Geometrical and Sequence Characteristics of α-Helices in Globular Proteins

Sandeep Kumar and Manju Bansal
Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India

ABSTRACT Understanding the sequence-structure relationships in globular proteins is important for reliable protein structure prediction and de novo design. Using a database of 1131 α-helices with nonidentical sequences from 205 nonhomologous globular protein chains, we have analyzed structural and sequence characteristics of α-helices. We find that geometries of more than 99% of all the α-helices can be simply characterised as being linear, curved, or kinked. Only a small number of α-helices ($\sim$4%) show sharp localized bends in their middle regions, and thus are classified as kinked. Approximately three-fourths ($\sim$73%) of the α-helices in globular proteins show varying degrees of smooth curvature, with a mean radius of curvature of 65 $\pm$ 33 Å; longer helices are less curved. Computation of helix accessibility to the solvent indicates that nearly two-thirds of the helices ($\sim$66%) are largely buried in the protein core, and the length and geometry of the helices are not correlated with their location in the protein globule. However, the amino acid compositions and propensities of individual amino acids to occur in α-helices vary with their location in the protein globule, their geometries, and their lengths. In particular, Gln, Glu, Lys, and Arg are found more often in helices near the surface of globular proteins. Interestingly, kinks often seem to occur in regions where amino acids with low helix propensities (e.g., β-branched and aromatic residues) cluster together, in addition to those associated with the occurrence of proline residues. Hence the propensities of individual amino acids to occur in a given secondary structure depend not only on conformation but also on its length, geometry, and location in the protein globule.

INTRODUCTION

Using basic stereochemical principles and recognizing the planarity of the peptide bond, Pauling et al. (1951) deduced the α-helical structure as one of the two possible hydrogen-bonded conformations for the polypeptide chains in proteins. Subsequent solution of the three-dimensional structure for myoglobin by x-ray crystallography (Kendrew et al., 1958) proved the validity of this model. Today it is well known that the α-helix is one of the most common secondary structural motifs in proteins. Because of the regular occurrence of (i, $i - 4$) hydrogen bonds in the main chain, the α-helices in proteins are generally quite uniform, with their helical parameters unit twist (t) and unit height (h) lying within well-defined ranges. However, α-helices have been observed to be distorted for a variety of reasons, for example, occurrence of proline residues (Barlow and Thornton, 1988), solvent-induced distortions (Blundell et al., 1983), and peptide bond distortions (Chakarabarti et al., 1986).

A systematic analysis of helix geometries found in globular proteins was first reported by Barlow and Thornton (1988). They found that most of the α-helices in globular proteins were smoothly curved. We had previously analyzed sequences and geometries of 64 long α-helices with lengths greater than 25 residues taken from 45 globular proteins (Kumar and Bansal, 1996). It was observed that long α-helices have unique amino acid composition characteristics that distinguish them from the short α-helices, taken from the data set of Richardson and Richardson (1988). The amino acid distribution in long α-helices could be correlated with their overall geometry and, in the case of curved helices, with the radius of curvature. We have now extended these studies to study variation in amino acid distribution in α-helices with respect to several parameters, with a data set that includes helices of all lengths in the June 1995 Protein Data Bank (PDB) (Bernstein et al., 1977). Using HELANAL (Kumar and Bansal, 1996) to characterize α-helix geometry in a database of 1131 α-helices of nonidentical sequences, from 205 nonhomologous globular protein chains, we found that the following rules are generally observed for all helices: 1) Almost all α-helices can be unambiguously characterized as being linear, curved, or kinked. 2) The majority of α-helices are smoothly curved and show varying degrees of curvature, with longer helices preferring a sharp kink to high smooth curvature. 3) Amino acid distribution and propensities of individual amino acids to occur in an α-helix depend on its location in the protein globule, on its length and geometry, and in the case where the helix is smoothly curved, on its radius of curvature.

Received for publication 1 December 1997 and in final form 8 June 1998.

Address reprint requests to Dr. Manju Bansal, Molecular Biophysics Unit, Indian Institute of Science, Bangalore-560 012, India. Tel.: 91-80-3092534; Fax: 91-80-3348535 or 91-80-3341683; E-mail: mb@mbu.iisc.ernet.in.

Dr. Kumar's present address is Bldg. 469, Rm. 147, NCI-FCRDC, P.O. Box B, Frederick, MD 21702.

0006-3495/98/10/1935/10 $2.00

MATERIALS AND METHODS

Composition of the database

We have used the June 1995 list of nonhomologous (sequence identity $\leq$ 25%) protein chains compiled by Hobohm and Sander (Hobohm et al.,

1992; Hobohm and Sander, 1994) to select a subset database of 205 nonhomologous globular protein chains whose three-dimensional structures have been solved by x-ray crystallography to a resolution of 2.5 Å or better. One thousand one hundred thirty-one (1131) α-helices of nonidentical sequences and nine or more amino acid residues found in these protein chains constitute the database for the studies presented here. We have only considered helices consisting of nine or more residues because these helices contain at least one residue in which both NH and C═O are hydrogen bonded. Besides this, because HELANAL characterizes the overall geometry of an α-helix from its local structural features (Kumar and Bansal, 1996, and briefly described below), the use of nine residues as the minimum helix length ensures that we have sufficient data points to deduce its overall geometry.

Helix boundaries

Initially, protein chain segments of length greater than eight residues and defined as helices by the Dictionary of Protein Secondary Structure (DSSP) (Kabsch and Sander, 1983) were taken for each of the 205 protein chains. Boundaries of these helices were checked and, if necessary, reassigned according to the following additional criteria:

1. The distance $|O_i—N_{i+4}| \leq 3.5$ Å at helix termini.
2. Angles between successive local helix axes at the helix termini region are less than 20° (Kumar and Bansal, 1996).

Helix geometry and structural parameters of α-helices

We have used an in-house developed computer program called HELANAL (Kumar and Bansal, 1996) to deduce the overall geometry of an α-helix from its local structural features, namely, local bending angles and local helix axes computed for every four consecutive C^α atoms. This window of four C^α atoms was slid over the length of the helix in steps of one C^α atom.

The geometry of the helix is assigned from the values of four parameters, namely, the maximum local bending angle in the helix, defining the largest distortion in the helix; the root mean square deviations for the line (rmsL) and the circle (rmsC) fitted to the local helix origins that trace out the three-dimensional path of helix axis; and r^2, the square of linear correlation coefficient for the line fit. If the maximum bending angle for the helix is $\geq 20°$, the helix is classified as kinked. If the maximum bending angle is $<20°$, then we look at the other three parameters. If both rmsC and rmsL are >1 Å, we do not assign any geometry type to the helix because this implies large and random distortions in the helix. If either one or both of the parameters rmsC and rmsL are less than 1 Å, then we compute the ratio, rmsL/rmsC. A value greater than 1 for the ratio indicates that the circle fits the local helix origins better than the line, and we assign helix geometry as curved. If the ratio is ≤ 0.7, it is implied that the line fits the local helix origins better than the circle. We validate the helix geometry to be genuinely linear if $r^2 \geq 0.8$. However, if the validation fails, that is, $r^2 < 0.8$, we do not assign any geometrical type to the helix. A value of 0.7–1.0 for the ratio defines a zone where line and the circle fit the local helix origins almost equally well. We have only a few α-helices in this zone. In this zone, if $r^2 \geq 0.8$, i.e., the local helix origins significantly correlate with line, the helix geometry is again classified as linear. But, if $r^2 \leq 0.5$, i.e., correlation of the local helix origins with the line is insignificant and could be expected by random chance, the helix geometry is classified as curved. For values of r^2 in the range between 0.5 and 0.8, the helix geometry type is ambiguous. Hence no geometrical type is assigned to the helix.

