Proc. Natl. Acad. Sci. USA
Vol. 77, No. 4, pp. 1736–1740, April 1980
Chemistry

Analytical approximation to the accessible surface area of proteins

(protein structure/domains)

SHOSHANA J. WODAK[*] AND JOËL JANIN

Service de Biochimie Cellulaire, Institut Pasteur 28, rue du Docteur Roux, 75724 Paris 15, France

Communicated by Frederic M. Richards, January 21, 1980

ABSTRACT We propose an analytical substitute to the geometrical construction that is commonly used in calculating the protein surface area that is accessible to the solvent. A statistical approach leads to an expression of accessible surface areas as a function of distances between pairs of atoms or of residues in the protein structure, assuming only that these atoms or residues are randomly distributed in space but not penetrating each other. This function gives good estimates of the accessible surface area and of the area buried in subunit contacts for a number of proteins. Its evaluation is very fast, and the function can be differentiated, which opens the way to new applications of accessibility measurements in the study of proteins. As an example, we show that the presence of domains is easily detected by an automatic procedure based on surface areas only.

The concept of accessible surface area, first proposed by Lee and Richards (1), has found many applications in the study of proteins (2). The accessible surface area of a protein atom is defined as the area of the surface over which a water molecule can be placed while making van der Waals contact with this atom and not penetrating any other protein atom. A geometrical construction (Fig. 1) leads to algorithms that calculate accessible surface areas from atomic coordinates derived from x-ray studies (1, 3, 4). These areas are linearly correlated to the free energies of transfer from polar to nonpolar solvents, or hydrophobic free energies, of hydrocarbons (5–8). Measurements of accessible surface areas and of area changes occurring in various biochemical processes may therefore give insights into the role of the solvent and of hydrophobicity in these processes. For instance, the evaluation of the surface area change when proteins fold (9) or associate (10) shows that hydrophobicity is the major driving factor in folding and in polymerization.

Because of the complexity of protein structures and of the many atoms present, the geometrical algorithms used in measuring accessible surface areas are costly in computer time. Moreover, it would be desirable to represent the surface area as an analytical function of the atomic coordinates because the function and its derivatives could be used in minimization procedures. An increase in computing efficiency can be achieved by simplifying the protein structure and representing each amino acid residue rather than each atom of the structure by a sphere (11). We have shown that good estimates of accessible surface areas are obtained in this way, and a systematic analysis of the trypsin–pancreatic trypsin inhibitor complex was made possible by the quickness of the calculation (12), even with the algorithm of Lee and Richards.

We develop here an *analytical approximation to the accessible surface area, expressed as a function of interatomic distances only*. The approximation is based on a statistical approach assuming that atoms or amino acid residues are randomly distributed. Measurements of accessible surface areas and of areas buried in protein structures by use of the analytical approximation are shown to be in close agreement with those of the geometrical procedure. Because the evaluation of the analytical function is very much faster, it opens the way to new applications of the accessibility measurements. We present as an example an automatic procedure that defines compact domains in proteins on the basis of accessibility criteria.

The surface area function

In defining the accessible surface areas according to Lee and Richards (1), we draw spheres of radii $r + r_w$ around each atom of the protein structure; r is the van der Waals radius of the atom and r_w is the radius of a sphere simulating a water molecule, typically 1.4 Å (Fig. 1). The spheres intersect, and the accessible surface area of the atom is:

$$A = S - B, \qquad [1]$$

in which S is the total surface area of the sphere $\underline{S}$ attached to that atom,

$$S = 4\pi(r + r_w)^2, \qquad [2]$$

and B (buried surface area) is the area cut out of the surface of $\underline{S}$ by all intersecting spheres. This area can be easily calculated when two spheres $\underline{S}_1$ and $\underline{S}_2$ only are present (Fig. 1). It is zero if the interatomic distance d is larger than the sum $r_1 + r_2 + 2r_w$ of the radii of the two spheres. Otherwise, the area cut out of $\underline{S}_1$ by $\underline{S}_2$ is:

$$b = \pi(r_1 + r_w)(r_1 + r_2 + 2r_w - d)\left(1 + \frac{r_2 - r_1}{d}\right). \qquad [3]$$

This simple function of the distance d represents the surface area of atom 1 buried by atom 2.

