. 5C; ribonuclease, Fig. 5D), or a branched structure (thioredoxin reductase, Fig. 5F) with ligand bound in one branch and crystallographic waters bound in the other. NADH peroxidase binding site has four branched grooves within the same pocket (Fig. 5G); two grooves bind FAD and NADH, while the others contain water.

Ligand binding sites vary widely in size. Among the sites with bound ligand in the set of monomeric enzymes listed in Electronic Supplementary Material, most are on the order of 10^2–10^3 Å^3. An exception is the bacteriochlorophyll A protein (Fig. 5H), which has an unusually large binding pocket, enclosed by anti-parallel β-sheet. This is the largest binding site in the set, with an MS volume of 10,438 Å^3 and a surface area 4,319 Å^2. The ligand is a complex of seven bacteriochlorophyll A molecules containing seven Mg^{+2} ions. The complex has a total volume of 7,281 Å^3 and occupies 70% of the binding site, while the other 30% contains water molecules. Figure 5I shows another protein with a large binding site (the 9th largest in the set). Glycogen phosphorylase binds pyridoxal 5′-phosphate; the binding site has a volume of 1,398 Å^3 and a surface area of 1,300 Å^2.

Shape complementarity, along with chemical complementarity, are underlying bases of molecular recognition. One way to evaluate the fit of a ligand shape to a binding pocket is to compare their volumes. Among the ligand binding sites listed in Electronic Supplementary Material, there is no simple correlation between size of binding sites and size of the protein. Larger proteins do not necessarily have larger binding sites; in fact, small and medium size proteins have some of the largest. This is illustrated by Figure 6A, a plot of the volume of binding sites formed by a single pocket/cavity against the volume of the protein (groups 1a and 2a, below). The size of the ligand is also not obviously correlated to the size of the protein (Fig. 6B). Figure 6C gives a plot of ligand volume vs. binding site volume; a linear correlation exists when pocket volumes are relatively small. However, there is significant variation in ligand volume when pocket size is greater than 700 Å^3. This could arise if part of the ligand surface is in contact with the protein while the rest is exposed to solvent, or, if there is significant unoccupied volume within the binding site, possibly providing space for motions of protein, ligand atoms, and/or water molecules. Unoccupied space in the active site could be useful in structure-based ligand design, as discussed below.

Is there a correlation of pocket/cavity size with ligand binding site?

It has been suggested that ligand binding sites tend to involve the largest pocket/cavity on the protein (DesJarlais et al., 1988; Laskowski et al., 1996). Our analysis reveals the same general trend, but with notable differences. Using the program SURFNET, Laskowski et al. (1996) found that ligand is bound in the largest cleft in 83% of monomeric enzymes. The table in Electronic Supplementary Material lists 51 of the 67 monomeric enzymes of the single-chain enzymes studied by Laskowski et al.; 14 of the omitted structures have binding sites with discrete flow to infinity, and two are not in current versions of PDB. For the majority of enzymes in our set, the binding site is the largest CAST-identified pocket/cavity in the molecule; for example, for enzymes with one binding site, 74% have ligand bound to the largest pocket/cavity (Table 1).

The set of monomeric enzymes may be divided into group 1 (45 examples), with one crystallographic ligand binding site, and group 2 (6 examples) with two ligand binding sites. Each group is further subdivided according to the number of pockets/cavities per binding site. Group 1a (39 examples) and group 1b (6 examples) have binding sites composed of a single pocket and several neighboring pockets, respectively. Examples of group 1a are an endonuclease (Fig. 5A), acetylcholinesterase (Fig. 5B), and thioredoxin reductase (Fig. 5F). Examples of group 1b are porin (Fig. 5C) and ribonuclease (Fig. 5D). An example of group 2 is anthranilate isomerase with two binding sites, each containing a phosphate ion (Fig. 5E).

For proteins with a single binding site composed of a single pocket/cavity (group 1a), 29 binding sites (74%) are the largest pocket/cavity (Table 1). An example of a group 1a protein in which the binding site is not the largest pocket is acetylcholinesterase, shown in Figure 5B bound to endrophonium (red) in its second largest pocket (blue), not the largest (green). For proteins with a single binding site composed of multiple pockets (group 1b), six binding sites (85%) include either the largest, or the second largest, pocket (Table 1), and often additional smaller ones. An example of the latter is the integral membrane protein, porin; the ligand binding site is a curved groove formed by three different pockets: the 2nd, 3rd, and 9th largest (Fig. 5C). For the six group 2 proteins (Table 1), 12 sites are subdivided into those formed by a single pocket/cavity (group 2a), and those involving multiple pockets/cavities (group 2b). Anthranilate isomerase (group 2a) is a bifunctional enzyme whose two nonoverlapping active sites are the largest and second largest pockets (Fig. 5E). Similarly, for penicillopepsin (1ppl, group 2a, structure not shown), the largest pocket contains one ligand, and its second largest pocket contains a different ligand. Group 2b binding sites are usually formed by several neighboring small pockets, none the largest (Table 1).

Although the trends are similar, there are significant differences between our findings and those of Laskowski et al. Among the proteins for which ligand binds to the largest cleft in our calculations, five are listed by Laskowski et al. as having ligand not bound to the largest cleft (1php, 1rnh, 1onc, 1add, 1cil). On the other hand, three proteins are listed by Laskowski et al. as binding in the largest, but which in our calculations bind to a smaller pocket. For example, in acetylcholinesterase (Fig. 5B) and mandelate rac-

