Protein Science (1996), *5*:2438–2452. Cambridge University Press. Printed in the USA.

Protein clefts in molecular recognition and function

ROMAN A. LASKOWSKI,[1] NICHOLAS M. LUSCOMBE,[1] MARK B. SWINDELLS,[2] AND JANET M. THORNTON[1]

[1]Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT, England
[2]Helix Research Institute, Inc., 1532-3 Yana, Kisarazu-shi, Chiba 292, Japan

(RECEIVED August 28, 1996; ACCEPTED September 8, 1996)

Abstract

One of the primary factors determining how proteins interact with other molecules is the size of clefts in the protein's surface. In enzymes, for example, the active site is often characterized by a particularly large and deep cleft, while interactions between the molecules of a protein dimer tend to involve approximately planar surfaces. Here we present an analysis of how cleft volumes in proteins relate to their molecular interactions and functions. Three separate datasets are used, representing enzyme–ligand binding, protein–protein dimerization and antibody–antigen complexes. We find that, in single-chain enzymes, the ligand is bound in the largest cleft in over 83% of the proteins. Usually the largest cleft is considerably larger than the others, suggesting that size is a functional requirement. Thus, in many cases, the likely active sites of an enzyme can be identified using purely geometrical criteria alone. In other cases, where there is no predominantly large cleft, chemical interactions are required for pinpointing the correct location. In antibody–antigen interactions the antibody usually presents a large cleft for antigen binding. In contrast, protein–protein interactions in homodimers are characterized by approximately planar interfaces with several clefts involved. However, the largest cleft in each subunit still tends to be involved.

Keywords: docking; enzymes; molecular recognition; protein binding sites; surface clefts

Proteins almost always interact with other molecules in performing their biological functions. These interactions include the binding of ligands in receptor sites, allosteric binding, the binding of antibodies to antigens, protein–DNA interactions, protein–protein interactions, multimerization, and protein–carbohydrate interactions. The key factors in all these interactions are the shape and chemical properties of a protein's surface. The surface is generally irregular, containing many clefts and grooves of varying shapes and sizes. These clefts are particularly important in certain types of interactions and are the subject of this paper. Clefts in protein surfaces have been studied principally because of their relevance to binding sites (for a review, see Lewis, 1991). A large cleft provides an increased surface area and, hence, increased opportunity for the protein to form interactions with other molecules, particularly small ligands. It has been suggested (Kuntz et al., 1982; DesJarlais et al., 1988) that the active site usually lies in the largest of all the protein's clefts or cavities. This tendency has also been reported by Colloc'h and Mornon (1988, 1990), though their detailed results have yet to be published (N. Colloc'h, pers. comm.). Smaller sites may also be important in some cases as in the binding of allosteric effectors (DesJarlais et al., 1988). However, despite these observations we have found no rigorous analysis of the importance of cleft size in recognition.

To perform such an analysis we need to know the volumes of every cleft in a given protein, which raises the problem of not only how to identify separate clefts in the surface, but also of how to measure a meaningful volume for each one. Internal cavities are fairly easy to define as they correspond to void regions that are completely bounded by the surrounding atoms. Several methods exist for identifying such cavities and computing their volumes, including "flood-filling" from a given starting point (Ho & Marshall, 1990), progressively "fattening" the protein's atoms until void regions are closed off from the outside world (Kleywegt & Jones, 1994), filling void regions with spheres (Smart et al., 1993; Williams et al., 1994), or as a by-product of molecular-surface generation (Voorintholt et al., 1989; Nicholls et al., 1993). Cavities have also been analyzed by groups interested in protein hydration (Rashin et al., 1986; Hubbard et al., 1994; Williams et al., 1994).

Surface clefts, on the other hand, are more difficult to define because of the problem of how far they should extend into open space; that is, where does the "sea level" of that part of the protein surface lie? A common solution is to consider all pairs of atoms in the structure and locate all void regions between them. This procedure, which restricts voids to within the outer limits of a protein's surface, manages to locate both internal cavities and surface

Reprint requests to: Janet M. Thornton, Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT, England; e-mail: thornton@bsm.bioch.ucl.ac.uk.

clefts. Void regions are often made solid either by filling them with spheres, or by marking grid-points between atoms. Either way, one obtains a "negative image" of the void regions in solid form, as is done in DOCK (Kuntz et al., 1982), POCKET (Levitt & Banaszak, 1992), SURFNET (Laskowski, 1995), and the method of Delaney (1992). Another method uses Connolly surfaces (Connolly, 1983), generated by probes of different sizes, as the probe's size gets larger so it becomes unable to slip into different-sized clefts in the surface, and the change in accessible protein surface from one probe to the next can indicate where the clefts are.

The method employed in the current study uses the SURFNET program (Laskowski, 1995), which is one of the sphere-filling methods mentioned above. An example of the cleft regions it generates is shown in Figure 1 for the enzyme thermolysin (Holden et al., 1987). The details are described in the Materials and methods section below. We used the program to compute the cleft volumes for three different categories of protein interactions: enzyme–ligand binding, protein–protein homodimers, and antibody–antigen interactions. The protein structures were all taken from the Brookhaven Protein Data Bank, PDB (Bernstein et al., 1977), and are identified by their four-character PDB codes throughout.

The bulk of the study concentrated on enzyme–ligand interactions. A representative dataset of enzyme structures was used which, for simplicity, only contained single-chain enzymes. We looked at how often the active site coincides with the largest of the clefts. We then looked at the role of clefts in other types of protein interactions, namely in protein–protein dimers and antibody–antigen complexes. And finally, we compared our results with those of Young et al. (1994) who demonstrated a close correspondence between protein binding sites and large hydrophobic patches on the protein surface.

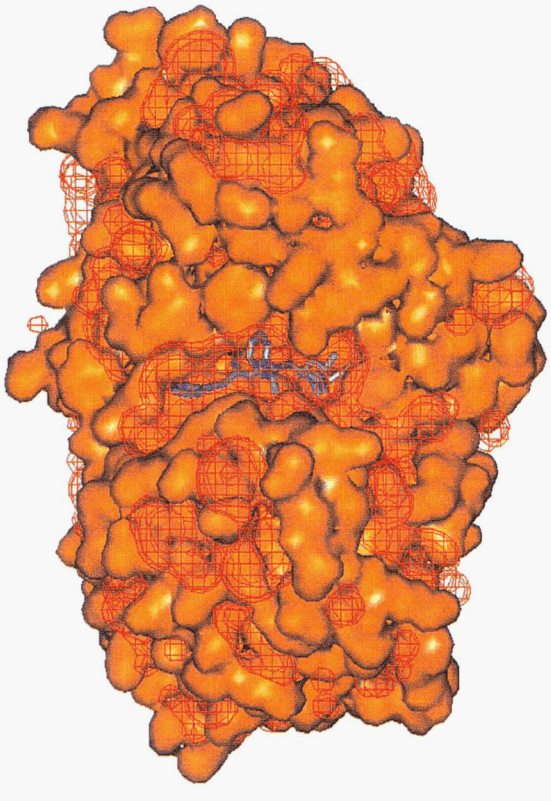

Fig. 1. The surface and clefts computed by SURFNET for the metalloproteinase thermolysin (PDB code *4tmn*, Holden et al., 1987). The contours defining the protein's clefts are shown as the red wire-cage regions lying between the protein's surface ridges (orange). The inhibitor molecule (in purple) can be seen sitting inside one of the cleft regions in the middle of the picture.

Results

Enzyme classes

After computing the clefts and their volumes for all 67 proteins in the enzymes dataset, the enzymes were classified according to which clefts ligands in the corresponding crystal structure were found to occupy. The classification was as follows: Class 1—ligand found in the largest cleft (cleft 1); Class 2—ligand in second largest cleft (cleft 2); Class 3—ligand in neither cleft 1 nor 2.

Where a protein had more than one ligand, the classification was based on the largest ligand and any others closely associated with it (i.e., involved in non-bonded contacts with it). The classes are shown in the final column of Table 1. In six cases (shown asterisked in the table) the class, initially assigned as Class 3, was changed to Class 1, as it was clear from the literature that the active site was located in the largest cleft even though a ligand was bound to a smaller cleft. For the most part, the ligands involved in these six cases were sulphate groups.

Table 2 gives a breakdown of the numbers of enzymes in each of the three classes, the majority (83.6%) falling into Class 1 where the active site corresponds to the largest of the protein's clefts (cleft 1). A further 9.0% fall into Class 2, with the active site in the second largest cleft, while the remaining 7.5% cases have it in neither the largest nor second largest clefts.

Class 1 and 2 enzymes

Given that most of the enzyme ligands are found in the largest cleft, how large is this cleft relative to the protein's other clefts? Figure 2 shows a highly schematic diagram of a selection of Class 1 enzymes, showing their clefts and ligands. In this diagram each protein is represented by a gray circle from which segments have been cut out to represent the protein's clefts. The area of each segment is proportional to the volume of the corresponding cleft, and the area of the remaining gray region is proportional to the volume of the whole protein. The clefts are placed clockwise around the circle, in decreasing order of size, from the 3 o'clock position until no more can be fitted within the circle's circumference. This means that some clefts may be lost in the schematic representation, but they will only be the smallest ones. All of the proteins in Figure 2 are drawn to the same scale. The ligands, also to the same scale, are shown as shaded diamonds with black regions, indicating the proportion of ligand atoms that are outside the cleft region.

