J. Mol. Biol. (1987) **196**, 641–656

Interior and Surface of Monomeric Proteins

Susan Miller[1], Joël Janin[2], Arthur M. Lesk[3,4,†] and Cyrus Chothia[1,3]

[1] *Christopher Ingold Laboratories, University College London*
20 Gordon Street, London WC1G 0AJ, England

[2] *Laboratoire de Biologie Physicochimique, Université Paris-Sud*
Bât. 433, 91405 Orsay, France

[3] *Medical Research Council, Laboratory of Molecular Biology*
Hills Road, Cambridge CB2 2QH, England

[4] *Biocomputing Programme, EMBL, Meyerhofstr. 1, Postfach 1022.09*
D-6900 Heidelberg, Federal Republic of Germany

(Received 6 October 1986, and in revised form 13 March 1987)

The solvent-accessible surface area (A_s) of 46 monomeric proteins is calculated using atomic co-ordinates from high-resolution and well-refined crystal structures. The A_s of these proteins can be determined to within 1 to 2% and that of their individual residues to within 10 to 20%.

The A_s values of proteins are correlated with their molecular weight (M_r) in the range 4000 to 35,000: the power law $A_s = 6{\cdot}3\ M^{0{\cdot}73}$ predicts protein A_s values to within 4% on average. The average water-accessible surface is found to be 57% non-polar, 24% polar and 19% charged, with 5% root-mean-square variations. The molecular surface buried inside the protein is 58% non-polar, 39% polar and 4% charged. The buried surface contains more uncharged polar groups (mostly peptides) than the surface that remains accessible, but many fewer charged groups.

On average, 15% of residues in small proteins and 32% in larger ones may be classed as "buried residues", having less than 5% of their surface accessible to the solvent. The accessibilities of most other residues are evenly distributed in the range 5 to 50%. Although the fraction of buried residues increases with molecular weight, the amino acid compositions of the protein interior and surface show no systematic variation with molecular weight, except for small proteins that are often very rich in buried cysteines.

From amino acid compositions of protein surfaces and interiors we calculate an effective coefficient of partition for each type of residue, and derive an implied set of transfer free energy values. This is compared with other sets of partition coefficients derived directly from experimental data. The extent to which groups of residues (charged, polar and non-polar) are buried within proteins correlates well with their hydrophobicity derived from amino acid transfer experiments. Within these three groups, the correlation is low.

1. Introduction

The polypeptide chains of globular proteins fold into compact shapes, in which large parts of the chain are shielded from contact with the solvent. It is believed that this shielding is the primary source of free energy stabilizing the native conformation. Kauzmann (1959) reviewed the forces that maintain the structure of globular proteins: covalent bonds, dispersion forces (van der Waals' forces), salt bridges, hydrogen bonds and "hydrophobic bonds", which arise from the preference of non-polar compounds for non-aqueous solvents. Kauzmann estimated the strength of hydrophobic bonds and pointed out that they are largely entropic in nature, in contrast to other bonds, which are mostly enthalpic. The contribution of hydrophobic bonds to the free energy of folding of a protein in water should be related to the partition coefficient of model organic compounds between water and

[†] Also associated with: Fairleigh Dickinson University, Teaneck-Hackensack Campus, Teaneck, NJ 07666, U.S.A.

0022-2836/87/150641-16 $03.00/0

organic solvents, because, as proteins fold, hydrophobic groups are transferred from an aqueous environment to the much less polar environment of the protein interior.

The crystal structure of myoglobin reported in 1960, and those of other proteins determined in the following years, confirmed much of Kauzmann's insight: protein interiors contain primarily nonpolar aliphatic and aromatic residues, with polar residues on the surface (Perutz *et al.*, 1965). Polar residues are generally defined as having hydrogen bond acceptor or donor groups in their side-chains. Lee & Richards (1971) were the first to attach numbers to this qualitative description of how amino acid residues distribute themselves between the protein interior and its surface. They used atomic co-ordinates derived from X-ray studies to measure the area of the protein surface in contact with solvent in the native structure. This is called the solvent-accessible surface area or ASA. Their analysis, applied to myoglobin, lysozyme and ribonuclease, gave a rather unexpected result: a large fraction of the protein surface (considered on an atom-by-atom rather than a residue-by-residue basis) was found to be non-polar, and a large fraction of the buried surface to be polar. Shrake & Rupley (1973) confirmed these results, noting however that non-polar residues had a markedly greater tendency to be buried than did polar ones.

The apparent contradiction between the facts that the protein interior is made primarily of nonpolar residues and that a large fraction of the buried surface is polar was eventually resolved. Chothia (1976) showed that the formation of the secondary structure stabilized by hydrogen bonds buries large amounts of polar surface, mostly from the peptide groups. In contrast, the assembly of α-helices and β-sheets into the tertiary structure buries almost exclusively non-polar surface, as aliphatic and aromatic side-chains come together to form close-packed interfaces with very few hydrogen bonds. Thus, the hierarchy of protein structure makes it possible to satisfy the requirement for the polar groups to form hydrogen bonds, while allowing a large fraction of the non-polar groups to be removed from contact with water more or less in the way Kauzmann had predicted.

The concept of accessible surface area provides convenient definition of the protein surface and interior and, thereby, permits a quantitative analysis of their chemical nature. We use here the *residue accessibility* (Lee & Richards, 1971), defined for each residue as the ratio of its ASA in the native protein to the ASA it would have in an unfolded and extended polypeptide chain, with side-chain conformations taken to be those typical of globular proteins. Amino acid residues having accessibilities above a threshold value are defined as accessible and constitute the protein surface; those having accessibilities lower than the threshold are defined as buried and constitute the interior. The amino acid compositions of the protein surface and interior can then be calculated. They are strikingly different, and the relative abundance of each residue type can be used to derive effective surface/interior partition coefficients for all 20 amino acid residues from a set of protein X-ray structures (Chothia, 1976; Janin, 1979).

There has been considerable interest in correlating the partition of residue types between the surface and interior of proteins with physicochemical properties of the amino acids. It can be compared, for instance, with the partition of amino acids or their derivatives between water and organic solvent (Nozaki & Tanford, 1971; Fauchère & Pliska, 1983), or between water and the vapour phase (Wolfenden *et al.*, 1981; Wolfenden, 1983). Hydropathy scales derived from partition coefficients are commonly used in the prediction of protein structure (Kyte & Doolittle, 1982; Eisenberg *et al.*, 1982). Many scales have been proposed; they are quite different, and often correlate poorly (Rose *et al.*, 1985*a*).

The present study makes use of the many new protein structures that have been determined in recent years, and of the better quality of the atomic co-ordinates derived from these X-ray studies after crystallographic refinement. We include here data from 37 structures of monomeric proteins that have been determined to 2·5 Å resolution or better, and have been well refined. Nine other closely related structures, of comparable quality, are used to assess the accuracy of ASA measurements. We have extended our previous work in two ways. We first determined how quantities such as the total protein ASA (A_s), the proportion of buried residues, the average accessible and buried surface areas per amino acid residue, and the fraction of residues that are buried, vary with the protein molecular weight. From these data we derived distribution parameters specific to each residue. We then utilized these results for reconsideration of the relationship between the interior/surface partitioning of residues and their hydrophobicity. This study is limited to monomeric proteins. Oligomeric proteins, which have the added complexity of subunit interfaces, will be discussed elsewhere.

2. Methods and Results

(a) *Accessible surface area measurements*

Accessible surface areas for individual atoms of 46 proteins (Table 1) were calculated from atomic co-ordinates deposited in the Protein Data Bank (Bernstein *et al.*, 1977) or given to us by the authors. We used the Shrake & Rupley (1973) algorithm as implemented by one of us (A.M.L.). For each protein atom A, a sufficiently large number of approximately evenly distributed points are placed on the solvation sphere of radius $R_A + R_W$ centered at the atomic position, where R_A is the van der Waals' radius of atom A, and R_W that of the solvent probe. In the absence of hydrogen atoms, we used group radii (Chothia, 1975) rather than atomic radii, and a solvent probe radius $R_W = 1·4$ Å.

