Protein Science (1994), *3*:1224–1235. Cambridge University Press. Printed in the USA.

Buried waters and internal cavities in monomeric proteins

MARK A. WILLIAMS,[1,2] JULIA M. GOODFELLOW,[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, United Kingdom
[2] Laboratory of Molecular Biology, Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, United Kingdom

(RECEIVED February 23, 1994; ACCEPTED June 1, 1994)

Abstract

We have analyzed the buried water molecules and internal cavities in a set of 75 high-resolution, nonhomologous, monomeric protein structures. The number of hydrogen bonds formed between each water molecule and the protein varies from 0 to 4, with 3 being most common. Nearly half of the water molecules are found in pairs or larger clusters. Approximately 90% are shown to be associated with large cavities within the protein, as determined by a novel program, PRO_ACT. The total volume of a protein's large cavities is proportional to its molecular weight and is not dependent on structural class. The largest cavities in proteins are generally elongated rather than globular. There are many more empty cavities than hydrated cavities. The likelihood of a cavity being occupied by a water molecule increases with cavity size and the number of available hydrogen bond partners, with each additional partner typically stabilizing the occupied state by 0.6 kcal/mol.

Keywords: buried water molecules; internal cavities; packing; thermostability

The interiors of proteins are well packed (Richards, 1977). The average volume occupied by peptide residues in the interior of a protein is similar to that in amino acid crystals or in solution. However, the packing of residues in the protein interior is not perfect. Atomic-sized cavities have been found in protein structures by crystallographic studies of xenon-protein complexes (Schoenborn, 1965; Tilton et al., 1984) and by theoretical methods that define accessible surfaces for proteins (Lee & Richards, 1971; Rashin et al., 1986; Tilton et al., 1986). These cavities frequently contain "buried" water molecules, which are isolated from the bulk solvent (Finney, 1977; Edsall & McKenzie, 1983; Baker & Hubbard, 1984).

Site-directed mutagenesis studies involving the replacement of a large side chain by a smaller one may produce an internal cavity. Such a mutation may destabilize folded proteins and their complexes by several kilocalories per mole (Fersht, 1985; Sandberg & Terwilliger, 1989; Eriksson et al., 1992). This destabilization has 3 components, which may be of similar magnitude: reduction in buried hydrophobic surface area, changes in intramolecular hydrogen bonding, and a "packing" term thought to be primarily due to loss of van der Waals contacts. A packing term can be associated with any cavity found in proteins, and consequently, cavities are regarded as destabilizing a folded protein. Although internal cavities have been referred to by the term "packing defect" (Connolly, 1985), there is evidence that the cavities in some proteins have a functional role, e.g., the diffusion of oxygen to the heme in myoglobin probably occurs through connected cavities within the protein molecule (Tilton et al., 1986).

Buried water molecules are often conserved between proteins belonging to the same homologous family, e.g., the serine proteases (Screenivasan & Axelsen, 1992). This suggests that they are structurally or functionally important. Buried water molecules may stabilize a protein structure by providing the otherwise missing van der Waals interactions for those atoms bordering a cavity and hydrogen bonding to otherwise unsatisfied protein hydrogen bonding groups. Despite their apparent isolation, buried water molecules exchange with bulk solvent, although they do so much more slowly than water hydrogen bonded to the surface of the protein (Otting et al., 1991). Meyer et al. (1988) have proposed that the movement of water through interior cavities facilitates the displacement of water from the active site of the serine proteases as a substrate binds.

In this paper, we present a survey of the properties of hydrated and nonhydrated cavities in protein structures. Previous surveys of buried water molecules (Finney, 1977; Vinogradov, 1980; Edsall & McKenzie, 1983; Baker & Hubbard, 1984) and protein cavities (Rashin et al., 1986) considered at most 15 high-resolution structures. Many more high-resolution protein struc-

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, UK; e-mail: thornton@uk.ac.ucl.bioc.bsm.

tures are now available. We have been able to accurately quantify earlier observations concerning the hydrogen bonding of buried water molecules. The increase in data has revealed previously unobserved features of the internal cavities within proteins of different size and structural class. By considering the local environments of both the hydrated and nonhydrated cavities, we have been able to estimate the energy involved in creating a cavity and the strength of the hydrogen bonds made by the buried water molecules.

Methods

High-resolution data set

Seventy-five high-resolution (<2.5 Å) protein structures were selected from the April 1993 Protein Data Bank (Bernstein et al., 1977) such that there was no sequence or structural homology between members of the set (sequence identity of <30% and SSAP analogy score of <80; Orengo et al., 1993). The set of proteins may be found in Table 1. Only protein monomers are considered in this study; the cavities and buried water molecules between the subunits of oligomers are being investigated separately as part of a broader survey of the character of protein-protein interactions.

Identifying buried water molecules and cavities

A buried water molecule is defined as one that cannot be connected by a continuous series of water-water hydrogen bonds to bulk water molecules. It is impossible to determine whether or not a water molecule is buried by simply calculating its solvent accessibility because a buried water molecule may have non-zero solvent accessibility if it is in an interior cavity that is large enough to contain other buried water molecules. A program, PRO_ACT (M.A. Williams), has been written that identifies buried water molecules by filling the solvent-accessible surface of a protein with experimental and computational water molecules and then identifying those experimental water molecules that are not connected to bulk water by a chain of hydrogen bonds.

In order to identify buried waters and interior cavities in a consistent manner, we have defined "solvent accessibility" and "interior cavity" in a manner consistent with a particular definition of atom-atom contact. This definition of contact does not use a single radius to describe each atom but allows the apparent radius of an atom to vary with the nature of its interacting partner, in accordance with observations of atom-atom interactions in protein structures.

Definitions of contact between atoms

The atom parameters used by PRO_ACT are based on studies of water-protein atom distances (Thanki et al., 1988; Walshaw & Goodfellow, 1993) and protein-protein atom distances observed in high-resolution protein structures. The water-atom and atom-atom distances have a strongly peaked distribution at short range due to the action of intermolecular forces. The observed distribution for each atom pair is usually used to define a single van der Waals radius. In our method a polar atom (O, N, or water) is assigned 2 characteristic radii (CR), a characteristic polar radius (CPR) that it exhibits in interactions with polar groups, and a characteristic apolar radius (CAR) that it exhibits in interactions with apolar groups. These radii are defined such that the sum of the appropriate radii of a pair of atoms equals the most probable value of the separation of that atom pair determined from a set of high-resolution protein structures. Each atom also has an associated maximum radius (MR) for polar (MPR) and apolar (MAR) interactions. The maximum radii are defined such that their sums equate with the largest values of the appropriate atom-atom distances for which the distribution is distinguishable from random. In defining these radii from crystallographic data on proteins, hydrogen atoms are not considered explicitly but are effectively included in the parameters of the other atoms. The characteristic and maximum radii of each atom type are given in Table 2 and some of the distributions used to obtain them are illustrated in Figure 1A.

These characteristic and maximum radii allow us to define contact between 2 atoms in a way that reflects the nature of the polar and apolar interactions observed in proteins. Solvent accessibility and interior cavity can then be defined in a way that is consistent with the definitions of contact and hence with the nature of interactions in proteins. Contact between atoms is defined as follows:

1. A pair of atoms are in contact if they are separated by less than the sum of their apolar maximum radii.
2. A pair of polar atoms are in polar contact if they are separated by less than the sum of their polar maximum radii.

The effects of these definitions on the classification of contacts between polar atoms are illustrated in Figure 1B.