The largest cleft varies markedly across different proteins ranging from 638 $Å^3$ in *1pek* (proteinase K) to 20,840 $Å^3$ in *1gpb* (T-state glycogen phosphorylase *b*), a factor of 32 difference. The case of the latter, *1gpb*, is exceptional in that the largest cleft is effectively a series of deep interconnected grooves in the exterior of the protein spanning most of its surface. These channels have been observed before with respect to their functional significance (Barford et al., 1988). The active form of glycogen phosphorylase *b* is a dimer, and its activity is controlled both allosterically (through metabolite binding to these channels) and through covalent modification (Johnson, 1992). The channels may have a further role in

Table 1. *Classification of the 67 proteins in the single-chain enzymes dataset*

E.C. number	PDB code	Protein	No. of domains	Folds(s)[a]	Ligand(s)	No. of atoms in ligand(s)[b]	Class[c]
Oxidoreductases							
1.1.1.21	1ads	Aldose reductase	1	TIM-barrel	NADPH	48	1
1.1.1.42	9icd	Isocitrate dehydrogenase	1	α & β	NADP+	27	1
1.1.1.44	1pgd	6-Phosphogluconate dehydrogenase	2	α & β, all-α	SO_4	5	2
1.1.1.85	1ipd	3-Isopropylmalate dehydrogenase	1	α & β	2 × SO_4	10	1*
1.1.3.15	1gox	Glycolate oxidase	1	TIM-barrel	Flavin mononucleotide prosthetic group	31	1
1.6.4.5	1tde	Thioredoxin reductase	2	2 × 3-layer $\beta\beta\alpha$	FAD	53	1
1.6.6.1	1cnd	Nitrate reductase	2	β-barrel, α/β	FAD	53	1
1.6.99.1	1oyb	Old yellow enzyme	1	TIM-barrel	Flavin mononucleotide prosthetic group	31	1
					P-hydroxybenzaldehyde	9	
1.11.1.1	2npx	NADH peroxidase	3	2 × 3-layer $\beta\beta\alpha$, α & β	FAD	53	1
					NADH	44	
1.11.1.5	1cca	Cytochrome *c* peroxidase	2	all-α, all-α	Heme	43	1
1.11.1.7	1arp	Peroxidase	2	all-α, all-α	Heme	43	1
					2 × NAG	28†	
1.14.13.2	1pbe	P-hydroxybenzoate hydroxylase	2	2 × 3-layer $\alpha\beta\alpha$	FAD	53	1
					P-hydroxybenzoic acid	10	
Transferases							
2.1.1.73	1hmy	HHAL DNA methyltransferase	2	α/β, all-β	S-adenosylmethionine	27	1
2.3.1.28	3cla	Chloramphenicol acetyltransferase	1	α & β	Chloramphenicol	20	1
2.4.1.1	1gpb	Glycogen phosphorylase *b*	1	α & β	Pyridoxal 5′-phosphate	15	1
2.4.2.1	1ula	Purine nucleoside phosphorylase	1	α & β	2 × SO_4	10	1
2.4.2.10	1sto	Orotate phosphoribosyltransferase	1	α/β	Orotidine 5′-monophosphate	24	1
2.7.1.1	2yhx	Yeast hexokinase B	3	α & β, α & β, α & β	Ortho-toluoylglucosamine	21	1
2.7.2.3	1php	3-Phosphoglycerate kinase	2	α/β, α/β	Adenosine diphosphate	27	2
2.7.4.8	1gky	Guanylate kinase	2	α/β, α & β	Guanosine 5 monophosphate	24	1
					SO_4	5	
Hydrolases							
3.1.1.0	2cut	Cutinase	1	α/β	Diethyl para-nitrophenyl phosphate	8	1
3.1.1.3	1thg	Lipase triacylglycerol hydrolase	1	α/β	2 × NAG + 2 × NAG	56	1*
3.1.1.7	1ack	Acetylcholinesterase	1	α/β	Edrophonium	12	1
3.1.3.2	1rpa	Prostatic acid phosphatase	1	α/β	Alpha-D-mannose + 2 × NAG	39	1
					NAG	14	
					Tartaric acid	10	
3.1.26.4	1rnh	Ribonuclease H	1	α & β	SO_4	5	3
3.1.27.0	1onc	P-30 protein	1	α & β	SO_4	5	2
3.1.27.3	1fut	Ribonuclease F1	1	α & β	Guanosine-2′-monophosphate	24	2
3.1.27.5	1rob	Ribonuclease A	1	α & β	Cytidine 2′-monophosphate	21	1
3.1.31.1	1snc	Staphylococcal nuclease	1	α & β	3′, 5′-Deoxythymidine bisphosphate	25	1
3.2.1.1	1cdg	Cyclodextrin glycosyltransferase	4	T IM-barrel, 3 × all-β	3 × Maltose	69	1
3.2.1.2	1byb	Beta-amylase	1	TIM-barrel	Maltotetraose	45	1
					SO_4	5†	
3.2.1.3	3bcl	Bacteriochlorophyll-A protein	1	all-β	7 × Bacteriochlorophyll A	462	1
3.2.1.8	1xnb	Xylanase	1	all-β	SO_4	5	1*
3.2.1.18	2sim	Sialidase	1	β-propeller	2,3-Dehydro-2-deoxy-*N*-acetyl neuraminic acid	20	2
3.2.1.73	1byh	Glucanohydrolase H	1	all-β	Epoxide inhibitor	27	1
3.2.2.22	1fmp	Ricin	1	α/β	Formycin-5′-monophosphate	23	1
3.4.11.1	1bll	Leucine aminopeptidase	2	α & β, α & β	Amastatin	33	1
3.4.17.1	2ctc	Carboxypeptidase A	2	all-β, all-β	L-phenyl lactate	12	1
3.4.21.1	2gmt	Gamma chymotrypsin	2	all-β, all-β	*N*-acetyl-L-alanyl-αlphal-phenylalanyl-chloroethyl	21	1

(*continued*)

Table 1. *Continued*

E.C. number	PDB code	Protein	No. of domains	Folds(s)[a]	Ligand(s)	No. of atoms in ligand(s)[b]	Class[c]
Hydrolases (*continued*)							
3.4.21.4	1ppc	Trypsin	2	all-β, all-β	Arginine-based inhibitors	38	1
3.4.21.12	2alp	Alpha-lytic protease	2	all-β, all-β	SO_4	5	1
					SO_4	5†	
3.4.21.36	1ela	Elastase	2	all-β, all-β	Trifluoroacetyl-L-lysyl-L-prolyl-P-isopropylanilide	32	1
					Acetate ion	4	
					SO_4	5†	
3.4.21.37	1hne	Human neutrophil elastase	2	all-β, all-β	Methoxysuccinyl-Ala-Ala-Pro-Ala chloromethyl ketone	31	3
3.4.21.64	1pek	Proteinase K	2	all-β, all-β	Substrate analogue	40	1
3.4.21.80	3sga	Proteinase A	2	all-β, all-β	Tetrapeptide inhibitor	33	1
3.4.22.2	1pip	Papain	2	all-β, all-β	Succinyl-Gln-Val-Val-Ala-Ala-P-nitroanilide	50	1
3.4.23.15	1smr	Renin	2	all-β, all-β	Ch-66 inhibitor	90	1
3.4.23.20	1ppl	Penicillopepsin	2	all-β, all-β	Phosphonate inhibitor	42	1
					SO_4	5†	
3.4.23.21	3apr	Rhizopuspepsin	2	all-β, all-β	Reduced peptide inhibitor	57	1
3.4.23.22	1epm	Endothiapepsin	2	all-β, all-β	PS2 inhibitor	74	1
					SO_4	5	
					$2 \times SO_4$	10†	
3.4.23.23	1mpp	Renin	2	all-β, all-β	SO_4	5	1*
3.4.24.27	1hyt	Thermolysin	2	α/β, all-α	L-benzylsuccinate	15	1
					Dimethyl sulfoxide	4†	
3.4.24.46	1iag	Adamalysin II	1	α/β	SO_4	5	1*
3.5.4.4	1add	Adenosine deaminase	1	TIM-barrel	1-Deaza-adenosine	19	3
3.8.1.5	2dhc	Haloalkane dehalogenase	1	α/β	Ethylene dichloride	4	3
Lyases							
4.1.1.48	1pii	Anthranilate isomerase	2	TIM-barrel, TIM-barrel	PO_4	5	1
					PO_4	5†	
4.1.1.64	2dkb	Decarboxylase	2	α & β, α & β	*N*-ethylsulfonic acid morpholine	12	1
4.1.3.7	1csh	Citrate synthase	1	all-α	Amidocarboxymethyldethia coenzyme A	51	1
					Oxaloacetate	9†	
4.1.3.27	2por	Porin	1	all-β	*N*-octyltetraoxyethylene	21	1
					$3 \times$ *N*-octyltetraoxyethylene	63†	
4.2.1.1	1cil	Carbonic anhydrase II	1	all-β	ETS inhibitor	19	2
4.2.1.3	8acn	Aconitase	3	α/β, α/β, β-barrel	Nitroisocitrate + Fe4-S4 cluster	21	3
4.2.1.11	5enl	Enolase	2	α/β, TIM-barrel	2-Phospho-D-glyceric acid	11	1
4.2.99.18	1abk	Endonuclease III	2	all-α, all-α	Fe4-S4 cluster	8	1*
4.3.1.8	1pda	Porphobilinogen deaminase	3	$3 \times$ α & β	Dipyromethane cofactor	30	1
					Acetate ion	4	
Isomerases							
5.1.2.2	1mns	Mandelate racemase	2	α & β, TIM-barrel	R-α phenyl glycidate	12	1
5.4.2.1	3pgm	Phosphoglycerate mutase	1	α & β	3-Phosphoglycerate	11	1
					$2 \times SO_4$	10	
Ligases							
6.3.4.15	1bib	Bira bifunctional protein	3	α & β, α & β, all-β	Biotin	16	1

Abbreviations:- NAG = *N*-acetyl-D-glucosamine; FAD = Flavin-adenine dinucleotide.

