PROTEINS: Structure, Function, and Genetics 8:6–13 (1990)

Hydrophobicity of Amino Acid Subgroups in Proteins

Glenn J. Lesser and George D. Rose
Department of Biological Chemistry, College of Medicine, Pennsylvania State University, Hershey, Pennsylvania 17033

ABSTRACT **Protein folding studies often utilize areas and volumes to assess the hydrophobic contribution to conformational free energy (Richards, F.M. Annu. Rev. Biophys. Bioeng. 6:151–176, 1977). We have calculated the mean area buried upon folding for every chemical group in each residue within a set of X-ray elucidated proteins. These measurements, together with a standard state cavity size for each group, are documented in a table. It is observed that, on average, each type of group buries a constant fraction of its standard state area. The mean area buried by most, though not all, groups can be closely approximated by summing contributions from three characteristic parameters corresponding to three atom types: (1) carbon or sulfur, which turn out to be 86% buried, on average; (2) neutral oxygen or nitrogen, which are 40% buried, on average; and (3) charged oxygen or nitrogen, which are 32% buried, on average.**

Key words: hydrophobicity, hydrophobic effect, protein folding, solvent accessibility

INTRODUCTION

Since the publication of a seminal review by Kauzmann[1] 31 years ago, the hydrophobic interaction has been viewed as a key factor and possibly the driving force in protein folding. As defined by Kauzmann and more recently by Baldwin,[2] the hydrophobic interaction refers to the process in which a hydrophobic group is removed from solvent access and buried within the solvent-shielded molecular interior upon folding of the protein. To quantify this effect, various scales have been proposed, as discussed earlier in a review.[3]

Scales may be classified into two types: solution measurements and empirical calculations. The thermodynamic basis for solution measurements was established by Cohn and Edsall[4] and Nozaki and Tanford,[5] who reckoned the hydrophobicity for solutes of interest as the free energy of transfer, ΔG^0_{xfer}, between an aqueous phase and a suitably chosen organic phase. The free energy of transfer is then given by the relationship

$$\Delta G^0_{xfer} = -RT \ln K_{aqueous \rightarrow organic},$$

where K is the partition coefficient between respective phases. The free energy of transfer of the amino acids between water and various organic solvents has been measured using this approach.[5–8]

These solution thermodynamics were applied to X-ray elucidated protein structures by Chothia[9] and Janin,[10] who treated the molecule's solvent-shielded interior (i.e., inside) and solvent-accessible exterior (i.e., outside) as distinct solvent phases. Using a well-known algorithm due to Lee and Richards[11] to distinguish inside from outside, Chothia[9] and Janin[10] calculated empirical distribution coefficients for residues within folded proteins.

Extending these earlier studies, we derived two new empirical scales for amino acid residues in proteins of known structure.[12] These scales measure

1. the mean area lost when a residue is transferred from a defined standard state to a folded protein. The area a residue buries upon folding is proportional to its hydrophobic contribution to the conformational free energy,[13] $\Delta G_{conformation}$.
2. the mean fractional accessibility of a residue, defined as its mean accessible area in protein molecules divided by the standard state area. The fractional accessibility is an intrinsic measure of hydrophobicity.

Although related, these two quantities are not equivalent. For example, a bulky arginine residue makes a large contribution to $\Delta G_{conformation}$ because its area loss upon folding is large, approximating that of leucine. Yet the fractional accessibility of an arginine is comparatively high because the remaining unburied area is also large.

Many workers have noted that residues which contain both hydrophobic and hydrophilic groups are not well-represented by a single hydrophobicity parameter.[14–19] Two of these studies introduced procedures for assessing such parameters group by group along the backbone or side chain of an amino acid. Eisenberg and McLachlan[15] use five atomic

Received February 22, 1990; accepted February 27, 1990.
Address reprint requests to George D. Rose, Department of Biological Chemistry, Hershey Medical Center, Pennsylvania State University, 500 University Drive, Hershey, Pennsylvania 17033.
G.J. Lesser's present address: North Carolina Baptist Hospital, Bowman Gray School of Medicine, Wake Forest University, Winston-Salem, NC 27103.

solvation parameters while Abraham and Leo[16] use a set of fragments. In each case, individual group contributions sum to residue partition coefficients that agree favorably with those obtained in solution studies by Fauchère and Pliska,[7] who measured the octanol–water partitioning of blocked derivatives of the 20 amino acids (i.e., of the *N*-acetyl amino acid amides).

In a similar vein, we now decompose our previous whole-residue scales into their constituent group contributions. These groups are then used to derive three characteristic parameters that represent:

1. groups containing carbon or sulfur (i.e., $-CH_n$, –SH, –S–)
2. groups containing a polar nitrogen or oxygen (i.e., $-CONH_2$, –OH, imidazole) and
3. groups containing a charged nitrogen or oxygen (i.e., $-NH_3^+$, $-COO^-$).

