Herskovits, T. T., Erhunmwunsee, L. J., San George, R. C., & Herp, A. (1981a) *Biochim. Biophys. Acta 667*, 44–58.

Herskovits, T. T., San George, R. C., & Erhunmwunsee, L. J. (1981b) *Biochemistry 20*, 2580–2587.

Herskovits, T. T., Jacobs, R., & Nag, K. (1983) *Biochim. Biophys. Acta 742*, 142–154.

Herskovits, T. T., Russell, M. W., & Carberry, S. E. (1984) *Biochemistry 23*, 1873–1881.

Herskovits, T. T., Carberry, S. E., & Villanueva, G. B. (1985a) *Biochim. Biophys. Acta 828*, 278–289.

Herskovits, T. T., Mazzella, L. J., & Villanueva, G. B. (1985b) *Biochemistry 24*, 3862–3870.

Kirkwood, J. G., & Auer, P. (1951) *J. Chem. Phys. 19*, 281–283.

Lips, D., Gielens, C., Preaux, G., & Lontie, R. (1982) *Arch. Int. Physiol. Biochim. 90*, B128.

Manwell, C. (1958) *J. Cell. Comp. Physiol. 52*, 341–352.

Manwell, C. (1960) *Arch. Biochem. Biophys. 89*, 194–201.

Miller, K. I., & van Holde, K. E. (1982) *Comp. Biochem. Physiol., B: Comp. Biochem. 73B*, 1013–1018.

Puett, D. (1973) *J. Biol. Chem. 248*, 4623–4633.

Redmond, J. R. (1962) *Physiol. Zool. 35*, 304–313.

Ryan, M., Terwilliger, N. B., Terwilliger, R. C., & Schabtach, E. (1985) *Comp. Biochem. Physiol., B: Comp. Biochem. 80B*, 647–656.

Salvato, B., Ghiretti-Magaldi, A., & Ghiretti, F. (1979) *Biochemistry 18*, 2731–2736.

Siezen, R. J., & van Driel, R. (1974) *J. Mol. Biol. 90*, 91–102.

Siezen, R. J., & Van Bruggen, E. F. J. (1974) *J. Mol. Biol. 90*, 77–89.

Svedberg, T., & Hedenius, A. (1934) *Biol. Bull. (Woods Hole, Mass.) 66*, 191–223.

Svedberg, T., & Pedersen, K. O. (1940) *The Ultracentrifuge*, Oxford University Press, Oxford, England.

Tanford, C. (1961) *Physical Chemistry of Macromolecules*, Chapters 4, 6, and 8, Wiley, New York.

Tanford, C. (1970) *Adv. Protein Chem. 24*, 1–95.

Tomimatsu, Y., & Palmer, K. J. (1963) *J. Phys. Chem. 67*, 1720–1722.

van Holde, K. E., & Cohen, L. B. (1964) *Biochemistry 3*, 1803–1808.

van Holde, K. E., & Miller, K. I. (1982) *Q. Rev. Biophys. 15*, 1–129.

Wood, E. G., & Peacocke, A. R. (1973) *Eur. J. Biochem. 35*, 410–420.

Internal Cavities and Buried Waters in Globular Proteins[†]

Alexander A. Rashin,[‡] Michael Iofin, and Barry Honig*

Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York 10032

Received August 30, 1985; Revised Manuscript Received January 27, 1986

ABSTRACT: A fast algorithm that detects internal cavities in proteins and predicts the positions of buried water molecules is described. The cavities are characterized in terms of volume, surface area, polarity, and the presence of bound waters. The algorithm is applied to 12 proteins whose structures are known to high resolution and successfully predicts the locations of over 80% of internal water molecules. Most proteins are found to have a number of internal cavities ranging in volume from 10 to 180 Å^3. Some of these cavities contain water and some do not, with the probability of containing a buried water increasing with cavity size. However, many large cavities are found to be empty (i.e., they do not contain a crystallographically determined water). For multidomain proteins over half of the total cavity volume is at the interdomain interface. Possible implications for the energetics of cavity formation and for the functional role of internal cavities are discussed.

It has been shown in a number of studies that the packing density in globular proteins is similar to that found in crystals of small organic molecules (Richards, 1974; Finney, 1975; Schultz & Schirmer, 1979). Since a random sequence of amino acids would not be expected to generate a tightly packed structure, close packing has been suggested as a criterion to be used in protein folding studies (Schulz & Schirmer, 1979; Rashin, 1980). On the other hand, internal cavities large enough to accommodate xenon atoms have been found to exist in proteins (Schoenborn, 1969; Richards, 1974; Tilton et al., 1984). Internal cavities have been associated with conformational flexibility (Lumry & Rosenberg, 1975), the mechanism of hydrogen exchange (Richards, 1979), and, more recently, the existence of multiple side chain conformations of a single protein (Smith et al., 1986). Despite their apparent importance there is little information available as to the detailed properties of cavities in proteins.

It would be of interest, for example, to determine the size of cavities that are permissible in a particular structure and to have a means of determining whether a particular cavity is likely to be filled with water molecules. These questions are likely to assume considerable importance in prediction algorithms for protein conformation. The type of problem that can be encountered in model building has been illuminated by the study of Novotny et al. (1984), who succeeded in generating apparently reasonable but nevertheless incorrectly folded structure for two proteins (for example, by using the C$^\alpha$ coordinates of hemerythrin, which is α-helical, to generate a hemerythrin-like structure with a sequence of an immunoglobulin domain, which is known to be primarily β-sheet). An analysis of cavity size in the incorrectly folded structures might, for example, show them to be unrealistic and thus provide criteria for their evaluation.

