Journal of Mathematical Chemistry 8(1991)229–242

PROTEIN SECONDARY STRUCTURE PREFERENCES
Dependence on medium-range steric interactions

Davor JURETIĆ* and Robert W. WILLIAMS
Department of Biochemistry, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814-4799, USA

Abstract

This paper addresses the question to what extent steric properties of sequence neighbors effect the preferences of an amino acid residue to assume the α-helical or some other secondary structure conformation. We find that an amino acid has increased tendency to be in α-helical conformation when its sequence neighbors are bulky. This result is an outcome of our automated method for finding conformational preferences as functions of physical parameters important for protein folding. The steric environment for a given residue in a protein is defined as an average of water-accessible surface areas of its primary structure neighbors in extended conformation for model tripeptides. For all amino acids, including non-helix formers like glycine and arginine, the preference for the helical structure increases if their primary structure neighbors form a larger steric environment.

1. Introduction

Most amino acids have a weak preference for one of the secondary structures: α-helix, β-sheet or reverse turn [1]. This result is based on statistical analysis of proteins with known secondary structure. The analysis assumes that the identity or physico-chemical properties of sequence neighbors are not important for secondary structure formation. However, neighboring residues in the primary structure can have some influence on the formation of a secondary structure of any given residue in the protein [2,3]. Steric interactions between side chains were recognized by Lim [4] as being important for predicting α-helices and β-sheet secondary structures in proteins. In our preliminary communication [5], we reported that isoleucine, a β-sheet former, prefers the α-helix when found between bulky primary structure neighbors. By using a combination of statistical analysis and physico-chemical considerations, one of us was able to predict the location of both trans- and extra-membrane helical segments for the photosynthetic reaction center M subunit [6].

The purpose of this paper is to explicitly examine the influence of the primary structure steric environment of the residue on its preference for the secondary structure. Both the formation of secondary structures and their association to form

*Present address: Filozofski Fakultet u Splitu, Nikole Tesle 12, 58000 Split, Croatia, Yugoslavia.

the closely packed protein interior are accompanied by a reduction in the surface accessible to solvent [7]. The formation of intramolecular hydrogen bonds and the reduction in the exposed surface are considered as major factors in enhancing the stability of protein structures [8].

As a convenient parameter to measure local medium-range steric interactions, we use the mean solvent-accessible surface area [9] of neighboring residues in the extended chain. The accessible surface area of protein segments was used previously for the minimization of the specific volume [10], maximization of solvent exclusion [11], calculation of the coefficient of compactness [12], roughness index [13], globularity index [14], protein alignment method [15], and solvation free energy of folding [16, 17]. In this work, we use only the standard twenty values for accessible surface areas of residues in model tripeptides, as calculated by Chothia [18] or Rose [19]. When accessible areas of neighboring residues are averaged, an initial steric environment is obtained, and we can explore the correlation between the final secondary structure of the residue and its initial steric environment.

Our approach is similar to the multipoint moving average method of Rose [20] and Kyte and Doolittle [21]. The difference is that we omit the central residue from the moving average, but at the same time take into account its secondary conformation. This approach can be used only with proteins of known secondary structure. However, the results of our statistical analysis (which is an extension of the Chou – Fasman [22] method for finding conformation preferences) can be used to predict secondary structure.

2. Methods

The program written by Kabsch and Sander [23] is used to assign secondary structures to 212 proteins of known X-ray structure from the Protein Data Bank at Brookhaven National Laboratory. Kabsch – Sander secondary structure assignments are further reduced to: α-helix (3-, 4- and 5-helix), β-sheet (including β-bridge), turn (including bend) and undefined structure (undefined structure is actually defined as a piece of low curvature not in H-bonded structure [23]). Only monomers are considered from the proteins containing two or more polypeptide chains. Although the Kabsch – Sander program, DSSP, identifies free cysteine side chains, we do not distinguish between cysteine and cystine.

Our computer program, written in FORTRAN, calculates the "initial" steric environment of each residue n (of the type i in the secondary conformation j) in the primary structure of each protein. The "initial" environment of the residue n is defined to be the average of solvent-accessible surface areas (in the extended conformation for model tripeptides [18]) for residues $n - m$ through $n + m$, *excluding* residue n. The number m is usually taken to be 4. This process is repeated for each residue of the protein excluding the first and last m residues in each polypeptide chain. When residues B (Glu or Gln), Z (Asp or Asn) and X (unknown) are found

in the protein data set, average values for their accessible surface areas are used, i.e. 1.85, 1.55 and 1.70 nm^2, respectively.