METHODS

The solvent-accessible surface area[11] was calculated for 82,599 nonhydrogen atoms in 10,937 residues from 61 proteins of known structure. The proteins used and their bracketed Brookhaven file names[20] are arabinose-binding protein [1ABP], actinidin [2ACT], alcohol dehydrogenase [4ADH], adenyl kinase [2ADK], α- lytic protease [1ALP], penecillopepsin [2APP], rhizopus acid protease [1APR], azurin [1AZU], cytochrome b_{562} [156B], cytochrome b_5 [2B5C], bovine phospholipase A_2 [2BP2], cytochrome c_{551} [351C], cytochrome c_{550} [155C], cytochrome c_2 [2C2C], carbonic anhydrase C [1CAC], concanavalin A [3CNA], carboxypeptidase A [3CPA], carp Ca-binding protein [1CPV], crambin [1CRN], α-cobratoxin [1CTX], cytochrome *c* [3CYT], erythrocruorin [1ECD], elastase [1EST], immunoglobulin Fab NEW [3FAB], ferredoxin [1FDX], ferredoxin (*S. platensis*) [3FXC], flavodoxin [3FXN], γ-chymotrypsin A [2GCH], glucagon [1GCN], D-glyceraldehyde-3-phosphate dehydrogenase [1GPD], glutathione reductase [2GRS], high potential iron protein [1HIP], insulin [1INS], lactate dehydrogenase [4LDH], leghemoglobin [1LH1], lamprey hemoglobin [1LHB], lysozyme [7LYZ], T_4 phage lysozyme [1LZM], sperm whale myoglobin [1MBN], horse hemoglobin [2MHB], snake neurotoxin [1NXB], ovomucoid, third domain [1OVO], prealbumin [2PAB], papain [8PAP], plastocyanin [1PCY], phosphoglycerate mutase [3PGM], avian pancreatic polypeptide [1PPT], pancreatic trypsin inhibitor [4PTI], trypsin (trigonal) [3PTN], trypsin [3PTP], Bence–Jones immunoglobulin [1REI], rhodanese [1RHD], ribonuclease A [4RSA], rubredoxin [3RXN], subtilisin [1SBT], *Streptomyces griseus* protease [2SGA], staphylococcal nuclease [2SNS], Cu, Zn superoxide dismutase [2SOD], *Streptomyces* subtilisin inhibitor [2SSI], triose phosphate isomerase [1TIM], and thermolysin [3TLN].

Atomic radii used were those of Richards[11]: tetrahedral C = 2.0 Å, trigonal C = 1.7 Å, carbonyl O = 1.4 Å, hydroxyl O = 1.6 Å, carboxyl O = 1.5 Å, tetrahedral N = 2.0 Å, trigonal N = 1.7 Å, divalent S = 1.85 Å, sulfhydryl S = 2.0 Å. The probe radius was 1.4 Å.

From the measured accessibilities, 173 accessibility vectors, $\Psi_{\text{residue}}^{\text{group}}$, were constructed, one for every group in the backbone or side chain of each of the 20 residues. For example, the 8 vectors for isoleucine represent N, C_α, C, O, C_β, $C_{\gamma 1}$, $C_{\gamma 2}$, and $C_{\delta 1}$. Each vector is 100-dimensional, with consecutive elements that correspond to the number of groups falling within semiopen intervals from 0–1% buried, 1–2% buried, . . . , 98–99% buried, and the closed interval from 99–100% buried. Returning to the example, the accessibility vector that represents all $C_{\delta 1}$ methylenes from the 542 isoleucine residues in the data base is

$$\Psi_{\text{Ile}}^{C\delta 1} = [25,1,0,2,1,0,1,0,0,0,1,0,0,1,0,1,3,3,1,0,\\ 0,0,0,2,1,2,5,2,1,3,3,1,1,0,1,3,0,3,0,2,2,\\ 0,2,2,1,0,1,0,0,1,0,5,0,3,0,2,2,1,2,1,6,5,\\ 1,2,0,1,3,1,0,3,2,1,4,2,0,0,4,3,2,2,1,1,2,\\ 4,2,3,3,7,5,6,2,11,8,6,11,7,27,19,31,249]$$

Two 100-dimensional column vectors, **b** and **e**, were also constructed. **b** contains the 100 midpoints of an accessibility vector, i.e., 0.5%, 1.5%, . . . , 99.5%. **e**, the identity vector, contains the multiplicative identity, i.e., 1,1, . . . , 1.

The *fractional accessibility* of a group or residue is defined as the measured solvent-accessible surface area divided by a standard state surface area. The fractional accessibility was used to construct the 173 accessibility vectors, Ψ.

Two similar procedures for defining a standard state area have been discussed in the literature.[21,22] The *extended standard state* for a residue,[21] X, is taken to be the surface area of that residue in the extended tripeptide Gly-X-Gly, with dihedral angles $\phi = -140°$, $\Psi = 135°$, $\chi^1 = -120°$ and $\chi^{2,\ldots,N} = 180°$. In the present study we use a stochastic standard state, similar to that of Shrake and Rupley.[22] The *stochastic standard state* is defined as the mean accessibility of an ensemble of tripeptides with dihedral angles that reflect the observed distribution in the data base. Using isoleucine as an example, 542 tripeptides, Gly-Ile-Gly, were constructed with angles

$$\phi_i, \Psi_i, \chi_i^{1,\ldots,N} \qquad \text{where } i = 1,2,\ldots,542$$

The stochastic standard state accessibility of Ile or its constituent groups is the corresponding ensemble average. Physically, this standard state reflects the degree to which a residue's groups are buried by backbone atoms from covalent neighbors.

The first, second, and third moments, $<b>$, $<b^2>$, and $<b^3>$, for each group were then calculated as

$$<b^i> = \frac{\Psi \cdot \boldsymbol{b}^i}{\Psi \cdot \boldsymbol{e}} \quad i = 1,2,3 \tag{1}$$

The first moment is the mean fractional accessibility, while the second moment is used to calculate the standard deviation, σ, as

$$\sigma = [<b^2> - <b>^2]^{1/2} \tag{2}$$

The third moment is a measure of skewness.

Linear regression analysis was performed using the subroutine RLONE and nonlinear regression using ZXSSQ, both from the International Mathematical Subroutine Library.[23]

In the charged residues Arg, Asp, and Glu, the most exposed side chain nitrogen or oxygen was taken to be the charged atom. In the ensuing analysis, this charged atom is designated as $N_{\eta 2}$ in Arg, $O_{\delta 2}$ in Asp, and $O_{\epsilon 2}$ in Glu, regardless of the original Brookhaven file label.

RESULTS

Table I lists several quantities of interest, as determined in this study. The mean accessible surface area, $<A>$, is the average solvent accessible surface area in folded proteins. Values of $<A>$ by group, residue, and side chain are included. The stochastic standard state accessibility, A^0, is also given. The difference, $A^0-<A>$, measures the mean area buried on folding; this quantity[13] is proportional to the hydrophobic contribution to $\Delta G_{\text{conformation}}$. The mean fraction buried, $(A^0-<A>)/A^0$, was calculated as $1 - <b>$, using Eq. (1); this quantity can be viewed as the area a moiety loses on folding normalized by the area it has to lose. The standard deviation, σ, and the skewness of the mean fraction buried were calculated using Eqs. (2) and (1), respectively.

