J. Mol. Biol. (1988) **204**, 155–164

Surface, Subunit Interfaces and Interior of Oligomeric Proteins

Joël Janin[1], Susan Miller[2] and Cyrus Chothia[2,3]

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

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

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

(Received 23 November 1987, and in revised form 28 April 1988)

The solvent-accessible surface area (A_s) of 23 oligomeric proteins is calculated using atomic co-ordinates from high-resolution and well-refined crystal structures. A_s is correlated with the protein molecular weight, and a power law predicts its value to within 5% on average. The accessible surface of the average oligomer is similar to that of monomeric proteins in its hydropathy and amino acid composition. The distribution of the 20 amino acid types between the protein surface and its interior is also the same as in monomers. Interfaces, i.e. surfaces involved in subunit contacts, differ from the rest of the subunit surface. They are enriched in hydrophobic side-chains, yet they contain a number of charged groups, especially from Arg residues, which are the most abundant residues at interfaces except for Leu. Buried Arg residues are involved in H-bonds between subunits. We counted H-bonds at interfaces and found that several have none, others have one H-bond per 200 Å^2 of interface area on average (1 Å = 0·1 nm). A majority of interface H-bonds involve charged donor or acceptor groups. which should make their contribution to the free energy of dissociation significant, even when they are few.

The smaller interfaces cover about 700 Å^2 of the subunit surface. The larger ones cover 3000 to 10,000 Å^2, up to 40% of the subunit surface area in catalase. The lower value corresponds to an estimate of the accessible surface area loss required for stabilizing subunit association through the hydrophobic effect alone. Oligomers with small interfaces have globular subunits with accessible surface areas similar to those of monomeric proteins. We suggest that these oligomers assemble from preformed monomers with little change in conformation. In oligomers with large interfaces, isolated subunits should be unstable given their excessively large accessible surface, and assembly is expected to require major structural changes.

1. Introduction

Specific interactions between macromolecules are responsible for the assembly of complex biological structures, a few examples of which are known to atomic level: protein–DNA complexes, viruses and oligomeric proteins. Biochemical data indicate that subunit association plays an important role in stabilizing the structure of large proteins and in generating new functions, ligand binding sites or regulation. A systematic comparison of the structure of oligomeric proteins with monomeric ones should therefore shed light on their specific properties.

We use here accessible surface area (ASA†: Lee & Richards, 1971) measurements applied to a sample of 23 dimers, tetramers and higher oligomers, for which high-quality X-ray structures have been established. These measurements lead to an analysis of the chemical and amino acid composition of the surface and interior of oligomeric proteins, and also of the subunit interfaces. We compare these results with those obtained in the same way on a sample of 46 monomeric proteins (Miller *et al.*, 1987*a*).

A simple relationship exists between ASA and molecular weight of oligomers, though the relationship is not exactly the same as in monomers (Miller *et al.*, 1987*b*). Chemical characteristics of the solvent-accessible surface and of the surface buried in the protein interior, including their hydropathy and amino acid composition, are similar in

† Abbreviation used: ASA, accessible surface area.

0022-2836/88/210155-10 $03.00/0

Table 1
Protein structures

Protein	File name	Reference
A. *Dimers*		
Avian pancreatic peptide	1PPT	Blundell *et al.* (1981)
Uteroglobin	†	Morizé *et al.* (1987)
Subtilisin inhibitor	2SSI	Mitsui *et al.* (1979)
Cytochrome *c'*	2CCY	Finzel *et al.* (1985)
Superoxide dismutase	2SOD	Tainer *et al.* (1983)
Fab KOL immunoglobulin fragment	1FB4	Marquart *et al.* (1980)
Triose phosphate isomerase	1TIM	Banner *et al.* (1975)
Alcohol dehydrogenase	4ADH	Eklund *et al.* (1981)
Aspartate aminotransferase	†	Ford *et al.* (1982)
Citrate synthase	3CTS	Remington *et al.* (1982)
Glycogen phosphorylase a	†	Sprang & Fletterick (1979)
B. *Tetramers*		
Mellitin	1MLT	Terwilliger & Eisenberg (1981*a*)
Prealbumin	2PAB	Blake *et al.* (1978)
Hemoglobin, human deoxy	3HHB	Fermi *et al.* (1984)
Glutathione peroxidase	1GP1	Epp *et al.* (1983)
Concanavalin A	2CNA	Reeke *et al.* (1975)
Phosphofructokinase	†	Evans & Hudson (1979)
Lactate dehydrogenase	†	White *et al.* (1976)
Glyceraldehyde-3-phosphate dehydrogenase (GPDH)	†	Skarzynski *et al.* (1987)
Catalase, beef liver	7CAT	Murthy *et al.* (1981)
C. *Hexamers*		
Insulin, 2 Zn	1INS	Blundell *et al.* (1972)
Phycocyanin C	†	Schirmer *et al.* (1987)
D. *Octomer*		
Hemerythrin	1HMQ	Stenkamp *et al.* (1983)

File names are from the Protein Data Bank (Bernstein *et al.* 1977).
† Coordinates are gifts from the authors.

monomers and oligomers. In oligomers, the surface involved in subunit contacts is significantly different from either the accessible or the buried surface. It is less polar than either of these two surfaces, yet it contains charged groups. Non-polar interactions (van der Waals' and hydrophobic) and polar interactions (H-bond) are present at subunit interfaces in proportions which can vary widely from one oligomer to another.

2. Methods and Results

(a) *Atomic co-ordinates and molecular symmetry*

Table 1 lists the 23 oligomeric proteins used in this study. Atomic co-ordinates were taken from the Protein Data Bank (Bernstein *et al.*, 1977) or given to us by the authors. These co-ordinates are derived from high-resolution X-ray studies (2·6 Å or better; 1 Å = 0·1 nm) and have all been subjected to crystallographic refinement under stereochemical restraints. References are given in Table 1.