Table 1
Protein structures

Protein	File name	Reference
Actinidin	2ACT	Baker & Dodson (1980)
Azurin, *Alcaligenes denitrificans*	1AZA	Norris *et al.* (1983)
Carbonic anhydrase B	2CAB	Kannan *et al.* (1984)
Carboxypeptidase A	5CPA	Rees *et al.* (1983)
Carboxypeptidase inhibitor (ICPA)	4CPA	Rees & Lipscomb (1982)
Chymotrypsin	4CHA	Tsukada & Blow (1985)
Crambin	1CRN	Hendrickson & Teeter (1981)
Cytochrome *b*5	2B5C	Matthews *et al.* (1972)
Cytochrome *c*, tuna, oxidized	3CYT	Takano & Dickerson (1981*a*)
reduced	4CYT	Takano & Dickerson (1981*b*)
Cytochrome *c*, rice	1CCR	Ochi *et al.* (1983)
Cytochrome *c*3, *Desulfovibrio vulgaris*	2CDV	Higuchi *et al.* (1984)
Cytochrome peroxydase	1CYP	Finzel *et al.* (1984)
Dihydrofolate reductase (DHFR), *E. coli*	4DFR	Bolin *et al.* (1982)
Dihydrofolate reductase (DHFR), *L. casei*	3DFR	Bolin *et al.* (1982)
Erythrocruorin, *Chironomus*, deoxy	1ECD	Steigemann & Weber (1979)
carbonmonoxy	1ECO	Steigemann & Weber (1979)
Ferredoxin, *P. aerogenes*	1FDX	Adman *et al.* (1973)
Ferredoxin, *S. platensis*	3FXC	Tsukihara *et al.* (1981)
Flavodoxin, *Clostridium MP*	3FXN	Smith *et al.* (1977)
High potential iron protein (HIPIP)	1HIP	Carter *et al.* (1974)
Intestinal calcium binding protein (ICBP)	1ICB	Szebenyi *et al.* (1981)
Lysozyme, human	1LZ1	Artymiuk & Blake (1981)
Lysozyme, hen	(*)	Grace (1979)
Lysozyme, phage T4	(*)	Weaver & Matthews (1987)
Myoglobin	1MBD	Phillips (1980)
Neurotoxin B, sea snake	1NXB	Tsernoglou & Petsko (1977)
Neurotoxin 3, scorpion	1SN3	Almassy *et al.* (1983)
Nuclease, *Staphylococcus aureus*	2SNS	Cotton *et al.* (1979)
Nuclease, *Bacillus amyloliquefaciens* (Barnase)	(*)	Mauguen *et al.* (1982)
Pancreatic trypsin inhibitor (PTI), bovine, form I	4PTI	Deisenhofer & Steigemann (1974)
form II	5PTI	Wlodawer *et al.* (1987)
Papain	(*)	Kamphuis *et al.* (1985)
Parvalbumin, carp	3CPV	Moews & Kretsinger (1975)
Pepsin, *Penicillium*	2APP	James & Sielecki (1983)
Phospholipase A2,	1BP2	Dijkstra *et al.* (1981)
Plastocyanin	1PCY	Guss & Freeman (1983)
Protease II, rat	3RP2	Reynolds *et al.* (1985)
Protease A, *S. griseus* (SGPA)	2SGA	Sielecki *et al.* (1979)
Protease B, *S. griseus* (SGPB)	3SGB	Read *et al.* (1983)
Retinol binding protein	(*)	Newcomer *et al.* (1984)
Ribonuclease A (X-ray)	1RN3	Borkakoti *et al.* (1982)
Ribonuclease (joint neutron and X-ray refinement)	5RSA	Wlodawer & Sjölin (1983)
Rubredoxin, *D. vulgaris*	3RXN	Adman & Jensen (1979)
Subtilisin BPN′	1SBT	Alden *et al.* (1971)
Thermolysin	3TLN	Holmes & Matthews (1982)

File names are from the Protein Data Bank (Bernstein *et al.*, 1977). Abbreviations used in the text are given in parentheses.
(*) Co-ordinates are gifts from authors.

The total ASA of residues in an extended polypeptide chain (Table 2) was estimated by calculating the ASA of Gly-X-Gly peptides in extended conformations ($\phi = -120^\circ$, $\psi = 140^\circ$). The side-chain conformations were the ones most frequently observed in proteins (Janin *et al.*, 1978). The values appearing in Table 2 are similar to those reported by Shrake & Rupley (1973) and Rose *et al.* (1985*b*), though we used slightly different atomic radii and peptide conformations.

(b) *Accuracy of accessible surface areas*

Some protein structures have been determined more than once, in different crystal forms or as independent molecules in the asymmetric unit of the same crystal. Slight differences between the structures are observed, which reflect small changes in conformation and experimental errors. Similar changes are expected to occur between molecules in crystalline states and in solution. The root-mean-square (r.m.s.†) displacement of main-chain atoms between optimally superposed pairs of highly refined X-ray structures of the same protein, is 0·25 to 0·4 Å, about twice the estimated error in atomic positions (Chothia & Lesk, 1986).

We calculated accessible surface areas for each set of atomic co-ordinates derived from six such pairs of duplicate structures (Table 3). The

† Abbreviations used: r.m.s., root-mean-square; see Table 1 also.

Table 2

Total accessible surface areas of amino acid residues

Residue	Total ASA (Å^2)	Main-chain ASA (Å^2)	Side-chain ASA		
			Total	Non-polar	Polar
Ala	113	46	67	67	
Arg	241	45	196	89	107
Asn	158	45	113	44	69
Asp	151	45	106	48	58
Cys	140	36	104	35	69
Gln	189	45	144	53	91
Glu	183	45	138	61	77
Gly	85	85			
His	194	43	151	102	49
Ile	182	42	140	140	
Leu	180	43	137	137	
Lys	211	44	167	119	48
Met	204	44	160	117	43
Phe	218	43	175	175	
Pro	143	38	105	105	
Ser	122	42	80	44	36
Thr	146	44	102	74	28
Trp	259	42	217	190	27
Tyr	229	42	187	144	43
Val	160	43	117	117	

ASA calculated for residue X in a Gly-X-Gly tripeptide, with the main chain in an extended conformation.

differences are an indication of the sensitivity of our measurements to the accuracy of atomic positions. The two sets of ribonuclease A co-ordinates are from the same crystal form and reflect experimental errors in recording and analysing diffraction data. They yield essentially the same ASA for the entire protein molecule and for most individual amino acid residues: the average ASA change for a residue is only 4·1 Å^2. However, two particular residues change by more than 20 Å^2, both lysines with highly exposed extended side-chains, with ASA values ≥ 100 Å^2.

Other differences reported in Table 3 may reflect the effects of small genuine conformational changes as well as experimental errors. The contribution of these effects to the protein A_s is very small, except perhaps in azurin, where two molecules in different crystalline environments differ by 2·8% in A_s. A_s values of individual amino acid residues change by no more than in ribonuclease A. Of 789 individual amino acid residues in our sample, 714 (90%) vary by less than 10 Å^2, which is 16% of their mean ASA (60 Å^2). Most of the 16 residues (2% of the total) that change by more than 20 Å^2 are highly exposed and polar.

The data of Table 3 suggest that meaningful limits on the accuracy of the numbers we calculate are that protein A_s values can be determined to within 1 to 2%, and individual residue ASA values to within 10 to 20%.

(c) *Comparison of homologous proteins*

A comparison of homologous proteins shows how the evolution of the amino acid sequence affects the accessible surface area (see Table 4). Overall protein A_s values vary little (1 to 4% in the five protein pairs of Table 4), as could be expected from the correlation of A_s with molecular weight. Individual residues vary more. In homologous proteins we define structurally equivalent residues as those occupying the same position in the three-dimensional structure (the "core residues" as defined by Chothia & Lesk, 1986). Table 4 shows that the ASA

Table 3

Accuracy of accessible surface areas

Protein	A_s (Å^2)	A_s change (%)	Mean ASA change per residue (Å^2)
Ribonuclease A (neutrons/X-rays)	6800	0·4	4·1
Cytochrome *c*, tuna (oxidized/reduced)	5600	1·3	3·9
Erythrocruorin (deoxy/met)	6600	0·0	0·7
PTI (form II/form I)	3940	1·2	4·5
Azurin†	6250	2·8	6·2
Rat protease†	10,320	0·3	

† Two protein molecules in the crystal asymmetric unit.

Table 4

Accessible surface areas of homologous proteins

Protein	Total number of residues	Core residues		Conserved residues	
		Number	ASA change ($Å^2$)	Number	ASA change ($Å^2$)
Lysozyme, human/hen	130/129	129	18·3	77	8·2
S. griseus proteases A and B	181/185	173	11·2	112	4·7
Papain–actinidin	212/218	211	16·5	100	5·8
Cytochrome *c*, tuna/rice	103/113	103	18·4	61	10·0
DHFR†, *E. coli*/*L. casei*	160/164	153	22·6	42	17·8

† See Table 1.

of equivalent residues differs by 10 to 20 $Å^2$ on average. This should be compared with a random value of 45 $Å^2$ obtained by comparing residues from unrelated protein structures.

Most of the ASA changes occur as residues on the protein surface are replaced by others having larger or smaller side-chains. As expected, the changes are smaller at conserved positions (where no side-chain substitution occurs), especially in closely related proteins. When the two proteases (A and B) from *Streptomyces griseus* are compared, the average conserved residue changes its ASA by only 4·7 $Å^2$. This is little more than the difference observed in Table 3 for two sets of ribonuclease A co-ordinates. The two proteases have 65% homology in their core of equivalent residues. Changes in ASA of conserved residues can be large in distantly related proteins; however, for the dihydrofolate reductases of *Escherichia coli* and *Lactobacillus casei*, the homology is only 27% and the mean ASA change of the conserved residues is almost as large as that of the mutated residues.

(d) *The accessible surface area of monomeric proteins*

Table 5 gives the accessible surface areas of 37 monomeric proteins together with their molecular weight (M_r), which ranges from 4000 to 35,000. The very good correlation of A_s with the M_r, first observed by Chothia (1975), is confirmed here with our larger sample of generally better determined protein structures. It has been expressed in various analytical forms. Janin (1976) and Teller (1976) noted that protein ASA values are approximately given by:

$$A = aM_r^{2/3}. \qquad (1)$$

This equation fits the data of Table 5 to within 5% with $a = 11{\cdot}4$ $Å^2$. The two-thirds power law in equation (1), relating the surface area to the volume or mass, would apply to any set of solid bodies having similar shapes and densities. However, the accessible surface area A_s is not the same as the area of the molecular surface. It is measured one probe radius R_w (1·4 Å) away from that surface, and encloses half a layer of solvent in addition to the molecular volume (Gates, 1979). For instance, the A_s is not zero but $4\pi R_w^2$ or 25 $Å^2$ for a single point in space, and, for small molecules, it is significantly larger than the area of the molecular surface. If the latter followed a two-thirds power law, the slope of the log–log plot of A_s *versus* M_r should be less than 2/3. Observed values are somewhat larger: a log–log plot of our data (Fig. 1) is linear with a slope of 0·73 instead of 0·67, and a correlation coefficient of 0·993. The equation:

$$A_s = 6{\cdot}3\ M_r^{0{\cdot}73}$$

fits observed values in Table 5 to within 4% on average, and to better than 7% for all but four of them. Two of the exceptional cases, SGPB (*S. griseus* protease B) and *Pseudomonas aerogenes* ferredoxin, have an ASA 10 to 12% smaller than predicted by equation (2). Iron accounts for 4% of the molecular weight of this ferredoxin. At the other extreme, *Staphylococcus* nuclease and *S. platensis* ferredoxin have ASA values 9% larger than predicted.