Solvent accessibility

For any atom exposed to solvent, the water molecules in contact with it are most likely to lie at a distance equal to the sum of the appropriate characteristic radii of the atom and water, e.g., a water interaction with oxygen is most likely to lie at a distance equal to the sum of the CPR of water and the CPR of oxygen, i.e., (1.35 + 1.50) = 2.85 Å. In a similar manner to Lee and Richards (1971), the solvent-accessible surface of a molecule is defined by the nonoverlapping surface produced by placing a sphere on each atom of the molecule. The radius of any particular atom's sphere is set equal to the sum of the appropriate characteristic radii of that atom and water.

The above definition and the atomic parameters in Table 2 produce atomic solvent accessibilities for polar and apolar atoms that are individually proportional to those produced by ACCESS (S. Hubbard, University College London), a program that implements the Lee and Richards (1971) algorithm. However, our definitions (implemented in PRO_ACT) slightly increase the polar atom accessibilities relative to those of the apolar atoms.

Hierarchical classification of the extent of burial of water molecules

We propose that the buried waters associated with a protein structure can be identified in the following manner (which is illustrated in Fig. 2). The solvent-accessible surface of the protein is covered with water molecules (represented by spheres of 1.5-Å radius) by first adding spheres at the crystallographically determined sites and then placing "computational" water mol-

Table 1. *The nonhomologous monomeric proteins*

Protein[a]		Resolution	Structural class[b]	No. of residues	X-ray waters/residue[c]	No. of buried waters	No. of probes[d]	No. of coincident waters[e]	Total probe volume (Å^3)
2ovo	Ovomucoid third domain	1.5	$\alpha+\beta$	56	0.6	0	1	0	4
5rxn	Rubredoxin	1.2	β	55	1.9	1	4	1	23
5pti	Pancreatic trypsin inhibitor	1.0	$\alpha+\beta$	60	1.1	1	5	1	25
3ebx	Erabutoxin	1.4	β	63	1.8	0	2	0	15
1sn3	Scorpion neurotoxin	1.8	β	66	1.1	1	5	1	25
1cseI	Subtilisin inhibitor	1.2	$\alpha+\beta$	65	1.3	1	2	0	10
1bovA	Verotoxin	2.2	$\alpha+\beta$	72	0.3	1	13	1	68
1utg	Uteroglobin	1.3	α	70	1.2	0	4	0	22
1hoe	Amylase inhibitor	2.0	β	74	0.3	1	7	1	36
2hipA	High potential Fe/S protein	2.5	β	72	0.7	0	3	0	16
4icb	Bovine calbindin D9K	1.6	α	78	0.7	0	9	0	53
1ubq	Ubiquitin	1.8	$\alpha+\beta$	76	0.8	0	13	0	67
1tpkA	Kringle 2	2.4	$\alpha+\beta$	86	0.4	1	7	1	37
3b5c	Cytochrome *b5*	1.5	$\alpha+\beta$	87	1.0	0	10	0	54
9rnt	Ribonuclease T1	1.5	$\alpha+\beta$	105	1.2	2	11	2	70
2sicI	Subtilisin inhibitor	1.8	$\alpha+\beta$	107	0.7	2	14	2	95
4cpv	Calcium binding parvalbumin	1.5	α	111	0.7	1	22	1	132
2rhe	Immunoglobulin	1.6	β	114	1.6	2	13	1	78
2trxA	Thioredoxin	1.7	α/β	116	0.6	1	17	1	99
1fkf	FK506 binding protein	1.7	$\alpha+\beta$	108	0.7	4	11	3	72
1msbA	Mannose b.p.	2.3	$\alpha+\beta$	117	0.5	3	14	3	76
1ycc	Cytochrome *c* (reduced)	1.2	α	111	1.0	4	16	4	94
4bp2	Prophospholipase A2	1.6	$\alpha+\beta$	120	0.5	0	8	0	46
1paz	Pseudoazurin	1.5	β	121	0.8	0	16	0	78
256bA	Cytochrome *b562*	1.4	α	110	0.8	0	15	0	82
3fgf	Basic fibroblast g.f.	1.6	β	126	0.6	9	35	8	202
7rsa	Ribonuclease A	1.3	$\alpha+\beta$	125	1.5	1	15	1	72
3chy	Chey protein	1.7	α/β	131	1.7	4	24	4	181
2cdv	Cytochrome *c*3	1.8	$\alpha+\beta$	111	0.4	7	16	7	121
2azaA	Azurin	1.8	β	134	1.0	2	15	2	81
1ifc	Intestinal fatty acid b.p.	1.1	β	131	1.8	4	19	4	110
1cobA	Superoxide dismutase	2.0	β	153	0.7	2	16	2	94
1f3g	Phosphocarrier III	2.1	β	150	0.1	3	25	3	144
9wgaA	Wheat germ agglutinin	1.8	β	171	0.7	1	1	1	4
1rnh	Ribonuclease H	2.0	$\alpha+\beta$	152	0.6	4	21	4	136
1mbc	Myoglobin	1.5	α	155	0.9	1	42	1	274
3lzm	Lysozyme	1.7	$\alpha+\beta$	164	0.9	5	30	4	189
4dfrA	Dihydrofolate reductase	1.7	α/β	164	1.3	1	22	1	115
5p21	Ras p21 protein	1.4	α/β	168	1.3	7	38	7	244
2fcr	Flavodoxin	1.8	α/β	174	0.6	9	32	8	232
1rbp	Retinol binding protein	2.0	β	176	0.9	2	22	2	127
3sdpA	Iron superoxide dismutase	2.1	α	187	0.5	3	51	1	385
1gcr	γ-Crystallin	1.6	β	174	0.7	3	16	3	95
4ptp	β-Trypsin	1.3	β	221	0.8	18	48	16	339
2fb4H	Immunoglobulin FAB	1.9	β	221	0.2	1	20	1	130
9pap	Papain	1.6	md	241	0.8	23	53	21	402
4cla	Chloramphenicol acetyltrf.	2.0	$\alpha+\beta$	216	0.9	7	38	7	250
2cna	Concanavalin A	2.0	β	239	0.02	1	61	1	383
1gst	Glutathione S-trf.	2.2	md	218	1.1	11	48	10	316
1cseE	Subtilisin carlsberg	1.2	α/β	276	1.3	23	40	18	244
3blm	β-Lactamase	2.0	md	257	0.8	11	50	10	337
2ca2	Carbonic anhydrase	1.9	md	258	0.7	6	53	6	321
2tscA	Thymidylate synthase	2.0	$\alpha+\beta$	266	0.7	10	66	9	409
2cyp	Cytochrome *c* peroxidase	1.7	md	294	0.9	23	52	21	362
2gbp	Glucose binding protein	1.9	α/β	311	0.7	6	58	5	344
2er7E	Endothelial aspartic protease	1.6	β	325	1.0	11	65	10	399
5cpa	Carboxypeptidase[f]	1.5	α/β	308	1.0	22	64	21	426
2tsl	Tyrosyl-transfer RNA synthase	2.3	α/β	317	0.3	3	51	3	310
6tmnE	Thermolysin	1.6	md	324	0.5	13	74	13	475

(*continued*)

Table 1. *Continued*

Protein[a]		Resolution	Structural class[b]	No. of residues	X-ray waters/residue[c]	No. of buried waters	No. of probes[d]	No. of coincident waters[e]	Total probe volume (Å^3)
6ldh	Lactate deh.	2.0	α/β	332	0.9	4	59	4	354
1gd1O	Glyceraldehyde phosphate deh.	1.8	α/β	335	0.5	18	70	16	475
2liv	l/i/v Binding protein	2.4	α/β	344	0.4	6	67	2	399
1ipd	Isopropylmalate deh.	2.2	α/β	347	0.2	9	75	9	464
3bcl	Bacteriochlorophyll protein	1.9	md	351	0.3	9	88	8	567
1ovaA	Ovalbumin	1.9	md	374	0.9	8	102	7	676
1nsbA	Neuraminidase sialidase	2.2	β	392	0.6	25	71	23	463
7aatA	Aspartate aminotrf.	1.9	α/β	402	0.8	15	98	14	627
1phh	*p*-Hydroxybenzoate hydroxylase	2.3	α/β	396	0.7	12	111	11	765
2cpp	Cytochrome P450cam	1.6	md	407	0.5	13	116	13	719
4enl	Enolase	1.9	α/β	438	0.8	34	109	31	776
1csc	Citrate synthase	1.7	md	431	0.2	16	113	15	751
3grs	Glutathione reductase	1.5	α/β	463	1.1	14	73	14	425
1cox	Cholesterol oxidase	1.8	md	503	1.1	60	146	52	1,112
1lfi	Lactoferrin	2.1	md	700	0.4	18	128	18	826
8acn	Aconitase	2.0	md	755	0.4	64	211	58	1,442
Totals for the set of proteins				16,228		608	3,160	552	20,471