Most oligomeric proteins have point group symmetries. Co-ordinate files usually contain only one asymmetric unit, from which co-ordinates for complete oligomers are generated by applying appropriate symmetry operations. Ten of the 11 dimers in Table 1 have 2-fold symmetry, either exact (crystallographic) or approximate. Eight out of nine tetramers have dihedral D_2 symmetry. The exceptions are the Fab immunoglobulin fragment and hemoglobin, which contain chemically different subunits. Table 1 also contains three higher oligomers: two hexamers, insulin and phycocyanin C, one octomer, hemerythrin. Hemerythrin has approximate dihedral D_4 symmetry. Both hexamers have exact 3-fold symmetry. Dihedral D_3 symmetry is only approximate in the 2-Zn crystal form of insulin used here, and absent in phycocyanin C, which contains different α and β chains.

(b) *Accessible surface areas*

We calculated accessible surface areas for individual atoms within these protein structures with the Shrake & Rupley (1973) algorithm implemented in a computer program by A. M. Lesk as described (Miller *et al.*, 1987*a*). The probe radius R_w was 1·4 Å. The accessible surface area (A_s) of a protein is the sum of that of its component atoms. A_t is the ASA of the denatured protein represented as an extended polypeptide chain. It was calculated as the sum of the residues' ASA quoted in Table 2 of Miller *et al.* (1987*a*).

Table 2 gives the values of A_s and the molecular weight (M_r) of the oligomers. The correlation between A_s and M_r can be seen on a log–log plot (Fig. 1), which includes data for monomeric proteins taken from Miller *et al.* (1987*a*). Separate

Table 2
Accessible surface area of oligomeric proteins

Protein	M_r	A_s (Å^2)	$\Delta A/A$ (%)	A_t (Å^2)
A. *Dimers*				
Avian pancreatic peptide	8500	5300	2	12,400
Uteroglobin	15,780	7500	−10	47,400
Subtilisin inhibitor	21,860	10,900	4	35,600
Cytochrome *c'*	27,800	12,400	−1	40,200
Superoxide dismutase	31,420	13,800	−1	46,400
Fab KOL	47,170	20,700	9	69,900
Triose phosphate isomerase	53,060	20,300	−2	80,000
Alcohol dehydrogenase	79,820	29,000	3	119,900
Aspartate aminotransferase	89,560	30,800	0	134,200
Citrate synthase	96,000	28,500	−11	140,600
Phosphorylase a	190,820	60,500	10	283,500
B. *Tetramers*				
Mellitin	11,400	6300	−3	17,500
Prealbumin	49,880	19,700	0	74,400
Hemoglobin	64,440	24,100	1	93,000
Glutathione peroxidase	83,600	28,600	2	124,500
Concanavalin A	102,240	33,200	2	151,600
Phosphofructokinase	141,160	40,600	−7	205,000
Lactate dehydrogenase	145,640	46,800	5	219,000
GPDH	146,000	43,200	−3	216,000
Catalase	231,720	60,900	−4	336,000
C. *Hexamers*				
Insulin	34,560	13,100	−14	50,800
Phycocyanin C	112,650	41,400	12	160,000
D. *Octomer*				
Hemerythrin	107,520	35,900	−4	118,000

Molecular weights (M_r) and accessible areas (A_s) include amino acid residues, prosthetic groups and bound ligands when present in co-ordinate files. Total accessible surface areas (A_t) are calculated for the extended polypeptide chains. $\Delta A/A$ is the deviation of A_s from the value given by eqn (2).

linear plots lead to power laws:

$$A_s = 6{\cdot}3\, M_r^{0{\cdot}73} \quad (1)$$

for monomers, and:

$$A_s = 5{\cdot}3\, M_r^{0{\cdot}76} \quad (2)$$

for oligomers.

These laws fit observed values of A_s to within 4% on average for monomers, 5% for oligomers (Miller *et al.*, 1987*a*,*b*). Deviations from equation (2) are quoted for individual oligomers in Table 2. Only six are greater than 7%. Three proteins, uteroglobin, citrate synthase and insulin, have an ASA 10 to 14%, smaller than predicted by equation (2). They contain very large subunit interfaces (see Table 4 below). The three oligomers with 9 to 12% excess ASA are the Fab fragment, phosphorylase and phycocyanin C†.

Equations (1) and (2) imply that an oligomeric protein has a larger ASA than a monomeric protein of the same molecular weight. The excess is 7 to 13% in the molecular weight range 4000 to 35,000 where monomeric proteins have been studied.

(c) *Amino acid composition of oligomeric proteins*

Table 3 lists molar amino acid compositions for monomeric and oligomeric proteins. Compositions

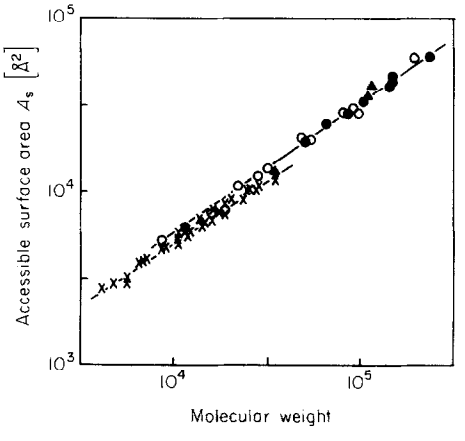

Figure 1. Accessible surface areas and molecular weights. ASA and molecular weight values are plotted on a log–log scale. (×) Monomers from Table 5 of Miller *et al.* (1987*a*); (○) dimers; (●) tetramers; (▲) hexamers and octomers. The lines are from eqns (1) and (2).

† We recently performed the same calculation on the Phycocyanin C hexamer of *Agmenellum quadruplicatum* (Schirmer *et al.*, 1986). Its ASA is within 1% of that predicted by eqn (2). This molecule occurs in the crystal as a true hexamer. The Phycocyanin C structure used in this paper is found in the crystal as a trimer and the hexamer was formed by model building on the basis of the *A. quadruplicatum* structure (Schirmer *et al.*, 1986, 1987).

