Biophysical Journal Volume 71 September 1996 1574–1586

Structural and Sequence Characteristics of Long α Helices in Globular Proteins

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

ABSTRACT Elucidation of the detailed structural features and sequence requirements for α helices of various lengths could be very important in understanding secondary structure formation in proteins and, hence, in the protein folding mechanism. An algorithm to characterize the geometry of an α helix from its C^{α} coordinates has been developed and used to analyze the structures of long α helices (number of residues $\geq$ 25) found in globular proteins, the crystal structure coordinates of which are available from the Brookhaven Protein Data Bank. All long α helices can be unambiguously characterized as belonging to one of three classes: linear, curved, or kinked, with a majority being curved. Analysis of the sequences of these helices reveals that the long α helices have unique sequence characteristics that distinguish them from the short α helices in globular proteins. The distribution and statistical propensities of individual amino acids to occur in long α helices are different from those found in short α helices, with amino acids having longer side chains and/or having a greater number of functional groups occurring more frequently in these helices. The sequences of the long α helices can be correlated with their gross structural features, i.e., whether they are curved, linear, or kinked, and in case of the curved helices, with their curvature.

INTRODUCTION

The prediction and characterization of secondary structures of proteins is an area of intense research activity because of its applications to the protein folding problem and the design of biotechnologically important peptides and proteins. The most common supersecondary motifs used for de novo protein design are coiled coils and four helix bundles whose major constituent is the α helix (DeGrado et al., 1989; Eisenberg et al., 1986; Hodges et al., 1988; Hecht et al., 1990; Myszka and Chaiken, 1994). Because of the 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 predicted as well as observed to be curved for a variety of reasons, e.g., sequence (Pauling and Corey, 1953), solvent-induced distortions in hydrogen bonds (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). The mean residue length of an α helix in their analysis was 10, and this as well as other early studies led to the general belief that the α helices in globular proteins are quite short (Creighton, 1993). However, an examination of the protein crystal structures in recent Brookhaven Protein Data Bank (PDB) releases and literature reveals that long α helices are being observed quite frequently, as more structures are being determined. This indicates that long α helices may occur more frequently in globular proteins than previously believed on the basis of a smaller and probably incomplete data set.

Long α helices are often found to constitute supersecondary structures, e.g., coiled coils (O'Shea et al., 1991; Lovejoy et al., 1993; Wilson et al., 1991; Banner et al., 1987) and helix bundles (Poulos et al., 1985; Lederer et al., 1981; Finzel et al., 1985; Lawson et al., 1991; Karplus and Schulz, 1987). Furthermore, it has been seen that long helices found in globular proteins are very often structurally and functionally important (Sundaralingam et al., 1985; Ludwig et al., 1991; Banner et al., 1987; O'Shea et al., 1991; Lovejoy et al., 1993).

It is of great interest to examine, in detail, the supersecondary structures adopted by long α helices to fully appreciate their role in protein architecture and function. In this report we describe an analysis of long α helices (number of residues $\geq$25) found in high-resolution protein crystal structures. We have used the method of Sugeta and Miyazawa (1967) to calculate local helix axes and origins for every four successive C^{α} atoms of a helix. We have developed an algorithm that uses these axes and origins to directly characterize and analyze the curvature of α helices. C^{α} atoms have also been used earlier for representing protein chains and for identifying various secondary structures (Srinivasan et al., 1975; Oldfield and Hubbard, 1994). The length of an α helix can be regarded as being determined by the equilibrium between helix promoting and helix destabilizing propensities of its constituent amino acids as well as its interaction with solvent and other secondary structure elements in the protein. Because, in general, the effect of environment around an α helix can be taken to be more or less the same for helices of all lengths, the question arises whether amino acid sequences of long helices are mere extensions of those found in short helices or they have unique characteristics that are distinct from those of short helices and which can be correlated with their length and/or

Received for publication 8 November 1995 and in final form 31 May 1996.

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

0006-3495/96/09/1574/13 $2.00

structure. The present PDB database is large enough to justify undertaking such a study. In this study, we have analyzed the amino acid sequences found in long α helices and demonstrate that these helices have characteristic sequence requirements that are different from those of short α helices and can be correlated with their structure and, in case of curved helices, with their curvature.

