Empirical Studies of Hydrophobicity. 3. Radial Distribution of Clusters of Hydrophobic and Hydrophilic Amino Acids[1]

H. Meirovitch and H. A. Scheraga*

Baker Laboratory of Chemistry, Cornell University, Ithaca, New York 14853.
Received August 7, 1980

ABSTRACT: Our previous classification of amino acids into three groups (hydrophobic, hydrophilic, and neutral–ambivalent) and the known structures of 19 proteins are used to examine in detail the existence of *local* hydrophobic and hydrophilic clusters in proteins; these structural features have been discussed recently by Krigbaum and Komoriya and by Kuntz and Crippen. For this purpose, four spherical layers are defined around the centers of mass of these proteins, and spheres of radius 8 Å around each of the amino acid residues define their close surroundings. For each amino acid in a given layer, we calculate the average fractions of hydrophobic, hydrophilic, and neutral–ambivalent residues (as defined above) located within its 8-Å sphere. In these calculations, an amino acid residue is represented by its C^α atom or, alternatively, by a remote side-chain atom, and accordingly two sets of results are presented. We scale the amino acids in each layer in decreasing order of the fractions of hydrophobic residues in their close surroundings and, alternatively, in increasing order of the fractions of hydrophilic residues. For the remote side-chain atoms in the three innermost layers, good agreement is obtained between our earlier classification and the present scale; i.e., for each of these layers, the fraction of hydrophobic (hydrophilic) residues in the immediate environment of hydrophobic amino acids is substantially higher (lower) than that in the neighborhood of hydrophilic amino acids. The neutral–ambivalent amino acids generally appear in the center of the scale. Disagreement between the two scales is detected only for the neutral amino acid His, which appears in the hydrophilic region, and for the hydrophilic amino acid Thr, which appears mostly in the center of the scale. These results strengthen the validity of our classification of amino acids which is based on criteria different (though not independent) from that based on the nature of the local surroundings and demonstrate that *local* hydrophobic and hydrophilic clusters exist in proteins. For the fourth (outermost) layer, the positions of several amino acids in the scale do not agree with our previous classification; this behavior is accounted for by an insufficient data base and by solvent effects. The degree of hydrophobicity (hydrophilicity) of the clusters is found to be much more pronounced for the innermost layer than for the outer ones. For the C^α atoms, such hydrophobic (hydrophilic) clusters are also detected but their degree of hydrophobicity (hydrophilicity) is much less marked than that detected for the side-chain atoms. This is accounted for by the larger flexibility of the side chains as compared to the backbone.

Introduction

In the first paper of this series[2] the 20 naturally occurring amino acids were classified into four groups: hydrophobic—Cys, Phe, Ile, Leu, Met, Val, and Trp; hydrophilic—Asp, Glu, Lys, Asn, Gln, Arg, Ser, Pro, and Thr; ambivalent—Ala, Gly, and Tyr; and one neutral amino acid—His. In the second paper,[3] we examined the radial distribution of these groups of amino acids around the center of mass of 19 proteins. One of the main conclusions of that paper was that the fraction of hydrophobic residues is substantially higher in the inner sphere of radius $0.75R_g$ (where R_g is the root-mean-square radius of gyration) than in the whole protein and that this fraction decreases gradually in the outer layers. An opposite behavior was detected for the hydrophilic residues. On the other hand, the fractions for the group of neutral and ambivalent amino acids in the various layers did not change in any particular direction but instead fluctuated randomly around the fraction of these amino acids in these proteins. These conclusions, which are in accord with results obtained by Krigbaum and Komoriya,[4] mean that *on the average over the whole protein* the close environment of a hydrophobic (hydrophilic) residue is substantially more hydrophobic (hydrophilic) than would be anticipated from random mixing. These expectations had been confirmed by several statistical studies of the preferred surroundings of amino acids, based on the known structures of proteins.[4-9]

