Influence of Water on Protein Structure. An Analysis of the Preferences of Amino Acid Residues for the Inside or Outside and for Specific Conformations in a Protein Molecule[1]

David H. Wertz[2a] and Harold A. Scheraga*[2b]

Department of Chemistry, Cornell University, Ithaca, New York 14853.
Received July 8, 1977

ABSTRACT: The x-ray structures of 20 proteins have been examined and each of the residues in these proteins was assigned to the inside or outside of the molecules and to a conformational state. The data obtained confirm that polar groups are generally found on the outside of proteins and nonpolar residues are generally found on the inside. Seven of the amino acids (Ala, Arg, Cys, His, Pro, Ser, Tyr) have inside/outside preferences which are not consistent with their usual assignment as *either* polar or nonpolar residues; explanations are given for these apparent inconsistencies. Of the three types of backbone structure considered here (extended, α helix, and nonregular), extended structures have the greatest preference for the inside of proteins, and nonregular structures have the greatest preference for the outside. It is suggested that differences in entropy play an important part in the inside/outside preferences of backbone structures. There are generally significant changes in the conformational preferences of the residues in going from the inside to the outside of proteins; environmental (rather than local) solute–solvent interactions seem to be the predominant cause of these changes in conformational preferences.

This paper summarizes information about the preferences of amino acid residues in proteins for the inside or outside *and* for specific conformations. Presumably, these preferences reflect, to some extent, the interactions of the residues with water. Studies related to the inside/outside preferences of residues have been reported previously.[3–6] However, as will be pointed out in the text, the analysis reported here was designed to provide somewhat different information.

Gō et al.[7] have summarized the various possible types of interaction of water with a polypeptide. The effect of solvent can also be examined from another point of view, in which water can affect the conformation of a protein in at least three possibly ways, viz., (1) the preferred conformation of a *protein* could be one which has as many nonpolar residues as possible on the inside of the protein while simultaneously having as many polar residues as possible on the outside; (2) water could act to stabilize or destabilize *entire* backbone conformations such as α helices and extended structures relative to nonregular structures; (3) the conformational preferences of each *residue* could depend on whether or not it is in contact with water. In an attempt to learn more about each of these three factors, we examined the x-ray structures of 20 globular proteins, essentially the same ones considered by Maxfield and Scheraga.[8] Each residue was assigned to a conformational state, using the criteria of Maxfield and Scheraga,[8] and also to either the inside or the outside of the proteins. It is *assumed* that the distribution of amino acids, of the backbone conformations, and of the conformations of the amino acids, between the inside and outside of the protein are affected by water–solute interactions; then, from an analysis of the conformational and inside/outside data, it is possible to obtain information about each of the above-mentioned three effects of water on the structure of proteins.

Method Used to Assign Residues to the Inside or Outside of a Protein Molecule

Each of the residues in the 20 proteins considered here was assigned to the inside or outside of the protein by means of the following seven-step algorithm.

(1) Using the Cartesian coordinates of each protein as supplied by the Brookhaven Data bank, two mutually perpendicular sets of parallel planes were passed through the protein molecules. One set of planes was parallel to the XY plane and the other was parallel to the XZ plane. Each set of planes was placed at 1-Å intervals. The first plane in the XY set was 1.7 Å below the most negative Z coordinate, and the first plane in the XZ set was 1.7 Å below the most negative Y coordinate.

(2) All planes passing through the protein molecule were examined, and the atoms whose van der Waals spheres[9] were intersected by each plane were identified.

(3) Three sets of lines were drawn through the protein, one set parallel to each of the axes. The lines in each set were parallel and 1 Å apart. The lines parallel to the X axis were drawn in each of the planes of the XY set, and the first line in each plane was 1.7 Å below the most negative Y coordinate in the protein. The lines parallel to the Y axis were drawn in each of the planes of the XY set, and the first line in each plane was 1.7 Å below the most negative X coordinate in the protein. The lines parallel to the Z axis were drawn in each of the planes of the XZ set, and the first line in each plane was 1.7 Å below the most negative X coordinate in the protein.

(4) The atoms that had been identified in step 2 were examined to determine which ones were intersected by the lines drawn in step 3. Also, the points of intersection of each line with the *surface* of the van der Waals sphere of an atom were recorded.

(5) The first and last atoms intersected by a line were considered to be on the surface of the protein; a "hit" was scored for both atoms found to lie on the surface by this criterion. To identify surface atoms in involuted folds (i.e., clefts), a 2-Å spacer was moved along the line that intersected the two previously located surface atoms; if this 2-Å spacer was placed in a position where it did not intersect any atoms, then the intersections of the line and the two atoms nearest to the ends of the 2-Å spacer in that position were also considered to be on the surface (of the cleft, in this case), and a "hit" was scored for both of these atoms.

(6) After all of the lines passing through the protein molecule were examined, an atom was assigned a score of 0 if it had fewer than 5 "hits", a score of 1 if it had 5 to 9 "hits", and a score of 2 if it had 10 or more "hits" (such a large number of "hits" is possible because spheres of 2.2 Å radii[9] can be intersected by many lines spaced 1 Å apart).

(7) The scores of all of the atoms in each residue (excluding hydrogens, which were not included in the x-ray coordinates used) were added, and the sum was divided by the number of nontertiary atoms (i.e., atoms bonded to only one or two

0024-9297/78/2211-0009$01.00/0

nonhydrogen atoms) in the residue. If this average score was equal to or greater than 1, the residue was considered to be on the outside of the protein; if the score was less than 1, it was considered to be on the inside. Tertiary atoms were not counted when determining the average score because experience showed that tertiary atoms rarely had much contact with water (i.e., they rarely had as many as 5 "hits"); if they had been included, they would have lowered the average score for the residue.