Table 3
Amino acid compositions and surface/interior transfer free energies

	Molar percentage composition						ΔG_h (kcal mol^{-1})	
	Total		Surface		Interior			
Residue	(O)	(M)	(O)	(M)	(O)	(M)	(O)	(M)
Ala	8·9	8·7	8·1	7·9	10·6	11·0	0·16 (0·05)	0·20 (0·06)
Arg	4·7	3·1	6·0	4·0	2·1	0·4	−0·63 (0·10)	−1·34 (0·25)
Asn	4·3	5·2	5·3	6·3	2·3	2·0	−0·51 (0·10)	−0·69 (0·12)
Asp	5·7	6·1	7·2	7·4	2·6	2·2	−0·61 (0·09)	−0·72 (0·11)
Cys	1·5	2·7	1·0	1·8	2·6	5·4	0·59 (0·13)	0·67 (0·10)
Gln	3·3	3·6	4·3	4·5	1·3	1·3	−0·71 (0·13)	−0·74 (0·15)
Glu	5·5	4·9	7·6	5·2	1·4	1·0	−1·00 (0·12)	−1·09 (0·17)
Gly	8·1	9·0	7·9	8·8	8·5	9·7	0·04 (0·06)	0·06 (0·06)
His	2·9	2·3	3·2	2·2	2·3	2·4	−0·20 (0·10)	0·04 (0·12)
Ile	5·2	4·9	3·1	3·0	9·3	10·5	0·67 (0·07)	0·74 (0·08)
Leu	8·7	6·5	5·3	4·3	15·3	12·8	0·63 (0·06)	0·65 (0·07)
Lys	6·1	6·7	8·8	8·9	0·8	0·3	−1·42 (0·15)	−2·00 (0·30)
Met	2·1	1·5	1·6	0·9	3·0	3·0	0·40 (0·11)	0·71 (0·14)
Phe	3·8	3·8	2·4	2·5	6·5	7·7	0·60 (0·08)	0·67 (0·09)
Pro	4·7	4·0	6·0	4·7	2·3	2·2	−0·58 (0·10)	−0·44 (0·12)
Ser	6·6	7·9	7·1	8·9	5·6	5·0	−0·15 (0·07)	−0·34 (0·08)
Thr	5·6	6·4	6·1	7·1	4·5	4·6	−0·19 (0·07)	−0·26 (0·09)
Trp	1·3	1·6	1·1	1·3	1·7	2·7	0·28 (0·14)	0·45 (0·13)
Tyr	3·2	4·4	2·9	4·8	4·0	3·3	0·20 (0·09)	−0·22 (0·10)
Val	7·9	6·6	5·1	4·6	13·3	12·7	0·58 (0·06)	0·61 (0·07)
Number	6101	5436	4030	4040	2071	1396		

Molar percentage amino acid compositions of 23 oligomeric proteins (O), of 37 monomeric proteins (M), of their surface (residues with more than 5% accessibility) and interior (less than 5% accessibility). Free energies are derived from the ratio of the last 2 values using eqn (4). Standard deviations (in parentheses) assume that the variance of each residue count N is equal to N. Monomer data are taken from Table 7 of Miller *et al.* (1987*a*).

are quoted for the proteins, for their surface and for their interior. Surface and interior are defined by comparing the ASA of each residue in the protein to that in the extended polypeptide chain. The ratio of the two is the residue accessibility, ranging from 0% for residues with no atom contact with the solvent, to 100% for fully accessible residues. Surface residues are defined as having accessibilities larger than 5%, interior residues as having smaller accessibilities. The 5% cut-off was used by Miller *et al.* (1987*a*) to define residues buried in monomeric proteins, based on a histogram of residue accessibilities which is discussed there. Changing the cutoff to 2% or 8% does not affect the conclusions drawn below.

The degree of similarity between two amino acid compositions c and c' is estimated from the value of the root-mean-square difference $\langle \Delta c \rangle$:

$$\langle \Delta c \rangle^2 = 1/20 \, \Sigma_i (c_i - c_i')^2. \quad (3)$$

The summation is over 20 amino acid types. Values of $\langle \Delta c \rangle$ between oligomers and monomers are: 1·0% for overall compositions; 1·1% for surface compositions; 1·1% for interior compositions. Between interior and surface compositions, $\langle \Delta c \rangle$ is 4·5% for oligomers and 4·7% for monomers. Thus, monomeric and oligomeric proteins have similar amino acid compositions on average, and the 20 amino acid types are distributed similarly between their surface and interior.

A difference is worth noting. Cys derivatives (cystines, heme or metal bound cysteine residues) are very abundant in the smaller monomeric proteins of the sample analyzed by Miller *et al.* (1987*a*): they form 18·6% of the interior in 16 proteins with molecular weight 4000 to 12,000. Cys derivatives are rare or absent in the larger monomers and in most of the oligomeric proteins studied here: 47% of the 96 Cys residues in oligomers have a free SH group. Their surface–interior distribution is similar to that of cysteine derivatives.

(d) *Surface–interior distribution of amino acid residues*

The ratio $f = c_{\text{interior}}/c_{\text{surface}}$ of the molar fractions of a given amino acid type can be viewed as a coefficient of partition between the protein interior and its surface. This ratio is high for non-polar residues: 50/19 for Ile, Leu, Phe, Val, Cys and Met, in oligomers; 52/17 in monomers. It is low for charged residues Asp, Glu, Lys and Arg: 7/30 in oligomers; 4/25 in monomers. Non-polar residues are thus 2·6 times more frequent inside proteins than on their surface, charged residues are five times less frequent.

Partition coefficients are converted to transfer free energies by:

$$\Delta G_t = -RT \ln f. \quad (4)$$

Values of ΔG_t derived from data on monomeric proteins are correlated to free energies of

water/organic solvent transfer and to the hydrophobicity of residues (Janin, 1979; Miller *et al.*, 1987*a*). We compare them in Table 3, to values which we calculate from the present set of proteins. Standard deviations, which are derived from the number of observed buried and accessible residues, are lower limits of the errors.