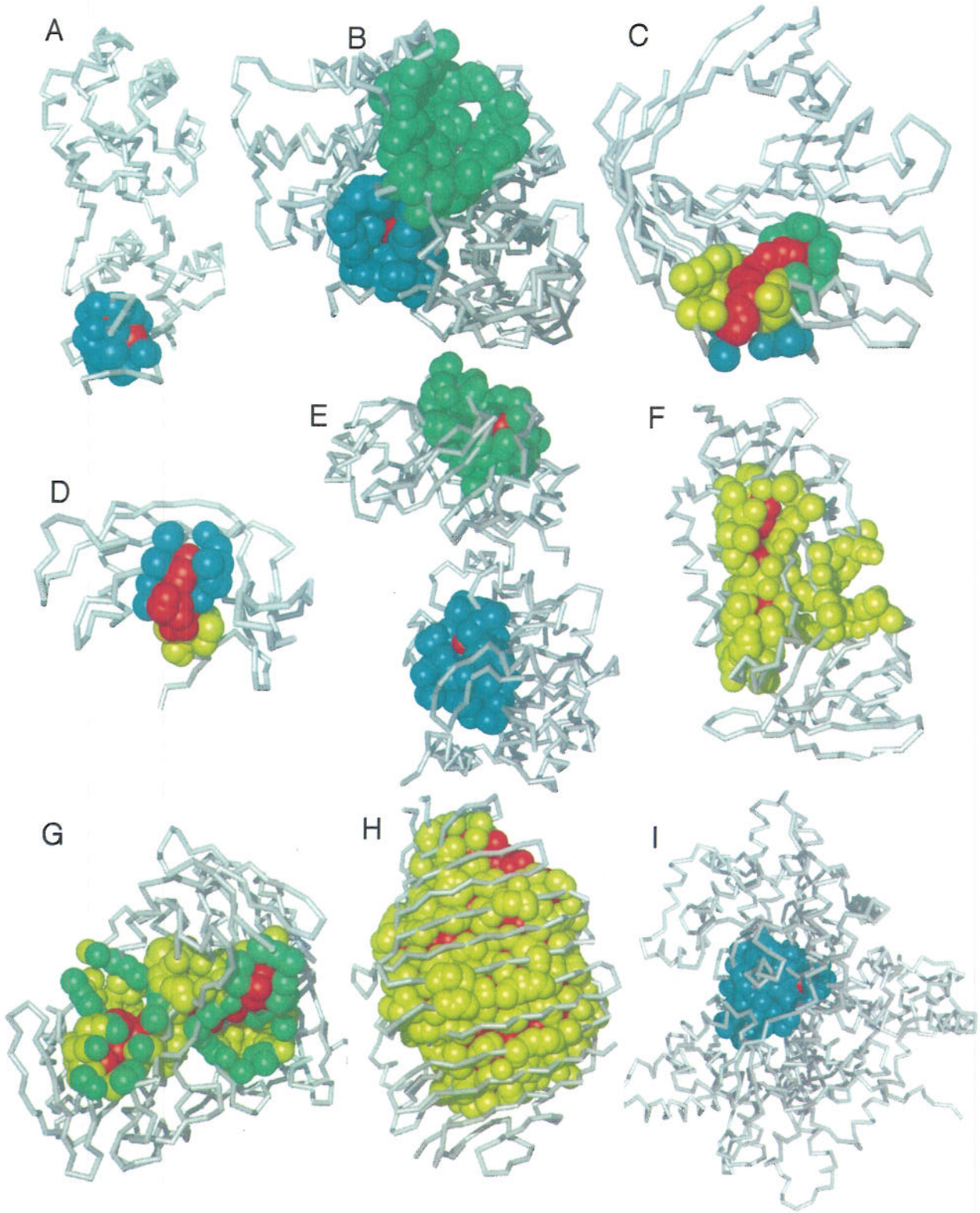

Fig. 5. Examples of protein binding sites. **A:** Endonuclease (2abk) has a simple spherical binding cavity (blue) for an iron/sulfur cluster (red). **B:** Acetylcholinesterase (1ack) binds to endrophonium (red) in its second largest pocket (blue), not the largest (green). **C:** Porin (2por) binds to N-octyltetraoxyethylene (red) in a curved groove formed by three pockets (yellow, green, and blue). **D:** Ribonuclease (1rob) has a binding site for cytidilic acid (red) formed by two pockets (blue and yellow). **E:** Anthranilate isomerase (1pii) binds two phosphates (red), each at a separate site (blue and green). Each binding pocket has unoccupied space. **F:** Thioredoxin reductase (1tde) has a binding pocket with two branches. FAD (red) binds to one branch; the other branch is ligand-free but has several water molecules (waters not shown). **G:** NADH peroxidase (2npx) has a large pocket (yellow and green) with four branches. NADH and FAD (red) occupy two branches; the other two branches vary in size and contain water molecules. The two large mouth openings (rim atoms in green), roughly in the shape of a butterfly, reveal the strong symmetry of the binding pocket. **H:** The binding pocket (yellow) of bacteriochlorophyll A protein (3bcl) is extraordinarily large (10,438 $Å^3$ in volume). Anti-parallel β-strands wrap the pocket containing the ligand, a complex of seven chlorophyll molecules and seven Mg^{+2} ions (red). **I:** The binding site (blue) of glycogen phosphorylase (1gpb) is a large inaccessible cavity that binds pyridoxal 5′-phosphate (red).

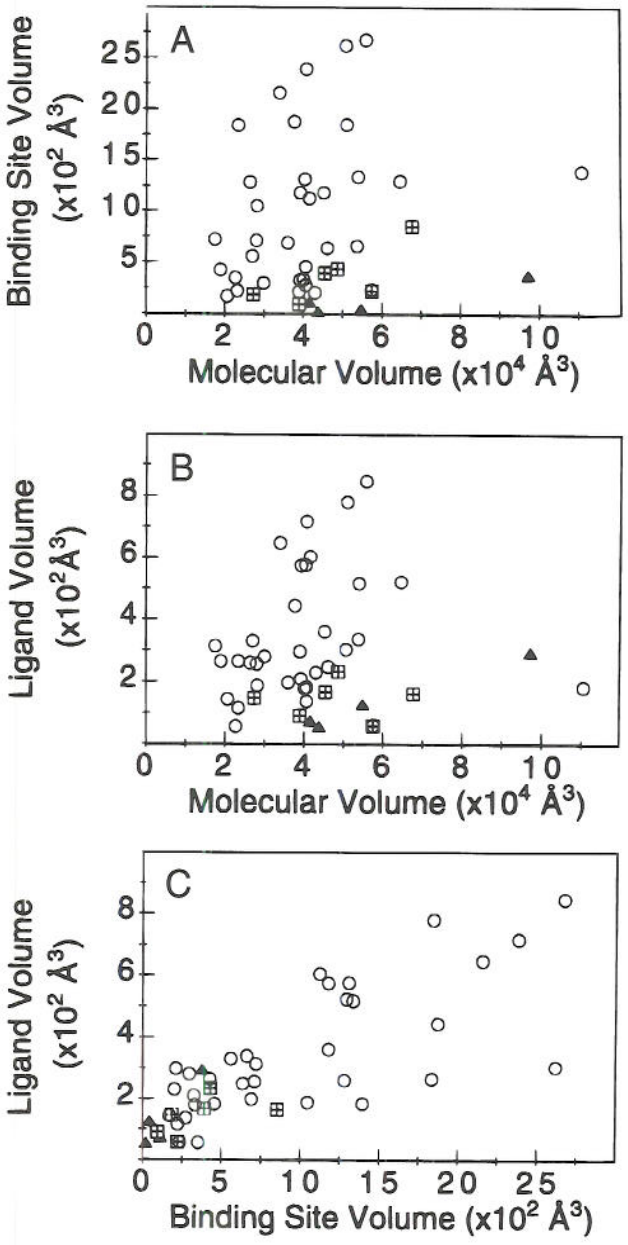

Fig. 6. The size of the protein binding sites and their ligands. Statistics are for binding sites formed by a single pocket or cavity (group 1a and group 2a proteins). Binding sites formed by the largest pocket/cavity (circles) or by the second largest pocket/cavity (divided squares) are distinguished; all others are indicated by solid triangles. **A:** A plot of binding site volume vs. protein volume. No simple correlation exists between the size of binding sites and the size of the protein. Larger proteins do not necessarily have larger binding sites. **B:** A plot of ligand volume vs. protein volume. The size of the ligand is not correlated to the size of the protein molecule. **C:** A plot of ligand volume vs. binding site volume; a linear correlation exists when pocket volumes are relatively small. Significant scattering is seen when pocket size is greater than 700 Å^3.