The total accessible surface area of unfolded proteins A_t, is estimated by taking the polypeptide chain in an extended conformation and adding up the surface areas of individual residues given in Table 2, ignoring prosthetic groups. A_t varies linearly with the molecular weight of the polypeptide chain. A linear regression yields:

$$A_t = 1{\cdot}48\ M_r + 21. \qquad (3)$$

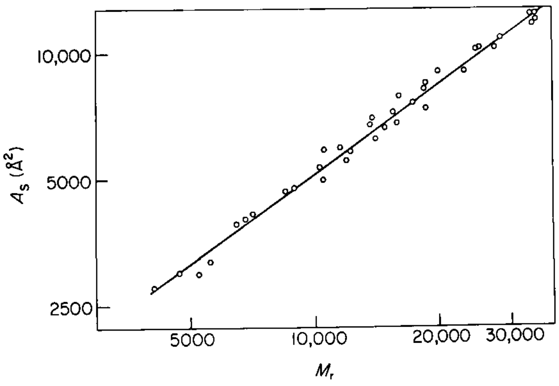

Figure 1. Accessible surface areas and molecular weight. A_s and molecular weight values from Table 5 are plotted on a log–log scale.

Table 5
Accessible surface areas of monomeric proteins

Protein	M_r	A_s (Å^2)	$(A_s-A)/A_s$ (%)	A_t (Å^2)
ICPA	4120	2760	2·3	5930
Crambin	4740	2990	0·0	6950
Ferredoxin, *P. aerogenes*	5550	2990	−12·4	7940
Rubredoxin	5610	3180	−6·6	8150
PTI	6520	3900	3·2	9650
Neurotoxin, snake	6870	4010	2·3	10,400
Neurotoxin, scorpion	7080	4140	3·2	10,380
ICBP	8500	4680	2·1	12,740
HIPIP	8960	4750	−0·3	13,230
Cytochrome *b*5	10,350	5330	0·9	14,500
Plastocyanin	10,480	4980	−7·0	15,570
Ferredoxin, *S. platensis*	10,550	5860	8·6	15,330
Parvalbumin	11,450	5930	4·0	17,060
Cytochrome *c*, tuna	11,930	5570	−5·3	17,080
Barnase	12,180	5910	−0·8	18,040
Ribonuclease A	13,690	6790	4·6	20,200
Cytochrome *c*3	13,760	6690	2·4	17,200
Phospholipase A2	13,800	7020	7·1	20,150
Azurin	14,050	6240	−5·9	20,930
Lysozyme, human	14,700	6620	−3·1	21,800
Erythrocruorin	15,340	6600	−6·5	22,150
Flavodoxin	15,790	6860	−4·8	22,800
Nuclease, *S. aureus*	15,990	7940	8·7	24,200
Myoglobin	17,300	7600	−1·1	25,200
DHFR, *E. coli*	18,460	8250	2·4	26,750
Lysozyme, phage T4	18,630	8530	4·9	27,950
SGPB	18,650	7410	−9·5	27,500
Retinol binding protein	20,050	9160	6·6	29,250
Papain	23,270	9140	−4·3	34,550
Protease II	24,530	10,300	3·8	36,800
Chymotrypsin	25,030	10,440	3·6	37,300
Subtilisin	27,540	10,390	−3·7	41,150
Carbonic anhydrase B	28,370	11,020	0·0	42,100
Penicillopepsin	33,460	12,640	1·7	49,200
Cytochrome peroxydase	33,930	11,920	−5·2	49,400
Carboxypeptidase A	34,450	12,110	−4·8	51,000
Thermolysin	34,500	12,650	−0·4	50,600

Molecular weights and accessible surface areas A_s include the contribution of prosthetic groups when present. A_t, the accessible surface area of unfolded proteins, is calculated without them. A is given by eqn (2).

This expression fits the values obtained for the proteins in Table 5 to within 1% on average. Note that the extrapolation to $M_r = 0$ yields a value close to the 25 Å^2 expected for the A_s of a single point in space.

As a consequence of equations (2) and (3), the A_s of native proteins increases more slowly with molecular weight than that of unfolded polypeptide chains. In the smaller proteins, 55% of the polypeptide accessible surface becomes buried upon folding. This fraction increases to 77% in the larger ones.

(e) *Hydropathy of accessible and buried surfaces*

We divided each protein into non-polar, polar and charged components and evaluated the accessible and buried surface area of each component. All carbon atoms are taken to be non-polar; nitrogen, oxygen and sulphur to be polar when they carry no electric charge, and charged in carboxylate, amino and guanidinium groups. Sulphur, which is much less polar in general than nitrogen or oxygen, and polar carbon atoms in carboxylate groups, contribute little to the surface areas compared to methyl and methylene groups. Thus, a very small error is introduced in taking S to be polar and C non-polar, independently of chemical environment.

The chemical character of the accessible surface of extended polypeptide chains depends only on the amino acid composition. Even though the amino acid compositions of the 37 monomeric proteins in our sample vary considerably, the fraction of the surface contributed by non-polar, polar and charged groups is almost constant in their unfolded states. On average, the accessible surface is 58% non-polar, 33% polar and 9% charged; the r.m.s. deviation from these average values is only 2%, and none deviates by more than 6%: the hydrophobicity of the unfolded polypeptides is the same for all these globular proteins.

The same fractions are plotted in Figure 2 for the native proteins. On average, their accessible surface is 57% non-polar, 24% polar and 19% charged. No

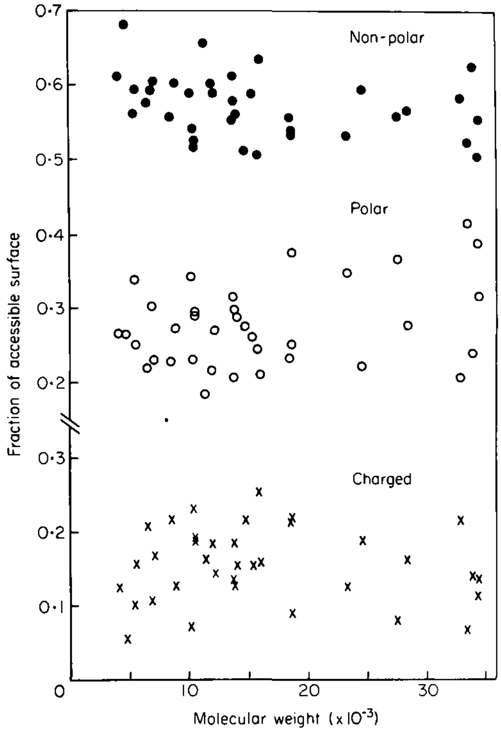

Figure 2. Hydrophobicity of the accessible surface of native proteins. The fraction of the A_s of each protein that is contributed by non-polar atoms, polar and charged groups is plotted against the protein molecular weight.

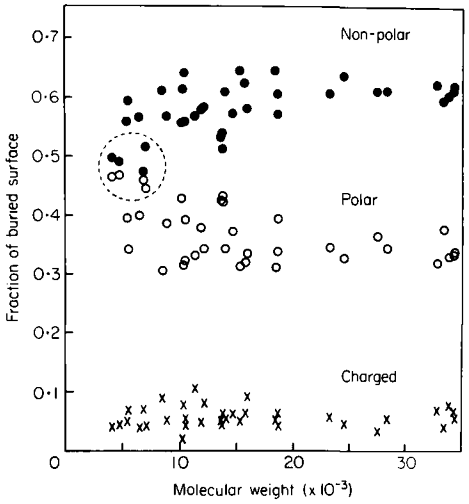

Figure 3. Hydrophobicity of the buried surface. The buried surface area is the difference between the ASA of the unfolded polypeptide chains and that of the native protein. The fraction of that surface that is contributed by non-polar, polar and charged atoms is plotted against the protein molecular weight. In 4 of the proteins of our sample, carboxy peptidase inhibitor (ICPA), crambin (1CRN) and snake and scorpion neurotoxins (1NXB and 1SN3) the contribution of polar atoms to the interior is unusually high and the contribution of non-polar atoms unusually low. The points corresponding to these structures are enclosed by a broken line. In these structures the interior contains cysteines that form disulphide bridges.

systematic dependence on molecular weight is observed. From one protein to another, the non-polar fraction varies between 50 and 68%, the largest value being that of crambin, a rather hydrophobic small protein. The r.m.s. fluctuation is 4%. The ratios of the uncharged polar fraction to the charged fraction vary more widely: uncharged polar groups form 41% of the accessible surface of penicillopepsin and charged groups only 7%, while charged groups contribute 25% to that of flavodoxin. The r.m.s. fluctuation of these fractions is 5%.

Non-polar groups contribute almost exactly as much to the accessible surface of native proteins as they do to that of unfolded polypeptides. Polar groups contribute significantly less, and charged groups twice as much. As a result, the surface that becomes buried when the chain folds has the same non-polar fraction as the surface that remains accessible, but contains more polar groups and very few charged ones (see also Lee & Richards, 1971). The mean values of the non-polar, polar and charged fractions of the buried surface are 58%, 39% and 4% (Fig. 3), and their r.m.s. fluctuations from the mean, 5%.

(f) *Interior and surface residues*

Few residues are buried to the extent that their ASA is effectively zero. The smaller proteins have only one or two completely buried residues. Even in the larger ones, only 15% of the residues are completely buried. Most of the protein interior is therefore composed of amino acid residues that make some contact with the solvent. We define the accessibility of individual residues as the ratio of their ASA measured in the protein structures to that given in Table 2 for an extended tripeptide.

A histogram of residue accessibilities is shown in Figure 4 for all 5436 residues of our sample. The origin peak, which drops sharply from 0 to 4% accessibility, contains buried residues. Beyond this value, residues are evenly distributed up to 50% accessibility. Most of the analysis below is done using a 5% accessibility cut-off as the definition of buried residues. The histogram of Figure 4 suggests that the definition of buried residues should not be sensitive to the cut-off value, within reasonable limits. We checked that very similar results are obtained with cut-offs of 2% and 8%, and also with a threshold applied to the ASA itself, as used by Janin (1979), instead of to the accessibility.