When more than two spheres are present, the surface area B buried on atom 1 is not simply the sum of the surface areas $b_2, b_3, \ldots b_n$, buried by all other atoms, for in general the surfaces cut out of $\underline{S}_1$ by the spheres $\underline{S}_2, \underline{S}_3, \ldots \underline{S}_n$ overlap. Because it is extremely difficult to give an exact analytical expression of B, geometric evaluations such as those of Lee and Richards (1) are used.

However, taking a statistical approach to the problem, we may say that *the probability for a point on the surface of* $\underline{S}_1$ *to be outside an intersecting sphere* $\underline{S}_i$ is simply $1 - b_i/S$, where b_i is given by Eq. 3. If we now assume the spheres to be randomly distributed, the probability for a point on the surface of $\underline{S}_1$ to be outside *all* intersecting spheres, and therefore to be accessible, is the product of individual probabilities. Then

$$A' = S \prod_{i=2}^{n} (1 - b_i/S) \qquad [4]$$

is an estimate of the accessible surface area of atom 1.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

[*] Present address: Laboratoire de Chimie Biologique, Université Libre de Bruxelles, 67 rue des Chevaux, 1640 Rhode-St-Genèse, Belgium.

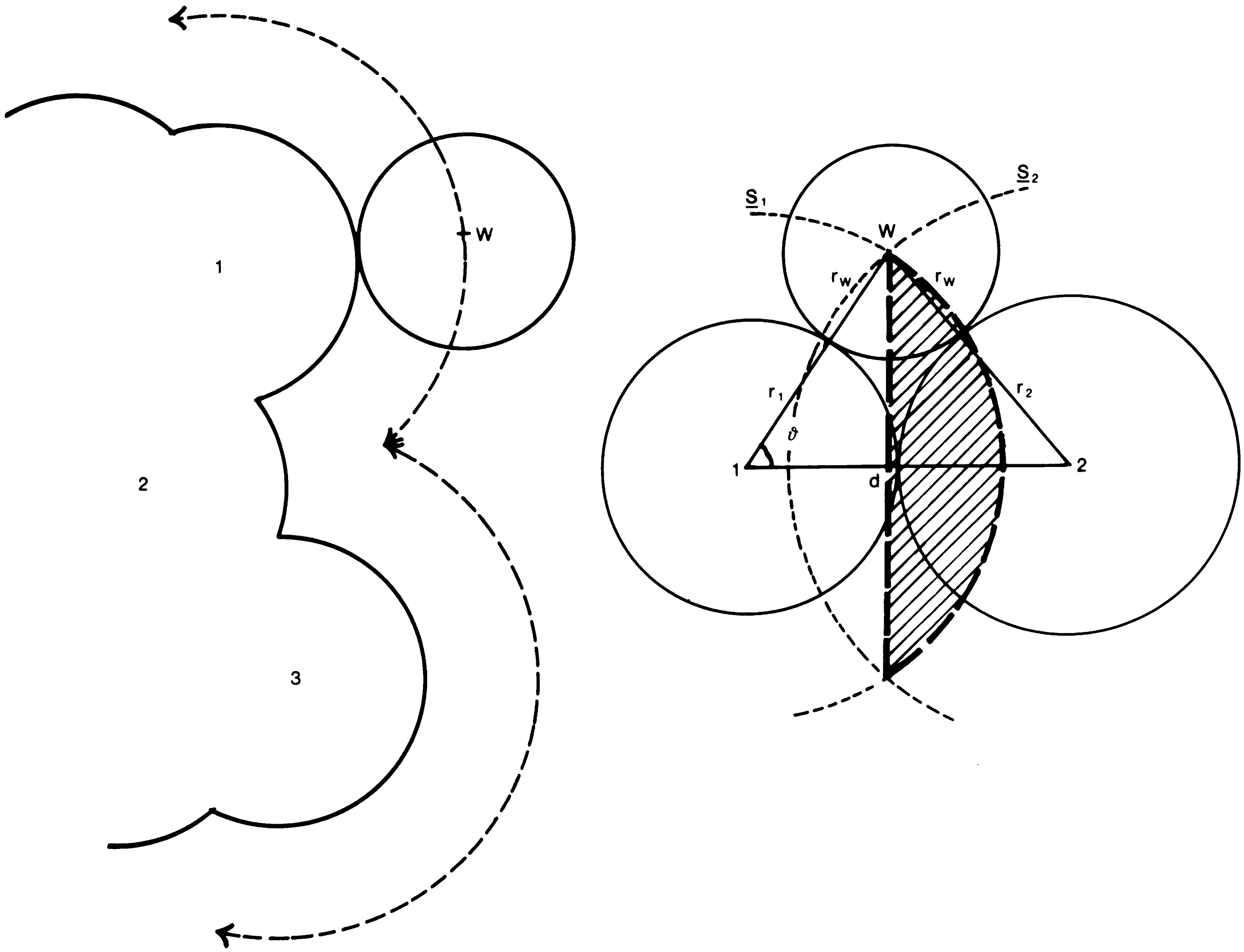

FIG. 1. Accessible surface areas (*Left*) and surfaces buried in contacts between atoms (*Right*) are represented in two dimensions. The region of the surface of the sphere $\underline{S}_1$ centered on atom 1 over which may be placed a water molecule W (taken to be a sphere for convenience) not penetrating atoms 2, 3, etc. is the accessible surface of atom 1 (dashed line). The hatched surface on the right is the region of $\underline{S}_1$ that is no longer accessible to the water molecule W due to the presence of atom 2. Its surface area is $b = \pi(r_1 + r_w)(1 - \cos\theta)$, where $\cos\theta$ may be derived from $(r_2 + r_w)^2 = d^2 + (r_1 + r_w)^2 - 2d(r_1 + r_w)\cos\theta$, leading to Eq. 2 in the text.

