J. Mol. Biol. (1996) **256**, 201–213

JMB

The Automatic Search for Ligand Binding Sites in Proteins of Known Three-dimensional Structure Using only Geometric Criteria

Klaus P. Peters, Jana Fauck and Cornelius Frömmel*

[1]*Humboldt-University of Berlin, Medical Faculty (Charité), Institute of Biochemistry, Hessische Straße 3-4, Berlin, D-10115 Germany*

**Corresponding author*

The biological function of a protein typically depends on the structure of specific binding sites. These sites are located at the surface of the protein molecule and are determined by geometrical arrangements and physico-chemical properties of tens of non-hydrogen atoms.

In this paper we describe a new algorithm called APROPOS, based purely on geometric criteria for identifying such binding sites using atomic co-ordinates. For the description of the protein shape we use an alpha-shape algorithm which generates a whole family of shapes with different levels of detail. Comparing shapes of different resolution we find cavities on the surface of the protein responsible for ligand binding.

The algorithm correctly locates more than 95% of all binding sites for ligands and prosthetic groups of molecular mass between about 100 and 2000 Da in a representative set of proteins. Only in very few proteins does the method find binding sites of single ions outside the active site of enzymes. With one exception, we observe that interfaces between subunits show different geometric features compared to binding sites of ligands. Our results clearly support the view that protein–protein interactions occur between flat areas of protein surface whereas specific interactions of smaller ligands take place in pockets in the surface.

Keywords: binding sites; protein structure; geometry; alpha-shape; protein surface

Introduction

Molecular recognition is one of the central questions in molecular biology. The ability of proteins to form specific stable complexes is fundamental to biological existence. Several aspects of protein–ligand interactions and the prediction thereof have been described (Kuntz *et al*., 1982; Goodsell & Olsen, 1990; Shoichet & Kuntz, 1991; Bacon & Moult, 1992; Mizutani *et al*., 1994; Norel *et al*., 1994). The interaction between ligand and protein takes place at the surface of the protein. This surface is very complex and convoluted. Furthermore, bound ligands vary greatly in size and properties. The smallest ligands such as O_2 and NO consist of two covalently linked atoms showing no or only partial atomic charges. Interactions between proteins and ligands of this type are defined by special arrangement of the electron systems of each participant. Likewise, small ions, e.g. calcium, sodium etc., form a complex compound or similar structure with few special charged atoms of the protein (McPhalen *et al*., 1991). A large number of known protein ligands are prosthetic groups, substrates and coenzymes. Their molecular masses lie between 100 and about 2000 Da and their binding sites are larger and more complex. The other end of the size scale is defined by the largest interacting partners of proteins: other biological polymers like nucleic acids, other proteins, and polysaccharides. They show molecular masses from 5000 up to 100,000 Da and more.

Generally, protein–ligand recognition is based on geometry as well as on properties of matching surfaces (charge, hydrophobicity; Goodsell & Olsen, 1990; Bacon & Moult, 1992; Jones & Thornton, 1995). There has recently been dramatic progress in molecular similarity research. Much of this interest is based on drug development. In order to design binding sites or ligands, the geometrical and

Abbreviations used: 3-D, three-dimensional; PDB, Protein Data Bank; ESA, enveloping surface area; DSA, detailed surface area; SSI, *Streptomyces* subtilisin inhibitor.

0022–2836/96/060201–13 $12.00/0

physico-chemical properties of the site of interaction and its complementarity to the ligand must be considered. The docking problem itself consists of three aspects: (1) finding binding sites on the protein surface; (2) evaluating molecular complementarity between ligand and binding site; and (3) searching for optimal binding modes of the ligand. If it were possible to define molecular complementarity, the two other questions would be classical optimisation problems.

A substantial problem in molecular docking is the representation of the surface of the participating molecules. There are several methods of characterising molecular surfaces focusing on fitting areas to each other, finding binding sites for special molecules or defining recognition sites for the immune system (Greer & Bush, 1978; Wodak & Janin, 1978; Kuntz *et al*., 1982; Conolly, 1986; Fanning *et al*., 1986; Novotny *et al*., 1986; Shoichet & Kuntz, 1991; Bacon & Moult, 1992; Kuhn *et al*., 1992; Norel *et al*., 1994; Yeates, 1995). Frequently, ‘‘critical’’ surface points are estimated. Such points can model local convex ‘‘knobs’’ and concave ‘‘holes’’ and can be described by sophisticated surface shape functions (Conolly, 1986) as well as by determination of the accessibility of protein surfaces using a test atom with a defined radius. This test atom is rolled over the van der Waals atomic surfaces. Using test atoms with larger radii, only protrusions of the protein molecule are contacted (Novotny *et al*., 1986). Smaller radii allow the test atom to come in contact with atoms in small grooves as well.

Other geometrically based approaches determine depressions and protrusions by comparing the general shape of a protein with local form (Wodak & Janin, 1978; Fanning *et al*., 1986). The global shape (‘‘sea level of the molecule’’) is described by spheres (Wodak & Janin, 1978) or more generally by ellipsoids (Fanning *et al*., 1986). In complex protein molecules, the use of such simple geometric shapes does not lead to an adequate description of the surface.

Another approach to finding holes is based on the calculation of surface-describing functions specifying topological properties. An example of this approach is the calculation of the fractal dimension around each atom (Kuhn *et al*., 1992). This method allows the determination of small pockets responsible for the binding of water molecules. Most methods based on this approach use Conolly’s description to specify the molecular surface. The resulting surface must be smoothed to avoid overly detailed protrusions and crevices. Recently another description of the protein surface was published (Yeates, 1995). It is based on the determination of the maximum contact radius for each atom and is sensitive to long-range accessibility of protein surfaces.

None of these methods have yet been applied to the determination of binding site pockets in proteins. Our objective in the development of APROPOS was to obtain an efficient and robust method for finding binding sites in proteins based solely on geometric criteria using atomic co-ordinates.

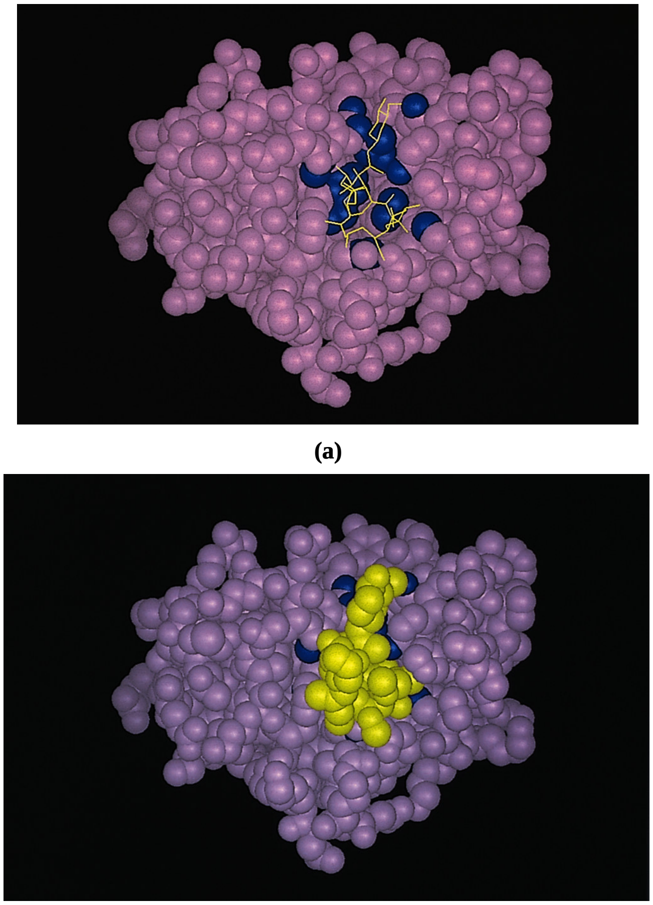
Here we describe a straightforward and powerful new method to identify pockets at protein surfaces that are likely to be responsible for interaction with other (small) molecules (substrates, coenzymes, prosthetic groups). The approach requires the three-dimensional structure of the protein as input and provides an experimentally testable prediction in the form of sets of atoms related directly to the binding site. On a large set of proteins with known 3-D structure and well described binding site(s), the reliability of the method is shown.