The number of occurrences N_{ijk} of each amino acid type i in each secondary structure j is counted in each class k of the environment ($k = 1, \ldots, 9$). The class limits are chosen so that a similar number of residues (approximately 1800 for the protein data set we had) fall into each of nine classes. Preference values are then calculated as

$$P_{ijk} = (N_{ijk}/N_{ik})(N/N_j) \tag{1}$$

where

$$N_{ik} = \sum_{j=1}^{4} N_{ijk} \tag{2}$$

and

$$N_j = \sum_{k=1}^{9} \sum_{i=1}^{20} N_{ijk} \,. \tag{3}$$

The total number of effective residues N is obtained by summing N_{ijk} over i, j and k. Summing over all 4 secondary conformations N_{ik} is obtained (eq. (2)) as the total number of residues of type i found in the environment of class k. The preference value (eq. (1)) is proportional to the probability

$$p_{ijk} = N_{ijk}/N_{ik} \tag{4}$$

that residue i is found in the secondary conformation j (out of 4 possible conformations) within the class k of the steric environments. The proportionality factor N/N_j is the inverse fraction of the conformation j in the protein data set.

Related proteins, such as the hemoglobins, cytochromes etc., were averaged as described by Levitt [1]. A list of 100 unrelated proteins was also created by taking only one or two proteins from each family of related proteins. Occasionally, two proteins from the same protein family were both included in the list of 100 unrelated proteins, because most of their steric environments, for the residues in the corresponding positions along the primary sequence, were different. The calculations with the input from a shorter (100 proteins) protein list gave essentially the same result as with a longer protein list (212 proteins with averaging of related proteins). When not stated otherwise, the averaging subroutine was used for weighing of related proteins. Both protein lists are available from the authors on request.

A linear regression line – preference versus steric environment (in nm^2) – was drawn through the nine points P_{ij1}, P_{ij2}, to P_{ij9} for each amino acid type i in each conformation j, and the sample regression coefficient (slope) was found by using the General Linear Models procedure (SAS Institute, Inc., Box 8000, Cary, NC 27511). The null hypothesis that the slope equals zero was tested in each case as

described in SAS User's Guide, Statistics (Version 5, 1985), p. 486, and in the legend of table 2.

The random local steric environment of each residue n was created in the following way. Eight residues, whose total surface area was to be averaged, were taken at random from a set of 147 residues. The frequencies of occurrence of the residues in that set were chosen so that they closely correspond to the frequencies found in the protein data set.

3. Results and discussion

3.1. PREFERENCES ARE FUNCTIONS OF SEQUENCE STERIC ENVIRONMENT

The upper panel in fig. 1A shows how the preference of the leucine for the α-helical conformation (circles) and for the β-sheet conformation (triangles) varies with the bulkiness of its initial environment. Leucines found in highly bulky environments (1.82 nm^2) prefer the α-helical to the β-sheet conformation. The opposite is true for the leucines found in the least bulky environments (1.28 nm^2). Leucine has been classified as an α-helix favoring amino acid in the protein data set containing 66 globular proteins [1]. β-former [1] valine (third panel from the top in fig. 1A) becomes indifferent between β-sheet and α-helix conformation in a very bulky environment. The preference of glutamate for the α-helix conformation (second panel from the top in fig. 1A) is enhanced in very bulky environments. Similarly, arginine (last panel in fig. 1A), indifferent between α-helix, β-sheet and reverse turn conformation on average [1], has a clear preference for the α-helix conformation in very bulky environments. The preferences of the same 4 residues for turn (triangles) and undefined conformation (circles) are given in fig. 1B.

Linear regression lines with positive slope for preference versus steric environment are found for *all* amino acids in the α-helix conformation (table 1 and fig. 1A). Negative slopes are found for amino acids in the β-sheet, reverse turn and undefined conformation (fig. 1). F values in table 1 test the null hypothesis that the slope of the linear regression line is equal to zero. A large F value indicates a large absolute value of the slope and/or a small standard error of the slope. A very small probability P of getting a larger F value (if the slope is truly equal to zero) leads to the conclusion that the independent variable (accessible surface) contributes significantly to the model (table 1). Choosing $P = 0.01$ as the cutoff significance level, we find a significant positive dependence of secondary structure preferences on local steric environment (accessible surface) for 14 amino acids in the α-helix conformation (table 1). In the case of the β-sheet, reverse turn and undefined conformation 5, 2 and 3 significant (negative) dependencies are found, respectively (not shown).

Table 2 illustrates the strong effect of a very high or very low steric environment on many preferences. For the data set of 212 proteins, the most bulky, the average, and the least bulky initial steric environment (per residue) is 1.82, 1.56 and