In our sample, the mean residue in an extended polypeptide chain has A_t = 162 Å^2. The mean ASA per residue in the native proteins is 48 Å^2, the mean residue accessibility 30%. Completely buried residues form 12% of our sample, buried residues with less than 5% accessibility 26%, and the remaining 74% of the residues form the protein

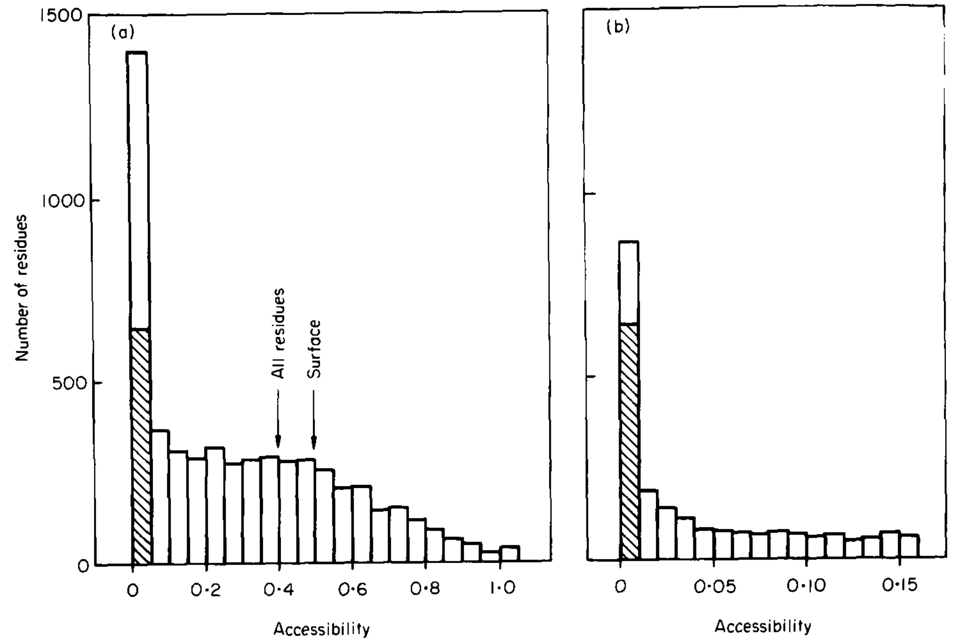

Figure 4. Histogram of residue accessibility. The accessibility of a given amino acid residue is the ratio of its ASA in a protein structure to that given in Table 2. The total number of residues in the 37 proteins is 5436. A detailed histogram is shown for residues with less than 15% accessibility. Residues with zero ASA are hatched. Mean values of the accessibility for all residues, and for all surface residues (with 5% accessibility or more), are indicated by arrows.

surface. The mean ASA of surface residues is 64 Å^2 and their mean accessibility 40%.

Table 6 shows that all these numbers except A_t depend on the protein molecular weight. In this Table, we divide the 37 proteins into four groups of average molecular weight 8500 (16 proteins), 16,000 (11 proteins), 25,000 (6 proteins) and 34,000 (4 proteins), in order to reduce the bias arising from the small number of residues in any individual protein. All four size groups contain approximately the same number of residues. Going from the first to the fourth size group, the proportion of buried residues doubles, while the mean residue ASA decreases by one-third from 58 to 40 Å^2. The mean ASA of surface residues decreases by only 15%.

The number N_b of buried residues has been shown to be related to the total number N_t of residues in the polypeptide chain by:

$$N_b^{1/3} = N_t^{1/3} - b. \tag{4}$$

This equation relates the volume of the protein interior, which is proportional to N_b, to the total volume of the protein, proportional to N_t (Janin, 1979). It makes the same assumptions on the protein shape as does equation (1), and it contains a single parameter b, which is related to the thickness of the layer of surface residues and to the accessibility "cut-off" used to define buried residues.

The 37 proteins of our sample fit equations with $b = 3·0 \pm 0·4$ (0% cut-off) or $b = 2·0 \pm 0·2$ (5% cut-off), as shown in Figure 5 where the lines are drawn according to equation (4). Observed and calculated values of N_b (5% cut-off) differ by less than 4 for all proteins with fewer than 100 residues except

Table 6
Buried and accessible residues

Mean M_r ($\times 10^{-3}$)	Fraction of buried residues			Mean ASA per residue (Å^2)		
	0%	2%	5%	A_t	A	A'
8	0·070	0·114	0·154	161	58·4	68·6
16	0·107	0·177	0·240	165	50·3	65·6
25	0·139	0·235	0·309	160	43·8	62·6
34	0·155	0·251	0·324	162	39·8	58·0
All	0·118	0·194	0·257	162	48·1	64·1

Groups of protein molecular weight are defined in the text. Buried residues have an accessibility lower than the limit indicated. A_t is the total ASA calculated on extended chains and divided by the total number of residues N_t. A is the ASA of the native proteins divided by N_t. A' is the same quantity divided by the number, $N_t - N_b$, of residues with more than 5% accessibility.

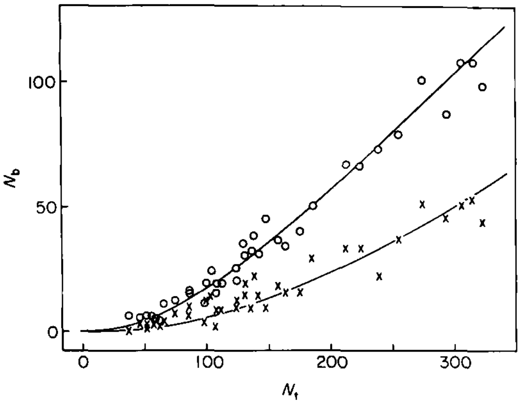

Figure 5. Buried residues. The number N_b of residues with zero accessibility (×) or with less than 5% accessibility (○) is plotted against the total number N_t of residues in each protein. The lines are calculated from eqn (4) with $b = 2{\cdot}0$ and $b = 3{\cdot}0$.

S. platensis ferredoxin, and by less than 12 for all proteins except penicillopepsin.

(g) *Very accessible residues: the N and C termini*

A few residues (39 out of 5436, or 0·7%) have accessibilities greater than unity. Some occur in sharp turns and have unusual conformations such that their ASA is slightly larger than in Gly-X-Gly extended tripeptides. The others are chain termini with free amino or carboxylate groups adding to their ASA. In most proteins, both the N and the C-terminal residues are very accessible to the solvent. The mean ASA of the N-terminal residues in our sample is 100 Å^2, that of the C-terminal residues 110 Å^2. In addition, three of the 37 proteins have disordered N termini and two have disordered C termini. Their accessibility is unknown, but it is certainly high.

Nearly all the excess accessible surface in the terminal residues belongs to the amino or the carboxylate group. The ASA of terminal amino groups is 33 Å^2 on average, or 62% of the maximum value observed (53 Å^2). Buried N termini do, however, occur in the serine proteases and in phospholipase A2, where they have functional roles. The *N*-acetyl-alanine terminus of parvalbumin is also largely buried. In contrast, no buried terminal carboxylate is found. The ASA values of the terminal carboxylate groups have a mean value of 53 Å^2, with a small standard deviation of 15 Å^2. A C-terminal glycine, with no side-chain preventing contacts between the carboxylate and the solvent, yields an isolated value of 96 Å^2 (Fig. 6).

The N termini of the proteins in our sample have an unusual amino acid composition, 76% of them being Ala, Ile, Lys or Met. The C termini are more evenly distributed, though Asn is in large excess (23%) at this position.

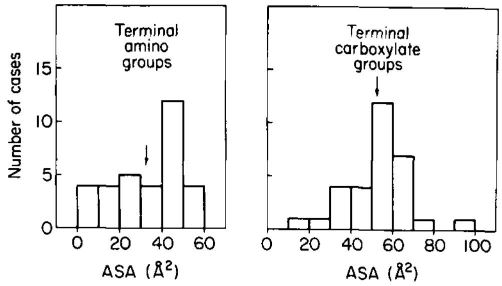

Figure 6. ASA of N and C-terminal groups. Histograms of the ASA of terminal amino and carboxylate groups. The mean values are indicated by arrows. No atomic coordinates are given in the Protein Data Bank for 3 N-terminal residues out of 37 and for 2 C-terminal ones, presumably because of disorder in the crystal structure.

(h) *Amino acid composition of the protein interior and surface*

The average amino acid composition of the protein interiors, formed by 1396 residues having less than 5% accessibility, is compared in Table 7 to that of the protein surface formed by 4040 residues having larger accessibilities. The data confirm the well-established fact that the protein interior is enriched in large aliphatic and aromatic residues, and the protein surface is enriched in charged residues.

The amino acids Val, Leu, Ile and Phe constitute 44% of the interiors of the proteins, but only 14% of their surfaces. Although the proportion of each varies widely from one protein interior to another, the total fraction comprising these four residues is quite constant in proteins larger than 15,000 molecular weight, the r.m.s. deviation from the mean being less than 7%. On average, Val, Leu, Ile and Phe are as abundant inside the smaller proteins (8×10^3 M_r sample) as in the larger ones. However, cystines form more than half of the residues buried in ICPA (see Table 1), crambin and pancreatic trypsin inhibitor (PTI), and cysteine–metal complexes form one-third of the buried residues in ferredoxins and rubredoxins. The cores of these proteins are therefore made of cysteine derivatives rather than of non-polar amino acid side-chains. Larger proteins contain few cystines or cysteines (Fig. 7(a)). Taken together, Val, Leu, Ile, Phe and Cys constitute more than 37% of the interior of all 37 proteins, and more than 50% in 24 of them (Fig. 7(b)).