METHODS

Composition of the data set

The January 1994 release of the PDB (Bernstein et al., 1977) contains 1640 entries for protein crystal structures solved at a resolution of 2.5 Å or better. Of these, HELIX records in 182 entries contain at least one α helix with 25 or more amino acid residues. Several of these 182 entries are similar, because they are for the same protein solved in different crystal forms, in the presence of different ligands, isolated from different sources, with point mutations and, in many cases, a combination of two or more such factors. Thus, a representative data set is selected using a strategy similar to the "Select until done" algorithm described by Hobohm et al. (1992), by choosing the best solved structures in case of multiple entries. The criteria used for defining a best solved structure are the following:

1. The structure with the lowest (i.e., best) value for resolution is selected.

2. If two structures have the same value for resolution, the one with the smaller *R* factor is chosen.

3. If two structures have the same value for both resolution and *R* factor, the one corresponding to the apoprotein is chosen.

The data set thus selected consists of 45 PDB entries, which together contain 98 long α helices. Initially all 98 long α helices have been used for structural analysis. However, because 17 of these proteins are multimeric (14 dimers and 3 trimers) in nature, the number of helices with unique sequences normalizes to 68. In four cases, the two or three helices with identical sequence, but occurring in different subunits, adopt dissimilar conformations and fall in different structural categories (described later). These four cases have been removed from the normalized data set used in subsequent structural and sequences analysis, which thus has 64 helices with unique sequences that can be unambiguously categorized in various structural classes.

Helix boundaries

The HELIX record in the PDB files is used as an initial searching tool for a long α helix. The helix boundaries are checked and reassigned using the following criteria:

1. The distances $|O_i \cdots N_{i+4}|$ are less than or equal to 3.5 Å at the ends of a helix.

2. The angle between successive local helix axes at the ends of the helix is less than 20° (see below).

This definition of helix boundaries is similar to the one widely used in the Dictionary of Protein Secondary Structure (Kabsch and Sander, 1983) and combines it optimally with that found to be the most useful by Richardson's group (Richardson and Richardson, 1988). The three implementations give similar results on helix boundaries. This enables us to directly compare the results of our analysis with those of earlier sequence analysis by Richardson and Richardson (1988).

Structural analysis

We have developed a software package called HELANAL to analyze the supersecondary structures adopted by long α helices. A long helix is characterized by fitting local helix axes and calculating local helix origins for four contiguous C^α atoms, using the procedure of Sugeta and Miyazawa (1967), and sliding this window over the length of the helix in steps of one C^α atom. The angles between successive local helix axes can identify local bends or kinks as well as the occurrence of smooth curvature in the helix. A matrix, whose elements $\mathbf{M}_{ij}$ are the bending angles between the local helix axes **i** and **j**, is obtained to get an idea about the overall geometry of the helix. Unit twist and unit height of a long helix are also calculated to analyze the uniformity of the α helix. The local helix origins trace out the path described by the helix in three-dimensional space. The local helix origins are reoriented in the *X-Y* plane, and the reoriented points are used to fit a circle and a line by the least-squares method.

The long α helices are classified as kinked, linear, or curved on the basis of the following criteria:

1. Kinked if the value of any of the local bending angles is >20°, accompanied by a large mean bending angle and a large standard deviation, for the complete helix.

2. Linear if it has a small mean bending angle (≤10°) with a small standard deviation (≤5°), all local bending angles are ≤20°, the root mean square deviation (r.m.s.d.) for a line fitted to the reoriented local helix origins is comparable to or better than the r.m.s.d for a circle fitted to the same points, and r^2 (the square of linear correlation coefficient) has a high (≥0.80) value.

3. Curved if it has a small mean bending angle (≤10°) with a small standard deviation (≤5°), all local bending angles are ≤20°, and the r.m.s.d. is much smaller (i.e., ≤ 0.5 Å) for a circle fitted to the reoriented local helix origins when compared to the r.m.s.d. for the line fitted to the same points.