[a] In the fold classifications, α/β refers specifically to doubly wound α/β proteins, and α & β is a general category for proteins with a mixed α and β composition.

[b] Additional ligands, not bound in the primary binding site, are indicated by a †.

[c] Enzymes assigned to Classes 1, 2, and 3 according to whether the largest ligand is in the largest cleft, the second largest cleft, or some other cleft, respectively. The asterisked entries indicate enzymes initially classed as Class 3 on this basis, but subsequently reclassified as Class 1 on account of the largest cleft corresponding to the true binding site as indicated in the literature.

Table 2. *Numbers of enzymes in each class*

Class	No. of proteins (percentage of total)
Class 1: Active site in largest cleft	56 (83.6%)
Class 2: Active site in second largest cleft	6 (9.0%)
Class 3: Active site in neither cleft 1 nor 2	5 (7.5%)
Proteins in dataset	67

enabling a large protein molecule, with a ratio of non-polar to polar residues similar to that of smaller molecules, to adopt a compact structure that allows access to solvent polar groups (L.N. Johnson, pers. comm.).

As might be expected, there is a trend for the larger clefts to be associated with the larger proteins (Fig. 2). This is confirmed by Figure 3, which shows that for all 67 proteins in the enzyme dataset the volume of cleft 1 tends to be correlated with protein size, although there is also considerable variation. What is even more striking, however, is the difference between the Class 1 pro-

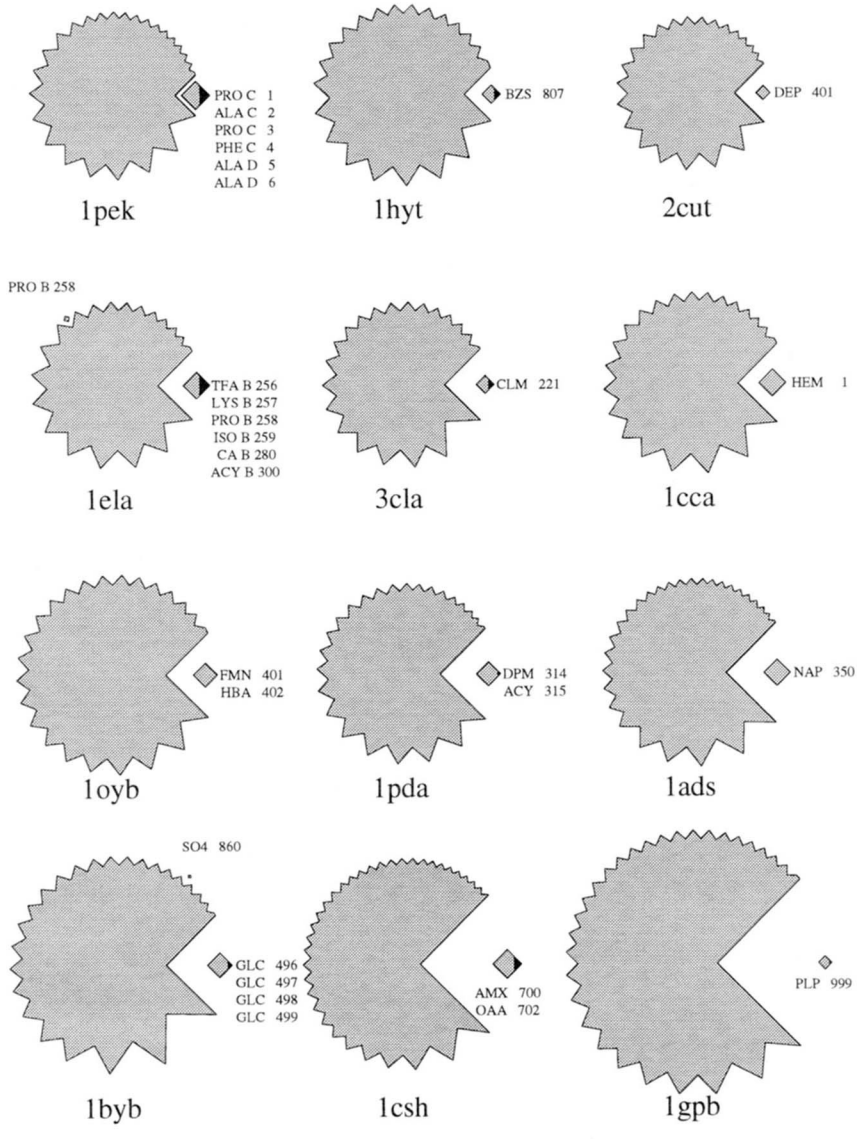

Fig. 2. Schematic diagram of 12 representative Class 1 enzymes (i.e., where the ligand is bound either fully or partially inside the protein's largest cleft). In this diagram, all areas are directly proportional to volumes and all proteins have been drawn to the same scale. Each plot is generated from a single gray circle with a number of segments cut out of it, each representing a cleft. The area of each segment is proportional to the volume of the cleft. The area of the original circle corresponds to the total volume of the protein plus its clefts. Thus, with the segments cut out, the remaining gray area is proportional to the protein volume. The clefts are cut out of the circle in order of decreasing size in a clockwise direction, starting at the 3 o'clock position. The ligands, shown within the largest cleft, are represented by diamonds. Again, the area of the ligand is proportional to its volume. Where part of the diamond is shaded black, the proportion of black corresponds to the proportion of the ligand atoms that are actually outside the given cleft. The 12 proteins shown here are plotted in increasing order of cleft 1 volume, from the smallest (*1pek*) to the largest (*1gpb*).

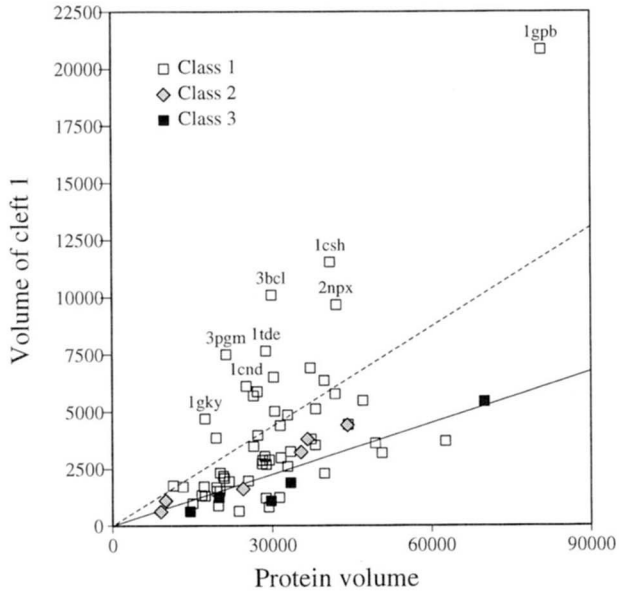

Fig. 3. Cleft 1 volume as a function of protein volume for the enzymes dataset. The three different classes are plotted with different markers, showing that Class 1 proteins (unshaded squares) tend to have larger ratios of cleft 1 volume to protein volume. The PDB codes of the most extreme cases are shown above the data points. The dashed line represents a best-fit line through the Class 1 enzymes, while the solid line represents a best-fit line through the enzymes in Classes 2 and 3. The volume units are Å^3.

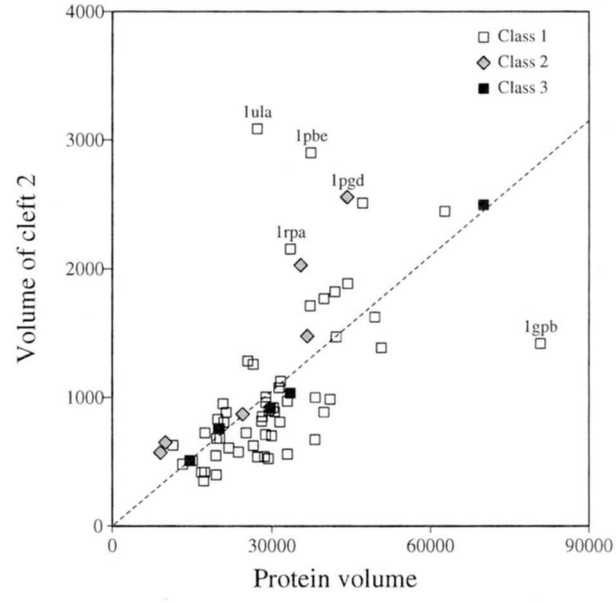

Fig. 4. Cleft 2 volume as a function of protein volume for the enzymes dataset. The three different classes are plotted with different markers. The dashed line gives a best-fit line to all the data points. Some of the outliers are labeled with the corresponding PDB codes. The volume units are Å^3.

teins and those in the other two classes. In many of the Class 1 proteins the ratio of cleft 1 volume to total protein volume is large, the most extreme cases being labeled with their PDB codes. This tendency for the Class 1 proteins to have larger ratios is best illustrated by the dashed line on the plot, which gives the best-fit line for the Class 1 enzymes. This is much higher than the solid line obtained by fitting to the data-points from enzymes in the other two classes.

Figure 4 compares the size of the second largest cleft, cleft 2, with the total protein volume. There is a stronger linear trend here, with fewer outliers and less difference across the three classes. However, all the Class 2 enzymes (shaded diamonds) do tend to lie slightly above the dashed line, which represents a best-fit to all the data points.