A closely related question is whether *local* clusters of hydrophobic and hydrophilic residues also exist in proteins. This question has been dealt with only in a few of the references cited above, while the others have studied the *average* environment of amino acids over the *whole* protein rather than the *local* environment. One would expect local clusters to exist mainly because of hydrogen-bonding and hydrophobic interactions. Many polar residues located in the interior of the native protein form hydrogen bonds with each other in order to compensate for their loss of hydrogen bonds with water in the unfolded state.[10] Creation of hydrophobic (and hence hydrophilic) domains is also enhanced by the aggregation of hydrophobic residues (to avoid contact with water)[16] which strongly dictates the process of folding as long as a compact structure has not been reached. It is reasonable to assume that these hydrophobic domains are not destroyed in the final stages of folding since substantial rearrangement of a compact structure is highly restricted by chain connectivity. Obviously other forces, such as electrostatic interactions, also take part in cluster formation but it is difficult to assess their contribution.[18] Local cluster formation in proteins has been studied by Kuntz and Crippen,[19] who found segregation of polar and nonpolar *atoms* (rather than *residues*). It should be noted that analyses with respect to atoms, on the one hand, and residues, on the other, can lead to substantially different results, as has been observed by Lee and Richards[20] and Chothia[12] in assessing the hydrophobicity of the surfaces of proteins. Krigbaum and Komoriya,[4] on the other hand, detected local clusters of polar and nonpolar *residues* by calculating the average value of their side-chain interaction parameter ξ, vis., $\langle \xi_N \rangle$, but for only two small groups of residues, those which are neighbors of Val and those which are neighbors of the (collective) four most polar residues, Asp, Pro, Lys, and Arg (these four being treated as a single polar-type residue). In contrast to other workers,[6-9] who examined the average distribution over the whole protein, Krigbaum and Komoriya focused attention on the local character of the clusters by calculating $\langle \xi_N \rangle$ for relatively narrow spherical

0024-9297/81/2214-0340$01.25/0

layers around the center of mass of each protein. They showed that, in every layer of their sample of larger proteins, the value of $\langle \xi_N \rangle$ for Val is smaller (i.e., less polar) than that for the four polar amino acids, i.e., that Val is preferentially surrounded locally by nonpolar residues and that the four polar ones locally by polar residues.

In the present paper, we study local hydrophobic and hydrophilic clusters, using a method similar to that of Krigbaum and Komoriya.[4] Our examination is extended to all 20 naturally occurring amino acids and to include C^α atoms as well as side chains, and the results are compared to the classification of our earlier paper.[2] We use the X-ray structure data of 19 proteins in our computations. The close surroundings around each residue are defined by a sphere of radius 8 Å (the choice of this radius as being the most adequate is discussed by Manavalan and Ponnuswamy[6]), and four concentric layers around the center of mass of each protein are searched. For each amino acid and for each one of the four layers, we calculate the average fractions of hydrophobic, hydrophilic, and the group of neutral–ambivalent residues located within the surrounding spheres. As in ref 2, 3, 21, and 22, we represent a residue either by its C^α atom or by a remote side-chain atom and, correspondingly, provide two sets of results. Since the side chains are much more flexible than the backbone, one would expect the formation of clusters to be more marked for the side-chain atoms than for the C^α's. For each layer, the results are scaled in decreasing order of the fractions of hydrophobic residues and, alternatively, in increasing order of the fractions of hydrophilic residues. These scales are compared with the hydrophobicity scale of ref 2 (based on the average distance of a residue from the center of mass of the proteins and on the average angle created by the two vectors, center of mass-to-C^α and C^α-to-a remote side-chain atom), and the correlations between them are discussed. In view of the above discussion, one would expect to find the close surroundings of hydrophobic residues to be more hydrophobic than those of the hydrophilic residues.

Finally, it should be pointed out that the results of this paper not only provide deeper insight into the problem of protein organization but also may be incorporated into protein folding algorithms.

Methods

In this work, as in ref 2 and 3, a sample of 19 proteins is used (see Table I of ref 2 for a listing). For each protein, we assign four spherical layers around the center of mass, defined in units of R_g—the root-mean-square radius of gyration. This definition is chosen in order to scale the results obtained from proteins with different molecular weights. The amino acids are divided into the three groups[2] cited in the Introduction (with the neutral and ambivalent amino acids combined into one group). As also mentioned in the Introduction, a residue is represented by its C^α atom or, alternatively, by a remote side-chain atom (for details see Table II of ref 2). Accordingly, we carry out two sets of calculations. In the first one, a sphere of radius 8 Å (see ref 6) is drawn around the C^α atom of each residue and the numbers of C^α atoms of hydrophobic, hydrophilic, and neutral–ambivalent residues located within this sphere are determined (in this calculation the two nearest neighbors on either side of a residue are not counted because their C^α-to-C^α distance is always smaller than 8 Å). For each layer and amino acid, we accumulate these numbers over the entire sample of proteins and compute the average fractions of the three groups of amino acids. The same procedure is also applied to the remote side-chain atoms. It should be pointed out that, in view of some differences in the behavior of smaller and larger proteins detected in ref 2 and 3, we first carried out our analysis separately for these two samples. The results, however, especially for the smaller proteins, were found to be statistically unreliable because of an insufficient data base and, therefore, we used the entire (combined) sample of 19 proteins.