Amino acid composition analysis

When one analyzes protein sequences it is necessary to take a data set containing only nonhomologous proteins to avoid bias in the results. Because we are analyzing the sequences found in long α helices, which constitute only a small fraction of the whole protein in most cases, it is important that the helices constituting the data set have nonidentical sequences, in addition to selecting just the nonhomologous proteins. However, because nonhomologous proteins can also contain helices of identical sequence and, conversely, homologous proteins can contain helices of nonidentical sequences, we obtained a data set of long helices of nonidentical sequences. Furthermore, it must be mentioned that the amino acid composition of our set of proteins containing long α helices is very similar to that of the data set of 1021 unrelated protein sequences analyzed by McCaldon and Argos (1988) (data not shown), indicating thereby that our data set contains mostly nonhomologous proteins, and results obtained from this data set are not biased on this account. This fact was further confirmed by the finding that all of the proteins used in this study, except for one pair, fall in the list of proteins (pdb_select.oct_1994 in embl database) with a sequence similarity of less than 25%. The only exceptions are 4XIS and 6XIA. Although these two proteins are quite homologous and the only long α helix present in both of them has a similar sequence in the middle, the two helices have different termini.

The distribution and frequency of occurrence of each amino acid in the proteins containing long α helices, in all of the long helices taken together; the middle regions of all long helices obtained by removing six residues from N-terminal (N-cap···N5) and C-terminal (C-5···C-cap) each, with long helices grouped into different structural families, viz. kinked, linear, and curved and according to their radii of curvature, are calculated by counting their number and percentage, respectively. The frequency of occurrence of different residue types, viz. apolar (A, F, I, L, M, V, Y), basic (K, R), acidic (D, E), and others (C, G, H, N, P, Q, S, T, W), are also computed for the long α helices and the middle regions of long α helices. χ^2 analysis has been carried out to determine whether calculated values of χ^2 between the overall amino acid distributions in various categories mentioned above are significant at the 5% level (i.e., probability of accepting the null hypothesis, $p < 0.05$). Similar calculations have been performed to identify individual amino acid residues whose frequencies show significant changes at the 5% level in various categories.

Statistical preferences for individual amino acids to occur in the middle regions of short and long α helices have been calculated according to the methods of Chou and Fasman (1978) and Williams et al. (1987).

RESULTS AND DISCUSSION

α helices in globular proteins are generally believed to be quite short and to be limited by the size of the protein globule (Creighton, 1993). Thus one might expect a correlation between helix length and the overall size of the protein, with longer helices occurring in larger proteins. In our data set we do not see any such correlation, and long α helices occur in proteins of all sizes. For example, a single polypeptide chain of ColE1 Rop consists of only 63 amino acids but it contains two long α helices forming an antiparallel coiled coil (Banner et al., 1987). Furthermore, it may be expected that short and long α helices reside in different portions of the protein. In our data set, we find long α helices to be present in all possible locations in the proteins, e.g., buried in the core, in the interface between two monomers of a multimer (ferricytochrome c′; Finzel et al., 1985) or on the surface. In a few cases, a long α helix may be buried along some portion of its length and exposed in the remaining length (citrate synthase; Wiegrand et al., 1984), or it may connect two different domains of a protein (troponin C; Satyshur et al., 1988). Also, as mentioned earlier, long α helices are often involved in supersecondary motifs and are important from both structural and functional aspects. Hence, it is important to have a good understanding of their structural characteristics.

Structural characteristics of long α helices

The average helical parameters, viz. unit height (h) and unit twist (t), for long α helices (including the kinked ones) lie in ranges between 1.48 and 1.55 Å and 95° and 101°, respectively. However, for the 49 linear and curved helices, the mean unit twist lies between 97° and 101°, and the mean unit height lies between 1.48 and 1.52 Å. Only the helices in the two subunits of manganese superoxide dismutase (PBD entry 3 MDS) have rather extreme values for mean unit twist and height, even though they are not kinked, thereby indicating unusual distorted conformations (described later). Thus, on average, long helices retain the characteristic α helical height and twist, and one cannot obtain any estimate of the axial curvature from these parameters alone.