ΔG_t values in oligomers and monomers are very similar. A seemingly large difference for Lys residues is not significant, given the very small number of buried Lys residues in monomers and in oligomers. Taken together, the two sets yield $\Delta G_t^{Lys} = -1{\cdot}64$ (0·17) kcal mol^{-1} (1 cal = 4·184 J). Buried Arg residues are rare inside monomers and inside the subunits of oligomers, yet they are fairly frequent at subunit interfaces, as we shall see below. They are heavily involved in H-bonds between subunits. Of 105 Arg residues located at interfaces, 26 are buried (less than 5% accessibility), hence the high value of ΔG_t^{Arg} in oligomers.

In our sample, Tyr residues are somewhat less frequent on the surface of oligomers than on the surface of monomers. As a result, ΔG_t^{Tyr} changes by 0·42 kcal mol^{-1}, which is twice the statistical standard deviation, but has no obvious chemical basis. For the 17 other amino acid types, the two sets of ΔG_t values differ by 0·1 kcal mol^{-1}, on average. Thus, they can be taken to be identical for all amino acids except arginine residues at subunit interfaces and, perhaps, tyrosine residues.

(e) *The subunit interface areas*

Subunit interface areas A_i are obtained as the difference between ASAs calculated on subunit in isolation and within the oligomer. Isolated subunits are assumed to retain their structure. Interface areas represent the surface area lost by a subunit in contacts with other subunits. In oligomers with more than one interface, calculations done on pairs of subunits yield interface areas for each type of contact (Table 4). Pairwise interface areas do not necessarily add up exactly to the total interface area, as a few atoms may be in contact with more than one neighboring subunit.

When the symmetry-relating subunits in oligomers are not exact, small differences may be observed from one subunit to another in surface area calculations. The largest relative difference was observed in 2-Zn insulin, where pairs of subunits

Table 4
Subunit interface areas

Protein subunit	Molecular weight	Surface area (Å^2) Accessible	Interface total	Pairwise		
A. *Dimers*						
Avian pancreatic peptide	4250	2630	690			
Uteroglobin	7890	3730	1500			
Subtilisin inhibitor	10,930	5470	750			
Cytochrome *c'*	13,900	6340	840			
Superoxide dismutase	15,710	6750	670			
Fab KOL, light chain	22,870	10,220	1830			
Fab KOL, heavy chain	24,300	10,500	1780			
TIM	26,530	10,080	1590			
Alcohol dehydrogenase	39,910	14,490	1630			
Aspartate aminotransferase	44,780	15,540	3150			
Citrate synthase	48,000	14,260	4890			
Phosphorylase a	95,410	30,270	3470			
B. *Tetramers*						
Mellitin	2850	1590	1040	440	410	260
Prealbumin	12,470	4890	1510	890	390	340
Hemoglobin, α-chain	15,740	5720	1730	820	670	270
Hemoglobin, β-chain	16,480	6350	1480	810	680	
Glutathione peroxidase	20,900	7130	1570	760	760	90
Concanavalin A	25,560	8300	2550	1410	1140	70
Phosphofructokinase	35,290	10,150	3600	2250	1260	90
Lactate dehydrogenase	36,410	11,700	5540	2750	1840	1080
GPDH	36,650	10,810	3800	2030	1450	490
Catalase	57,930	15,220	10,570	4630	4570	2060
C. *Hexamers*						
Insulin	5760	2250	1430	640	700	†90
Phycocyanin C, α-chain	18,030	6600	2200	1620	†610	
Phycocyanin C, β-chain	19,520	7200	2400	1540	710	†210
D. *Octomer*						
Hemerythrin	13,440	4490	1700	890	130	†710

Molecular weights, accessible surface areas and interface areas are given per subunit and include prosthetic groups or bound ligands when present in co-ordinate files. Interface areas calculated between pairs of subunits are quoted for tetramers and higher-order oligomers.

† Heterologous interfaces.

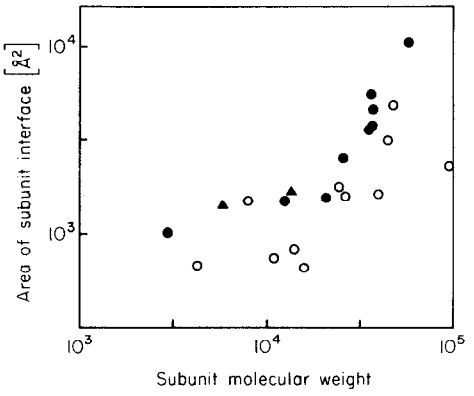

Figure 2. Interface areas and molecular weights. Total subunit interface areas are plotted against their molecular weights on a log–log scale. (○) Dimer subunits; (●) tetramer subunits; (▲) subunits of hexamers and octomers.

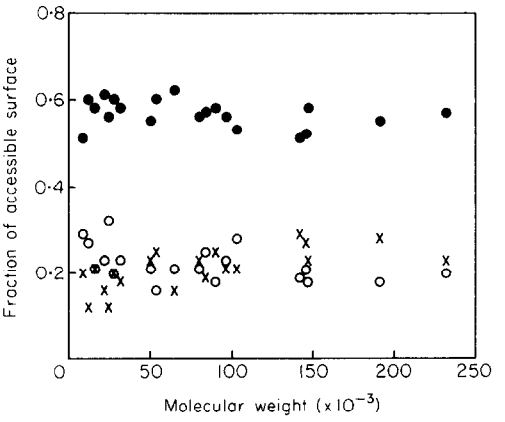

Figure 3. Hydropathy of the accessible surface of oligomeric proteins. The fraction of A_s in each protein that is contributed by non-polar (●), uncharged polar (○), or charged (×) groups, is plotted against its molecular weight.

related by a non-crystallographic 2-fold axis have somewhat different conformations (Blundell *et al.*, 1972). Their interface areas differed by 8%, other differences being less than 5%. Values quoted in Table 4 are averages.