The starting point of our approach is the observation that binding sites of smaller ligands seem to be little caves (grooves, pockets, cavities, depressions) at the surface of proteins. Such pocket-like structures are described for several binding sites, e.g. for haem (first observed in myoglobin, Kendrew *et al*., 1961) and for substrates (e.g. lysozyme, proteases; Blake *et al*., 1965; Robertus *et al*., 1972). The deduction of binding pockets without bound ligands in crystal structures is currently achieved by inspecting 3-D models of proteins or by comparing structural data with the results of chemical modification experiments, spectroscopic analysis, site-directed mutagenesis, evolutionary considerations, etc. (see for example Cooperman *et al*., 1992).

Until now there has been no automated approach using exclusively computational geometry. The method given here is based on unbiased analyses of geometrical features of protein surfaces. For the generation of a molecular envelope we used the alpha-shape algorithm (Edelsbrunner & Mücke, 1994), which generates a one-parametric family of shapes. These envelopes preserve more or less molecular detail as the parameter becomes smaller or larger, respectively.

For our approach, we selected two envelopes of the molecule: one which allows the detailed description of binding sites of a protein molecule and a second preserving a representation of the global form of the molecule. We then analysed the differences between these two envelopes to determine sets of neighbouring atoms which build larger depressions at the surface of the molecule. As a result the method generates a list of atoms arranged in clusters which represent pockets at the molecule envelope. The program is called APROPOS standing for Automatic PROtein POcket Search.

The algorithm correctly identifies more than 95% of all binding sites for ligands and prosthetic groups with molecular masses from 100 up to 2000 Da in a representative set of more than 300 proteins. The method indicated an interface between subunits to be a binding site in only one multimeric protein among 82 and rarely identified binding sites for single ions outside the active sites of enzymes.



(a)

(b)

Figure 1. The FK506 binding protein (PDB code 1FKF, 107 amino acids) with ligand FK506 (Van Duyne *et al*., 1993). Comparison of the binding region for the inhibitor FK506 proposed by APROPOS (blue, distinct atoms of residue Tyr26; Phe36; Phe46; Val55; Ile56; Arg57; Trp59; Ala81; Tyr82; Phe99) and the region covered by the ligand. (a) The thin wire model of FK506 (yellow) and (b) space filling model of FK506.

Results

The FK506 binding protein

We first illustrate the results of the method by applying it to the FK506 binding protein, which has only one binding site for one substrate. The FK506 binding protein catalyses the isomerisation of the proline imide bond and is involved in signal transduction of immune stimulation (Fischer, 1994, Galat & Metcalfe, 1995). A three-dimensional structure is also available for a complex with the ligand FK506 (Van Duyne *et al*., 1993), permitting direct evaluation of the accuracy of prediction. For the FK506 binding protein (three-letter PDB code; Bernstein *et al*., 1977: 1FKF; Figure 1) a perfect match between the predicted region and the binding region derived from X-ray structure analysis of the 1:1 complex can be seen. Only this single binding site per molecule was found by APROPOS. The ligand consists of 57 heavy atoms. The predicted site contains 18 atoms. The method proposed residues Tyr26, Phe36, Phe46, Val55, Ile56, Arg57, Trp59, Ala81, Tyr82 and Phe99 as members of the binding site. Comparing contacts less than 4.0 Å^2 observed in X-ray analysis (Van Duyne *et al*., 1993) only five amino acids (Asp37, Arg42, Glu54, His87 and Ile91) were missed. The four polar residues (marked by underlining) are situated at the margin of the pocket. Since the algorithm lists only the deeper parts of the pocket they are not indicated. For another data set of the isomerase (PDB code (Bernstein *et al*., 1977): 1YAT, 113 amino acids; Rotonda *et al*., 1993) the results were almost identical. In the following all PDB codes refer to Bernstein *et al*. (1977).

Recognition of binding sites in proteinases, e.g. subtilisin family

As an example of more complex binding sites we compared results obtained for a group of proteinases. Polypeptide hydrolysing proteinases (endopeptidases) convert polymeric substrates of remarkable size. In these enzymes the binding sites are composed of several subsites. For proteinases, seven to nine subsites are generally described, each accommodating one amino acid residue of the peptide substrate (Schechter & Berger, 1967; Bode *et al*., 1987; Phillips & Fletterick, 1992). The subsites are located on both sides of the catalytic residues. As an example of proteinases we considered the prediction of binding sites in 23 different experimental structures of subtilisins (Table 1). As constituents of binding sites, APROPOS indicated from 13 to 22 heavy atoms in 8 to 14 different amino acid residues at the surface of subtilisin molecules. All of them are in close proximity. Generally, APROPOS correctly mapped in all enzyme structures the main constituents of binding sites S1, S2, S3, and S'1 but only sporadically the subsites S4, S5, S6, and S'2, S'3 situated distantly to the catalytic triad (Bode *et al*., 1987). The reason for this failure was that these subsites are quite flat and do not form a pocket as observed for S1, S2, S3 and S'1. The described difference of the number of included atoms and amino acids is not produced by large conformational changes induced by the ligands (see below) but seems to be caused by subtle structural variations reflecting different surrounding in the crystals or experimental errors. In more shallow regions such small changes of co-ordinates have stronger consequences on the result than in a deep pocket.

Upon binding, ligands influence the structure of proteins (induced fit; Koshland, 1958). Therefore, APROPOS may produce different results for proteins which are resolved as complexes compared with data sets derived from uncomplexed proteins. In the case of the subtilisins, we could compare X-ray structures of 13 enzymes with ligands and ten without. We found that the prediction by APROPOS is not influenced by the conformational changes

Table 1. The determination of amino acid residues as constituents of substrate binding sites by APROPO in the subtilisin protease family

Site:		S'1							S1 S3	S2							
	S'3	S1		S1			S1 S2 S3		S4	triad	Triad	S2	S4		S4 S6	S5 S6	
Amino	Asn	Asn	Ala	Gly	Thr	Gly	Ser	Leu	Gly	His	Ser	Leu	Gly	Ile / Val	Tyr / Trp	Gly	
acid:	mc sc	sc	mc sc		sc		mc	mc sc		sc	sc	sc		sc	sc		n_{atom}
1SBC		155			220	100	125	126	127	64	221	96				128	15
1CSE		155	152	154	220	100	125	126	127	64							16
1SBT		155	152	154	220	100	125	126	127	64		96					21
1SUB		155				100	125	126	127	64	221[a]	96	102	107 /		128	17
1SUC		155		154		100	125	126	127	64	221[a]						20
1S01		155	152	154		100	125	126	127	64	221	96	102	107 /		128	17
1S02		155	152			100	125	126	127	64	221	96		107 /		128	17
1SBN		155	152	154	220	100	125	126	127	64	221	96					18
1SIB		155	152	154	220	100	125	126	127	64	221	96		107 /			19
2SIC		155	152		220	100	125	126	127	64	221	96		107 /			19
3SIC		155	152	154	220	100	125	126	127	64	221	96		107 /			21
5SIC		155	152	154	220	100	125	126	127	64	221	96					18
1ST2		155	152			100	125	126	127	64	221	96	102	107 /	104 /	128	19
2ST1		155				100	125	126	127	64	221	96		107 /		128	15
2SNI		155[b]	152[b]			100	125[b]	126	127	64[b]				107 /			21
1MEE		155	152	154	220	100	125	126	127	64	221	96		107 /	104 /	128	22
1THM		163	160	162		108	133	134	135	71	225	104					19
1TEC	69	163[b]	160[b]	162[b]		108	133	134	135[b]	71	225	104			/ 112[b]		22
2TEC		163	160	162	224	108	133	134	135	71	225				/ 112		20
3TEC		163	160	162	224	108	133	134	135	71	225	104		/ 115	/ 112		20
2PRK	67	161[b]	158[b]	160[b]	223[b]	100	132	133	134	69	224	96					16
1PEK		161	158			100	132	133	134	69	224	96					13
3PEK	67	161	158			100	132	133	134	69	224	96					15

In the Table the number of the amino acid is given if at least one atom of it was indicated by APROPOS.

mc, sc, predicted atoms of the given amino acid are part of the main chain and side-chain, respectively. $n_{\mathbf{atom}}$, total number of atoms indicated as members of the binding site by APROPOS.