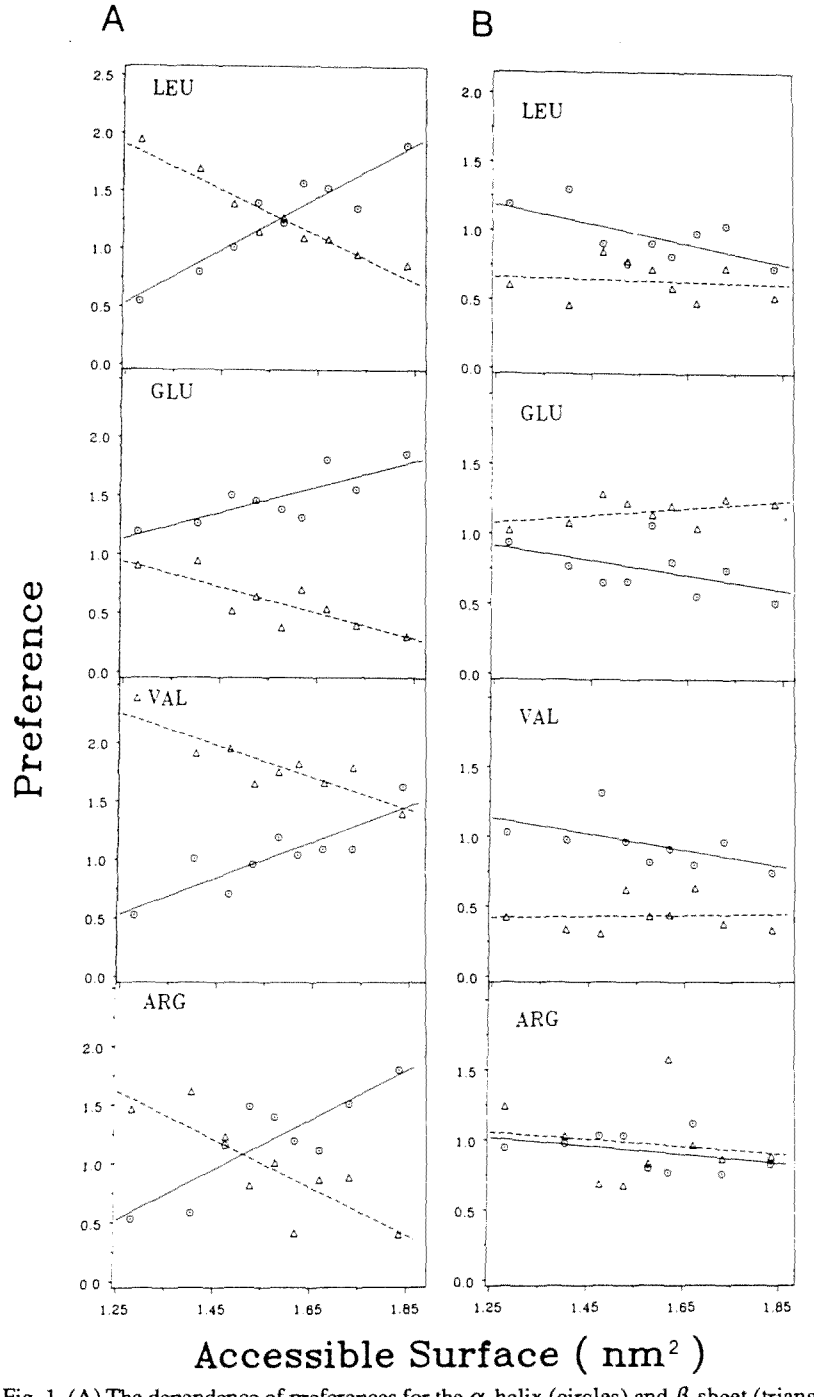

Fig. 1. (A) The dependence of preferences for the α-helix (circles) and β-sheet (triangles) conformation and (B) for the reverse turn (triangles) and undefined (circles) conformation on local steric environment. The preferences for Leu, Glu, Val and Arg are calculated as described in section 2 (by using the data set of 212 proteins) and the linear regression line drawn through the nine points ($k = 1, \ldots, 9$).

Table 1

Linear regression analysis[a] of amino acid preferences for α-helix

Amino acid	Intercept	Slope[b] b	Standard[c] error of b: s_b	F value[d]	Probability[e] $Pr(t^2 > F)$
Ala	− 1.47	1.85	0.39	22.26	0.0022
Cys	− 0.14	0.65	0.61	1.14	0.3203
Leu	− 2.37	2.32	0.34	45.65	0.0003
Met	− 3.64	3.19	0.75	18.16	0.0037
Glu	− 0.28	1.12	0.30	13.59	0.0078
Gln	− 2.61	2.45	0.33	56.30	0.0001
His	− 1.02	1.33	0.53	6.30	0.0404
Lys	− 1.78	1.91	0.40	23.09	0.0020
Val	− 1.44	1.58	0.35	19.74	0.0030
Ile	− 1.88	1.85	0.43	18.49	0.0036
Phe	− 2.65	2.45	0.59	17.18	0.0043
Tyr	− 1.31	1.34	0.47	8.29	0.0237
Trp	− 2.44	2.23	0.75	8.86	0.0206
Thr	− 1.66	1.59	0.21	55.34	0.0001
Gly	− 0.91	0.90	0.09	105.71	0.0001
Ser	− 1.31	1.30	0.21	39.62	0.0004
Asp	− 1.16	1.37	0.36	14.65	0.0065
Asn	− 1.55	1.47	0.59	6.22	0.0413
Pro	− 0.59	0.68	0.45	2.35	0.1695
Arg	− 2.18	2.16	0.49	19.25	0.0032

[a]The preference value (eq. (1)) is a dependent variable Y, while the average surface area in nine bins is an independent variable X. Their average values are $\bar{Y}$ and $\bar{X}$, respectively.