The ratio of cleft 1 to cleft 2 volume for a given protein provides an indication of how unusually large cleft 1 is, irrespective of the protein's size, and Figure 5 shows the distribution of this ratio for all three classes. It is clear that the Class 1 proteins have higher ratios of cleft 1 to cleft 2 volume than the other classes, going up as high as 14.7 in the case of *1gpb* (T-state glycogen phosphorylase *b*), which has already been discussed above. For Classes 2 and 3 the ratios are all below 2.6, whereas 36 of the 56 Class 1 proteins (64%) have a ratio above this value.

Class 3 enzymes

In Class 3 proteins, the ligand is in neither the largest nor second largest clefts. Indeed, in some of these cases the ligands are associated with fairly small clefts, well down in the volume ranking. The ratios of the cleft 1 to cleft 2 volumes tend to be close to 1.0, indicating that here again there is no obviously dominant cleft. Figure 6 shows a schematic plot of the five examples in this class.

It can be seen that most of the ligands are actually binding in relatively small clefts. Below, we consider each of these proteins in turn and find that two of them can be rationalized, while three appear to be genuine exceptions.

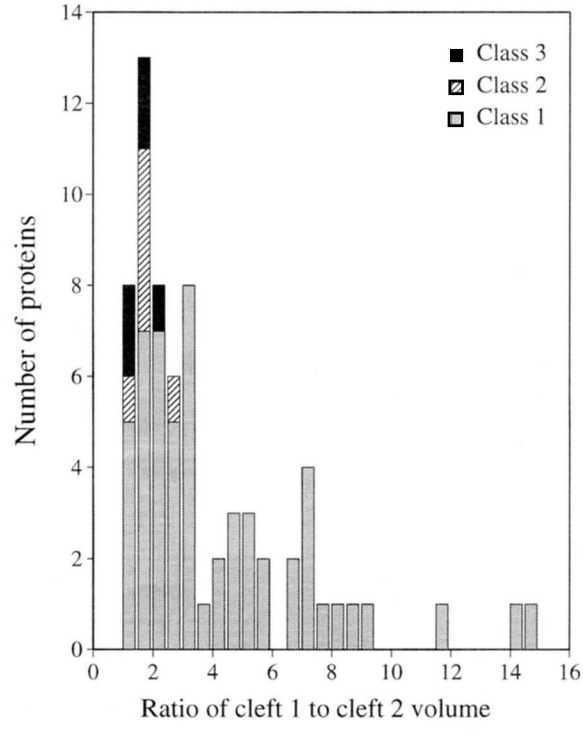

Fig. 5. Histogram of the ratios of largest to second largest cleft volumes in the enzymes dataset. The proteins in Class 1 tend to have larger cleft 1 to cleft 2 volume ratios than the proteins in Classes 2 and 3.

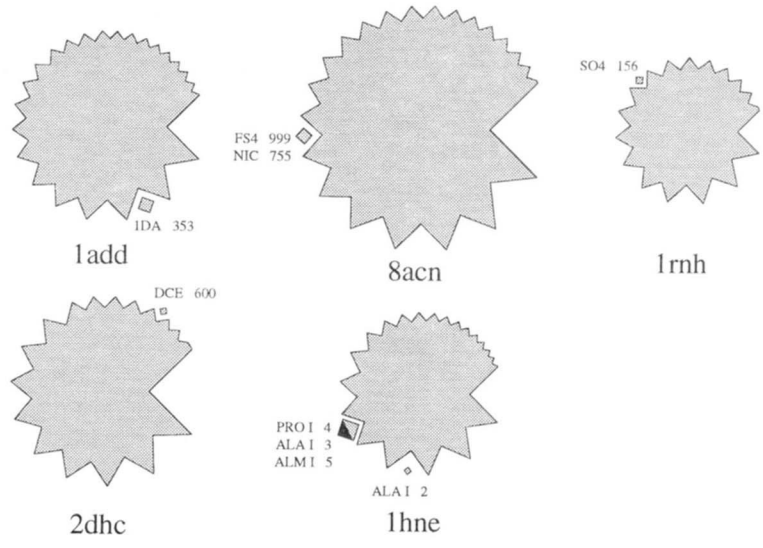

Fig. 6. Schematic diagram of the five Class 3 proteins in the enzymes dataset. In each case the ligand is bound in one of the smaller clefts. The plots are described in the legend to Figure 2.

In the first case, *1add*, the ligand fits snugly into the bottom of a deep cleft. The enzyme is adenosine deaminase (Wilson & Quiocho, 1993) the ligand is 1-deaza-adenosine (DAA), and the occupied cleft is the third largest on the protein's surface. In fact, in this case, the actual binding site is very wide—too wide to be detected by SURFNET, as it is wider than the maximum 4.0 Å-radius spheres used for packing into the protein voids. Increasing the maximum limit for the sphere size does eventually locate this wide cleft. Indeed, the cleft then becomes by far the largest, in effect promoting this enzyme from Class 3 to Class 1. However, this is a special case, and the reasons for not using larger sphere sizes throughout will be discussed later.

The next Class 3 enzyme is an aconitase (*8acn*) containing both an iron–sulphur cluster, [4 Fe–4 S] in which the eight atoms are at the corners of a cube, and a nitroisocitrate bound to it. The cluster is part of the catalytic mechanism that catalyzes the stereospecific dehydration-rehydration of citrate to isocitrate (Lauble et al., 1992). The cluster is held in place by three cysteines from the protein with the remaining bond coming from the nitroisocitrate inhibitor. Both the iron cluster and the inhibitor are completely buried inside the protein. Examination of the structure suggests that the protein must have closed around the ligand after binding, hence reducing the apparent size of the cleft.

In *1rnh*, a ribonuclease H (Yang et al., 1990), the crystal structure contains a sulphate group bound in the 10th largest cleft. As mentioned above, it is quite common to find sulphates not binding in enzyme active sites. However, in this example (in contrast to the asterisked cases in Table 1) the active site does not correspond to the largest cleft. Indeed, the active site is rather flat. The function of ribonuclease H is to degrade the RNA of RNA–DNA hybrids. Its substrate is thus a very large molecule and, as we shall see later, interactions with large molecules tend to be via more planar surfaces than through deep surface clefts.

The final two Class 3 cases appear to have genuinely small active sites. The first of these is *2dhc*, a haloalkane dehalogenase (Verschueren et al., 1993), which converts 1-haloalkanes into primary alcohols and a halide ion. The ligand is a small molecule, ethylene dichloride, which is bound in the protein's small active site (corresponding to the 15th largest cleft). The second enzyme is *1hne*, human neutrophil elastase (Navia et al., 1989). Its ligand is a five-residue peptide, methoxysuccinyl-Ala-Ala-Pro-Ala chloromethyl ketone. The peptide appears to lie on the surface of the protein in a very shallow depression, with residues 2 and 5 (Ala 2 and the Ala chloromethyl) dipping into smaller clefts, which hydrogen bond it to the protein. The remaining residues make hydrophobic contacts with the protein's surface. The binding site is fairly small, being only the sixth largest of the protein's clefts. It is interesting to compare this structure with that of a homologous elastase structure in our dataset, namely porcine pancreatic elastase (*1ela*), which is a Class 1 enzyme with a large, elongated groove at the binding site. Although the binding sites of the two proteins are very similar, neutrophil elastase has two sidechains (Leu 99 and Ile 151), which block the groove on either side of the active site, making it much shorter than the elongated groove of pancreatic elastase.

Originally, we had included two further proteins in our dataset of enzymes, selected from the PDB on the basis of the E.C. number. Both were classified as Class 3 proteins. However, both turned out not to be enzymes at all, but rather the recognition domains of larger enzymes, with no enzymatic activity of their own. The two proteins were: *1sha*, a SH2 domain of tyrosine kinase, and *2pk4*, a kringle 4 domain of human plasminogen. In both cases the ligand lies on the protein's surface, dipping into small clefts, rather than being bound inside a single large cleft. The SH2 domain is responsible for the recognition and binding of phosphorylated tyrosines and the *1sha* structure is a complex with phosphopeptide A—a hexapeptide that contains a phosphorylated tyrosine (Waksman et al., 1992). The recognition site for the tyrosine is a relatively small cleft, being the fifth largest on the protein's surface, but contains all the residues necessary for phosphate binding. The ligand molecule actually arches over the protein's surface, burying either end in a separate cleft.

In the second case, the kringle domain is complexed with aminocaproic acid (Wu et al., 1991). Human plasminogen is responsible for the removal of fibrin deposits from the walls of blood vessels and it comprises a serine proteinase domain plus five kringle do-

mains. The kringle domains are responsible for recognizing and binding fibrin. In the crystal structure the ligand molecule is small (nine non-hydrogen atoms) and fits snugly in a fairly small cleft on the protein's surface (cleft 5) corresponding to the lysine-binding subsite of the fibrin-binding kringle. It is held in place by hydrogen bonds at either end and by hydrophobic interactions along its length. Thus, unlike enzyme active sites, neither the SH2 domain nor the kringle domain requires a large cleft.

Protein dimers

The second dataset for which cleft volumes were computed was a dataset of protein homodimers (Table 3). Here, one chain was taken as the primary chain and the other as the ligand. The interactions of the latter with the clefts of the former were examined. The analysis was complicated by the fact that 22 of the 31 structures had one or more hetero groups complexed with the dimer in the crystal structure. These ranged in size from metal ions to a 20 base-pair fragment of DNA. Although the decision to consider these groups as part of the primary chain or not had an impact on the resultant cleft volumes, it had only a minimal influence on the overall findings, as will be seen shortly.