The method for assigning a residue to the inside or outside of a protein molecule, described above, differs significantly from the "rolling ball" method of Lee and Richards[3] since the latter does not assign small voids to the outside. However, the results of our method (when expressed on an atom-by-atom basis) are qualitatively the same as those of Lee and Richard,[11] who considered the exposure of atoms (not residues) to the outside. By considering residues, rather than atoms, the degree of exposure of a residue to the outside can be correlated with its conformational state.

Shrake and Rupley[4] assessed the exposure of atoms to water by assigning 92 points to ths surface of the van der Waals sphere of each atom and determining the number of such points which were not inside the van der Waals sphere of any other atom. As in our method, voids between atoms too small to accommodate a water molecule were assigned to the inside.

The method described in this paper was chosen because the algorithm is simple and easy to program and requires very little computer time. The 1.5 s of CPU time (IBM 370/168) required by this method to assign the residues of lysozyme to the inside or the outside is substantially less (estimated to be about 15 times less) than the time required by the procedure of Shrake and Rupley.[4] Lee and Richards did not publish estimates of the time required by their method. However, the method of Shrake and Rupley appears to be much simpler than that of Lee and Richards and hence is probably faster.

Tanaka and Scheraga[6] examined 25 proteins and determined whether the side chain of each residue made contact with other side chains or not. Residues which make no contacts are on the outside. However, since a residue can make contact with another residue and still be on the outside, their data cannot be compared with the data in this paper.

Relative Preferences of Amino Acids for the Inside and Outside of Proteins

We define a quantity P_{in} for any residue X as the number of interior X residues divided by the total number of X residues, in all of the proteins examined. The values of P_{in} for each of the 20 naturally occurring amino acids are given in Table I. Chothia[5] has determined the fraction of each residue that is completely buried (i.e., has less than 5% of its surface area exposed to water). This is a much more stringent condition than our requirement that most of the atoms of the residue have little or no contact with water for the residue to be considered inside. Given these differences in definition as to what constitutes "inside" ("buried"), it is not surprising that Chothia's values for the fraction of each residue that is buried and our values of P_{in} differ significantly. For example, Chothia's average value of P_{in} is 0.28, compared to ours of 0.57. We did not use Chothia's stringent requirement for assigning residues to the inside because such a requirement assigns many residues to the outside even though they have little contact with the solvent.

As is well known, polar residues are generally found outside and nonpolar residues are generally found inside the protein molecule. However, as pointed out earlier,[3,6] there are exceptions to this generalization.

The nonpolar residues Ala and Pro have values of P_{in} that are *less* than the average value of 0.57. Presumably the (small) methyl side-chain group of Ala is not sufficiently nonpolar (compared to the two polar peptide groups linking the Ala residue to the rest of the polypeptide chain) to force the Ala residue preferentially to the inside; the net effect is that Ala has no preference for either the inside or the outside. Considering that the pyrrolidine ring of Pro contains four methylene groups, it might seem surprising that Pro has a strong preference to be on the outside of protein molecules (only Lys has a stronger preference for the outside than Pro). However, recalling that pyrrolidine itself is miscible with water,[12] that polyproline is one of the most water-soluble polyamino acids,[13] and that crystalline polyproline has a strong affinity for water,[14] the preference of Pro residues for contact with water is consistent with its behavior in the model systems referred to above.[15]

The strong preference for Cys to be inside is accounted for by the fact that almost all Cys residues in proteins occur as cystine (with a disulfide bond of low polarity), and that its role as a bridge between two different parts of the polypeptide chain would shield it from contact with water no matter what its polarity were.

The absence of a *strong* preference of Ser for the outside is at first surprising, especially since Thr (even with an extra methyl group) does exhibit a strong preference for the outside. However, conformational energy calculations show that most of the low-energy conformations of Ser have a side chain–backbone hydrogen bond,[16] while the methyl group of Thr causes most of the conformations of Thr with side chain–backbone hydrogen bonds to be relatively high in energy.[16] Since the intramolecular hydrogen bond would have to be broken for the OH group to be completely hydrated, the net solvation energy will be more negative for Thr than for Ser. It would be desirable to test this explanation by examining x-ray structures of proteins for such side chain–backbone hydrogen bonds. However, since the positions of hydrogens are not determined at the present resolution of protein x-ray structures, such a test cannot be carried out at this time.

His and Tyr both contain groups that are polar and easily ionized and yet Table I shows that these residues are generally found inside proteins. The reason for this apparent anomaly is that the polar atoms in each of these residues constitute only one part of the side chain; the remainder of the side chains are nonpolar. An examination of our data on an atom-by-atom basis showed that these residues are usually placed so that the

Table I
Values of P_{in} for the Naturally Occurring Amino Acid Residues

Residue	P_{in}	Residue	P_{in}
Ala	0.52	Leu	0.77
Arg	0.49	Lys	0.31
Asn	0.42	Met	0.76
Asp	0.37	Phe	0.87
Cys	0.83	Pro	0.35
Gln	0.35	Ser	0.49
Glu	0.38	Thr	0.38
Gly	0.41	Trp	0.86
His	0.70	Tyr	0.64
Ile	0.79	Val	0.72
Av P_{in}[a]			0.57
P_{in}^{total}[b] = no. of residues inside / total no. of residues			0.54

