Empirical Studies of Hydrophobicity. 1. Effect of Protein Size on the Hydrophobic Behavior of Amino Acids[1a]

H. Meirovitch, S. Rackovsky, and H. A. Scheraga*[1b]

Baker Laboratory of Chemistry, Cornell University, Ithaca, New York 14853.
Received November 16, 1979

ABSTRACT: The hydrophobic properties of the 20 naturally occurring amino acids are studied in a sample of 19 proteins, using the average reduced distance from the center of mass, $\langle r \rangle$, and average side-chain orientational angle, $\langle \theta \rangle$, previously defined by Rackovsky and Scheraga. Separate hydrophobicity scales are established in terms of each variable, and the classifications resulting from the two scales are compared. The behavior of each amino acid with respect to each of the two parameters is also studied in two subgroups of proteins, consisting of smaller and larger proteins, respectively. In general, consistency is observed between the two types of hydrophobicity scales, and, within a given scale, between smaller and larger proteins. Certain discrepancies are observed, however. On the basis of a detailed study of these discrepancies, three amino acids—Gly, Ala, and Tyr—are shown to fall into an ambivalent category. It is suggested that this effect arises because of competition between hydrophobic and hydrophilic groups on the same residue. Study of $\langle r \rangle$ and $\langle \theta \rangle$ leads to certain general conclusions about protein structure: (1) Hydrophobic behavior, measured in terms of $\langle r \rangle$ and $\langle \theta \rangle$, manifests itself more strongly in smaller proteins, probably because of greater geometric constraints and larger surface-to-volume ratio in the smaller proteins. (2) In the larger proteins, there is an inner sphere of radius $\sim 0.7R_g$ (where R_g is the root-mean-square radius of gyration) in which there is no average orientational preference, since the degree of hydrophobicity is essentially constant within this sphere. (3) The *orientational* preferences of C^α–C^β bonds are qualitatively similar to those of the side chains as a whole; this indicates strong correlation between backbone and side-chain orientations. The results of this work are compared with the hydrophobicity scales of Nozaki, Tanford, and Jones, of Manavalan and Ponnuswamy, of Wertz and Scheraga, and of Chothia and with the side-chain interaction parameters of Krigbaum and Komoriya. Consistency is found with the results of Manavalan and Ponnuswamy, of Wertz and Scheraga, and of Krigbaum and Komoriya. Possible sources of discrepancy with the other two scales are discussed.

Introduction

The hydrophobic interaction is one of the principal driving forces responsible for the folding of globular proteins to their native (biologically active) structures.[2,3] In the model of Némethy and Scheraga,[3] the free energy of the hydrophobic interaction (i.e., the association of nonpolar groups in water) has contributions primarily from the balance of van der Waals interactions between all components of the system and from changes in the structure of water around the interacting nonpolar moieties.[4] Such an interaction may be regarded as the *reversal* of the process in which nonpolar groups are transferred from a nonpolar solvent to water.[3]

An early view of protein organization[2] postulated that nonpolar amino acids are buried to form a hydrophobic core, which is surrounded by an outer layer of polar residues in contact with solvent.[4] It was shown, however, by Klotz[5] and by Lee and Richards[6] that, on an atomic basis, a substantial portion of the exposed solvent-accessible surface of proteins is composed of nonpolar groups. They also showed that many polar groups are inaccessible to solvent in the molecular interior. Kuntz[7] and Chothia[8] demonstrated that a large proportion of these polar groups are involved in hydrogen bonding.

These observations emphasize the importance of the distinction that should be drawn between hydrophobicity of the entire residue and polarity, which is a property that is definable on an atomic basis. Hydrophobicity has been defined either experimentally, by studying the free energy of transfer of amino acids from water to organic solvents,[9,10] or empirically, by examination of X-ray structures of proteins. The latter approach can be carried out in various ways[6,11-16] and is itself different in principle from the experimental approach, *which does not reflect the influence of chain connectivity and other interactions.* For example, on an atomic basis, proline is one of the least polar of the amino acids. By the criterion of free energy of transfer, it is also highly hydrophobic.[9,10] By any protein structural criterion, however, it is one of the amino acids most likely to occur on the protein exterior, in contact with solvent. On an empirical basis, therefore, Pro is a hydrophilic amino acid.

Clearly, hydrophobicity scales derived on an atomic basis and those resulting from the various whole-residue approaches can be expected to differ, at least in some details. Furthermore, solvent accessibility and location of the molecular surface are factors which are difficult to incorporate into a practical protein-folding algorithm, in which one would like to include some representation of the hydrophobic effect. In a recent publication,[11] Rackovsky and Scheraga examined the radial distribution of amino acids about the center of mass of proteins as a means of representing the hydrophobic effect. This approach was actually used, in a simplified manner, in folding algorithms by Kuntz et al.[17] and by Ycas et al.[18] It should be noted that, because of the irregular shape of proteins, there is only an approximate correlation between the radial distribution and solvent accessibility or surface location. Nevertheless, it is to be expected that amino acids will tend to distribute themselves radially in keeping with their relative hydrophobicities, so that examination of the distribution functions for a representative sample of proteins should lead to an appropriately detailed, practically useful hydrophobicity scale.

It was found[11] that there was indeed a correlation between the average reduced distance $\langle r \rangle$ (where r is the actual distance from the center of mass divided by the radius of gyration of the protein) of a given amino acid and generally accepted notions of hydrophobicity. Another important result was obtained by comparing *radial* distribution functions of C^α atoms with those of appropriate side-chain atoms for each amino acid. It was found that, in general, these distributions differ significantly. This implies that it is necessary to include the information contained in both distributions in order to represent the hydrophobicity of a given amino acid properly.

A new parameter was therefore defined, viz., θ, the angle between the center-of-mass-to-C^α and C^α-to-side-chain-atom vectors, and its distribution was investigated. There proved to be a correlation between θ and hydrophobicity.

0024-9297/80/2213-1398$01.00/0

Amino acids which are generally considered hydrophilic tend to have distributions of θ peaked at $\theta < 90°$ (side chain pointing outward). Hydrophobic residues have distributions which peak at $\theta > 90°$ (side chain pointing inward). It should be noted that θ is a local parameter, which can be used to characterize side-chain behavior as a function of position.

It is of interest to investigate the correlation between r and θ. Since local environment varies with r (presumably becoming more polar on the average as r increases), θ is expected to change accordingly. It is also of interest to compare the behavior of the two parameters for each amino acid in smaller and larger proteins. In smaller proteins the surface-to-volume ratio is larger, and one expects that solvent effects will be more marked. As the radius of gyration increases, one expects to find a "dead" volume around the center of mass which is effectively shielded from solvent. It will be of interest to see whether this effect exists, and in what form it is manifested.

The answers to these questions will provide important insights into the organization of proteins. They are also of very practical importance in formulating protein-folding algorithms. None of the folding algorithms published heretofore[17-19] take side-chain orientation into account in treating hydrophobic interactions, nor do they consider the relationship between r and θ and protein size. It is to be expected that these factors will have an important influence on the final structure that results from the folding of the polypeptide chain.