The assumption of a random distribution of spheres is incorrect unless excluded volumes are taken into account. With three atoms, Fig. 2 shows that, when the spheres $\underline{S}_3$ and $\underline{S}_4$ are not allowed to penetrate $\underline{S}_2$ by more than a certain distance s (for hard sphere atoms, s is equal to $2r_w$), some of the surface cut out of $\underline{S}_1$ by $\underline{S}_2$ cannot overlap with that cut out by $\underline{S}_3$ or any other sphere. Thus, whereas b in Eq. 3 represents the *maximum* buried surface area, the *minimum* surface area b' buried by atom 2 on atom 1 is the area cut out of $\underline{S}_1$ by a sphere of radius $r_2 + r_w - s$, which, according to Eq. 3, should be:

$$b' = \pi(r_1 + r_w)(r_1 + r_2 + 2r_w - s - d)\left(1 + \frac{r_2 - s - r_1}{d}\right) \qquad [5]$$

($b' = 0$ if negative).

Considering all neighboring atoms, we shall take as an approximation to the accessible surface area of atom 1:

$$A_c = A' - B' \quad (A_c = 0 \text{ if } A' < B'), \qquad [6]$$

where

$$A' = S \prod_{i=2}^{n} \left(1 - \frac{b_i - b_i'}{S}\right) \qquad [7]$$

and

$$B' = \sum_{i=2}^{n} b_i'. \qquad [8]$$

The components b_i and b_i' of the buried surface area are calculated for each neighbor i of atom 1 by Eqs. 3 and 5, respectively. They are functions of the distance d_i between these two atoms and not of the positions of other atoms. In the absence of neighbor, the accessible surface area of atom 1 is $A_c = S$, the total area of sphere $\underline{S}_1$. When the number of neighbors increases, B' increases and A' decreases until eventually A_c becomes zero or negative. Negative values, which have no physical meaning, are taken to be zero. The accessible surface areas being additive, the calculation of A_c may be repeated for all atoms in the structure; individual atomic values are summed to yield the total accessible surface area of a molecule.