Total interface areas go from 670 Å^2 per subunit in superoxide dismutase, 9% of the subunit surface, to 10,570 Å^2 in catalase, 40% of the subunit surface. Though A_i tends to increase both with the subunit size (Fig. 2) and the number of subunits in the oligomer, there is no simple correlation with either M_r or A_s. Interfaces covering more than 1000 Å^2 are found even in small dimers: uteroglobin has an interface that is almost as large as alcohol dehydrogenase, a protein five times larger.

In tetramers, three different interfaces relate subunits in pairs. When one pairwise interface area is much larger than the other two, the tetramer appears as a "dimer of dimers": examples from Table 4 are prealbumin and phosphofructokinase. However, not all of tetramers are like that: mellitin, glutathione peroxidase or catalase have two pairwise interfaces of equivalent size.

In hexamers and octomers, subunits related by 2-fold axes form isologous interfaces, subunits related by a 3-fold or 4-fold axis form heterologous interfaces, which dimers and tetramers do not have. The largest pairwise interfaces are isologous in all three examples. The very small heterologous interface of insulin is limited to Zn-bound histidine residues.

(f) *Hydropathy of accessible and buried surfaces*

We divide the protein surface into non-polar, neutral polar and charged components, taking for simplicity all carbon atoms to be non-polar, nitrogen, oxygen and sulfur to be polar, or charged in carboxylate, amino and guanidinium groups. The hydropathy of the surface is estimated from their relative contributions to the surface area.

The hydropathy of the accessible surface (A_s) of oligomeric proteins is described in Figure 3. The non-polar, polar and charged fractions show no systematic variation with molecular weight. Mean values and root-mean-square departures of individual protein surfaces from the mean are quoted in Table 5. Monomeric and oligomeric proteins are identical in the hydropathy of their native (A_s) and total (A_t) accessible surfaces, and therefore also in that of the surface buried upon folding (A_b), which is the difference between the two.

As for monomeric proteins, the accessible and buried surfaces are equally non-polar (57 to 58%), the latter being enriched in neutral polar groups and depleted in charged groups. Buried polar groups are mostly peptide groups, which H-bond when α-helices and β-sheets form. The association of

Table 5
Hydropathy of protein surface

Surface area (%)	Non-polar	Polar	Charged
A. *Accessible surface* (A_s)			
Oligomers	57 (3)	22 (3)	21 (5)
Monomers	57 (4)	24 (5)	19 (5)
B. *Total unfolded surface* (A_t)			
Oligomers	59 (2)	32 (2)	9 (2)
Monomers	58 (2)	33 (2)	9 (2)
C. *Buried surface*			
In monomers (A_b)	58 (5)	39 (5)	4 (–)
Between secondary structure elements	70 (5)	24 (6)	6 (2)
At interfaces (A_i)	65 (4)	22 (7)	13 (5)

Mean values of the non-polar, uncharged polar and charged surface area fractions are calculated on 37 monomeric proteins (Miller *et al.*, 1987*a*) and on the 18 dimers and tetramers listed in Table 1. Values quoted for the surface buried between secondary structure elements (helix–helix, helix–sheet and sheet–sheet packings) refer to 6 monomeric proteins analyzed by Chothia (1976). Root-mean-square departures of individual values from the means are given in parentheses.

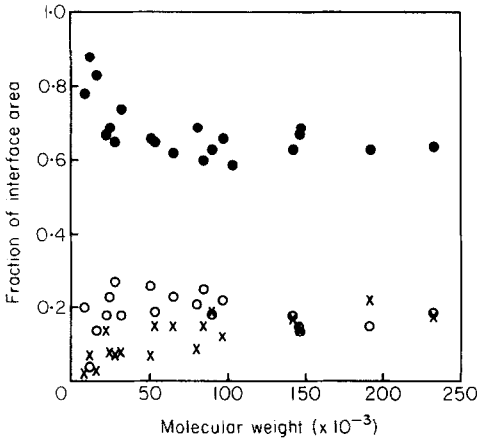

Figure 4. Hydropathy of subunit interfaces. The fraction of A_i in each oligomeric protein that is contributed by non-polar (●), uncharged polar (○), or charged (×) groups is plotted against its molecular weight.

secondary structure elements involves surfaces that are much less polar (Chothia, 1976).

In oligomers, the buried surface includes the surface buried within subunits and the interfaces. Figure 4 and Table 5 indicate that subunit interfaces are 65% non-polar on average, more than either the accessible or the buried surface, and almost as much as the surface buried between secondary structure elements. The non-polar fraction is fairly constant from one interface to another; its root-mean-square value is only 4%. The most non-polar interfaces are found in small oligomers: 90% in mellitin, 80% in avian pancreatic peptide and uteroglobin. The lowest value is 54% in hemerythrin. Other interfaces are 60 to 73% non-polar.

(g) *Amino acid composition of interfaces*

The non-polar character of interfaces must result in part from the amino acid composition of the subunit surfaces involved in contacts. We define interface residues as those that have a smaller ASA in the oligomer than in the isolated subunit. The subunits of the 26 proteins contain 6101 residues, 26% of which are interface residues in this sense. Each interface residue contributes 40 Å^2 on average to the interface area A_i and 32 Å^2 to the accessible surface area A_s. Only one-third have less than 5% accessibility. Therefore, the majority of interface residues also belong to the protein surface as we define it. In addition, 1% of the interface area involve prosthetic groups and bound ligands.