The present work is aimed at the investigation of these questions. We use a larger sample of proteins than was used in ref 11, together with more detailed statistical procedures. We shall examine the behavior of $\langle r \rangle$ and $\langle \theta \rangle$ in smaller and larger proteins, and, within the latter group, we shall investigate $\langle \theta \rangle$ in inner and outer layers. The results will be compared with the hydrophobicity scales of Nozaki and Tanford,[9] Jones,[10] Manavalan and Ponnuswamy,[12] Wertz and Scheraga,[13] and Chothia[14] and with the side-chain interaction parameters of Krigbaum and Komoriya.[15,16,20]

Methods

In the present study 19 proteins were examined. Their coordinates were obtained from the Protein Data Bank at Brookhaven National Laboratory. The proteins are listed in Table I, ordered according to the number of amino acid residues. Their radii of gyration are also given in the table. This sample was chosen to satisfy three criteria: (i) that reliable X-ray coordinates be available; (ii) that the members of the set not be excessively homologous or closely related; and (iii) that the members of the set be mainly single unit proteins. This last criterion was established because of the possibility that multisubunit proteins might have rather specialized distributions of amino acids in order to bring about the necessary association of subunits. For each amino acid we are interested in two parameters (see ref 11). The first one is the average distance of a particular atom of the amino acid from the center of mass of the protein. In order to scale the results obtained from a sample of proteins of different molecular weights, we define the reduced distance, r, by dividing the actual distance R by the root-mean-square radius of gyration R_g, of the protein

$$r = R/R_g \tag{1}$$

For practical reasons, the reduced distance r is divided into discrete intervals of length 0.05. Reduced distances which fall within the limits of each interval are equated to the reduced distance at the center of the interval. It should be pointed out that, even for the largest protein in the sample, $0.05R_g$ is less than the experimental resolution of the coordinates, so that no serious errors should arise from this procedure. For a given amino acid, the average value of r is therefore estimated[11] by

$$\langle r \rangle = \sum_k P_k[0.05(k-1) + 0.025] \equiv \sum_k P_k r_k \tag{2}$$

where

$$P_k = N_k/N \tag{3}$$

N is the total number of residues of the given type in the sample of proteins, N_k is the number of residues of the same type in the interval k, and r_k is the reduced radius at the center of the kth interval.

Small values of $\langle r \rangle$ indicate a tendency for the amino acid to appear in the interior of a protein, which is interpreted as indicating hydrophobic character. On the other hand, hydrophilic amino acids are characterized by larger values of $\langle r \rangle$. In this study (as in ref 11 and 15),[20] we compute two values of $\langle r \rangle$ for each of the amino acids, one for C^α and the other for a remote side-chain atom (for details see Table II). In this context, we also define the estimated standard deviation of r, $\langle \Delta r^2 \rangle^{1/2}$, which measures the distribution of the reduced distances around the average value $\langle r \rangle$

$$\langle \Delta r^2 \rangle^{1/2} = \{\sum_k P_k(r_k - \langle r \rangle)^2\}^{1/2} \tag{4}$$

The estimated average $\langle r \rangle$, for each one of the amino acids, is obtained from a relatively small sample of size N. Therefore, $\langle r \rangle$ is subject to statistical fluctuation about the "correct" average value which would have been obtained from a very large sample. This fluctuation, which we denote by $\langle \Delta r^2 \rangle_N^{1/2}$, can be estimated easily if one assumes that the amino acids are sampled independently; in this case[21]

$$\langle \Delta r^2 \rangle_N^{1/2} = \langle \Delta r^2 \rangle^{1/2}/N^{1/2} \tag{5}$$

Table I
Proteins Used in This Work and Their Root-Mean-Square Radii of Gyration

	protein	no. of amino acids	R_g,[a] Å
	Group A[b]		
1	rubredoxin	54	10.00
2	ferredoxin	54	9.61
3	bovine pancreatic trypsin inhibitor	58	10.96
4	oxidized chromatium high-potential iron protein	85	11.65
5	tuna cytochrome *c* (oxidized)	103	12.68
6	calcium-binding parvalbumin	108	12.73
7	ribonuclease S	124	14.56
8	hen egg white lysozyme	129	13.96
9	flavodoxin (oxidized)	138	13.84
10	staphylococcal nuclease	142	14.82
11	sperm whale myoglobin	153	15.25
	Group B[b]		
12	adenylate kinase	194	17.34
13	papain	212	16.33
14	concanavalin A	237	17.23
15	chymotrypsinogen A	245	16.14
16	subtilisin BPN'	275	16.77
17	carboxypeptidase A	307	17.98
18	thermolysin	316	19.59
19	lactate dehydrogenase apoenzyme M4	329	21.65

[a] Root-mean-square radius of gyration. [b] These two groups constitute an arbitrary classification into small and large proteins, respectively.

Table II
Values of $\langle r_\alpha \rangle$ and $\langle r_s \rangle$ for Smaller Proteins (Group A), Larger Proteins (Group B), and Total Sample (Group C)

amino acid	side-chain atom[a]	group A		group B		group C[b]		
		$\langle r_\alpha \rangle_A$	$\langle r_s \rangle_A$	$\langle r_\alpha \rangle_B$	$\langle r_s \rangle_B$	$\langle r_\alpha \rangle_C$	$\langle r_s \rangle_C$	$\langle \Delta r_\alpha^2 \rangle_C^{1/2}$
Ala	C^β	0.99 ± 0.03[c]	1.00 ± 0.03	0.89 ± 0.03	0.89 ± 0.03	0.93 ± 0.02	0.94 ± 0.02	0.31
Asp	O^{δ_1}	1.06 ± 0.03	1.15 ± 0.03	0.99 ± 0.03	1.04 ± 0.03	1.01 ± 0.02	1.08 ± 0.03	0.27
Cys	S^γ	0.88 ± 0.04	0.81 ± 0.05	0.88 ± 0.06	0.90 ± 0.06	0.88 ± 0.03	0.84 ± 0.04	0.26
Glu	O^{ϵ_1}	1.02 ± 0.03	1.15 ± 0.03	1.01 ± 0.04	1.09 ± 0.04	1.02 ± 0.02	1.12 ± 0.03	0.27
Phe	C^ζ	0.71 ± 0.04	0.62 ± 0.04	0.82 ± 0.04	0.81 ± 0.04	0.78 ± 0.03	0.73 ± 0.03	0.26
Gly		1.09 ± 0.03		0.95 ± 0.02		1.00 ± 0.02		0.31
His	N^{ϵ_2}	0.87 ± 0.06	0.91 ± 0.07	0.91 ± 0.06	0.93 ± 0.06	0.89 ± 0.04	0.92 ± 0.05	0.33
Ile	C^{δ_1}	0.80 ± 0.04	0.74 ± 0.04	0.79 ± 0.03	0.78 ± 0.03	0.79 ± 0.02	0.76 ± 0.02	0.27
Lys	N^ζ	1.04 ± 0.02	1.24 ± 0.02	1.07 ± 0.03	1.23 ± 0.03	1.05 ± 0.02	1.23 ± 0.02	0.22
Leu	C^{δ_1}	0.78 ± 0.04	0.73 ± 0.04	0.88 ± 0.03	0.86 ± 0.03	0.85 ± 0.02	0.82 ± 0.03	0.31
Met	C^ϵ	0.80 ± 0.05	0.77 ± 0.06	0.87 ± 0.04	0.88 ± 0.04	0.84 ± 0.03	0.83 ± 0.04	0.21
Asn	C^γ	0.93 ± 0.04	1.02 ± 0.04	1.01 ± 0.03	1.05 ± 0.03	0.98 ± 0.02	1.04 ± 0.02	0.26
Pro	C^γ	1.05 ± 0.05	1.10 ± 0.05	0.97 ± 0.04	1.00 ± 0.04	1.00 ± 0.03	1.04 ± 0.03	0.28
Gln	C^δ	0.96 ± 0.04	1.09 ± 0.04	1.04 ± 0.03	1.12 ± 0.03	1.02 ± 0.03	1.11 ± 0.03	0.23
Arg	N^{η_1}	1.00 ± 0.05	1.12 ± 0.06	0.98 ± 0.03	1.07 ± 0.04	0.98 ± 0.03	1.09 ± 0.03	0.24
Ser	O^γ	1.06 ± 0.04	1.10 ± 0.04	1.00 ± 0.03	1.03 ± 0.03	1.02 ± 0.02	1.04 ± 0.02	0.32
Thr	C^{γ_2}	0.95 ± 0.04	0.99 ± 0.04	1.01 ± 0.03	1.03 ± 0.03	0.99 ± 0.03	1.02 ± 0.03	0.33
Val	C^{γ_1}	0.84 ± 0.03	0.87 ± 0.04	0.79 ± 0.03	0.79 ± 0.03	0.81 ± 0.02	0.81 ± 0.02	0.27
Trp	C^{η_2}	0.83 ± 0.07	0.83 ± 0.08	0.83 ± 0.04	0.90 ± 0.04	0.83 ± 0.04	0.87 ± 0.04	0.24
Tyr	O^η	0.86 ± 0.05	0.97 ± 0.05	0.95 ± 0.03	1.06 ± 0.03	0.93 ± 0.03	1.03 ± 0.03	0.25