Different types of amino acid side-chains in a homologues position are seperated by ‘‘/’’.

Structures considered (three-letter code in first row of the Table): subtilisin Carlsberg: 1SBC; 1CSE (+ eglin); subtilisin BPN', 1SBT; 1SUB (mutant); 1SUC (mutant); 1S01 (mutant); 1S02 (mutant); 1SBN (+ eglin); 1SIB (+ eglin); 2SIC (+ SSI); 3SIC (+ SSI); 5SIC (+ SSI); 1ST2 (oxidised); 2ST1 (oxidised); 2SNI (+ chymotrypsin inhibitor 2); mesenteric endopeptidase; 1MEE (+ eglin); thermitase, 1THM; 1TEC (+ eglin); 2TEC (+ eglin); 3TEC (+ eglin); proteinase K, 2PRK; 1PEK (+ peptide substrate); 3PRK (+ peptide chloromethyl-ketone); SSI, *Streptomyces* subtilisin inhibitor (protein structures determined as complexes (n = 13) are marked by underlining).

The definition of the subsites (S1, S2. . . , S'1, S'2 . . .) and the active site residues (triad) is according to Schechter & Berger (1967), Robertus *et al*. (1972), Bode *et al*. (1987), and Phillips & Fletterick (1992).

[a] Instead of a serine residue this mutant of subtilisin BPN' contains cysteine at this position.

[b] Amino acids determined are described as a second cluster.

Table 2. Putative active site residues of inorganic pyrophosphatase from yeast (Cooperman *et al*., 1992) and predicted atoms from APROPOS on the basis of X-ray data (PDB code 1PYP; Arutiunian *et al*., 1981)

Active site residues	Atoms indicated by APROPOS	Residual activity after mutation (%)	Remarks
Lys56	Lys56 CE, NZ	2 (Lys → Arg)	(cons), confirmed by chemical modification studies
Glu58	Glu58 OE2	6 (Glu → Asp)	(cons)
	Gln70, CD, AE1, 2		
	Phe79, CB		(cons) exchange Phe, Val, Tyr
Tyr93	Tyr93, OH	7 (Tyr → Phe)	(cons)
Asp115	Asp115, OD2	6 (Asp → Glu)	(cons)
Asp117	Asp117, CB, CG	1 (Asn → Glu)	(cons)
Asp120	Asp120 OD1,2	0 (Asp → Glu)	(cons)
	Leu144, CD1		(cons) exchange Leu, Val only
Asp152	Asp152 OD1,2	0 (Asp → Val)	(cons)
	Phe189 CD1, CE1		(cons)
Tyr192	Tyr192 CZ, OH	22 (Tyr → Phe)	(cons)

Numbering of amino acids is according to Cooperman *et al*. (1992).
(cons), amino acids are conserved in different primary structures of inorganic pyrophosphatases.

caused by the attachment of a ligand during structure determination. We observed no significant differences when comparing the amino acids indicated for proteins alone with those indicated for complexes. In the case of complexes, we noticed a slightly greater (10%) number of atoms defined as being members of the binding region. Binding sites seem to become slightly more open as a result of ligand attachment.

Among the proteins considered in this paper, only one case was found (lipase:triacylglycerol acylhydrolase, PDB code: 3TGL/4TGL, 269 amino acids) in which ligand binding had a significant influence, leading to the prediction of an incorrect substrate binding site in the absence of ligand (see below).

Binding site for inorganic pyrophosphatase (1PYP)

The efficiency of the method could be tested with proteins for which no experimentally determined 3D structures of complexes are available. Inorganic pyrophosphatase (pyrophosphate phosphohydrolase, EC 3.6.1.1) catalyses the hydrolysis of inorganic pyrophosphate and binds several magnesium ions which are important for its activity (Cooperman *et al*., 1992). A preliminary 3-D X-ray crystallographic structure at 3 Å resolution has been published and deposited in the protein data bank (PDB code: 1PYP; Arutiunian *et al*., 1981). A divalent metal-ion binding cavity has been proposed, based on visual inspection of the structure, that contains several acid residues which appear to interact with bound metal ions. In the neighbourhood of acid residues there are two lysine side-chains and one arginine residue that could plausibly interact with PP_i (Cooperman *et al*., 1992).

APROPOS identified one pocket (Table 2) in which the binding of Mg^{2+}-pyrophosphate seems to be possible. Furthermore, most residues essential for catalysis are identified as constituents of this site. The method proposed seven acidic side-chains (Glu58, Tyr92, Asp115, Asp117, Asp120, Asp152 and Tyr192) presumably responsible for Mg^{2+} and one basic residue (Lys56) as a member of the pocket likely to be responsible for interaction with negatively charged PP_i. From chemical modification studies and sequence comparison of distantly related pyrophosphatases a very similar set of participating side-chains was derived (Cooperman *et al*., 1992). Note that most residues including hydrophobic side-chains defined by APROPOS as part of the active site are conserved throughout evolution. The conclusions are supported by site-directed mutagenesis experiments (column 3, Table 2). Some polar residues situated nearby were missed in the APROPOS prediction (Glu48, Asp147, Glu148, 150, Lys154, 193, Arg78, Tyr89) because most of them (six from eight, underlined) are located at the margin of the pocket and the remaining two residues were hidden by other atoms.

Overview of binding pockets in a large set of protein structures

To estimate the reliability of the method we considered a larger set of protein structures for which clear results were obtained for binding sites. The analysis of about 300 proteins (sets Ia, Ib, II, and III; see Materials and Methods, Database) showed an excellent agreement between experimental results and the outcome by APROPOS in a distinct range of protein and ligands sizes (Table 3).

We considered a prediction to be successful if at least seven atoms of the site were indicated by APROPOS as constituents. If the structure of the protein-ligand complex was known, atoms of the given protein within 4.0 Å distance between centres of atoms to the ligand were defined as binding site members. In other examples we used the active/binding site definition from the literature.

The coincidence of prediction and experimental results was impressive, not only for different structures of one protein (see e.g. subtilisin (Table 1), hemoglobins, etc.) but also for all protein families equally. The data sets included protein structures showing both higher (1.6 Å) and lower resolution (3.0 Å). The quality of the prediction of a binding site did not depend on the resolution of X-ray experiments.

The influence of the protein size on the results of APROPOS

Smaller proteins are likely to serve as ligands to larger molecules and do not as a rule show ligand binding sites themselves (Goodsell & Olsen, 1993). Furthermore, small proteins have a reduced capacity to build larger pockets due to the small number of residues. Therefore, the number of binding sites found by the method should depend on the molecular mass of the proteins. The smallest protein for which APROPOS correctly predicted the binding site was ferredoxin (1FDX, 54 amino acids long with covalently attached iron sulphur complex, FeS). We could not find any binding site in proteins (peptides) consisting of less than 50 amino acids. For small proteins with known ligands (mainly ferredoxins and cytochromes, e.g. 1FDX, 1FXD, 2FXB, 351C, 1CC5, 1ABA, 3B5C, 1FXI, 1FXA, 3FXC, 1CYC, 5CYT, 1CTH, 2CDV, 1CY3, 1UTG) up to about 100 amino acids in length, we obtained correct predictions in only about one-half of the cases (underlined PDB codes). One reason for these errors is the open and rugged structure of the apoprotein, after removal of the ligand, which does not form significant pockets.