Results and Discussion

Results for the fractions (in percents) of hydrophobic, hydrophilic, and the neutral–ambivalent residues located around each of the 20 amino acids (for each of the layers) are presented in Tables I and III for the side-chain and C^α atoms, respectively. For each layer, the amino acids appear in these tables in decreasing order of the fractions of hydrophobic residues around them; when two amino acids have the same fractions of hydrophobic residues, they are ranked in increasing order of the fractions of hydrophilic residues. For comparison, we also rank them in Tables II and IV, respectively, in decreasing order of the fractions of hydrophilic residues. We have found[3] that the fractions of hydrophobic and hydrophilic residues decrease and increase, respectively, in going from the interior toward the exterior of the protein. Therefore, a relatively wide layer should also be radially inhomogeneous: its outer region would be more hydrophilic and less hydrophobic than the inner one. For such a layer, our analysis would therefore show higher (lower) fractions of hydrophobic residues around hydrophobic (hydrophilic) amino acids than around hydrophilic (hydrophobic) ones, and this might be interpreted incorrectly as evidence for the existence of local clusters. In order to eliminate this effect, we therefore selected our layers to be thin. The innermost layer, however, was chosen to be relatively wide ($0–0.75R_g$) in view of the radial homogeneity that we found[2] in this sphere for the larger proteins which provide most of the residues of our sample.

Results for the Side-Chain Atoms. Results for the side-chain atoms are presented in Table I. In the fifth column of each layer, we provide the total number of remote side-chain atoms that occur in the vicinity (i.e., in a sphere of radius 8 Å) of a given type of amino acid for the given layer in the entire sample of proteins; obviously, the larger these numbers, the better is the statistical significance of the results. In the last column of the table, the amino acids are listed in the same order as they appear in column 2 of Table IV of ref 2, i.e., according to our earlier classification;[2] two dashed lines are used to divide the residues in the last column into the three groups (from top to bottom), hydrophobic, neutral–ambivalent, and hydrophilic. In order to compare our results in ref 2 with those of the present paper, these dashed lines are extended to the left, dividing each column into three sections. The line above the bottom line gives the fractions (in percents) of the three groups of amino acids in the various layers together with the total number of remote side-chain atoms located in the layer. We also define a parameter Δ which is the difference between the largest and smallest values of the fractions in a given column. The values of Δ appear in the bottom line of the table.

If hydrophobic and hydrophilic clusters do exist in proteins, one would expect in general to find the hydrophobic amino acids in the upper section of the table (since their close environment would be hydrophobic) and the neutral–ambivalent and hydrophilic amino acids in the intermediate and lower sections, respectively, of the table. Deviations from this picture, however, are also expected, especially for some of the ambivalent amino acids. This is because, by including them in the same group with the

Table I
Results for the Side-Chain Atoms[a]

spherical layer in units of R_g																			
I 0-0.75					II 0.75-0.95					III 0.95-1.15					IV 1.15-				
amino acid	A	B	C	D	amino acid	A	B	C	D	amino acid	A	B	C	D	amino acid	A	B	C	D	classification of amino acids (ref 2[b])
Phe	62	19	19	543	Cys	50	26	24	224	Cys	45	30	25	86	Cys	33	37	30	60	Phe
Ile	58	21	21	950	Phe	50	30	20	261	Trp	41	39	20	100	Trp	33	42	25	33	Ile
Leu	57	22	21	960	Ile	49	29	22	383	Met	40	36	24	77	Val	32	48	20	132	Val
Cys	57	25	18	232	Val	45	31	24	512	Leu	37	34	29	273	Leu	28	47	25	109	Trp
Met	52	22	26	221	Leu	45	32	23	494	Ile	36	39	25	155	Met	27	59	14	22	Met
Val	52	24	24	1044	Ala	41	34	25	535	Phe	35	43	22	49	Arg	26	44	30	147	Leu
Ala	51	26	23	765	Met	39	31	30	137	Val	34	43	23	287	Gly	26	50	24	312	Cys
Pro	50	25	25	198	Trp	39	37	24	145	Gln	33	42	25	182	Asn	25	46	29	206	His
Trp	48	29	23	172	Thr	37	41	22	319	Ala	32	41	27	387	Gln	24	49	27	192	Tyr
Thr	44	31	25	401	Tyr	34	43	23	224	Gly	32	42	26	434	Thr	24	52	24	208	Ala
Asn	42	29	29	153	Ser	33	41	26	404	Tyr	31	42	27	248	Phe	23	47	30	53	Gly
Gly	40	32	28	604	His	33	44	23	134	Glu	29	46	25	86	Ser	23	49	28	360	Asn
Tyr	38	36	26	207	Gly	32	40	28	418	Lys	28	50	22	281	Ala	23	56	21	233	Arg
Ser	36	37	27	486	Pro	31	45	24	157	Asn	27	46	27	277	Ile	22	48	30	96	Thr
His	33	41	26	191	Asn	27	43	30	187	Pro	27	49	24	170	Asp	22	55	23	238	Pro
Gln	33	42	25	93	Arg	27	51	22	119	Arg	27	51	22	133	Glu	20	51	29	203	Asp
Glu	32	44	24	124	Gln	27	51	22	88	Thr	26	48	26	250	Lys	20	58	22	416	Glu
Asp	26	44	30	214	Asp	25	43	32	198	Ser	26	49	25	267	Tyr	19	55	26	194	Gln
Lys	24	59	17	54	Lys	25	47	28	81	His	23	51	26	65	Pro	18	53	29	118	Ser
Arg	22	56	22	92	Glu	24	54	22	177	Asp	21	52	27	229	His	17	58	25	81	Lys
fraction of all residues in layers, %	52	25	23	779[c]		37	38	25	702		20	53	27	741		10	67	23	1011	
Δ[d]	40	40	12	7704[e]		26	28	10	5197		24	29	9	4036		16	17	16	3413	