[†] This work has been supported by grants from the NIH (GM30518) and NSF (PCM82-07145 and DMB85-03484).

[‡] Present address: Department of Physiology and Biophysics, Mount Sinai School of Medicine, New York, New York 10029.

0006-2960/86/0425-3619$01.50/0

Another area where the presence of cavities is likely to be important concerns the assignment of specific atoms, or residues, as being on the surface or inside of proteins [see, e.g., Chothia (1976) and Rashin and Honig (1984)]. Without a clear definition of cavities, a group deep inside of a protein might be assigned to the external surface on the basis of standard accessibility criteria (Lee & Richards, 1971), even though it was only adjacent to an internal cavity.

Finally, it is possible that in certain proteins a predesigned cavity might play an important functional role. One example where this question arises concerns visual pigments and bacteriorhodopsin whose retinal chromophores exhibit subpicosecond isomerization times [see, e.g., Dinur et al. (1981) and Doukas et al. (1984)]. How a cis–trans isomerization within a protein can occur so rapidly is unclear; conceivably, the presence of an internal cavity facilitates the motion. In order to evaluate this type of hypothesis, it is necessary to determine the size of a cavity that a particular protein can tolerate.

The first computational study of internal cavities in proteins was included as part of the pioneering surface accessibility paper of Lee and Richards (1971). Lysozyme, ribonuclease, and myoglobin were considered in that work. A recent paper by Tilton et al. (1984) reports a study of cavities in metmyoglobin–xenon complexes using an algorithm developed by Connolly. The goal of this paper is to report the results of a survey of cavities and to infer characteristic structural patterns that may be of general applicability.

Methods

Surface Area Algorithm. Our method for location of cavities and of possible sites of water binding is based on the modified algorithm of Shrake and Rupley (1973), (Rashin, 1984). The algorithm represents the surface of a solvated atom by a uniform distribution of points on a sphere of radius $r_v + r_p$, where r_v is the van der Waals radius of the atom and r_p is a probe radius. Specifically, the van der Waals radii used are as follows: tetrahedral carbon, tetrahedral nitrogen, or sulfur with hydrogens attached, 2.0 Å; trigonal carbon and trigonal NH, 1.7 Å; trigonal CH, CH_2, and sulfur, 1.85 Å; hydroxyl and trigonal nitrogen, 1.5 Å; trigonal NH_2, 1.8 Å; oxygen and water, 1.4 Å (Rashin, 1984). The points on the surface of a solvated atom that are inside the solvated spheres of other atoms are treated as being buried, with the remainder being accessible to the probe. We represent the spherical surface of each atom by a set of 500 points almost uniformly distributed on the surface (an exactly uniform distribution is impossible because regular polyhedra with high numbers of vertices do not exist). We have found that no significant improvement is obtained by using a larger number of points. A further increase in the number of points does in some cases detect a few additional small cavities or channels to the surface which lead to the disappearance of small cavities. However, the total cavity volume and the number of cavities remain constant to within a few percent.

Surfaces as Sets of Connected Points. Two accessible points on the same sphere are considered as being connected by an edge if the distance between them is less than 1.5 times the smallest distance between any two points on the sphere. Two accessible points are regarded as connected if there exists a sequence of edges connecting them. A set of accessible points connected to one another constitutes a connected surface. For example, all the points on the surface of an isolated sphere are connected. However, only accessible points form surfaces. Therefore, if some of the points turn out to be buried due to overlaps with neighboring atoms, these points and the edges connected to them do not contribute to the connectivity. As a result, in some cases there may be accessible points on a single sphere with no sequence of edges connecting them. Thus there will be two or more connected surfaces on the sphere that are not connected to one another.

Connectivity Algorithm. In order to determine the number of connected surfaces on each sphere and to assign accessible points to each of these surfaces, the following algorithm is used. An accessible point is chosen and used to initiate a list. All accessible points on the sphere that are connected to it by an edge are then added to the list. This procedure is repeated for each of these newly added points. (Only points not yet in the list are added.) A description of a connected surface is complete when no new accessible points connected by an edge to accessible points already in the list can be found. Additional connected surfaces are defined by repeating the procedure, beginning with another point on the sphere not included in the previous lists. The process is completed when all accessible points are included. During the process every accessible point is assigned the index of the surface to which it belongs. This routine is applied to the surface of each accessible atom.

Now, two points belonging to accessible surfaces of two different atoms can be at a distance of less than the connectivity distance described above, and therefore these points are connected by an edge. If there exists one such edge, all the points of the two surfaces are connected and form a larger connected surface. Thus the problem of connecting two surfaces is reduced to the search for one connectivity distance (edge) between two points belonging to different surfaces. To accelerate the search, neighbor lists for each atom are created in the course of accessibility calculations. If each connected surface on an individual atom is considered as a generalized "point" and connectivity between neighboring surfaces as an "edge", then the algorithm for the connectivity of different atomic surfaces is essentially identical with that for the connectivity of the points on the surface of a single atom. The largest connected surface obtained is the outer surface of the protein, and all other connected surfaces belong to cavities.