The partial derivative of A_c relative to the distance d_i of atom 1 to atom i is

$$\frac{\partial A_c}{\partial d_i} = \frac{\partial A'}{\partial d_i} - \frac{\partial B'}{\partial d_i}, \qquad [9]$$

where

$$\frac{\partial A'}{\partial d_i} = -A'\left(\frac{db_i}{dd_i} - \frac{db_i'}{dd_i}\right) \Big/ (S - b_i + b_i') \qquad [10]$$

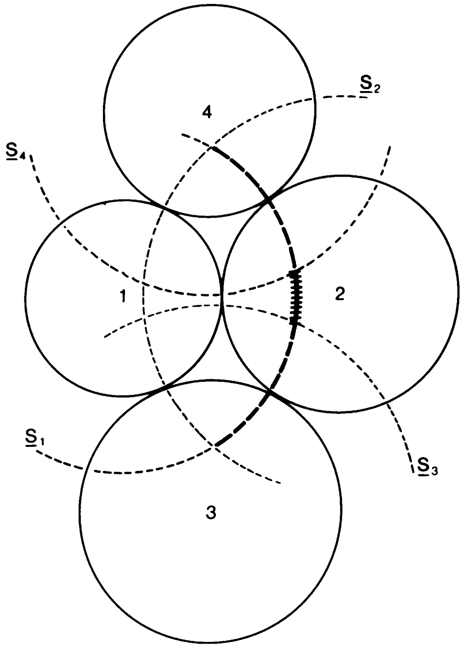

FIG. 2. Effect of volume exclusion on buried surface areas. The dashes represent the surface of the sphere $\underline{S}_1$ attached to atom 1 that is buried in contacts with atom 2. It overlaps strongly with the surfaces buried by atoms 3 and 4. However, no overlap can occur in the hatched region if atoms 3 and 4 are forbidden to penetrate atom 2.

and

$$\frac{\partial B'}{\partial d_i} = \frac{\mathrm{d}b_i'}{\mathrm{d}d_i}. \qquad [11]$$

The derivatives of b_i and b_i' are obtained by differentiation of Eqs. 3 and 5.

Assessment of the function

We chose to apply our statistical approximation of the surface area function to a simplified model of protein structures rather than to all atoms in these structures. In the simplified model (11), each amino acid residue is reduced to a sphere centered on its center of mass. Residues may therefore be treated as single atoms, with appropriate van der Waals radii for the 20 types of amino acids. Accessible surface areas calculated on simplified protein models with the geometrical procedures of Lee and Richards agree well with the values obtained with all atoms present, but the calculation is much faster (12). The assumption of a random distribution in space, which is at the basis of our statistical approach, is probably better for residues than for individual atoms. Moreover, application of the function to atoms rather than residues may require a different treatment for covalently bound atoms, which do interpenetrate, and noncovalently bound atoms, which do not. These points are presently under investigation.

In Table 1, we compare the results of accessibility measurements made on phage T4 lysozyme with the procedure of Lee and Richards applied either to all atoms or to the "simplified" protein and the values obtained by using the analytical function. The agreement is very good (2–4%) both on the "native" (crystallographic) structure and on a "denatured" protein with the polypeptide chain in an extended configuration. The analytical function contains an adjustable parameter, the distance s over which spheres representing the surface of residues are allowed to penetrate; s should be nearly equal to the diameter of a water molecule, or 2.8 Å. Preliminary tests showed that a value of s = 2.5 Å gave the best results; it was used all through this work.

Table 1. Accessible surface area of phage T4 lysozyme

Lysozyme	Surface area, Å^2	Ratio	R factor
Native structure			
All atoms	8,962	(1.0)	
Simplified model	8,604	0.96	0.13
Analytical	8,767	0.98	0.19
Denatured			
All atoms	23,005	(1.0)	
Simplified model	22,927	1.0	0.07
Analytical	23,588	1.02	0.07