Barlow and Thornton (1988) were the first to systematically characterize the structural geometries of α helices in globular proteins by comparison with a probe helix. This approach was, thus, an indirect way of characterizing the geometry of a helix. Our procedure does not require a reference helix and directly characterizes the geometry of helices in terms of the three-dimensional path described by the local helix origins and the angles between local helix axes. It computes a local helix axis and a local helix origin for every set of four contiguous C^α atoms and is ideally suited to characterize the 3.6_{13} α helices, because this corresponds approximately to one turn of the α helix. The algorithm is computationally very efficient, is based on simple mathematics, and requires only C^α atoms to characterize the geometry of the α helix. Recently, an analysis of the C^α geometry of protein structures has demonstrated the suitability and advantages of using only C^α atoms to characterize regular features of protein structures, especially α helices and β sheets (Oldfield and Hubbard, 1994). Richardson and Richardson (1988) also obtained the best results when helix ends were defined using C^α atom positions.

The results of our analysis of long α helices are summarized in Table 1. Most of the 98 α helices could be classified unambiguously using the criteria outlined in the Methods section. Only in three cases (indicated by an asterisk in Table 1) do the line and circle fit almost equally well, but because the r.m.s.d. for the circle fit is marginally better and the value of r^2 is low (< 0.4), these helices have been classified as curved. Helices spanning residues Asp 21–Glu 45 in both of the subunits in manganese superoxide dismutase (entry 3 MDS in PDB) (Ludwig et al., 1991) are considerably distorted, with several 5→1 hydrogen bonds being broken and the mean unit twist and mean unit height being (95.1 ± 6.8)° and (1.45 ± 0.17) Å, respectively. Hence these helices have not been included in our analysis and the unnormalized data set for structural analysis consists of 96 long α helices. After normalization, our data set contains 64 long α helices, which have been selected on the basis of the unique sequences and the best fit to various structural classes defined in the Methods section. Of these 64 helices, 9 (14.1%) are linear, 15 (23.4%) are kinked, and 40 (62.5%) are smoothly curved. Barlow and Thornton's data set contained only one helix in common with our data set, viz. myoglobin helix, spanning the residues Gly 124—Leu 149 (1 MBD in PDB), and this has been identified as curved by both methods.

Thus we can unambiguously classify about 95% of the long α helices as linear, curved, and kinked. In the earlier analysis by Barlow and Thornton (1988), use of a rather arbitrary cutoff for radius of curvature, to classify an α helix as curved or linear, had led to the α helices forming a coiled coil motif being classified as linear (see the discussion in Barlow and Thornton, 1988). We classify an α helix as curved or linear on the basis of a better circle or line fit to the path traced by local helix origins of the helix, irrespective of the value of its radius of curvature. Hence, our classification of the helix geometries is based on statistically sound criteria and does not lead to such apparent paradoxes. The present method unambiguously describes the helices in the model coiled coil for paramyosin (Parry and Suzuki, 1969) as well as the α helices in the more recently reported leucine zipper motifs as smoothly curved, with the circle fit to the local helix origins being much better than the line fit (2ZTA and MCC entries in Table 1).

Fig. 1 (*A*, *B*, and *C*) shows an example for each of the three classes of long α helices (viz., linear, curved, and

TABLE 1 Structural analysis of 98 long α helices in high-resolution protein crystal structures