Helical parameters of α-helices, viz., unit height and unit twist, are also computed by HELANAL for every four C^α atoms. For each of the 205 nonhomologous protein chains in our database, there exists a corresponding DSSP file containing secondary structure assignments. Among several other parameters, these DSSP files contain values of virtual torsion angles and ϕ, ψ for each residue in the protein chain. We have used the values corresponding to the residues in 1131 α-helices for structural analysis of the helices.

Estimation of kink angle in the case of kinked helices

As mentioned above, the geometry of an α-helix was classified as kinked if the maximum local bending angle, computed by HELANAL, exceeds 20°. The kinks are localized in a small region in the middle of the kinked helices. The region of kink is characterized by the large values of the local bending angles ($>20°$) and hydrogen bond distances (>3.5 Å) at several consecutive residues, and is identified from plots of the local bending angle at the residue number i and the hydrogen bond distance (O_{i-2}–N_{i+2}) versus the residue number i in the helix (Kumar and Bansal, 1996). Each kinked α-helix was then divided into two segments, consisting of fragments before and after the region of kink, and the kink angle was calculated as the angle between the least-squares lines fitted to the local helix origins in these two segments after the kink region was excluded.

Computation of helix accessibility to the solvent

Solvent accessibility (Vorobjev and Hermans, 1997) of an α-helix is calculated as the ratio of accessible surface area of the helix within the protein to its accessible surface area upon its isolation from the protein. The accessible surface area of the helix is the sum of the accessible surface areas of all of the atoms in the helix. The accessible surface area for each atom in the helix is calculated by the method of Lee and Richard (1971), using a probe radius of 1.4 Å. For each helix, the whole protein in the crystal asymmetrical unit including all subunits has been considered for computation of its helix accessibility. We have used helix accessibility as a measure of its location in the protein globule.

Statistical methods

The database of 1131 α-helices is subdivided into various categories according to helix location in tertiary structure (i.e., their solvent accessibility) and geometry (viz., linear, curved, and kinked). Furthermore, the curved helices are classified according to the radius of curvature ranges. Helices in the database are also subdivided into two broad length groups, viz., short helices (9–21 residues) and long helices (>21 residues), i.e., those that are significantly longer than one full repeat (18 residues) for α-helices. In each of these classes at least five amino acid residues of each type are present in the database, and propensities for individual amino acid are computed according to the formula given by Williams et al. (1987). The χ^2 test has been used to determine whether the differences in the amino acid distributions of α-helices in any two classes are significant (Medhi, 1992). To measure the differences in the sequence compositions of helices in any two classes, Euclidean and Hamming distances are also computed in 20-dimensional amino acid composition space (Chou and Zhang, 1995). Because these two parameters showed identical trends in all calculations, only the values of Hamming distances have been discussed in detail. Change in proportion of an amino acid in a given helix class is considered to be significant if it is greater than 2σ, where σ is the estimated standard deviation (Medhi, 1992).

RESULTS AND DISCUSSION

The commonly assigned structural parameters, unit height and unit twist for the α-helices, do not show large variations, as also noted in our earlier analysis of long α-helices (Kumar and Bansal, 1996). Mean unit twist for the 1131 α-helices ranges between 94° and 103°, with an average value of $(99 \pm 1)°$. The mean unit height for these helices ranges between 1.3 Å and 1.7 Å, with an average value of (1.51 ± 0.03) Å. The average values for the backbone torsion angles (ϕ, ψ) in the middle regions of 1131 α-helices, i.e., the residues other than those in the first and last

turns in the helices, are (−63 ± 7)° and (−43 ± 8)°, respectively, whereas the average value for the virtual torsion angles computed for every four successive C^{α} atoms in the 1131 helices is (50 ± 6)°. These observations indicate that structural parameters of α-helices are quite uniform, and deviations from regularity (linearity) in α-helices cannot be estimated from these parameters alone.

Geometrical and length distribution of α-helices in globular proteins

Of 1131 α-helices in our database, with helix lengths ranging between 9 and 37 residues, 1122 (99.2%) can be characterized as being linear, curved, or kinked. In only nine cases (0.8%) the helices are either irregular or their geometries cannot be assigned unambiguously. A majority (821 of 1131, 72.6%) of the α-helices in globular proteins are smoothly curved, as also found in the earlier analyses by Barlow and Thornton (1988) and Kumar and Bansal (1996). The proportions of linear and kinked helices are 22.6% (256 of 1131) and 4% (45 of 1131), respectively. Fig. 1 shows the distribution of α-helices with respect to helix length (number of residues in an α-helix) in each of the three geometric classes, as well as in the whole database of 1131 α-helices. The number of crystallographically observed helices decreases rapidly with the increase in helix length. This observation is in conformity with the observations of Barlow and Thornton (1988), Blundell and Zhu (1995), and Zhu and Blundell (1996). However, our database does not show maxima corresponding to helix lengths of 7, 11, 15.5, and 22 residues, as observed by Srinivasan (1976). The mean helix length in our database of 1131 α-helices is (14 ± 5) residues, whereas the mean helix lengths for linear, curved, and kinked helices are (12 ± 3), (14 ± 5), and (20 ± 6), respectively, indicating that kinked helices are usually longer than the average.

The database of 1131 α-helices is also subdivided into two length classes, namely short α-helices (containing 9–21 residues) accounting for 1046 (92.5%) of 1131 α-helices and 85 (7.5%) long α-helices (> 21 residues). The 1046 short α-helices include 758 (72.5%) curved, 251 (24%) linear, and 28 (2.7%) kinked α-helices, whereas the 85 long α-helices include 63 (74.1%) curved, 5 (5.9%) linear, and 17 (20%) kinked α helices. Thus, among long α-helices, the proportion of kinked helices increases at the expense of linear helices.