[a] Abbreviations: b.p., binding protein; deh., dehydrogenase; g.f., growth factor; trf., transferase.
[b] The structural class of the protein is taken from Orengo et al. (1993) (md, multidomain protein).
[c] The number of water molecules per protein residue given in the PDB file.
[d] The number of probes of radius ≥0.95 Å that can be accommodated within the structure.
[e] The number of buried water molecules coincident with a probe (i.e., probe-water distance ≤1.7 Å).
[f] Two water molecules were removed from the 5cpa structure because they clashed with other water molecules.

ecules at random points on the surface until no gap remains in which a sphere can be placed without it overlapping another. The solvent accessibility of each water molecule is then calculated and those found to be accessible are marked 1. The solvent-accessibility calculation is then repeated, but those molecules marked 1 are excluded from the structure. Molecules found to be accessible are marked 2. The accessibility calculation is then repeated, but this time molecules marked 1 or 2 are excluded. This process is repeated until no surviving water molecules are found to be solvent accessible. Of the remaining unmarked crystallographic water molecules, those that are determined to be in polar contact with marked molecules are themselves marked. The remaining unmarked molecules are the buried water molecules. Molecules marked 3 or above occupy clefts in the protein structure.

Clearly the results of the above procedure, like any accessibility calculation, depend on the atomic parameters used therein. Reducing or increasing the water molecule radius by 0.2 Å typically alters the number of buried water molecules determined for a protein by 1.

Table 2. *Characteristic and maximum atomic radii*[a]

Atom type	CPR	MPR	CAR	MAR	VDWR
Carbon			1.93	2.23	2.00
Sulfur			1.68	1.94	1.80
Oxygen	1.35	1.62	1.62	2.06	1.70
Nitrogen	1.45	1.67	1.82	2.41	1.85
Water	1.50	1.63	1.97	2.59	1.40

[a] All radii are given in Å. CAR is the characteristic apolar radius of the atom; CPR is its characteristic polar radius; MAR and MPR are the corresponding maximum radii of the atom. Given for comparison is the van der Waals radius (VDWR) ascribed to each atom by Tilton et al. (1984).

Identifying cavities

Having defined contact between atoms and filled the solvent-accessible surface of the protein with crystallographic and computational water molecules, a cavity can be defined: a cavity exists at any point between protein atoms, or between protein and water molecules, that does not lie within the apolar maximum radius of a protein atom or water molecule. PRO_ACT identifies interior cavities in the following way.

First, an "interior surface" of the protein is defined by the nonoverlapping surface associated with a protein molecule, produced by placing a sphere on each atom or nonburied water (crystallographic or computational) of radius equal to the MAR of that atom. A small probe sphere is then placed at a random point on this interior surface. This probe sphere is compelled to move a small step in a direction that increases the distance to the nearest atom's "apolar" surface (defined by placing a sphere on each atom with a radius equal to the atom's CAR). The move is accepted provided that the probe's center does not cross the interior surface of the protein. If the move is not accepted, the step length is halved and a new move attempted. After each accepted move, the probe's radius is increased to equal

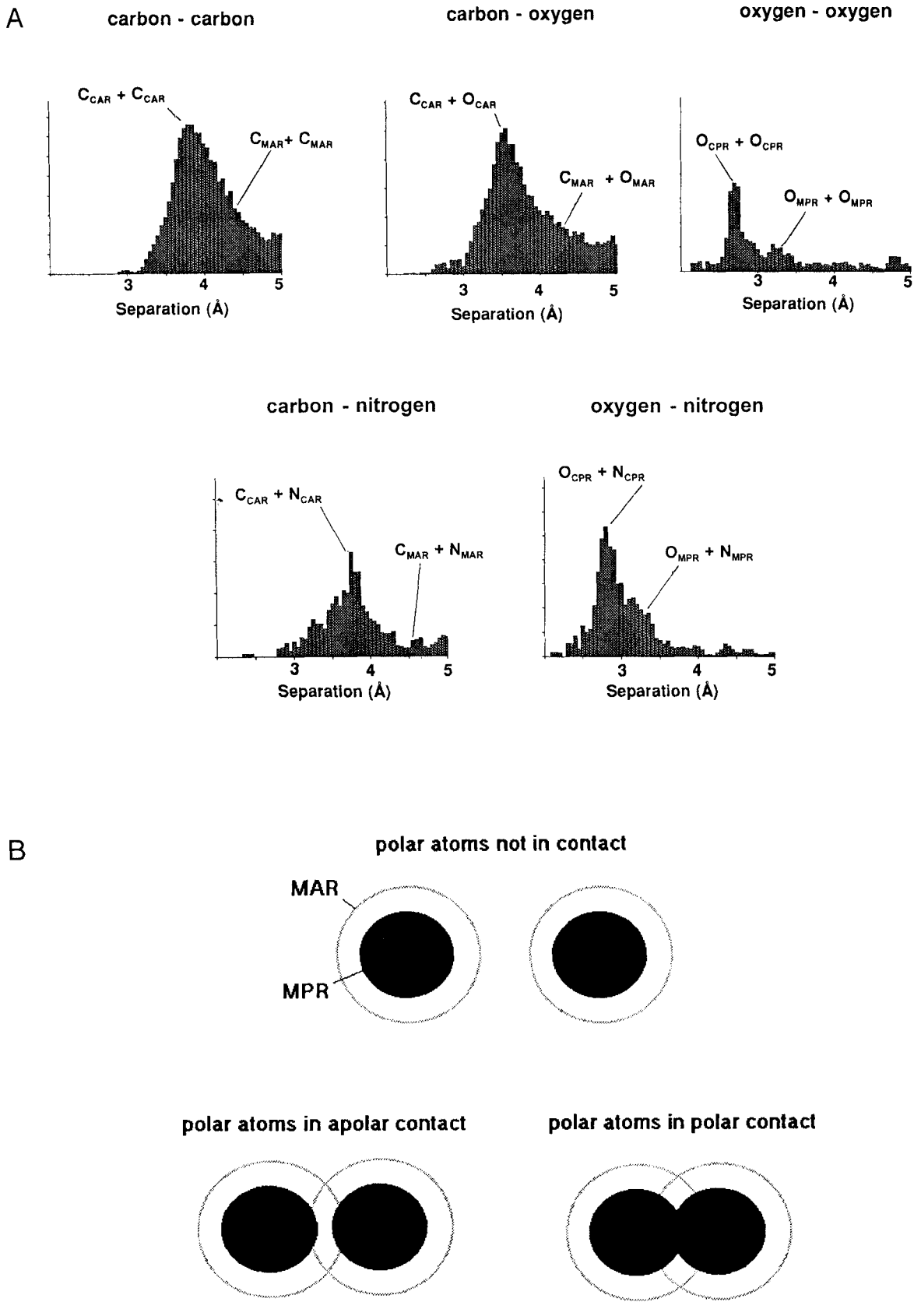

Fig. 1. A: Normalized distributions of the shortest inter-side-chain atom–atom distances for several pairs of atoms. The characteristic and maximum radii for these atoms (Table 2) determined from each distribution are marked. **B:** The classification of contact between polar atoms, illustrating the definitions given in the text.