Table 1. *Size ranking of pockets/cavities forming ligand binding sites*

	Size rank[a]	Number of sites
Group 1a[a]	1	29
	2	4
	4	2
	5	2
	17	1
	20	1
Group 1b[a]	1	3
	2	3
	5	1
Group 2a[b]	1	4
	2	3
Group 2b[b]	2	1
	6	1
	23	1
	24	2

[a]Sizes are ranked with 1 as the largest.
[b]Group 1a sites are composed of a single pocket/cavity and on proteins with a single ligand binding site. Group 1b sites composed of multiple pockets/cavities and on proteins with a single ligand binding site. Group 2a sites are composed of a single pocket/cavity and on proteins with two ligand binding sites. Group 2b sites are composed of multiple pockets/cavities and on proteins with two ligand binding sites.

emase (1mns), ligand binds to the second largest pocket. Our results for glycogen phosphorylase are very different from Laskowski et al. The phosphorylase ligand binding cleft is the largest in their data set, and composed of an interconnected set of deep grooves on the protein exterior and spanning much of its surface. Although cleft volume in Laskowski et al. is described as having "no absolute meaning," the size they compute for glycogen phosphorylase is 20,840 Å^3. We find that the binding site of glycogen phosphorylase is a completely buried, interior cavity (Fig. 5I), with a volume of 1,398 Å^3 and a surface area of 1,300 Å^2. In the CAST calculation, even the huge binding site of bacteriochlorophyll A (Fig. 5H) is only 10,438 Å^3.

Pocket mouth statistics

The pocket analysis obtained with CAST permits for the first time a quantification of pocket mouth parameters. The number of mouths per pocket and mouth rim atoms are identified rigorously; mouth opening area as well as circumference are computed analytically. Pocket mouth atoms are color coded in Figures 3B and 5G.

Of 46 binding sites formed by a single pocket or cavity (groups 1a and 2a, Table 1), only five are cavities (Fig. 7A). The majority of binding sites are pockets with one or two mouths. For binding sites with more than two mouths, most have larger volumes, >1,800 Å^3 (Fig. 7B). Larger ligands tend to be associated with pockets of multiple mouths (Fig. 7C). However, the number of mouths of a binding site are not correlated with the volume of the protein (data not shown). Binding sites of very large proteins may be cavities or have few mouths (e.g., glycogen phosphorylase, Fig. 5I), whereas binding sites of moderate sized proteins may have many mouth openings (e.g., thioredoxin reductase, Fig. 5F, with 6 mouths). Examination of the absolute size of the mouth openings in terms of area, rather than number of mouths, leads to similar conclusions: large pockets tend to have large mouth opening area, and the total area of the molecule is not a good indicator for the total mouth opening area of its binding site(s).

The mouth openings of ligand binding pockets are fairly regular, and the area of the mouth is well correlated with the circumference