As expected, the pattern of binding was quite different from that of the enzyme–ligand complexes. Homodimers usually have two-fold symmetry and the gross surface of the interface between the dimer's two molecules tends to be fairly flat with an approximately circular region of contact (Jones & Thornton, 1995a). In fact, the interaction region tends to be the flattest part of the surface of each molecule (Jones & Thornton, 1995b). We found that, on average, around four clefts are involved in the interaction. Figure 7 shows schematic diagrams of three examples. Note that because the clefts in Figure 7 are shown in order of decreasing size, the occupied clefts may appear to be on opposite sides of the protein, whereas in fact, they are more likely to be in close proximity.

The first two examples in Figure 7, *1pp2* and *2tsc*, are cases where the largest cleft is involved. In fact, the largest cleft tends to be involved in most dimer interfaces, and this is often to accommodate a large side chain, or group of side chains, from the other molecule. Table 4 shows the numbers of homodimers in Classes 1, 2 and 3, with the Class 1 cases in the majority. The table presents two separate cases: in Case 1 the results were obtained by ignoring all hetero groups in the crystal structures during cleft volume calculation, while in case 2 the hetero atoms were included as part of the chain. The results are similar in both cases. In case 1, 24 of the 31 proteins (77.4%) are in Class 1, with the largest cleft involved at the dimer interface. In 16 of these 24 dimers (52%) this largest cleft contains more of its partner chain than any other cleft. The corresponding figures for case 2 (final column) are: 21 dimers in Class 1 (68%), with 13 of these (42%) having cleft 1 more occupied than any other.

The third example in Figure 7 addresses the A and B chains of the lectin domain from mannose binding protein A, *1msb* (Weis et al., 1991). Only two clefts are involved, both of which are fairly small. Figure 8 shows a flat interface between the two chains. Just two side chains from each chain protrude into clefts in the other in this homodimeric complex.

In other words, although there is a tendency for the interface to be approximately flat overall, the interaction usually involves the largest of the surface clefts. The flatness might be a by-product of the required symmetry of the homodimers to achieve the best fit between two molecules; namely, that any large hollows in one chain need to be complemented by large protrusions in the other, and vice versa. This might be to ensure the correct dimerization of the two subunits as well as providing a larger area of contact between them.

Antibody–antigen complexes

Our third dataset represented antibody–antigen complexes. Antibody structures (light and heavy chains) were aligned on four conserved cysteine residues by a least-squares fitting procedure to give all the fragments a common orientation (A. Martin and R. McCallum, pers. comm.), as shown in Figure 9. As can be seen from the figure, the upper part of the structure contains the complementarity determining regions, which is where the antigen binds, whereas the lower half is where the remainder of the antibody

Table 3. *The dataset of 31 non-homologous protein dimers*

PDB code	Protein	PDB code	Protein
1cdt	Cardiotoxin	2rve	ECO RV endonuclease
1fc1	FC fragment (immunoglobulin)	2sod	Superoxide dismutase
1msb	Mannose binding protein	2ssi	Subtilisin inhibitor (*streptomyces*)
1phh	P-hydroxybenzoate hydrolase	2ts1	Tyrosyl transferase RNA synthase
1pp2	Phospholipase	2tsc	Thymidylate synthase
1pyp	Inorganic pyrophosphatase	2wrp	Trp repressor
1sdh	Hemoglobin (clam)	3aat	Aspartate aminotransferase
1utg	Uteroglobin	3enl	Enolase
1vsg	Variant surface glycoprotein	3gap	Catabolite gene activator protein
1ypi	Triose phosphate isomerase	3grs	Glutathione reductase
2ccy	Cytochrome C3	3icd	Isocitrate dehydrogenase
2cts	Citrate synthase C	3sdp	Iron superoxidase
2gn5	Gene 5 DNA binding protein	4mdh	Cytoplasmic malate dehydrogenase
2or1	434 Repressor	5adh	Alcohol dehydrogenase
2rhe	Bence-Jones protein	5hvp	HIV protease
2rus	Rubisco		

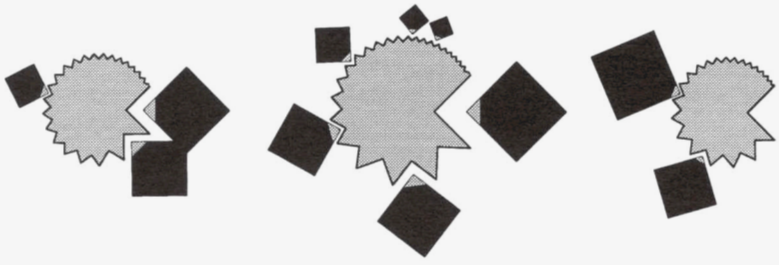

Fig. 7. Schematic diagrams of three of the proteins in the protein dimers dataset (Table 3). The diagrams are similar to those in Figure 2, except that here the first chain of the protein is represented by the gray circle, with the cut out segments corresponding to its clefts, while the protein's other chain is taken to be the ligand. This second chain is represented by the diamonds, whose gray areas correspond to the volume lying in each cleft and whose black areas correspond to the volume not in any cleft.

would normally be. As we were not interested in the lower part of the structure a common cutoff line was defined for all the structures, as shown in Figure 9. The cleft regions were generated as before for each fragment, but only those above the cutoff were considered.

In this dataset, the ligands varied greatly in size, from small haptens of 11 atoms, to whole proteins. Consequently, the results obtained followed both of the patterns described above. For the small and medium-sized antigens in Table 5, nearly all (i.e., 15 of the 18 cases, or 83%) followed the trend shown by the enzyme–ligand complexes, with the antigen binding in the largest cleft. Two of the three exceptions, *2mcp* and *1ind*, had the antigen bound in cleft 2, while in the third case, *1cbv*, the antigen is a fragment of DNA, which appears not to bind in any clefts.

On the other hand, where the antigen is very large, the pattern of binding is very similar to that in the protein dimers. That is, several clefts are involved, often including the largest one (in four of the eight examples).

Relative importance of cleft size

So far we have shown that interactions between proteins and other molecules often involve a cleft in the protein's surface that is much larger than others. It would seem, therefore, that having a large cleft is important for binding, particularly at an enzyme's catalytic site. But how important is cleft size when compared with other factors?

The importance of hydrophobicity has recently been demonstrated by Young et al. (1994). Using a dataset of enzymes, antibody fragments, and other proteins, they computed the sizes of clusters of hydrophobic residues on the surface of each protein and ranked them according to size. They found that the location of the co-crystallized ligand tended to correspond to one of the strongest of the hydrophobic clusters.

To compare the relative importance of hydrophobicity with cleft size, we used the Young et al. (1994) dataset, with the minor

Table 4. *Numbers of dimers in each class*

	Case 1	Case 2		
Class	Hetero atoms ignored	Hetero atoms taken as part of chain	No hetero atoms at interface	Total
Class 1	24	11	10	21
Class 2	4	3	1	4
Class 3	3	4	2	6
Totals	31	18	13	31

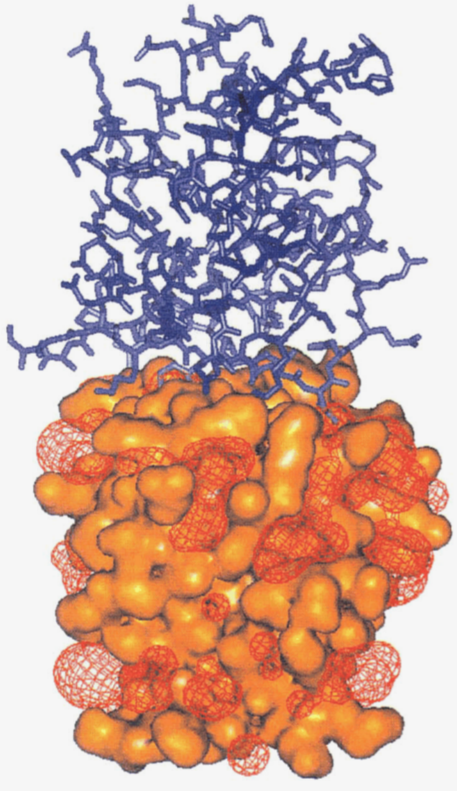

Fig. 8. The interface between chains A and B of the lectin domain of mannose binding protein A, PDB code *1msb* (Weis et al., 1991). The orange surface at the bottom represents the van der Waals surface of chain A, with its clefts delineated by the red wire-cage contour surfaces. The blue stick representation in the top half of the picture shows the covalent bonds of chain B. The interface between chains A and B is largely flat, with just the two sidechains (Phe 112 and Asn 115) from chain B protruding into clefts in chain A (and, similarly, the corresponding sidechains in chain A protruding into clefts in chain B).