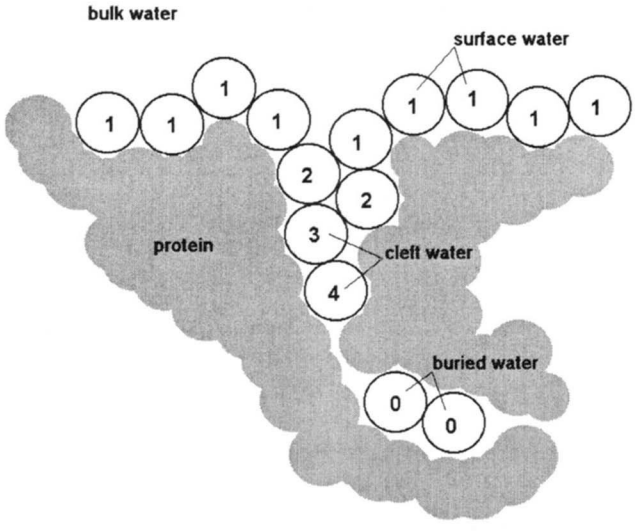

Fig. 2. Sketch illustrating the process of determining whether or not a water molecule is buried. The molecules are marked with a number that indicates their degree of separation from bulk solvent—those marked 0 are buried, those marked 3 or higher are within a cleft on the protein's surface.

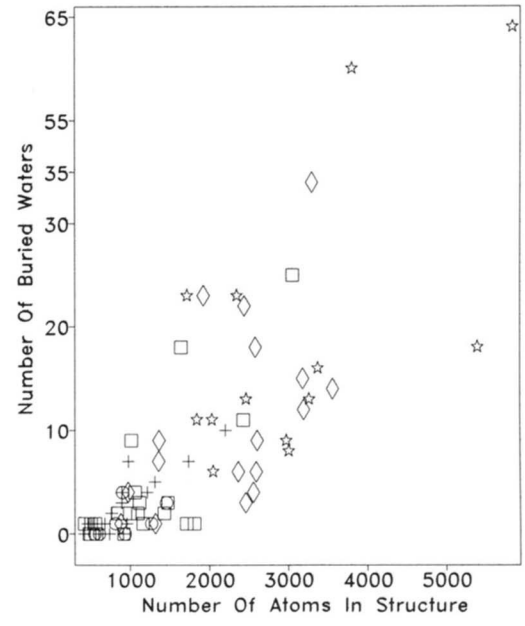

Fig. 3. Number of buried water molecules found in each of the proteins listed in Table 1 plotted against the size of the structure (including HETATOMs) and distinguishing proteins of different structural class. ○, An all α protein; □, an all β protein; ◇, an α/β protein; +, an $\alpha+\beta$ protein; ☆, a multidomain protein. The structural class of each protein is taken from Orengo et al. (1993).

the new distance to the nearest atom's apolar surface. The probe continues to move until its radius cannot be increased by at least 0.025 Å. If this final probe radius is greater than a specified value, the probe is retained; if not, it is discarded. An accepted probe fills a small spherical cavity in the structure and subsequent probes may not approach it more closely than the sum of their radii. The process is repeated until no more probes are accepted.

The result of this procedure is that the cavities in the protein are filled with spherical probes each having a radius greater than some specified radius. As this specified radius is reduced, more and smaller probes fill narrower cavities. In the studies presented here we consider only "large" cavities that are "wide" enough to contain probes of radius greater than 0.95 Å (slightly less than half the CAR of water). Although many narrower cavities exist within proteins, only the wider cavities can contain water molecules. A probe radius of 0.95 Å is a convenient, though somewhat arbitrary, dividing line between small (unlikely to be hydrated) cavities and large cavities with a reasonable (5% or greater) chance of being hydrated. In order that we can make comparisons between the buried water molecules determined crystallographically and the probes generated by PRO_ACT, we also discount those few probes that are less than the water–water polar contact distance from an unburied water molecule.

Results

Buried water molecules and their interactions

The number of experimentally determined water molecules that we have identified as buried for each protein structure shows a tendency to increase with protein size (Fig. 3). There is on average 1 buried water molecule per 27 residues, although there is a wide variation in the number of buried waters per residue for individual proteins. The total number of waters (surface, cleft, and buried) per residue that is reported for a protein structure also varies widely (Table 1), and it might be supposed that the variation in the number of buried waters is mainly a reflection of this. However, there is, in general, no correlation between the total number of water molecules found experimentally per residue and the number of buried waters reported per probe site. The number of buried waters seems to generally reflect real structural features of the protein. Aconitase (Lauble et al., 1992) and cholesterol oxidase (Vrielink et al., 1991) have by far the largest number of buried water molecules—both having nearly twice the third highest number (Table 1). This is partly explained by the presence of a buried binding site in these proteins, which in aconitase contains 17, and in cholesterol oxidase 13, buried water molecules. There is no clear dependence of the number of buried waters per residue on structural type. However, the average ratio for the all α-helical proteins—1 buried water per 91 residues—is the lowest of the 5 classes of proteins distinguished in Table 1.

The distribution of the distances of each buried water to its nearest unburied water molecule is shown in Figure 4, together with the distribution of all protein atoms to their nearest unburied water molecule (dotted line). Buried water molecules occur at all distances from the surface of the protein and occur most frequently about 1 atom deep (i.e., with 1 layer of protein atoms between the buried waters and the surface waters).

The buried water molecules show a wide spread in the number of polar contacts that they make, with 45% of waters making 3 polar contacts, 37% making 4 or more, and only 18% making 2 or fewer (Fig. 5A, and examples in the kinemages). The average number of polar contacts made by each buried water molecule is 3.23. Previous studies (Finney, 1977; Edsall &

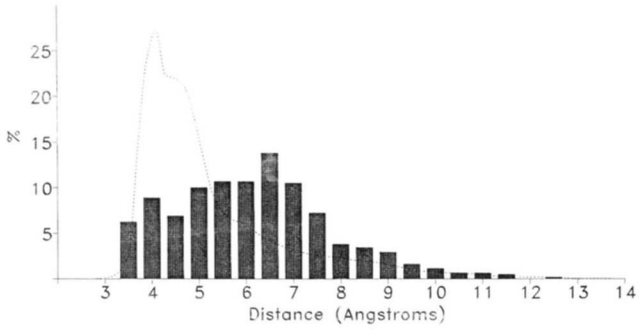

Fig. 4. Distribution of the distances of each of the buried water molecules found in the proteins listed in Table 1 to their nearest surface water. The dashed line is the distribution of distances of the protein atoms to their nearest surface water (on the same scale).