[a] Average of the listed values of P_{in} of the 20 residues. [b] This value pertains to *all* residues in the 20 proteins, without distinguishing one kind of residue from another.

polar atoms of the side chain are outside the protein and the nonpolar atoms are inside. This same placement of the residues can be seen in the data of Lee and Richards.[3] Since the nonpolar parts of the His and Tyr side chains contain more atoms than the polar parts, the result is that, on the "average", these residues are inside the protein.[17]

Lys, like His and Tyr, has a polar group at the end of a larger nonpolar side chain, and yet Lys has the greatest preference for the outside (lowest P_{in}) of all the residues in Table I. This means that the entire side chain of Lys, not just the polar part, is usually outside the protein. We believe the reason that the whole Lys side chain is usually outside is that (when Lys is on the outside) each backbone conformation of Lys has an exceptionally large number of side-chain conformations, which are close in energy.[16] When Lys is on the inside, the side chain is fixed in one conformation. Entropy of mixing calculations, based on the calculated energies of *N*-acetyl-*N'*-methylamides of the naturally occurring amino acids,[16] show that when the average backbone conformation of Lys goes from a state where the side chain is fixed in one conformation (i.e., inside) to a state where the side chain exists as a statistical ensemble of conformations (i.e., outside) the entropy of mixing for the average backbone conformation increases by 4.9 eu. This is 2.8 eu greater than the entropy increase for the average residue when it goes from one side-chain conformation to an ensemble of conformations. [The *N*-acetyl-*N'*-methylamides of the residues are crude models for residues in a protein, but they are adequate for estimating the relative increase in the entropies of the residues when they go from one side-chain conformation (i.e., inside) to an ensemble of conformations (i.e., outside).] The entropy of mixing of the typical residue (that has a greater preference for the outside than the average residue) is 1.2 eu greater than the entropy of mixing of the typical residue (that has a greater preference for the inside than the average residue).

Arg, like Lys, is a charged residue, and we estimate that the entropy of mixing of the average backbone conformation of Arg is 1.5 eu greater than the entropy of mixing of the average residue. Yet Arg, unlike Lys, does not have a strong preference for the outside. The entropy of mixing of the average Arg conformation is significantly less than the entropy of mixing of the average Lys conformation, but one would still expect Arg to have a strong preference for the outside. However, Arg does not have a strong preference for the outside because it is often part of acid–base pairs. Maxfield and Scheraga (ref 8 and unpublished results) found that, when Arg was part of an α helix, 50% of such Arg residues were part of an intrahelical acid–base pair. Maxfield and Scheraga did not look for the presence of nonintrahelical acid–base pairs, but the data of Tanaka and Scheraga[6] imply that such acid–base pairs are quite common. Since water weakens the interactions between the groups of an acid–base pair, one would expect to find almost all acid–base pairs on the inside of the protein.

Relative Stabilities of Helical, Extended, and Nonregular Structures on the Inside and the Outside of Proteins

The criteria of Maxfield and Scheraga[8] were used to assign each residue of the 20 proteins to one of five conformational states (α_R, α_L, extended, ζ_R, ζ_L). Residues in the α_R conformation were subdivided into one of two conformational states. An α_{Rh} residue is part of a run of four or more residues in the α_R conformation (i.e., part of a right-handed α helix), and an α_{Ri} residue is an isolated α_R residue or part of a run of two or three α_R residues. Residues in the extended conformation were also subdivided into one of two conformational states: (1) to the ϵ_{ex} state if it was part of a run of four or more residues in the extended conformation (i.e., an extended structure) or (2) to the ϵ_i conformation if it was an isolated extended residue or if it was part of a run of two or three extended residues. This gave a total of seven conformational states (α_{Rh}, α_{Ri}, ϵ_{ex}, ϵ_i, α_L, ζ_R, and ζ_L). In Table II, these seven conformational states have been consolidated into three types of backbone structure, nonregular structure (all residues except α_{Rh} and ϵ_{ex} residues), helix (α_{Rh} residues), and extended structure (ϵ_{ex}).

Table II
Fraction of Inside and Buried Residues in Various Backbone Structures

Backbone structure	P_{in}	P_{bur}
Nonregular	0.43	0.20
Helix	0.60	0.28
Extended	0.68	0.43

Not all of the residues in surface nonregular, helix, and extended structures are on the exterior side of the structure and thus on the outside of the protein. One would expect up to half of the residues in backbone structures that lie on the surface of the protein to be on the side of the structure that faces the interior of the protein; therefore, such residues would be inside the molecule. In order to obtain a more accurate count of the residues that are part of backbone structures, *all* of whose sides are inside the protein, we introduce the concept of a "buried" residue. A buried residue is defined as one that (together with both residues on either side of it) is inside the protein. We may define a quantity P_{bur} (analogous to the definition of P_{in}) as the fraction of X residues that are buried. The values of P_{in} and P_{bur} for the three types of backbone structure are listed in Table II.

Both the α helices and β sheets, which are made up of extended structures, have a network of hydrogen bonds. Yet, the values of P_{bur} for these two structures differ greatly. Therefore, the inability of the atoms in these networks of hydrogen bonds to be solvated cannot be the reason for the larger relative preference of extended structures for the inside. In fact, if hydrogen bonds were the origin of this behavior, one would expect extended structures to prefer the surface more than helices because a significant fraction of extended structures are not part of a parallel or antiparallel hydrogen-bonded β structure and, therefore, their amide groups could be accessible to the solvent.