The accessible surface area of phage T4 lysozyme was calculated by using: (*i*) a computer program of M. Levitt (12), which implements the procedure of Lee and Richards, with all atoms present; (*ii*) the same procedure and the simplified model of the protein structure; and (*iii*) with the analytical function and the simplified model. The "native" structure refers to atomic coordinates of the Cambridge Data Bank as determined by Remington *et al.* (13). The "denatured" protein is obtained by artificially setting all ϕ and ψ dihedral angles to −140° and 140°, respectively. The residue radii in the simplified model are listed in ref. 12. The water radius r_w is 1.4 Å, the s parameter, 2.5 Å. The ratios given are those of the approximate surface areas (*ii* and *iii*) to the exact ones (*i*). The average deviation R between values obtained for each of the 164 residues of phage T4 lysozyme is defined in the text.

The good agreement observed in Table 1 shows that *the analytical function applied to simplified proteins represents correctly the accessible surface area* and the area buried in the globular structure (the difference between the "denatured" and "native" values) of phage T4 lysozyme. Similar results are obtained with other proteins (Table 2) and with protein complexes in which the accessible surface area and *the area buried in subunit contacts are correctly evaluated by the analytical function*. Not surprisingly, the fit between the geometrical procedure and the analytical function is less good when the accessibilities of individual amino acid residues are compared (Fig. 3). Most of the discrepancy results from random errors introduced by the replacement of residues by single spheres. Still, the average discrepancy,

$$R = \frac{\sum |A_c{}^i - A^i|}{\sum A^i}, \qquad [12]$$

between accessible surface areas of individual residues calculated with the geometric procedure and all atoms present (A^i) or with the single sphere representation and the analytical function ($A_c{}^i$) is only about 20%, a reasonable value for such a crude model. The correlation between $A_c{}^i$ and A^i being linear (Fig. 3), the discrepancy observed with individual residues averages out on proteins with tens or hundreds of residues.

Defining compact domains in proteins

We propose to define compact domains in protein structures purely on the basis of surface area criteria. They are groups of residues having a *minimum surface-to-volume ratio, which implies compactness, and a minimum surface of contact with the remainder of the structure.* In usual terms, these criteria define relatively autonomous regions with most interactions within the region and least without. Domains including a single chain segment can easily be recognized by the following procedure: the chain is assumed to be cut at some point, leading to two artificial subunits whose accessible surface area is calculated; the sum of the two surface areas minus that of the whole protein represents the interface area between the two subunits (10). When the cutting point is moved along the polypeptide chain, limits of domains will appear as minima of the interface area. At these points, the artificial subunits have

Table 2. Accessible and buried surface areas calculated by using the analytical function

Protein	Accessible surface areas, Å^2			Buried surface areas, Å^2		
	A	A_c	Ratio	A	A_c	Ratio
IgG fragment V$_{REI}$						
Monomer	5,528	5,376	0.97			
Dimer	9,625	9,400	0.93	1,431	1,352	0.95
Human deoxyhemoglobin						
α chain	7,829	7,833	1.0			
β chain	8,226	8,257	1.0			
$\alpha\beta$ dimer	14,425	14,660	1.02	1,630	1,430	0.88
Trypsin–BPTI complex						
Trypsin	8,902	9,530	1.07			
BPTI	3,556	3,635	1.02			
Complex	11,075	11,930	1.08	1,383	1,235	0.89
Lobster GPDH						
Monomer	15,470	16,120	1.04			
Red/blue dimer	28,020	29,290	1.04	2,920	2,950	1.01
Red/green dimer	28,814	30,100	1.04	2,126	2,140	1.0
Tetramer	50,310	53,880	1.07	11,570	10,600	0.92
Dogfish apo-LDH						
Monomer	17,120	17,867	1.04			
Red/blue dimer	31,640	33,120	1.05	2,600	2,614	1.0
Red/yellow dimer	28,700	30,520	1.06	5,540	5,214	0.94
Red/green dimer	30,015	31,860	1.06	4,225	3,874	0.92
Tetramer	45,170	49,330	1.09	23,310	22,140	0.95
Concanavalin A						
Monomer	10,600	11,350	1.07			
Dimer I-II	18,620	20,248	1.08	2,580	2,452	0.95
Dimer I-III	18,745	20,124	1.07	2,455	2,576	1.04
Tetramer	32,060	35,293	1.10	10,340	10,107	0.98