Prediction of Water Positions in Cavities. The following algorithm is used to predict internal waters. For each accessible point in a cavity a line is constructed through this point and the center of the atom to which it belongs. The center of a test water molecule is placed on this line at a distance of the sum of the radii of the atom and the water molecule from the atomic center. All atoms with which this test water molecule can form hydrogen bonds are then found, and angles between the possible hydrogen bonds formed by the water molecule are calculated. A tentative position for a water molecule is defined if the hydrogen bonds it forms satisfy assigned criteria for acceptable bond lengths and bond angles. The range of the allowable bond lengths was set between 2.2 and 3.5 Å; for bond angles, a "soft" range was defined as between 60° and 150°, and a "hard" range was defined as between 80° and 130°. These somewhat arbitrary criteria are intended to allow for the inaccuracy of our method, for errors in crystallographic coordinates, and for deviations of hydrogen bonding angles from the expected tetrahedral angle of about 110° (Finney, 1979; Edsall & McKenzie, 1983).

For each tentative position of a water molecule, this procedure generates a matrix with dimensions of the number of polar atoms within the allowed hydrogen bonding distance. A matrix element is assigned a value of unity for any pair of potentially hydrogen bonding atoms if the angle between the lines connecting their centers with the center of the water oxygen is within the allowed limits. Otherwise, an element is assigned a value of zero. The number of hydrogen bonds that can be simultaneously formed is then equal to the dimension of the largest submatrix that contains only nonzero off-diagonal elements. A simple algorithm finds this submatrix. As a probe radius smaller than that of water is often used in defining the cavity surfaces (in order to account for the inaccuracy of the point representation), the full-sized water molecules used in the present algorithm are tested for severe overlaps with protein atoms. Those with overlaps of more than 0.25 Å are rejected.

Buried water molecules in proteins almost invariably form hydrogen bonds with polar atoms of the protein (Finney, 1979; Edsall & McKenzie, 1983). It is assumed that a water molecule can be buried within a protein only when it can simultaneously form at least three hydrogen bonds with protein atoms to compensate for the loss of hydrogen bonds with bulk water. Since the cavity surface often has many points at distances of a few tenths of an angstrom from one another, overlapping water molecules are often predicted. Two predicted water molecules are regarded as overlapping if the distance between their centers is less than 2.2 Å. To select a nonoverlapping set in each cavity, two criteria are used. For two overlapping water molecules the one that conforms to the hard tetrahedrality criterion is selected. If both conform to the same tetrahedrality criterion, the one with a lower value for the sum of the lengths of its three hydrogen bonds is selected. Predicted waters with the value of this sum below 8.4 Å are rejected to eliminate the geometrically improbable case where water forms simultaneous hydrogen bonds with NH and CO groups of the same peptide unit.

Comparison with Experimentally Observed Waters. Predicted waters are compared to the crystallographic positions of bound waters. To check whether all experimentally determined internal waters have been identified, these have to be selected in a consistent way from the entire list of waters in the crystallographic data set. It is impossible to determine whether an experimentally observed water is internal by accessibility measurements alone because waters in large cavities may appear accessible. To avoid this problem, protein atoms in close proximity to an experimentally observed water are compared with lists of atoms that form different connected surfaces. A water is assigned to the cavity (or the surface) that contains the maximum number of the water's neighbors. Waters that cannot be associated with any calculated surface are also identified. These are generally found to be crystallographic waters bound in the second solvation shell, or in rare cases, they correspond to waters in very small cavities that are not detectable with a particular probe radius.

Cavity Size and Probe Radius. Cavity size is characterized by accessible area and volume. However, neither of these properties describes the shape of a cavity nor predicts whether it can accommodate a probe of a given size. The maximum radius of the probe a given cavity can contain is determined by increasing the probe radius in increments of 0.2 Å until the cavity disappears. The final probe radius is a measure of the cavity's shape. Due to the assumption of spherical atoms, each cavity becomes connected to the outer protein surface at some small probe radius. The radius at which this occurs is taken as the radius of the channel connecting the cavity to the protein surface. Cavities found at one probe radius are identified with cavities found at another probe radius by counting numbers of identical atoms forming their surfaces. Due to the point representation of the surface and to the occurrence of short hydrogen bonds, cavities containing water are better identified with a probe radius of 1.2 Å rather than 1.4 Å, and therefore results using the former value are reported. We list cavities located with a probe radius of 1.2 Å in terms of the atoms forming them, their volumes and areas, limiting probe radii at which they disappear, and waters they contain.

Volume Calculations. To calculate volumes, a rectangular box surrounding each cavity is defined and is filled with a fine cubic grid of points. In order to identify those grid points that are inside the cavity, each grid point inside of a probe sphere centered at each accessible point is marked. The cavity volume is then calculated by counting the number of marked points. The smallest cavity detectable corresponds to the case of a single accessible point, and thus the minimum cavity volume is that of the probe sphere centered at this accessible point. A rapid algorithm was devised that avoids the time-consuming calculation of the distances between grid points and centers of probe spheres. For very small probe radii and larger cavities the probe spheres form a shell near the accessible points of the surface of the cavity while the center of the cavity is left empty. This will underestimate the cavity volume. However, because no cavity we found could be detected with a probe radius of >2 Å, we concluded that all cavities are completely filled if probes of radius >1 Å are used. Our method outlined here is quite similar to that of Pavlov and Fedorov (1983), which in turn yields volumes identical to within a few percent with those produced by the Connolly algorithm (Connolly, 1985).