We obtained nearly perfect results for all proteins with between 150 and 350 amino acids (1300 to 3500 non-hydrogen atoms). Slightly better results could be obtained in the case of enzymes. The smallest enzymes in our database had about 100 amino acids (e.g. ribonucleases 1RDS, 1RAT . . . 7RAT, 9RNT, 1FUD, 1BSR, 7RSA; lysozmes 135L, 1HEL, 1HHL, 1GHL; prolyl isomerases 1YAC, 1FKF, 1FKE). APROPOS correctly indicated the active site (binding site and catalytic site) for all of them also.

For proteins, which are themselves known as ligands of other proteins (set III), such as small proteases and amylase inhibitors (human pancreatic secretory trypsin inhibitor (1HPT), bovine pancreatic trypsin inhibitor (1BTI), pancreatic trypsin inhibitor (5PTI), alpha-amylase inhibitor 1HOE, etc.) and toxins (cardiotoxin 1CDT; erabutoxin 3EBX; verotoxin 1BOV) APROPOS did not predict any meaningful binding sites. The reason could be that these proteins have too few amino acids to allow pocket formation. Furthermore, these proteins do not require a ligand for function. The significance of this fact was illustrated when larger proteins of this type were considered. Even though the number of amino acids is more than 140 we found no binding site for low molecular mass ligands in basic fibroblast growth factor 4FGF (146 amino acids), Kunitz inhibitor 1TIE (172 amino acids) or leukocyte elastase inhibitor 1HLE (346 amino acids).

The influence of the ligand size on the results of APROPOS

Considering the outcome of APROPOS in relation to the size of the ligand we obtained an excellent prediction for substrates of enzymes larger than four non-hydrogen atoms. The binding site of the non-covalent specifically attached sulphate anion in the sulphate binding protein (1SBP) could also be determined by the algorithm. If metal ions (calcium, zinc, manganese, etc.) are constituents of the active site (e.g. thermolysin 3TLN, metallo-protease 1EZM, superoxide dismutase 1SOS, 2SOD, 2SDY, carbonic anhydrase 2CAB, carboxypeptidase 3CPA, 4CPA, 5CPA, alpha-amylase 6TAA) their binding location was predicted as a part of the active site. Outside the active site of enzymes APROPOS correctly indicated binding sites of metal ions only

Table 3. Discovering binding sites in known protein–ligand complexes (sets Ia, Ib, II, and III, totally 309 proteins, defenition see Materials and Methods) (several proteins were mentioned repeatedly because they have more than one ligand)

Set	Ligands	*p*	*ip*	*np*	Remarks
Ia	Coenzyme, substrates, prosthetic	222	1	3	Smallest substrates: H_2CO_3, H_2O_2,
Ib	groups (not included haem, FeS)				Mostly: molecular mass 600–1000 Da
II	Haem group	29	5		—
	FeS groups	4	1	4	FeS
	Ions (mostly Ca^{2+})	4	—	43	Only ions considered which are not located in active sites of enzymes
I, II	Unknown	12	—	—	Unknown function of the proposed additional sites
III	Unknown	3	—	—	Unknown function
I–III	Binding regions between subunits of multimeric proteins	1	—	**81**	Each subunit is counted only once

p, number of known binding sites found by APROPOS.

ip, incompletely predicted (only few atoms of significantly larger binding site were predicted or less than the total number of ligand binding sites were found by APROPOS).

np, number of proteins for which the binding site for ligands or interfaces to other subunits were not indicated by APROPOS.

Table 4. Comparison of the number of heavy atoms of some ligands (n_L) and the number of atoms indicated as constituents of a binding site by APROPOS (n_A) in some proteins

Protein	Code	Ligand	n_L	n_A
Transferrin	1TFD	Fe^{2+}	1	24
Superoxide dismutase	1SOS	$H_2O_2/Zn^{2+}/Cu^{2+}$	4	11
Acetylcholinesterase	2ACE	Acetylcholine	10	17
Xylose isomerase	3XIS	Xylose	10	32
Fatty acid binding protein	1IFB	Fatty acid	10–20	26
Inorganic pyrophosphatase	1PYP	$PP_i, 3 * Mg^{2+}$	12	37
Liver alcohol dehydrogenase	6ADH	NAD, ethanol	12 + 3	20
Isocitrate dehydrogenase	5ICD	Isocitrate	13	12
Cytochrome *c*553	1C53	Haem, Fe^{2+}	43	2 ∗ 7
Myoglobin	1MBA	Haem, Fe^{2+}	43	31
Flavocytochrome *B* 2	1FCB	Flavin mononucleotid + haem, Fe^{2+}	31 + 4	18 + 19
Ribonuclease F1	1FUS	RNA	∼50[a]	14
Subtilisin	2SNI	Chymotrypsin inhibitor	20–40[b]	19
FK506 binding protein	1YAT	FK506	57	18
Phospholipase A2	1BBC	Phospholipids	∼60	21

[a] Assuming two nucleotides are specifically bound.

[b] As maximum we assume nine binding subsites for amino acid residues of peptides (Schechter & Berger, 1967; Bode *et al*., 1987; Phillips *et al*., 1992). The lowest numbers are two or three amino acids.

in two proteins. The identification of the binding sites of transferrin (1TFD) and ferritin (1FHA) was straightforward due to the deep pocket responsible for the specific Fe^{2+} binding. In ion ‘‘storage’’ proteins such as parvalbumin (Ca^{2+}-binder: 1PAL, 1RTP, 4CPV, 5PAL) APROPOS found no binding sites of ligands. At the surface of these molecules APROPOS indicated only very small depressions (less than five atoms size), none of which were identical to the Ca^{2+} binding site. The method also failed quite often in cases where small ligands, e.g. FeS, are covalently bound to small proteins (see above). The binding site of the iron sulphur cluster was found only in one-half of the cases (see above; Table 3).

The relation between ligand size and number of atoms predicted by APROPOS

To obtain information about potential ligands for a given binding site, it is useful to seek for a relation between features of the estimated regions and properties of the ligands. As a first step we searched for a correlation between size of the ligand and number of atoms mediated by APROPOS. The lowest number of atoms indicated for one definitive binding site was seven observed in several proteins. In inorganic pyrophosphatase (1PYP; Arutiunian *et al*., 1981) we obtained the largest number of atoms (37) defined as members of an open pocket. There was no correlation whatsoever between the molecular mass of a ligand and the number of atoms indicated by APROPOS as part of the binding sites (Table 4). The number of atoms in the pocket depends not on the overall size of the ligand but rather on the size of its interacting portion which may be quite small (e.g. myoglobin 1MBA, hemoglobins 1MHB). The number of pocket atoms may exceed that of the ligand if the binding pocket is a deep cavity (e.g. acetylcholine esterase 2ACE, superoxide dismustase 1SOS, inorganic pyrophosphatase 1PYP).

Interfaces between subunits of multimeric proteins as binding sites

In our test set 82 subunits of multimeric proteins were included. In addition to dimeric proteins, there were several examples of proteins with more than two subunits. Therefore, the number of interfaces between protein subunits was more than 100. Only for one interface the shape was similar to a binding site for a small ligand (human transforming growth factor 1TGF; Schlunegger & Grütter, 1992). The 3-D structure of this dimeric protein was described by the authors as containing a ‘‘special new fold’’. The interaction between the two subunits is stabilised by a disulphide bridge. The structural features of the interaction include an alpha-helix from one subunit and a curved beta-sheet from the other. The beta-sheet partially covers the helical structure and was predicted by our method to be a binding site for a ligand. All other interfaces were not identified as binding sites for ligands.

In summary, the method worked very successfully on nearly all types of enzymes and provided correct results for proteins with more than 150 amino acids.

Additional binding sites with unknown function found by APROPOS

In about 5% of all proteins considered (a total of 15 out of about 309: 1AAN, 1ALD, 1CGI; 1CGJ; 1CHO, 1GPR, 1LLA, 1PGX, 1SIL, 1UBQ, 2SNS, 3BLM, 3TGL, 8RUB, 9RNT) we predicted additional binding regions with unknown function. No data regarding the role of these sites could be found in the literature. In all cases the shapes of the indicated regions were comparable to that of the other binding site. At present we have no basis to determine the extent to which these findings indicate real binding sites or the extent to which the method failed.