Values of $<A>$ and $(A^0-<A>)/A^0$ in Table I differ slightly from those measured in our previous analysis of 23 proteins.[12] The former study, which was limited to whole residues, serves as a control for the sensitivity of these parameters to the size of the data base. The difference between $(A^0-<A>/A^0)$ in the former 23-protein study and in the present 61-protein study is rarely greater than 2% and never greater than 5%. Thus, these parameter values appear to be characteristic of globular proteins in general and not merely a function of the particular proteins chosen for inclusion in the data base.

The standard deviations in Table I are large. The accessibility vector for $\Psi_{\text{Ile}}^{C\delta 1}$, shown explicitly in the previous section, is a typical example. While 69% of all groups are at least 90% buried, the remaining 31% of the groups are distributed almost uniformly between complete exposure and 90% burial, resulting in the listed value of σ.

Despite large standard deviations, the accessibility vectors, particularly those of carbon atoms, are markedly skewed. For carbon atoms, the majority of values are tightly "bunched" around 95% or more burial, consistent with observations of other investigators.[24] Visual impressions of skewed distributions, evident in the $\Psi_{\text{Ile}}^{C\delta 1}$ example, were quantified by calculating the third moment about the mean, a test of skewness,[25] using Eq. (1). For a completely unskewed distribution, the third moment would be zero. The large values of $<b^3>$ observed for the hydrophobic groups indicate that such groups have a distinct preference for the buried interior of the protein.

Accessibility vectors having large standard deviations but highly skewed distributions are consistent with Richards' earlier observation that "the relevant forces and final structure require more careful definition than is implied by the common feeling that inside equals nonpolar and outside equals polar."[13]

In this study, cysteine residues are distinguished from 1/2-cystine residues and listed separately in Table I. Surprisingly, the –SH in cysteine loses almost as much surface area on folding as the –S– in 1/2-cystine. Apparently, Cys tends to be buried, regardless of whether it forms a disulfide bridge.

As described in the previous section, two similar procedures for calculating a standard state accessibility are found in the literature. In Table II, the stochastic standard state, from Table I, is compared with the extended standard state, from Chothia.[9] The areas are similar in both states, differing by a few percent in most cases, and by no more than 15% in the worst case.

Side chain data from Table I were subdivided into (1) hydrophobic groups from hydrophobic residues, (2,3) hydrophobic groups from polar or charged residues, (4) polar groups, and (5) charged groups. Specifically, these groups include:

1. carbon atoms from hydrophobic residues (Val, Ile, Leu, Met, Phe, and Trp) and sulfur atoms from Cys and 1/2-Cys;
2. carbon atoms from polar and charged residues (Arg, Asn, Asp, Gln, Glu, Tyr), excluding those adjacent to the polar or charged terminal group, and Ala;
3. carbon atoms from polar or charged residues (Arg, Asn, Asp, Gln, Glu, His, Ser, and Thr) adjacent to a polar or charged terminal group;
4. uncharged oxygen or nitrogen atoms (from Arg, Asn, Asp, Gln, Glu, His, Ser, and Thr), and
5. charged oxygen or nitrogen atoms (from Arg, Glu, and Asp).

For each group, the straight line of best fit was

TABLE I. Mean and Standard State Surface Areas for the Amino Acid Residues, Their Constituent Atomic Groups, and Their Side Chains*

Residue	A^0	$\langle A \rangle$	$(A^0-\langle A \rangle)/A^0$	σ	Skewness
1. Ala (944)	118.1	32.1	.73	.28	.54
N	6.5	1.3	.84	.35	.78
C_α	13.6	4.0	.71	.40	.61
C	1.1	0.3	.77	.49	.72
O	25.0	5.5	.78	.33	.68
C_β	71.9	20.9	.71	.32	.55
S-C	71.9	20.9	.71	.32	.55
2. Arg (315)	256.0	99.2	.62	.23	.33
N	5.1	1.0	.84	.37	.79
C_α	9.4	2.5	.74	.42	.67
C	1.2	0.3	.84	.42	.80
O	23.4	4.6	.81	.31	.71
C_β	25.9	6.4	.76	.32	.63
C_γ	23.9	8.0	.67	.41	.55
C_δ	29.6	12.0	.60	.39	.44
N_ϵ	13.3	5.5	.59	.42	.47
C_ζ	2.2	1.0	.57	.38	.40
$N_{\eta 1}$	58.6	20.4	.66	.30	.45
$N_{\eta 2}$	63.4	37.7	.41	.33	.21
S-C	216.9	90.9	.58	.25	.31
3. Asn (497)	165.5	62.6	.63	.25	.36
N	5.5	0.7	.87	.31	.82
C_α	10.1	2.8	.73	.40	.62
C	1.5	0.3	.79	.47	.75
O	23.1	6.0	.74	.36	.64
C_β	32.1	12.0	.63	.35	.46
C_γ	3.0	1.3	.56	.44	.45
$O_{\delta 1}$	31.1	14.3	.54	.42	.40
$N_{\delta 2}$	59.1	25.1	.58	.34	.38
S-C	125.3	52.7	.58	.27	.33
4. Asp (612)	158.7	62.5	.61	.25	.34
N	5.1	0.9	.84	.36	.77
C_α	10.8	3.7	.66	.43	.54
C	1.1	0.3	.76	.49	.70
O	23.5	5.4	.78	.32	.66
C_β	33.5	14.1	.58	.37	.42
C_γ	5.1	2.3	.56	.38	.40
$O_{\delta 1}$	38.7	11.2	.72	.30	.54
$O_{\delta 2}$	40.9	24.8	.40	.37	.24
S-C	118.2	52.3	.56	.28	.31
5. Cys (263)	146.1	17.0	.89	.14	.75
N	5.2	0.5	.94	.26	.91
C_α	10.5	2.0	.82	.37	.77
C	1.3	0.2	.86	.35	.82
O	25.6	4.2	.84	.30	.76
C_β	38.5	3.3	.92	.16	.83
S_γ	65.0	6.9	.90	.17	.79
S-C	103.5	10.2	.90	.14	.81
6. 1/2-Cys (173)	146.1	16.2	.89	.14	.76
N	5.2	1.0	.85	.34	.79
C_α	10.5	1.4	.87	.28	.80
C	1.3	0.4	.78	.48	.74
O	25.6	4.9	.81	.30	.71
C_β	38.5	4.7	.88	.21	.78
S_γ	65.0	3.9	.94	.12	.87
S-C	103.5	8.6	.92	.14	.82
7. Gln (384)	193.2	72.0	.63	.22	.35
N	5.3	0.6	.89	.27	.85
C_α	8.2	2.6	.69	.45	.61
C	0.9	0.2	.79	.47	.75
O	23.4	4.6	.81	.31	.71
C_β	26.6	7.3	.73	.29	.55
C_γ	30.9	11.3	.64	.37	.49