Chothia[5] determined the exposed surface area of 12 proteins and found, as we have, that residues in extended structures are more likely to be inside the protein than residues in α helices. Given the difference between Chothia's and our criteria for assigning residues to the inside, and given the difference between his and our criteria to assign residues to helical and extended structures,[18] the qualitative agreement between his results and ours is fortuitous. In addition, Chothia did not distinguish, as we have, between residues inside the protein, but part of a structure on the surface of the protein, and residues that are part of structures buried inside the protein.

Chothia[5] ascribed his observation (that, generally, the outer strands of a β sheet were less exposed to solvent than the residues in an α helix, and that the residues in the central strand of a three-stranded β sheet were much less exposed to solvent than the outer strands) to the assumption that the residues in β sheets are better able to shield each other from solvent. For the reasons given in the next paragraph, we believe that these observations do not result because β sheets are better at shielding each other from solvent but are due to the fact that extended structures are generally buried inside the protein and thus shielded from the solvent by other residues.

Table III
Change in the Apparent Value of $\Delta G°$ of Conformational Interconversion in Going from the Inside to the Outside

Residue	$\epsilon_i \rightleftarrows \epsilon_{ex}$		$\alpha_{Ri} \rightleftarrows \alpha_{Rh}$		$\epsilon_i \rightleftarrows \alpha_{Rh}$	
	Unadjusted[a]	Adjusted[b]	Unadjusted[a]	Adjusted[c]	Unadjusted[a]	Adjusted[c]
Ala	0.78	0.16	0.56	0.15	0.34	−0.07
Arg	0.42	−0.20	0.04	−0.37	0.01	−0.40
Asn	−0.41	1.03	1.10	0.69	−0.16	−0.57
Asp	0.38	−0.24	0.19	−0.22	−0.39	−0.80
Cys	0.50	−0.12	0.22	−0.19	0.58	0.17
Gln	0.07	−0.55	0.35	−0.06	0.15	−0.26
Glu	0.17	−0.45	0.55	0.14	−0.22	−0.63
Gly	0.46	−0.16	0.77	0.36	0.68	0.27
His	0.44	−0.18	0.16	−0.25	−0.08	−0.49
Ile	0.43	−0.19	0.43	0.02	0.47	0.06
Leu	0.18	−0.44	0.47	0.06	0.24	−0.17
Lys	0.50	−0.12	0.26	−0.16	−0.04	−0.45
Met	−0.17	−0.79	*d*	*d*	0.44	0.03
Phe	0.37	−0.25	1.59	1.18	0.81	0.40
Pro	0.03	−0.59	0.52	0.11	−0.06	−0.47
Ser	0.61	−0.01	0.54	0.13	0.30	−0.11
Thr	0.67	0.05	0.69	0.28	0.50	0.09
Trp	0.29	−0.33	0.29	−0.12	−0.20	−0.61
Tyr	0.20	−0.42	0.60	0.19	−0.20	−0.61
Val	0.16	−0.46	0.33	−0.08	0.30	−0.11
Av $\lvert\Delta(\Delta G°)\rvert$[e]	0.36	0.34	0.51	0.25	0.31	0.34

[a] Values of $\Delta(\Delta G°)$ from eq 6, expressed in kcal/mol. [b] Values of $\Delta(\Delta G°)$ from eq 10, with $[\Delta(\Delta G°)]_{group} = 0.62$. [c] Values of $\Delta(\Delta G°)$ from eq 13 with $[\Delta(\Delta G°)]_{group} = 0.41$. [d] Not enough data were available to obtain these values. [e] Average of the absolute values of the numbers in each column.

The conclusion that extended structures have a greater preference for the inside than the other types of structures is consistent with the observation that many of the homopolymers which adopt either the random coil or helical conformations in solution adopt the extended structure in the crystal.[19] For the reasons discussed in the Appendix, it appears that there is a greater increase in the entropy of nonregular and helical structures than of extended structures in going from the inside of the protein (where the librational motions of all types of residues are highly restricted and, therefore, the entropy is low) to the outside of the protein (where the restrictions on the librational motions are less severe and, therefore, the entropy is higher). This means that a protein will usually be more stable if it has nonregular or helical structures on the surface instead of extended structures.

The preference of nonregular structures for the surface of proteins is probably the result of three factors: (1) The absence of a network of amide hydrogen bonds implies that many of the amide groups in a nonregular structure are free to hydrogen bond with water. (2) The absence of a network of hydrogen bonds also implies that the entropy of a nonregular structure (on the surface of a protein) is higher than that of a helical or extended structure because intramolecular hydrogen bonds restrict the librations of a residue and cause the entropy of conformations with hydrogen bonds to be low;[16,20] on the inside, all types of structures have restricted librational motions. (3) While helical and extended structures are somewhat flexible, they cannot make sharp turns;[21] yet sharp changes in the direction of the polypeptide chain (which are classified here among the nonregular structures) seem necessary for the polypeptide chain to fold into a compact, globular structure. If the main role of nonregular structure were to join lengths of regular structure together, then most of the sections of nonregular structure should occur on the surface of the protein. We cannot ascertain the relative importance of these three possible causes for the preference of nonregular structure to lie on the surface of proteins. However, by the conformational criteria used here, we found more residues of the 20 proteins in nonregular structures than in either helical or extended structures. This large amount of nonregular structure would seem to be much more than would be necessary if the predominant role of nonregular structure were to supply turns to link lengths of regular structures together; i.e., factor no. 3 above cannot be the dominant one.