Longer helices are generally less curved than shorter ones

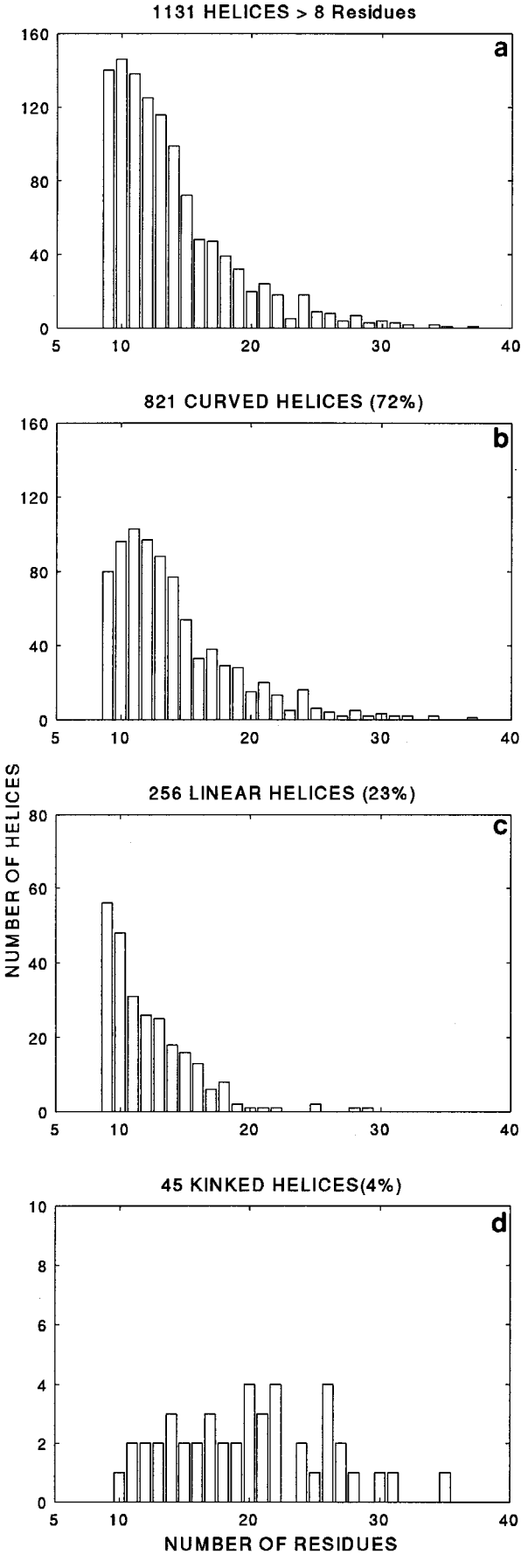
The mean radius of curvature for 821 curved helices is (65 ± 33) Å (range 8–373 Å). Fig. 2 *a* shows the distribution of helices with respect to the radius of curvature. The average radii of curvature in the two length classes of curved helices, viz., short α-helices with 9–21 residues and long helices with >21 residues, are (63 ± 34) Å and (96 ± 32) Å, respectively. Fig. 2 *b* shows a plot of mean helix length for helices in different curvature ranges versus the radius of curvature. It can be seen that the mean helix length increases linearly with the radius of curvature (square of linear correlation coefficient, $r^2 = 0.96$). It is interesting to find that the minimum radius of curvature for each helix length also increases linearly with the helix length (square of linear correlation coefficient, $r^2 = 0.93$). These observations indicate that longer helices tend to be less curved than shorter ones. Barlow and Thornton (1988) had also noted that long helices in coiled coils are less curved than the short helices in globular proteins. Thus it appears that a high degree of curvature cannot be sustained or propagated over a long polypeptide helix.

Do α-helices with different lengths or geometries occur in different regions of the protein globule?

Most of the α-helices in globular proteins are quite short; the mean helix length in the present data set is 14 residues. It is generally believed that the length of α-helices is limited by the size of the protein (Creighton, 1993). Thus one might expect a correlation between the helix length and the size of the protein as defined by the number of amino acid residues in the protein. In our database, we do not observe any such correlation. For example, a small 56-residue polypeptide chain of ColE1 Rop contains two long α-helices, forming an antiparallel coiled coil (Banner et al., 1987).

It may also be expected that part or the whole of longer helices may be exposed to solvent, whereas shorter helices can be readily buried in the protein core, i.e., long and short helices may occur in different regions of the protein globule. We have tested this hypothesis by computing solvent accessibility for each of the 1131 α-helices. Accessibility of an α-helix defines its location in the protein tertiary structure (Blundell and Zhu, 1995; Zhu and Blundell, 1996) and in quaternary structure, in the case of proteins containing more than one subunit. Mean solvent accessibility for the 1131 α-helices is (32 ± 15)%. The majority of the helices (751 of 1131, 66.4%) are "largely buried" in the protein core, i.e., have less than 40% solvent accessibility. 356 (31.5%) helices are "partially buried" (40–60% solvent accessibility), and only 24 (2.1%) are "largely exposed" (> 60% solvent accessibility) to the solvent. The mean helix lengths for largely buried, partly buried, and largely exposed helices are (14 ± 4), (14 ± 5), and (16 ± 6), respectively. Fig. 3 shows a plot of solvent accessibilities of 1131 α-helices versus their lengths. There is no significant correlation between length and solvent accessibility, indicating that an α-helix of any length can occur anywhere in the protein globule, and there is no segregation between long and short α-helices into different regions of the protein globule.

A kink in an α-helix, containing a proline residue after the first four positions, facilitates compensatory hydrogen bond formation between solvent and residues $i - 3$ and $i -$



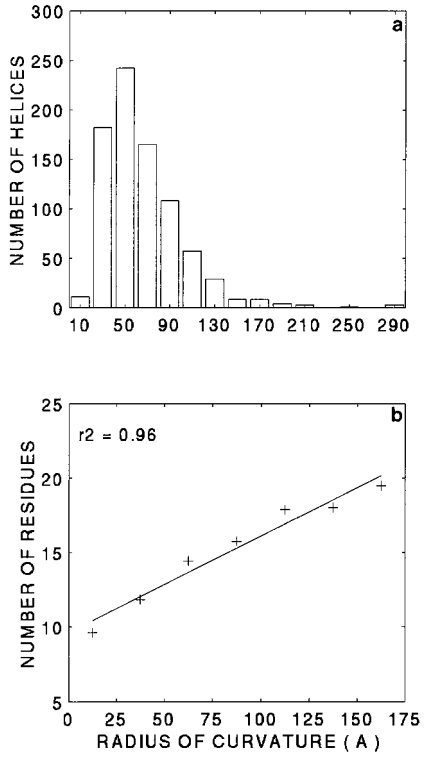


FIGURE 2 (*a*) Bar diagram showing distribution of 821 curved α-helices with respect to their radius of curvature. The *x* axis denotes the radius of curvature, and the *y* axis denotes the number of helices. The mean radius of curvature for all 821 curved α-helices is (65 $\pm$ 33) Å. (*b*) Mean number of residues for α-helices in each of the seven classes, viz. radius of curvature ranges of 0–25 Å, 25–50 Å, 50–75 Å, 75–100 Å, 100–125 Å, 125–150 Å, and 150–175 Å. The mean helix length (number of residues) (*y* axis) is plotted against the midvalues of radii of curvature ranges in the above seven ranges (*x* axis).