(continued)

TABLE I. Continued

Residue	A^0	$\langle A \rangle$	$(A^0-\langle A \rangle)/A^0$	σ	Skewness
C_δ	3.7	1.8	.53	.44	.38
$O_{\epsilon 1}$	34.2	17.8	.48	.39	.32
$N_{\epsilon 2}$	60.0	25.9	.57	.33	.36
S-C	155.4	64.0	.59	.25	.32
8. Glu (510)	186.2	78.3	.58	.23	.29
N	4.9	0.9	.84	.37	.78
C_α	9.2	3.0	.68	.43	.58
C	1.0	0.3	.75	.53	.71
O	22.7	5.1	.78	.33	.67
C_β	26.8	9.8	.64	.35	.47
C_γ	33.2	14.2	.58	.38	.41
C_δ	5.5	3.1	.45	.41	.29
$O_{\epsilon 1}$	41.4	14.4	.66	.30	.45
$O_{\epsilon 2}$	41.5	27.6	.34	.35	.18
S-C	148.4	69.0	.54	.26	.27
9. Gly (1004)	88.1	27.5	.69	.29	.49
N	11.0	2.1	.82	.32	.73
C_α	46.8	17.0	.64	.35	.47
C	3.0	0.8	.77	.33	.65
O	27.3	7.6	.72	.36	.61
10. His (236)	202.5	54.3	.74	.23	.51
N	5.7	0.8	.90	.28	.86
C_α	9.4	2.2	.77	.37	.68
C	1.4	0.3	.83	.40	.80
O	23.9	4.3	.82	.30	.73
C_β	34.0	6.9	.80	.29	.68
C_γ	1.6	0.3	.80	.31	.70
$N_{\delta 1}$	11.4	2.9	.75	.36	.65
$C_{\epsilon 1}$	51.6	15.4	.71	.30	.52
$N_{\epsilon 2}$	30.5	11.3	.64	.35	.47
$C_{\delta 2}$	33.0	10.0	.70	.34	.55
S-C	162.1	46.7	.71	.25	.49
11. Ile (542)	181.0	24.2	.87	.18	.73
N	3.6	0.6	.90	.33	.87
C_α	4.8	1.0	.81	.44	.78
C	0.8	0.2	.88	.41	.87
O	21.7	2.7	.88	.26	.81
C_β	11.0	1.0	.91	.22	.86
$C_{\gamma 1}$	37.0	3.2	.92	.17	.84
$C_{\gamma 2}$	58.4	7.6	.87	.23	.77
$C_{\delta 1}$	43.7	8.1	.82	.33	.74
S-C	150.1	19.8	.87	.19	.75
12. Leu (770)	193.1	26.6	.87	.18	.72
N	4.9	0.6	.91	.27	.87
C_α	5.9	0.9	.85	.36	.80
C	1.1	0.2	.87	.39	.84
O	23.4	3.2	.87	.26	.78
C_β	23.1	2.6	.89	.22	.81
C_γ	9.6	1.2	.87	.31	.83
$C_{\delta 1}$	63.4	8.6	.87	.23	.77
$C_{\delta 2}$	61.7	9.3	.85	.25	.75
S-C	157.8	21.7	.86	.19	.73
13. Lys (704)	225.8	112.4	.51	.20	.19
N	5.5	0.8	.88	.30	.83
C_α	8.9	3.2	.65	.46	.55
C	1.1	0.3	.77	.50	.73
O	23.2	6.0	.75	.35	.63
C_β	25.9	7.5	.71	.31	.53
C_γ	22.6	9.9	.57	.42	.42
C_δ	28.0	14.2	.50	.38	.32
C_ϵ	36.0	20.7	.43	.36	.24
N_ζ	74.6	49.8	.34	.30	.14
S-C	187.1	102.2	.45	.22	.16
14. Met (165)	203.4	34.5	.84	.22	.68
N	5.2	1.0	.90	.33	.86
C_α	8.2	1.8	.79	.40	.72

(continued)

TABLE I. Mean and Standard State Surface Areas for the Amino Acid Residues, Their Constituent Atomic Groups, and Their Side Chains* (Continued)