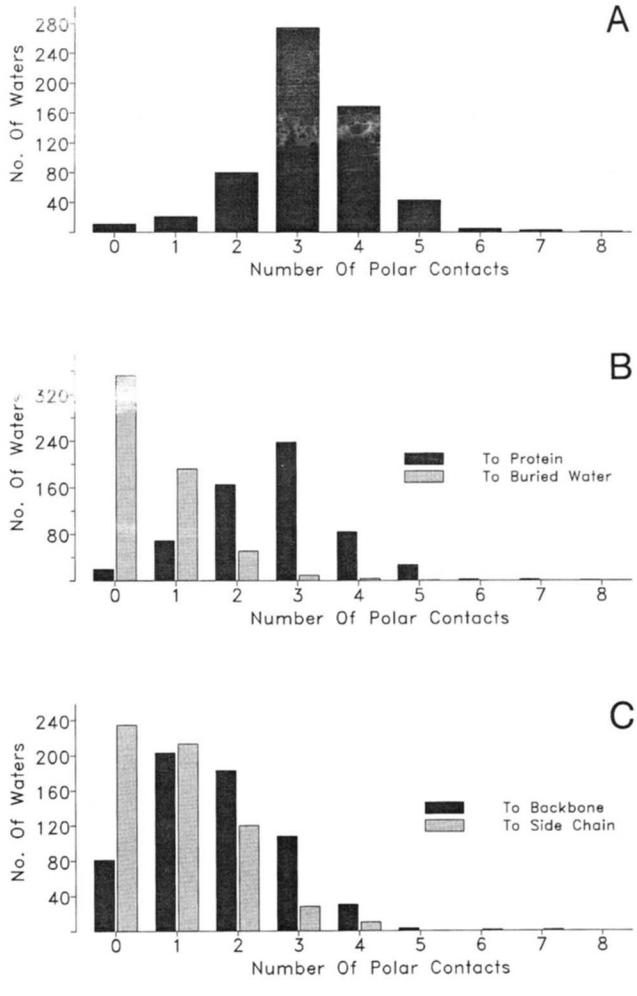

Fig. 5. Distributions of the number of polar contacts made by the buried water molecules. **A:** Total number of polar contacts. **B:** Number of polar contacts made with protein atoms and the number of polar contacts made with other buried water molecules. **C:** Number of polar contacts made with protein side-chain or protein backbone atoms.

McKenzie, 1983; Baker & Hubbard, 1984; Rashin et al., 1986) have suggested that a buried water molecule typically makes 3 hydrogen bonds with the protein. Although polar contacts are only potential hydrogen bonds (whose enumeration would require a complex analysis of the possible local hydrogen bonding networks), the distributions of the numbers of polar contacts made to the protein and to other buried water molecules (Fig. 5B) essentially confirm the results of the earlier studies in that the most common number of polar contacts to the protein is 3. However, they also show that 42% of the buried water molecules make 2 or fewer polar contacts with the protein and usually compensate for this shortage of hydrogen bonds to protein by make polar contacts with other buried water molecules—42% of buried waters make a polar contact to at least 1 other buried water. The total polar contacts are comprised of 53% to protein backbone, 30% to side chain, and 17% to other buried water molecules (Fig. 5B,C and examples in the kinemages). Only 3% of the water molecules make no polar contact to the protein. Polar contacts to protein are made to hydrogen bond acceptors and donors in the ratio 3:2.

We call a set of buried water molecules that are connected by a continuous network of polar contacts a "cluster." The distribution of water molecules among clusters of different sizes is shown in Figure 6. We find that 58% of water molecules are isolated from other buried water molecules, while 22% belong to clusters containing 2 molecules and 20% to larger clusters. The largest observed clusters are the 17- and 13-molecule clusters found, respectively, in the active sites of aconitase (Lauble et al., 1992) and cholesterol oxidase (Vrielink et al., 1991), and a 6-molecule cluster observed in papain (Kinemage 3; Kamphuis et al., 1984). Large clusters (4 or more water molecules) are generally elongated rather than globular.

Interior cavities

The number of large probes, which define wide cavities, that can be accommodated within each of the protein structures is given in Table 1. It can be seen that, contrary to conclusion of a previous study on a smaller set of proteins (Rashin et al., 1986), there is, on average, a monotonic increase in the volume of the wide cavities (i.e., cavities with a diameter of at least 1.9 Å) associated with a protein as the size of the protein increases (Fig. 7A). If we take the average residue volume to be 143 $Å^3$—calculated from data in Tables 1.1 and 6.3 of Creighton (1993)—the data in Table 1 imply that wide cavities form between 0.002% and 1.55% of the volume of a protein (note that the total volume of all cavities is roughly 4 times that of the wide cavities that are the focus of this study).

The proportion of a globular protein's residues that are buried increases with protein size, in accordance with the relationship (number of buried residues)$^{1/3}$ = (number of residues)$^{1/3}$ − constant (Janin, 1979). Consequently, it is expected that the ratio of cavity volume to protein size will increase slowly with protein size. Our data clearly show such an increase. We cannot, however, reliably determine a quantitative relationship between average cavity volume and protein size because the cavity volumes associated with proteins of a similar size may differ substantially. This variation might be supposed to be due to different packing in different structural classes of protein, to differences in the accuracy of the experimental information used to determine the structures, or to the relative globularity of the structures. We have found no general difference in the cavity volume of proteins of the same size but different structural class

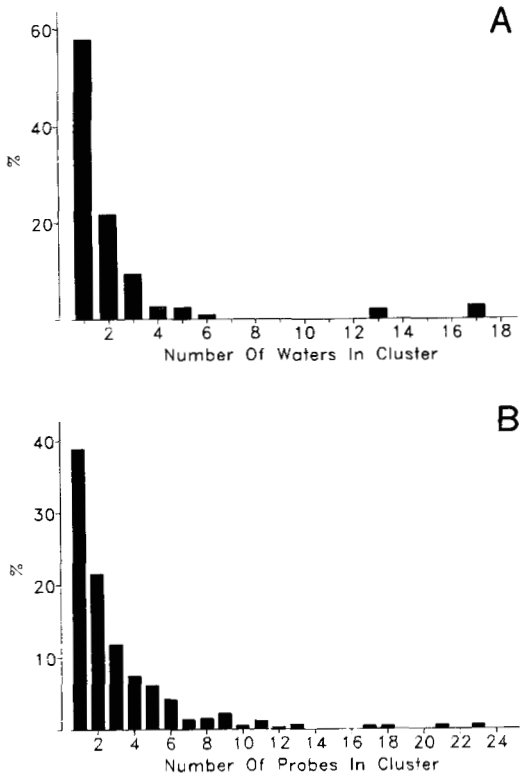

Fig. 6. Distribution of buried water molecules (**A**) and probes (**B**) among clusters of increasing size.

(α, β, α/β, $\alpha+\beta$, multidomain), nor any discernible dependence of the total cavity volume of proteins of similar size on the resolution of the structure. However, it is clear that several unusual structures have low interior cavity volumes in comparison to normal globular proteins of similar size, e.g., wheat germ agglutinin (Wright, 1990). In order to allow for variation in globularity, we have investigated the relationship between cavity volume and the number of buried (zero solvent accessibility) atoms in the structure. Although the observed trend is consistent with the hypothesis that, on average, the cavity volume increases linearly with the number of buried residues in the structure, substantial variation exists, and it is not possible to derive any reliable quantitative expression of this trend.

There are many more sites capable of accommodating 0.95-Å probes in each protein than there are buried water molecules in that protein (Fig. 7B). The distribution of the distance between a probe and its nearest buried water molecule is nonrandom for distances less than 1.7 Å. This fact allows us to separate the probes into 2 classes—those that are less than 1.7 Å from a buried water molecule (hydrated probe sites) and those that are not (nonhydrated probe sites). In total, 18% of the probe sites can be regarded as hydrated. Conversely, 91% of buried waters are less than 1.7 Å from a probe site. There are several causes of the high proportion of nonhydrated probe sites. The nonhydrated probe sites have a much smaller average number of polar contacts than the hydrated sites (Fig. 8A). This suggests that burial of a water molecule within many of the cavities within a protein is thermodynamically unfavorable because they lack sufficient hydrogen bonds to compensate for the loss of bonds to bulk water as the water is buried. There is also a strong tendency for the proportion of probe sites hydrated to be reduced as the radius of the cavity is reduced (Fig. 8B), presumably due to the greater van der Waals repulsion between water and protein in the smaller cavities. There is no clear relationship between the proportion of the probe sites that are hydrated and their distance from the surface (Fig. 8C).