Other residues abundant inside proteins are Ala and Gly (21% of the interior residues), and Ser and Thr (10%), but these are also very common on the surface. On the other hand, the protein surface is rich in charged residues, which are excluded from the interior. On average, Asp, Glu, Lys and Arg constitute 27% of the protein surface and only 4% of the interior (Fig. 7(c)).

Table 7

Amino acid composition of protein interior and surface

	All			8			16			25			34		
Residue	Total	Inside	Surface	Total	Inside	Surface	Total	Inside	Surface	Total	Inside	Surface	Total	Inside	Surface
Ala	8·7	11·0	7·9	9·3	8·2	9·5	9·0	10·1	8·7	9·1	13·1	7·3	7·3	11·0	5·6
Arg	3·1	0·4	4·0	2·5	0·5	2·8	3·9	0·3	5·1	3·2	0·5	4·4	2·5	0·5	3·5
Asn	5·2	2·0	6·3	4·8	3·1	5·1	5·2	1·9	6·2	5·2	1·4	6·9	5·6	2·2	7·2
Asp	6·1	2·2	7·4	6·8	1·5	7·8	6·1	1·6	7·5	4·6	1·9	5·9	6·9	3·5	8·6
Cys	2·7	5·4	1·8	5·4	18·6	3·0	2·8	6·1	1·8	2·1	3·5	1·5	0·4	0·2	0·5
Gln	3·6	1·3	4·5	3·2	1·0	3·6	2·7	0·5	3·4	4·2	1·6	5·4	4·7	1·7	6·1
Glu	4·9	1·0	6·2	6·6	0·5	7·7	5·6	0·0	7·4	3·5	1·2	4·5	3·7	2·0	4·5
Gly	9·0	9·7	8·8	8·2	5·2	8·7	8·8	9·3	8·6	9·3	11·5	8·3	9·9	10·2	9·8
His	2·3	2·4	2·2	1·7	2·1	1·7	2·8	3·2	2·7	2·3	2·3	2·3	2·0	1·7	2·1
Ile	4·9	10·5	3·0	4·9	9·8	4·0	5·1	13·1	2·6	4·8	9·6	2·6	4·9	9·2	2·9
Leu	6·5	12·8	4·3	5·7	16·0	3·9	6·6	14·1	4·2	6·7	11·0	4·7	6·9	11·7	4·5
Lys	6·7	0·3	8·9	7·9	0·0	9·4	8·5	0·0	11·1	5·4	0·7	7·6	4·4	0·2	6·4
Met	1·5	3·0	0·9	0·9	1·5	0·8	2·6	4·8	1·9	1·3	2·8	0·6	0·8	2·2	0·1
Phe	3·8	7·7	2·5	3·9	9·3	2·9	3·6	9·1	1·9	2·9	3·8	2·5	5·2	9·7	3·0
Pro	4·0	2·2	4·7	4·7	3·6	4·9	2·9	1·3	3·4	5·0	3·1	5·9	3·7	1·5	4·8
Ser	7·9	5·0	8·9	7·1	1·5	8·1	5·9	3·2	6·8	9·4	6·6	10·7	9·6	6·7	11·0
Thr	6·4	4·6	7·1	6·3	3·1	6·9	6·3	4·5	6·8	5·7	4·0	6·4	7·7	6·0	8·5
Trp	1·6	2·7	1·3	1·1	3·1	0·8	1·6	2·4	1·4	2·1	2·8	1·8	1·6	2·5	1·2
Tyr	4·4	3·3	4·8	4·1	2·6	4·3	3·5	3·5	3·5	4·2	2·6	4·9	5·0	4·2	6·8
Val	6·6	12·7	4·6	4·9	8·8	4·2	6·5	10·9	5·1	8·9	16·0	5·8	6·1	12·7	2·9
Number	5436	1396	4040	1258	194	1064	1560	375	1185	1379	426	953	1239	401	838

Percentage amino acid composition of the proteins, their interior (residues with less than 5% accessibility), and their surface (residues with more than 5% accessibility). The first data set contains all 37 proteins in our sample. The size groups around molecular weights ($\times 10^{-3}$) 8, 16, 25 and 34 are defined in the text.

Table 8

Partition of residues between the protein surface and interior

	Surface/interior transfer free energies (kcal mol^{-1})					
	Set 1					
Residue	All	S.D.	Set 2 (8)†	Set 3 (16)	Set 4 (25)	Set 5 (34)
Ala	0·20	0·06	−0·1	0·1	0·3	0·4
Arg	−1·34	0·25	−1·0	−1·8	−1·3	−1·3
Asn	−0·69	0·12	−0·3	−0·7	−1·0	−0·7
Asp	−0·72	0·11	−1·0	−0·9	−0·7	−0·5
Cys	0·67	0·10	1·1	0·7	0·5	−0·4
Gln	−0·74	0·15	−0·7	−1·1	−0·7	−0·7
Glu	−1·09	0·17	−1·6	−2·0	−0·8	−0·5
Gly	0·06	0·06	−0·3	0·0	0·2	0·0
His	0·04	0·12	0·1	0·1	0·0	−0·1
Ile	0·74	0·08	0·5	1·0	0·8	0·7
Leu	0·65	0·07	0·9	0·7	0·5	0·6
Lys	−2·00	0·30	−2·0	−2·0	−1·4	−2·0
Met	0·71	0·14	0·4	0·6	0·9	1·8
Phe	0·67	0·09	0·7	1·0	0·2	0·7
Pro	−0·44	0·12	−0·2	−0·6	−0·4	−0·7
Ser	−0·34	0·08	−1·0	−0·4	−0·3	−0·3
Thr	−0·26	0·09	−0·5	−0·2	−0·3	−0·2
Trp	0·45	0·13	0·8	0·3	0·3	0·4
Tyr	−0·22	0·10	−0·3	0·0	−0·4	−0·3
Val	0·61	0·07	0·4	0·5	0·6	0·9
N terminus	−1·25	0·32				
C terminus	(−2·0)					

Free energies are calculated from eqn (6) using data of Table 7 for all residues in 37 proteins (set 1) or in 4 size groups (sets 2 to 5). The value given for the N terminus is equal to $-RT \ln (N_a/N_b)$, where N_a and N_b are the number of accessible and buried terminal amino groups. As we have no buried carboxylate, the value given for the C terminus is arbitrarily set to −2·0. Standard deviations for set 1 are derived from statistics by assuming that the variance of each number N is equal to that number. Standard deviations for values in the other sets are about twice as large.

† M_r ($\times 10^{-3}$).

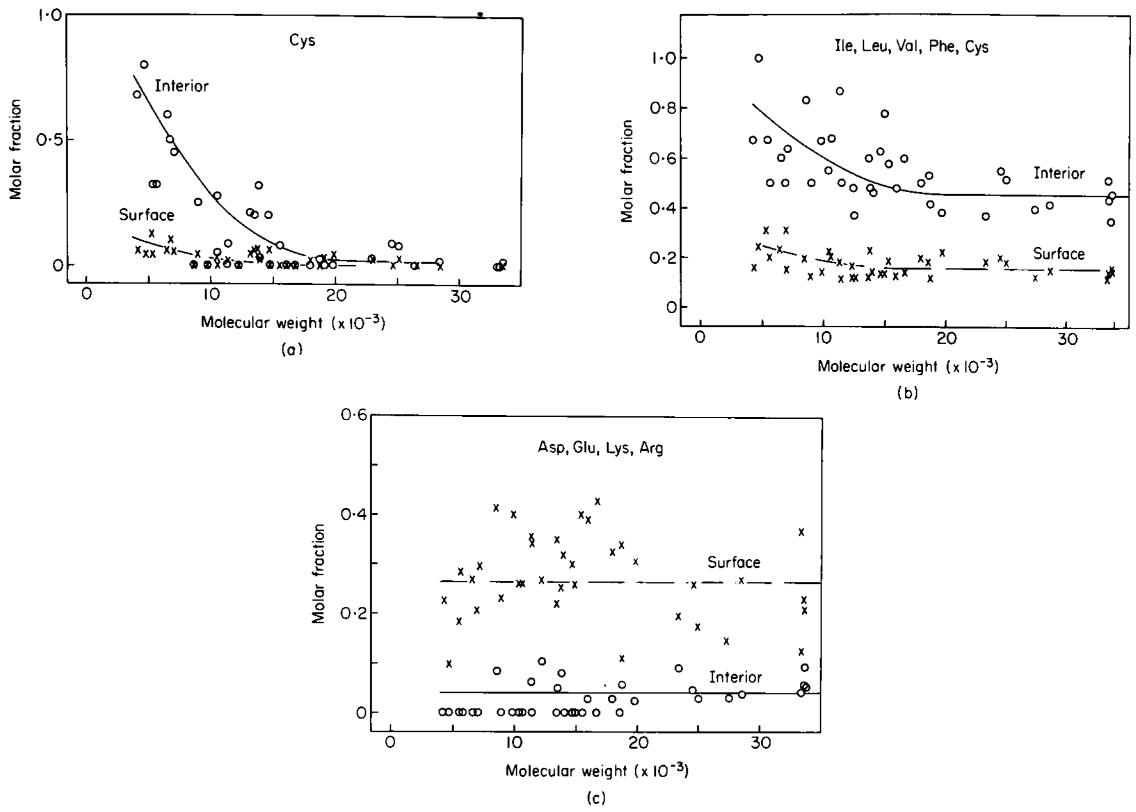

Figure 7. Amino acid compositions of the protein interior and surface. The molar fraction of Cys among buried or accessible residues (5% accessibility threshold) is plotted in (a) against the protein molecular weight. Cysteine residues and derivatives are much less abundant in proteins of more than 15,000 molecular weight than in smaller proteins, where they contribute largely to the interior. The plot in (b) represents the sum of the molar fractions of Ile, Leu, Val, Phe and Cys residues, which, together, represent about half of the protein interior. The plot in Fig. 7(c) represents the sum of the molar fractions of Asp, Glu, Lys and Arg residues, which are found almost exclusively on the surface.