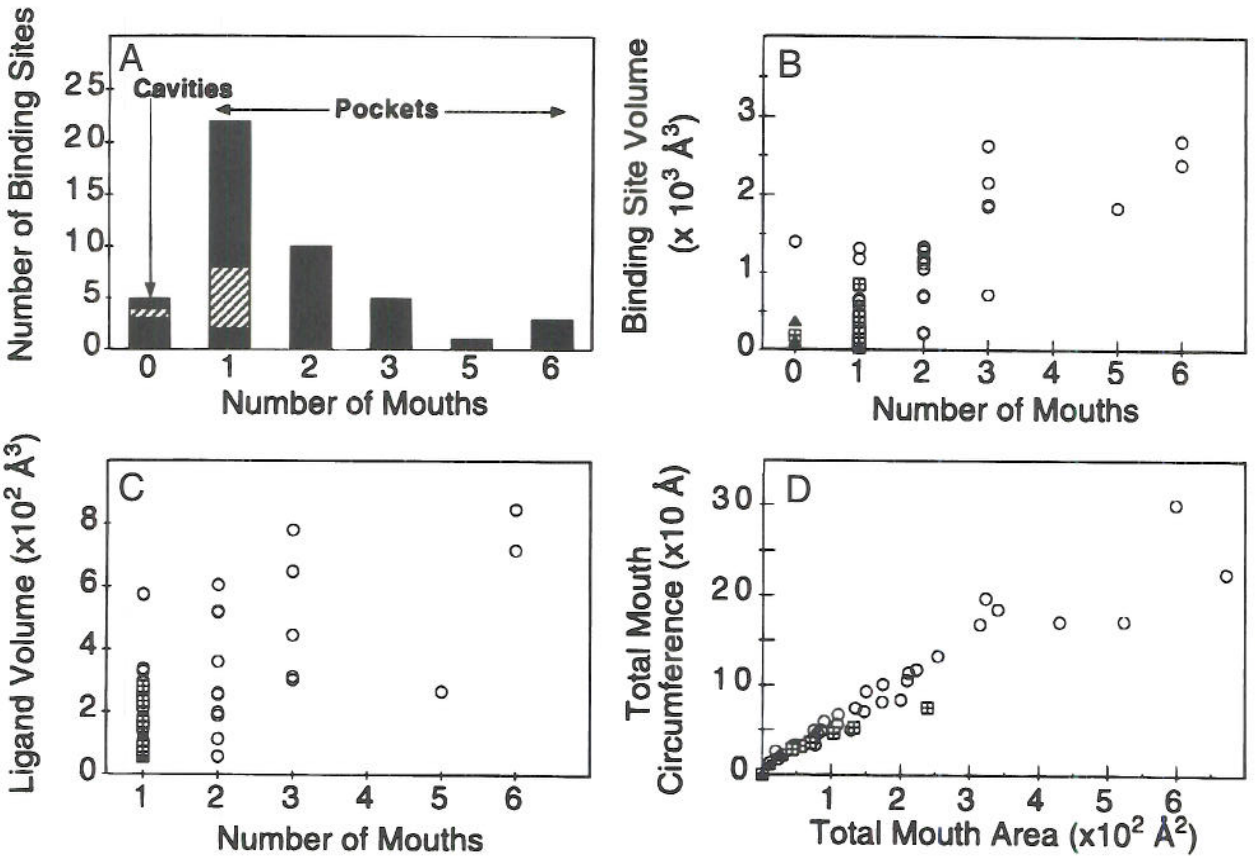

Fig. 7. The mouth openings of protein binding sites. Statistics are for binding sites formed by a single pocket or cavity (group 1a and group 2a proteins). **A:** The distribution of binding sites according to the number of mouth openings. Cavities have zero mouths. Binding sites formed by the largest pocket/cavity are in black, those formed by the second largest pocket/cavity are hatched, and the rest of the binding sites are in gray. Most binding sites have one to two mouths. **B:** The volume of binding sites plotted against the number of mouths. Binding sites with more than two mouths tend to have larger pocket volumes (>1,800 Å^3). Bacteriochlorophyll A (3bcl) is not included because its large size puts it off scale. In **B–D**, binding sites formed by the largest pocket/cavity (circles) or by the second largest pocket/cavity (divided squares) are distinguished; all others are indicated by solid triangles. **C:** Ligand volume plotted against the number of mouths of the binding site. Larger ligands tend to be associated with pockets with multiple mouths. Bacteriochlorophyll A (3bcl) is not included. **D:** Mouth perimeter plotted against mouth area of binding sites.

of the rims (Fig. 7D). This indicates that mouth openings have similar "jaggedness"; that is, a fixed mouth opening area is expected to be associated with a fixed rim boundary length. The correlation of mouth circumference with the area of the pocket, on the other hand, is rather poor (data not shown).

Comparison of CAST to other methods for characterizing protein pockets

Several popular methods for locating pockets are based on cavity-identification algorithms employing variable sized probes, such as the Connolly (1993) programs and VOIDOO in the "O" package (Kleywegt & Jones, 1994). The optimal probe radius for locating each pocket depends on local geometry, and the choice of probe size is empirical and to a large extent, arbitrary. In CAST there is no adjustment of probe radius, and no spheres of varying size are generated to fill or mold the pocket. In the computational geometry methods employed by CAST, once solvent probe radius is specified, pockets are identified and quantified without iteration.

In addition to use of adjustable probe size, computational methods for locating and measuring pockets often involve discretization steps. Discretization methods are numerical methods relying on analysis of grid points or uniformly spaced dots for identification and quantification of volume and area. For example, several methods place dots on the surface of atoms at a distance equal to the van der Waals plus probe radii, and surface components are then derived from the collection of these dots (Shrake & Rupley, 1973; Rashin et al., 1986). Alternatively, a volume box containing the protein is divided into cubic grids, and the collection of the grid points that lie outside any atom is processed for identification and quantification of cavities and pockets (Voorintholt et al., 1989; Ho & Marshall, 1990; Delaney, 1992; Kleywegt & Jones, 1994). In general, the quality of such measurement of pocket volume and area depends on the level of spacing, because a count of the number of dots/grids is the basis of the measurement. It is also influenced by the orientation of the molecule, which is mapped onto discrete grid points. Cavities or pockets of the same or slightly larger size than the grid are often not identified. Very fine grids (0.1 Å), often computationally burdensome, are necessary if pock-

ets and cavities are not to be missed. In CAST, analytical formulas are used throughout for calculating area and volume of pockets as well as cavities. No dots or grids are placed to approximate molecular shape. It should be noted that one widely used method for cavity identification (Connolly, 1983) is analytical in nature, and a number of investigators use the Connolly cavity program, along with an adjustable probe size approach, to locate and measure pockets.

A fundamental issue in pocket computation is defining where the pocket begins and the outside solvent space ends. This is related to the "can-of-worms problem of molecular speleologists," described by Kleywegt and Jones (1994). Briefly, for grid-based methods, the pocket-bounding cubic box often contains part of the outside bulk solvent space, because the pocket mouth opening is rarely the planar bounding face of the box. This may happen with either inflated atomic radii or the van der Waals atomic radii. This problem is often dealt with by repeated computation on multiple copies of the molecule in different orientations, where computed volume may differ by 20%. For example, a recommended recipe for the VOIDOO program in the "O" package is 10 different computations on randomly rotated copies of the molecule (Kleywegt & Jones, 1994). Random errors may average out, and a standard deviation of the computed pocket/cavity volume can provide estimation of the precision of the calculation. In CAST, the extent of the pocket is rigorously defined by the mouth triangles, a unique subset of the unique Delaunay triangles determined only by the relative positions of atoms, the probe radius, and the atom radii. The boundary of pocket and bulk solvent is objectively defined in CAST, and there is no can of worms.

Peters et al. (1996) use alpha shape for automatic identification of pockets in the program APROPOS, built on the Delaunay triangulation program DELCX (REGTRI) and the alpha shape program MKALF. This method is equivalent to computing the difference between two alpha shapes, each with a different α value (one shape with α = 20.0 Å, and another with α between 3.5 and 4.5 Å). Shapes with different α values are analogous to dual complexes obtained with different sized solvent probes. CAST does not involve more than one alpha value. This difference in methodology leads to a number of differences in results. First, CAST identifies all atoms for each computed pocket, as well as all rim atoms lining the mouth openings. APROPOS usually cannot determine all atoms of a pocket: it identifies the deeper region, in some cases about half of the atoms of a pocket. Second, CAST computes the volume and area of pockets and their mouths, parameters not computed in APROPOS.

FKBP and other proteins analyzed by Peters et al. (1996) offer a convenient frame of reference for comparison of CAST and APROPOS calculations. In CAST calculations, FKBP has four pockets and two cavities. The largest is the FK506 binding site, lined by 38 atoms; the second largest, near the active site, is lined by 14 atoms (Fig. 8D, and see below). APROPOS locates the FK506 binding site with 18 atoms; the auxiliary pocket near the binding site (blue in Fig. 8D) is not identified. In a group of 16 small proteins, mainly ferredoxins and cytochromes (5cyt; 1cth; 2cdv; 2cy3; 3fxc; 1fxd; 1fxa), binding sites for seven proteins are not identified by APROPOS, but are by CAST. Among the nine small proteins where APROPOS succeeds, binding sites of eight are also computed by CAST; the exception is glutaredoxin (1aba), whose binding site flows to infinity. Peters et al. also report that APROPOS does not detect pockets in small proteins, some of which function as inhibitors of other proteins and so do not have obvious concave binding pockets/cavities. CAST consistently identifies pockets in these proteins (BPTI, 6pti, and uteroglobin, 1utg); often, the pockets contain crystallographic waters. For larger inhibitors, APROPOS does not detect any of the binding sites of smaller ligands. For example, the sulfate binding site of basic fibroblast growth factor (4fgf) was not detected by APROPOS. However, CAST locates and measures this site, and many pockets in the other two large inhibitors (1tie and 1hle) mentioned by Peters et al. as having no binding sites. Lectin (1lte) has binding sites calculated by CAST, but not by APROPOS; these include a small, flat pocket in lectin that binds lactose, Ca^{+2} and Mn^{+2}. The head of another sugar molecule, the N-linked carbohydrate anchors vertically to a small pocket in lectin; this is also computed by CAST. Peters et al. (1996) suggest a generalization that binding sites become slightly more open as a result of ligand attachment. This is not the case for the elastase complex (1ela), for which the CAST-identified ligand pocket is somewhat more constricted than in ligand-free elastase (3est) (see Fig. 8A–C).