Results and Discussion

General Description of Cavities. Table I shows that all but one of the dozen well-refined proteins for which coordinates were available from the 1983 edition of the Protein Data Bank (Bernstein et al., 1977) contain cavities. The total cavity volume of a given protein varies from 0 to about 600 $Å^3$ and with a few deviations shows a tendency to increase with molecular weight and to constitute less than 2% of the total protein volume (Pavlov & Fedorov, 1983; Connolly, 1985). It is found that no single cavity can host a probe of radius 2 Å or more and that all cavities become connected to the protein surface by channels of 0.4-Å radius [see also Tilton et al. (1984)]. A comparison of the volume of the largest probe a cavity can accommodate to cavity volume reveals that many cavities and almost all large ones are highly nonspherical.

Most of the cavities are found to contain water although some are empty in the sense that they do not contain any crystallographically determined waters [it should be noted, however, that most crystallographers do not identify water molecules with occupancies below 0.3 (Finney, 1979)]. Total cavity volume, in particular that of empty cavities, does not exhibit any simple dependence on molecular weight. The total accessible area of protein cavities is negligible (less than a few tenths of 1% of the total accessible or buried area of a corresponding protein) and thus is unlikely to be an important factor in protein stability (Rashin, 1984).

The volume of a particular cavity (or its existence, for that matter) is a sensitive function both of the probe size and of the radii assigned to the individual atoms in the protein. The radii used in this work (see above) were taken from a recent version of the Lee and Richards (1971) algorithm (as provided by J. Matthew, private communication). These radii are almost identical with those given by Pauling (1960), which were adopted by Shrake and Rupley (1973). Using somewhat larger atomic radii, Tilton et al. (1984) obtain an internal cavity volume for myoglobin ranging between 489 and 141 $Å^3$, depending on probe radius. For the probe radius of 1.2 Å used in this work, Tilton and co-workers find a volume of 241 $Å^3$

Table I: Summary of Cavities and Internal Waters in 12 Proteins

proteins[a]	M_r	no. of cavities[b]	total vol[b] (Å^3)	total area[b] (Å^2)	no. of internal waters		mean deviation (Å)
					exptl	predicted[c]	
4PTI	6000	2 (0)	70 (0)	20 (0)	4	4/4	0.37
1NXB	7000	0	0	0	0	0	
1SN3	7000	2 (1)	32 (14)	6 (3)	1	1/1	0.24
3CYT	11000	5 (3)	34 (21)	2 (1)	2	2/2	0.60
1RN3	13000	5 (3)	41 (27)	3 (1)	1	2/1	0.78
2LYZ	14000	8 (4)	190 (94)	53 (26)	5	5/5	0.82
1ECA	14000	9 (9)	401 (401)	69 (69)	0	0	
2MBN	17000	23 (22)	391 (326)	85 (60)	1	5/1	0.72
2ACT	23000	21 (11)	449 (127)	130 (18)	15	15/12	0.56
2PTN	23000	13 (4)	494 (40)	168 (5)	17	19/17	0.57
2CHA	25000	26 (18)	571 (247)	120 (50)	10	11/9	0.89
3TLN	34000	30 (17)	528 (188)	117 (27)	16	20/16	0.79

[a] The protein abbreviations are from the Protein Data Bank (Bernstein et al., 1977): 4PTI, pancreatic trypsin inhibitor; 1NXB, neurotoxin B; 1SN3, scorpion toxin; 3CYT, cytochrome *c*; 1RN3, pancreatic ribonuclease; 2LYZ, egg lysozyme; 1ECA, erythrocruorin; 2MBN, myoglobin; 2ACT, actinidin; 2PTN, pancreatic trypsin; 2CHA, α-chymotrypsin; 3TLN, thermolysin. [b] Values in parentheses correspond to cavities that do not contain crystallographically determined waters. [c] The value after the slash indicates the number of predicted waters that correspond to crystallographically determined waters.

as compared to the value of 391 Å^3 given in Table II. However, since their radii are larger than ours by about 0.1 Å, a more appropriate comparison should be based on the volumes they obtain by using a probe radius that is smaller than ours by 0.1 Å, i.e., 1.1 Å. In this case, the calculated cavity volumes for myoglobin agree quite well: 369 Å^3 in Tilton et al. (1984) as compared to the 391-Å^3 value calculated in this work.

The great sensitivity of cavity volume to atomic and probe radii would appear to raise serious uncertainties as to the validity of the analysis presented in this work. However, the detection of internal bound waters provides an important test of both the volume algorithms and the combination of atomic and probe radii that are used. As shown in the next section, we successfully predict the location of most bound waters. In contrast, we have found that many of the smaller water-filled cavities are missed if a 1.2-Å^3 probe and the atomic radii given in Tilton et al. (1984) are used [whereas the radii of Tilton et al. (1984) and a 1.1-Å probe yield results similar to those presented here]. It thus appears that the combination of probe and atomic radii used in this work is a reasonable one and that the total internal cavity volume cannot be much smaller than we have calculated.