Taking into account a hint from one referee we checked whether the additional binding sites occur in regions involved in crystal contact. But in the 15 proteins mentioned above one observes no crystal contacts near the additional sites. Only in two cases (2SNS, 9RNT, both are nucleases) does a contact in the crystal occur near the active site.

Particularities in several proteins

In some proteins with more than 350 amino acids and clearly separated domains a deep but functionally unimportant cleft between two domains is formed. APROPOS consequently identified residues in a pocket between two domains for *N*-5'(phospho-ribosyl) anthranilate isomerase (1PII) and, e.g., for complete immunoglobulins. When the domains were examined separately, APROPOS correctly determined both binding sites, e.g. for anthranilate isomerase (1PII).

There are some proteins which exhibit binding sites in the form of holes inside the molecule (Banaszak *et al*., 1994). This class is exemplified by the fatty acid binding proteins (2HMB, 1IFB, 1IFC, and homologous proteins 1TTA, 1OPB), and biotin binding protein (streptavidin 1PTS), which deposit the ligand in the interior of the molecule. APROPOS always found the internal binding hole of the protein.

In enzymes which catalyse reactions between two or more substrates bound simultaneously to the active site, APROPOS predicted mostly one site consisting of quite a large number of atoms covering both binding sites. We obtained results suggesting two or more distinct binding sites in only a few cases (see e.g. Table 1 subtilisins).

Several proteins are known to inhibit enzymes by direct interaction with the active site. If these inhibitors attach to the enzymes like substrates (Bode & Huber, 1992) the binding pockets of enzymes are filled by the ‘‘knobs’’ of the inhibitors. The specificity of this inhibitor type is defined by the same interactions observed between enzyme subsites and substrates. The main regions of interaction for this type of inhibitor were correctly estimated by APROPOS.

Another type of proteinases inhibitor covers partially the binding pockets (mostly S1) and catalytic sites of proteinases. Such inhibitors show a high specificity for special enzymes based on protein–protein interactions outside the binding pocket (e.g. hirudin in the case of interaction with thrombin, Bode & Huber, 1992). Their specific binding is the result of a very large set of interactions between atoms of both proteins and is comparable to protein subunit interactions. The site of interaction between such inhibitors could not be mediated by APROPOS.

Another class of high molecular mass protein ligands consists of nucleic acids. DNA binding proteins are specially adapted to the form of double-stranded DNA (major and minor groove and walls between both). In both small and larger proteins one observes a groove which contacts the DNA backbone (Pabo & Sauer, 1992) and which was correctly determined by APROPOS (endonucleases: 1END, 1RVE, DNA-binding protein 2GN5, deoxyribonuclease I (DNase I) 3DNI, catabolite gene activator protein 3GAP).

Failure of APROPOS in three proteins

In the range between 150 and 350 amino acids, the method failed to correctly identify binding sites for three proteins in the chosen data set of about 275. These were a monomeric lectin (1LTE, 239 amino acids, binds sugar residues near the margin of the molecule in a small and flat cleft), the achromobacter protease (1ARB, 268 amino acids, very flat structure around the active site residues) and triacylglycerol acylhydrolase (3TGL, 269 amino acids, free enzyme). For lectins it is known that the monosaccharide–lectin interactions are relatively weak and show only modest specificity (Drickamer, 1995). We presume that a deep pocket is not necessary for such a binding mode.

The observation of a flat binding site in the achromobacter protease was unique for all 62 proteases considered (all known endo-protease families were included, Phillips & Fletterick, 1992): 1PPL, 1PPM, 3APP, 2ER7, 2APR, 3APR, 5APR, 4PEP, 1SMR, 1RNE, 1HNE, 2SGA, 3SGB, 1LPR, 2ALP, 1P12, 1SGT, 3RP2, 2TGA, 1TON, 5CHA, 1TRM, 2PTC, 4PTP, 1TGS, 2TGP, 1ACB, 1CGI, 1CGJ, 1CHO, 8GCH, 1GCT, 2ACT, 1PPO, 9PAP, 1LYB, 1NPC, 1EST, 1EZM (23 different structures of subtilisins, see Table 1). For all of them APROPOS worked properly and indicated substantial elements of the active site. To our knowledge there is no explanation for the special geometric feature of achrombacter protease.

In the ligand free lipase (3TGL) the catalytic site is covered by a helix (the ‘‘lid’’, Derewenda, 1994). The method indicated a pocket as a binding site that closes after activation of the enzyme. In contrast, in 4TGL (lipase with bound inhibitor) the binding pocket was correctly estimated by APROPOS. Therefore, the reason for failure in the case of lipase was not the method itself but the conformational change in the protein.

Discussion

We were surprised that our simple geometric approach was successful in most cases because we applied only a description of the local surface shape with a ‘‘low resolution’’. APROPOS used only the centres of atoms and did not explicitly take into consideration the radii of heavy atoms which range from 1.4 Å to 2.2 Å (united atoms) in proteins. Furthermore, the properties of the atoms participating in ligand binding were not reflected here.

Like any predictive method, there were margins of error. The method worked best when the protein considered contained more than 100 amino acids and the ligand was larger than four heavy atoms. In simple cases, the results of APROPOS agree with intuitive analyses of 3-D structures. However, the method may be superior to visual inspection and sequence considerations (Casari *et al*., 1995) in the identification of multiple and complex binding sites. Furthermore, it could be shown that it is a powerful objective algorithm independent of the experience of the researcher. The small number of errors (less than 2% of known binding sites were not found and binding sites not previously described were found in less than 5% of the proteins) allows the highly reliable prediction of binding sites. Our approach is notable for its simplicity and high speed.

Several shortcomings of APROPOS are directly related to the algorithm itself. The method usually could not determine all participating atoms of the binding site. Complete binding sites for low molecular mass and non-covalently bound ligands are pockets consisting of a base and a surrounding wall. As a rule, the two parts blend into one another. Due to its algorithm APROPOS indicated only the atoms forming the base of the pocket. Generally our method correctly indicated about one-half of the protein atoms which are in van der Waals contact with the ligand. APROPOS will be beneficial in computational docking.

Theoretically, docking of a ligand to a protein can be divided into three steps: searching the binding site, generating different orientation of the ligand in the binding site and evaluating the given binding mode. Searching consists of matching the ligand shape to that of the protein. There are several methods to dock molecules based on an extensive random search for binding sites (Greer & Bush, 1978; Kuntz *et al*., 1982; Conolly, 1986; Goodsell & Olsen, 1990; McPhalen *et al*., 1991; Bacon & Moult, 1992; Kuhn *et al*., 1992; Mizutani *et al*., 1994; Norel *et al*., 1994) or on experimental knowledge of the location of binding sites in proteins.

Knowledge of the location of the active site speeds up and improves docking prediction. It significantly reduces the degrees of freedom in computational docking. Until now there has been no automated procedure to find active sites in proteins (Bacon & Moult, 1992). In order to find binding sites of proteins and to support computationally docking one should have an algorithm that objectively deduces binding sites. The algorithm itself should be influenced as little as possible by investigator preconceptions and should yield a clear result when the ligand is not known. The method we used to achieve this is based on a discrete description of the geometry of the protein surface by means of an alpha-shape algorithm. This approach is related in some aspects to the Voronoi binding site models (Boulu *et al*., 1990; Bradley & Crippen, 1993). The high efficiency of the method is a result of the striking geometrical feature of binding sites for smaller ligands. Such sites in proteins are formed by conspicuous pockets at the surface of the protein molecule or by holes in the protein interior. This is in marked contrast to the structure of multimeric protein subunit interfaces, which are quite flat (Jones & Thornton, 1995). The holes (and knobs) described for such interfaces (Conolly, 1986) are obviously much smaller than pockets of specific ligand binding sites.