CAST identifies ancillary pockets for recruitment into ligand design

The recruitment of unoccupied pockets near the active site has been proposed for development of structure-based ligand design methods. In this context, "unoccupied" usually means "not occupied by substrate analog atoms." Based on studies of elastase–inhibitor complexes, Ringe, Petsko, and colleagues (Mattos et al., 1994; Mattos & Ringe, 1996) suggested that pockets near the active site, but having no interactions with substrate, may provide additional contact surface for designed ligands. Similarly, SAR by NMR (Shuker et al., 1996) identifies sites near the active site for linked-fragment ligand design. We expect that the use of CAST for automatic identification and measurement of proximal, unoccupied pockets will aid recruitment strategies. To explore this application of CAST, we examined pockets and binding sites in complexes of elastase–ligand, FKBP–ligand, and HIV-1 protease–inhibitor. In each case, there are clear indications of the potential of CAST for locating and characterizing neighboring pockets/cavities for active site ligand design.

Elastase

In ligand-free elastase, CAST-identified pockets in or near the active site are the large pocket (blue, Fig. 8A) and an ancillary pocket (green). Visualization of the pockets in Figure 8A, showing the molecular surface of the heavy atoms lining the pocket, is not the same as other pocket representations in this paper. It is included for comparison to the other figures, in which united atoms lining the pocket/cavity are color coded spheres. The observation of two different ligand binding modes in elastase–inhibitor complexes, shown in Figures 8B and 8C, was the basis for the suggestion of proximal pocket recruitment (Mattos et al., 1994, 1995; Mattos & Ringe, 1996). In the elastase–ligand complexes, the blue pocket in free elastase becomes two pockets, blue and yellow in Figures 8B and 8C. In a substrate-like mode, inhibitor binds to the blue and yellow pockets, leaving the green one unoccupied (Fig. 8B). In the other mode, the smaller green pocket is recruited; inhibitor binds the yellow and green pockets, leaving the blue one unoccupied (Fig. 8C). The ability of CAST to identify in free elastase the major blue binding pocket, and the nearby green pocket suggests its use in a priori identification of unoccupied pockets in and near