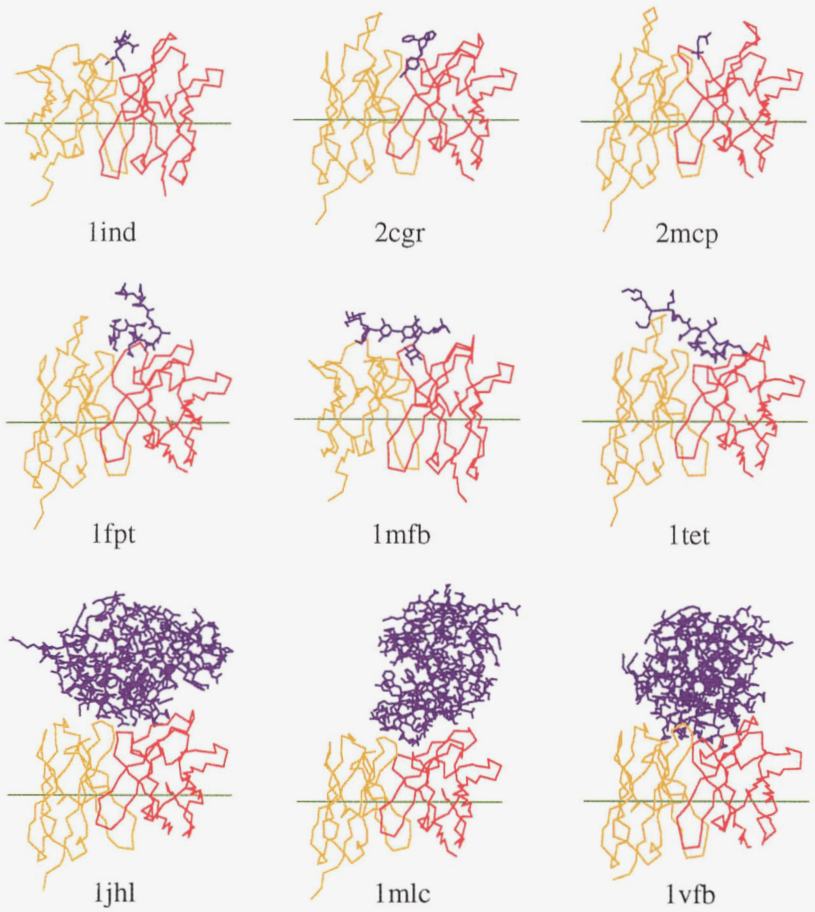

Fig. 9. C-alpha traces for nine of the structures in the antibody–antigen dataset (Table 5). The orange and red lines represent the C^{α} traces of the antibody's light and heavy chains, respectively. The purple lines at the top of each structure represent the bonds of the antigen molecule. In the top three structures the antigens are haptens, in the middle three the outer two are peptides, while *1mfb* is a heptasaccharide, and in the bottom three the antigens are proteins. The horizontal green line in each structure marks the cutoff above which the gap regions were taken into account.

exceptions noted in the Materials and methods section below, to compute the clefts for each protein. Table 6 compares the results of the hydrophobic cluster ranking of Young et al. (1994) with the ranking of cleft volumes. For the enzyme complexes (Table 6a), the majority of the binding sites (17 out of 20 = 85%) have a rank 1 cleft volume. That is, the ligand is found in the largest cleft. The hydrophobic cluster rankings tend to be lower, with only eight examples (40%) having rank 1. The three rightmost columns show how the two properties, hydrophobicity and cleft volume, compare with one another.

For the protein complexes (Table 6b) the picture is less clear cut. Both properties show a poorer correspondence with the binding site, but this is to be expected given that the region of interaction in these cases is spread over a larger surface area.

Effects of ligand binding

The strong correspondence reported here between the largest cleft and the binding site in enzymes is almost too good. Could it be an artefact of ligand binding? Might the presence of the ligand distend the binding site and make it appear unusually large? After all, there

Table 5. *The 26 entries in the antibody–antigen dataset*[a]

Small antigens—*Haptens*			
1baf	AN02	1igj	26-10
1dbb	DB3	1ind	CHA255
1eap	17E8	2cgr	Igg2b(κ)
1fig	1F7	2mcp	McPC603
1ibg	40-50	4fab	4-4-20
Medium antigens—*Peptides/carbohydrates/DNA*			
1acy	59.1	1him	17/9
1cbv	BV01-01	1mfb	SE155-4
1fpt	C3	1tet	TE33
1ggi	50.1	2igf	B13I2
Large antigens—*Proteins/cyclic peptides*			
1ikf	Igg1 (κ)	1ncd	NC41
1jel	JE142	1vfb	D1.3
1jhl	D11.15	2hfl	HyHEL-5
1mlc	D44.1	3hfm	HyHEL-10

[a]Each antibody is denoted by its four-letter PDB code and name.

Table 6. *Comparison of cleft volume ranking with the hydrophobic cluster ranking of Young et al. (1994)*

PDB code	Protein	No. of residues in ligand	Hydrophobic cluster rank[a]	Cleft volume rank	Comparison[b]		
a. Enzyme complexes							
1tpa	Anhydro-trypsin	58	3	1	+		
2ptc	β-Trypsin	58	3	1	+		
1tgs	Trypsinogen	56	2	1	+		
2tgp	Trypsinogen	58	3	1	+		
4tpi	Trypsinogen	58	1	1		=	
4sgb	Serine protease B	51	1	1		=	
2kai	Kallikrein A	56	2	1	+		
1hne[c]	Human neutrophil elastase	5	2	4			−
2est	Elastase	4	1	1		=	
1cho	α-Chymotrypsin	53	2	1	+		
2sec	Subtilase Carlsberg	64	4	1	+		
2sni	Subtilisin novo	64	4	1	+		
1tec	Thermitase	63	3	1	+		
2er9	Endothelial aspartic protease	6	3	1	+		
5apr	Acid proteinase	6	3	1	+		
6tmn	Thermolysin	3	1	2			−
1tlp	Thermolysin	3	1	2			−
2cpk	CAMP-dependent protein kinase	20	1	1		=	
4hvp	HIV-1 protease	6	1	1		=	
4cpa	Carboxypeptidase A	37	1	1		=	
	Average rank:		2.10	1.25			
	Counts:				11	6	3
b. Protein complexes							
1fdl	IgG1 Fab fragment	129	6	4	+		
2hfl[d]	IgG1 Fab fragment	129	4	1	+		
3hfm[d]	IgG1 Fab fragment	129	1	6			−
1him[d]	IgG2a fragment	8	2	3			−
2igf[d]	IgG1 Fab′ fragment	19	3	1	+		
1ncb	Fab′ complex	389	2	3			−
1fc2	IgG FC fragment	43	1	6			−
3hhr	Portion of growth hormone	194	4	1	+		
1rbp	Retinol binding protein	1	1	3			−
	Average rank:		2.67	3.11			
	Counts:				4	—	5

[a]Taken from Table 2 of Young et al. (1994).
[b]The comparison column shows where the cleft volume rank is higher (+), lower (−), or the same as (=) the hydrophobic cluster rank of Young et al. (1994).
[c]Protein also in the enzymes dataset of Table 1.
[d]Protein also in the antibody–antigen dataset of Table 5.

are a number of well-known examples where the protein undergoes substantial conformational change on ligand binding, the best-known example being hexokinase (Anderson et al., 1979), in which the two domains rotate and close over a glucose molecule when it enters the binding site. However, a recent study has shown that such cases tend to be in the minority. An examination of 195 pairs of protein structures, one native and one complexed with a ligand, showed that in fewer than 10% of proteins did the residues within 5.0 angstroms of a bound ligand exhibit a *rms* movement of more than 2.0 Å, while the majority of the remaining 90% showed an *rms* movement of less than 1.0 Å (E. Croft and R.E. Hubbard, pers. comm.). Furthermore, the major motions tend to result in the protein closing around the ligand, and so would be expected to reduce the size of the ligand-binding cleft, rather than increase it. The closed form of yeast hexokinase is in the enzyme dataset of Table 1 (*2yhx*), and despite the domains being in their closed conformation, the resultant cleft that remains for the binding of ATP is still the largest cleft on the surface.

To explore the effects of ligand size on cleft volume we took a protein that has several structures in the PDB, each complexed with a different ligand. The protein we looked at was carboxypeptidase A, for which there were 10 different structures in the January 1995 release of the PDB, two having no ligands and eight with different ligands bound, ranging in size from a small molecule of 12 atoms, to a 38-residue peptide. Figure 10 shows the cleft 1 regions and bound ligands of each of these 10 structures. The shape of the binding site in all these cases appears to have a fairly constant central region. In some of the structures this has one or

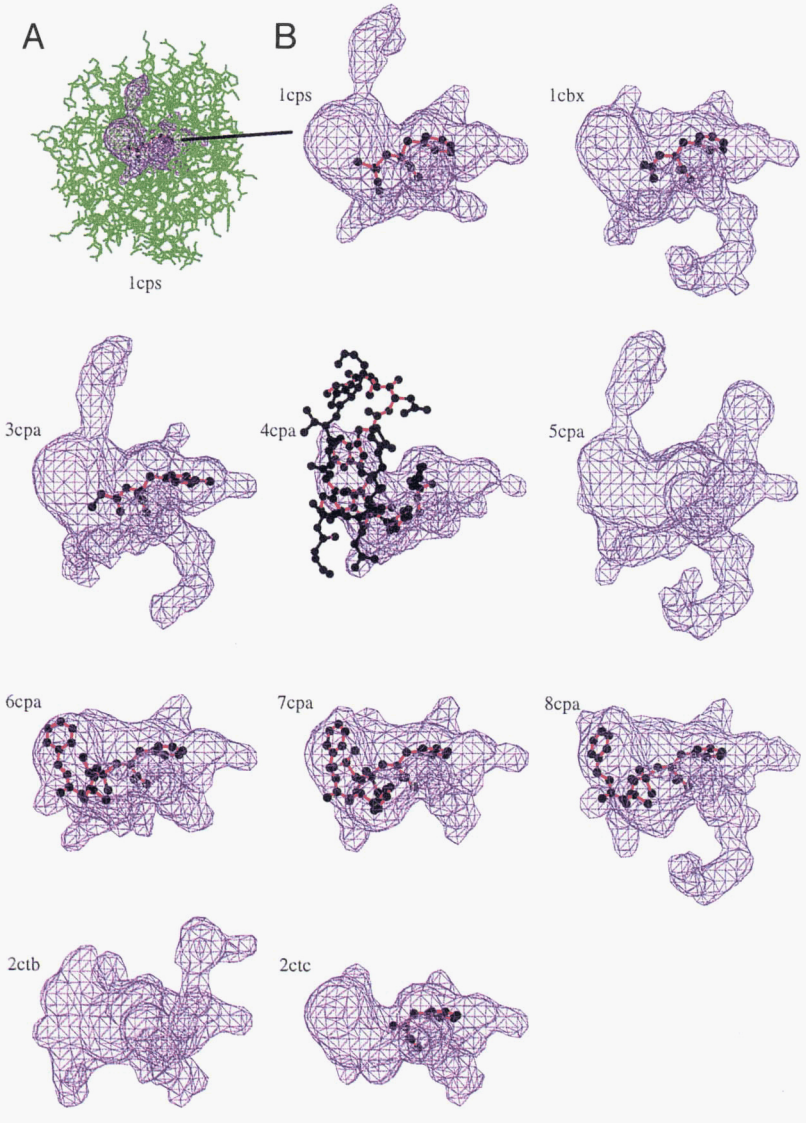

Fig. 10. The binding clefts of 10 different structures of carboxypeptidase A, each complexed with a different ligand. (**A**) A stick representation (in green) of one of the 10 protein structures, PDB code *1cps* (Cappalonga et al., 1992). The protein's binding cleft is outlined by the purple wire-cage contour surface with the ligand inside it shown by red bonds and black atoms. (**B**) The binding clefts and ligands of the 10 structures, each labeled with its PDB code. The first, *1cps*, corresponds to the structure in (**A**). Note that only part of the ligand in *4cpa* is shown; the ligand itself is a 37-residue peptide. There is no ligand in either *5cpa* or *2ctb*.