[a] For each of layers I-IV, the amino acids are listed in decreasing order of the fractions of hydrophobic residues (based on column A in each layer I). The dotted lines divide the table into three sections. For details see text. In each layer, the column headings have the following meaning: A, hydrophobic (%); B, hydrophilic (%); C, neutral and ambivalent (%); D, number of residues in spheres of radius 8 Å around *all* residues of a given type. [b] The amino acids are listed in the same order as in column II of Table IV (ref 2). [c] This is the total number of residues in the given layer. [d] Δ is the difference between the largest and smallest values of the fractions in a given column. [e] This is the sum of the numbers in column D.

Table II
Reordering of the Positions of the Amino Acids Based on the Results of Table I[a]

I	II	III	IV
Phe	Cys	Cys	Cys
Ile	Ile	Leu	Trp
Leu	Phe	Met	Arg
Met	Val	Trp	*Asn*
Val	Met	Ile	Leu
Cys	Leu	*Ala*	*Phe*
Pro[b]	Ala	*Gln*	Val
Ala	Trp	Gly	*Ile*
Trp	*Gly*	Tyr	Gln
Asn	Thr	*Phe*	*Ser*
Thr	Ser	*Val*	*Gly*
Gly	*Tyr*	Glu	Glu
Tyr	Asn	Asn	*Thr*
Ser	Asp	Thr	Pro
His	His	Pro	Asp
Gln	Pro	Ser	Tyr
Glu	Lys	Lys	Ala
Asp	Arg	Arg	Lys
Arg	Gln	His	His
Lys	Glu	Asp	*Met*

[a] Amino acids are listed in increasing order of the fractions of hydrophilic residues (based on column B in each layer of Table I). When two amino acids have the same fractions of hydrophilic residues, they are listed in decreasing order of the fractions of hydrophobic residues. The dotted lines divide the results into three sections as in Table I. For details see text. [b] Amino acids in italics do not belong to the same sections in Tables I and II.

neutral residues, their different behavior in the interior and exterior of proteins is not taken into account (see ref 2). For example, in ref 2, Ala was shown to behave as a hydrophobic and a hydrophilic amino acid in the interior and outer parts of proteins, respectively. Therefore, one would expect Ala to appear in the upper section for the inner layers and in the intermediate or even the lower sections for the outer layers.

Let us first examine the results for the three inner layers. The table reveals that six of the seven amino acids classified as hydrophobic in ref 2 (i.e., on the basis of the average distance $\langle r \rangle$ from the center of mass and the average side-chain orientation $\langle \theta \rangle$) indeed appear in the upper section of the table. These amino acids are Phe, Ile, Leu, Cys, Met, and Val. In ref 2, Trp was also classified as hydrophobic. In Table I, however, it falls in the upper section only for the third layer whereas, for the two inner layers, it appears in the intermediate section, close to the upper one. The present results for Trp agree with our earlier results[2] for $\langle \theta \rangle$ which showed hydrophobic behavior for Trp in the smaller proteins (and in the outer layer of the larger ones) but slightly hydrophilic behavior for the inner layer of the larger ones. This hydrophilic tendency of Trp is probably due to the polar NH group of its indole ring which, while inside the protein molecule, tends to hydrogen bond with other polar groups. Four of the nine amino acids classified as hydrophilic in ref 2 always appear in the lower section of the table, i.e., surrounded by substantially lower and higher fractions of hydrophobic and hydrophilic residues, respectively. They are Glu, Asp, Lys,