The errors in empirical protein conformation prediction algorithms based on short-range (intra-residue) interactions alone[8] are generally ascribed to the fact that medium- and long-range intraprotein interactions and protein–water interactions are not included in the algorithm. The fact that empirical conformational prediction algorithms, based on short-range interactions, have moderate success shows that the short-range interactions are more important than the long-range and protein–water interactions in determining the conformation of a residue.[22] However, since each of the three types of backbone structure in Table II have quite different preferences for the inside of proteins, the effects of protein–water interactions are not negligible, and the accuracy of protein prediction algorithms might be improved significantly if the effects of protein–water interactions could be accounted for. Using the values of P_{in} in Table I, it should be possible to predict whether or not a section of a peptide chain is inside or outside, irrespective of its conformation. Then, the data in Table II (which show that the conformational preferences differ on the inside and the outside) could be used to modify the predictions, depending on whether the section of the protein whose conformation is to be predicted is inside or outside.

Changes in Conformational Preferences in Going from the Inside to the Outside

Given the partial success of protein conformation prediction algorithms based on the assumption that short-range intraprotein interactions predominate, it would be useful to learn if a similar assumption about residue–water interactions would also be successful. Short-range intraprotein interactions may be defined as those within a residue, i.e., by ignoring interactions with neighboring residues. Let us define residue–water inreractions in an analogous manner: local solute–sol-

vent interactions are the interactions between a residue and water with no influence from the neighboring residues. The *changes* in the local solute–solvent interactions caused by the *presence* of other residues will be attributed to the environmental solute–solvent interactions. If the analogy between the short-range intraprotein interactions and the local solute–solvent interactions is valid, then the local solute–solvent interactions would predominate over the environmental interactions. The following discussion is concerned with the validity of this assumption.

The data of Table III were obtained in an attempt to determine the effects of the local solute–solvent interactions on the conformational preferences of the 20 naturally occurring amino acids. Consider the three conformational changes

$$\epsilon_i \rightleftarrows \epsilon_{ex} \tag{1}$$

$$\alpha_{Ri} \rightleftarrows \alpha_{Rh} \tag{2}$$

and

$$\epsilon_i \rightleftarrows \alpha_{Rh} \tag{3}$$

Equation 1 pertains to a reaction in which an isolated residue in the extended conformation is transferred from a nonregular structure to an extended structure; similarly, eq 2 pertains to helical residues. Equation 3 pertains to a reaction in which a residue in the ϵ_i conformation is transferred from a nonregular structure to an α-helical structure. In general, for the equilibrium

$$a \rightleftarrows b \tag{4}$$

between two conformational states a and b, the value of $\Delta G°$ may be written as

$$\Delta G° = -RT \ln (\text{no. of b/no. of a}) \tag{5}$$

To express how the preference of a residue for conformation a, relative to its preference for conformation b, changes in going from the inside to the outside of a protein molecule, we consider $\Delta(\Delta G°)$, the change in the values of $\Delta G°$, where

$$\Delta(\Delta G°) = -RT \ln (\text{no. of b outside/no. of a outside}) + RT \ln (\text{no. of b inside/no. of a inside}) \tag{6}$$

A positive value of $\Delta(\Delta G°)$ means that b/a is greater on the inside. The values of $\Delta(\Delta G°)$ for each residue, computed from eq 6, are listed in the "unadjusted" columns in Table III.

Since the only difference between the ϵ_i and ϵ_{ex} (or between the α_{Ri} and α_{Rh}) conformations is the conformations of the neighbors of the residue, the local solute–solvent interactions are by definition identical for the species on both sides of eq 1 and 2, respectively. Thus, if local solute–solvent interactions were to predominate (as the short-range intraprotein interactions do), then $|\Delta(\Delta G°)|$ for reactions 1 and 2 should be ~0 and, in any event, much *smaller* than $|\Delta(\Delta G°)|$ for a reaction (such as reaction 3) that involves a conformational change of the given residue. In fact, however, the opposite is observed; i.e., the average "unadjusted" values of $|\Delta(\Delta G°)|$ are greater for reactions 1 and 2 (0.36 and 0.51 kcal/mol, respectively) than for reaction 3 (0.31 kcal/mol). It thus appears that, while the sum of the local and environmental solute–solvent interactions has an important effect on the conformational preferences of the residues, it is the environmental solute–solvent interactions that predominate, *not* the local solute–solvent interactions.

It was shown in Table II that nonregular, helix, and extended structures have different inside/outside preferences. Thus, it could be argued that the values of $\Delta(\Delta G°)$ are the sums of two parts, the contribution of the individual residue, and a "group" contribution from the type of backbone structure of which the residue is a part. In order to remove this "group" contribution to $\Delta(\Delta G°)$, we can adjust the values computed by eq 6, as follows:

$$[\Delta(\Delta G°)]_{adjusted} = [\Delta(\Delta G°)]_{unadjusted} - [\Delta(\Delta G°)]_{group} \tag{7}$$

Since ϵ_i is part of a nonregular structure, the "group" equilibrium corresponding to equation 1 is:

$$\text{nonregular} \rightleftarrows \epsilon_{ex} \tag{8}$$

The $[\Delta(\Delta G°)]_{group}$ for this equilibrium can be obtained from eq 6 as:

$$[\Delta(\Delta G°)]_{group_1} = -RT \ln [(1 - P_{in})_{\epsilon_{ex}}/(1 - P_{in})_{nonregular}] + RT \ln [(P_{in})_{\epsilon_{ex}}/(P_{in})_{nonregular}] \tag{9}$$

$$[\Delta(\Delta G°)]_{group_1} = 0.62 \tag{10}$$

when $(P_{in})_{\epsilon_{ex}}$ and $(P_{in})_{nonregular}$ are taken from column 2 of Table II.