Another question concerning the analysis of cavity volumes is the influence of the extent of refinement of a particular structure on the quality of the results. However, we found few differences between thermolysin at 2.3-Å resolution refined with Diamond's real space procedure (Matthews et al., 1974) and the 1.6-Å structure refined with the restricted least-squares method (Holmes & Matthews, 1982). There was an approximately 10% decrease in cavity volume in the better refined structure as well as a small decrease in the number of cavities. Since we have restricted our analysis to refined structures, it does not appear that coordinate inaccuracies are affecting our general conclusions.

Bound Waters. As is evident from the last three columns of Table I, at least 80% of the experimentally observed internal water molecules are successfully predicted with an average accuracy of about 0.8 Å. However, for some proteins a few water molecules that are not experimentally observed are predicted. Further details about the volumes of individual cavities and the experimentally observed water molecules they contain are given in Table II. It appears that for larger cavities the volume per internal water molecule is often near 30 Å^3, the value of the partial specific volume of water in the bulk phase. In smaller cavities, the volume per water molecule is close to the van der Waals volume of a single water molecule of 11.5 Å^3. It is of interest that some cavities which appear to be completely nonpolar can contain a water molecule forming three hydrogen bonds. This is due to the fact that a hydrogen bond can be longer than the sum of the van der Waals radii of a polar atom and a probe, and thus some polar atoms with which water forms hydrogen bonds may not contribute to the accessible surface.

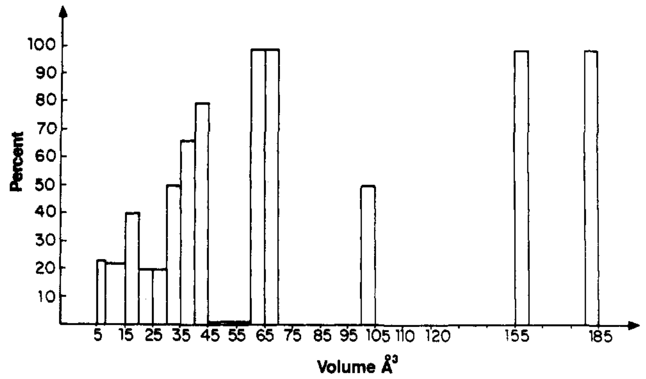

FIGURE 1: Percentage of water-filled cavities as a function of cavity volume.

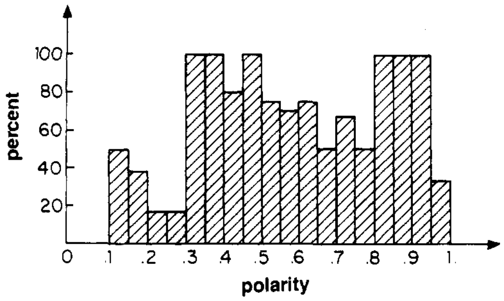

FIGURE 2: Percentage of water-filled cavities as a function of polarity. Polarity is defined as the ratio of the accessible area of all nitrogens and oxygens in a cavity (except for nitrogens in proline and in tryptophan side chain) to the total accessible area of the cavity.

Figure 1 demonstrates another feature hidden in Table II, i.e., that the percentage of cavities filled with water (according to crystallographic data) increases with cavity size. This, for example, would result if larger cavities were more polar or if large empty cavities are energetically more unfavorable than small ones. Our results suggest that the first factor is probably not the most important one. Figure 2 shows that, for cavities with polarities above 0.3 (polarity is defined in the figure caption), the percentage of water-filled cavities is essentially constant. However, about half of the cavities have polarities

Table II: Number of Cavities and Internal Waters as a Function of Cavity Volume in 12 Proteins[a]

	vol (Å³)																			
protein	8	15	20	25	30	35	40	45	50	55	60	65	70	75	100	105	155	160	180	185
4PTI	1											1								
	1											3								
1NXB	0																			
	0																			
1SN3	0	1	1																	
	0	0	1																	
3CYT	5																			
	2																			
1RN3	4	1																		
	1	1																		
2LYZ	2	2	2								1	1								
	2	0	1								0	1								
1ECA	5	1	3		1	1		1	1	1						1				
	0	0	0		0	0		0	0	0						0				
2MBN	6	9	2	2	1	1			1				1							
	0	0	0	0	0	0			0				1							
2ACT	7	5	5	2								1						1		
	1	2	4	0								3						5		
2PTN	4	2	2			1	1					1				1				1
	3	0	1			1	1					2				3				6
2CHA	12	3	1	1	1		2	3		1		1	1							
	0	1	0	1	0		1	3		0		2	2							
3TLN	7	13	4		2	1		1					1	1						
	2	4	1		1	1		1					1	3						
total	53	37	20	5	5	4	3	5	2	2	1	5	3	1		2		1		1
filled	12	8	8	1	1	2	2	4	0	0	0	5	3	1		1		1		1

[a] The first row for each protein indicates the number of cavities within the volume interval defined by the number at the top of each column and the top of the preceding column. For example, column 3 contains the number of cavities of volume between 8 and 15 Å³. The second row indicates the number of the crystallographically determined waters in the cavities in a given volume interval.