(i) *Partition coefficients and transfer free energies*

The distribution of the amino acid residues between the protein interior and its surface can be described by an effective partition coefficient. However, we have seen that the ratio N_b/N_s of the number of buried to accessible residues changes with the protein size. In contrast, the average amino acid compositions of the protein interior and surface are essentially the same for all four groups of molecular weight (Table 7), with the exception of Cys. Thus, we use these amino acid compositions to estimate partition coefficients in the following way: if there are N_b buried residues of a given type out of a total of ΣN_b buried residues of all types, and N_s surface residues of the type out of a total of ΣN_s surface residues, the partition coefficient is:

$$f = \frac{N_s/\Sigma N_s}{N_b/\Sigma N_b}. \tag{5}$$

It is converted into a notional free energy of transfer through the usual expression:

$$\Delta G_t = -RT \ln f, \tag{6}$$

where $RT = 0{\cdot}596$ kcal mol^{-1} at 27 °C. ΔG_t can be viewed as the free energy of transfer from inside the protein to its surface. Values of ΔG_t derived from the data of Table 7 are reported in Table 8. Set 1 is obtained with all data in our sample; the other sets with the four protein size groups taken individually. The five different sets of ΔG_t values are closely correlated: the standard deviation of the ΔG_t values for individual amino acids among the sets is 0·27 kcal mol^{-1}. This variation in the values of ΔG_t obtained for the same residue type is probably a more realistic estimate of the errors affecting these values than standard deviations calculated on purely statistical grounds and reported in Table 8. The largest discrepancy between size groups affects Cys: its ΔG_t value drops from 1·1 kcal mol^{-1} in the smaller proteins to a negative value in larger ones. The reason for that behaviour is that several of the small proteins have interiors rich in disulphide bridges (see above, and Fig. 3). Certain other residues tend to do the opposite (Ala, Glu and Gly), but the effects are much smaller.

We have also calculated surface–interior partition coefficients and free energies of transfer for different values of the threshold chosen to distinguish buried from accessible residues. A set of ΔG_t values obtained with a 2% accessibility cut-off is linearly correlated to set 1 with a correlation coefficient of 0·994. Individual values change by 0·12 kcal mol^{-1} on average; none changes by more than 0·2 kcal mol^{-1}. With a 10% cut-off, the correlation coefficient is 0·988 and the changes are of the same order. A scale calculated with an ASA cut-off of 5 Å^2, instead of an accessibility cut-off, yields a correlation coefficient of 0·987 with set 1. Thus, the scale of surface/interior transfer free energies is insensitive to the precise definition of buried residues.

(j) *Comparison with other scales of transfer free energies*

We compared the ΔG_t values in set 1 to data previously published by Chothia (1976) and Janin (1979). Chothia (1976) used a 5% accessibility threshold as we do here. However, he calculated partition coefficients from the fraction of buried residues, not from the amino acid compositions. Janin (1979) used amino acid compositions, and a threshold of 20 Å^2 ASA. After linear rescaling to set 1, his set of ΔG_t values yields a correlation coefficient of 0·97 and shows no difference larger than 0·3 kcal mol^{-1}. Values calculated from the data published by Chothia (1976) correlate less well to set 1 (correlation coefficient 0·89) and show significant differences for Glu, Lys and Arg.

There is a correlation between the accessible surface area of amino acids and their free energy of transfer from water to organic solvents (Chothia, 1974; Frommel, 1984; Eisenberg & McLachlan, 1985). Figure 8 compares the behaviour of residues in proteins (set 1 in Table 8) with free-energy scales derived from transfer experiments between water and non-polar organic solvents or between water and the vapour phase. Fauchère & Pliska (1983) quote values of water/solvent ΔG_t for the more hydrophilic residues, which were missing in earlier data collected by Nozaki & Tanford (1971). Figure 8(a) shows the general similarity between their values, put on a free-energy scale, and set 1. For hydrophilic residues, the two sets can be considered as identical within their respective error limits, except for Tyr and Pro, which appear much more frequently on the protein surface than expected from their solution properties. For hydrophobic residues the correlation is poor: Cys, Val, Ile, Leu, Phe, Met and Trp have nearly the same ΔG_t value in set 1, but their water/solvent transfer free energies vary from 1·5 to 3·1 kcal mol^{-1} (Fig. 8(a)).

Figure 8(b) compares set 1 to the results of water/vapour transfer experiments (Wolfenden *et al.*, 1981; Wolfenden, 1983). A linear correlation coefficient of 0·85 is obtained for 18 data points excluding Lys, for which there is an obvious discrepancy, and Pro, absent from Wolfenden's data. Again if the hydrophobic residues Cys, Val, Ile, Leu, Met, Phe and Trp are considered by themselves, the correlation is poor. It should also be noted that the free energies of water/vapour transfer vary by more than 20 kcal mol^{-1},

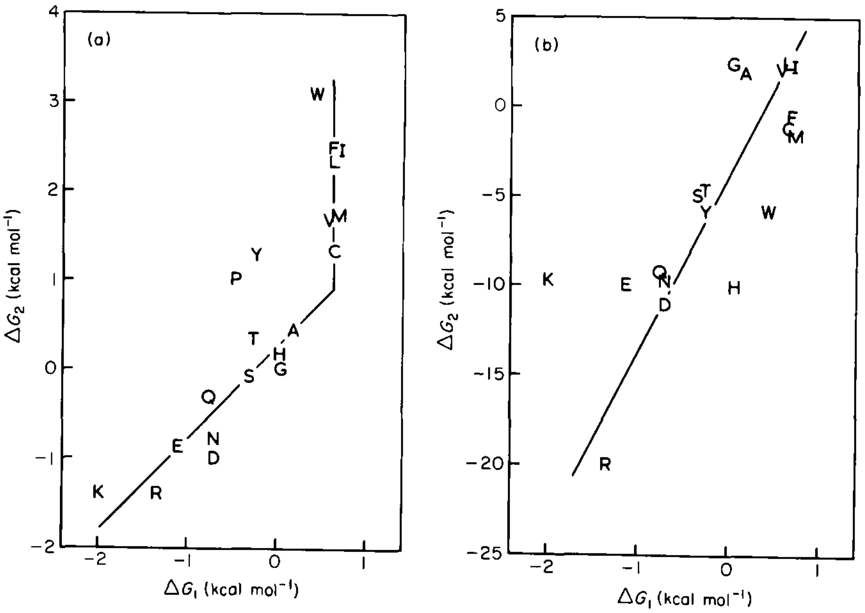

Figure 8. Correlation between transfer free energies. Water/organic solvent transfer free energies ΔG_2, quoted by Eisenberg & McLachlan (1985), are derived from the work of Fauchère & Pliska (1983). They are plotted in (a) against surface/interior transfer free energies ΔG_1 from Table 8 (set 1). The straight line corresponds to the equation: $\Delta G_2 = \Delta G_1 + 0{\cdot}22$ kcal mol^{-1}, with a linear correlation coefficient of 0·93 for 13 data points. A linear regression on all 20 data points yields a correlation coefficient of 0·87. Eisenberg & McLachlan (1985) propose their own scale of ΔG_t, which is closely related to that of Fauchère & Pliska (1983) and somewhat less to ours (correlation coefficient of 0·79). Amino acids are designated by the one-letter code. In (b), water/vapour phase transfer free energies from Wolfenden *et al.* (1981) are plotted against the same ΔG_1 values. The regression line yields: $\Delta G_2 = 10{\cdot}9\ \Delta G_1 - 4{\cdot}4$ kcal mol^{-1}, with a correlation coefficient of 0·853.

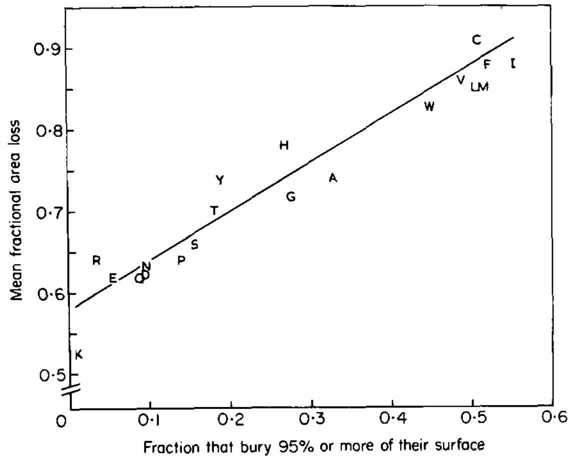

Figure 9. For each residue we plot the mean fractional area loss (F) as determined by Rose *et al.* (1985*b*) against the fraction that bury 95% or more of their surface (B) (see Table 7). The correlation coefficient between the 2 measures is 0·96. The relationship is expressed by the formula: $F = 0{\cdot}58B + 0{\cdot}58$.

compared with less than 3 kcal mol^{-1} for the free energies of surface/interior transfer and 4·5 kcal mol^{-1} for the solvent transfer. The relationships between the values are therefore unclear, despite the correlations.

(k) *Comparison with the results of Rose* et al. *(1985*b*)*

Rose *et al.* (1985*a*,*b*) have argued that the definition of buried residues used here and in our previous work (those that bury 95% or more of their surface) is too stringent. They proposed two different measures of residue accessibility, which, they assert, better illuminate the relationship between the behaviour of residues in proteins and experimental free energies of transfer. In this section we show that (1) one of their proposed scales is in fact so similar to ours (correlation coefficient 0·96) that it could not yield significant new insights, and (2) that the claimed correlation of their other scale with free energies of transfer is not observed unless the data are arbitrarily restricted to those 11 side-chains included in the Nozaki & Tanford (1971) study.

Rose *et al.* (1985*b*) propose two measures of residue accessibility: (1) the mean area lost by residues when proteins go from the unfolded state to the folded. For any particular residue this quantity is the difference between the average accessible surface area of the residue in folded proteins and that calculated for the residue in unfolded proteins. (2) The fractional accessibility of a residue. This is the mean accessible surface area of a residue in native proteins divided by its accessible surface area in the unfolded state.