From the results presented here several implications can be drawn for molecular modelling. Firstly, the design of specific binding sites for low molecular mass substances requires the creation of a groove. Secondly, if the binding site of a given protein is a pocket, low molecular mass substances or other proteins may be preferred as inhibitor of the regular ligand binding. The interaction between a protein and its genuine ligand can, for example, be prevented by filling the binding pocket by another small ligand (inhibitor in enzymes), which results in a competitive binding mode. In this case there is a high probability that all proteins showing identical ligand (substrate) specificity would be influenced by this inhibitor. But the active site pocket can also be covered by another protein which could show a large number of interactions with the enzyme outside the active centre. This binding type results in a selective binding to a specific enzyme and could give a non-competitive binding mode. Both possibilities are observed in nature, e.g. in proteinase inhibition (Bode & Huber, 1992). Thirdly, it seems to us that the prevention of protein–protein interaction, e.g. between subunits of multimeric proteins, will only be possible using larger ligands, due to the lack of cavities situated in the interface large enough to bind low molecular mass ligands. Therefore, the design of small ligands to specifically and at low concentration prevent the association of subunits will be difficult to achieve.

For the method described here further developments will concentrate on implementing a similarity search between ligand and protein binding site. For this purpose the inclusion of atoms situated in the ‘‘wall’’ around the binding site is also necessary. In addition to geometric similarity, the atomic properties of protein and ligand will be considered as well. A further topic of research is concerned with oligomeric proteins and their binding sites consisting of more than one peptide chain. This work is in progress and will be reported elsewhere.

Materials and Methods

Database

In this paper only monomeric proteins or subunits of multimeric proteins were considered. We describe here binding sites of ligands which consist of single peptide chains. Therefore, in this analysis, e.g. antigen binding regions of antibodies, viral acid proteases or the binding sites of bisphosphoglycerate in haemoglobin were not included. The bisphosphoglycerate molecule is bound between the four monomers of the haemoglobin molecule (Perutz, 1970). We also excluded membrane proteins from our present consideration. These special cases will be analysed in detail in a subsequent paper.

We applied the method to 309 protein co-ordinate sets deposited in the Brookhaven Data Bank (PDB) version October 1993 (Bernstein *et al*., 1977). The primary selection criterion was the knowledge of the location of ligand binding sites and high confidence in the experimental data. Unless indicated otherwise, the highest resolution data were preferred for identical proteins with several sets of co-ordinates (sequence homology 100%) and for experimentally generated mutant proteins (sequence homology > 98%). To test the reproducibility of the method, different experimental resolutions of a protein were used in several cases. Only co-ordinate data sets derived from X-ray crystallographic experiments were included in this study.

Data from 189 proteins (set Ia) were available as protein–ligand complexes in the Protein Data Bank (arranged according to chain length starting with 54 amino acids and ending with 842, given as PDB code): 1FDX, 1FXD, 1RPE, 2OR1, 2FXB, 351C, 1CC5, 1ABA, 3B5C, 1FXI, 1FXA, 3FXC, 1CYC, 1SHA, 5CYT, 1RDS, 256B, 5FD1, 1CTH, 1FKF, 1YCC, 2CDV, 2FKE, 1RNB, 1CCR, 1YEA, 3C2C, 1NCO, 1YAT, 2HMQ, 1C2R, 1CY3, 1PTS, 1PPA, 1BP2, 5RAT, 2CCY, 4BP2, 1BBH, 1POC, 155C, 1OPB, 1ECA, 4FXN, 1ITH, 1HDS, 1THB, 2MHB, 4SDH, 1FDH, 1MBA, 2HBG, 1FX1, 2SNS, 2SOD, 1LH1, 1MBC, 1MBD, 1MBS, 1MYG, 1SDY, 1SOS, 4DFR, 3DFR, 1MUP, 5P21, 1FLV, 1OFV, 2FCR, 1BBP, 1RBP, 1FHA, 3SGB, 1DRF, 1GKY, 2HMB, 1DR1, 8DFR, 3SDP, 1P12, 7LPR, 3GAP, 3CLA, 1AKE, 1PPO, 4GST, 1HNE, 1PPF, 1PPG, 1TRM, 2PTC, 4PTP, 1AK3, 1TGS, 2TGP, 1LTE, 1EST, 1ACB, 1CGI, 1CGJ, 1CHO, 8GCH, 1GCT, 4HTC, 2TSC, 1ARC, 4TGL, 1DRI, 1CSE, 1SNI, 3SIC, 5SIC, 1SBN, 1MEE, 2SNI, 1PEK, 1TEC, 3TEC, 3PRK, 1RHD, 2CYP, 1TFD, 1ABE, 3CPA, 3GBP, 4CPA, 1BBR, 1GCA, 2GBP, 3AT1, 8ATC, 1SBP, 2CMD, 1FNR, 1ADS, 1LLD, 4PFK, 1PFK, 2PIA, 1TRB, 1PPL, 1PPM, 5APR, 3APR, 1LDM, 2ER7, 9LDT, 1GD1, 4MDH, 1SMR, 5FBP, 1RNE, 1LGA, 1LYB, 1ADD, 1APM, 1MNS, 1GOX, 1DMB, 1ATN, 2OHX, 5ADH, 6ADH, 1SIL, 3XIS, 1XIM, 4XIA, 1AAW, 7AAT, 1CP4, 2CPP, 3PGK, 5ICD, 1CSC, 2CTS, 1NPX, 1LVL, 2HPD, 8RUB, 3LAD, 3GRS, 2TPR, 1BTC, 8CAT, 1FCB, 1ACE, 1GAL, 1LLA, 8ACN, 1GPB. Monomers of multimeric proteins (62 proteins) are marked by underlining.

In set Ib, 51 proteins (7 monomers of oligomeric proteins) which are homologous to other proteins with known binding sites were collected: 9RNT, 1FUS, 1POA, 1PP2, 1ALC, 1BBC, 1BSR, 1POD, 4P2P, 9RAT, 1HEL, 1HHL, 1GHL, 2BP2, 1ALB, 1IFB, 3LZM, 2SGA, 3ADK, 1ABM, 2ALP, 9PAP, 2ACT, 1SGT, 3RP2, 2TGA, 1TON, 5CHA, 1RVE, 1ARB, 1RTC, 1SBC, 1ST2, 1S01, 1S02, 1SBT, 1SUB, 1SUC, 1THM, 2PRK, 1EZM, 5CPA, 4TMS, 1NPC, 1CMS, 3APP, 2APR, 4PEP, 6LDH, 1ALD, 6XIA.

In a further, smaller set (set II, 34 proteins, 13 monomers of oligomeric proteins) we took the definition of functional residues from references given in the PDB data set itself if not described otherwise: 2MCM, 4RAT, 1END, 1NDK, 1GPR, 2CPL, 2GCR, 1GP1, 1ABK, 2CNA, 5TIM, 3BLM, 3SC2, 3DNI, 2CAB, 4BLM, 3TGL, 2HHM, 1PYP, 3TLN, 1ALA, 1BIA, 1MRR, 2LIV, 1IPD, 2LBP, 1FBA, 2TS1, 4ENL, 1PII, 6TAA, 2AAA, 1THG, 1AOZ.

As a ‘‘control group’’ (set III) we selected a set of 32 small proteins of up to 120 amino acids in length that, to the best of our knowledge, did not contain any binding sites. They are mostly ligands of other biological macromolecules or storage proteins (e.g. calcium ions): 6RLX, 1DFN, 6RXN, 1CBN, 8RXN, 1CAD, 1RDG, 1HPT, 1BTI, 5PTI, 1DTX, 1CDT, 1FAS, 4MT2, 3EBX, 1PI2, 2SN3, 1BOV, 1R69, 3IL8, 1HOE, 4ICB, 1UBQ, 1PGX, 2CI2, 1TEN, 2PLT, 7PCY, 1AAN, 1RTP, 4CPV, 1TGF. Furthermore, three larger proteins with no binding sites for low molecular mass ligands were considered: basic fibroblast growth factor 4FGF (146 amino acids), Kunitz proteinase inhibitor 1TIE (172 amino acids) and leukocyte elastase inhibitor 1HLE (346 amino acids).