The data in Table 1 show that α-helical proteins have a relatively small average number of buried water molecules per residue but have similar cavity volumes to proteins of other structural classes. The average number of polar contacts made by a water placed at each probe site of a protein is plotted for each protein in Figure 9. The small number of buried waters in the all α proteins is probably explained by the low average number of polar contacts that could be made between buried water

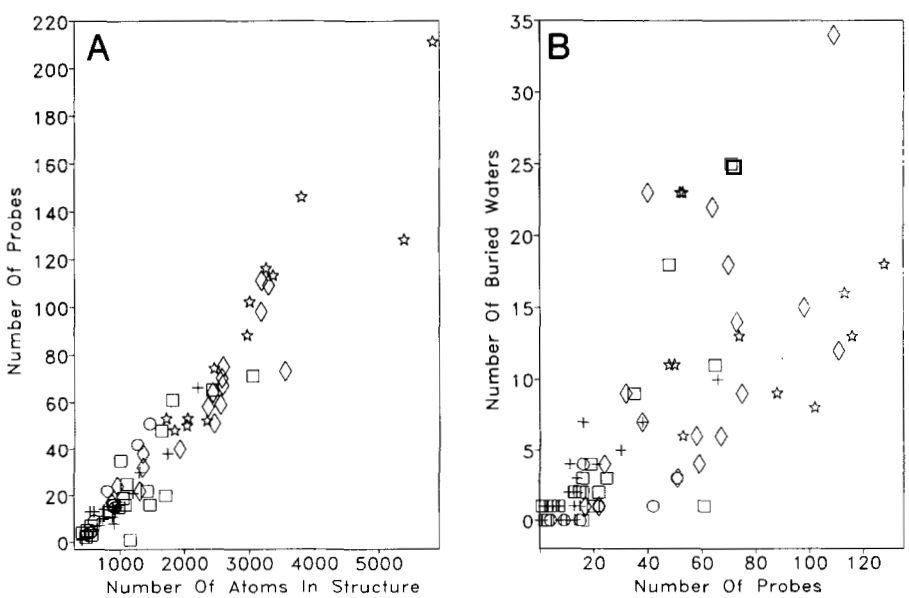

Fig. 7. A: Total number of probes of radius ≥0.95 Å that can be accommodated within each of the proteins listed in Table 1. **B:** Number of buried water molecules found within each protein plotted against the number of probes that can be accommodated within that protein. The structural class of each protein is indicated by the symbols as in Figure 3. For clarity, the cholesterol oxidase and aconitase structures are omitted from this plot.

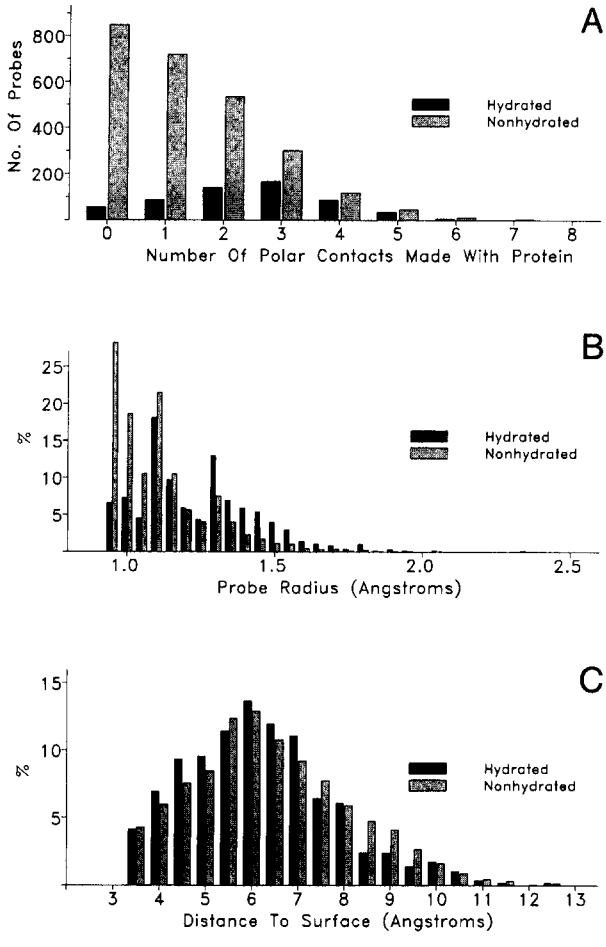

Fig. 8. **A:** Distribution of polar contacts that are made by computational water molecules placed at each of the probe sites. The sites that are hydrated by a crystallographically determined buried water are distinguished from those that are not. **B:** Distribution of the radii of the probes associated with hydrated and nonhydrated sites. **C:** Distribution of the distances from each probe to its nearest surface water molecule.

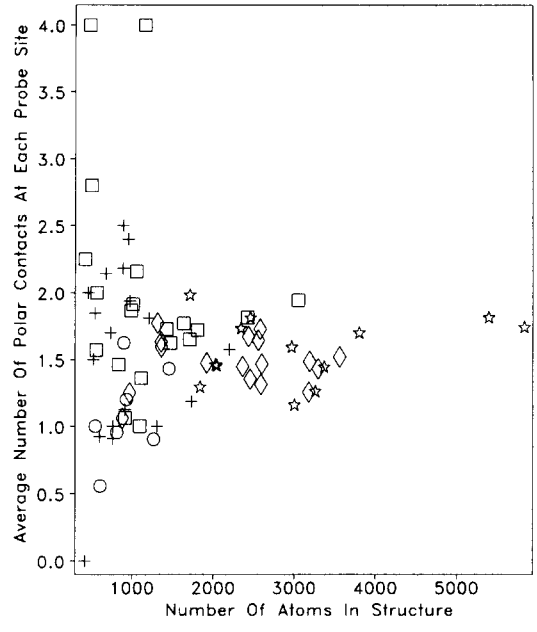

Fig. 9. Average number of polar contacts made by a computational water molecule placed at each of the probe sites determined for each of the proteins in Table 1.

protein because the main-chain donor and acceptor groups are nearly all involved in intrahelical hydrogen bonds.

The distribution of probes among clusters of different size is shown in Figure 6B (clusters are defined as for buried water molecules using the same distance criterion). It appears that the large clusters of probes (>3 members) are typically rather larger than the large clusters of buried water molecules. Six of the 10 largest clusters of probes are found in cytochrome P450cam (Poulos et al., 1987), where these large cavities are only sparsely hydrated (Fig. 10). Like myoglobin, this is a heme-binding protein that requires oxygen to bind to the heme for its function. The large clusters of probes found in cytochrome P450cam may indicate multiple pathways along which oxygen can diffuse through the protein to and from the heme, in a similar way to the proposed diffusion of oxygen through cavities within myoglobin (Tilton et al., 1986).

Discussion

A novel procedure has been developed that defines solvent accessibility, contacts between protein atoms, and cavities within proteins in a consistent way, and that also takes into account the apparent change in an atom's radius depending on its interaction partner. The solvent accessibilities found by this procedure are broadly consistent with the Lee and Richards (1971) approach, although exposure of polar atoms to solvent is increased—in line with their ability to hydrogen bond to water molecules.

The position and width of a cavity are defined in terms of the properties of, and the interactions between, surrounding atoms. Cavities are defined to occur at positions where the local van der Waals interactions are relatively weak, and the width of a cavity is determined by growing a probe atom in the cavity until the average packing observed within protein structures is restored. The use of a growing spherical probe to locate cavities is a more satisfactory approach than previously proposed methods based on multiple application of surface-accessibility calculations with probes of different sizes. The cavity is defined by spherical probes that have similar properties to atoms, are simple to manipulate, and may be easily identified with possible sites for bound water or other small molecules. A similar approach has recently been used to define pores through and ion binding sites in membrane proteins (Smart et al., 1993).