Since α_{Ri} and ϵ_i are part of nonregular structure, the "group" equilibrium corresponding to eq 2 and 3 is:

$$\text{nonregular} \rightleftarrows \alpha_{Rh} \tag{11}$$

The value of $\alpha(\Delta G°)$ for this equilibrium can be obtained from eq 6 as:

$$[\Delta(\Delta G°)]_{group_{2,3}} = -RT \ln [(1 - P_{in})_{\alpha_{Rh}}/(1 - P_{in})_{nonregular}] + RT \ln [(P_{in})_{\alpha_{Rh}}/(P_{in})_{nonregular}] \tag{12}$$

$$[\Delta(\Delta G°)]_{group_{2,3}} = 0.41 \tag{13}$$

when $(P_{in})\alpha_{Rh}$ and $(P_{in})_{nonregular}$ are taken from column 2 of Table II.

The values of $[\Delta(\Delta G°)]_{adjusted}$ for each residue are listed in the "adjusted" columns of Table III.

From an examination of the adjusted average values of $|\Delta(\Delta G°)|$ in Table III, it can be seen that the value for reaction 3 (0.34 kcal/mol) is *slightly* larger than the value for reaction 2 (0.25 kcal/mol) but that it is the same as the value for reaction 1 (0.34 kcal/mol). Again, if local solute–solvent interactions predominated, the average value of $|\Delta(\Delta G°)|$ should be *much* larger (not equal to or only *slightly* larger) for reaction 3 than for reactions 1 and 2. This result means that, even when the effects of the inside/outside preferences of the backbone structures are taken into account, local solute–solvent interactions do not predominate; this prevents us from determining the effects of local solute–solvent interactions on the conformational preferences of the residues from an examination of the conformational preferences of the individual residues in proteins.

The observation that local solute–solvent interactions do not dominate the total solute–solvent interaction, while the short-range intraprotein interactions do dominate the total intraprotein interactions, would seem to imply that the local/environmental solute–solvent interaction ratio is smaller than the short-range/long-range intraprotein interaction ratio. However, since the effects of short-range intraprotein interactions and the local solute–solvent interactions are assessed (from x-ray data on proteins) by examining each type of residue in an "average" environment in a protein, the implication that these ratios differ is not necessarily true. Depending on the environment of a residue, the medium- and long-range intraprotein interactions can either reinforce or diminish the conformational preference caused by the short-range interactions. When an average is taken over many environments, the medium- and long-range interactions tend to cancel, leaving only the invariant short-range interactions. This is presumably the reason that empirical prediction algorithms (based only on short-range interactions) work as well as they do. The environmental solute–solvent interactions are the changes in the local solute–solvent interactions caused by the

presence of other residues around a residue. Since the presence of other residues always serves to shield a residue from water, the environmental interactions, unlike the long-range interactions, always act in one direction. When an average is taken, the environmental interactions do not cancel; as a result, it becomes difficult to separate the effects of the local solute–solvent interactions from the environmental ones in their influence on the conformational preferences of the residues.

Thus, while short-range prediction algorithms work fairly well because of the dominance of short-range interactions, these algorithms could be improved by including the effect of solvent. However, one would have to include the *total* solvent effect, not just the local one.

Conclusion

This study confirmed the importance of the well-known preference for polar groups to be on the outside and for nonpolar groups to be on the inside of protein molecules. One of the most interesting results obtained here is that entropic effects, arising from the difference between the restricted crystalline-like environment of the inside of a protein molecule and the fluid less-restricted environment of groups on the outside (in water), play a major role in determining the placement of residues and of different types of backbone structures in the protein molecule.

Acknowledgment. We are indebted to Drs. L. G. Dunfield, G. Nemethy, and S. S. Zimmerman for helpful discussions and comments on this manuscript.

Appendix

Change in Entropy upon Restriction of Backbone Segments. Our hypothesis that extended structures gain less entropy than helical or nonregular structures in going from the inside of a protein to the surface was *originally* based on two factors: (1) In an extended structure, the end-to-end distance of each residue is close to the maximum possible length. Thus, if a residue in the extended conformation undergoes librational motion, or changes its conformation, its end-to-end distance will usually decrease. (2) The highly folded nature of a stable protein would fix the ends of any section of the peptide chain. With the ends of a section fixed, a residue in a protein can decrease its end-to-end distance only if the end-to-end distances of the neighboring residues increase. However, the neighbors of a residue in an extended structure (being themselves in the extended conformation) cannot increase their end-to-end distances easily. Thus, the librational motions of an extended structure will be restricted (no matter in which part of the globular protein it lies), and the entropy of the extended structure will be low. In helical or nonregular structures (even those with fixed ends), the residues are at neither their maximum nor minimum distances, and the end-to-end distance of a residue can adapt easily to the effects of librational motions; as a result the entropy will be greater when it is in an environment in which these librational motions can occur (viz., the surface).