4 (relative to a proline residue at the *i*th position) (Woolfson and Williams, 1990). Besides, water is known to induce bends and kinks in α-helices (Blundell et al., 1983). According to these concepts, one would expect helix geometry to be correlated with its solvent accessibility and kinked helices to occur predominantly on the protein surface. Fig. 4 shows distributions of linear, curved, and kinked helices with helix accessibility, and it is clear that there is no significant correlation between helix geometry and solvent accessibility. In addition, it is found that solvent accessibilities of Pro and Gly residues at kink regions are similar to the mean residue accessibility of that kinked helix, indicating that there is no noticeable preference for Pro and Gly at kinks to be exposed on the protein surface.

FIGURE 1 Histograms showing the length distribution of all α-helices (>8 residues in length) in globular proteins. The *x* axis denotes the number of residues in the α-helix (helix length). The *y* axis denotes the number of helices. (*a*) All 1131 α-helices. (*b*) 821 curved α-helices. (*c*) 256 linear α-helices. (*d*) 45 kinked α-helices. In the case of nine α-helices, no helix geometry is assigned.

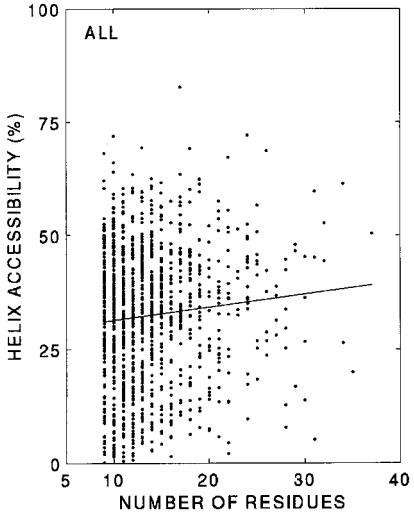


FIGURE 3 Plot showing solvent accessibility of all 1131 α-helices (helix accessibility) with respect to number of residues in the helices. The best fit line indicates that there is no significant correlation between helix accessibility and its length.

Factors affecting the helix-forming propensities of amino acids

Helix-forming propensities of various amino acids in barnase have been analyzed by Fersht and co-workers (Serrano and Fersht, 1989; Serrano et al., 1992) with site-directed mutagenesis experiments, and they concluded that assigning a unique and generally applicable helix-forming propensity to an amino acid may not be valid. Chakrabartty et al. (1991) also found a strong positional dependence for Ala→Gly substitutions in alanine-based peptides. The amino acid distributions in the middle of helices are different from those at helix termini (Argos and Palau, 1982; Richardson and Richardson, 1988; Kumar and Bansal, 1998). Helix propensities of amino acids vary with position of the helix in protein tertiary structure (Blundell and Zhu, 1995; Zhu and Blundell, 1996). In this report, we present further evidence for the variations observed in amino acid propensities in α-helices with respect to the length and geometries of the helices, as well as their location in the protein globule.

Amino acid compositions of α-helices vary with their location in the protein globule

In general, we find that all 20 amino acids are less accessible to the solvent when present in α-helices as compared to their average accessibilities in the whole protein globule. This finding is consistent with the observation in the previous section that most of the α-helices are buried in the protein core. The question arises whether the amino distributions in the α-helices vary with their location in the protein globule. Table 1 shows χ^2 comparison and Hamming distances among amino acid distributions in various helix accessibility classes. It can be seen that differences among the amino acid compositions of helices in different accessibility classes are highly significant. Large and consistently increasing values of Hamming distances reflect that amino acid compositions of these classes are well separated in the 20-dimensional amino acid composition space. Propensities of several amino acids vary with helix accessibility. Fig. 5 shows the variations in propensities of Ala, Leu, Gln, Glu, Lys, and Arg as a function of helix accessibility. Propensities of aliphatic amino acids Ala and Leu decrease linearly with increase in helix accessibility. Proportions of Ala and Leu decrease from 15.4% and 13.0%, respectively, in helices with less than 10% helix accessibility, to 11.6% and 10.4%, respectively, in helices with 50–60% accessibility. The changes in proportions of Ala and Leu are significant at the 95% level of confidence. These observations indicate that Ala and Leu occur more frequently in a buried helix than in a helix near or on the protein surface, as expected on the basis of their hydropho-

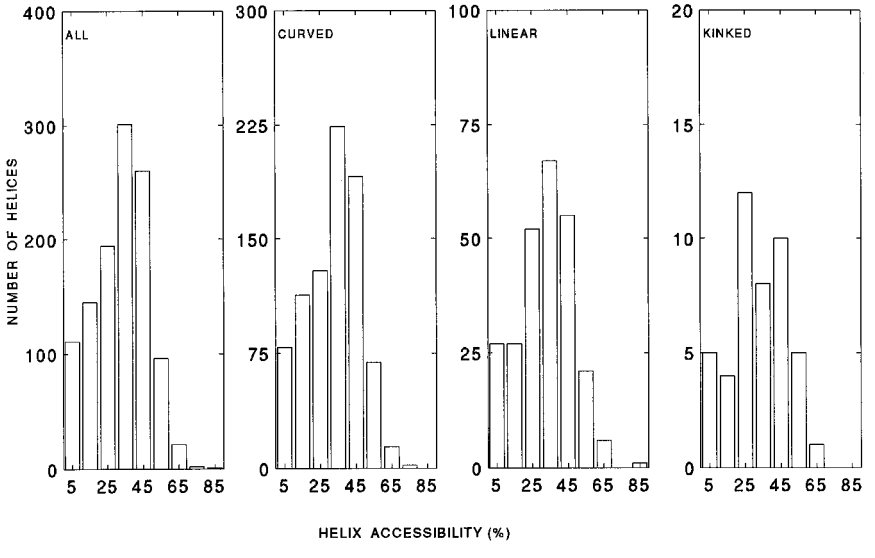


FIGURE 4 Bar diagrams showing the distributions of α-helices with respect to helix accessibility in various helix geometries. The *x* axis denotes helix accessibility, and the *y* axis denotes the number of helices in each case.

TABLE 1 Values of χ^2 comparison and Hamming distances for amino acid distributions in various helix accessibility classes

Helix accessibility (%)	0–10 (111)	10–20 (145)	20–30 (194)	30–40 (301)	40–50 (260)	50–60 (96)	60–70 (21)
0–10		135.0	399.2	1063.1	1516.7	1095.5	305.7
(111)		15.5	26.1	32.5	39.1	52.0	58.5
10–20			98.8	311.3	515.5	484.2	162.9
(145)			14.0	21.2	27.4	40.7	47.6
20–30				98.3	185.3	227.0	80.0
(194)				12.8	17.2	30.1	37.0
30–40					80.9	147.6	57.4
(301)					11.8	25.1	33.8
40–50						70.2	38.7
(260)						17.9	28.4
50–60							26.1
(96)							28.1

The number of helices in each accessibility class is given in parenthesis. In each cell the first number represents the value of χ^2, and the second number represents the Hamming distance (D. U.) between two helix classes whose amino acid distribution are being compared. D. U. stands for distance units. For a 19-parameter system such as amino acid distribution in each helix class, the null hypothesis is rejected at the 95% level of confidence ($p < 0.05$) if $\chi^2 > 30.14$.