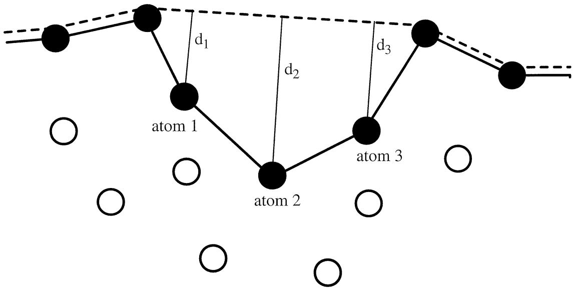


Figure 2. Definition of enveloping surface area and detailed surface area: (●) surface atoms; (○) inner atoms; (- - -), enveloping surface area (ESA); (—), detailed description of the surface area (DSA); d_i, distance between atom i and ESA. The mean distance $\bar{d}_i$ is defined as the average of the distance d_i and of all distances to ESA of the neighbours, e.g. for atom 2: $\bar{d}_2 = d_1 + d_2 + d_3 / 3$.

In total, the data sets used here represents about 75 different families of protein folds (Orengo *et al*., 1993; Lessel & Schomburg, 1994).

The co-ordinates of ligands were discarded from those data sets which contained them. Generally, co-ordinates of heteroatoms were not considered.

Method

To describe the shape of the molecule, the alpha-shape algorithm implemented by Edelsbrunner *et al*. (Mücke, 1993; Edelsbrunner & Mücke, 1994) was used. Their program is available *via* anonymous ftp from ftp.ncsa.uiuc.edu.

The starting-point for this alpha-shape algorithm is the 3-D structure of the molecule given as a set of points (the centres of the atoms) in three-dimensional Euclidean space. The alpha-shape is a one-parametric family of polytops, derived from the Delaunay triangulation of the set of points. This type of triangulation is characterised by the property that for each tetraeder its surrounding sphere contains no point of the set other than the four vertices of the tetraeder. This triangulation is uniquely defined.

Each component of the triangulation (tetraeder, triangle, edge) will be related to a size defined as the radius of its smallest circumsphere. Now, in dependence of a parameter α, the respective shape of the set of points is obtained by omitting all objects (tetraeder, triangle, edge) larger than α. For illustration, assume that the points of the set are built of some solid material and that the edges, triangles and tetraeders are made of some soft erasable material. Now one uses a spherical eraser with radius α. This eraser will be stopped only by the points. Using this eraser, all edges, triangles and tetraeders which are directly accessible to it are removed. The resulting surface is the surface of the alpha-shape.

The alpha-shape algorithm describes these surfaces as lists of adjacent triangles, called the face lists. Depending on α, one gets a more or less detailed description of molecule shape (Figure 2). For α = ∞ the convex envelope is obtained. For α-values smaller than one-half of the smallest inter-atomic distance or for α = 0 one obtains the

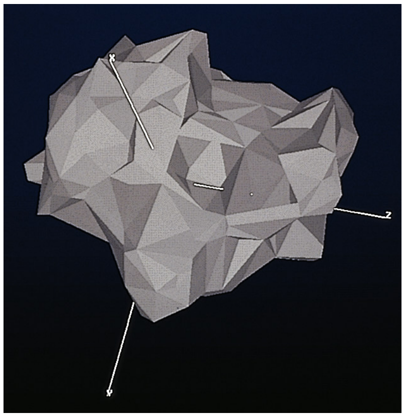


Figure 3. Alpha-shape ($\alpha = 4.0$ Å) of FK506 binding protein (Van Duyne *et al*., 1993); PDB code 1FKF (Bernstein *et al*., 1977).

set of points itself. α-Values between these two extremes give a more or less detailed description of the form of the set of points. Figure 3 shows the alpha-shape of the FK506 binding protein 1FKF (Van Duyne *et al*., 1993) without ligand for $\alpha = 4.0$ Å.

To locate cavities on the solvent accessible surface, we considered such cavities as significant deviations from the global form. This approach led us to the following steps: (1) estimation of the global geometrical form (envelope) of the molecule (ESA); (2) obtaining a suitably detailed description of the surface of the molecule (DSA); (3) comparison of these two surfaces.

The global form of the molecule (ESA)

For the first step, a ‘‘smoothed’’ surface was required that disregarded uneven regions. To obtain such an envelope, we selected alpha-shapes with a relatively large α parameter. For more or less spherical and compact molecules one can take $\alpha = \infty$ leading to the convex hull. For large and branched protein molecules with several clearly distinct domains, the convex envelope results in a cover quite different from the protein molecule due to the ‘‘arms’’ of the molecule and does not yield a useful global form. As a result, one has to limit α to suitable values. We found that alpha-shapes with $\alpha = 20.0$ Å best reflect the global shape of spherical protein molecules and globular parts of most multidomain proteins.

The detailed description of the molecule (DSA)

Here we had to choose a value of α that yields a shape reflecting the local structure of binding sites. In a binding pocket there has to be empty space between opposite protein atoms to accommodate at least one atomic layer of the ligand. The diameter of relevant atoms of ligand lies between 2.8 Å (oxygen) and about 4.5 Å (united carbon, e.g. CH_3-). Taking into account that the sum of van der Waals radii of neighbouring atoms of the protein lies also between 2.8 Å and 4.5 Å the smallest distance between two centres of atoms situated at both margins of the binding pocket is between 5.6 Å and 9.0 Å. Hence for $\alpha<4.5$ Å the resulting shape will trace all such pockets. On the other hand, we had to ensure that our shape is not too detailed and that it does not include all small pockets. Testing different α-values between 2.5 Å and 10.0 Å, we obtained the best results with α-values about 3.5 Å to 4.5 Å for the localisation of binding sites specific for ligands with molecular masses between 100 and 1000 Da.

Comparing detailed and global forms

After defining a detailed surface and an envelope surface area of the molecule we determined for each atom i of DSA its Euclidean distance d_i to ESA, called its deepness. Since the ESA is described as a list of triangles in the three-dimensional space we calculate this distance by taking the minimum over the distances from the given atom (defined as a point in three-dimensional space) to each triangle of the ESA. This deepness associated to each atom describes locally the detailed form (DSA) of the protein with respect to its global form (ESA). But binding pockets consist of several atoms and therefore we are not interested in very small (possibly deep) cavities. Therefore, this local description was not sufficient and we looked for a description including the neighbourhood of an atom of the DSA. Also, since the DSA is given as a list of triangles we could interpret the DSA as graph or net in the Euclidean space where the edges are straight lines. Hence the neighbourhood (or adjointness) of an atom of DSA is naturally given in a graph-theoretical sense. Now a slightly more global description was achieved by introducing the so-called mean deepness $\overline{d}_i$ to each atom i which is defined as the arithmetic mean of d_i and the deepnesses d_j of all neighbouring atoms j. This procedure results in a smoothing of the rugged DSA.

Then we got a list of those atoms at the surface which have large mean deepness. To get a connected binding region the next step was the clustering of this set of atoms. We used an agglomerative method with a cut-off of about 12.0 Å. For small proteins (chain length < 100 amino acids) the problem may arise that the clustering puts atoms together which lie on different sides of the protein. It is because the clustering takes the Euclidean distance between the atoms. This can be avoided by changing the parameter that controls the clustering or by making a visual inspection. The output can be coupled to interactive graphics programs which visualise the determined atoms.

The program APROPOS is written in C and is available for Sun and SGI machines on request from the authors. The average run time for a protein containing about 3000 non-hydrogen atoms is about three minutes.

Acknowledgements

We would like to express our thanks to Professor Edelsbrunner and his colleagues for the alpha-shape program and their help with the installation. This work was partially supported by grants of BMBF, Germany (K.P.P. and J.F.).