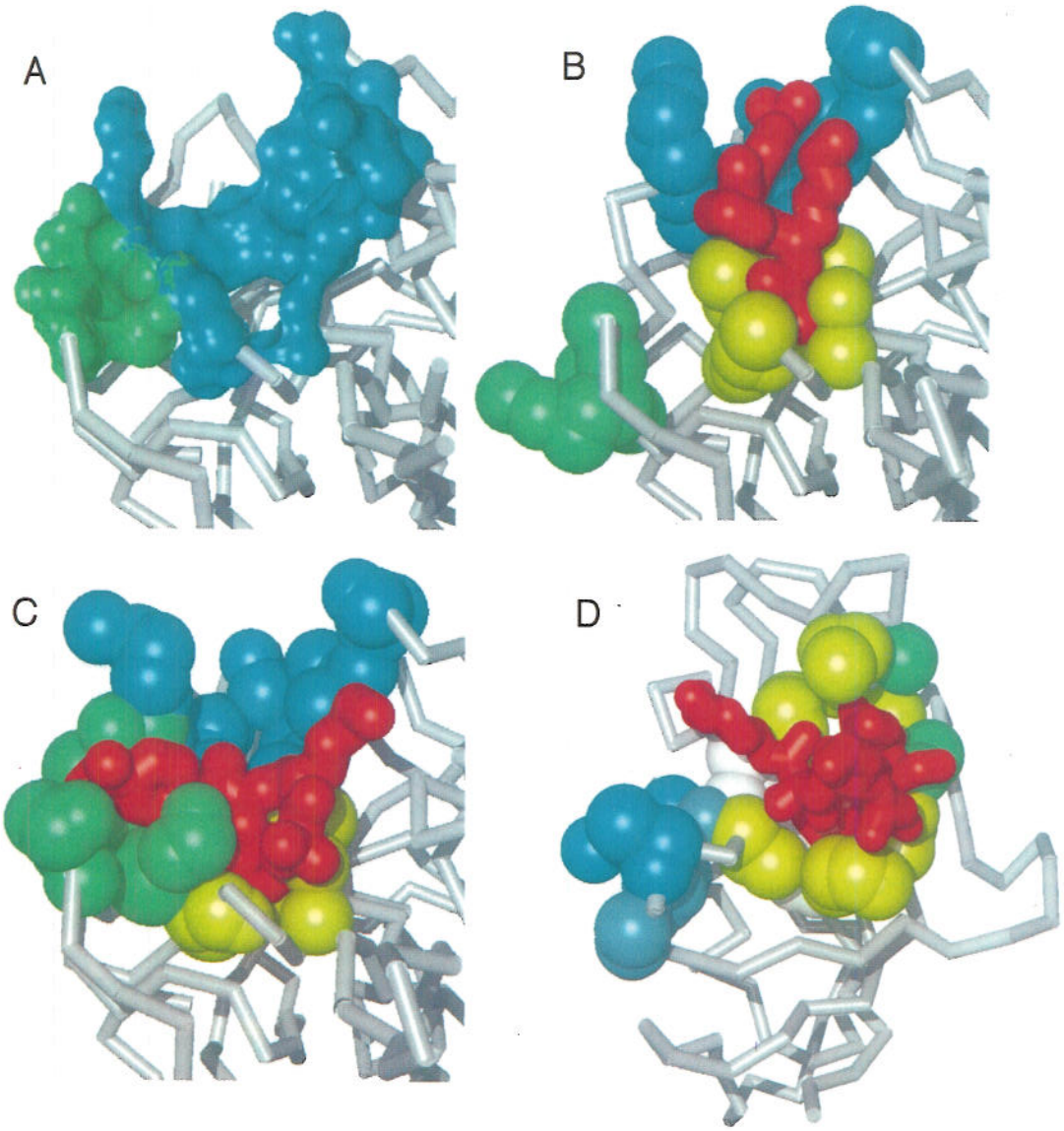

Fig. 8. Pockets in and near the active site of elastase and FKPB. Ligands are red. **A:** Ligand-free porcine pancreatic elastase (3est). The largest pocket (blue, 715 Å^3) contains the S1 and S4 subsites. A smaller pocket (green, 84 Å^3) is in the immediate vicinity of the active site. In this figure, the molecular surface (Connolly, 1983) of the pockets atoms is generated and displayed by SURF in the VMD package (see Methods). In other color figures, pocket/cavity atoms are shown as spheres representing 'united' heavy atoms. **B:** Elastase complexed with TFA-Lys-Pro-ISO (1ela). The ligand binds in a mode similar to substrate. The large active site pocket in ligand-free elastase closes somewhat to give two pockets; one binds TFA (yellow, 64 Å^3) and roughly corresponds to site S1, while the other binds ISO (blue, 218 Å^3) and is roughly site S4. Lys and Pro residues in the ligand have extensive exposed area. The green pocket similar to the one identified by CAST in the ligand-free structure is also observed. **C:** Elastase complexed with TFA-Lys-Phe-ISO (1elc) binds in a different mode. The Phe ring fills the CAST pocket corresponding approximately to the S1 subsite (yellow, 57 Å^3). The ISO group binds to the green pocket (88 Å^3). The blue pocket is not computed by CAST because it flows to infinity; blue atoms are identified on the basis that they are pocket atoms in **B** (1ela). **D:** The two largest pockets of the protein FKBP in the complex FKBP-FK506 (1fkf). The largest pocket (yellow, green, white) is the binding pocket for FK506 (red). The second largest pocket (blue) is in close proximity. Atoms shown by NMR experiments (Shuker et al., 1996) to be in residues making strong NOEs with the first compound, the second compound, and both compounds are green, dark blue, and white, respectively (see text).

the active site to provide extra interaction surface for designed ligands.

FK506 binding protein

In SAR by NMR (structural activity relationship by NMR), two compounds that bind weakly to each of two different sites near or in the active site are identified, then covalently linked to form a strongly binding inhibitor. Success has been reported for FK506 binding protein (FKBP) (Shuker et al., 1996). To examine whether CAST may be useful in a priori identification of suitable target proteins for a linked-fragment strategy, we analyzed the crystal structure of FKBP. The coordinates are available only for a com-

plex of FKBP with a different ligand, the immunosuppressant FK506 (Van Duyne et al., 1993). In the ligand-removed FKBP, there are two CAST-identified pockets in or near the active site (Fig. 8D). One is the active site pocket (yellow, green, and white), the largest in the molecule (182 Å^3). Close by is a pocket unoccupied by ligand (blue), the second largest (46 Å^3). The red ligand is FK506, not the linked-fragment ligand. In the NMR experiment, residues are reported which have strong NOEs with the individual free ligands (Shuker et al., 1996); all of these residues have atoms in the two CAST-identified pockets. Atoms from residues observed by NMR to be in closest contact only with the first weakly bound compound, only with the second weakly bound compound, and with both compounds, and are colored green, dark blue, and white, respectively, in Figure 8D. As with elastase, analysis of FKBP active-site and neighboring pockets suggests strongly that CAST may be useful in a priori identification of proteins with unoccupied pockets near the active site that are suitable for recruitment strategies, including linked-fragment approaches.

HIV-1 protease–inhibitor complexes

To explore the potential of CAST in identifying pockets near the active site for inhibitor design, 30 HIV-1 protease–inhibitor complexes were examined. HIV-1 protease has perhaps the largest number of independently determined protein–inhibitor crystal structures (Vondrasek & Wlodawer, 1997), and design and analysis of its inhibitors is an area of vigorous study (Navia et al., 1989; Wlodawer et al., 1989; Wlodawer & Erickson, 1993; Gustchina et al., 1994; Vondrasek et al., 1997). The active site is symmetrically formed by both subunits of the homodimer (green, Fig. 3C,D). Two pockets near the active site, red and dark blue, are candidates for unoccupied surface that may be incorporated into a ligand design strategy. In most protease–inhibitor complexes, these pockets are composed of atoms in residues 30, 31, 74, 76, and 88 in each chain, giving symmetrical pockets in the dimer (but different from the S1–S4 subsites of Gustchina et al., 1996). We suggest that red and dark blue ancillary pockets might usefully be incorporated into protease inhibitor design. Apparently, the protein in this region is flexible enough to accommodate the ancillary pockets as part of the active site, because, in the structure of protease–inhibitor complex 1hvj, one of the symmetry-related red and blue pockets is incorporated into the main ligand binding pocket, while the other is not.

Several additional aspects of the HIV-1 protease active site are illuminated by CAST analysis. First, active site plasticity is captured by the wide distribution of binding site volumes and areas (Fig. 9A,B). The flexibility of the active site results in rearrangements of binding site groups in the presence of different ligands (Wlodawer & Erickson, 1993). The smallest binding pocket is 853 Å^3 (1bvg); the largest is 1,566 Å^3 (Fig. 3C), close to the 1,613 Å^3 in a ligand-free protease structure (Fig. 3D). Protease inhibitor binding sites usually have two mouths, giving a tunnel-like topology to the binding pocket, and the mouth opening area is correlated with binding site volume (Fig. 9C). Second, even though active site volume varies widely with ligand, non-active site pockets identified by CAST in HIV-1 protease–inhibitor complexes tend to remain in roughly the same location, and are apparently not highly sensitive to crystal lattice contacts that might affect surface side-chain orientation. The protease–inhibitor structures analyzed contain 11–20 pockets/cavities; a representative complex in Figure 3C has 18 pockets and no cavities. Occasionally, a larger