Accessible surface areas were calculated by the procedure of Lee and Richards and a computer program of M. Levitt (A) or by the analytical function on simplified protein models (A_c). The surface areas buried in subunit interfaces are defined as the sum of the subunit accessible surface areas minus that of the complex (10). The ratios given are A_c/A. All atomic coordinates are from the Cambridge Data Bank. When two subunits were independently determined (V$_{REI}$ fragment and GPDH), the values quoted are averages. For the larger multimeric proteins, LDH, GPDH, and concanavalin A, the simplified model was used in calculating A. The LDH and GPDH dimers are defined in ref 14; the concanavalin dimers in ref 15. GPDH, glucose-6-phosphate dehydrogenase; LDH, lactate dehydrogenase; BPTI, pancreatic trypsin inhibitor.

a minimum of interactions and most closely resemble actual protein subunits. Fig. 4 illustrates the efficiency of this procedure applied to an immunoglobulin light chain, where the division into two domains is obvious owing to their very small interface.

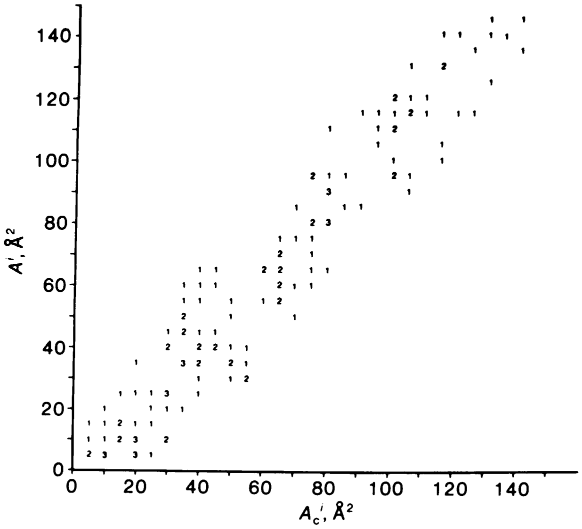

FIG. 3. Accessibility of amino acid residues in phage T4 lysozyme. The values $A_c{}^i$ of the accessible surface areas of the residues calculated with the analytical function and the simplified model are plotted against the corresponding values A^i measured by the procedure of Lee and Richards with all atoms present.

Conclusion

The importance of the solvent in processes involving proteins and other biological molecules has long been recognized. Because of the complexity of the physics in the liquid state, and especially of liquid water, it has proved very difficult to estimate quantitatively the interactions of water with proteins. Though actual interactions of water molecules with chemical groups (for instance, hydrogen bonds) can be demonstrated, the effect of organic solutes on bulk water, generally but im-

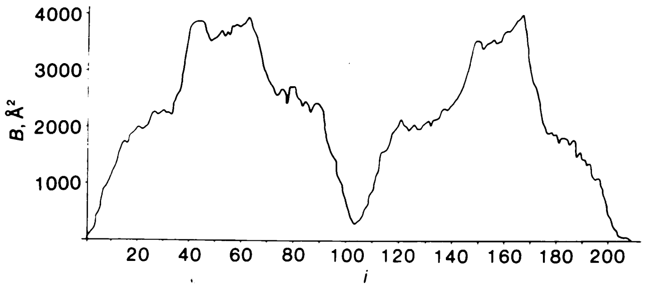

FIG. 4. Compact domains in immunoglobulin light chains. The light chain moiety of an immunoglobulin Fab fragment (16) was cut successively at each residue i. The accessible surface areas A_1 and A_2 of the two fragments (1 to i) and (i+1 to 208) were calculated. The sum $A_1 + A_2$ is larger than the accessible surface area A_0 of the light chain itself (A_0 = 12,218 Å^2); the difference $B = A_1 + A_2 - A_0$ is plotted here against i. B is minimum at residue 103, defining two domains (1 to 103) and (104 to 208). Their accessible surface areas are A_1 = 5952 Å^2 and A_2 = 6566 Å^2; B = 300 Å^2 is the area of their interface.

properly known as the hydrophobic effect (17), may not be counted simply as a sum of interatomic interactions. A number of different approaches have been proposed, including powerful molecular dynamics (18, 19) or Monte Carlo simulations (20) of water surrounding protein molecules. However, these calculations are very complex even with the smallest proteins.