[a] This is the side-chain atom used to calculate $\langle r_s \rangle$ for each amino acid. [b] Values of the standard deviation $\langle \Delta r_\alpha^2 \rangle^{1/2}$ are given only for group C. [c] These are the fluctuations $\langle \Delta r^2 \rangle_N^{1/2}$ defined in eq 5.

where $\langle \Delta r^2 \rangle^{1/2}$ is estimated with the help of eq 4. We shall use eq 5 because of its simplicity even though the sampling cannot be proved to be independent.

The second useful parameter is the angle θ, the angle between vectors from the center of mass to C^α and from C^α to a remote side-chain atom (see Figure 3 of ref 11). Clearly, when $\theta > 90°$, the side chain is oriented toward the interior of the protein, whereas $\theta < 90°$ means that the side chain tilts to the outside. For each one of the amino acids, one can calculate, from the sample of proteins, the average $\langle \theta \rangle$. Amino acids having $\langle \theta \rangle > 90°$ will be considered as hydrophobic since their side chains prefer to orient toward the inside of the proteins. From similar arguments, the hydrophilic side chains will be those with $\langle \theta \rangle < 90°$. For practical reasons, we divide the range 0–180° into 50 intervals of 3.6° each. Angles which fall within the limits of such an interval are equated to the angle at the center of the interval. For a given amino acid we therefore obtain

$$\langle \theta \rangle = \sum_{k=1}^{50} q_k[3.6(k-1) + 1.8] \equiv \sum_{k=1}^{50} q_k \theta_k \qquad (6)$$

where

$$q_k = n_k/N \qquad (7)$$

n_k is the number of occurrences of the side-chain angle θ in the kth interval.

As in the case of the distances, we can estimate the standard deviation $\langle \Delta\theta^2 \rangle^{1/2}$

$$\langle \Delta\theta^2 \rangle^{1/2} = \{\sum_{k=1}^{50} q_k[\theta_k - \langle \theta \rangle]^2\}^{1/2} \qquad (8)$$

and also the fluctuation in $\langle \theta \rangle$, $\langle \theta^2 \rangle_N^{1/2}$

$$\langle \Delta\theta^2 \rangle_N^{1/2} = \langle \Delta\theta^2 \rangle^{1/2}/N^{1/2} \qquad (9)$$

Results and Discussion

A. Classification by $\langle r \rangle$. In Table II we present the values of $\langle r \rangle$ for the 20 amino acids, calculated with the help of eq 2 and 3. These values are computed twice: for the C^α atom ($\langle r_\alpha \rangle$) and for a remote side-chain atom—$\langle r_s \rangle$ (the choice of this atom for each amino acid appears in the second column of Table II). The sampling is carried out over three groups of proteins: 1–11, 12–19, and 1–19, of Table I. We denote these groups by A, B, and C, respectively. The first group contains small proteins (R_g < 15.5 Å) and the second, larger proteins. The estimated error in $\langle r \rangle$, $\langle \Delta r^2 \rangle_N^{1/2}$, was calculated by means of eq 4 and 5. We also give values of $\langle \Delta r_\alpha^2 \rangle_C^{1/2}$ (eq 4) for the group 1–19.

We shall use the values of $\langle r_\alpha \rangle_C$ obtained for the sample 1–19 to classify the amino acids into two main groups—hydrophobic and hydrophilic. The hydrophobic range is defined by $\langle r \rangle \leq 0.88$, and the hydrophilic range by $\langle r \rangle > 0.95$. Values of $\langle r \rangle$ in the interval $0.88 < \langle r \rangle \leq 0.95$ will be regarded as neutral. The values of $\langle r \rangle$ chosen to define the two groups are arbitrary to some extent. According to the above definition, the following seven amino acids are considered hydrophobic: Cys, Phe, Ile, Leu, Met, Val, and Trp. The hydrophilic group includes ten amino acids: Asp, Glu, Gly, Lys, Asn, Pro, Gln, Arg, Ser, and Thr. Three amino acids, Ala, His, and Tyr, fall into the neutral range. The above analysis applied to $\langle r_s \rangle_C$ leads to the same results with the exception that, by the side-chain criterion, Tyr falls into the hydrophilic rather than the neutral range. Table II shows that, for the hydrophobic amino acids, in most cases $\langle r_s \rangle < \langle r_\alpha \rangle$, and the opposite occurs for the hydrophilic residues, viz., $\langle r_s \rangle > \langle r_\alpha \rangle$. This behavior, which has been noted previously,[11] reflects the tendency of a hydrophobic side chain to orient toward the interior hydrophobic core of the protein and of hydrophilic side chains to orient toward the outer hydrophilic shell and the surrounding water. The difference between $\langle r_s \rangle$ and $\langle r_\alpha \rangle$ is larger in group A than in group B. This is because this difference is measured in units of R_g, which is smaller in group A.

As was remarked in the Introduction, a larger fraction of the residues in small proteins is accessible to the surrounding water than in large ones, and this might affect the relative distances from the center of mass of hydrophobic and hydrophilic residues. It is therefore interesting to compare the values of $\langle r_\alpha \rangle$ for groups A and B. As can be seen from Table II, for most of the amino acids there is no difference in classification between the two samples. This suggests that the classification of these amino acids