Table III
Results for C^{α} [a]

spherical layer in units of R_g																			
I 0-0.75					II 0.75-0.95					III 0.95-1.15					IV 1.15-					
amino acid	A	B	C	D	amino acid	A	B	C	D	amino acid	A	B	C	D	amino acid	A	B	C	D	classification of amino acids (ref 2)
Ile	47	32	21	627	Met	45	31	24	97	Trp	42	33	25	55	Cys	46	35	19	57	Phe
Leu	45	32	23	662	Cys	43	35	22	115	Met	42	40	18	92	His	38	40	22	45	Ile
Phe	45	33	22	353	Phe	41	43	16	174	Ile	39	38	23	211	Arg	32	39	29	90	Val
Cys	45	37	18	232	Ala	39	38	23	493	Val	39	39	22	255	Ile	32	40	28	68	Trp
Ala	43	32	25	659	Val	38	38	24	599	Phe	38	44	18	119	Trp	32	43	24	37	Met
Arg	43	35	22	103	Trp	38	41	21	103	Cys	36	38	26	112	Gln	31	42	27	130	Leu
Val	42	31	27	805	Leu	35	39	26	523	Leu	35	41	24	293	Val	28	51	21	152	Cys
Thr	41	32	27	363	Lys	35	42	23	330	Glu	31	47	22	270	Leu	28	39	33	109	His
Pro	41	33	26	175	Asn	33	43	24	299	Ala	30	47	23	441	Met	27	46	27	11	Tyr
His	40	37	23	176	Arg	32	42	26	215	Asn	29	47	24	224	Asn	27	50	23	153	Ala
Ser	39	33	28	439	Pro	31	40	29	160	Lys	28	50	22	454	Gly	27	51	22	297	Gly
Tyr	39	37	24	276	Thr	31	45	24	334	Arg	27	46	27	136	Phe	25	46	29	28	Asn
Lys	39	39	22	98	Ile	30	39	31	417	Thr	27	48	25	250	Lys	25	55	20	236	Arg
Trp	37	36	27	176	Gln	30	42	28	194	Gly	26	47	27	443	Pro	24	49	27	89	Thr
Gly	36	37	27	558	His	30	49	21	80	Gln	25	46	30	202	Asp	22	49	29	160	Pro
Asn	35	33	32	184	Tyr	29	43	28	233	Tyr	25	47	28	302	Tyr	22	50	28	103	Asp
Glu	35	38	27	181	Gly	29	45	26	456	Ser	24	47	29	294	Ser	22	50	28	355	Glu
Gln	34	37	29	91	Ser	29	45	26	427	Asp	24	50	26	335	Glu	22	60	18	130	Gln
Met	34	40	26	151	Glu	27	49	24	191	Pro	22	56	22	210	Ala	20	54	26	215	Ser
Asp	33	40	27	218	Asp	24	47	29	188	His	21	53	26	121	Thr	18	60	22	213	Lys
fraction of all residues in layers, %	46	29	25	792		34	44	21	702		20	54	26	889		14	61	25	753	
Δ	14	8	14	6527		21	16	13	5628		20	23	12	4819		28	25	11	2678	

[a] See footnotes of Table I.

Table IV
Reordering of the Positions of the Amino Acids Based on the Results of Table III[a]

I	II	III	IV
Val	Met	Trp	Cys
Ile	Cys	Ile	Arg
Leu	Ala	Cys	*Leu*
Ala	Val	Val	His
Thr	Leu	Met	Ile
Phe	*Ile*	Leu	Gln
Pro	*Pro*	Phe	Trp
Ser	*Trp*	*Arg*	Met
Asn	Lys	*Gln*	*Phe*
Arg	Arg	Glu	*Pro*
Trp	*Gln*	Ala	*Asp*
Cys	*Phe*	*Asn*	*Asn*
His	*Asn*	Gly	Tyr
Tyr	Tyr	Tyr	Ser
Gly	*Thr*	Ser	*Val*
Gln	Gly	Thr	*Gly*
Glu	Ser	*Lys*	Ala
Lys	Asp	Asp	Lys
Met	His	His	Glu
Asp	Glu	Pro	Thr

[a] See footnotes of Table II.

and Arg. Ser, Asn, Pro, and Gln can also be relegated to the hydrophilic category since they all fall in the lower section for two of the layers and appear in the intermediate section only once (Ser and Asn are positioned at the bottom of the intermediate section). Some disagreement with the results of ref 2 is detected, however, for Thr which is hydrophilic according to ref 2 but located in the intermediate section of the two innermost layers, and falls in the lower section only in the third layer.