	Residue	A^0	$<A>$	$(A^0-<A>)/A^0$	σ	Skewness
	C	1.1	0.2	.84	.41	.79
	O	24.1	3.2	.87	.24	.79
	C_β	24.0	3.9	.84	.28	.75
	C_γ	29.6	4.3	.86	.26	.76
	S_δ	36.4	5.6	.85	.27	.75
	C_ϵ	74.8	14.6	.81	.30	.70
	S-C	164.8	28.4	.83	.24	.69
15.	Phe (389)	222.8	28.8	.88	.17	.73
	N	4.8	0.6	.90	.30	.86
	C_α	9.2	1.4	.86	.29	.78
	C	1.2	0.2	.86	.41	.84
	O	23.2	2.8	.88	.26	.82
	C_β	29.3	3.3	.89	.23	.82
	C_γ	0.6	0.0	.93	.23	.90
	$C_{\delta1}$	22.1	2.5	.89	.24	.81
	$C_{\epsilon1}$	36.7	4.5	.88	.23	.79
	C_ζ	37.7	5.5	.86	.24	.76
	$C_{\epsilon2}$	36.1	5.1	.86	.24	.76
	$C_{\delta2}$	21.9	2.9	.87	.25	.78
	S-C	184.4	28.8	.84	.18	.74
16.	Pro (448)	146.8	56.7	.62	.26	.36
	N	0.4	0.1	.78	.54	.76
	C_α	13.2	4.5	.66	.43	.58
	C	0.6	0.2	.68	.57	.64
	O	21.4	5.9	.73	.37	.62
	C_β	39.1	16.7	.58	.36	.41
	C_γ	44.3	19.8	.56	.36	.38
	C_δ	27.6	9.5	.66	.34	.48
	S-C	111.0	45.9	.59	.30	.36
17.	Ser (874)	129.8	47.9	.64	.28	.40
	N	6.0	1.2	.83	.34	.74
	C_α	12.3	4.6	.63	.44	.54
	C	1.2	0.3	.77	.49	.73
	O	24.5	6.5	.74	.36	.63
	C_β	44.3	18.8	.58	.37	.42
	O_γ	41.5	16.5	.61	.36	.44
	S-C	85.8	35.3	.59	.32	.38
18.	Thr (704)	152.5	49.5	.68	.24	.43
	N	3.8	0.7	.84	.39	.79
	C_α	9.0	2.9	.69	.44	.61
	C	1.1	0.2	.81	.49	.79
	O	24.0	5.1	.79	.33	.70
	C_β	18.0	5.1	.72	.34	.59
	$O_{\gamma1}$	33.5	11.9	.65	.37	.49
	$C_{\gamma2}$	63.1	23.6	.63	.33	.43
	S-C	114.6	40.6	.65	.27	.41
19.	Trp (157)	266.3	34.9	.87	.17	.73
	N	5.1	0.3	.93	.16	.87
	C_α	7.3	1.4	.81	.42	.76
	C	1.2	0.1	.90	.37	.88
	O	23.8	2.6	.90	.23	.83
	C_β	30.1	3.3	.89	.24	.82
	C_γ	1.3	0.1	.94	.18	.91
	$C_{\delta1}$	30.2	5.0	.84	.26	.72
	$N_{\epsilon1}$	29.8	4.8	.84	.23	.71
	$C_{\epsilon2}$	3.3	0.3	.90	.21	.82
	$C_{\zeta2}$	38.5	6.0	.85	.23	.73
	$C_{\eta2}$	37.9	5.1	.87	.24	.77
	$C_{\zeta3}$	34.1	3.6	.90	.22	.83
	$C_{\epsilon3}$	21.5	2.0	.91	.21	.85
	$C_{\delta2}$	2.2	0.3	.88	.28	.84
	S-C	228.9	30.4	.87	.18	.73
20.	Tyr (388)	236.8	54.9	.77	.20	.54

(continued)

TABLE I. Continued

	Residue	A^0	$<A>$	$(A^0-<A>)/A^0$	σ	Skewness
	N	4.8	0.7	.88	.32	.84
	C_α	8.9	1.7	.81	.35	.74
	C	1.3	0.3	.83	.46	.81
	O	23.7	4.3	.82	.30	.73
	C_β	29.2	4.8	.84	.30	.76
	C_γ	0.4	0.1	.80	.44	.79
	$C_{\delta1}$	22.6	3.9	.83	.28	.73
	$C_{\epsilon1}$	34.8	7.8	.78	.27	.62
	C_ζ	2.6	0.6	.78	.34	.68
	O_η	52.9	17.7	.67	.31	.48
	$C_{\epsilon2}$	34.1	8.5	.76	.29	.60
	$C_{\delta2}$	21.5	4.7	.79	.32	.68
	S-C	198.1	48.0	.76	.21	.54
21.	Val (858)	164.5	24.6	.86	.20	.71
	N	4.3	0.6	.90	.30	.86
	C_α	7.3	1.2	.84	.34	.78
	C	1.0	0.1	.90	.34	.88
	O	23.5	3.1	.87	.26	.79
	C_β	11.1	1.5	.87	.30	.81
	$C_{\gamma1}$	57.4	9.6	.84	.26	.72
	$C_{\gamma2}$	59.9	8.5	.86	.23	.75
	S-C	128.4	19.6	.85	.21	.72

*A^0 is the standard state accessibility, $<A>$ the mean accessibility in proteins, $(A^0-<A>)/A^0$ the mean fraction buried, and σ the standard deviation of $(A^0-<A>)/A^0$. Skewness is the third moment of $(A^0-<A>)/A^0$ about the mean. The values in parentheses are the number of residues of each type in the data base. For Arg, Asp, and Glu, the most exposed side chain is taken to be the charged atom and is designated $N_{\eta2}$, $O_{\delta2}$, and $O_{\epsilon2}$, respectively.

TABLE II. Stochastic Standard State Area Versus Extended Standard State Area*

Residue	A^0	Extended	Difference (%)
Ala	118.1	115	3
Arg	256.0	225	14
Asn	165.5	160	4
Asp	158.7	150	6
Cys	146.1	135	8
Gln	193.2	180	7
Glu	186.2	190	2
Gly	88.1	75	15
His	202.5	195	4
Ile	181.0	175	3
Leu	193.1	170	12
Lys	225.8	200	11
Met	203.4	185	9
Phe	222.8	210	6
Pro	146.8	145	1
Ser	129.8	115	12
Thr	152.5	140	9
Trp	266.3	255	4
Tyr	236.8	230	3
Val	164.5	155	6

*Column A^0 contains each residue's stochastic standard state area from Table I. Column Extended contains the extended standard state areas due to Chothia,[9] and column Difference contains the percentage difference between A^0 and extended.

derived as described in the previous section. The linear equations together with standard errors and correlation coefficients, ρ, are, respectively:

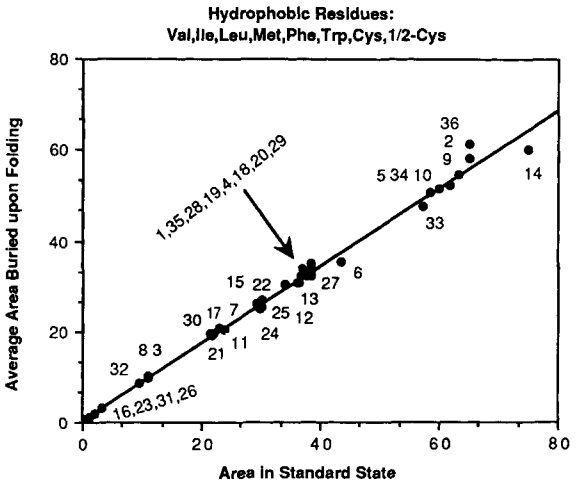

Fig. 1. The stochastic standard state area is plotted against mean area buried for hydrophobic groups from Table I. Labels on the points correspond to the following:

1.	Cys	C_β	13.	Met	S_δ	25.	Trp	$N_{\epsilon 1}$
2.	Cys	S_γ	14.	Met	C_ϵ	26.	Trp	$C_{\epsilon 2}$
3.	Ile	C_β	15.	Phe	C_β	27.	Trp	$C_{\zeta 2}$
4.	Ile	$C_{\gamma 1}$	16.	Phe	C_γ	28.	Trp	$C_{\eta 2}$
5.	Ile	$C_{\gamma 2}$	17.	Phe	$C_{\delta 1}$	29.	Trp	$C_{\zeta 3}$
6.	Ile	$C_{\delta 1}$	18.	Phe	$C_{\epsilon 1}$	30.	Trp	$C_{\epsilon 3}$
7.	Leu	C_β	19.	Phe	C_ζ	31.	Trp	$C_{\delta 2}$
8.	Leu	C_γ	20.	Phe	$C_{\epsilon 2}$	32.	Val	C_β
9.	Leu	$C_{\delta 1}$	21.	Phe	$C_{\delta 2}$	33.	Val	$C_{\gamma 1}$
10.	Leu	$C_{\delta 2}$	22.	Trp	C_β	34.	Val	$C_{\gamma 2}$
11.	Met	C_β	23.	Trp	C_γ	35.	1/2-Cys	C_β
12.	Met	C_γ	24.	Trp	$C_{\delta 1}$	36.	1/2-Cys	S_γ

The straight line of best fit for these points is given by Eq. (3) in the text.

$$(A^0-\langle A\rangle) = 0.86\ (\pm 0.01)\cdot A^0 + 0.30\ (\pm 0.48) \quad \rho = .99 \quad (3)$$

$$(A^0-\langle A\rangle) = 0.69\ (\pm 0.03)\cdot A^0 + 0.28\ (\pm 1.60) \quad \rho = .97 \quad (4)$$

$$(A^0-\langle A\rangle) = 0.65\ (\pm 0.02)\cdot A^0 - 0.20\ (\pm 0.70) \quad \rho = .99 \quad (5)$$

$$(A^0-\langle A\rangle) = 0.61\ (\pm 0.05)\cdot A^0 + 0.39\ (\pm 1.95) \quad \rho = .97 \quad (6)$$

$$(A^0-\langle A\rangle) = 0.48\ (\pm 0.09)\cdot A^0 - 4.75\ (\pm 4.83) \quad \rho = .98 \quad (7)$$

By way of illustration, the data corresponding to Eq. (3) have been plotted in Figure 1. Each point in the figure represents the standard state area versus area lost on folding for a single side chain hydrophobic group from Table I.

Equations (3)–(7) are of the form

$$\text{Average area buried} = m \cdot \text{Standard state area} + b$$

The slope, m, represents the increment of area that is buried, on average, for each corresponding increment of area in the standard state—i.e., how much a group buries versus how much it has to bury. All but the charged groups extrapolate back through 0 (i.e., $b \approx 0$), as would be expected for a series of alkanes.[13] The nonzero intercept for charged groups in Eq. (7), which results from fitting a line to three points, is probably due to experimental error.

Equation (3) is interpreted to mean that hydrophobic groups in hydrophobic residues bury 86% of their available surface area, on average. Equations (4)–(7) measure the composite tendencies of hydrophobic groups with polar or charged covalent neighbors. These four equations reflect the positional compromise that results when a surface-seeking polar or charged group is coupled to an interior-seeking hydrophobic group. On average, methylene groups that are immediately adjacent to a polar end-group bury 65% of their available surface [Eq. (5)], while nonadjacent methylenes bury 69% [Eq. (4)]. Polar end-groups bury 61% of their available surface [Eq. (6)], while charged end-groups bury only 48% [Eq. (7)].

Lysine appears more exposed than other charged residues, with atoms C_γ, C_δ, C_ϵ, and N_ϵ exhibiting anomalous values of $(A^0-\langle A\rangle)/A^0$, as shown in Table I. However, these values are probably due to side chain disorder, not anomalous exposure, and, for this reason, lysine was not included when calculating Eq. (7).

Three Characteristic Parameters for Globular Proteins

Figure 1 and Eq. (3) can be used to derive a set of self-consistent parameters, akin to those of Eisenberg and McLachlan.[15] A three-parameter fit to the side chain accessibility data summarized in Table I was made by assuming three classes of atoms: (1) carbon or sulfur (C/S), (2) neutral oxygen or nitrogen (O/N), and (3) charged oxygen or nitrogen (O^-/N^+). A representative parameter for atoms from class (1) can be extracted directly from Eq. (3); i.e., hydrophobic atoms are 86% buried on average. We define the *mean fractional area loss, f,* as

$$f \equiv (A^0 - \langle A\rangle)/A^0$$

and the *mean fractional area, f′,* as

$$f' \equiv \langle A\rangle/A^0 = 1 - f$$

and write either f(C/S) = 0.86 or f'(C/S) = 0.14.