Table III
Values of $\langle\theta\rangle$ for the Three Groups of Proteins[a]

amino acid	group A $\langle\theta\rangle_A$, deg	group B $\langle\theta\rangle_B$, deg	group B $\langle\theta\rangle_{in}$, deg	group B $\langle\theta\rangle_{out}$, deg	group C $\langle\theta\rangle_C$, deg	group C $\langle\theta_\beta\rangle_C$, deg
Ala	81 ± 4[b]	91 ± 3	91 ± 5	91 ± 4	87 ± 3	87 ± 3
			97 ± 4	74 ± 6		
Asp	69 ± 5	72 ± 4	82 ± 7	68 ± 4	71 ± 3	72 ± 3
Cys	113 ± 6	88 ± 8	83 ± 11	93 ± 10	104 ± 5	103 ± 5
Glu	71 ± 4	73 ± 4	77 ± 8	72 ± 5	72 ± 3	73 ± 3
Phe	120 ± 6	100 ± 5	86 ± 8	111 ± 6	108 ± 5	109 ± 4
His	92 ± 8	90 ± 6	89 ± 10	90 ± 8	90 ± 5	93 ± 5
Ile	114 ± 5	100 ± 4	94 ± 5	107 ± 5	105 ± 3	106 ± 3
Lys	67 ± 3	62 ± 3	89 ± 11	60 ± 3	65 ± 2	71 ± 3
Leu	112 ± 5	101 ± 4	96 ± 5	104 ± 4	104 ± 3	103 ± 3
			103 ± 4	89 ± 8		
Met	107 ± 8	95 ± 7	80 ± 11	103 ± 8	100 ± 5	95 ± 5
Asn	64 ± 5	73 ± 4	86 ± 10	70 ± 4	70 ± 3	68 ± 3
Pro	78 ± 6	78 ± 4	85 ± 9	76 ± 5	78 ± 4	71 ± 3
Gln	63 ± 5	68 ± 4	90 ± 8	64 ± 4	66 ± 3	72 ± 4
Arg	85 ± 6	79 ± 5	85 ± 10	76 ± 6	81 ± 4	85 ± 4
Ser	87 ± 5	82 ± 3	95 ± 6	77 ± 6	83 ± 3	79 ± 3
Thr	80 ± 5	84 ± 4	91 ± 7	82 ± 4	83 ± 3	81 ± 3
Val	88 ± 5	96 ± 3	100 ± 4	92 ± 4	94 ± 3	93 ± 3
			98 ± 3	83 ± 8		
Trp	104 ± 9	90 ± 6	83 ± 8	97 ± 8	94 ± 5	96 ± 6
Tyr	90 ± 6	80 ± 4	74 ± 8	82 ± 4	83 ± 3	96 ± 3

[a] For group A, only $\langle\theta\rangle_A$ is given because this group does not contain enough residues to provide statistical significance for $\langle\theta\rangle_{in}$ and $\langle\theta\rangle_{out}$. For group B, $\langle\theta\rangle_B$, $\langle\theta\rangle_{in}$, and $\langle\theta\rangle_{out}$ (see text) are given. For group C, $\langle\theta\rangle_C$ and $\langle\theta_\beta\rangle_C$ (see text) are given. For Ala, Leu, and Val, values of $\langle\theta\rangle_{in}$ and $\langle\theta\rangle_{out}$ are also given with $R = 1.1R_g$ (see text). [b] These are the fluctuations $\langle\Delta\theta^2\rangle_N^{1/2}$ defined in eq 9.

into hydrophobic, hydrophilic, and neutral groups, using the criterion of $\langle r_\alpha\rangle$, is consistent in the sense that it is not dependent on the size of the protein.

There is, however, a group of amino acids which do show significantly different behavior in the two groups. Two amino acids, Ala and Gly, exhibit a substantial decrease of $\langle r_\alpha\rangle$ on passing from group A (small proteins) to group B (larger proteins). Both move from the hydrophilic range in A (Gly very strongly so) to the neutral range (Ala almost hydrophobic) in B. Two other residues, Tyr and Asn, exhibit an increase of $\langle r_\alpha\rangle$ on going from A to B, with resulting transfers from one range to another. We shall defer classifying these residues until further data are analyzed in sections B and C.

It is interesting to point out that $\langle r_\alpha\rangle_A > \langle r_\alpha\rangle_B$ for Asp, Pro, and Ser, although they fall in the same class in both groups. This might be accounted for by the relatively frequent occurrence of these amino acids in bends.[22-24] Bends are known to appear more often on the surface of proteins than in the interior.[22,25] There is evidence to indicate that the number of bends per amino acid is greater in smaller proteins than in larger ones.[26] This might be the reason that larger fractions of these residues are located on the surface in small proteins than in large ones.

It should be noted that the relation $\langle r_\alpha\rangle_A < \langle r_\alpha\rangle_B$, which holds for the three hydrophobic amino acids Phe, Leu, and Met, is compatible with results for $\langle\theta\rangle$ which will be discussed later, suggesting that hydrophobicity is more marked in the smaller proteins than in the larger ones.

To conclude the discussion of Table II, we point out that the data in the table can be used in theoretical procedures for protein folding. Recently, several such procedures have been applied to simplified models of proteins,[17-19] two of them[17,18] based on optimizing an "objective" function which incorporates the available experimental data, such as limits on distances between C^α atoms, S atoms in disulfide bonds, and also the hydrophobic character of the amino acids. This last effect was taken into account[17-19] by imposing restrictions that forced the hydrophobic residues to be located, on the average, closer to the center of mass of the protein than the hydrophilic ones. The values for $\langle r_\alpha\rangle$ provided in Table II can be useful for such studies. Clearly, for a small protein one would use $\langle r_\alpha\rangle_A$ and for a large protein $\langle r_\alpha\rangle_B$.[27] We also provide the values of $\langle\Delta r_\alpha^2\rangle_C^{1/2}$ (eq 5), which can be incorporated in the objective function. These values are given only for the entire sample C because they do not differ much from the values for the other two samples but are more statistically reliable because of the larger sample size.

B. Classification by $\langle\theta\rangle$. In Table III we present the results for $\langle\theta\rangle$ calculated by eq 6 and 7. For each of the amino acids, the angle θ is defined with respect to the remote side-chain atom appearing in the second column of Table II. The statistical error $\langle\Delta\theta^2\rangle_N^{1/2}$ was estimated by means of eq 8 and 9. Table III contains results for the small proteins ($\langle\theta\rangle_A$), for the large ones ($\langle\theta\rangle_B$), and for the entire sample ($\langle\theta\rangle_C$). For this last sample we also provide results for $\langle\theta_\beta\rangle$, where θ_β is computed with respect to the β carbon instead of a remote side-chain atom. We also want to examine the changes occurring in $\langle\theta\rangle$ as a function of the distance r_α for the set of larger proteins (group B). Therefore we define inner and outer regions of the protein and calculate for them $\langle\theta\rangle_{in}$ and $\langle\theta\rangle_{out}$, respectively (the actual definitions of these regions will be given later).

From the results for $\langle\theta\rangle_C$ for sample C we shall again classify (somewhat arbitrarily) the amino acids into the three classes, hydrophobic, hydrophilic, and neutral. The hydrophobic amino acids are those with $\langle\theta\rangle_C > 95°$, which means that their side chains prefer to orient toward the center of mass of the protein. This group includes the five amino acids Cys, Phe, Ile, Leu, and Met. The hydrophilic group consists of the ten amino acids with $\langle\theta\rangle_C < 85°$ (i.e., those which prefer to orient toward the outside). They are Asp, Glu, Lys, Asn, Pro, Gln, Arg, Ser, Thr, and Tyr. The rest of the amino acids, Ala, His, Val, and Trp, are considered here neutral since their average θ does not show any preferred direction, $85° < \langle\theta\rangle_C < 95°$. Gly is not classified here since it does not have a side chain. It should be noted that the present classification differs from the previous one based on $\langle r_\alpha\rangle_C$, in which Val and Trp were

considered hydrophobic and Tyr was in the neutral range. We shall show below that this discrepancy can be removed by carrying out a more careful analysis which takes into account the results obtained for the two subgroups. Before going into this analysis, it should be pointed out that except for Tyr the values of $\langle\theta_\beta\rangle_C$ are almost equal to $\langle\theta\rangle_C$. This behavior was also observed in groups A and B. Since the orientation of C^β is determined by that of the backbone, this observation suggests that the side-chain *orientation* is correlated strongly to the backbone structure. The case of Tyr will be discussed later.