[b]Writing $x = X - \bar{X}$ and $y = Y - \bar{Y}$, the sample regression coefficient is calculated as:

$$b = \sum xy / \sum x^2.$$

[c]Writing $d_{xy} = Y - \hat{Y}$ for the $n = 9$ deviations from the regression line $\hat{Y} = \bar{Y} + bx$, the sample standard deviation of the regression coefficient is

$$s_b = \sqrt{\sum d_{xy}^2 / \left(\sum x^2\right)(n-2)}.$$

[d]The F value is calculated as:

$$F = (b/s_b)^2 = t^2,$$

where t is Student's t value for testing the null hypothesis that the parameter (slope) equals zero.

[e]The level of significance: $Pr(t^2 > F)$, associated with the observed F ratio.

1.28 nm^2, respectively, when roughly 10% of the residues are included in the high and 10% in the low category. As an example, to find the conformational preferences of alanine in such environments (table 2), we inserted these numbers (as x) into a straight line equation for the alanine in helix, which is from table 1: $1.85x - 1.47$.

Table 2

Preferences in 10% high, average, and 10% low steric environments

Amino acid	Conformation					
	α-helix			β-sheet		
	HH[a]	HA	HL	SH	SA	SL
Ala	1.90	1.42	0.90	0.73	0.73	0.73
Cys	1.04	0.87	0.69	1.49	1.47	1.45
Leu	1.85	1.25	0.60	0.73	1.25	1.82
Met	2.17	1.34	0.44	0.92	1.27	1.65
Glu	1.76	1.47	1.15	0.29	0.59	0.89
Gln	1.85	1.21	0.53	0.77	0.96	1.17
His	1.40	1.05	0.68	0.48	1.00	1.55
Lys	1.70	1.20	0.66	0.60	0.78	0.98
Val	1.44	1.02	0.58	1.42	1.79	2.18
Ile	1.49	1.01	0.49	1.41	1.62	1.84
Phe	1.81	1.17	0.49	1.00	1.34	1.70
Tyr	1.13	0.78	0.41	1.60	1.59	1.58
Trp	1.62	1.04	0.41	0.90	1.34	1.81
Thr	1.23	0.82	0.38	0.93	1.19	1.46
Gly	0.73	0.49	0.24	0.57	0.66	0.75
Ser	1.06	0.72	0.35	0.82	0.97	1.13
Asp	1.33	0.98	0.59	0.42	0.42	0.42
Asn	1.13	0.74	0.33	0.54	0.65	0.76
Pro	0.65	0.47	0.28	0.09	0.40	0.74
Arg	1.75	1.19	0.58	0.43	0.97	1.54

[a] The notation used is HH, HA and HL for high (10%), average, and low (10%) steric environment, respectively, of the residues found in the α-helical conformation. SH, SA and SL is the analogous notation for β-sheet residues.

The same procedure was repeated for other amino acids in helical and sheet conformation. A large number of previously unknown strong preferences or dislikes of amino acids for particular secondary conformation becomes apparent when their initial steric environment in the primary structure is taken into account. For instance, the underlined preferences are for the amino acids that are from 7 (Pro) to 2 (Phe) times more likely to assume helical than sheet conformation when found in a high steric environment. In a low steric environment, these *same* amino acids are from 4 (Met) to 2 (Gln) times more likely to assume sheet conformation. For an average steric environment, the preferences are very similar to the ones found recently by Lundeen [24].

The disadvantage of the straight line approximation for preference functions can be seen clearly in the proline β-sheet preference (table 2), which becomes negative

for a high enough steric environment. This happens, for instance, with proline 71 from cytochrome c, whose local steric environment is very bulky – 1.94 nm^2. The lack of a dependence on steric environment in the case of Ala and Asp preference for β-sheet conformation (not shown) may be due to the linear approximation as well. Such cases are rare, so that the linear regression line provides a very good first approximation for preference functions. "Natural" preference functions, associated with the scanning window averaging procedure, were derived by one of us [25], and shown to be well approximated with straight line functions in most cases.

As another test for our observations, a new database has been developed in collaboration with Dr. B. Lee from NIH, which contained ninety different polypeptide chains with a resolution of less than 0.3 nm (unpublished data). Also, we used a different computer environment (Apollo Domain System at NIH, Bethesda, MD), but the results, i.e. the dependence of conformational preferences on sequence steric environment, were almost identical.