Let us now examine the results for the ambivalent and neutral amino acids. Ala has been classified as ambivalent[2] since its average distance from the center of mass, $\langle r \rangle$, was found to be substantially larger (and fell in the hydrophilic range) for the smaller proteins than for the larger ones (for which Ala appeared in the neutral range). Also, in the smaller proteins, and in the outer layers of the larger ones, it behaved as a hydrophilic residue with respect to $\langle \theta \rangle$ (i.e., on the average its side chain orients more toward the outside); on the other hand, in the inner layer of the larger proteins, it behaved as a slightly hydrophobic residue. This phenomenon has been accounted for[2] by competition between the opposing behavior of the backbone polar peptide groups and the *small* nonpolar methyl group. When Ala is on the surface of the protein, it tends to hydrogen bond with water; when it is inside the protein, the effect of the hydrophobic side chain becomes dominant. The positions of Ala in Table I are in accordance with these results. In the two innermost layers, Ala appears in the upper section whereas, in the third layer, it is located in the intermediate one. The statistical significance of these results, however, is not certain since, in Table II, where amino acids are ranked in increasing order of the fractions of hydrophilic amino acids, Ala appears in the intermediate, upper, and upper sections for layers I, II and III, respectively. Tyr has also been classified as ambivalent in ref 2 but, in contrast to Ala, it belongs to the intermediate section for layers II and III but to the lower section for layer I in both Tables I and II [this last result agrees with the low value of $\langle \theta \rangle$ (i.e., hydrophilic) detected for Tyr in the inner layer of the larger proteins[2]]. In ref 2, Gly was defined as ambivalent since its average distance from the center of mass, $\langle r \rangle$, was found to fall well in the hydrophilic range for the smaller proteins but decreased to the neutral range for the larger ones. Gly is less hydrophobic than Ala since it does

not have the hydrophobic methyl group. On the other hand, its two polar backbone peptide groups have less freedom to form intraprotein hydrogen bonds than the more flexible polar side chains. Gly should, therefore, appear in lower positions than Ala and would be expected to lie in the intermediate or in the higher positions of the lower section. This expectation is indeed confirmed by both Tables I and II.

His is the only amino acid for which a complete discrepancy between the classification of ref 2 and the present results is detected. According to all the criteria of ref 2, His behaved as a neutral amino acid whereas here it always appears in the lower section, i.e., prefers a hydrophilic environment. We do not known how to account for this discrepancy.

The results discussed above show that the side chains form hydrophobic and hydrophilic clusters. The substantially higher values of Δ obtained for the innermost layer than for the outer ones mean that cluster formation is much more pronounced in the interior than in the exterior part of the proteins. This conclusion, however, should be examined cautiously since the large values of Δ for the innermost layer may result from the radial inhomogeneity detected in ref 3 for the *smaller* proteins in this layer; i.e., the fractions of hydrophobic and hydrophilic residues gradually decrease and increase, respectively, in going from the inside to the outside (in spheres with radius $<0.75R_g$) for small proteins (but remain the same for large proteins). We have therefore applied our analysis to the group of larger proteins (for details see Table I of ref 2) and obtained somewhat smaller values of Δ for the first layer; these values, however, were still much higher than those obtained for the other layers. This means that hydrophobic and hydrophilic clusters exist in the inner layer of the larger proteins which were found[3] to be radially homogeneous.

In this context, we comment that the fractions for the neutral–ambivalent residues, unlike those of the hydrophobic and hydrophilic residues, do not change in any definite direction from amino acid to amino acid. Rather, they fluctuate slightly in a random manner around their average fraction in the layer. This expected behavior is also reflected by the relatively small values of Δ for these amino acids as compared to the values of Δ obtained for the hydrophobic and hydrophilic ones.

The results for the outermost layer (layer IV) have been excluded from the above discussion for two reasons: (1) A high fraction of its side chains interacts with the surrounding water, and this might affect cluster formation. (2) The data base for this layer is the smallest as compared with those of the other three layers (3413 neighbors vs. 7704 for the innermost layer); the values of Δ for the hydrophobic and hydrophilic amino acids are also the lowest and that of the neutral–ambivalent amino acids is the highest among the corresponding values for the other layers. This makes the results for the fourth layer statistically the most unreliable. Indeed, the positions of several amino acids in this layer differ substantially from their positions in the other layers. For example, Gly and Arg appear here in the upper section whereas the two hydrophobic amino acids Phe and Ile fall in the intermediate and lower ones, respectively. Also, Tyr is located close to the bottom of the lower section, whereas it belongs to the intermediate section in layers II and III and appears near the top of the lower section in layer I.

As has already been pointed out, in Table I we have chosen to arrange the amino acids in decreasing order of the fractions of hydrophobic residues and, when two amino acids had the same fractions of hydrophobic residues, they were ranked in increasing order of the fractions of hydrophilic residues. Obviously, this criterion for ordering has no preference over the alternative one, i.e., that which ranks the amino acids in increasing order of the fractions of hydrophilic residues (and those with the same fractions of hydrophilic residues in decreasing order of the fractions of hydrophobic residues). If the order of amino acids determined by this last criterion differed substantially from that of the former one, it would mean that the results of Table I are not reliable. In Table II, we therefore list the amino acids according to the alternative criterion. The table reveals that the amino acid composition of the three sections in the first and second layers is changed very little from that of Table I. For the first layer, Ala exchanges places with Pro and, for the second one, Gly moves to the intermediate section (it is located in the lower section in Table I), and Tyr appears in the lower section, whereas in Table I it belongs to the intermediate one. In the third layer, four amino acids, Ala, Gln, Phe, and Val, change sections and, in the fourth layer, a discrepancy with Table I is detected for the seven amino acids Asn, Phe, Ile, Ser, Gly, Thr, and Met. This increase in the discrepancy between the results of the two tables, going from the innermost layer toward the outermost one, results from the corresponding decrease in the values of Δ and the increase in the statistical fluctuations, as has been discussed previously.