We now examine the results for the hydrophobic amino acids in the two subgroups. The values of $\langle\theta\rangle_A$ for Cys, Phe, Ile, and Leu are considerably higher (i.e., more hydrophobic) than the corresponding $\langle\theta\rangle_B$. The same tendency is exhibited by Met and Trp (the latter falling in the neutral range according to its value of $\langle\theta\rangle_C$), but no definite statement can be made for these because the values of $\langle\theta\rangle_A$ and $\langle\theta\rangle_B$ are equal within statistical error. This difference in the values of $\langle\theta\rangle$ for the two samples can be accounted for by geometric effects and hydrophobic interactions. Most of the hydrophobic side chains are large enough so that, in the smaller proteins, the need to avoid unfavorable contacts at the protein–solvent interface imposes more stringent geometrical requirements. They therefore tend to orient themselves more strongly toward the interior of the molecule. In the larger proteins, on the other hand, the side chains have more volume available to occupy without encountering solvent. The net result is that $\langle\theta\rangle_A > \langle\theta\rangle_B$.

We shall now attempt to confirm this hypothesis by examining the results obtained for $\langle\theta\rangle_{in}$ and $\langle\theta\rangle_{out}$—in the inner and outer domains of the protein, respectively, for the group of larger proteins. We begin by describing some features of the structure of the large proteins. In other work[29] it was found that, inside a sphere of radius $\sim 0.75R_g$ around the center of mass, the frequency of occurrence of the hydrophobic residues is much higher, and that of the hydrophilic residues is much lower, than their average frequencies of occurrence in the protein. These frequencies of occurrence are almost unchanged in smaller spheres around the center of mass ($R < 0.75R_g$) and in this sense we call the sphere with radius $0.75R_g$ "homogeneous". Outside this sphere, the proportion of hydrophobic residues decreases with increasing distance from the center of mass whereas the frequency of occurrence of the hydrophilic residues increases above their frequency of occurrence in the protein. The implication of the homogeneity discussed above is that a side chain of a residue which is located deep enough in the homogeneous sphere feels roughly the same environment in any direction and therefore should not show any preferred orientation. Therefore, for the homogeneous sphere, one would expect $\langle\theta\rangle_{in} \sim 90°$. Higher values can be expected for $\langle\theta\rangle_{out}$, for hydrophobic residues. The results in Table III for group B were obtained by dividing the protein into two layers with the boundary (for C^α) at $R = 0.7R_g$ (0.7 was chosen rather than 0.75 to reduce overlap of side chains with the outer layer). They appear to confirm the above hypothesis—i.e., that for hydrophobic amino acids, $\langle\theta\rangle_{out} > \langle\theta\rangle_{in} \sim 90°$ (for Cys and Leu the values of $\langle\theta\rangle_{in}$ and $\langle\theta\rangle_{out}$ are equal within the range of statistical error). It follows from this observation that, in large proteins, $\langle\theta\rangle_{out}$ more accurately reflects the hydrophobic or hydrophilic nature of the amino acid than $\langle\theta\rangle_B$ or $\langle\theta\rangle_{in}$. This conclusion applied to Trp shows that it behaves as a hydrophobic amino acid rather than a neutral one. This is in accord with the values of $\langle\theta\rangle_A$, $\langle r_\alpha\rangle_A$, and $\langle r_\alpha\rangle_B$ (Table II) for Trp, which are found to be in the hydrophobic range. We have therefore resolved the discrepancy between the $\langle r\rangle$ and $\langle\theta\rangle$ criteria with respect to Trp.

$\langle\theta\rangle_{in}$ and $\langle\theta\rangle_{out}$ have also been calculated for larger values of R. Our purpose was to follow the changes in $\langle\theta\rangle_{out}$ while approaching the surface of the protein. Significant changes in $\langle\theta\rangle_{out}$ were detected for only a few amino acids. In these cases, the results are shown in Table III (for $R = 1.1R_g$) under the values for $\langle\theta\rangle_{in}$ and $\langle\theta\rangle_{out}$ obtained with $R = 0.7R_g$.

The behavior of Val is of special interest. As we mentioned previously, values of $\langle r_\alpha\rangle$ in Table II clearly define it as a hydrophobic amino acid. On the other hand, the value of $\langle\theta\rangle_A$ falls in the neutral region. In the interior of the large proteins, valine behaves as a markedly hydrophobic amino acid ($\langle\theta\rangle_{in} \sim 100°$). However, on the extreme periphery of the large proteins, in a more polar environment (with R increased to $1.1R_g$) $\langle\theta\rangle_{out}$ decreases to 83°, in the hydrophilic region. Apparently the hydrophobic side chain of Val is not sufficiently flexible to completely block favorable interactions between the neighboring peptide groups and water. This might account for the surprisingly low values of $\langle\theta\rangle_A$ and $\langle\theta\rangle_{out}$. We shall see this phenomenon in more marked fashion in Ala. Despite this tendency to interact with solvent, we feel justified in classifying Val as hydrophobic on the basis of the hydrophobic values of $\langle r_\alpha\rangle$ and $\langle\theta\rangle_{in}$; clearly, though, both manifestations of the behavior of Val must be incorporated in a protein-folding algorithm.

By all criteria of $\langle\theta\rangle$, Asn appears to be hydrophilic. Though $\langle r_\alpha\rangle$ falls in the upper part of the neutral range for group A, it is well within the hydrophilic range for group B, but the difference is relatively small and subject to some question because of the statistical uncertainty. We therefore classify Asn as hydrophilic.

To summarize, according to the distance analysis (Table II), we have defined seven hydrophobic amino acids: Phe, Ile, Leu, Met, Trp, Cys, and Val. The first six also show consistent hydrophobic behavior with respect to $\langle\theta\rangle$ in all the samples studied, even though the values of $\langle\theta\rangle_A$ are significantly higher than those of $\langle\theta\rangle_{out}$. We explain this fact as arising from the more stringent geometrical and energetic constraints acting in the smaller proteins. Only Val behaves in an inconsistent manner (with respect to $\langle\theta\rangle$) in that, when located near the surface of the protein, it behaves as a slightly hydrophilic amino acid.

For the hydrophilic amino acids as defined by $\langle\theta\rangle_C$ (except for Tyr, which will be discussed later), we find that the relation $\langle\theta\rangle_{out} < \langle\theta\rangle_{in} \sim 90°$ in group B generally holds (in some cases these values are equal within statistical error). The relation $\langle\theta\rangle_{in} \sim 90°$ stems from the homogeneity of the protein core, which we have discussed previously. The hydrophilic residues with C^α located outside this sphere of radius $0.7R_g$ shows $\langle\theta\rangle_{out} < 90°$, and this expresses the attraction exerted on their side chains by the surrounding water and the highly polar environment of the outer shell of the proteins. We also investigated the effect of increasing R to $1.2R_g$, but the values obtained for $\langle\theta\rangle_{out}$ in the various shells did not change significantly.