Rose *et al.* (1985*b*) used the fractional loss of accessibility by a residue as a measure of the extent to which residues are buried. The equivalent measure in our work is the fraction that bury 95% or more of their surface. In Figure 9 we plotted for each residue the fractional accessibility as determined by Rose *et al.* (1985*b*) and the fraction that bury 95% or more of their surfaces as determined in the work reported here. The correlation coefficient of the two sets of measurements is very high: 0·96. If in place of our data we use that of Chothia (1976), who also defined buried residues as those

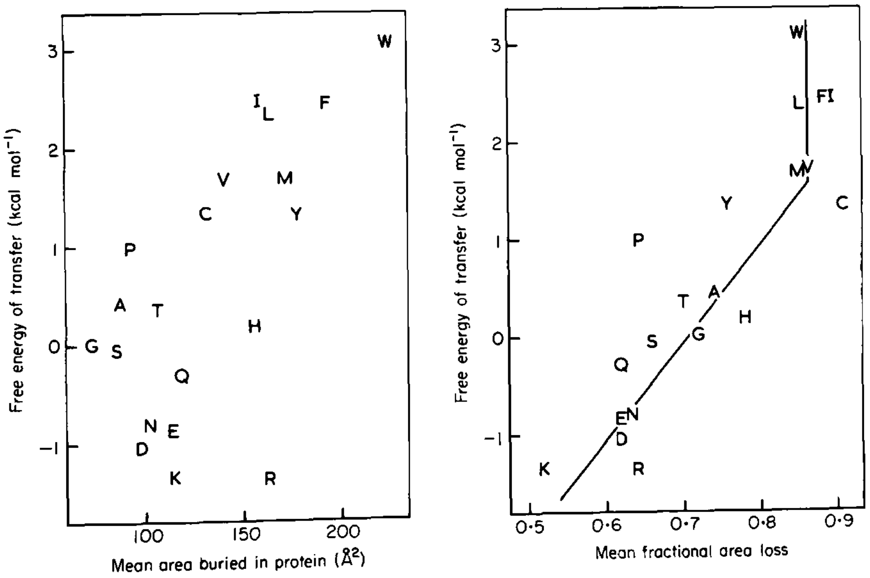

Figure 10. The free energy change of the transfer of amino acid side-chains from water to organic solvent (Fauchère & Pliska, 1983) are plotted against (a) the mean area residues bury in proteins and (b) the mean fractional area loss of residues (Rose *et al.*, 1985*b*). See the text. Note the similarity of (b) to Fig. 8(a).

with 95% or more of their surface buried, or that of Janin (1979), who defined buried residues as those with 20 Å or less accessible surface, the correlation coefficients are 0·86 and 0·91.

The high correlation between these two measurements of the extent to which residues are buried implies that they must lead to similar descriptions of the behaviour of residues in proteins. For example, from the values they obtain for mean fractional area loss, Rose *et al.* (1985*b*) showed that if residues are divided into hydrophobic, moderately polar and very polar groups the residues within each group are buried to a similar extent. The same result was obtained from our measure of the extent to which residues are buried (Chothia, 1976; Fig. 9).

Rose *et al.* (1985*b*) also claimed that the mean area lost by residues as proteins go from the unfolded to the folded state is strongly correlated with the experimental hydrophobicities derived from transfer experiments. However, if the values from all 20 residues are used, rather than the 11 used by Rose *et al.* (1985*b*), the correlation is clearly poor (Fig. 10(a)). (In producing this graph we use the free energy data of Fauchère & Pliska (1983), which are in good agreement with the data of Nozaki & Tanford (1971) for the 11 amino acids common to both studies.) Moreover, the residue property that is logically related to the transfer experiments is not the loss of area by residues but the distribution of residues between the surface and interior of proteins. Rose *et al.* (1985*b*) use the mean fractional area loss to measure this latter quantity; and, indeed, a fit of their values to the experimental scale of Fauchère & Pliska (1983) gives a correlation coefficient of 0·88 (Fig. 10(b)). As with our data, the correlation is good for the hydrophobic, moderately polar and very polar groups of residues but poor for residues within these groups (compare Figs 8(a) and 10(b)).

3. Discussion and Conclusion

(a) *The accessible surface area of proteins*

The data presented here are limited to monomeric proteins in the molecular weight range 4000 to 35,000. This limitation is imposed by our having confined our analysis to high-resolution X-ray structures that have been submitted to crystallographic refinement.

The A_s of these monomeric proteins deviates by only 4% on average from values calculated from molecular weights using equation (2), although we estimate that they are measured to within 1 to 2%. Some of the residuals arise from inaccuracies in atomic positions, which tend to increase the observed A_s. For example, the A_s of phage T4 lysozyme calculated from 2·4 Å resolution unrefined co-ordinates (file 1LZM of the Protein Data Bank) is 5% larger than the one reported here, where we used a structure refined by Weaver & Matthews (1987) at 1·7 Å resolution. For a given molecular weight, the observed A_s values are nearly constant in spite of the wide variety of shapes, amino acid compositions and general properties of the proteins studied. Remarkably, proteins such as phage T4 lysozyme, papain or thermolysin, which fold into well-defined domains, have A_s values that are not significantly different from those of single-domain proteins of the same size.

In multi-domain proteins, the correlation observed here between the A_s of native proteins and their molecular weight is found to valid for individual domains (Wodak & Janin, 1981), although smaller fragments, or fragments of single domain proteins, have larger A_s values than predicted from the correlation. Globular domains with low A_s values are close-packed and, in some cases, they have been shown to be stable when dissected from the rest of the protein (Vita *et al.*, 1984).

The basis for the correlation is assumed to be the hydrophobic effect and the close packing of protein interiors, which tend to reduce the polypeptide surface in contact with the solvent to stabilize a compact globular structure. In computer simulations of protein folding where this effect is not taken into account, incorrect structures with low conformational energies can be obtained. These incorrect structures have large A_s values (Novotny *et al.*, 1984).

The hydrophobicity of the accessible and buried surfaces is fairly constant from one protein to another, fluctuations about the average values being less than 5% r.m.s. Both surfaces are formed mostly (57 to 58%) by non-polar aliphatic and aromatic groups. The accessible surfaces of native proteins contain both charged and uncharged polar groups, in proportions that vary widely, yet the balance of non-polar groups on the protein surface is maintained rather strictly. It probably governs the solubility of proteins.

A large amount of uncharged polar surface is buried when proteins fold, and very little charged surface. Charged side-chains and chain terminal residues are generally excluded from the protein interior. In contrast, more uncharged polar surface is buried than remains accessible. It is formed mainly by peptide groups that hydrogen bond to form secondary structures.

The A_s of native proteins increases more slowly with their molecular weight than that of unfolded polypeptide chains. As a consequence, the number of amino acid residues that become buried when the chain folds increases with the protein size. An interior is formed. It contains only a few residues in the smaller proteins, and over 100 residues in the larger ones, about one-third of the total.

(b) *Residue accessibilities and hydrophobicities*

In this paper we have defined buried residues as those that bury within the protein interior 95% or more of their surface. This particular definition is related to the observation, derived from the analysis of homologous protein structures, that a discrete set of residues forms the conserved hydrophobic interior, as Figure 4 suggests. In the different globins, for example, there are 35 sites at which the residues are consistantly hydrophobic. The residues at these sites have accessible surface areas of less than 10 Å^2 (Lesk & Chothia, 1980). However, quite different definitions, for example the use of mean accessibilities by Rose *et al.* (1985*b*), give similar descriptions of the behaviour of residues in proteins (see Methods and Results, section (k)). Thus our general results and conclusions do not depend upon the exact nature of the definition used.

From the amino acid compositions of the protein surface and interior, we derive the hydrophobicity scale presented in Table 8 (set 1). In this scale, Gly residues have a ΔG_t value near zero, not because the scale has been adjusted to bring it to zero, as usual with hydrophobicity scales, but because Gly is equally frequent (9 to 10%) inside proteins and at their surface. The actual value of ΔG_t for Gly shifts by about 0·07 kcal mol^{-1} in either direction when the threshold of accessibility distinguishing surface from buried residues is changed from 5% to 2% or 10%. ΔG_t values of other residues shift by the same amount, so that scales obtained with different cutoffs are essentially identical except for a small origin shift.

In set 1, the less-polar residues are Ile, Leu, Val, Phe, Met and Cys, at $\Delta G_t \approx 0{\cdot}7$ kcal mol^{-1}. Residues with hydroxyl groups (Ser, Thr, Tyr) cluster at about $-0{\cdot}25$ kcal mol^{-1}. Proline is placed between these residues and the two amide residues (Asn and Gln), which are near $-0{\cdot}7$ kcal mol^{-1}. All charged residues (Asp, Glu, Arg, Lys, N and C termini) are below $-0{\cdot}7$ kcal mol^{-1}.

Trytophan is peculiar. In the set published by Janin (1979), its free energy of transfer was close to that of Gly. In this set it is 0·4 kcal mol^{-1}, making Trp non-polar, though less so than aliphatic residues. This discrepancy can be attributed partly to effects of sampling, the number of Trp residues in our set of proteins being small. It also reflects a change in the definition of buried residues, from an ASA threshold of 20 Å^2 (Janin, 1979) to a definition based on accessibility. Note that 20 Å^2 is equivalent to a 23% accessibility for Gly, but only 8 to 9% for Trp. There are fewer buried Gly with the present 5% cut-off, and about the same fraction of buried Trp residues. Gly and Trp are extreme cases. The effect of other residues is much smaller than our estimate of the error affecting ΔG_t values. This point being made, set 1 of Table 9 can be considered as an improved version of the previous set (Janin, 1979), with which it is well correlated.