Table III: Description of Two Largest Cavities and the Waters They Contain[a]

				waters				
				X-ray		predicted		
protein	vol (Å³)	polarity	atoms[b]	no.[c]	H bonds[b]	no.[c]	H bonds[b]	deviation[d]
2PTN	180	0.56	G23O; N25N, CA, OD1; T26N; V27N, O; P28CA, C, O; Y29CA, C; Q30O, CB, CG; V31CA, CG2; S32N; L46CB, CD2; R66O; L67CA, CB, CD2; G69N, CA; E70N, O, CB; D71CA, OD1; R117O; V118CG1; W141CH2; L155CD1	604	N25N; E70O; D71OD1; HOH717	1	N25N; E70O; D71OD1	0.12
				708	Q30O; R66O; G69N; E70N; HOH709	2	Q30O; R66O; E70N	0.35
				715	P28O; Q30O; G69N	3	P28O; Q30O; G69N	0.61
				516	N25OD1; R117O; HOH709; HOH717	4*	P28O; G69N; R117O	1.50
				709	V27O; Q30O; Q70O, HOH516; HOH708, HOH717	5*	N25OD1; E70O; HOH1	1.98
						6*	Q30O; HOH5; HOH15	0.81
				717	N25N; T26N; V27N; V27O; HOH516; HOH604; HOH709	7*	G23O, V27N; T26N	1.42
1ECA	103	0.0	I23CG2, CD1; L24CA, CD1; V27CG2; I62CG2; V63CA, CG1, CG2; F66CB, CG, CD2, CE2; F100CE1, CE2, CZ; F104CD2, CE2	none		none		

[a] The largest-water filled and largest empty cavity detected in our sample. These can contain probes with radii less than 2 and 1.8 Å, respectively. [b] Protein atoms and crystallographically determined waters are listed in the format of the Protein Data Bank. The one-letter code is used for amino acids. Thus, for example, G23O denotes the carbonyl oxygen of Gly-23. [c] Arbitrary numbers are assigned to predicted waters. The predicted waters are listed next to the closest crystallographically determined waters. An asterisk denotes waters that were predicted only when waters 1–3 were included as protein atoms. [d] Distance in Å between the oxygens of the predicted and experimental waters.

above 0.3 whether they are large or small (where a 30-Å³ volume is used to differentiate large and small cavities). Since the percentage of the water-filled cavities is about twice as large for cavities with volumes above 30 Å³ (Figure 1), a difference in polarity cannot explain this large effect.

Table III contains a description of the largest water-filled and empty cavities. (A detailed description of any of the other cavities is available upon request.) In agreement with the experimental observation, the largest empty cavity in erythrocruorin (as well as in other globins) is predicted to contain no waters. Some successes as well as possible sources of failure of the water prediction algorithm are illustrated with the example of the trypsin cavity. Three experimentally determined water molecules (604, 708, 716), each bound to at least three polar protein atoms, are accurately predicted. In contrast, the presence of waters 516, 709, and 717 is predicted only after the algorithm is applied with the first three waters included as protein atoms. Even in this case, the coordinates of the second group are not as accurately predicted as those of the first. The complication associated with the second group

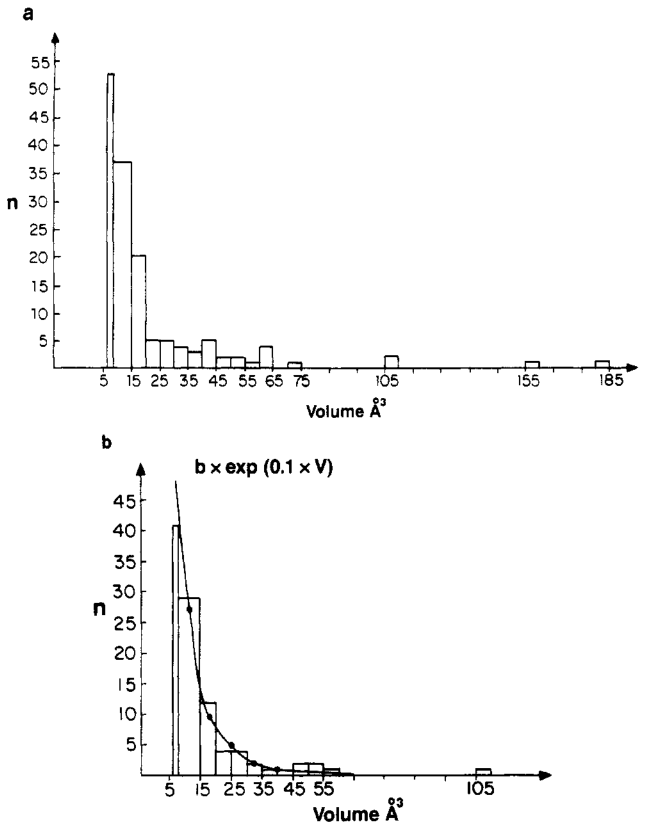

FIGURE 3: Number of cavities as a function of their volume: (a) all cavities; (b) empty cavities.

results from the fact that they form hydrogen bonds with one another, rather than primarily to protein atoms. In general, the largest deviations from experimentally observed locations are for waters that appear to bind cooperatively as a cluster. However, the overall success of our predictions (see Table I) suggests that most internal waters in proteins bind noncooperatively and that the factors that determine their binding have been defined in our algorithm.