We did not calculate actual accessibility values in fully folded proteins. Such calculations have been done recently, although the authors did not distinguish between different secondary structures [26]. Importantly, they found that clustering for accessibility does exist along the sequence.

We also modified the sliding window procedure so that only residues in the α-helical conformation (4 left and 4 right of the central residue) are averaged. This procedure collected environments from the middle of longer helical segments. The slope of the linear regression line – preference for middle-helix conformation versus environment (average solvent-accessible surface area) – increased for all amino acids except histidine, aspartate, proline and glutamine.

When the mean volume of the residues buried in the proteins is used (table 2 in [8]) to define steric environments along the sequence, the observed dependence of the secondary structure preferences on the average volume is very similar (results not shown) to the results presented above. This is not surprising in view of the correlation between volume and surface area that must exist for each residue.

The observed strong dependence of the preferences for the secondary structure on local steric environment (fig. 1 and table 1) should disappear if residues to be averaged are not neighboring residues. We tested this conjecture by creating a random steric environment for each residue as described in section 2. As an example, the results for alanine in the α-helical configuration are shown in fig. 2. The dependence on local steric environment (fig. 2A) disappears (fig. 2B) when "neighboring" residues are picked up at random from the total set of residues representative of our protein data set. The same result (the disappearance of the dependence of preferences on steric environment) is obtained with nineteen other amino acids in a random steric environment (results not shown).

Preference (eq. (1)) is proportional to the probability (eq. (4)) of finding residue i in secondary conformation j. In fig. 3, we plot the average probability for finding the α-helix conformation of the residue as a function of its local steric environment. The average was taken over twenty probabilities for individual amino

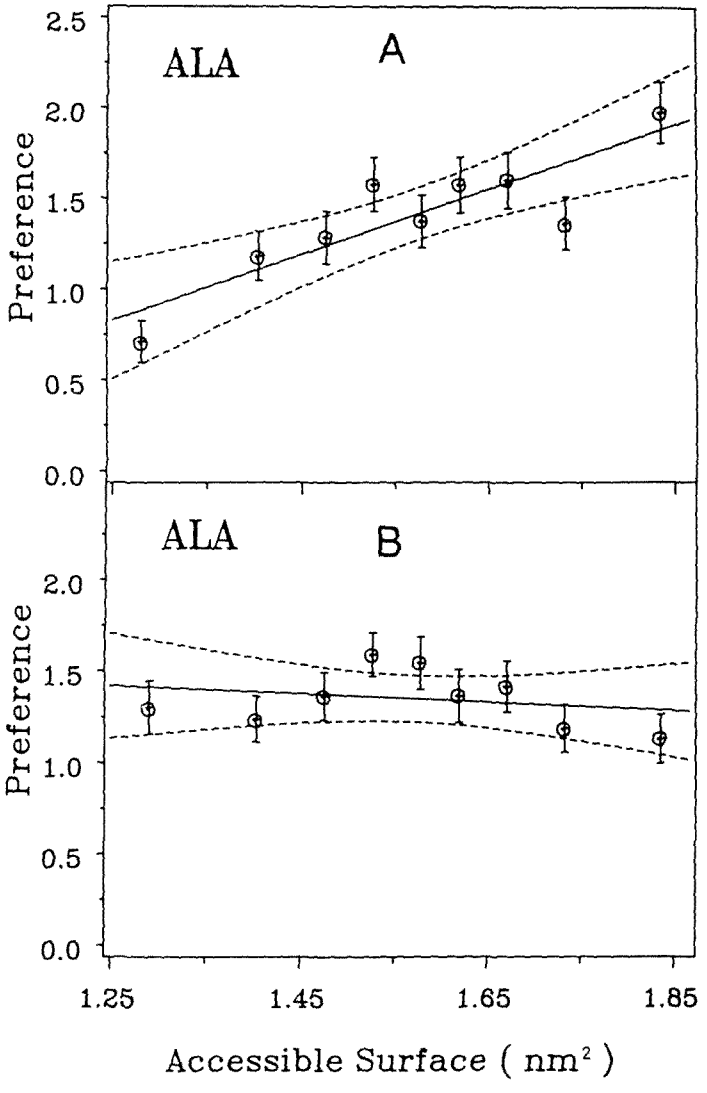

Fig. 2. The preference for the α-helical conformation of alanine when (A) alanine neighbors are sequence neighbors and (B) alanine "neighbors" are chosen at random (see section 2). An error analysis for the scatter of the preference values is also included. Vertical bars are interval estimations (one standard deviation) based on the assumption that preference values are approximately normally distributed in the y direction. Dashed lines are 95% confidence limits for linear regression. One hundred different proteins were sampled.