The total volume of the spherical probes that fill the wide cavities (cavities that have a diameter of 1.9 Å or greater) within a protein typically constitutes about 1% of the total volume of the protein. The total volume of these probes increases (possibly linearly) with the number of atoms of the protein that are inaccessible to solvent. The total cavity volume of a protein of a given size shows no dependence on structural class. However, the cavity volume of proteins of similar size may differ substantially, possibly reflecting functional differences between the proteins (Rashin et al., 1986; Tilton et al., 1986).

Approximately 18% of the probes of radius >0.95 Å can be considered to be coincident with buried water molecules deter-

Fig. 10. Stereo plot of the extensive cavities in cytochrome P450cam. The surface of the cavities is determined by PRO_ACT and displayed using SURFNET (R. Laskowski, University College London).

mined by crystallography. The likelihood that the cavity associated with a particular probe is hydrated increases with probe radius and the number of atoms that would be able to make polar contact with a water molecule placed at the probe's center. Those water molecules that are observed to hydrate the interior cavities of protein molecules typically make 3, often more but rarely fewer, polar contacts. These contacts may all be with the protein molecule, though a large proportion of the buried waters make 2 or fewer polar contacts with protein and compensate for the loss of interaction energy by making polar contacts with other buried water molecules.

The ability of buried water molecules to make hydrogen bonds with any otherwise unsatisfied protein hydrogen bonding group results in their having many different structural roles, which have been described in many reports of protein structures. Often buried water molecules act as a bridge between secondary structural elements or as a splint applied to reinforce elements of distorted secondary structure (Thanki et al., 1990). For example, of the buried waters found in cytochrome P450cam (Poulos et al., 1987; Fig. 10), two bridge between the strands at the open end of β-hairpins, two tie a loop to a β-strand, five occur at ends of, or at kinks in, α-helices, and two mediate binding of the heme. The large number of buried water molecules identified in this study will allow us to carry out a statistical analysis of their many roles and determine the relative importance of each to protein structure and function.

It is interesting to consider the results of the study of the hydrated and nonhydrated cavities in terms of the energetics of cavity formation and hydration. Hydration of a protein cavity requires transfer of water from bulk solvent to the cavity. For cavities of a particular size and offering a particular number of hydrogen bonding partners, the fraction that are hydrated is related to the energy associated with transferring a water molecule to that particular environment. The likelihood of a cavity being hydrated increases with its size (Fig. 8B) and the number of hydrogen bonding groups that surround it (Fig. 8A). If we minimize the effects due to cavity size by considering only cavities described by probes of radius 1.1–1.4 Å, we can plot the number of hydrated and nonhydrated cavities with given numbers of potential polar contacts to a water placed at the center of the cavity (Fig. 11). This plot indicates a rapid increase in the fraction of cavities that are hydrated with increasing number of polar contacts. This increasing fraction reflects an increase in the stability of the hydrated state due to the increasing number of hydrogen bonds made by the buried water molecule. The change in the observed ratio of hydrated to empty cavities due to the addition of a single polar contact can be converted to an estimate of the average increase in stability associated with the addition of a single hydrogen bond (ΔF) using the Boltzmann relation, i.e., $\Delta F = -0.592 \ln(p_{i+1}/p_i)$ kcal/mol, where p_i is the ratio of hydrated to nonhydrated cavities with i polar contacts. Implicit in the above procedure is the assumption that the proportion of the actual buried waters that are experimentally detected is independent of the number of hydrogen bonds made by those waters.

Applying this procedure to the data in Figure 11 implies that increasing the number of polar contacts from 0 to 1 appears to stabilize the hydrated state by 0.7 kcal/mol, thereafter the second polar contact is on average worth 0.9 kcal/mol, the third 0.5 kcal/mol, and the fourth 0.4 kcal/mol. The relatively large apparent stabilization arising from the first 2 polar contacts may reflect the greater average strength of these bonds because they are more likely to be able to adopt good hydrogen bonding geometry, whereas it is unlikely that 3 or 4 hydrogen bonding partners will all be in optimal positions. Alternatively, it may reflect the greater difficulty of the experimental identification of a presumably less firmly located molecule or be a statistical aberration resulting from the smaller number of examples of buried water molecules with no or 1 hydrogen bond.

It has been observed that the B-factor of surface water molecules is reduced with increased number of hydrogen bonds to

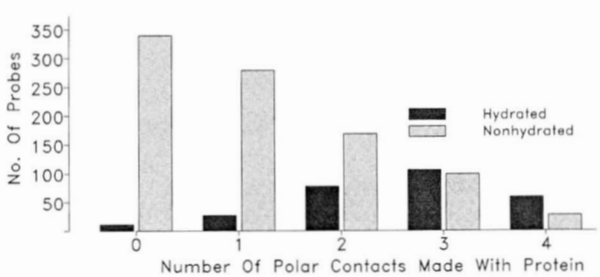

Fig. 11. Distribution of polar contacts that are made by water molecules placed at each of the sites of midsize probes (radius >1.1 Å and <1.4 Å). A maximum of 2 contacts with hydrogen bond donors and 2 with acceptors are counted in producing this distribution. The sites that are coincident with a crystallographically determined buried water are distinguished from those that are not.

the protein. Because an arbitrary upper limit is imposed on the *B*-factor of solvent molecules included in the PDB file, it is generally the case that surface water molecules making fewer and particularly those making no H-bonds are substantially underrepresented. We have examined the distribution of *B*-factors of the buried water molecules found in this study in relation to the distribution of *B*-factor limits applied to the deposited data. We see that the *B*-factors of the vast majority of buried waters making 2 or more hydrogen bonds are well below even the lowest *B*-factor limits imposed on the data. Consequently, the estimates of the observed number of such waters are not significantly affected by the use of such limits. However, buried waters making 0 or 1 hydrogen bond have higher average *B*-factors, and we estimate from the distributions that their reported numbers are 10–20% too low. If we take this under-reporting into account, our estimates for the energy of the first and second polar contacts are reduced by 0.05–0.1 kcal/mol.

The above results suggest that a buried water–protein hydrogen bond stabilizes a folded protein by, on average, 0.6 kcal/mol. This energy is similar to the values of 0.5–1.8 kcal/mol determined for hydrogen bonds between uncharged polar groups in proteins via mutagenesis experiments (Fersht et al., 1985) and suggests that water is rather effective at bridging between polar groups in the interior of proteins. Both of these values for hydrogen bond strengths are substantially smaller than the 4 kcal/mol per hydrogen bond determined by free energy perturbation calculations of water molecules in protein cavities (Wade et al., 1990).

Vinogradov (1980) observed that the number of hydrogen bonds made by water molecules to main-chain hydrogen bond acceptors is greater than that made to unambiguous donors and suggested that this implied that the water···O=C interaction was energetically more favorable than the water···H−N interaction. For buried water molecules, we find that the ratio of polar contacts in which the waters act as hydrogen donors (to protein main chain or side chain) to those in which they act as acceptors is 1.5:1. However, we do not consider that this offers support for Vinogradov's idea because computational water molecules placed at each nonhydrated probe site have a very similar donor:acceptor ratio of 1.6:1. Thus, we suggest that donor:acceptor hydrogen bond ratios of buried water molecules principally reflect the atomic composition of the protein and not the strength of the protein–water interaction.

The free energy associated with the process of hydrating an apolar cavity can be considered to have 2 principal components: the free energy of removing a water molecule from bulk water and the packing free energy resulting from van der Waals interactions of the water molecule with the protein. Transferring a water molecule to a polar cavity involves an additional favorable component due to hydrogen bonding. We observe that there is rapid increase in the fraction of midsize probe sites that are hydrated with increasing number of polar contacts. Because the small increase in stability associated with additional hydrogen bonds would have little effect on occupancy of the cavities if the overall transfer energy were very negative or very positive, this observation implies that the overall free energy associated with transferring a water from bulk to a midsize cavity must be near 0. Because the free energy of removing a water molecule from bulk has been determined experimentally to be +6.3 kcal/mol (Ben-Naim & Marcus, 1984) and the contribution due to the hydrogen bonds made by buried water molecules in polar cavities is estimated by us to be typically between −0.6 and −2.3 kcal/mol, in order that the overall free energy of transfer does not differ substantially from 0, it must be that the packing free energy contributes −4 to −7 kcal/mol.