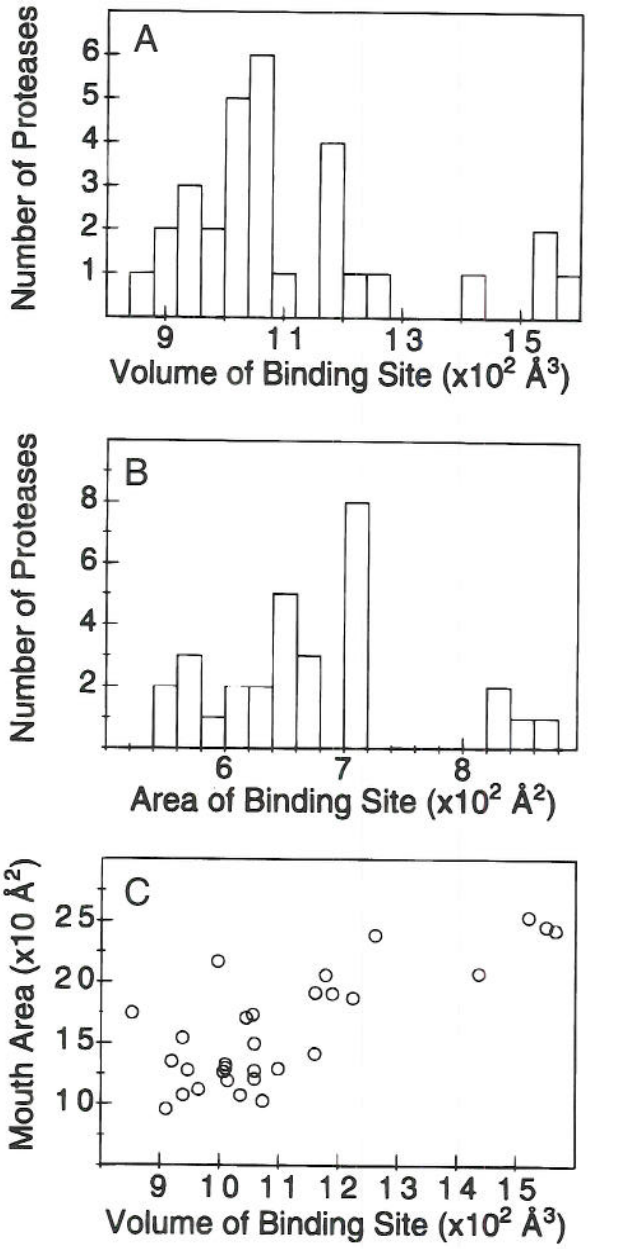

Fig. 9. The flexibility of the HIV-1 active site is illustrated by the wide distribution of binding site volumes, area, and mouth openings observed for different protease–inhibitor complexes. **A:** A histogram of the volume of the inhibitor binding site. Binding sites have a volume ranging from 853 Å^3 (9hvp) to 1,566 Å^3 (1hvi),with an average around 1,000 Å^3. **B:** The binding site area of HIV-1 proteases complexed with different inhibitors also ranges over a broad distribution (540–875 Å^2). **C:** The mouth area of the binding sites are correlated with the volume of the binding sites. Binding sites with large volume tend to have large mouths. The protease–inhibitor complexes analyzed in **A–C** are 1aaq, 1bvg, 1cpi, 1dif, 1gno, 1hbv, 1hih, 1hiv, 1hos, 1hps, 1hpv, 1hpx, 1hte, 1htf, 1htg, 1hvi, 1hvj, 1hvk, 1hvl, 1hvr, 1hxb, 1sbg, 4phv, 4hvp, 7hvp, 9hvp, 2upj, 5upj, 6upj, and 7upj.

pocket in one complex is identified as several in another complex. Many pockets/cavities involve roughly the same atoms in all complexes and in ligand-free protease.

In summary, our results suggest that CAST analysis of protein pockets and cavities will aid auxiliary pocket recruitment strategies in ligand design. Close proximity of two pockets, and their

nontrivial size, are the hallmarks of favorable targets. With further analysis of the physical-chemical properties of the pocket atoms (electrostatics, hydrophobicity, and H-bonding capabilities), we hope to incorporate prediction of optimal size, polarity, and geometry of candidate ligands. Although we emphasize the recognition of ancillary pockets for incorporation into designed ligand contact surfaces, the same principles apply to identification and utilization of "unoccupied" regions of the active site pocket itself.

Electronic supplementary material

A table of protein–ligand complexes used in statistical analyses in Figures 4, 6, and 7 and Table 1 is provided in the electronic edition of Protein Science available over Internet. The table contains the pdb name, protein name, bound ligand, and whether the bound ligand is bioactive. Proteins are divided into groups 1 and 2, as described in the text.

Acknowledgments

This work is supported by NIH Grant GM26242 to CW and a grant from NSF Institute of Mathematics and its Applications to JL.

References

Akkiraju N, Edelsbrunner H. 1996. Triangulating the surface of a molecule. *Discrete Appl Math 71*:5–22.

Alard P, Wodak SJ. 1991. Detection of cavities in a set of interpenetrating spheres. *J Comput Chem 12*:918–922.

Connolly ML. 1983. Analytical molecular surface calculation. *J Appl Crystallogr 16*:548–558.

David EE, David CW. 1982. Voronoi polyhedra as a tool for studying solvation structure. *J Chem Phys 76*:4611–4614.

Delaney JS. 1992. Finding and filling protein cavities using cellular logic operations. *J Mol Graph 10*:174–177.

DesJarlais RL, Sheridan RP, Seibel GL, Dixon JS, Kuntz ID. 1988. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *J Med Chem 31*:722–729.

Edelsbrunner H. 1995. The union of balls and its dual shape. *Discrete Comput Geom 13*:415–440.

Edelsbrunner H, Facello M, Fu P, Liang J. 1995. Measuring proteins and voids in proteins. In: *Proc. 28th annual Hawaii international conference system sciences*. Los Alamitos, California: IEEE Computer Society Press. pp 256–264.

Edelsbrunner H, Facello M, Liang J. 1996. On the definition and the construction of pockets in macromolecules. In: Hunter L, Klein T, eds. *Biocomputing: Proceedings of the 1996 Pacific symposium*. Singapore: World Scientific Publishing. pp 272–281.

Edelsbrunner H, Mucke EP. 1994. Three-dimensional alpha shapes. *ACM Trans Graph 13*:43–72.

Edelsbrunner H, Shah NR. 1996. Incremental topological flipping works for regular triangulations. *Algorithmica 15*:223–241.

Facello MA. 1995. Implementation of a randomized algorithm for Delaunay and regular triangulations in three dimensions. *Comput Aided Geomet Design 12*:349–370.

Finney JL. 1975. Volume occupation, environment and accessibility in proteins. The problem of the protein surface. *J Mol Biol 96*:721–732.

Gellatly BJ, Finney JL. 1982. Calculation of protein volumes: An alternative to the Voronoi procedure. *J Mol Biol 161*:305–322.

Gerstein M, Tsai J, Levitt M. 1995. The volume of atoms on the protein surface: Calculated from simulation, using Voronoi polyhedra. *J Mol Biol 249*:955–966.

Gibson K, Scheraga H. 1987. Exact calculation of the volume and surface area of fused hard-sphere molecules with unequal atomic radii. *Mol Phys 62*:1247–1265.

Gustchina A, Sansom C, Prevost M, Richelle J, Wodak SY, Wlodawer A, Weber IT. 1994. Energy calculations and analysis of HIV-1 protease–inhibitor crystal structures. *Protein Eng 7*:309–317.