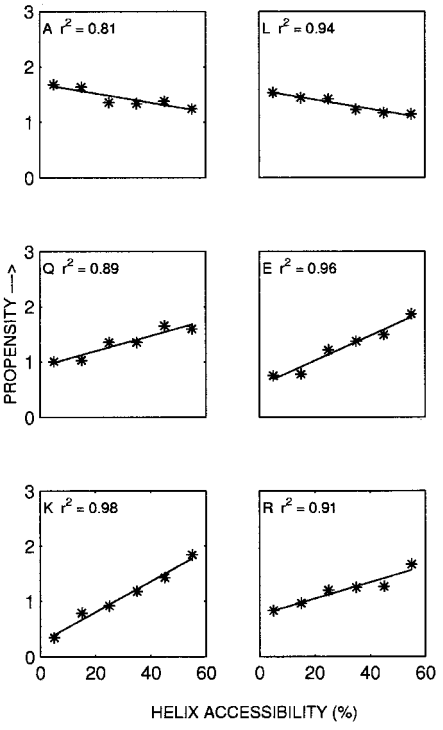


FIGURE 5 Helix propensities of amino acid residues Ala, Leu, Glu, Gln, Lys, and Arg, which show variation with the location of helix in the protein globule. In each diagram, helix accessibility is along the *x* axis, and the propensity of the amino acid is along the *y* axis. The lines fitted by the least-squares method are also shown. The value of the square of linear correlation coefficient (r^2) and the single-letter amino acid code are indicated in the upper left corner of each box.

bic character. Propensities of the charged amino acids Glu, Lys, and Arg and the polar neutral amino acid Gln increase linearly with helix accessibility, implying that they are more frequent in helices that are near or on the protein surface. Changes in proportion all four of these amino acids, Glu (4.6% in helices with less than 10% accessibility to 12.5% in helices with 50–60% accessibility, $\Delta = +7.9\%$), Gln (3.5% in helices with less than 10% accessibility to 5.9% in helices with 50–60% accessibility, $\Delta = +2.4\%$), Lys (2.0% in helices with less than 10% accessibility to 11.6% in helices with 50–60% accessibility, $\Delta = +9.6\%$), and Arg (3.8% in helices with less than 10% accessibility to 8.2% in helices with 50–60% accessibility, $\Delta = +4.4\%$), are large and highly significant, even at the 99.99% level of confidence.

Does amino acid composition of an α-helix encode its geometry?

As shown above, α-helices in proteins do not always follow a straight or linear path. Rather, they bend and kink to optimize their interactions with rest of the protein core and/or solvent. These helix geometries are not a consequence of thermal fluctuations and inaccuracies in the protein crystal structure data, because we find that the mean B-factor calculated for the C^α atoms of the helices remains nearly constant for all helices and does not depend on their geometries or their lengths. Thus helix geometries found in globular proteins are inherent properties of α-helices and may be encoded in their amino acid distributions.

Characteristics of kinked α-helices

Our database of 1131 α-helices contains 45 kinked α-helices, with a total of 46 kinks. On an average, two main-chain hydrogen bonds are broken in the regions of kink in these helices (defined in Materials and Methods). The average value for the kink angle in the 45 kinked helices is (33 ± 11)°, and that of the largest local bending angle in their regions of kink is (34 ± 11)°. It is interesting to note that mean values of torsion angles (ϕ, ψ) in kinked helices are (−65 ± 11)° and (−41 ± 10)°, respectively, which are very similar to those in linear (ϕ = (−63 ± 8)°, ψ = (−42 ± 10)°) and curved (ϕ = (−63 ± 6)°, ψ = (−42 ± 7)°) helices. However, the standard deviation values for kinked helices are slightly higher because of larger deviations in the regions of kink. In 25 of 45 kinked helices, the kink is caused by the presence of the imino acid residue proline, and the average value of the kink angle in these helices is (32 ± 8)° (range 17–58°). Nine helices contain Gly residue in the region of the kink; the average kink angle in these helices is (37 ± 14)° (range 17–60°). Both Pro and Gly residues are present at four of the kinked regions, and one kink contains two successive prolines. Thus 29 kinks in our database are accounted for by Gly and Pro.

The remaining 17 kinks are caused by residues other than Pro and Gly in 16 α-helices. The kink angles for these helices vary in the range of 8–66°; the average is (33 ± 15)°. In these 17 kinked regions, β-branched residues (Val, Ile, or Thr) were found to occur eight times, aromatic amino

acids (Tyr, Trp, or Phe) five times, and His once. The regions of kink in three helices contain both β-branched and aromatic residues.

Table 2 shows χ^2 comparisons and Hamming distances for amino acid distributions in various helix classes. The differences between the amino acid distributions of linear and curved helices are not significant, whereas those between linear and kinked helices as well as between curved and kinked helices are significant, even at the 99.99% level of confidence ($p < 0.001$, χ^2 values > 43.82). These results indicate that amino acid distribution in 45 kinked helices is significantly different from the amino acid distributions in linear and curved helices.

Propensity values for the amino acids that show large differences among the various geometrical classes of helices are shown in Fig. 6. In addition to proline, several other amino acids also show appreciable changes in their propensities to occur in kinked helices. Propensities of Pro, Ile, Trp, Tyr, and Gln show an increase for kinked helices as compared to linear and curved helices, whereas propensities of Ala, His, and Glu show an appreciable decrease. Amino acid distribution in the regions of kink in the 45 kinked α-helices can be useful in identifying kink-causing residues. Proportions of amino acid residues Ala (13.0% in all helices, 10.7% in kinked helices, 8.7% in the region of kink) and Glu (8.2% in all helices, 6.2% in kinked helices, 3.8% in the region of kink) decrease, whereas the proportion of Pro (1.6% in all helices, 3.7% in kinked helices, 14.1% in the region of kink) increases by large amounts. Changes in proportions of Pro and Glu in the 45 kinked helices as well as in their kinked regions with respect to their proportions in all of the 1131 α-helices are significant at the 95% level of confidence. Pro is well known to cause kinks in α-helical structure (Woolfson and Williams, 1990; Von Heijne, 1991; Sankararamakrishnan and Vishveshwara, 1992), but a decrease in the proportion of Glu in kinked helices is interesting and suggests that Glu may have a tendency to resist the occurrence of sharp bends/kinks in α-helices. Given the conformational flexibility of Gly, it may be expected to have a higher propensity to occur in kinked regions of helices. In our studies, the propensity of Gly to occur in kinked α-helices remains more or less equal to that in curved or linear helices, but the proportion of Gly increases from 3.3% in the complete kinked helices to 6.1% in the regions of kink, indicating a tendency for Gly to be sequestered in the kinked regions, when it does occur in kinked helices. However, the (ϕ, ψ) values for these Gly residues are similar to those for other amino acids in the kink regions.