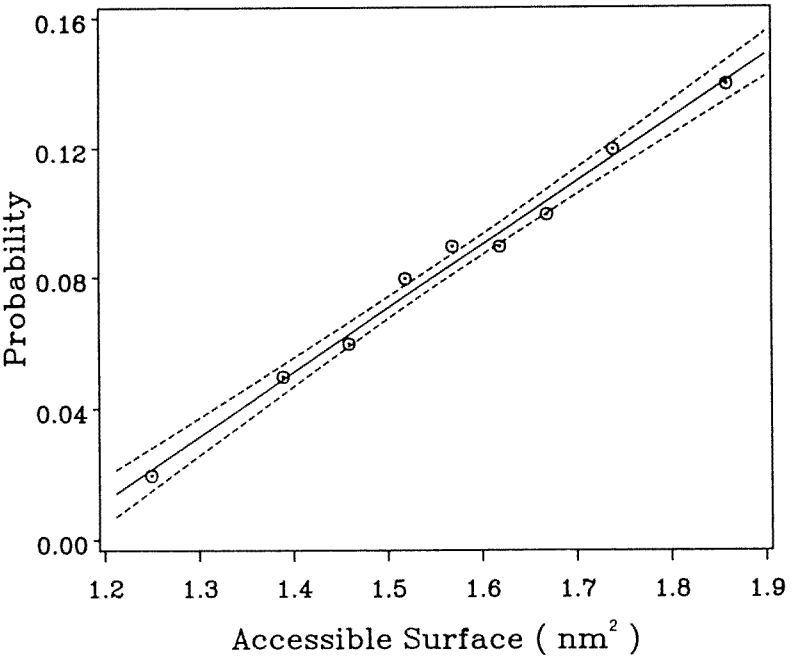

Fig. 3. Increased probability of α-helix conformation in bulkier primary structure environments. A sliding window of seven residues is used in the computer experiment, in which only longer helical segments and only residues from the middle helix region are examined with respect to their steric environment. Dashed lines are 95% confidence limits for linear regression analysis of the data from 100 proteins.

acids. Only residues from the middle of the longer helical segments were sampled as described above. An increased preference for the α-helical con-formation in the neighborhood of higher initial solvent-accessible surface areas suggests an increased probability that α-helix conformation will be formed in such neighborhoods during the folding process.

The results presented up to now indicated that the average steric environment of each residue in the α-helix conformation might be bulkier (in nm^2 of solvent-exposed surface area) than the steric environment for the same residue in the β-sheet conformation. This is indeed the case. Each amino acid type has highest (average) steric environment when in the α-helical conformation. When these 20 numbers (not shown) were averaged, the result was that for the central residue in the α-helix conformation, the mean solvent-accessible area of its sequence neighbors is significantly higher (1.601 ± 0.018 nm^2) than when such a residue is in the β-sheet conformation (1.549 ± 0.020 nm^2).

We next examined the average area of all nonapeptide segments *without* excluding the central residue (the data from 100 different proteins were used). When the central residue is in the α-helix conformation, the mean solvent-accessible

area is higher (1.614 ± 0.053 nm^2) than when the central residue is in the β-sheet conformation (1.549 ± 0.047 nm^2). This is true for each of twenty amino acid types (results not shown), as when environments are averaged. When the accessible surface area (or some other physico-chemical attribute) of the central residue in the sliding window is not excluded from the averaging procedure, this is equivalent to adding a constant (1/9)th contribution to the average. One undesirable result of adding a constant term to the moving average is that the distinction between helix and sheet conformation becomes less clear. Preference functions calculated by the inclusion of the central residue attribute are similar to the ones shown in fig. 1 (results not shown). The correlation between secondary structure and different chemical attributes of the amino acid sequence has been considered by Dunker et al. (private communication).

There is no apparent correlation between solvent-accessible surface areas of individual amino acids and of its eight neighbors. The preferred steric environment (in the primary structure) does not reflect the physical attribute of a given residue in a central position. A similar observation was made about the preferred hydrophobic environment of residue in protein interior [27], which was not very well correlated with the hydrophobic index of given residue. On the other hand, the bulkiness of its primary structure environment is clearly correlated with the secondary structure of the residue.

We also examined the influence of increased window size (increased size of local environment). Averaging any number of neighbors between 8 and 16 (m = 4 to 8) does not change the observed association of the α-helix conformation with bulkier neighbors and that of the β-sheet conformation with less bulky neighbors (not shown).

3.2. THE IMPLICATIONS FOR FOLDING AND SECONDARY STRUCTURE PREDICTION

The scales of accessible surface areas that are used in this paper (from Rose [19] and Chothia [18]) are relevant for the initial situation when polypeptide is kept in the extended form and before folding starts. Our results indicate that a helix (i.e. 3-, 4- or 5-helix) can nucleate more easily in a local primary structure environment of higher initial solvent-accessible surface area. α-helix formation may require steric protection offered by sequence neighbors of the residues with helix propensity. The hydrophobic contribution to the folding process, due to the buried surface, is indeed very large. For hydrophobic amino acids, Chothia [7] suggested a linear correlation between the accessible surface area of amino acid side chains and the free energy of transfer from water to organic solvent. For each square nm removed from contact with water there was a gain of 2.5 kcal/mol of hydrophobic free energy. Recently, the empirical correlation with hydrophobic free energy was extended to the apolar surface area of amino acids as well [28], so that a linear relation was established between surface area and transfer energy of all twenty coded amino acids.