The amino acid composition of subunit interfaces in oligomeric proteins can be calculated on a molar basis as described above. However, many interface residues contribute only marginally to subunit contacts, in which 30% of them lose less than 10% of their ASA. Statistics based on the area contributed to interfaces by each type of residue are

Table 6

Amino acid composition of protein surfaces and subunit interfaces

Residue	Surface area composition (%) A_s	A_b	A_i
Ala	5·9	6·3	4·1
Arg	8·4	6·3	9·9
Asn	5·2	3·8	4·6
Asp	7·8	4·4	4·8
Cys	0·4	1·6	0·8
Gln	5·4	3·3	3·5
Glu	10·3	4·9	4·1
Gly	4·8	4·0	4·2
His	3·5	3·4	4·5
Ile	2·2	6·8	4·6
Leu	3·8	11·4	10·5
Lys	14·9	5·6	5·4
Met	1·5	2·9	3·9
Phe	1·9	5·9	6·0
Pro	5·6	3·6	5·3
Ser	6·3	4·5	4·1
Thr	5·5	4·8	4·7
Trp	0·8	2·5	2·4
Tyr	2·7	5·1	5·4
Val	3·2	9·1	7·3

Contribution of each type of residues to A_s, the solvent-accessible surface area of oligomeric proteins in Table 1, A_b, the surface area buried upon folding these proteins, and A_i, the subunit interface area. Note that the area compositions are not the same as the molar fractions quoted in Table 4. Each residue contributes to the surface area in proportion of its molar fraction, but also of its size.

therefore more representative of their participation than are molar fractions. Therefore, Table 6 quotes amino acid compositions as percentage contributions to interfaces areas (A_i), accessible surface areas (A_s) and the buried surface area (A_b).

The data clearly indicate that the interfaces are like the protein interior in their amino acid composition, and unlike the protein-accessible surface. The root-mean-square composition difference ($\langle \Delta c \rangle$) is 1·3% between A_i and A_b, 3·6% between A_i and A_s, and 3·8% between A_b and A_s. Non-polar residues contribute largely to interfaces. Ile, Leu, Phe, Val, Cys and Met contribute 33% to A_i, 38% to A_b and only 13% to A_s. Conversely, charged residues Asp, Glu and Lys contribute much more to the accessible surface (33%) than to interfaces (14%) or the buried surface (15%). Yet, Arg residues make the second largest contribution to interfaces (10%), the largest being Leu.

(h) *Hydrogen bonds between subunits*

We counted H-bonds and ionic interactions connecting subunits to their neighbors in 20 dimers and tetramers. Acceptor and donor groups less than 3·4 Å apart and with acceptable angular geometry were assumed to be H-bonded. We included bound ligands and prosthetic groups in the search and tested two possible orientations of amide groups in Asn and Gln, and of the imidazole group of His. We found 264 H-bonds, each present twice in dimers, four times in tetramers.

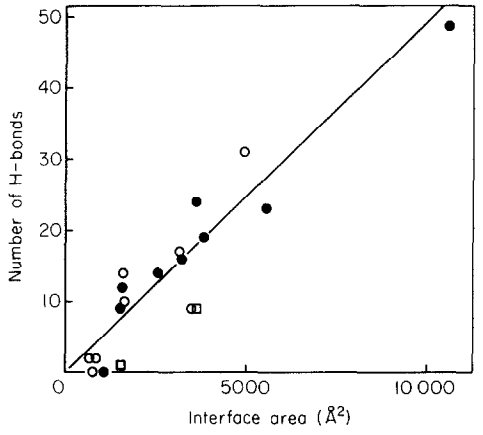

Figure 5. Hydrogen bonds at subunit interfaces. The number of intersubunit H-bonds is plotted against the corresponding interface area. H-bonds are assumed to exist between acceptor X-A groups and donor groups DH less than 3·4 Å apart, provided the DH. . .A-X angle is greater than 120°. Each bond is present twice in dimers, 4 times in tetramers, but only once in the Fab fragment and twice in hemoglobin. (○) Dimer subunits; (●) tetramer subunits. The subunits of uteroglobin and the Fab fragment (□) are linked by disulfide bridges.

The smaller interfaces have very few H-bonds (Fig. 5). Mellitin and the subtilisin inhibitor have none, other interfaces covering less than 1000 Å^2 per subunit have only one or two per subunit. Uteroglobin has one very weak H-bond (3·4 Å long) for 1500 Å^2 of interface area, and, in addition, it has a disulfide bridge between subunits. Above 1500 Å^2 per subunit, the number of polar interactions is roughly proportional to the buried surface, with one H-bond per 200 Å^2 on average.

Main-chain to main-chain H-bonds represent only 22% of intersubunit polar interactions. Thus, side-chains are involved in 78% of cases. The vast majority are charged: 22% of all polar interactions are salt bridges and 35% are H-bonds between one charged and one neutral group. The guanidinium group of Arg is the H-bond donor in 33% of all polar interactions. It is by far the most important side-chain group in this respect, being present in 42% of all polar interactions involving side-chains. Other important polar groups are the carboxylates of Asp and Glu (present in 36% of side-chain H-bonds), the amide groups of Asn and Gln (24%), and the amino group of Lys (11%).

3. Discussion

The correlation between accessible surface area and molecular weight is a general feature of proteins. Different amino acid sequences folding into different three-dimensional structures achieve the same ASA, as folding tends to reduce the surface of the polypeptide chain in contact with water, replacing solvent–solute interactions with solvent–solvent interactions that are thermodynamically more favorable. This is the structural basis of the hydrophobic contribution to protein folding, an idea that Kauzmann (1959) put forward years before any three-dimensional structure was known.

The low ASA of native proteins is due to their globular shape and to the close-packing of atoms. Monomeric proteins of up to 35,000 molecular weight, for which the correlation was first established, are globular and close-packed. Oligomeric proteins, which are often much larger, are also less globular. They can be described as multi-globular assemblies of subunits or domains separated by clefts. Yet, we established that the ASA of these assemblies is also strictly correlated to its molecular weight, implying that the hydrophobic effect contributes to the folding of oligomeric proteins as much as it does to that of monomeric ones.