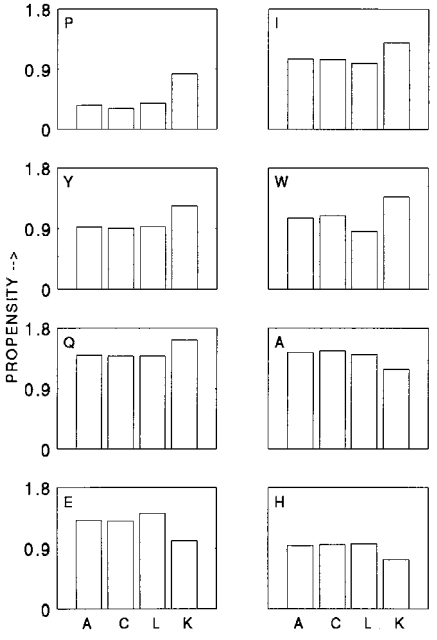


FIGURE 6 Bar diagrams showing variation in propensities with respect to helix geometry for those amino acids that show a large increase or decrease in kinked helices. Amino acids are denoted by a single-letter code in the upper left corner of each box. Bars along the horizontal axis correspond to A: all 1131 α-helices; C: 821 curved α-helices; L: 256 linear α-helices; and K: 45 kinked helices, respectively.

TABLE 2 Values of χ^2 comparison and Hamming distances in the amino acid compositions of various classes of α-helices, characterized by their geometry and length

Helix classes	χ^2 value	Hamming distance (D. U.)
Geometrical classes in α-helices		
Curved (821) vs. linear (256)	27.01	7.50
Curved (821) vs. kinked (45)	58.85	16.70
Linear (256) vs. kinked (45)	48.71	16.00
821 curved helices classified according to radius of curvature		
0 Å–25 Å (36) vs. 25 Å–50 Å (288)	188.12	19.40
25 Å–50 Å (288) vs. 50 Å–75 Å (245)	43.98	8.90
50 Å–75 Å (245) vs. 75 Å–100 Å (139)	37.42	9.70
75 Å–100 Å (139) vs. 100 Å–125 Å (67)	16.84	9.40
100 Å–125 Å (67) vs. >125 Å (46)	41.47	13.40
Helices classified according to helix length (only middle regions of helices were considered)		
Short (1046) vs. long (85)	44.98	14.80

The number of helices in each class is given in parenthesis, along with the helix class name. For a 19-parameter system such as amino acid distribution, the null hypothesis is rejected at the 95% ($p < 0.05$) level of confidence, if $\chi^2 > 30.14$. D. U. stands for distance units. The middle region of an α-helix corresponds to the residues remaining after the first and last four residues are removed.

In the case of smoothly curved α-helices, amino acid composition varies with the radius of curvature

Significant differences are observed in amino acid distributions for helices in radii of curvature ranges 0–25 Å and 25–50 Å, 25–50 Å and 50–75 Å, 50–75 Å and 75–100 Å, 100–125 Å, and >125 Å, as seen from Table 2. Differences in amino acid distributions are insignificant between the classes 75–100 Å and 100–125 Å. Variation in propensities

of amino acids that show appreciable differences in five radius of curvature classes, namely, 0–25 Å, 25–50 Å, 50–75 Å, 75–100 Å, and 100–125 Å are shown in Fig. 7. The propensity to occur in the complete dataset of 821 curved α-helices is also plotted alongside in each figure. The helix-forming propensities of Ala, Asn, and Glu remain nearly constant for helices in all radius of curvature classes,

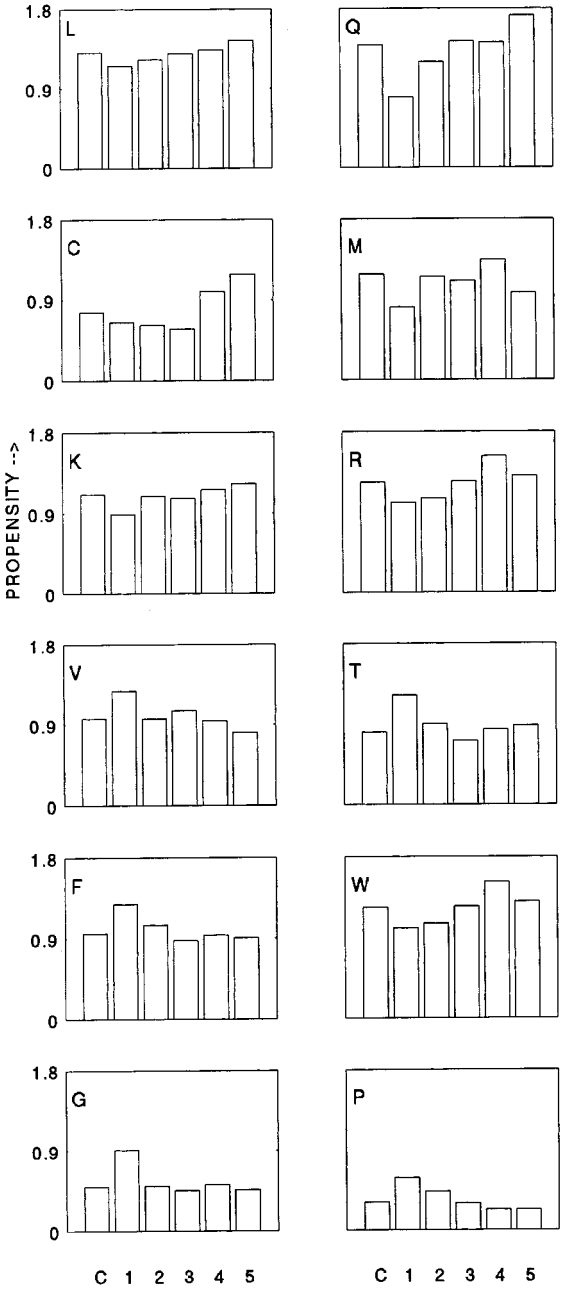


FIGURE 7 Bar diagrams showing the variation in propensities of amino acids in curved α-helices with respect to radius of curvature. Only amino acids that show interesting features have been shown. They are identified by a single-letter code in the upper left corner of each box. Bars along the horizontal axis correspond to (C) all 821 curved α-helices, followed by ranges of radius of curvature between (1) 0 Å and 25 Å; (2) 25 Å and 50 Å; (3) 50 Å and 75 Å; (4) 75 Å and 100 Å; and (5) 100 Å and 125 Å, respectively.