PDB entry	Helix	Mean B.A. (°)	Max. B.A. (°)	r.m.s.d. (c) (Å)	r.m.s.d. (l) (Å)	r^2	Helix type	Rad. curv. (Å)
1ALK	A. N334-G360	7.5 ± 3.8	18.1	0.20	0.69	0.977	C	80
1ALK	B. N334-G360	7.5 ± 3.4	14.9	0.22	0.68	0.928	C	76
1BGC	Q 12-H 40	4.0 ± 2.0	7.8	0.21	0.22	0.998	L	209
1BGC	A101-L125	3.6 ± 2.1	6.9	0.08	0.18	0.997	L	249
1BGC	A144-R170	6.0 ± 3.0	11.7	0.19	0.89	0.765	C	59
1BTC	K258-A284	8.2 ± 3.1	17.4	0.42	1.61	0.976	C	79
1CDE	S161-G187	10.5 ± 4.3	21.1	0.33	2.55	0.929	K	41
1COS	A. W 2-E 27	5.2 ± 1.5	6.7	0.15	0.48	0.997	C	166
1COS	B. W 2-E 27	4.8 ± 3.2	12.5	0.09	1.03	0.987	C	88
1COS	C. W 2-E 27	6.7 ± 2.2	11.0	0.20	0.60	0.988	C	90
1CPT	D254-R285	7.3 ± 5.1	23.5	2.63	0.92	0.995	K	28
1CSC	N 5-Q 26	11.0 ± 5.8	24.4	0.32	1.31	0.955	K	35
1CSC	Y167-R195	10.7 ± 9.6	38.4	0.53	1.51	0.761	K	39
1CSC	Y393-G416	12.4 ± 3.7	19.2	0.26	1.49	0.931	C	34
1DHR	S 95-K120	8.2 ± 5.8	21.3	0.43	0.66	0.619	K	59
1END	D 14-G 38	9.3 ± 7.4	26.7	0.25	0.69	0.966	K	69
1FHA	Q 14-Y 40	4.7 ± 2.9	12.9	0.17	0.43	0.997	C	157
1FHA	N 50-R 76	8.1 ± 2.6	12.2	0.08	0.15	0.997	L	319
1FHA	G 96-D123	4.1 ± 1.4	7.4	0.08	0.34	0.536	C	168
1FHA	L129-M158	8.0 ± 8.5	29.7	0.28	0.58	0.806	K	99
1GPB	T 47-D 78	11.6 ± 8.1	30.9	0.56	2.78	0.923	K	41
1GPB	K289-S314	6.7 ± 2.2	10.2	0.23	1.50	0.979	C	140
1GPB	D527-Y553	6.4 ± 2.4	10.5	0.27	0.35	0.998	L	140
1GSR	A. Q 81-T107	5.1 ± 3.1	11.2	0.12	0.82	0.988	C	84
1GSR	A. G112-Q133	16.6 ± 13.5	44.9	0.37	0.95	0.905	K	37
1GSR	B. Q 81-T107	5.6 ± 3.4	12.8	0.14	0.61	0.173	C	82
1GSR	B. G112-Q133	16.0 ± 13.8	44.5	0.20	2.70	0.906	K	43
1HSB	P 57-G 83	6.3 ± 3.8	15.6	0.41	0.86	0.986	C	68
1HUW	L 6-Y 35	4.0 ± 1.6	6.9	0.18	0.31	0.652	C	182
1HUW	A155-S184	5.6 ± 1.7	8.2	0.07	2.65	0.950	C	123
1LIS	K 13-R 36	6.0 ± 2.0	9.8	0.10	0.46	0.045	C	87
1LIS	T 44-D 74	12.3 ± 14.2	54.4	1.36	10.00	0.295	K	25
1LPE	S 54-E 79	5.2 ± 1.7	8.4	0.19	0.33	0.316	C*	128
1LPE	T 89-V122	4.1 ± 2.1	8.8	0.24	1.02	0.957	C	88
1LPE	T130-Q163	5.4 ± 3.2	12.3	0.24	1.14	0.967	C	82
1LTS	D197-Q221	5.0 ± 2.2	10.7	0.17	1.03	0.987	C	91
1LVL	C 48-R 66	5.4 ± 2.4	9.0	0.14	2.57	0.877	C	83
1MBD	L124-L149	4.6 ± 3.1	12.1	0.10	9.06	0.244	C	89
1OSA	F 65-F 92	5.5 ± 2.4	11.0	0.07	1.07	0.985	C	83
1PGD	A178-V207	4.8 ± 2.0	8.5	0.25	0.68	0.996	C	194
1PGD	K315-E347	5.6 ± 2.4	9.6	0.20	0.38	0.999	L	280
1PHC	T237-S267	9.5 ± 6.7	26.5	0.26	0.57	0.965	K	144
1PYA	B. N206-A232	6.1 ± 2.9	13.6	0.25	0.94	0.909	C	59
1PYA	D. N206-A232	5.4 ± 3.0	11.8	0.24	0.93	0.882	C	58
1PYA	F. N206-A232	5.8 ± 3.3	15.8	0.23	5.10	0.780	C	58
1RCB	A 70-A 94	8.3 ± 4.0	16.2	0.24	0.31	0.993	L	117
1RHG	A. Q 11-Y 39	6.4 ± 2.6	11.1	0.13	0.27	0.994	L	211
1RHG	A. A143-A172	7.4 ± 2.4	12.0	0.09	5.68	0.756	C	67
1RHG	B. Q 11-Y 39	4.0 ± 1.5	7.7	0.12	1.22	0.985	C	97
1RHG	B. A143-A172	6.1 ± 2.8	12.8	0.12	8.58	0.487	C	67
1RHG	C. Q 11-Y 39	6.5 ± 3.0	10.9	0.19	0.22	0.214	C*	209
1RHG	C. A143-A172	8.4 ± 4.1	18.1	0.16	1.25	0.956	C	59
1RNR	A. P 66-G 88	6.8 ± 4.0	16.2	0.05	0.42	0.997	C	144
1RNR	A. I101-I129	5.2 ± 3.0	11.5	0.11	0.19	0.973	L	293
1RNR	A. S185-E222	9.6 ± 5.9	23.1	3.94	0.84	0.987	K	26
1RNR	A. M224-S254	6.6 ± 3.6	17.2	0.42	1.42	0.957	C	56
1RNR	B. E 67-G 88	6.5 ± 3.8	12.9	0.67	0.27	0.999	L	49
1RNR	B. I101-I129	5.2 ± 2.4	9.9	0.10	4.43	0.860	C	318
1RNR	B. S185-E222	10.8 ± 7.9	29.2	1.20	2.14	0.978	K	76
1RNR	B. M224-S254	7.3 ± 3.8	18.6	0.35	1.83	0.959	C	53
1ROP	A. T 2-D 30	4.5 ± 2.7	9.8	0.17	0.47	0.993	C	138
1ROP	A. A 31-F 56	4.8 ± 2.4	10.3	0.19	6.20	0.642	C	89
1TOP	F 75-F105	4.1 ± 1.8	8.9	0.05	0.96	0.992	C	132
1TRO	A. Y 7-Q 31	5.2 ± 1.5	8.2	0.14	8.97	0.232	C	109