Ho CMW, Marshall GR. 1990. Cavity search: An algorithm for the isolation and display of cavity-like binding regions. *J Comput Aided Mol Design 4*:337–354.

Hubbard SJ, Gross KH, Argos P. 1994. Intramolecular cavities in globular proteins. *Protein Eng 7*:613–626.

Humphrey W, Dalke A, Schulten K. 1996. VMD-visual molecular dynamics. *J Mol Graphics 14*:33–38.

Kim SY, Liang J, Barry B. 1997. Chemical complementation identifies a proton acceptor for redox-active tyrosine D in photosystem II. *Proc Natl Acad Sci USA 94*:14406–4411.

Kleywegt GJ, Jones TA. 1994. Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr D50*:178–185.

Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. 1982. A geometric approach to macromolecule–ligand interactions. *J Mol Biol 161*:269–288.

Laskowski RA. 1995. SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph 13*:323–330.

Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. 1996. Protein clefts in molecular recognition and function. *Protein Sci 5*:2438–2452.

Lee B, Richards FM. 1971. The interpretation of protein structures: Estimation of static accessibility. *J Mol Biol 55*:379–400.

Levitt DG, Banaszak LJ. 1992. POCKET: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J Mol Graph 10*:229–234.

Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S. 1998a. Analytical shape computation of macromolecules. I. Molecular area and volume through alpha shape. *Proteins Struct Funct Genet*. In press.

Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S. 1998b. Analytical shape computation of macromolecules. II. Identification and computation of inaccessible cavities in proteins. *Proteins Struct Funct Genet*. Forthcoming.

Liang J, McGee MP. 1998. Correlation between hydration structure of antithrombin and water transfer during reactive-loop insertion. *Biophys J 75*:573–582.

Liang J, Subramaniam S. 1997. Computation of molecular electrostatics with boundary element methods. *Biophys J 73*:1830–1841.

Lin SL, Nussinov R, Fischer D, Wolfson HJ. 1994. Molecular surface representations by sparse critical points. *Proteins Struct Funct Genet 18*:94–101.

Mattos C, Giammona DA, Petsko GA, Ringe D. 1995. Structural analysis of the active site of porcine pancreatic elastase based on the X-ray crystal structures of complexes with trifluoroacetyl–dipeptide–anilide inhibitors. *Biochemistry 34*:3193–3203.

Mattos C, Rasmussen B, Ding XC, Petsko GA, Ringe D. 1994. Analogous inhibitors of elastase do not always bind analogously. *Nat Struct Biol 1*:55–58.

Mattos C, Ringe D. 1996. Locating and characterizing binding sites on proteins. *Nat Biotechnol 14*:595–599.

McGee MP, Teuschler H, Liang J. 1998. Effective electrostatic charge of coagulation factor X in solution and on phospholipid membranes: Implications for activation mechanisms and structure–function relationships of the Gla domain. *Biochem J 330*:533–539.

Merritt EA, Bacon DJ. 1997. Raster3D: Photorealistic molecular graphics. *Methods Enzymol 277*:505–524.

Munson PJ, Singh RK. 1997. Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence–structure alignment. *Protein Sci 6*:1467–1481.

Navia M, Fitzgerald PMD, McKeever BM, Leu C-T, Heimbach JC, Herbert WK, Sigal IS, Darke PL, Springer JP. 1989. Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1. *Nature 337*:615–620.

Nicholls A, Bharadwaj R, Honig B. 1993 GRASP: Graphical representation and analysis of surface properties. *Biophys J 64*:A166.

Peters KP, Fauck J, Frommel C. 1996. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J Mol Biol 256*:201–213.

Procacci P, Scateni R. 1992 A general algorithm for computing Voronoi volumes: Application to the hydrated crystal of myoglobin. *Int J Quant Chem 42*:1515–1528.

Rashin A, Iofin M, Honig B. 1986. Internal cavities and buried waters in globular proteins. *Biochemistry 25*:3619–3625.

Richards FM. 1977. Areas, volumes, packing, and protein structures. *Annu Rev Biophys Bioeng 6*:151–176.

Richards FM. 1985. Calculation of molecular volumes and areas for structures of known geometries. *Methods Enzymol 115*:440–464.

Richmond T. 1984. Solvent accessible surface area and excluded volume in proteins: Analytical equations for overlapping spheres and implications for the hydrophobic effect. *J Mol Biol 178*:63–89.

Sayle RA, Milner-White EJ. 1995. RasMol: Biomolecular graphics for all. *Trends Biochem Sci 20*:374–376.

Shrake A, Rupley J. 1973. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J Mol Biol 79*:351–371.

Singh RK, Tropsha A, Vaisman II. 1996. Delaunay tessellation of proteins: Four body nearest-neighbor propensities of amino acid residues. *J Comput Biol 3*:213–221.

Smart OS, Goodfellow JM, Wallace BA. 1993. The pore dimensions of gramicidin A. *Biophys J 65*:2455–2460.

Shuker SB, Hajduk PJ, Meadows RP, Fesik SW. 1996. Discovering high-affinity ligands for proteins: SAR by NMR. *Science 274*:1531–1534.

Van Duyne GD, Standaert RF, Karplus PA, Schreiber SL, Clardy J. 1993. Atomic structures of the human immunophilin FKBP-12 complexes with FK506 and Rapamycin. *J Mol Biol 229*:105–124.

Varshney A, Brooks FP, Wright WV. 1994. Linearly scalable computation of smooth molecular surfaces. *IEEE Comp Graphics Appl 14*:19–25.

Vondrasek J, van Buskirk CP, Wlodawer A. 1997. Database of three-dimensional structures of HIV proteinases. *Nat Struct Biol 4*:8.

Vondrasek J, Wlodawer A. 1997. Database of HIV proteinase structures. *Trends Biochem Sci 22*:183.

Voorintholt R, Kosters MT, Vegter G, Vriend G, Hol WGJ. 1989. A very fast program for visualizing protein surfaces, channels and cavities. *J Mol Graph 7*:243–245.

Williams MA, Goodfellow JM, Thornton JM. 1994. Buried waters and internal cavities in monomeric proteins. *Protein Sci 3*:1224–1235.

Wlodawer A, Erickson JW. 1993. Structure-based inhibitors of HIV-1 protease. *Annu Rev Biochem 62*:543–585.

Wlodawer A, Miller M, Jaskolski M, Sathyanarayana BK, Baldwin E, Weber IT, Selk LM, Clawson L, Schneider J, Kent SBH. 1989. Conserved folding in retroviral proteases: Crystal structure of a synthetic HIV-1 protease. *Science 245*:616–621.