The use of the accessible surface areas provides a bypass to the problem. Rather than trying to represent the details of the interactions occurring between the many protein atoms and the fluctuating solvent surrounding, it is reasonable to assume that their free energy (which includes a strong entropic component) is proportional to the surface of contact between the protein molecule and water. The assumption has been checked experimentally with small organic molecules, including amino acids. If it is valid for macromolecules, it becomes easy to evaluate the solvent contribution to the stability of proteins or protein complexes when their three-dimensional structures are known. The procedure developed in this paper removes the major difficulties linked to the geometric construction used previously in accessibility measurements. Accessible surface areas are calculated simply and efficiently as a combination (but not a sum) of pairwise interactions between residues, these interactions being now represented by the *b* and *b′* components of the buried surface areas for each pair of residues. As a consequence, accessible surface areas can be used in energy minimization or other procedures that operate on analytical functions of the atomic positions.

We thank M. Levitt for the gift of computer programs. Part of this work was completed during the 1979 Woıkshop of the Centre Européen de Calcul Atomique et Moléculaire on Protein Folding in Orsay, France. S.J.W. thanks the Fonds National de la Recherche Scientifique Suisse for support.

1. Lee, B. K. & Richards, F. M. (1971) *J. Mol. Biol.* **55,** 379–400.
2. Richards, F. M. (1977) *Annu. Rev. Biophys. Bioeng.* **6,** 151–176.
3. Shrake, A. & Rupley, J. A. (1973) *J. Mol. Biol.* **79,** 351–371.
4. Finney, J. L. (1978) *J. Mol. Biol.* **119,** 415–441.
5. Hermann, R. B. (1972) *J. Phys. Chem.* **76,** 2754–2759.
6. Harris, M. J., Higushi, T. & Rytting, J. H. (1973) *J. Phys. Chem.* **71,** 2694–2703.
7. Chothia, C. (1974) *Nature (London)* **248,** 338–339.
8. Reynolds, J. A., Gilbert, D. B. & Tanford, C. (1974) *Proc. Natl. Acad. Sci. USA* **71,** 2925–2927.
9. Chothia, C. (1976) *J. Mol. Biol.* **105,** 1–14.
10. Chothia, C. & Janin, J. (1975) *Nature (London)* **256,** 705–708.
11. Levitt, M. (1976) *J. Mol. Biol.* **104,** 59–107.
12. Wodak, S. J. & Janin, J. (1978) *J. Mol. Biol.* **124,** 323–342.
13. Remington, S. J., Anderson, W. F., Owen, J., Ten Eyck, L. F., Graigner, C. T. & Matthews, B. W. (78) *J. Mol. Biol.* **118,** 81–98.
14. Rossmann, M. G., Adams, M. J., Buehner, M., Ford, G. C., Hackert, M. L., Liljas, A., Rao, S. T., Banaszak, L. J., Hill, E., Tsernoglou, D. & Webb, L. (1973) *J. Mol. Biol.* **76,** 533–537.
15. Reeke, G. N., Jr., Becker, J. W. & Edelman, G. M. (1975) *J. Biol. Chem.* **250,** 1525–1547.
16. Saul, F. A., Amzel, L. M. & Poljack, R. J. (1978) *J. Biol. Chem.* **253,** 585–597.
17. Hildebrand, J. H. (1979) *Proc. Natl. Acad. Sci. USA* **76,** 194.
18. McCammon, J. A., Gelin, B. R. & Karplus, M. (1977) *Nature (London)* **267,** 585–590.
19. Rosky, P. J., Karplus, M. & Rahman, A. (1979) *Biopolymers* **18,** 825–854.
20. Hagler, A. T. & Moult, J. (1978) *Nature (London)* **272,** 222–226.