The correlation being non-linear, it cannot apply for oligomers and for their subunits at the same time. For instance, the ASA of a subunit of molecular weight m in a tetramer should be from equation (2):

$$A_s/4 = 5{\cdot}3/4\,(4m)^{0{\cdot}76} = 3{\cdot}8m^{0{\cdot}76}. \qquad (5)$$

The ASA of the isolated subunit is larger by A_i. Assuming that it is equal to that of a monomeric protein of the same molecular weight given by equation (1), we get:

$$A_i = 6{\cdot}3m^{0{\cdot}73} - 3{\cdot}8m^{0{\cdot}76}. \qquad (6)$$

Real tetramer interfaces are larger than expected from equation (6), and therefore their isolated subunits have a larger ASA than do monomeric proteins of the same size. The excess ASA reaches 28% in some of the tetramers. A similar equation derived for dimers also predicts values of A_i that are far too small. This raises the question of how oligomers assemble from subunits that should not be stable, given their large ASA.

Observed interface areas can also be compared to a value that we derived in the following manner (Chothia & Janin, 1975; Janin & Chothia, 1978). We estimated the free energy of the degrees of translational and rotational freedom lost by two large molecules upon association, and assumed that the hydrophobic effect should compensate for it. Taking the hydrophobic free energy to be 25 cal mol^{-1} per square ångström unit of accessible surface (Chothia, 1975), we concluded that at least 1200 Å^2 should be buried upon association, or 600 Å^2 per subunit in dimers. The insulin dimer, the trypsin–trypsin inhibitor complexes and the hemoglobin dimer were found to have interfaces of about that size.

An inspection of Table 4 shows that the rule holds for many other systems. The smallest dimer interfaces, in superoxide dismutase and avian pancreatic peptide, are just above 600 Å^2 per subunit. These proteins have rather compact subunits with little excess ASA. We therefore suggest that the structure of monomer intermediates in their assembly is similar to that of the

subunits in the dimer. Examples of assembly of preformed components with little conformation change are known from the X-ray studies of several protease–protease inhibitor complexes (Huber *et al.*, 1974; Hirono *et al.*, 1984) and of a lysozyme–antibody complex (Amit *et al.*, 1986). These complexes all have interface areas of the order of 1400 Å^2, or 700 Å^2 per subunit (Janin & Chothia, 1976; Amit *et al.*, 1986).

On the other hand, most dimers have interfaces which are much larger than that. The structure of their monomers is likely to be strongly affected by dimerization. Uteroglobin, for instance, has two closely associated subunits. Each subunit is an open structure made of four α-helices. The fourth α-helix makes very few contacts within its subunit, and many with the other subunit (Morizé *et al.*, 1987). A similar situation is encountered in the tryptophan repressor (Schevitz *et al.*, 1985). A monomeric molecule having this structure would be highly unstable and completely insoluble in water, given the non-polar character of the protein surface exposed upon dissociation.

The assembly of tetramers is expected to proceed through monomer and dimer intermediates. The remarks made above apply to the first step. Each tetramer in Table 4, except mellitin, has at least one pairwise interface larger than 600 Å^2 per subunit. Dimers formed through this interface should therefore be stable. Tetramerization could then proceed through dimer–dimer interfaces, which are twice as large as given in Table 4. This mechanism is supported by renaturation studies carried out on several oligomeric enzymes, including triose phosphate isomerase, alcohol dehydrogenase, lactate dehydrogenase and glyceraldehyde-3-phosphate dehydrogenase, reviewed by Jaenicke (1984) and Jaenicke & Rudolph (1986). Dimeric intermediaries should, however, be excluded for mellitin, which has small pairwise interfaces, even though they cover 1000 Å^2 per subunit in total. The tetramer should form in one step, presumably from partly unfolded monomers observed in solution studies (Terwilliger & Eisenberg, 1981*b*).

The data of Table 4 also suggest that the assembly of hexamers and tetramers proceeds through dimer intermediates, rather than through trimers or tetramers. In the three cases studied here at least, isologous interfaces are large enough to stabilize dimers and heterologous interfaces are small.

Our study of the hydropathy of the protein surface and of its amino acid composition reveals no significant difference between monomeric and oligomeric proteins. None was expected, since the distribution of amino acids between the protein surface and interior is governed by physicochemical properties of the side-chains. However, the surfaces that are involved in subunit association are unlike the accessible surface and the surface buried within subunits. They are also unlike interfaces between components of several protein–protein complexes such as protease–inhibitor complexes, and complexes between antibodies and protein antigens. The hydrophobicity of the interfaces in these complexes is similar to that of the average protein surface, and therefore lower than that of subunit interfaces in oligomers (Janin & Chothia, 1976; our unpublished results).

Protease–inhibitor and antibody–antigen contacts occur between preformed proteins. Subunit interfaces, which form during folding, are more like the surfaces involved in packing α-helices and β-sheets, being made mostly of hydrophobic side-chains. Main-chain groups contribute only 18% of the interface areas. They form 22% of the accessible surface, which is also largely side-chains, but with many polar or charged groups and very few H-bonds (Chothia, 1976). At subunit interfaces, charged groups are relatively abundant and polar interactions cannot be ignored. Yet, several of the smaller oligomers studied here have highly non-polar interfaces with no H-bonds at all. The extensive subunit interface of uteroglobin contains at most two per dimer. In larger proteins, about one-third of the interface area is polar, and H-bonds are formed at the rate of about one H-bond per 200 Å^2 of buried surface.

These 200 Å^2 are worth some 5 kcal mol^{-1} of hydrophobic free energy on a basis of 25 cal mol^{-1} per square ångström unit, contributing to the free energy of dissociation. The contribution of an interface H-bond A . . . B is equal to its free energy plus that of a water–water H-bond, minus the free energy of two H-bonds made to water by A and B in the dissociated state. Estimates based on model systems fit rather well with measurements done in solution, and suggest that the balance is 0·5 to 1·8 kcal mol^{-1} in favor of association when A and B are neutral, 3 to 6 kcal mol^{-1} when A or B is charged (Fersht, 1987). A majority of interface H-bonds involve charged groups, the rest being mostly main-chain to main-chain H-bonds. Thus, optimal use is made of the relatively few polar side-chains present at interfaces. The selection of arginine among other polar residues is remarkable from this point of view. Its side-chain contributes four times as many H-bonds as Lys does, and interfaces are the only places where Arg residues are found buried in significant numbers.