The unsolved folding problem is how to find, from the primary sequence alone, accurate folding instructions for the creation of α-helices. A "consensus" pathway of folding [29] suggests that locally folded regions of the polypeptide chain appear first in microseconds. These regions are most likely α-helical structures that involve only a few amino acids very close to one another in the sequence. In attempts to predict the location of helical segments, one can apply different folding codes of a statistical or stereochemical nature [30,31], including the results described in this paper.

For instance, helices 4–17 in myoglobin (sperm whale), 94–105 in flavodoxin (clostridium), 17–30 in L-arabinose binding protein (E. coli), 96–120 in glutathionine reductase (human erythrocyte), and 108–117 in rhodanase (bovine liver) all have a broad maximum in the profile of steric environments associated with their residues. The height of the maximum (around 2 nm^2) is such that only Gly and Pro would prefer not to be in the helical conformation. To see this, one can use table 1 as descibed in section 3.1. Therefore, it would be straightforward to predict a helical conformation for these segments even if they include many residues that nominally are not helix formers.

The Chou and Fasman method [32–34] does not include information about interactions with neighboring residues (except in averaging preferences). Therefore, we can expect improvements when medium range steric interactions are taken into account. For example, it is easy to see that the number of underpredicted helical regions in the original Chou and Fasman procedure [22] can be reduced by 50% by using our preference functions instead of constant preferences (table 3). This improvement is minimal in the sense that refinements such as preferences for helical boundaries [30] are not applied.

Table 3

Underpredicted helical regions in the original Chou–Fasman procedure (C&F) [22]: Improvements (underlined) achieved by averaging our steric-dependent preferences (J&W)

Underpredicted helical regions	Average preference		Average steric environment (nm^2)
	J&W	C&F	
α-Hemoglobin 36 – 42	<u>1.05</u>	0.84	1.67
β-Hemoglobin 35 – 41	<u>1.50</u>	0.93	1.86
Cytochrome c 49 – 54	0.89	0.93	1.53
Cytochrome c 71 – 75	<u>1.14</u>	0.92	1.72
Cytochrome b_5 80 – 86	<u>1.08</u>	0.92	1.66
α-Chymotrypsin 164 – 173	<u>1.04</u>	0.83	1.70
Elastase 164 – 170	0.84	0.91	1.49
Elastase 237 – 245	0.98	0.97	1.58
Subtilisin BPN′ 5 – 10	0.73	0.81	1.52
Subtilisin BPN′ 103 – 110	0.93	0.87	1.65
Subtilisin BPN′ 242 – 252	<u>1.04</u>	0.94	1.62

As expected, high (average) helix preference correlates with high (average) steric environment of the segment (table 3). Correct predictions are achieved when the average steric environment of the segment (last column in table 3) is higher than the overall average of 1.56 nm^2. For instance, the β-hemoglobin 35 – 41 segment is found to be in the 3-helix conformation by the Kabsch – Sander program [23]. This segment has residues Tyr – Pro – Trp – Gln – Arg – Phe that favor β-sheet or turn conformation (with the exception of Gln), and is predicted to be non-helical by Chou and Fasman due to its low average helix preference (0.89). However, the very high average steric environment (1.86 nm^2) of this segment and, accordingly, the high average helix preference (1.50) calculated from the table 1 data, does not leave any doubt that helix conformation is the most probable for these residues. This is one example of how steric interactions can change the overall preferences of the polypeptide segment for secondary structures. Several amino acids that have high β-sheet preference values, as determined by Chou and Fasman [22] or Levitt [1], prefer a helix structure when they are found grouped together. Polypeptide segments rich in Phe, Trp, Ile, Leu, Met and Val may be underpredicted as helical segments in the Chou – Fasman algorithm that averages high β-sheet preferences of these amino acids. Such segments are often found as trans-membrane helices of membrane proteins.

Rose and collaborators [19] found a proportionality between the mean area buried on transfer from the standard state to the folded protein and the area in the standard state (as defined in their paper). The mean area buried is proportional to the hydrophobic contribution to the conformational free energy [13]. Therefore, the initial solvent-accessible surface area (in the standard state) must also be proportional to the free energy change due to hydrophobic contributions. When we used Rose's hydrophobicity scale (column IV in table 1 of [19]) for the average *buried* solvent-accessible area per residue, we found an even stronger positive correlation between the helical conformation of the residue and the potential to bury its local primary structure environment [25]. The buried surface scale [19] turned out to be better than other hydrophobicity scales [16,35] in locating both trans- and extra-membrane helical segments for the photosynthetic reaction center M subunit [6].