The above arguments seem reasonable for a completely extended chain, i.e., one with all backbone dihedral angles equal to 180°. However, real extended structures are not completely extended (i.e., their dihedral angles are close to, but differ somewhat from, 180°), and it is not certain how valid the above arguments are for such "distorted" structures. To test the validity of these arguments on real extended structures, vibrational frequencies of *N*-acetyl-(Ala)$_5$-amide were calculated by the method of Wilson et al.[23] The entropy of the chains was computed from the vibrational frequencies. The flexible geometry force field used in these calculations will be described in another publication. The blocked penta-L-alanine chains were taken in the extended conformation (ϕ = −153°, ψ = 155°) and in the helical conformation (ϕ = −60°, ψ = −41°), respectively. The ends of the chain were fixed in three different ways: (1) A stiff bond was placed between the acetyl and amide end groups and given an equilibrium bond length equal to the actual distance between the end groups. (2) One atom in each end group was fixed in space. (3) Two atoms in each of the end groups were fixed in space.

Table IV
Vibrational Entropy of Extended and Helical Chains of *N*-Acetyl-(Ala)$_5$-amide with Different Restrictions

Type of restriction	$\Delta S_{\epsilon-\alpha}$, eu	Decrease in $\Delta S_{\epsilon-\alpha}$ upon restriction of ends, eu
Unrestricted	21	
Ends bonded (case 1)	21	0
1 atom at each end of chain fixed (case 2)	19	2
2 atoms at each end of chain fixed (case 3)	16	5

In the three restricted cases, and for the chain with unrestricted ends, the hydrogen bonds in the helical structures made their entropies *lower* than those of the extended structures. However, extended structures in proteins are often part of hydrogen-bonded β sheets, and helices in proteins are often irregular and thus lack good intrachain hydrogen bonds. For these reasons, the effects of the hydrogen bonds on the entropies of the models used in these calculations (viz., *single* extended chains and perfect helices) must be subtracted out before the entropies of the models can be used to answer the question under consideration here. This can be accomplished by examining how the difference in entropy between the helix and the extended structure changes when the ends go from the free to the restricted state. While the helix has a lower entropy both when the ends are free and when they are restricted, our hypothesis requires that the difference in entropy between the helix and the extended structure be smaller when the ends are restricted.

Table IV shows that fixing the atoms at the ends of the chains (cases 2 and 3) does cause the entropy difference between the extended and helical structures to decrease, as predicted. However, the fact that the presence of a *stiff* bond between the ends of the chains (case 1) does not cause the entropy difference to decrease means that our original hypothesis (that restriction of the end-to-end distance of the chains would result in a decrease in the entropy difference between the chains) is not correct. It appears that the entropy difference between the extended and helical structures changes only when the twisting and rotational motions of the ends of the chains are inhibited. It is worth pointing out that, as the restrictions on the movements of the atoms at the ends of the chains are increased (i.e., in going from case 2 to case 3), the entropy difference between the conformations decreases; also, the typical atom in even the most restricted of the (Ala)$_5$ chains studied here can move more freely than the atoms in a chain on the surface of a protein.

In conclusion, our original hypothesis (that entropy favors helices over extended structures on the outside) seems to be confirmed. These entropic effects arise from the fact that the polypeptide chain is folded; this restricts the rotational motion of the ends of the sections of the chain. This restriction causes a greater reduction of the entropy of extended structures (compared to helices) on the outside.

References and Notes

(1) This work was supported by research grants from the National Science Foundation (PCM75-08691) and from the National Institute of General