We thank Drs P. R. Evans, R. Huber, H. Jansonius, J. P. Mornon, M. G. Rossmann, T. Schirmer and A. J. Wonacott for the gift of atomic co-ordinates. We are grateful to Pr A. M. Lesk for discussion and the gift of computer programs, and to J. Cresswell for the Figures.

References

Amit, A. G., Mariuzza, R. A., Phillips, S. E. V. & Poljack, R. (1986). *Science*, **233**, 747–753.

Banner, D. W., Bloomer, A. C., Petsko, A. G., Phillips, D. C., Pogson, C. I., Wilson, I. A., Corran, P. H., Furth, A. J., Milman, J. D., Offord, R. E., Priddle, J. D. & Waley, S. G. (1975). *Nature (London)*, **255**, 609–614.

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.

Blake, C. C. F., Geisow, M. J., Oatley, S. J., Rérat, B. & Rérat, C. (1978). *J. Mol. Biol.* **121**, 339–356.

Blundell, T. L., Dodson, G. G., Hodgkin, D. C. & Mercola, D. (1972). *Advan. Protein Chem.* **26**, 279–402.

Blundell, T. L., Pitts, J. E., Tickle, I. J., Wood, S. P. & Wu, C. W. (1981). *Proc. Nat. Acad. Sci., U.S.A.* **78**, 4175–4179.

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

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

Chothia, C. & Janin, J. (1975). *Nature (London)*, **256**, 705–708.

Eklund, H., Samama, J. P., Wallen, L., Bränden, C. I., Åkeson, Å. & Jones, T. A. (1981). *J. Mol. Biol.* **146**, 561–587.

Epp, O., Ladenstein, R. & Wendel, A. (1983). *Eur. J. Biochem.* **133**, 51–69.

Evans, P. R. & Hudson, P. J. (1979). *Nature (London)*, **279**, 500–504.

Fermi, G., Perutz, M. F., Shaanan, G. & Fourme, R. (1984). *J. Mol. Biol.* **175**, 159–174.

Fersht, A. R. (1987). *Trends Biochem. Sci.* **12**, 301–304.

Finzel, B. C., Weber, P. C., Hardmann, K. D. & Salemme, F. R. (1985). *J. Mol. Biol.* **186**, 627–643.

Ford, G. C., Eichele, G. & Jansonius, J. N. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 2559–2563.

Hirono, S., Akagawa, H., Mitsui, Y. & Iitaka, Y. (1984). *J. Mol. Biol.* **178**, 389–413.

Huber, R., Kukla, D., Bode, W., Schwager, P., Bartels, K., Deisenhofer, J. & Steigemann, W. (1974). *J. Mol. Biol.* **89**, 73–101.

Jaenicke, R. (1984). *Angew. Chem. Int. Ed. Engl.* **23**, 395–413.

Jaenicke, R. & Rudolph, R. (1986). *Methods Enzymol.* **131**, 218–250.

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

Janin, J. & Chothia, C. (1976). *J. Mol. Biol.* **100**, 197–211.

Janin, J. & Chothia, C. (1978). *Biochemistry*, **17**, 2943–2948.

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

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

Marquart, M., Deisenhofer, J. & Huber, R. (1980). *J. Mol. Biol.* **141**, 369–391.

Miller, S., Lesk, A. M., Janin, J. & Chothia, C. (1987*a*). *J. Mol. Biol.* **196**, 641–656.

Miller, S., Lesk, A. M., Janin, J. & Chothia, C. (1987*b*). *Nature (London)*, **328**, 834–836.

Mitsui, Y., Satow, Y., Watanabe, Y., Hirono, S. & Iitaka, Y. (1979). *J. Mol. Biol.* **131**, 697–724.

Morizé, I., Surcouf, E., Vaney, M. C., Epelboin, Y., Buehner, M., Fridlansky, F., Milgrom, E. & Mornon, J. P. (1987). *J. Mol. Biol.* **194**, 725–739.

Murthy, M. R. N., Reid, T. J. III, Sicignano, A., Tanaka, N. & Rossmann, M. G. (1981). *J. Mol. Biol.* **152**, 465–499.

Reeke, G. N., Becker, J. N. & Edelman, G. M. (1975). *J. Biol. Chem.* **250**, 1525–1547.

Remington, S., Wiegand, G. & Huber, R. (1982). *J. Mol. Biol.* **158**, 111–152.

Schevitz, R. W., Otwinowski, Z., Joachimiak, A., Lawson, C. L. & Sigler, P. B. (1985). *Nature (London)*, **317**, 782–786.

Schirmer, T., Huber, R., Schneider, M., Bode, W., Miller, M. & Hackert, M. L. (1986). *J. Mol. Biol.* **188**, 651–676.

Schirmer, T., Bode, W. & Huber, R. (1987). *J. Mol. Biol.* **196**, 677–695.

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

Skarzynski, T., Moody, P. C. E. & Wonacott, A. J. (1987). *J. Mol. Biol.* **193**, 171–187.

Sprang, S. & Fletterick, R. J. (1979). *J. Mol. Biol.* **131**, 523–551.

Stenkamp, R. E., Sieker, L. L. & Jensen, L. H. (1983). *Acta Crystallogr. sect. B*, **39**, 697–703.

Tainer, J. A., Getzoff, E. D., Richardson, J. S. & Richardson, D. C. (1983). *Nature (London)*, **306**, 284–287.

Terwilliger, T. C. & Eisenberg, D. (1981*a*). *J. Biol. Chem.* **257**, 6010–6015.

Terwilliger, T. C. & Eisenberg, D. (1981*b*). *J. Biol. Chem.* **257**, 6016–6022.

White, J. L., Hackert, M. L., Buehner, M., Adams, M. J., Ford, G. C., Lentz, P. J., Smiley, I. E., Steindel, S. J. & Rossman, M. G. (1976). *J. Mol. Biol.* **102**, 759–779.

Edited by R. Huber