This rough estimate of the packing energy is consistent with that from calculations of hydrating a water-sized cavity within a protein (−4.3 to −4.5 kcal/mol; Wade et al., 1990), but somewhat larger than experimental values for the packing energy of solid hydrocarbons (−2 kcal/mol per methylene group). This packing energy results from the interaction of 2 surfaces, that of water and that of protein. Assuming that the van der Waals energy of this interaction is the same as that of normal protein–protein interactions, we can estimate the destabilization of protein structure due to the formation of a midsize apolar cavity to be equal to the loss of half this packing energy, i.e., +2 to +3 kcal/mol, an estimate that is in accord with those inferred from the results of site-directed mutagenesis experiments (Eriksson et al., 1992; Lee, 1993).

The above observations concerning the hydration of cavities within proteins have implications for the interpretation of protein engineering experiments that create cavities within the protein structure. Such cavities are likely to be hydrated, a likelihood that increases with the size of the cavity and the number of potential hydrogen bonding partners surrounding the cavity. Attempts to relate experimental data on the relative unfolding free energy of cavity-forming mutants must consider this possibility in addition to the burial of hydrophobic surface and loss of van der Waals interactions. Indeed, it seems possible that if the cavity formed is sufficiently large and presents sufficient hydrogen bonding groups in a suitable orientation, it may, when hydrated, actually increase the stability of a folded protein.

Note added in proof

Since this paper was submitted, a methodologically distinct study of internal cavities has appeared whose results are in broad agreement with our own (Hubbard et al., 1994).

Acknowledgments

We thank Roman Laskowski for help with his SURFNET graphics software. This work was supported by Science and Engineering Research Council grants GR/H37051 and GR/H63678.

References

Baker EN, Hubbard RE. 1984. Hydrogen bonding in globular proteins. *Prog Biophys Mol Biol 44*:97–179.

Ben-Naim A, Marcus Y. 1984. Solvation thermodynamics of nonionic solutes. *J Chem Phys 81*:2016–2027.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JD, 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.

Connolly ML. 1985. Atomic size packing defects in proteins. *Int J Pept Protein Res 28*:360–363.

Creighton TE. 1993. *Proteins: Structures and molecular properties*. New York: W.H. Freeman and Company.

Edsall JT, McKenzie HA. 1983. Water and proteins. 2. The location and dynamics of water in protein systems and its relation to their stability and properties. *Adv Biophys 16*:53–183.

Eriksson AE, Baase WA, Zhang XJ, Heinz DW, Blaber M, Baldwin EP, Matthews BW. 1992. Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science 255*:178–183.

Fersht AR. 1985. *Enzyme structure and mechanism, 2nd ed.* New York: Freeman. pp 306–307.
Fersht AR, Shi JP, Knill-Jones J, Lowe DM, Wilkinson AJ, Blow DM, Brick P, Carter P, Waye MMY, Winter G. 1985. Hydrogen bonding and biological specificity analysed by protein engineering. *Nature 314*:235–238.
Finney JL. 1977. The organization and function of water in protein crystals. *Philos Trans R Soc Lond B 278*:3–32.
Hubbard SJ, Gross KH, Argos P. 1994. Intramolecular cavities in globular proteins. *Protein Eng 7*:613–626.
Janin J. 1979. Surface and inside volumes in globular proteins. *Nature 277*:491–492.
Kamphuis IG, Kalk KH, Swarte MBA, Drenth J. 1984. Structure of papain refined at 1.65 Å resolution. *J Mol Biol 179*:233–256.
Lauble H, Kennedy MC, Beinert H, Stout CD. 1992. Crystal structures of aconitase with isocitrate and nitroisocitrate bound. *Biochemistry 31*:2735–2748.
Lee B. 1993. Estimation of the maximum change in stability of globular proteins upon mutation of a hydrophobic residue to another of smaller size. *Protein Sci 4*:733–738.
Lee B, Richards FM. 1971. The interpretation of protein structures: Estimation of static accessibility. *J Mol Biol 55*:379–400.
Meyer E, Cole G, Radhakrishnan R, Epp O. 1988. Structure of native porcine pancreatic elastase at 1.65 Å resolution. *Acta Crystallogr B 44*:26–55.
Orengo CA, Flores TP, Taylor WR, Thornton JM. 1993. Identification and classification of protein fold families. *Protein Eng 6*:485–500.
Otting G, Liepinsh E, Wüthrich K. 1991. Proton exchange with internal water molecules in the protein BPTI in aqueous solution. *J Am Chem Soc 113*:4363–4364.
Poulos TL, Finzel BC, Howard AJ. 1987. High-resolution structure of cytochrome P450cam. *J Mol Biol 195*:687–700.
Rashin AA, Iofin M, Honig B. 1986. Internal cavities and buried waters in globular proteins. *Biochemistry 25*:3619–3625.
Richards FM. 1977. Areas, volumes, packing and protein structure. *Annu Rev Biophys Bioeng 6*:151–176.
Sandberg WS, Terwilliger TC. 1989. Influence of interior packing and hydrophobicity on the stability of a protein. *Science 245*:54–57.
Schoenborn BP. 1965. Binding of xenon to horse haemoglobin. *Nature 208*:760–762.
Screenivasan U, Axelsen PH. 1992. Buried water in homologous serine proteases. *Biochemistry 31*:12785–12791.
Smart OS, Goodfellow JM, Wallace BA. 1993. The pore dimensions of gramicidin A. *Biophys J 65*:2455–2460.
Thanki N, Goodfellow JM, Thornton JM. 1988. Distribution of water around amino acid residues in proteins. *J Mol Biol 202*:636–657.
Thanki N, Thornton JM, Goodfellow JM. 1990. Influence of secondary structure on the hydration of serine, threonine and tyrosine residues in proteins. *Protein Eng 3*:495–508.
Tilton RF, Kuntz ID, Petsko GA. 1984. Cavities in proteins: Structure of a metmyoglobin-xenon complex solved to 1.9 Å. *Biochemistry 23*:2849–2857.
Tilton RF, Singh UC, Weiner SJ, Connolly ML, Kuntz ID, Kollman PA, Max N, Case DA. 1986. Computational studies of the interaction of myoglobin and xenon. *J Mol Biol 192*:443–456.
Vinogradov SN. 1980. Structural aspects of hydrogen bonding in amino acids, peptides, proteins and model systems. In: Ratajczak H, Orville-Thomas WJ, eds. *Molecular interactions, vol 2*. New York: John Wiley and Sons. pp 179–229.
Vrielink A, Lloyd LF, Blow DM. 1991. Crystal structure of cholesterol oxidase from *Brevibacterium sterolicum* refined at 1.8 Å resolution. *J Mol Biol 219*:533–554.
Wade RC, Mazor MH, McCammon JA, Quiocho FA. 1990. Hydration of cavities in proteins: A molecular dynamics approach. *J Am Chem Soc 112*:7057–7059.
Walshaw J, Goodfellow JM. 1993. Distribution of solvent molecules around apolar side-chains in protein crystals. *J Mol Biol 231*:392–414.
Wright CS. 1990. 2.2 Ångstroms resolution structure analysis of two refined *n*-acetylneuraminyl-lactose-wheat germ agglutinin isolectin complexes. *J Mol Biol 215*:635–651.