whereas Met and Arg show an increase followed by a sharp decrease for large a radius of curvature class (100–125 Å). The amino acid residues Leu, Gln, Lys, and Cys generally show increasing propensities for helices with larger radii of curvature, whereas Val, Thr, Pro, Phe, and Trp show an opposite trend. It is interesting to note that the amino acid residues Pro, Val, Phe, Thr, and Gly, which have all been found to be favored at kink regions, also have considerably higher propensities to occur in the highly curved helices (radius of curvature $\leq$ 25 Å) than in the helices with larger radii of curvature, as seen from Fig. 7. However, Ile, Tyr, and Gln, which are also often found at kinks, have very low propensities for such highly curved helices. In particular, proportions of Thr (4.6% in all 821 curved helices, 6.9% in the 36 highly curved helices, $\Delta = +2.3\%$) and Gly (3.9% in all 821 curved helices, 7.2% in the 36 highly curved helices, $\Delta = +3.3\%$) increase significantly in the case of highly curved helices. The proportion of Thr also decreases significantly in the case of the less curved helices with radius of curvature $>$ 125 Å (4.6% in all 821 curved helices, 2.7% in the case of the 46 less curved helices, $\Delta = -1.9\%$). Given that Gly has greater conformational flexibility than other amino acids, it is possible for Gly residues to cause local bends (less severe than kinks) in α-helices, thus leading to their being highly curved. It is interesting to note that Pro is also found to occur in highly curved helices, whereas it is rare in helices with large radii of curvature. The above observations clearly demonstrate curvature-dependent sequence variations in curved α-helices.

Amino acid composition of long α-helices differs from that of short α-helices

α-Helices in globular proteins show length-dependent sequence variations, with amino acids having longer side chains and/or a greater number of side-chain functional groups occurring more frequently in longer α-helices (Kumar and Bansal, 1996). Whereas N-cap and C-cap regions (Presta and Rose, 1988) in α-helices are expected to be similar for helices of all lengths, length-dependent variations may occur in the middle regions of α-helices. Table 2 shows χ^2 comparisons as well as Hamming distances for the middle regions of helices classified into two length classes, viz., short helices (1046 helices with 9–21 residues) and long helices (85 helices with $>$21 residues). It can be seen that the differences in the distributions of amino acids in the middle regions of helices in these two length classes are significant even at the 99.99% ($\chi^2 > 43.82$) level of confidence. Proportions of amino acid residues Leu ($\Delta = +1.3\%$), Gln ($\Delta = +1.1\%$), and Arg ($\Delta = +1.4\%$) increase substantially, whereas those of Ala ($\Delta = -2.8\%$), Val ($\Delta = -2.0\%$), and Lys ($\Delta = -0.9\%$) show an appreciable decrease for longer helices. Changes in the proportions of Ala, Arg, and Val are significant at the 95% level of confidence. Propensities of individual amino acids to occur in the middle regions of short and long α-helices are given in Table 3.

TABLE 3 Statistical propensities for individual amino acids to occur in middle regions (as defined in Table 2) of all 1131 α-helices, along with the propensities in the short and long classes separately

Amino acid	All helices (9–37 residues)	Short α-helices (9–21 residues)	Long α-helices (>21 residues)
Ala (A)	1.56 ± 0.05	1.63 ± 0.05	1.32 ± 0.10
Cys (C)	0.87 ± 0.09	0.76 ± 0.10	1.24 ± 0.24
Asp (D)	0.74 ± 0.04	0.72 ± 0.05	0.81 ± 0.09
Glu (E)	1.16 ± 0.05	1.18 ± 0.06	1.08 ± 0.11
Phe (F)	0.94 ± 0.06	0.98 ± 0.07	0.82 ± 0.11
Gly (G)	0.43 ± 0.03	0.41 ± 0.03	0.49 ± 0.06
His (H)	0.76 ± 0.07	0.74 ± 0.08	0.81 ± 0.15
Ile (I)	1.16 ± 0.05	1.17 ± 0.06	1.14 ± 0.12
Lys (K)	1.12 ± 0.05	1.15 ± 0.06	1.01 ± 0.10
Leu (L)	1.35 ± 0.05	1.32 ± 0.05	1.47 ± 0.11
Met (M)	1.41 ± 0.09	1.37 ± 0.10	1.51 ± 0.21
Asn (N)	0.78 ± 0.05	0.80 ± 0.06	0.72 ± 0.10
Pro (P)	0.10 ± 0.02	0.08 ± 0.02	0.13 ± 0.05
Gln (Q)	1.40 ± 0.07	1.33 ± 0.08	1.64 ± 0.17
Arg (R)	1.33 ± 0.06	1.26 ± 0.07	1.57 ± 0.15
Ser (S)	0.72 ± 0.04	0.71 ± 0.05	0.78 ± 0.09
Thr (T)	0.84 ± 0.05	0.84 ± 0.05	0.85 ± 0.10
Val (V)	1.05 ± 0.05	1.11 ± 0.05	0.82 ± 0.09
Trp (W)	0.98 ± 0.10	0.94 ± 0.11	1.10 ± 0.23
Tyr (Y)	0.94 ± 0.06	0.92 ± 0.07	1.03 ± 0.13

Amino acid propensities and their standard deviations are calculated according to formulae given by Williams et al. (1987). The 1131 α-helices contain a total of 6629 amino acid residues, of which 5125 residues constitute 1046 short helices and 1504 residues are present in the 85 long helices.

These results, in general, confirm the conclusions from our earlier analysis (Kumar and Bansal, 1996), which used a smaller database of short helices, i.e. among similar residues, those with the longer side chains are preferred by long α-helices over others with shorter side chains (e.g., Leu is preferred over Ala and Val, Gln is preferred over Asn). In addition, residues with a greater number of functional groups have higher propensities for longer helices (e.g., Arg has a higher propensity than Lys, Tyr has a higher propensity than Phe).

Potential applications of the results

The results presented above are based on a comprehensive analysis of a large database of α-helices found in nonhomologous globular proteins. Because of the sufficiently large size of the database and the use of more than one statistical measure to characterize various properties, this study is able to avoid pitfalls and small data biases. It provides an improved understanding of the fine features of α-helical structures and can lead to more accurate secondary structure prediction and better de novo design strategies.

Although several previously known structural properties of α-helices have been reaffirmed in this analysis, there are some surprises, too. For example, this study establishes for the first time a strong structural relationship between helix length and curvature, as well as amino acid composition. This result is invaluable from the point of view of de novo design, because it implies that one can regulate the curvature range of an α-helix by varying its length and composition. Furthermore, one should be able to predict the approximate curvature of an α-helix based on its length and sequence. This, in turn, requires an accurate estimation of helix length. This can be done by recognizing some "sequence punctuation marks" in the form of characteristic amino acid residues found to occur with high propensities at the helix termini (Kumar and Bansal, 1998). The development of sensitive prediction methods that can incorporate variations in amino acid propensities with respect to various structural parameters described here as well as with respect to the N- and C- terminal regions (Kumar and Bansal, 1998) should result in improvement in secondary as well as higher order structure prediction. For example, a high content of polar amino acid residues can indicate that the helix may lie at or near the protein surface, whereas the presence of an excess of apolar aliphatic residues like Leu can be an indication that the helix is long and is probably located in the core of the protein globule. The presence of β-branched and aromatic residues, in addition to Gly and Pro in the middle of helices, would suggest that the helix either has high curvature or is kinked. Similarly, these results may be very useful in the de novo design of protein motifs based on α-helical structures. If the designed helix should be long, one may consider using Leu, Arg, and Tyr instead of Ala, Val, Lys, and Phe. To generate α-helix with high curvature, residues like Thr and Gly should be preferred.