Medical Sciences of the National Institutes of Health, U.S. Public Health Service (GM-14312).
(2) (a) NIH Postdoctoral Fellow, 1975–1977. (b) Author to whom requests for reprints should be addressed.
(3) B. Lee and F. M. Richards, *J. Mol. Biol.*, **55**, 379 (1971).
(4) A. Shrake and J. A. Rupley, *J. Mol. Biol.*, **79**, 351 (1973).
(5) C. Chothia, *J. Mol. Biol.*, **105**, 1 (1976).
(6) S. Tanaka and H. A. Scheraga, *Macromolecules*, **9**, 945 (1976).
(7) M. Gō, N. Gō, and H. A. Scheraga, *J. Chem. Phys.*, **52**, 2060 (1970).
(8) F. R. Maxfield and H. A. Scheraga, *Biochemistry*, **15**, 5138 (1976).
(9) A *uniform* van der Waals radius of 2.2 Å was assumed for all nonhydrogen atoms for programming simplicity. A large radius was chosen to account for the volume not only of each nonhydrogen atom but also of the hydrogens that are attached to it (since hydrogen atoms are not reported in x-ray structures of proteins). A value of 2.2 Å is an intermediate one for the effective radii of several functional groups.[10] This radius was used in the calculation that assigns atoms to one of three classes according to their exposure to solvent. It is convenient (and adequate for such a qualitative assignment) to use a uniform radius.
(10) L. G. Dunfield, A. W. Burgess, and H. A. Scheraga, *J. Phys. Chem.*, submitted.
(11) Our method (on an atom-by-atom basis) was compared with that of Lee and Richards, using Lysozyme as an example. For this purpose, we determined the percent accessible area represented by 10 "hits" (which is our criterion to place an atom on the outside) and compared it with the corresponding quantity of Lee and Richards. The 10 "hits" in our criterion is equivalent to ~11% of the total surface area of a spherical atom, according to our method for determining hits. In all cases, the atoms assigned to the outside by our method were found by Lee and Richards[3] to have accessibility to the outside. Quantitatively, 82% of the atoms that we assigned to the outside had at least 11% accessibility, according to Lee and Richards. The remaining 18% of the atoms that we assigned to the outside had less than 11% accessibility. About half of the cases that differed are attributable to the use of different atomic radii in the two methods (if the same radii had been used, we estimate that agreement would have been obtained for ~90% of the atoms). The remaining 10% difference undoubtedly arises from differences in criteria for designating exposed surface area of the atoms.
(12) A. Pictet and G. Court, *Chem. Ber.*, **40**, 3771 (1907).
(13) A. Berger, J. Kurtz, and E. Katchalski, *J. Am. Chem. Soc.*, **76**, 5552 (1954).
(14) I. D. Kuntz, *J. Am. Chem. Soc.*, **93**, 514 (1971).
(15) A referee has suggested that the high occurrence of proline on the outside might be due to the fact that it is found often in β bends. β bends are distinct structures from α helices and extended structures and, therefore, in the context of this paper, are part of nonregular structure (see, also, the next section on relative stabilities of helical, extended, and nonregular structures). In the next section, it will be shown that nonregular structures have the greatest preference for the outside; therefore, β bends would have a large preference for occurrence on the outside. Proline appears to have a higher probability of occurrence in β bends than any other residue (S. S. Zimmerman and H. A. Scheraga, unpublished results) because (in contrast to other residues) it occurs almost exclusively in nonegular structures. However, if one calculates the probability for any residue in a *nonregular* structure to occur in β bends, then proline has an *average* (not an overwhelmingly large) probability to occur in β bends. Therefore, the greater-than-average preference of proline for the outside cannot be due to its occurrence in β bends.
(16) S. S. Zimmerman, M. S. Pottle, G. Nemethy, and H. A. Scheraga, *Macromolecules*, **10**, 1 (1977).
(17) G. Nemethy, I, Z. Steinberg, and H. A. Scheraga, *Biopolymers*, **1**, 43 (1963).
(18) Chothia[5] assigned essentially all residues in the proteins that he studied, apparently by visual inspection, as being in either α helices or extended structures. By the conformational criteria used here, there are more residues in nonregular structures than in α helices or extended structures.
(19) S. S. Zimmerman and L. Mandelkern, *Biopolymers*, **14**, 567 (1975).
(20) P. N. Lewis, F. A. Momany, and H. A. Scheraga, *Isr. J. Chem.*, **11**, 121 (1973).
(21) In this context, a turn is a section of the polypeptide chain (that connects two regular structures) in which the general direction of the polypeptide chain is significantly different at the end of the section from what it is at the beginning. Our use of the term "turn" here is to be distinguished from the general usage of the term "β turn" or "β bend."
(22) H. A. Scheraga, *Pure Appl. Chem.*, **36**, 1 (1973).
(23) E. B. Wilson, Jr., J. C. Decius, and P. C. Cross, "Molecular Vibrations", McGraw-Hill, New York, N.Y., 1955, Chapter 2.

Changes in Unperturbed Dimensions Accompanying Helix–Coil Transitions in Cross-Linked Homopolypeptides, with Special Reference to Poly(hydroxybutyl-L-glutamine)

Wayne L. Mattice

Department of Chemistry, Louisiana State University, Baton Rouge, Louisiana 70803.
Received August 8, 1977

ABSTRACT: Mean-square unperturbed radii of gyration, $\langle s^2\rangle_0$, have been computed as functions of degree of polymerization and helix content for cross-linked polypeptides of the poly(L-alanine) type. Zimm–Bragg statistical weights were assigned values appropriate for poly(hydroxybutyl-L-glutamine) in water. The g ($\equiv\langle s^2\rangle_{0,\text{ branched}}/\langle s^2\rangle_{0,\text{ linear}}$) for polypeptides with a finite degree of polymerization fall between the limits defined using random-flight statistics and rigid-rod behavior. For polypeptides of the molecular weight usually encountered, g varies strongly with helix content when helicity exceeds 20%. Partially helical polypeptides require substantially higher degrees of polymerization than do completely disordered polypeptides in order to attain the limiting g obtained from random-flight statistics.

Proteins frequently contain interchain cross-links. The most prevalent naturally occurring cross-link results from disulfide bond formation by two cysteinyl residues.[1] Covalent cross-links have also been chemically induced in protein complexes to determine which polypeptide chains are neighbors. This approach has been applied to chromatin,[2–5] ribosomes,[6] and membranes.[7] Treatment of mouse LA-9 cells with tetranitromethane, for example, produces a cross-link between the C-terminal half of histone H2B and the C-terminal half of histone H4.[5] Analysis of products of cross-linking reactions frequently includes gel permeation chromatography or polyacrylamide gel electrophoresis in either aqueous sodium dodecyl sulfate or aqueous acid–urea solution. Typical proteins are denatured under these conditions.[8] They may, however, still have a substantial fraction (up to ~50%) of their amino acid residues present in α helices.[9] Analysis of the transport properties of these proteins would be facilitated by knowledge of the effect of cross-linking on their unperturbed dimensions.

One approach to an estimate of the unperturbed dimensions would be to utilize random-flight statistics to compute parameters denoted by g[10] and f_i.[11]

$$g \equiv \frac{\langle s^2\rangle_0 \text{ for cross-linked molecule}}{\langle s^2\rangle_0 \text{ for analogous linear molecule}} \quad (1)$$

$$f_i \equiv \frac{\langle s^2\rangle_0^{1/2} \text{ for } i\text{th uncross-linked polypeptide chain}}{\langle s^2\rangle_0^{1/2} \text{ for cross-linked molecule}} \quad (2)$$

0024-9297/78/2211-0015$01.00/0