To conclude, the eight hydrophilic amino acids Asp, Glu, Lys, Pro, Gln, Arg, Ser, and Thr behave in a consistent way with respect to $\langle r\rangle$ and $\langle\theta\rangle$ in all the samples; only Asn has a slightly lower value of $\langle r_\alpha\rangle_A$ (in the neutral range) but is defined as hydrophilic according to the values of the two parameters in the other samples.

C. Some Special Cases. We shall discuss now the behavior of the four remaining amino acids—His, Ala, Gly, and Tyr. For His the values of $\langle\theta\rangle$ in all the samples are

very close to 90°. This fact defines His as a neutral amino acid, in accordance with our previous conclusions based on Table II. Ala falls in the neutral range of $\langle\theta\rangle_C$, but, as can be seen from Table III, it does not behave in a consistent way in the various samples. In group B Ala has the neutral value $\langle\theta\rangle_B = 91°$ whereas in group A the value of $\langle\theta\rangle_A$ lies in the hydrophilic regime. These results are in accordance with the results obtained for $\langle r_\alpha\rangle$ (Table II) for those two samples, which were found to be in the neutral and hydrophilic regions, respectively. This ambivalent nature of Ala is demonstrated even more strongly by the values of $\langle\theta\rangle_{in}$ and $\langle\theta\rangle_{out}$. In group B we obtained $\langle\theta\rangle_{in} = 91°$ as expected from the homogeneity of the protein core. When R was raised to $1.1R_g$, $\langle\theta\rangle_{in}$ increased to 97°, which is in the hydrophobic range, and $\langle\theta\rangle_{out}$ decreased considerably, to the hydrophilic value 74° (see lower line for Ala in Table III). Further increasing of R to $1.2R_g$ lowered $\langle\theta\rangle_{out}$ even more—to 67°. Similar behavior of Ala was also found for the small proteins of group A. We obtained (for $R = R_g$) $\langle\theta\rangle_{in} = 89 \pm 5°$ and $\langle\theta\rangle_{out} = 72 \pm 6°$ (not in Table III).

We suggest that this behavior of Ala reflects a competition between the small, nonpolar side chain and the two essentially unhindered polar peptide groups which flank it. When an Ala residue is located on the surface of the protein, exposed to water, the tendency of the two peptide groups to be hydrated is apparently dominant over the unfavorable interaction of the methyl group with water (because the small side chain does not have enough freedom to bury itself). Inspection of models indicates that, in the low-energy conformations of blocked Ala,[30] the methyl group [whose direction is determined completely by (ϕ, ψ)] lies roughly in the same direction as the NH and CO groups in the adjacent peptide units. The net effect is that the side chain points outward to facilitate peptide group hydration. Thus, $\langle\theta\rangle_{out} = 74°$ for $R > 1.1R_g$.

In the shell given by $0.7R_g < R < 1.1R_g$, the hydrophobic character of the methyl side chain is apparently dominant. This is the reason for the high value of $\langle\theta\rangle_{in} = 97°$ for the sphere $R < 1.1R_g$. In the inner sphere given by $R < 0.7R_g$, the isotropic environment is manifest, and the residue exhibits no strong orientational preference ($\langle\theta\rangle_{in} = 91°$).

We therefore regard Ala as an ambivalent amino acid, in the sense that opposing hydrophilic and hydrophobic tendencies seem to be finely balanced. The orientational behavior is therefore a function of the environment. This ambivalence is also reflected in the values of $\langle r\rangle_A$ and $\langle r\rangle_B$, as noted above.

These considerations suggest that Gly should behave like a less hydrophobic analogue of Ala, since it possesses the neighboring peptide groups but no hydrophobic side chain. Indeed, the results of Table II indicate that this is the case. In group A, where surface (solvent) effects are more marked, $\langle r_\alpha\rangle_A$ for Gly is very hydrophilic and larger than $\langle r_\alpha\rangle_A$ for Ala. In group B, where solvent effects are less marked, $\langle r_\alpha\rangle_B$ for Gly decreased to a neutral value as Gly enters into more intramolecular hydrogen bonds. Nevertheless, $\langle r_\alpha\rangle_B$ for Gly is greater than $\langle r_\alpha\rangle_B$ for Ala, reflecting the absence of the hydrophobic methyl side chain. Because of the adaptability of Gly to differing environments, we place it also in the ambivalent category.

Of special interest are the results obtained for Tyr. The Tyr side chain consists of three components, the CH_2 group and the phenyl ring, which are highly nonpolar, and a polar hydroxyl group connected to the ring. In what follows we shall discuss the opposite effects of these polar and nonpolar parts on the orientation of the side chain. First of all, it should be noted that, for the small proteins, the value of $\langle\theta\rangle_A$ for Tyr falls in the neutral range and for the larger proteins of group B, $\langle\theta\rangle_B$ falls in the hydrophilic range. We also calculated the values of $\langle\theta\rangle_{in}$ and $\langle\theta\rangle_{out}$ (with $R = R_g$) for group A and found $\langle\theta\rangle_{in} = 84 \pm 8°$, $\langle\theta\rangle_{out} = 102 \pm 11°$. These results put Tyr in the ambivalent category, hydrophobic when close to water and hydrophilic while in the interior of the protein.

These observations probably express once again the interaction between two different parts of the residue—in this case the bulky, nonpolar benzyl group and its polar hydroxyl substituent, which together form a relatively rigid unit. Tyrosine residues whose C^α are located in the interior are able to dispose their side chains so that the ring is buried and the OH exposed simultaneously. Tyr residues whose C^α is located farther from the center of mass are more constrained by the need to bury the aromatic ring. It is therefore reasonable that $\langle\theta\rangle_A > \langle\theta\rangle_{out} > \langle\theta\rangle_{in}$. It should be emphasized that Tyr is the only amino acid for which $\langle\theta_\beta\rangle_C$ is substantially higher than $\langle\theta\rangle_C$. The high values of $\langle\theta_\beta\rangle_C$ express the tendency for the backbone to force the CH_2 group to point inward, whereas the tendency of the OH group to point outward decreases $\langle\theta\rangle_C$.[31]

This picture correlates well with the values of $\langle r_\alpha\rangle$ of Tyr, which are hydrophobic in group A and slightly hydrophilic in group B. These variations are apparently adapted to locate C^α at a distance from the surface which enables the ring to be buried and the OH exposed. We place Tyr in the ambivalent group on the basis of its differing behavior in groups A and B.

It is essential to point out, as implied in the Introduction, that the polarity of an amino acid *residue*, as measured by its free energy of transfer from a nonpolar solvent to water, has contributions from both its backbone and side chain. Thus, the polarity (or hydrophobicity and/or hydrophilicity) of a residue is an intrinsic property of a residue. The quantities $\langle r\rangle$ and $\langle\theta\rangle$, however, or for that matter any other quantity resulting from an analysis of protein structural data, reflect *not only* the intrinsic polarity of the residue but also the geometrical constraints imposed on the spatial arrangements of the residues because they are connected in a chain molecule; i.e., a nonpolar residue tends to seek the nonpolar interior of the protein, subject to its ability to do so by such constraints. Since smaller proteins have greater surface-to-volume ratios than larger ones, there is a greater need for the nonpolar groups of the smaller proteins to seek the inside (and avoid the relatively large surface) than there is in larger proteins; this greater hydrophobic tendency for small compared to large proteins is borne out by the observations reported here.

The values for $\langle\theta\rangle$ given in Table III might be incorporated in protein-folding procedures. In such procedures the values of the standard deviation $\langle\Delta\theta^2\rangle^{1/2}$ (eq 8) might also be useful. We therefore point out that for most of the amino acids we found that $\langle\Delta\theta^2\rangle_C^{1/2}$ lies around 38°; only Lys and Gln have a lower standard deviation, ~28°.