References

Arutiunian, E. G., Terzian, S. S., Voronova, A. A., Kuranova, I. P., Smirnova, E. A., Vainstein, B. K., Höhne, W. E. & Hansen, G. (1981). X-ray diffraction study of inorganic pyrophophatase from Baker’s yeast at the 3 Å resolution (russian). *Doklad. Akad. Nauk SSSR*, **258**, 1481–1485.

Bacon, D. J. & Moult, J. (1992). Docking by least-squares fitting of molecular surface patterns. *J. Mol. Biol.* **225**, 849–858.

Banaszak, L., Winter, N., Xu, Zh., Bernlohr, D. A., Cowan, S. & Jones, T. A. (1994). Lipid-binding proteins: a family of fatty acid and retinoid transport proteins. *Advan. Protein Chem.* **45**, 89–151.

Bernstein, F., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., 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 (Version October 1993).

Blake, C. C. F. Koenig, D. F., Mair, G. A., North, A. C. T. Phillips, D. C. & Sarma, V. R. (1965). Structure of hen egg-white lysozyme. *Nature*, **206**, 757–761.

Bode, W. & Huber, R. (1992). Natural protein inhibitors and their interaction with proteinases. *Eur. J. Biochem.* **204**, 433–451.

Bode, W., Papamokos, E. & Musil, D. (1987). The high-resolution X-ray crystal structure of the complex formed between subtilisin Carlsberg and eglin c, an elastase inhibitor from the leech Hirudo medicinalis. *Eur. J. Biochem.* **166**, 673–692.

Boulu, L. G., Crippen, G. M., Barton, H. A., Kwon, H. & Marletta, M. A. (1990). Voronoi binding site model of a polycyclic aromatic hydrocarbon binding protein. *J. Med. Chem.* **33**, 771–775.

Bradley, M. P. & Crippen, G. M. (1993). Voronoi modeling: the binding of triazines and pyrimidines to *L. casei* dihydrofolate reductase. *J. Med. Chem.* **36**, 3171–3177.

Casari, G., Sander, C. & Valencia, A. (1995). A method to predict functional residues in proteins. *Struct. Biol.* **2**, 171–178.

Connolly, M. L. (1986). Shape complementarity at the hemoglobin $\alpha_1\beta_1$ subunit interface. *Biopolymers*, **25**, 1229–1247.

Cooperman, B. S., Baykov, A. A. & Lahti, R. (1992). Evolutionary conservation of the active site of soluble inorganic pyrophosphatase. *Trends Biochem. Sci.* **17**, 262–266.

Derewenda, Z. S. (1994). Structure and function of lipases. *Advan. Protein Chem.* **45**, 1–52.

Drickamer, K. (1995). Multiplicity of lectin-carbohydrat interactions. *Nature Struct. Biol.* **2**, 437–439.

Edelsbrunner, H. & Mücke, E. P. (1994). Three dimensional alpha-shapes. *ACM Transact. Graph.* **13(1)**, 43–72.

Fanning, D. W., Smith, J. A. & Rose, G. D. (1986). Molecular cartography of globular proteins with application to antigenic sites. *Biopolymers*, **25**, 863–883.

Fischer, G. (1994). About PPIs and their effectors. *Appl. Chem.* **106**, 1479–1501.

Galat, A. & Metcalfe, S. M. (1995). Peptidylproline *cis/trans* isomerases. *Prog. Biophys. Mol. Biol.* **63**, 67–118.

Goodsell, D. S. & Olson, A. J. (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins: Struct. Funct. Genet.* **8**, 195–202.

Goodsell, D. S. & Olsen, A. J. (1993). Soluble proteins: size shape and function. *Trends Biochem. Sci.* **18**, 65–68.

Greer, J. & Bush, L. B. (1978). Macromolecular shape and surface maps by solvent exclusion. *Proc. Natl Acad. Sci. USA*, **75**, 303–307.

Jones, S. & Thornton, J. M. (1995). Protein–protein interactions: a review of protein dimer structures. *Progr. Biophys. Mol. Biol.* **63**, 31–65.

Kendrew, J. C., Watson, H. C., Strandberg, B. E. & Dickerson, R. E. (1961). A partial determination by X-ray methods, and its correlation with chemical data. *Nature*, **190**, 666–670.

Koshland, D. E., Jr (1958). Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl Acad. Sci. USA*, **44**, 98–104.

Kuhn, L. A., Siani, M. A., Pique, M. E., Fisher, C. L., Getzoff, E. D. & Tainer, J. A. (1992). The interdependence of protein surface topography and bound water molecules revealed by surface accessibility and fractal density measures. *J. Mol. Biol.* **228**, 13–22.

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **161**, 269–288.

Lessel, U. & Schomburg, D. (1994). Similarities between protein 3-D structures. *Protein Eng.* **7**, 1175–1187.

McPhalen, C. A., Strynadka, N. C. J. & James, M. N. G. (1991). Calcium-binding sites in proteins: a structural perspective. *Advan. Protein Chem.* **42**, 77–144.

Mizutani, M. Y., Tomiaka, N. & Itai, A. (1994). Rational automatic search method for stable docking models of protein and ligand. *J. Mol. Biol.* **243**, 310–326.

Mücke, E. P. (1993). Shapes and implementation in three dimensional geometry. PhD thesis, Department for Computer Sciences, University of Illinois at Urbana-Champaign, USA.

Norel, R., Fischer, D., Wolfson, H. J. & Nussinov, R. (1994). Molecular surface recognition by computer vision-based technique. *Protein Eng.* **7**, 39–46.

Novotny, J., Handschumacher, M., Haber, E., Bruccoleri, R. E., Carlson, W. B., Fanning, D. W., Smith, J. A. & Rose, G. D. (1986). Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains). *Proc. Natl Acad. Sci. USA*, **83**, 226–230.

Orengo, C. A., Flores, T. P., Taylor, W. R. & Thornton, J. M. (1993). Identification and classification of protein fold families. *Protein Eng.* **6**, 485–500.

Pabo, C. A. & Sauer, R. T. (1992). Transcribtion factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.* **61**, 1053–1095.

Perutz, M. F. (1970). Stereochemistry of cooperative effects in haemoglobin. *Nature*, **228**, 726–739.

Phillips, M. A. & Fletterick, R. J. (1992). Proteases. *Curr. Opin. Struct. Biol.* **2**, 713–720.

Robertus, J. D., Alden, R. A., Birktoft, J. J., Kraut, J., Powers, J. C. & Wilcox, P. E. (1972). An X-ray crystallographic study of binding of peptide chloromethyl ketone inhibitors to subtilisin BPN'. *Biochemistry*, **11**, 2439–2449.

Rotonda, J., Burbaum, J. J., Chan, H. K., Marcy, A. I. & Becker, J. W. (1993). Improved calcineurin inhibition by yeast FKBP12-drug complexes, crystallographic and functional analysis. *J. Biol. Chem.* **268**, 7607–7609.

Schechter, I. & Berger, A. (1967). On the size of the active site in proteases I. Papain. *Biochem. Biophys. Res. Commun.* **27**, 157–162.

Schlunegger, M. P. & Grütter, M. G. (1992). An unusual feature revealed by the crystal structure at 2.2 Å resolution of human transforming growth factor-b2. *Nature*, **358**, 430–434.

Shoichet, B. K. & Kuntz, I. D. (1991). Protein docking and complementarity. *J. Mol. Biol.* **221**, 327–246.

Van Duyne, G. D., Standaert, R. F., Karplus, P. A., Schreiber, S. L. & Clardy, J. (1993). Atomic structures of the human immunophilin FKBP-12 complexes with FK506 and Rapamycin. *J. Mol. Biol.* **229**, 105–124.

Wodak, S. J. & Janin, J. (1978). Computer analysis of protein–protein interaction. *J. Mol. Biol.* **124**, 323–342.

Yeates, T. O. (1995). Algorithms for evaluating the long-range accessibility of protein surfaces. *J. Mol. Biol.* **249**, 804–815.

Edited by R. Huber

(Received 11 September 1995; accepted 6 November 1995)