Energetics of Cavity Formation. An estimate for the free energy required for the formation of a spherical cavity can be obtained from scaled particle theory (Reiss, 1966), which assumes a liquid consisting of hard spheres. The energy of cavity formation is found to increase linearly with cavity surface for cavities larger than the solvent molecule and to become nonlinear for smaller sizes. The free energies of formation of a methane-sized cavity in cyclohexane, benzene, or carbon tetrachloride calculated with scaled particle methods are 3.8–4.1 kcal/mol (Lee, 1985a,b). However, proteins are not hard-sphere liquids, and cavities are often nonspherical and smaller than amino acid side chains. Thus it is of interest to find a more direct estimate for the free energy of cavity formation in proteins.

Figure 3 plots the distribution of cavity size for all cavities (Figure 3a) and for empty cavities (Figure 3b). While Figure 3a indicates a general tendency toward a decrease in the number of cavities with increasing size, this decrease does not follow any simple function. The distribution of empty cavities in Figure 3b is, in contrast, well described by a single exponential function $b \exp(0.1V)$, where b is a constant and V is a cavity volume. If it is assumed that this function corresponds to a Boltzmann distribution, the cost of forming an empty cavity is found to be 60 cal/(mol·Å^3), which corresponds to only slightly more than kT for an empty cavity of the size of a water molecule (11.5 Å^3). Given the small sample size and inaccuracy in the method, this number can at best be viewed as highly approximate. However, it does compare rather well with the values obtained with scaled particle theory. These values are of about a factor of 2 larger than would be predicted on the basis of our analysis, but given the approximations, the agreement is not bad. Regardless of which of the two numbers is used, they both suggest that an energetic cost of forming a cavity the size of water molecule is on the order of $(1\text{–}2)kT$.

Assuming a value of kT per 10 Å^3, the energetic cost of the large unfilled cavity in erythrocruorin (Table III) is about 6 kcal/mol while the cost of total unfilled volume in erythrocruorin exceeds 24 kcal/mol. This number for the entire protein is probably not very meaningful and is subject to a number of uncertainties in addition to those resulting from inaccuracies in the coordinates. First, it is conceivable that the cavities are not really empty and that they contain disordered waters that cannot be detected crystallographically. Second, the energetics of forming asymmetric cavities may be very different than for the formation of symmetric spherical cavities. At the present time it is impossible to approach these issues in more than a qualitative way. Nevertheless, our results show that the existence of fairly large empty cavities in proteins is not uncommon and thus the energy required to form them, while significant, is not unacceptably large for a single cavity.

Functional Significance. It is of interest that the largest empty cavity in our sample (Table III) is located between helices B, E, and G, i.e., between the two domains of the globin molecule [as identified in Rashin (1981)]. Empty cavities at approximately the same location and ranging in size between 50 and 180 Å^3 are also present between domains in all six myoglobin-like structures in the Protein Data Bank (myoglobin, erythrocruorin, lamprey hemoglobin, leghemoglobin, and two subunits of human hemoglobin). In some cases these cavities extend to the heme, suggesting that the heme might in some way play a role in their formation. This might account for the appearance of interdomain motions in hemoglobin which are absent in apohemoglobin [Sassaroli et al., 1982; Oton et al., 1981). The presence of this empty cavity in all globins suggests that it may serve some purpose, for example, facilitating the relative motions of helices or domains. It is interesting in this regard that the structural differences between the deoxy and carbonmonoxy forms of human hemoglobin involve the backbone motions of helices A and E relative to helices G and H (Baldwin & Chothia, 1979).

Cavities occur between domains in a number of other cases although these cavities are not always empty. Over $^2/_3$ of the total cavity volume in trypsin is between its two domains, a pattern that is repeated in chymotrypsin. In thermolysin only about 40 Å^3 of cavities is found in the stable C-terminal domain 235–316 (Vita et al., 1984) with the rest of 490 Å^3 of total cavity volume being either at its interface with the rest of the structure or within the domain known to undergo a significant conformational change upon ion binding (Titani et al., 1972; Vita et al., 1979). Table I shows that the ratio of the total cavity volume to molecular weight is much smaller for single-domain proteins (4PTI, 1NXB, 1SN3, 2CTY, 1RN3) than for multidomain proteins [see, e.g., Rashin (1981) for location of domains]. For example, hen lysozyme which possesses domain structure has more than 4 times the cavity volume of the single-domain ribonuclease having approximately the same molecular weight. These observations suggest that cavities may serve to facilitate interdomain or other motions.

For the case of the retinal-containing pigments mentioned above, consideration of models suggests that an empty cavity with a volume of approximately 100 Å^3 would allow a cis–trans isomerization of the chromophore to take place without a significant displacement of atoms in the protein. This may account, in part, for the fact that isomerization in the protein is both more efficient and faster than it is in solution, where large cavities would not be expected to exist.

SUMMARY

We have described an algorithm that detects internal cavities in proteins and predicts the locations of buried water molecules. The application of the algorithm to 12 proteins leads to the following conclusions. (1) Proteins may contain a significant number of cavities of sufficient size to hold one or more water molecules. The cavities range in volume from about 10 to 180 Å^3 and, together, may account for approximately 2% of the total volume of a given protein. A predicted structure, obtained in a protein folding scheme, is probably unreasonable if its cavity volume exceeds these limits. (2) The cavities may or may not contain one or more crystallographically determined water molecule(s). The tendency to contain internal waters increases somewhat with cavity size, but large empty cavities may also be present. (3) Internal waters tend to be located in positions where they can form at least three hydrogen bonds. Most buried waters hydrogen bond primarily to polar atoms of the protein, but some form clusters and hydrogen bond to one another. (4) As a crude estimate, empty cavities can be formed at an energetic cost of kT per 10 Å^3. (5) The frequent occurrence of cavities near interdomain boundaries suggests that cavities may be present in some proteins to enhance internal flexibility or interdomain motion.