D. Comparison with Other Hydrophobicity Scales. We shall now compare our empirical results with scales of hydrophobicity determined by other methods. They are the average surrounding hydrophobicity scale obtained by Manavalan and Ponnuswamy,[12] the experimental hydrophobicity index determined by Nozaki and Tanford[9] and by Jones,[10] scales determined by Wertz and Scheraga[13] and by Chothia,[14] both giving the fraction of buried residues for each amino acid over a sample of proteins, and the side-chain interaction parameters reported by Krigbaum and Komoriya.[16] The results are summarized in Table IV; in each column they are listed in order of decreasing hy-

Table IV
Comparison of Hydrophobicity Scales[a]

I hydrophobicity index of Nozaki, Tanford, and Jones[9,10]		II $\langle r_\alpha \rangle_C$		III av surrounding hydrophobicity of Manavalan and Ponnuswamy[12]		IV Wertz and Scheraga[13]		V Chothia[14]		VI side-chain interaction parameter ξ_i of Krigbaum and Komoriya[16] [b]	
	kcal/mol				kcal/mol		fraction buried		fraction 95% buried		
Trp	3.77	Phe	0.78	Val	15.71[c]	Phe	0.87	Ile	0.60	Cys	1.73
Ile	3.15	Ile	0.79	Ile	15.67	Trp	0.86	Val	0.54	Ile	2.31
Phe	2.87	Val	0.81	Leu	14.90	Cys	0.83	Phe	0.50	Met	2.44
Pro	2.77	Trp	0.83	Cys	14.63	Ile	0.79	Cys	0.50	Phe	2.59
Tyr	2.67	Met	0.84	Met	14.39	Leu	0.77	Leu	0.45	Trp	2.78
Leu	2.17	Leu	0.85	Phe	14.00	Met	0.76	Met	0.40	Val	3.31
Val	1.87	Cys	0.88	Trp	13.93	Val	0.72	1/2 Cys	0.40	Tyr	3.58
Met	1.67	----------		----------		His	0.70	Ala	0.38	----------	
Lys	1.64	His	0.89	Tyr	13.42	Tyr	0.64	Gly	0.36	Leu	3.93
Cys	1.52	Tyr	0.93	Ala	12.97	Ala	0.52	Trp	0.27		
----------		Ala	0.93	Gly	12.43	Ser	0.49	Thr	0.23	Ala	4.32
Ala	0.87	Gly	1.00	His	12.16	Arg	0.49	Ser	0.22	Thr	5.16
His	0.87	----------		----------		Asn	0.42	Glu	0.18	Ser	5.37
Arg	0.85	Asn	0.98	Glu	11.89	Gly	0.41	Pro	0.18	----------	
Glu	0.67	Arg	0.98	Gln	11.76	Thr	0.38	His	0.17	His	5.66
Asp	0.66	Thr	0.99	Arg	11.72	Glu	0.38	Asp	0.15	Asp	6.04
Gly	0.10	Pro	1.00	Thr	11.69	Asp	0.37	Tyr	0.15	Gly	6.09
Asn	0.09	Asp	1.01	Asn	11.42	Pro	0.35	Asn	0.12	Gln	6.13
Ser	0.07	Glu	1.02	Pro	11.37	Gln	0.35	Gln	0.07	Glu	6.17
Thr	0.07	Gln	1.02	Lys	11.36	Lys	0.31	Lys	0.03		
Gln	0.00	Ser	1.02	Ser	11.23			Arg	0.01	Asn	6.24
		Lys	1.05	Asp	10.85					Arg	6.55
										Pro	7.19
										Lys	7.92

[a] Within each column the amino acids are listed in order of decreasing hydrophobicity. See text for a fuller discussion of the position of Gly in column II. [b] See text for discussion and categorization of residues.

drophobicity. Our results (column II) are arranged in three groups—hydrophobic, neutral and ambivalent, and hydrophilic (which are separated in the table by dashed lines). In each of these groups we chose to order the results according to the value of $\langle r_\alpha \rangle_C$.[32] This is for comparison with the results of the other methods, which were obtained by averaging over samples which consist of both small and large proteins. In column III we also divided the amino acids (somewhat arbitrarily) by dashed lines into three groups (from top to bottom)—hydrophobic, neutral, and hydrophilic. For the Nozaki–Tanford–Jones scale in the first column only two groups, hydrophobic and hydrophilic, are defined by a dashed line. For comparison, dashed lines have also been added to column VI in the same places as in columns II and III; the solid lines in column VI indicate the separation into nonpolar, intermediate, and polar groups, as defined by Krigbaum and Komoriya.[16]

We now compare our results (column II) to those obtained by calculating the average surrounding hydrophobicity (column III). Table IV shows that the makeup of the three groups—hydrophobic, neutral, and hydrophilic—as determined by these two methods is identical, although the order within each group differs.[32] This is not surprising since the average surrounding hydrophobicity is obtained by averaging the Nozaki–Tanford–Jones index contributed by residues located within a radius of 8 Å around each residue. Hydrophobic amino acids defined by $\langle r_\alpha \rangle$ are concentrated around the center of mass of the protein and therefore one expects them to be surrounded by a large fraction of hydrophobic residues. The hydrophilic amino acids, according to the $\langle r_\alpha \rangle$ criterion, are located mainly toward the periphery of the protein and therefore are likely to have a more hydrophilic environment. It is interesting to point out that, even though Gly is located far from the center of mass, i.e., $\langle r_\alpha \rangle_C = 1.00$, its average surrounding hydrophobicity[12] falls in the neutral range. The results of Wertz and Scheraga,[13] obtained by calculating the fraction of buried residues for each amino acid, are also in accord with our results (the fact that Gly appears in the hydrophilic range is consistent with the value $\langle r_\alpha \rangle_C = 1$). There are discrepancies between Chothia's[14] results (column V) and those of Wertz and Scheraga[13] (column IV), in particular, in the relative positions of Trp, His, Tyr, Arg, and Gly. A possible reason for this might be the different samples of proteins used (20 by Wertz and Scheraga, 12 by Chothia), but the main reason lies in the different ways that the surface of the protein is defined by the two methods.[13] Some discrepancies also occur between our results (column II) and those of Krigbaum and Komoriya (column VI), mainly in the central group between dashed lines in column VI (Leu, Ala, Thr, and Ser). Only Ala appears in this central group in both columns. Krigbaum and Komoriya, however, also presented a slightly different ordering of the residues of column VI, based mainly on free energy of transfer data; this alternative order is in good agreement with the data in column II. Discrepancies exist also between our empirical results and the Nozaki–Tanford–Jones hydrophobic index (column I). The most substantial differences are that Pro, Tyr, and Lys appear to be hydrophobic in Tanford's scale whereas, according to our classification, they are hydrophilic or ambivalent. On the other hand, Gly appears in the hydrophilic region of column I but in the ambivalent region of column II.