The correlation of set 1 with free energies derived from water/organic solvents or water/vapour experiments is high: the correlation coefficient with the data of Fauchère & Pliska (1983) is 0·87 and that with the data of Wolfenden *et al.* (1981) is 0·85. These correlation coefficients are high because the extent to which charged, polar and non-polar residue groups are buried within a protein reflects their hydrophobicity. There is not a good correlation between the extent to which residues *within* these groups are buried and their hydrophobicity (Fig. 8 and above). This is why the correlation coefficient with the data of Nozaki & Tanford (1971), which do not include the charged and more polar amino acids, is only 0·52. The partition of residues between the interior and the surface of a protein is governed by factors other than just their relative solubility in the two different environments, even though that plays the major role (Janin & Chothia, 1980; Chothia, 1984).

We thank John Cresswell for the figures, Professors Guy Dodson, J. Drenth, T. A. Jones, Brian Matthews and David Phillips for sets of atomic co-ordinates, and the Royal Society, the National Institute of General Medical Sciences (GM25435) and the National Science Foundation (PCM83-20171) for support.

References

Adman, E. T. & Jensen, L. H. (1979). *Amer. Crystallogr. Assoc.* Abstr., vol. 6, p. 65.

Adman, E. T., Sieker, L. C. & Jensen, L. H. (1973). *J. Biol. Chem.* **248**, 3987–3996.

Alden, R. A., Birktoft, J. J., Kraut, J., Robertus, J. D. & Wright, C. S. (1971). *Biochem. Biophys. Res. Commun.* **45**, 337–344.

Almassy, P. J., Fontecilla-Camps, J. C., Sudath, F. L. & Bugg, C. E. (1983). *J. Mol. Biol.* **170**, 497–527.

Artymiuk, P. J. & Blake, C. C. F. (1981). *J. Mol. Biol.* **152**, 737–762.

Baker, E. N. & Dodson, E. J. (1980). *Acta Crystallogr. sect. A*, **36**, 559–572.

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

Bolin, J. T., Filman, D. J., Matthews, D. A., Hamlin, R. C. & Kraut, J. (1982). *J. Biol. Chem.* **257**, 13650–13662.

Borkakoti, N., Moss, D. S. & Palmer, R. A. (1982). *Acta Crystallogr. sect A*, **38**, 2210–2217.

Carter, C. W., Jr, Kraut, J., Freer, S. T., Xuong, N. H., Alden, R. A. & Bartsch, P. G. (1974). *J. Biol. Chem.* **249**, 4212–4225.

Chothia, C. (1974). *Nature (London)*, **248**, 338–339.

Chothia, C. (1975). *Nature (London)*, **254**, 304–308.

Chothia, C. (1976). *J. Mol. Biol.* **105**, 1–12.

Chothia, C. (1984). *Annu. Rev. Biochem.* **53**, 537–572.

Chothia, C. & Lesk, A. M. (1986). *EMBO J.* **5**, 823–826.

Cotton, F. A., Hazen, E. E., Jr & Legg, M. J. (1979). *Proc. Nat. Acad. Sci., U.S.A.* **76**, 2551–2555.

Deisenhofer, J. & Steigemann, W. (1974). In *Proteinase Inhibitors* (Fritz, H., Tschesche, H., Greene, L. J. & Truscheit, E., eds), Bayer Symp., Vol. 5, pp. 484–496, Springer-Verlag, Berlin.

Dijkstra, B. W., Kalk, K. H., Hol, W. G. J. & Drenth, J. (1981). *J. Mol. Biol.* **147**, 97–123.

Eisenberg, D. & McLachlan, A. D. (1985). *Nature (London)*, **319**, 199–203.

Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. (1982). *Nature (London)*, **299**, 371–374.

Fauchère, J. L. & Pliska, V. (1983). *Eur. J. Med. Chem.-Chim. Ther.* **18**, 369–375.

Finzel, B. C., Poulos, T. L. & Kraut, J. (1984). *J. Biol. Chem.* **259**, 13027–13036.

Frommel, C. (1984). *J. Theor. Biol.* **111**, 247–260.

Gates, R. (1979). *J. Mol. Biol.* **127**, 345–351.

Grace, D. C. P. (1979). D. Phil. thesis, Oxford University.

Guss, J. M. & Freeman, H. C. (1983). *J. Mol. Biol.* **169**, 521–563.

Hendrickson, W. A. & Teeter, M. M. (1981). *Nature (London)*, **290**, 107–113.

Higuchi, Y., Kusunoki, M., Matsuura, Y., Yasuoka, N. & Kakudo, M. (1984). *J. Mol. Biol.* **172**, 109–139.

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

James, M. N. G. & Sielecki, A. R. (1983). *J. Mol. Biol.* **163**, 299–361.

Janin, J. (1976). *J. Mol. Biol.* **105**, 13–14.

Janin, J. (1979). *Nature (London)*, **277**, 491–492.

Janin, J. & Chothia, C. (1980). *J. Mol. Biol.* **143**, 95–128.

Janin, J., Wodak, S., Levitt, M. & Maigret, B. (1978). *J. Mol. Biol.* **125**, 357–386.

Kamphuis, I. G., Drenth, J. & Baker, E. N. (1985). *J. Mol. Biol.* **182**, 317–329.

Kannan, K. K., Ramanadham, M. & Jones, T. A. (1984). *Ann. N.Y. Acad. Sci.* **429**, 49–61.

Kauzmann, W. (1959). *Advan. Protein Chem.* **16**, 1–63.

Kyte, J. & Doolittle, R. F. (1982). *J. Mol. Biol.* **157**, 105–132.

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

Lesk, A. M. & Chothia, C. (1980). *J. Mol. Biol.* **136**, 225–270.

Matthews, F. S., Argos, P. & Levine, M. (1972). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 387–395.

Mauguen, Y., Hartley, R. N., Dodson, E. J., Dodson, G. C., Bricogne, G., Chothia, C. & Jack, A. (1982). *Nature (London)*, **297**, 162–164.

Moews, P. C. & Kretsinger, R. H. (1975). *J. Mol. Biol.* **91**, 201–228.

Newcomer, M. E., Jones, T. A., Åqvist, J., Sundelin, J., Eriksson, U., Rask, L. & Pederson, P. A. (1984). *EMBO J.* **3**, 1451–1454.

Norris, G. E., Anderson, B. F. & Baker, E. N. (1983). *J. Mol. Biol.* **165**, 501–521.

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

Nozaki, Y. & Tanford, C. (1971). *J. Biol. Chem.* **246**, 2211–2217.

Ochi, H., Hata, Y., Tanaka, N., Kakudo, M., Sakurai, T., Aihara, S. & Morita, Y. (1983). *J. Mol. Biol.* **166**, 407–418.

Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965). *J. Mol. Biol.* **13**, 669–678.

Phillips, S. E. V. (1980). *J. Mol. Biol.* **142**, 531–554.

Read, R. J., Fujinaga, M., Sielecki, A. R. & James, M. N. G. (1983). *Biochemistry*, **22**, 4420–4433.

Rees, D. C. & Lipscomb, W. N. (1982). *J. Mol. Biol.* **160**, 475–498.

Rees, D. C., Lewis, M. & Lipscomb, W. N. (1983). *J. Mol. Biol.* **168**, 367–387.

Reynolds, R. A., Remington, S. J., Weaver, L. H., Fisher, R. G., Anderson, W. F., Ammon, L. H. & Matthews, B. W. (1985). *Acta Crystallogr. sect. B*, **41**, 139–147.

Rose, G. D., Gierasch, L. M. & Smith, J. A. (1985*a*). *Advan. Protein Chem.* **37**, 1–109.

Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. & Zehfus, M. H. (1985*b*). *Science*, **229**, 834–838.

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

Sielecki, A. R., Hendrickson, W. A., Broughton, C. G., Delbaere, L. T. J., Brayer, G. D. & James, M. N. G. (1979). *J. Mol. Biol.* **134**, 781–804.

Smith, W. W., Burnett, R. M., Darling, G. D. & Ludw M. L. (1977). *J. Mol. Biol.* **117**, 195–225.

Steigemann, W. & Weber, E. (1979). *J. Mol. Biol.* **127**, 309–338.

Sweet, R. M. & Eisenberg, D. (1983). *J. Mol. Biol.* **171**, 479–488.

Szebenyi, D. M. E., Obendorf, S. K. & Moffat, K. (1981). *Nature (London)*, **294**, 327–332.

Takano, T. & Dickerson, R. E. (1981*a*). *J. Mol. Biol.* **153**, 79–94.

Takano, T. & Dickerson, R. E. (1981*b*). *J. Mol. Biol.* **153**, 95–115.

Teller, D. C. (1976). *Nature (London)*, **260**, 729–731.

Tsernoglou, D. & Petsko, G. A. (1977). *Proc. Nat. Acad. Sci., U.S.A.* **74**, 971–974.

Tsukada, H. & Blow, D. M. (1985). *J. Mol. Biol.* **184**, 703–711.

Tsukihara, T., Fukuyama, K., Nakamura, M., Katsube, Y., Tanaka, N., Kakudo, M., Wada, K., Hase, T. & Matsubara, M. (1981). *J. Biochem. (Tokyo)*, **90**, 1763–1773.

Vita, C., Dalzoppo, D. & Fontana, A. (1984). *Biochemistry*, **22**, 5512–5519.

Weaver, L. H. & Matthews, B. W. (1987). *J. Mol. Biol.* **193**, 189–200.

Wlodawer, A. & Sjölin, L. (1983). *Biochemistry*, **22**, 2720–2728.

Wlodawer, A., Deisenhofer, J. & Huber, R. (1987). *J. Mol. Biol.* **193**, 145–156.

Wodak, S. J. & Janin, J. (1981). *Biochemistry*, **20**, 6544–6552.

Wolfenden, R. (1983). *Science*, **222**, 1087–1093.

Wolfenden, R., Andersson, L., Cullis, P. M. & Southgate, C. C. B. (1981). *Biochemistry*, **20**, 849–855.

Edited by R. Huber