Corresponding parameters for classes (2) and (3) cannot be extracted from Eqs. (4)–(7) because the slopes in these cases reflect a weighted average between coupled polar and apolar groups, as discussed above. The intrinsic accessibility of polar atoms can be deconvoluted from the influence of apolar covalent neighbors by assuming f(C/S) and calculating f(O/N) in a one-parameter fit to the observed data. That is, δ in the equation

$$\delta = \Sigma\ [\text{obs}(i)\text{-calc}(i)] \quad (8)$$

is minimized, where obs(i) is the observed (i.e., measured) surface area of the ith residue in the data

base and calc(i) is calculated from the standard state areas for each group within the ith residue as

$$\text{calc}(i) = \Sigma(A^0_{\text{carbon+sulfur}} \cdot f'(\text{C/S}) + \Sigma(A^0_{\text{oxygen+nitrogen}}) \cdot f'(\text{O/N})$$

For example, if the ith residue is an Asn, then calc(i) is written as

$$\text{calc}(i) = (C_\beta + C_\gamma) \cdot f'(\text{C/S}) + (O_{\delta 1} + N_{\delta 2}) \cdot f'(\text{O/N})$$

and, using $f'(\text{C/S}) = 0.14$ together with the standard state areas from Table I, is calculated as

$$\text{calc}(i) = (32.1 + 3.0) \cdot 0.14 + (31.1 + 59.1) \cdot f'(\text{O/N})$$

The value of $f'(\text{O/N})$ that minimizes Eq. (8) was computed for all Asn, Gln, Ser, Thr, and Tyr in the data base, a total of 2847 residues. (His residues, which may be charged, were excluded.) Calculated in this way, $f'(\text{O/N}) = 0.60$.

These values of $f'(\text{C/S})$ and $f'(\text{O/N})$ can then be used to derive $f'(O^-/N^+)$. The value of $f'(O^-/N^+)$ that minimizes Eq. 8 was computed for all 1437 Asp, Glu, and Arg residues in the data base. (Lys was excluded for reasons discussed previously.) Calculated in this way, $f'(O^-/N^+) = 0.68$.

Summarizing these three parameters:

1. $f(\text{C/S}) = 0.86$, i.e., hydrophobic groups are 86% buried, on average;
2. $f(\text{O/N}) = 0.40$, i.e., uncharged, polar groups are 40% buried, on average; and
3. $f(O^-/N^+) = 0.32$, i.e., charged groups are 32% buried, on average.

How well can this crude three-parameter fit estimate the measured mean accessibilities of amino acid residues? Table III and Figure 2 compare the mean accessibilities of residue side chains, $\langle A \rangle$, from Table I, with the sum of atomic accessibilities calculated as

$$\text{calc}(i) = \Sigma(A^0_{C+S}) \cdot f'(\text{C/S}) + \Sigma(A^0_{O+N}) \cdot f'(\text{ON}) + \Sigma(A^0_{O^-+N^+}) \cdot f'(O^-/N^+) \quad (9)$$

For all but four residues, the difference between the measured value and the three-parameter estimate is no more than 7 Å^2. Within the group of four outliers, Ala and Thr side chains are each ~10 Å^2 more exposed than would be expected from the three-parameter estimate, possibly because exposure of a single methyl group is only slightly disfavored. Pro is considerably more exposed (~30 Å^2) than would be anticipated from the three-parameter estimate, presumably because steric constraints override hydrophobic considerations in placing proline residues within reverse turns, which are typically situated at the solvent-accessible surface of proteins.[3] Finally, Lys is ostensibly more exposed than expected (~36 Å^2) due to the technical problems discussed previously.

TABLE III. Measured Versus Calculated Side Chain Areas*

Residue type	Measured area	Calculated area	Difference
Ala	20.9	10.1	−10.8
Arg	90.9	97.3	+6.4
Asn	52.7	58.6	+5.9
Asp	52.3	56.3	+4.0
Cys	10.2	14.5	+4.3
1/2-Cys	8.6	14.5	+5.9
Gln	64.0	64.7	+0.6
Glu	69.0	62.0	−7.0
Gly	0	0	0
His	46.7	41.8	−4.9
Ile	19.8	21.0	+1.2
Leu	21.6	22.1	+0.4
Lys	102.2	66.5	−35.7
Met	28.4	23.1	−5.3
Phe	23.8	25.8	+2.0
Pro	45.9	15.5	−30.4
Ser	35.3	30.9	−4.4
Thr	40.6	31.3	−9.3
Trp	30.4	32.0	+1.6
Tyr	48.0	51.8	+3.9
Val	19.6	18.0	−1.6

*Measured areas are the mean accessibilities of residue side chains from Table I. Calculated areas are computed by multiplying the standard state area for each group by an appropriate f'-value: $f'(\text{C/S}) = 0.14$; $f'(\text{O/N}) = 0.60$; $f'(O^+/N^-) = 0.68$. All areas are given in Å^2.

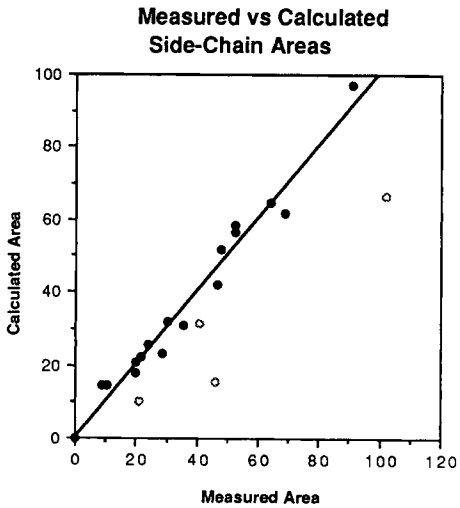

Fig. 2. Measured areas are the mean side chain surface areas, <A>, from Table I. Calculated areas were computed using Eq. (9) with the three atomic accessibility parameters: $f'(\text{C/S})$, $f'(\text{O/N})$, $f'(O^-/N^+)$. Filled circles correspond to the 17 best-fit residues: Arg, Asn, Asp, Cys, 1/2-Cys, Gln, Glu, Gly, His, Ile, Leu, Met, Phe, Ser, Trp, Tyr, and Val. For each, the difference between the measured and calculated area is not more than 7 Å^2. The straight line of best fit through these 17 points is shown. Ideally, the line would have a slope of 1 and an intercept of 0. The actual line is given by the equation:

Calculated area = 1.03 · Measured area (±0.04) + 0.07 (±1.67)

with $\rho = .99$. Open circles correspond to the four outliers: Ala, Lys, Pro, and Thr, as discussed in the text.