more arms that extend into an adjacent groove on the surface. The presence or absence of these connected grooves is governed by minor differences in the specific packing of the sidechains in the vicinity of the binding site. Nevertheless, by comparing just the common cores of the cleft regions in Figure 10 one can see that they have a similar size and shape, irrespective of the size of the ligand or whether they even contain one.

This is further illustrated by the cleft 1 volumes listed in Table 7, which show a wide range, from 812 $Å^3$ to 1,331 $Å^3$, depending on the presence or absence of extra grooves. Irrespective of the volume range, cleft 1 corresponds to the binding site in all 10 cases. In addition, the presence or absence of a bound ligand does not appear to influence the cleft's volume. In other words, ligand binding does not significantly distend the binding site to make it artificially coincide with the largest cleft.

Discussion

Our results quantitatively show that for most enzyme–ligand complexes there is a close correspondence between the largest cleft and the protein's binding site. What is more, this largest cleft tends to be significantly larger than the protein's other clefts, suggesting a functional significance. It would appear that these proteins have evolved to incorporate such large binding sites. There are several advantages in having such a large cleft. First, it provides a means of maximizing the number of complementary interactions (hydrogen bonds and hydrophobic contacts) that the protein can make with its substrate. Second, it enables the precise positioning of the substrate to facilitate catalysis. Finally, the burial of the substrate in a cleft shields it from bulk polarizable water molecules, which effectively decreases the dielectric constant and allows the enzyme

Table 7. *Comparison of cleft sizes for different ligand complexes of carboxypeptidase A*

PDB code	Protein volume (Å^3)	Cleft 1 volume (Å^3)	Cleft 2 volume (Å^3)	Ration volume 1:2	Ligand name	No. of residues	No. of atoms	No. of atoms in cleft 1
1cps	29,194	1,110	746	1.49	Sulfodiimine inhibitor	1	16	All
1cbx	29,292	1,069	683	1.57	L-benzylsuccinate inhibitor	1	15	All
3cpa	29,325	1,162	749	1.55	Glycyl-L-tyrosine	2	17	All
4cpa	29,173	872	577	1.51	Potato carboxypeptidase A inhibitor	38	291	30
5cpa	29,396	1,331	744	1.79	None	—	—	No ligand
6cpa	29,147	996	647	1.54	Phosphonate	1	33	All
7cpa	29,051	862	707	1.22	Phosphonate	1	41	37
8cpa	29,118	1,028	739	1.39	Phosphonate	1	32	31
2ctb	29,387	1,033	699	1.48	None	—	—	No ligand
2ctc	29,375	812	520	1.56	L-phenyl lactate	1	12	All

to generate the strong electrostatic forces necessary for catalysis (Fersht, 1985). In contrast, when the ligand binds in the second largest cleft, the volumes of the largest and second largest clefts tend to be very similar (ratio <2.0).

In a few enzymes ligands bind to neither the largest nor second largest clefts. Here the binding tends to be in some way unusual, as when covalent binding or recognition of a large molecule is involved.

These results show that the calculation of cleft volume alone provides a simple means of automatically locating binding sites, with the ratio of cleft 1 to cleft 2 volume indicating the likelihood of the largest cleft corresponding to the binding site. However, the problem of identifying critical residues within the cleft remains and, for this, one needs to use other complementary methods.

Although the current analysis considered only single-chain enzymes, we have briefly looked at the relationship between the domains in the structures and the location of the clefts. Nearly half the enzymes in our dataset (31 out of 67) consist of only a single domain, so here the protein's fold is solely responsible for the creation of a large cleft. It has been known for some time that in α/β proteins the binding sites are very easy to predict. For example, in α/β-barrel structures (TIM-barrels) the active site is formed by the eight loops connecting the carboxy ends of the β strands to the amino ends of the α-helices in the barrel (Brändén & Tooze, 1991). On the other hand, the active sites of doubly wound α/β-proteins (e.g., Rossmann folds) nearly always occur at the "topological switch point" (Brändén, 1980) where the protein's strand order is reversed. Of the single domain proteins that contain this fold (identified by α/β in Table 1), such as *1sto* orotate phosphoribosyltransferase (Scapin et al., 1994), we have found that the topological switch point is closely associated with the largest cleft. From this we may conclude that switch points represent an ideal way of producing large clefts for ligand binding.

In the multi-domain proteins the active site frequently occurs in the cleft generated by two interfacing domains. For example, in the structure of NADH peroxidase, *2npx* (Stehle et al., 1993), there are three domains, two of which are three-layer $\beta\beta\alpha$ sandwich domains that are both associated with the active site. Not surprisingly, the largest cleft is situated between these two domains, in the region that contains the active site. However, if we now split the protein into its three constituent domains and recalculate the cleft volumes separately, we find that each three-layer sandwich domain still contributes its largest cleft to the active site. In one domain the protein binds to the substrate (NADH) and in the other it binds the coenzyme (FAD). Thus, not only do the two domains contribute their largest clefts for ligand binding, but these individual clefts combine, using additional space from the domain interface to create a deep groove that dominates the multidomain structure.

Our results for the homodimers and antibody–antigen complexes confirm the observations made by Jones & Thornton (1995a, 1995b) that interactions between large molecules tend to occur across a roughly planar interface. We find that on average about four clefts on either surface are involved and that the largest cleft is usually one of them. The latter finding implies that cleft volume remains important even in these relatively planer interactions, both to ensure correct binding and to increase the contact surface area of the mutual interface.

Finally, we found that, for enzyme–ligand complexes the binding site showed a closer correspondence with cleft volume than it did with hydrophobic cluster size, as defined by Young et al. (1994), and is therefore more generally applicable to the detection of active sites.

Since this work was completed, a separate study, using an entirely different method of detecting and delineating binding clefts has appeared (Peters et al., 1996). The method uses an alpha-shape algorithm to define both a global and a detailed description of the shape of the protein, the difference between the two giving the clefts at the surface and holes in the interior. The method appears to be particularly sensitive to deeper rather than, as here, larger clefts, but overall achieves similar results to our own. A direct comparison, however, is impractical as their dataset was generated in a significantly different manner. The method, like ours, was unable to identify binding sites for covalently bound ligands such as iron–sulphur clusters. It also found no correlation between the size of the ligand and the size of the binding pocket. The principal conclusions of the study are in close agreement with our own; deepest clefts correlated closely with protein binding sites and interactions between larger molecules, such as protein–protein interactions, tended to involve flat areas of protein surface.

Materials and methods

The datasets

Three main datasets were used, representing enzyme–ligand complexes, protein–protein dimers, and antibody–antigen complexes. A fourth dataset, based on that of Young et al. (1994), was used to compare the association of binding sites with, on the one hand, large clefts on the protein's surface and, on the other, large hydro-

phobic patches, as reported by Young et al. (1994). All protein structures were obtained from the Protein Data Bank, PDB (Bernstein et al., 1977).

The enzyme–ligand dataset consisted of 67 representative enzyme structures. Only single-chain proteins were used, each having one or more ligands bound. Multimeric enzymes were excluded purely for the sake of simplicity, so as not to complicate the analysis. Proteins were selected from 1,575 enzymes in the January 1995 release of the PDB on the basis of their Enzyme Classification (E.C.) number (Bielka et al., 1992). Only a single structure was taken from each E.C. group and this was the complex (i.e., over five atoms where possible) with the best resolution and R-factor. Resolutions ranged from 1.50 to 2.80 Å over the entire dataset.

The resultant dataset is given in Table 1. Most structures have one or two ligands ranging in size from very small ligands of only five atoms, such as sulphate ions, to peptide chains of five or six residues in length. Although all the examples are single-chain enzymes, some comprise two or more domains. The numbers of domains given in Table 1 were identified using an automated domain assignment method (Swindells, 1995) allied with visual inspection on the graphics. The fold type of each domain was identified by visual inspection.

The intention was to obtain a dataset that was representative of enzymes in terms of their catalytic function, as defined by the E.C. number. One consequence was that some of the proteins in the dataset are structurally similar, most notably the trypsins (E.C. numbers 3.4.21.*) and the aspartic proteases (E.C. numbers 3.4.23.*).