ACKNOWLEDGMENTS

We thank Drs. J. L. Smith, W. Hendrickson, B. K. Lee, and E. Bucci for stimulating discussions.

Registry No. 4PTI, 9087-70-1; 3CYT, 9007-43-6; 1RN3, 9001-99-4; 2LYZ, 9001-63-2; 2ACT, 39279-27-1; 2PTN, 9002-07-7; 2CHA, 9004-07-3; 3TLN, 9073-78-3; H_2O, 7732-18-5.

REFERENCES

Baldwin, J., & Chothia, C. (1979) *J. Mol. Biol. 129*, 175–220.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Jr., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasuma, M. (1977) *J. Mol. Biol. 112*, 535–542.

Chothia, C. (1976) *J. Mol. Biol. 105*, 1–14.

Connolly, M. L. (1985) *J. Am. Chem. Soc. 107*, 1118–1124.

Doukas, A., Junnarkar, M., Alfano, R., Callender, R., Kakitani, T., & Honig, B. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 4790–4794.

Edsall, J. T., & McKenzie, H. A. (1983) *Adv. Biophys. 16*, 53–183.

Finney, J. L. (1975) *J. Mol. Biol. 96*, 721–732.

Finney, J. L. (1979) in *Water: A Comprehensive Treatise* (Franks, F., Ed.) Vol. 6, pp 47–122, Plenum Press, New York and London.

Holmes, M. A., & Matthews, B. W. (1982) *J. Mol. Biol. 160*, 623–639.

Honig, B., Ebrey, T., Callender, R., Dinur, U., & Ottolenghi, M. (1979) *Proc. Natl. Acad. Sci. U.S.A. 76*, 2503–2507.

Lee, B. K. (1985a) *Biopolymers 24*, 813–823.

Lee, B. K. (1985b) in *Mathematics and Computers in Biomedical Applications* (Eisenfeld J., & Delisi, C., Ed.) pp 3–11, Elsevier, Amsterdam.

Lee, B. K., & Richards, F. M. (1971) *J. Mol. Biol. 55*, 379–400.

Lumry, R., & Rosenberg, A. (1975) *Colloq. Int. C. N. R. S. 246*, 53–62.

Matthews, B. W., Weaver, L. H., & Kesler, W. R. (1974) *J. Biol. Chem. 243*, 8030–8044.

Novotny, J., Bruccoleri, R., & Karplus, M. (1984) *J. Mol. Biol. 177*, 787–818.

Oton, J., Bucci, E., Steiner, R. F., Fronticelly, C., Franchi, D., Montemarano, J., & Martinez, A. (1981) *J. Biol. Chem. 256*, 7248–7256.

Pauling, L. (1960) *The Nature of the Chemical Bond*, 3rd ed., Cornell University Press, Ithaca, NY.

Pavlov, M. Yu., & Fedorov, B. A. (1983) *Biopolymers 22*, 1507–1522.

Rashin, A. A. (1980) *Biomolecular Structure, Conformation, Function and Evolution* (Srinivasan, R., Ed.) Vol. 2, pp 133–149, Pergamon Press, Oxford and New York.

Rashin, A. A. (1984) *Biopolymers 23*, 1605–1620.

Rashin, A. A., & Honig, B. (1984) *J. Mol. Biol. 173*, 515–521.

Reiss, H. (1966) *Adv. Chem. Phys. 9*, 1–84.

Richards, F. M. (1974) *J. Mol. Biol. 82*, 1–14.

Richards, F. M. (1979) *Carlsberg Res. Commun. 44*, 47–63.

Sassaroli, M., Bucci, E., & Steiner, R. F. (1982) *J. Biol. Chem. 257*, 10136–10140.

Schoenborn, B. P. (1969) *J. Mol. Biol. 45*, 297–303.

Schulz, G. E., & Schirmer, R. H. (1979) *Principles of Protein Structure*, (Cantor, C. R., Ed.) Springer-Verlag, New York, Heidelberg, and Berlin.

Shrake, A., & Rupley, J. A. (1973) *J. Mol. Biol. 79*, 351–371.

Smith, J. L., Hendrickson, W. A., Honzatko, R. B., & Sherif, S. (1986) *Biochemistry* (in press).

Tilton, R. F., Jr., Kuntz, I. D., Jr., & Petsko, G. A. (1984) *Biochemistry 23*, 2849–2857.

Titani, K., Hermodson, M. A., Ericsson, L. H., Walsh, K. A., & Neurath, H. (1972) *Biochemistry 11*, 2427–2435.

Vita, C., Fontana, A., Seeman, J. R., & Chaiken, I. M. (1979) *Biochemistry 18*, 3023–3031.

Vita, C., Dalzoppo, D., Fontana, A., & Rashin, A. A. (1984) *Biochemistry 23*, 5512–5519.