SUMMARY

The role of the hydrophobic effect in protein folding remains controversial.[2,26–28] The work of many investigators is based on the underlying assumption that the hydrophobic contribution to $\Delta G_{conformation}$ scales linearly with buried surface area.[9,10,12,24,29] Others have questioned the validity of such an approach.[27,30]

Our intent in this study has been to provide detailed measurements of group accessibilities from X-ray elucidated proteins and to demonstrate that the area such groups bury upon folding exhibits linear scaling by atom size and type. These observations may prove useful in subsequent quantitative studies of protein folding and energetics.

ACKNOWLEDGMENTS

We thank Ken Dill for useful discussion and helpful suggestions. This work was supported by grants from the National Institutes of Health (AG 06084 and GM 29458).

REFERENCES

1. Kauzmann, W. Some factors in the interpretation of protein denaturation. Adv. Prot. Chem. 14:1–64, 1959.
2. Baldwin, R.L. Temperature dependence of the hydrophobic interaction in protein folding. Proc. Nat. Acad. Sci., U.S.A. 83:8069–8072, 1986.
3. Rose, G.D., Gierasch, L.M., Smith, J.A. Turns in peptides and proteins. Adv. Prot. Chem. 37:1–109, 1985.
4. Cohn, E.J., Edsall, J.T. "Proteins, Amino Acids, and Peptides as Ions and Dipolar Ions." Princeton, NJ: Van Nostrand-Reinhold, 1943.
5. Nozaki, Y., Tanford, C. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. J. Biol. Chem. 246:2211–2217, 1971.
6. Wolfenden, R., Andersson, L., Cullis, P.M., Southgate, C.B. Affinities of amino acid side chains for solvent water. Biochemistry 20:849–855, 1983.
7. Fauchère, J.-L., Pliska, V.E. Hydrophobic parameters π for amino acid side chains from partitioning of N-acetyl-amino-acid amides. Eur. J. Med. Chem.-Chem. Ther. 18: 369–375, 1983.
8. Radzicka, A., Wolfenden, R. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and the neutral aqueous solution. Biochemistry 27:1664–1670, 1988.
9. Chothia, C. The nature of the accessible and buried surface in proteins. J. Mol. Biol. 105:1–14, 1976.
10. Janin, J. Surface and inside volumes in globular proteins. Nature (London) 277:491–492, 1979.
11. Lee, B.K., Richards, F.M. The interpretation of protein structures: Estimation of static accessibility. J. Mol. Biol. 55:379–400, 1971.
12. Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., Zehfus, M.H. Hydrophobicity of amino acid residues in globular proteins. Science 229:834–838, 1985.
13. Richards, F.M. Areas, volumes, packing and protein structure. Annu. Rev. Biophys. Bioeng. 6:151–176, 1977.
14. Guy, H.R. Amino acid side chain partition energies and distribution of residues in soluble proteins. Biophys. J. 47: 61–70, 1985.
15. Eisenberg, D., McLachlan, A.D. Solvation energy in protein folding and binding. 319:199–203, 1986.
16. Abraham, D.J., Leo, A.J. Extension of the fragment method to calculate amino acid zwitterion and side chain partition coefficients. Proteins: Structure, Function, Genet. 2:130–152, 1987.
17. Lawrence, C., Auger, I., Mannella, C. Distribution of accessible surfaces of amino acids in globular proteins. Proteins: Structure, Function, Genet. 2:153–161, 1987.
18. Roseman, M.A. Hydrophobicity of polar amino acid side chains is markedly reduced by flanking peptide bonds. J. Mol. Biol. 200:513–522, 1988.
19. Rose, G.D. Protein hydrophobicity: Is it the sum of its parts? Proteins: Structure, Function, Genet. 2:79–80, 1987.
20. Bernstein, F.C., Koetzle, T.G., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rogers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The protein data bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
21. Chothia, C. Structural invariants in protein folding. Nature (London) 254:304–308, 1975.
22. Shrake, A., Rupley, J.A. Environment and exposure to solvent of protein atoms: Lysozyme and insulin. J. Mol. Biol. 79:351–372, 1973.
23. International Mathematical Subroutine Library, 7500 Bellaire Blvd., Houston, Texas 77036.
24. Miller, S., Janin, J., Lesk, A.M., Chothia, C. Interior and surface of monomeric proteins. 196:641–656, 1987.
25. Snedecor, G.W., Cochran, W.G. "Statistical Methods." Ames: Iowa State Univ. Press, 1978.
26. Privalov, P.L., Gill, S.J. Stability of protein structure and hydrophobic interaction. Adv. Prot. Chem. 39:193–234, 1988.
27. Wood, R.H., Thompson, P.T. Differences between pair and bulk hydrophobic interactions. Proc. Natl. Acad. Sci. U.S.A., 87:946–949, 1990.
28. Spolar, R.S., Jeung-Hoi, H., Record, J.T., Jr. Hydrophobic effect in protein folding and other noncovalent processes involving proteins. Proc. Natl. Acad. Sci. U.S.A. 86:8382–8385, 1989.
29. Kyte, J., Doolittle, R.F. A simple method for displaying the hydropathic character of a protein sequence. J. Mol. Biol. 157:105–132, 1982.
30. Pratt, L.R., Chandler, D. J. Theoretical and computational studies of hydrophobic interactions. Chem. Phys. 73:3430–3433, 1980.