It should again be noted that there is a fundamental difference between the Nozaki–Tanford and Jones scale and the other hydrophobicity scales. The latter are based on empirical inspection of protein structures.[33] They therefore reflect in a complex way the interaction between the hydrophobic interaction and other factors, such as local chain structure preference dictated by short-range interactions. The Nozaki–Tanford–Jones scale, which is based

on experiments performed on isolated amino acids, should not be expected to give identical results and indeed does not. The results for Lys, for example, probably reflect the long hydrophobic side chain to which N^{ζ} is attached and whose effect is more manifest in water than in a protein environment. The results for Pro do not reflect the tendency of Pro to occur in bends in proteins, which are located toward the molecular exterior. The high hydrophobicity of Tyr in the Nozaki–Tanford–Jones scale does not reflect the tendency of Tyr to be located so that the OH can be exposed.

The discrepancy observed for Gly (columns I and II) also reflects differences in behavior between the free and the connected amino acid. As a free amino acid, Gly has complete freedom to hydrogen bond with the surrounding water molecules. As a part of a chain, on the other hand, the freedom of the two polar peptide groups to form hydrogen bonds with water is highly reduced, and this is probably the reason the values of $\langle r \rangle$ for Gly fall in the neutral range for the larger proteins. In smaller proteins, however, Gly is forced outward by the hydrophobic amino acids, as has already been discussed; therefore, Gly is classified as ambivalent by our criteria.

Summary

We have examined the behavior of the average reduced distance from the center of mass, $\langle r \rangle$, and the average side-chain orientational angle, $\langle \theta \rangle$, for the naturally occurring amino acids in a sample of 19 proteins. The behavior of these parameters was compared in two subsamples composed of small and larger proteins, respectively. The following results were obtained.

(1) For most of the amino acids consistent qualitative results for hydrophobicity are obtained from comparison of the two groups. Hydrophobic behavior, as measured in terms of $\langle r \rangle$ and $\langle \theta \rangle$, is manifested more strongly in the small proteins. This is apparently due to greater geometrical constraints imposed on residues by the smaller volume and larger surface-to-volume ratio in the small proteins.

(2) Three amino acids—Tyr, Ala, and Gly—show distinctly different behavior in the two groups. These ambivalent results can be explained in terms of competition between hydrophobic and hydrophilic groups on the residue.

(3) The behavior of $\langle \theta \rangle$ was investigated in inner and outer layers of the larger proteins. It was observed that in these proteins a "dead volume" exists, consisting of a sphere of radius $\sim 0.7R_g$, centered on the center of mass. Within this sphere, the degree of hydrophobicity is essentially constant, and the effect of solvent is small. As a result, residues in this volume see an essentially isotropic environment, and no orientational preference is to be found.

(4) The orientational preferences of C^{α}–C^{β} bonds were also calculated and compared to those of the side chain as a whole. It was found that the orientational preferences are qualitatively similar in most cases, indicating correlation between side-chain and backbone orientations.

(5) The results of this approach are compared with several other hydrophobicity scales. Consistency is found with those of Manavalan and Ponnuswamy,[12] Wertz and Scheraga,[13] and Krigbaum and Komoriya.[16] Discrepancies between these scales and those of Nozaki and Tanford[9] and Jones,[10] and Chothia,[14] are discussed.

Acknowledgment. We thank Dr. George Némethy for helpful discussions.

References and Notes

(1) (a) This work was supported by research grants from the National Institute of General Medical Sciences of the National Institutes of Health, U.S. Public Health Service (GM-14312), and from the National Science Foundation (PCM79-20279). (b) To whom requests for reprints should be addressed.
(2) Kauzmann, W. *Adv. Protein Chem.* **1959**, *14*, 1.
(3) Némethy, G.; Scheraga, H. A. *J. Phys. Chem.* **1962**, *66*, 1773.
(4) The formation of a compact conformation of a protein, with nonpolar groups in its interior, can be thought of as arising from an initially denatured conformation in which all parts of the chain are highly exposed to water. Thus, the free energy for the conversion of the denatured conformation to the compact conformation contains contributions from *both* the van der Waals interactions and the (partial or complete) dehydration of the groups that become (partially or completely) buried.[3]
(5) Klotz, I. M. *Arch. Biochem. Biophys.* **1970**, *138*, 704.
(6) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379.
(7) Kuntz, I. D. *J. Am. Chem. Soc.* **1972**, *94*, 8568.
(8) Chothia, C. *Nature (London)* **1975**, *254*, 304.
(9) Nozaki, Y.; Tanford, C. *J. Biol. Chem.* **1971**, *246*, 2211.
(10) Jones, D. D. *J. Theor. Biol.* **1975**, *50*, 167.
(11) Rackovsky, S.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1977**, *74*, 5248.
(12) Manavalan, P.; Ponnuswamy, P. K. *Nature (London)* **1978**, *275*, 673.
(13) Wertz, D. H.; Scheraga, H. A. *Macromolecules* **1978**, *11*, 9.
(14) Chothia, C. *J. Mol. Biol.* **1976**, *105*, 1.
(15) Krigbaum, W. R.; Rubin, B. H. *Biochim. Biophys. Acta* **1971**, *229*, 368.
(16) Krigbaum, W. R.; Komoriya, A. *Biochim. Biophys. Acta* **1979**, *576*, 204.
(17) Kuntz, I. D.; Crippen, G. M.; Kollman, P. A.; Kimelman, D. *J. Mol. Biol.* **1976**, *106*, 983.
(18) Ycas, M.; Goel, N. S.; Jacobsen, J. W. *J. Theor. Biol.* **1978**, *72*, 443.
(19) Tanaka, S.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1975**, *72*, 3802.
(20) After completion of the manuscript, we became aware of the paper by Krigbaum and Komoriya.[16] We, therefore, include here a comparison with their parameters.
(21) Feller, W. "An Introduction to Probability Theory and Its Application"; Wiley: New York, 1968; Chapter 9.
(22) Lewis, P. N.; Momany, F. A.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1971**, *68*, 2293.
(23) Chou, P. Y.; Fasman, G. D. *J. Mol. Biol.* **1977**, *115*, 135.
(24) Isogai, Y.; Némethy, G.; Rackovsky, S.; Leach, S. J.; Scheraga, H. A. *Biopolymers* **1980**, *19*, 1183.
(25) Kuntz, I. D. *J. Am. Chem. Soc.* **1972**, *94*, 4009.
(26) Rose, G. D.; Wetlaufer, D. B. *Nature (London)* **1977**, *268*, 769.
(27) Krigbaum and Komoriya[16] studied the effect of protein size on pair frequencies and also pointed out that, in protein-folding procedures such as those of Tanaka and Scheraga,[19,28] different sets of these frequencies should be used for smaller and larger proteins.
(28) Tanaka, S.; Scheraga, H. A. *Macromolecules* **1976**, *9*, 945.
(29) Meirovitch, H.; Scheraga, H. A. *Macromolecules* **1980**, *13*, 1406.
(30) Zimmerman, S. S.; Pottle, M. S.; Némethy, G.; Scheraga, H. A. *Macromolecules* **1977**, *10*, 1.
(31) Krigbaum and Komoriya[16] also noted that the behavior of Tyr is attributable to its polar and nonpolar parts.
(32) It should be emphasized that the behavior of the amino acid residues according to the various criteria ($\langle r_\alpha \rangle_A$, $\langle r_\alpha \rangle_B$, $\langle \theta \rangle_A$, etc.) enabled us to classify them into four groups (hydrophobic, hydrophilic, neutral, and ambivalent, with the last two combined into one category in Table IV). We could not, however, find enough correlations among these parameters to scale them more precisely *within* each of these groups.
(33) The results of Manavalan and Ponnuswamy[12] and Krigbaum and Komoriya[16] are based *both* on the data of Nozaki and Tanford[9] and on X-ray coordinates.