CONCLUSIONS

The present analysis of 1131 α-helices, with lengths varying from 9 to 37 residues, shows that geometries of nearly all α-helices in the globular proteins can be simply characterized as being linear, curved, or kinked. Most of the α-helices (~73%) show varying degrees of smooth curvature, with longer helices generally tending to be less curved. A localized sharp kink appears to be the preferred mode of severe/high bending for longer helices, as indicated by the fact that mean length of the kinked helices is 20 ± 6, whereas that of linear and curved helices is 12 ± 3 and 14 ± 5 residues, respectively.

α-Helices of various lengths or geometries do not segregate into different regions of a protein globule. Structural features of α-helices, such as length and geometry, do not correlate with the thermal fluctuations in the coordinates of their constituent atoms. Rather, these features are inherent properties of the helices, encoded in their amino acid sequences. Sequence compositions and propensities of individual amino acids to occur in α-helices vary with helix length, geometry, curvature, and location in the protein globule. Residues with polar side chains, viz., Gln, Glu, Lys, and Arg, are all significantly more frequent in helices that occur at or near the protein surface. Very significant differences are seen between the amino acid composition of 45 kinked α-helices and those of linear or smoothly curved

helices. Helix-breaking residues (such as β-branched and aromatic residues, in addition to Gly and Pro) are often found in the region of kink, suggesting a thermodynamic correlation. Gly, Pro, Val, Thr, and Phe are also found to occur, with more than their expected frequencies, in the highly curved helices. Among amino acids with similar side chains, those with longer and/or a greater number of functional groups are found much more frequently in longer helices (viz. Leu > Ala/Val, Gln > Asn, and Arg > Lys). Hence propensities of individual amino acids to occur in a given secondary structure depend not only on its conformation but also on the length, geometry, curvature, and location of the secondary structure in the protein globule. The sequence-structure correlation observed in this analysis, along with those observed in a previous paper (Kumar and Bansal, 1998), if incorporated in protein structure prediction algorithms, should lead to better prediction of secondary structure as well as de novo design of novel motifs containing α-helices with predefined length and geometry.

The authors are grateful to Prof. N. V. Joshi for advice on statistical analysis and to Dr. M. RaviKiran for several useful suggestions during the course of this investigation. The authors also thank Mr. Anirban Ghosh for help in preparation of the manuscript.

SK acknowledges the Council for Scientific and Industrial Research of India for a fellowship.

REFERENCES

Argos, P., and J. Palau. 1982. Amino acid distribution in protein secondary structures. *Int. J. Protein Pept. Res.* 19:380–393.

Banner, D. W., M. Kokkinidis, and D. Tsernoglou. 1987. Structure of the ColE1 Rop protein at 1.7 Å resolution. *J. Mol. Biol.* 196:657–675.

Barlow, D., and J. M. Thornton. 1988. Helix geometry in proteins. *J. Mol. Biol.* 201:601–619.

Bernstein, F. C., T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and Tasumi, M. 1977. The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* 112:535–542.

Blundell, T. L., D. Barlow, N. Barkakoti, and J. M. Thornton. 1983. Solvent induces distortions and curvature of α-helices. *Nature.* 306: 281–283.

Blundell, T. L., and Z.-Y. Zhu. 1995. The α helix as seen from the protein tertiary structure: a 3-D structural classification. *Biophys. Chem.* 55: 167–184.

Chakarabarti, P., M. Bernard, and D. C. Rees. 1986. Peptide bond distortion and curvature of α helices. *Biopolymers.* 25:1087–1093.

Chakrabartty, A., J. A. Schellman, and R. L. Baldwin. 1991. Large differences in the helix propensities of alanine and glycine. *Nature.* 351: 586–588.

Chou, K. C., and C. T. Zhang. 1995. Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* 30:275–349.

Creighton, T. E. 1993. Proteins: Structure and Molecular Properties, 2nd Ed. W. H. Freeman and Company, New York.

Hobohm, U., and C. Sander. 1994. Enlarged representative set of protein structures. *Protein Sci.* 3:522–524.

Hobohm, U., M. Scharf, R. Schneider, and C. Sander. 1992. Selection of representative protein data sets. *Protein Sci.* 1:409–417.

Kabsch, W., and C. Sander. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers.* 22:2577–2637.

Kendrew, J. C., G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wycoff, and D. C. Philips. 1958. A three dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature.* 181:662–666.

Kumar, S., and M. Bansal. 1996. Structure and sequences characteristics of long α helices in globular proteins. *Biophys. J.* 71:1574–1586.

Kumar, S., and M. Bansal. 1998. Dissecting α helices: position specific analysis of α helices in globular proteins. *Proteins Struct. Funct. Genet.* 31:460–476.

Lee, B., and F. M. Richards. 1971. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol* 55:379–400.

Medhi, J. 1992. Statistical Methods: An Introductory Text. Wiley Eastern Limited, New Delhi.

Pauling, L., R. B. Corey, and H. R. Branson. 1951. The structure of proteins: two hydrogen bonded helical configurations of the polypeptide chains. *Proc. Natl. Acad. Sci. USA.* 37:205–211.

Presta, L. G., and G. D. Rose. 1988. Helix signals in proteins. *Science.* 240:1632–1641.

Richardson, J. S., and D. C. Richardson. 1988. Amino acid preferences of specific locations at the ends of α helices. *Science.* 240:1648–1652.

Sankararamakrishnan, R., and S. Vishveshwara. 1992. Geometry of proline-containing alpha-helices. *Int. J. Pept. Protein Res.* 39:356–363.

Serrano, L., and A. R. Fersht. 1989. Capping and α helix stability. *Nature.* 342:296–299.

Serrano, L., J. L. Neira, J. Sancho, and A. R. Fersht. 1992. Effect of alanine versus glycine in α helices on protein stability. *Nature.* 256:453–455.

Srinivasan, R. 1976. Helical length distribution from protein crystallographic data. *Ind. J. Biochem. Biophys.* 13:192–193.

Von Heijne, G. 1991. Proline kinks in transmembrane alpha-helices. *J. Mol. Biol.* 218:499–503.

Vorobjev, Y. N., and J. Hermans. 1997. SIMS: computation of a smooth invariant molecular surface. *Biophys. J.* 73:722–732.

Williams, R. W., A. Chang, D. Juretic, and S. Loughran. 1987. Secondary structure predictions and medium range interactions. *Biochim. Biophys. Acta.* 916:200–204.

Woolfson, D. N., and D. H. Williams. 1990. The influence of proline residues on alpha-helical structure. *FEBS Lett.* 277:185–188.

Zhu, Z.-Y., and T. L. Blundell. 1996. The use of amino acid patterns of classified helices and strands in secondary structure prediction. *J. Mol. Biol.* 260:261–276.