To summarize: The results for the side-chain atoms show that, in all regions of the protein, amino acids defined as hydrophobic in ref 2 are surrounded by substantially higher fractions of hydrophobic residues and lower fractions of hydrophilic residues than amino acids defined as hydrophilic in ref 2. This means that *local* hydrophobic and hydrophilic clusters occur in proteins. The correspondence, however, between the amino acid composition of the three sections of Tables I and II (for the three inner layers) and the composition of the hydrophobic, neutral–ambivalent, and hydrophilic groups (ref 2) is not perfect, especially as far as His and Thr are concerned. For the outermost layers, we found disagreements with our classification of ref 2, which are attributed to an insufficient data base and to interactions of these residues with the surrounding water.

Results for the C^α Atoms. In the second set of results, a residue is represented by its C^α atom rather than by a remote side-chain atom. These results are summarized in Tables III and IV, which are identical in structure with Tables I and II, respectively. Since the spatial position of a remote side-chain atom is generally less restricted by chain connectivity than that of a backbone atom, and also since C^α atoms are less accessible to each other, one would expect the formation of clusters to be much less pronounced for the C^α's than for the remote side-chain atoms. This is indeed borne out by the substantially lower values of Δ (for the two innermost layers) in Table III than in Table I. Because of these smaller values of Δ, the difference between the fractions of hydrophobic and hydrophilic residues of adjacent amino acids on the average should also be smaller than those of Table I. We therefore expect the order of amino acids given in Table III to be statistically less significant than that in Table I. The results of Tables III and IV indeed reveal that, for some hydrophobic (hydrophilic) amino acids, a diffusion into the lower (upper) section occurs. For the first layer of Table III, Arg appears in the upper section whereas Trp and Met fall in the lower one. In Table IV, on the other hand, Thr and Pro are located in the upper section (of layer I) and

Cys and Met in the lower one. Seven amino acids (in italics in Table IV), rather than two in Tables I and II, appear in different sections (in layer I) in the two tables. This is apparently because of the stronger effect of statistical fluctuation for the C^{α} atoms, as discussed above. In the second layer, Lys and Arg appear in the intermediate section of Table III while Ile appears in the lower one. In Table IV, Pro appears in the upper section and Phe in the lower one. Here also, seven amino acids do not belong to the same sections in the two tables. In layer III, all the amino acids defined as hydrophobic in ref 2 appear in the upper section of Tables III and IV; on the other hand, Lys and Glu in Table III and Arg and Glu in Table IV appear in the intermediate section. In layer III, only four amino acids do not occur in the same sections in the two tables. The data base for the fourth layer is the poorest among the four layers, as is demonstrated by its smallest number of neighbor residues (2678 vs. 6527 for the first layer). We therefore do not discuss this layer. It is of interest to examine the behavior of the neutral–ambivalent group. For the three innermost layers of both Tables III and IV, Ala behaves as expected, i.e., appears in the upper section of the two first layers and in the intermediate one of the third layer (see discussion of Table I). The other two ambivalent amino acids, Tyr and Gly, always belong to the lower section of layers I, II, and III of Tables III and IV (in contrast, in Table I Gly appeared once and Tyr twice in the intermediate section). As in Tables I and II, the neutral amino acid His always appears in the lower section.

To summarize: in spite of the deviations in amino acid positions from the expected ones, most of the hydrophobic amino acids appear in the upper section and hydrophilic amino acids in the lower one, which means that hydrophobic and hydrophilic clusters of C^{α} atoms also occur. As discussed above, these clusters are much less pronounced than those detected for the remote side-chain atoms.

As pointed out in the Introduction, some of the information in Tables I–IV can be incorporated into approximate protein folding algorithms. For example, empirical attractive interactions (or distance restrictions) can be imposed on pairs of residues of the same class (e.g., hydrophobic or hydrophilic) to ensure formations of local clusters [see, e.g., the distance geometry approach to protein conformation[23,24]]. Such interactions are needed in addition to geometrical restrictions for hydrophobic (hydrophilic) residues to be located preferentially in the interior (exterior) since we have seen in the present paper that the formation of local clusters depends only weakly on the radial distance from the center of mass. The parameters for these interactions should be optimized to lead to the observed local clusters.