Chothia [18] and Richards and Richmond [36] have shown that a residue going into a β-sheet conformation loses more of its accessible surface than a residue acquiring an α-helix conformation. However, one must take into account that two helices coming together in a protein, in a surface–surface recognition process, are responsible for much further reduction in a solvent-accessible surface area [37]. In effect, comparing β-sheets with isolated helix structures is not instructive for the closely-packed situation in a fully folded protein, and may be misleading in the initial situation where an α-helix can form in less than 10^{-6} seconds [38]. A higher solvent-accessible surface area of nearest neighbors to the residue destined for the α-helical conformation will contribute the thermodynamic drive for the creation of a scaffold of the α-helices. As autonomous folding units [31], α-helices are the most likely controllers of the fast initial folding process. More stable structures can be selected between these "seeds for folding" [39] in the latter stages of folding.

Acknowledgements

The authors thank Dr. B. Lee, Dr. K. Dunker, and Dr. H.V. Westerhoff for helpful discussions, and S. Louchran for help in programming. This work was supported by National Science Foundation Grant PCM-8443154 and USUHS Grant GM7160.

References

[1] M. Levitt, Biochemistry 17(1978)4277.
[2] J. Garnier, D.J. Osguthorpe and B. Robson, J. Mol. Biol. 120(1978)97.
[3] N. Qian and T.J. Sejnowski, J. Mol. Biol. 202(1988)865.
[4] V.I. Lim, J. Mol. Biol. 88(1974)857.
[5] D. Juretić and R.W. Williams, Biophys. J. 51(1987)235a.
[6] R.W. Williams and S. Loughran, Biophys. J. 51(1987)234a.
[7] C. Chothia, Nature (London) 248(1974)338.
[8] C. Chothia, Ann. Rev. Biochem. 53(1984)537.
[9] B. Lee and F.M. Richards, J. Mol. Biol. 55(1971)379.
[10] A.M. Lesk and G.D. Rose, Proc. Nat. Acad. Sci. USA 78(1981)4304.
[11] A.A. Rashin, Nature (London) 291(1981)85.
[12] M.H. Zehfus and G.D. Rose, Biochem. 25(1986)5759.
[13] F.M. Richards, Ann. Rev. Biophys. Bioeng. 6(1977)151.
[14] S.J. Wodak and J. Janin, Biochem. 20(1981)6544.
[15] A. Sali and T.L. Blundell, J. Mol. Biol. 212(1990)403.
[16] D. Eisenberg and A.D. McLachlan, Nature (London) 319(1986)199.
[17] L. Chiche, L.M. Gregoret, F.E. Cohen and P.A. Kollman, Proc. Nat. Acad. Sci. USA 87(1990)3240.
[18] C. Chothia, J. Mol. Biol. 105(1976)1.
[19] G.D. Rose, A.R. Geselowitz, G.J. Lesser, R.H. Lee and M.H. Zehfus, Science 229(1985)834.
[20] G.D. Rose, Nature (London) 272(1978)586.
[21] J. Kyte and R.F. Doolittle, J. Mol. Biol. 157(1982)105.
[22] P.Y. Chou and G.D. Fasman, Biochem. 13(1974)222.
[23] W. Kabsch and C. Sander, Biopolymers 22(1983)2577.
[24] M. Lundeen, J. Inorg. Biochem. 27(1986)151.
[25] D. Juretić and B. Lee, Biophys. J. 55(2/2)(1989)354a.
[26] D.J. Lipman, R.W. Pastor and B. Lee, Biopolymers 26(1987)17.
[27] P. Manavalan and P.K. Ponnuswamy, Nature (London) 275(1978)675.
[28] C. Frommel, J. Theor. Biol. 111(1984)247.
[29] J. Szulmajster, Biosci. Rep. 8(1988)645.
[30] J.S. Richardson and D.C. Richardson, Science 240(1988)1648.
[31] L.G. Presta and G.D. Rose, Science 240(1988)1632.
[32] P.Y. Chou and G.D. Fasman, Ann. Rev. Biochem. 47(1978)251.
[33] P.Y. Chou and G.D. Fasman, Adv. Enzymol. 47(1978)45.
[34] R.W. Williams, A. Chang, D. Juretić and S. Loughran, Biochim. Biophys. Acta 916(1987)200.
[35] J.L. Fauchere and V. Pliska, Eur. J. Med. Chem. Chim. Ther. 18(1983)369.
[36] F.M. Richards and T. Richmond, in: *Molecular Interactions and Activity in Proteins*, ed. P. Porter and D.W. Fitzimons (Excerpta Medica, Amsterdam, 1978), p. 23.
[37] A.M. Lesk and C. Chothia, Biophys. J. 32(1980)35.
[38] B. Gruenewald, C.U. Nicola, A. Lustig and G. Schwartz, Biophys. Chem. 9(1979)137.
[39] R.L. Baldwin, Trends Biochem. Sci. 11(1986)6.