The second dataset consisted of a non-homologous set of 31 protein dimers obtained from the dataset of Jones & Thornton (1995a). Each dimer contained two homologous subunits. Proteins were chosen so that no two had a sequence identity of more than 35% and a SSAP score of >80 (Jones & Thornton, 1995a). The latter is a measure of the structural similarity of two proteins (Orengo et al., 1993), and this restriction ensured that the dataset contained only structurally dissimilar proteins. The resultant dataset is shown in Table 3, where the resolutions of the structures range from 1.34 to 2.90 Å.

The third dataset consisted of a set of 26 unique antibody–antigen complexes. In each case, the portion corresponding to the antibody structure was a dimer of a light and heavy chain. The fragments were chosen from all the complete light/heavy chain dimers in the PDB such that, for each unique antibody, the structure with the best resolution and R-factor was retained (R. McCallum, pers. comm.). The 26 complexes are listed in Table 5, grouped according to the size of the bound antigen, which ranges in size from small haptens to full proteins. Resolutions lie between 1.8 and 3.1 Å.

An additional dataset was that taken from Young et al. (1994) in order to compare our results directly. The Young et al. (1994) dataset consists of 30 structures, comprising 20 enzyme structures taken from nine different enzyme classes, seven antibody fragments, hirudin, a growth hormone and a retinol-binding protein. All are X-ray structures with a bound ligand. Two amendments had to be made: *2hhr* was replaced by *3hhr*, which has superceded it in the PDB, and *3htc* was excluded from our dataset as it was a C^{α}-only structure.

Locating clefts in a protein structure

Protein clefts in the various datasets were located using the program SURFNET (Laskowski, 1995), which works as follows. The program considers all possible pairs of protein atoms in turn, but all ligands and water molecules are ignored so that only the voids in the protein structure itself are computed. A sphere is placed midway between the surfaces of each pair of protein atoms so that it just touches each one. Then, any clashes between the sphere and other nearby protein atoms are removed by reducing the size of the sphere accordingly. After checking all neighbouring atoms, the program retains the resultant sphere only if it is between a predefined minimum and maximum size; here, sphere radii used were 1 Å and 4 Å, respectively. The result of this procedure is a number of separate groups of interpenetrating spheres, both inside the protein and on its surface, which correspond to the protein's cavities and clefts. A surface is then generated around each of these clusters to give a solid representation, or "negative image," of the separate clefts and cavities in the protein (see Fig. 1).

The values of 1 Å and 4 Å for the minimum and maximum sphere radii are a somewhat arbitrary compromise between two extremes: too large a maximum sphere size or too small a minimum sphere size results in the shape of each cleft becoming very extended and filamentous in nature, as all the surface clefts join up via small channels and coalesce; on the other hand, too small a maximum sphere size results in many clefts not being picked up at all, as the spheres cannot span the gap between atoms at the cleft's edges. As the values are arbitrary, the resultant cleft volumes have no absolute meaning. However, throughout this work only the relative volumes of the different clefts are of interest, so it is only important that they be computed in a consistent manner.

Cleft volumes are also sensitive to the exact disposition of side chains defining a cleft. In fact, the displacement of any side chain, one way or another, might either result in two separate cleft regions coalescing, or one region being split into two. This becomes relevant where the volumes of the largest and second largest clefts are similar as an extra channel or two on either cleft can tilt the balance between which appears the larger of the two.

The protein surfaces and the surfaces defining the clefts (e.g., Fig. 1) were stored as 3D grids of density values (Laskowski, 1995), allowing them to be viewed interactively using standard molecular graphics packages such as QUANTA™ (Molecular Simulations Inc., Burlington, MA, USA). In this work, a grid-spacing of 1.0 Å for the grids was used throughout.

The volume of each cleft was calculated simply by counting the numbers of grid points within the cleft surface. Clefts were then ranked according to their volume, with the largest cleft being referred to as "cleft 1," the second largest as "cleft 2," and so on. The ligand atoms were then considered in turn to see in which cleft region, if any, they were located.

Acknowledgments

We would like to thank Juswinder Singh for suggesting this line of research, Susan Jones for the protein dimer dataset, and Robert McCallum and Andrew Martin for the dataset of antibody–antigen complexes. This work was partly funded by a vacation scholarship for N.M.L. from the Wellcome Trust. R.A.L. would like to thank Parke Davis Pharmaceutical Research for financial support.

References

Anderson CM, Zucker FH, Steitz TA. 1979. Space-filling models of kinase clefts and conformation changes. *Science 204*:375–380.

Barford D, Schwabe JWR, Oikonomakos NG, Acharya KR, Hajdu J, Papageorgiou AC, Martin JL, Knott JCA, Vasella A, Johnson LN. 1988. Channels at

the catalytic site of glycogen phosphorylase *b*: Binding and kinetic-studies with the β-glycosidase inhibitor D-gluconohydroximo-1,5-lactone *N*-phenylurethane. *Biochemistry 27*:6733–6741.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Bielka H, Dixon HBF, Karlson P, Liebécq C, Sharon N, Van Lenten EJ, Velick SF, Vliegenthart JFG, Webb EC. 1992. *Enzyme nomenclature 1992*. London, UK: Academic Press, Inc.

Brändén C-I. 1980. Relation between structure and function of α/β-proteins. *Q Rev Biophys 13*:317–338.

Branden C, Tooze J. 1991. *Introduction to protein structure*. New York: Garland.

Cappalonga AM, Alexander RS, Christianson DW. 1992. Structural comparison of sulfodiimine and sulfonamide inhibitors in their complexes with zinc enzymes. *J Biol Chem 267*:19192–19197.

Colloc'h N, Mornon J-P. 1988. Protein surface analysis: Qualitative approach using B-spline functions and quantitative comparison. *J Mol Graph 6*:220.

Colloc'h N, Mornon J-P. 1990. A new tool for the qualitative and quantitative analysis of protein surfaces using B-spline and density of surface neighbourhood. *J Mol Graph 8*:133–140.

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

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, Venkataraghavan R. 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.

Fersht A. 1985. *Enzyme structure and mechanism*. New York: Freeman.

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

Holden HM, Tronrud DE, Monzingo AF, Weaver LH, Matthews BW. 1987. Slow- and fast-binding inhibitors of thermolysin display different modes of binding: Crystallographic analysis of extended phosphonamidate transition-state analogues. *Biochemistry 26*:8542–8553.

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

Johnson LN. 1992. Glycogen-phosphorylase: Control by phosphorylation and allosteric effectors. *FASEB J 6*:2274–2282.

Jones S, Thornton JM. 1995a. Protein–protein interactions: A review of protein dimer structures. *Prog Biophys Mol Biol 63*:31–65.

Jones S, Thornton JM. 1995b. Principles of protein–protein interactions. *Proc Natl Acad Sci USA 93*:13–20.

Kleywegt GJ, Jones TA. 1994. Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Cryst 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.

Lauble H, Kennedy MC, Beinert H, Stout CD. 1992. Crystal structures of aconitase with isocitrate and nitroisocitrate bound. *Biochemistry 31*:2735–2748.

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.

Lewis RA. 1991. Clefts and binding sites in protein receptors. *Methods Enzymol 202*:126–156.

Navia MA, McKeever BM, Springer JP, Lin T-Y, Williams HR, Fluder EM, Dorn CP, Hoogsteen K. 1989. Structure of human neutrophil elastase in complex with a peptide chloromethyl ketone inhibitor at 1.84-Å resolution. *Proc Natl Acad Sci USA 86*:7–11.

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

Orengo CA, Flores TP, Taylor WR, Thornton JM. 1993. Identification and classification of protein fold families. *Protein Eng 6*:485–500.

Peters KP, Fauck J, Frömmel 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.

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

Scapin G, Grubmeyer C, Sacchettini JC. 1994. Crystal structure of orotate phosphoribosyltransferase. *Biochemistry 33*:1287–1294.

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

Stehle T, Claiborne A, Schulz GE. 1993. NADH binding-site and catalysis of NADH peroxidase. *Eur J Biochem 211*:221–226.

Swindells MB. 1995. A procedure for detecting structural domains in proteins. *Protein Sci 4*:103–112.

Verschueren KHG, Seljée F, Rozeboom HJ, Kalk KH, Dijkstra BW. 1993. Crystallographic analysis of the catalytic mechanism of haloalkane dehalogenase. *Nature 363*:693–698.

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.

Waksman G, Kominos D, Robertson SC, Pant N, Baltimore D, Birge RB, Cowburn D, Hanafusa H, Mayer BJ, Overduin M, Resh MD, Rios CB, Silverman L, Kuriyan J. 1992. Crystal structure of the phosphotyrosine recognition domain SH2 of v-*src* complexed with tyrosine-phosphorylated peptides. *Nature 358*:646–653.

Weis WI, Kahn R, Fourme R, Drickamer K, Hendrickson WA. 1991. Structure of the calcium-dependent lectin domain from a rat mannose-binding protein determined by MAD phasing. *Science 254*:1608–1615.

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

Wilson DK, Quiocho FA. 1993. A pre-transition-state mimic of an enzyme: X-ray structure of adenosine deaminase with bound 1-deazaadenosine and zinc-activated water. *Biochemistry 32*:1689–1694.

Wu T-P, Padmanabhan K, Tulinsky A, Mulichak AM. 1991. The refined structure of the ε-aminocaproic acid complex of human plasminogen kringle 4. *Biochemistry 30*:10589–10594.

Yang W, Hendrickson WA, Crouch RJ, Satow Y. 1990. Structure of ribonuclease H phased at 2Å resolution by MAD analysis of the selenomethionyl protein. *Science 249*:1398–1405.

Young L, Jernigan RL, Covell DG. 1994. A role for surface hydrophobicity in protein–protein recognition. *Protein Sci 3*:717–729.