Conclusions

The results for the remote side-chain atoms presented in Tables I and II show that, on a local basis, a strong correlation exists between the type of the amino acid (i.e., hydrophobic, etc.) and the average fraction of hydrophobic (and hydrophilic) residues located in its vicinity. In most cases, residues defined as hydrophobic in ref 2 are found to have substantially larger fractions of hydrophobic residues in their neighborhood than do the hydrophilic residues (in the same layer). Similarly, the hydrophobic residues have smaller fractions of hydrophilic residues in their neighborhood than do the hydrophilic residues. Discrepancies occur for His and Thr. For the C^{α} atoms, this correlation is less perfect, and some hydrophobic and hydrophilic amino acids appear in the lower and upper layers, respectively, rather than vice versa. This is accounted for by the lower flexibility of the backbone as compared to that of a side chain. Hydrophobic and hydrophilic local clusters are still detected for the C^{α} atoms but they are less pronounced than those for the side chains. Several points are of interest: (1) The order of amino acids determined in Tables I and II does not constitute an alternative scale for hydrophobicity (see Table IV of ref 2) since it depends on the classification of ref 2. The fact, however, that this order is highly correlated with the classification of ref 2 means that the hydrophobicity of amino acids can be expressed by several different parameters (which are not independent) such as the average distance from the center of mass of proteins, the average side-chain orientation, and the tendency to appear in certain clusters. (2) In this paper, we use the notions of hydrophobic and hydrophilic clusters since we assume that hydrophobic interactions as well as hydrogen bonding between residues are the most important interactions affecting the formation of these clusters. It should be pointed out, however, that other interactions such as electrostatic interactions also play a role in this process.

References and Notes

(1) 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).
(2) Meirovitch, H.; Rackovsky, S.; Scheraga, H. A. *Macromolecules* **1980**, *13*, 1398.
(3) Meirovitch, H.; Scheraga, H. A. *Macromolecules* **1980**, *13*, 1406.
(4) Krigbaum, W. R.; Komoriya, A. *Biochim. Biophys. Acta* **1979**, *576*, 204.
(5) Tanaka, S.; Scheraga, H. A. *Macromolecules* **1976**, *9*, 945; **1977**, *10*, 291.
(6) Manavalan, P.; Ponnuswamy, P. K. *Arch. Biochem. Biophys.* **1977**, *184*, 476.
(7) Manavalan, P.; Ponnuswamy, P. K. *Nature (London)* **1978**, *275*, 673.
(8) Crippen, G. M. *Biopolymers* **1977**, *16*, 2189.
(9) Warme, P. K.; Morgan, R. S. *J. Mol. Biol.* **1978**, *118*, 273.
(10) Kuntz[11] and Chothia[12] have shown that a large proportion of the inaccessible polar groups are involved in hydrogen bonding. From the available experimental data, the strengths of interpeptide hydrogen bonds and peptide–water hydrogen bonds are comparable.[13-15]
(11) Kuntz, I. D. *J. Am. Chem. Soc.* **1972**, *94*, 8568.
(12) Chothia, C. *Nature (London)* **1975**, *254*, 304.
(13) Schellman, J. A. *C. R. Trav. Lab. Carlsberg* **1955**, *29*, 230.
(14) Gill, S. J.; Huston, J.; Clopton, J. R.; Downing, M. *J. Phys. Chem.* **1961**, *65*, 1432.
(15) Klotz, I. M.; Franzen, J. S. *J. Am. Chem. Soc.* **1960**, *82*, 5241.
(16) In the model of Némethy and Scheraga,[17] the free energy of formation of a hydrophobic bond between nonpolar groups in water has two main contributions, the balance of the van der Waals interaction between all components of the system, and the changes in water structure accompanying association, with the latter being a larger contribution than the former.
(17) Némethy, G.; Scheraga, H. A. *J. Phys. Chem.* **1962**, *66*, 1773.
(18) Our view of cluster formation, in which the clusters form in the initial and subsequent stages of folding (due to hydrophobic and hydrogen bonding), differs from that of Krigbaum and Komoriya,[4] who attribute cluster formation to van der Waals forces that are operative in the final stages of folding.
(19) Kuntz, I. D.; Crippen, G. M. *Int. J. Pept. Protein Res.* **1979**, *13*, 223.
(20) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379.
(21) Rackovsky, S.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1977**, *74*, 5248.
(22) Krigbaum, W. R.; Rubin, B. H. *Biochim. Biophys. Acta* **1971**, *229*, 368.
(23) Havel, T. F.; Crippen, G. M.; Kuntz, I. D. *Biopolymers* **1979**, *18*, 73.
(24) Goel, N. S.; Ycas, M. *J. Theor. Biol.* **1979**, *77*, 253.