TABLE 1 Continued

PDB entry	Helix	Mean B.A. (°)	Max. B.A. (°)	r.m.s.d. (c) (Å)	r.m.s.d. (l) (Å)	r^2	Helix type	Rad. curv. (Å)
1TRO	G. Y 7-Q 31	4.6 ± 1.9	7.0	0.17	9.58	0.109	C	96
256B	A. P 56-N 80	6.6 ± 3.7	17.1	1.68	0.34	0.987	L	26
256B	B. P 56-N 80	7.0 ± 3.8	17.0	0.64	0.47	0.995	L	62
2CCY	A. P 5-A 30	11.5 ± 8.7	33.7	0.33	1.38	0.956	K	43
2CCY	B. P 5-A 30	10.2 ± 8.7	29.3	0.68	1.22	0.984	K	56
2CMD	S285-K312	10.7 ± 10.4	38.5	0.79	1.61	0.911	K	37
2GST	A. E 90-C114	6.4 ± 3.8	13.7	0.35	0.35	0.970	L	92
2GST	B. E 90-C114	6.3 ± 3.6	13.2	0.23	0.37	0.932	L	103
2HHM	A. W 5-I 27	6.5 ± 3.2	10.4	0.23	0.28	0.302	C*	98
2HHM	B. W 5-I 27	5.9 ± 2.7	10.6	0.11	0.54	0.816	C	68
2HPD	A. Y198-S226	6.0 ± 2.4	12.4	0.29	1.40	0.982	C	88
2HPD	A. D250-K282	9.2 ± 7.3	33.8	0.35	0.92	0.315	K	83
2HPD	B. Y198-S226	3.8 ± 2.6	9.1	0.35	1.35	0.983	C	86
2HPD	B. D250-K282	11.1 ± 8.7	39.2	0.35	0.89	0.285	K	85
2LH1	N 57-T 81	6.7 ± 3.3	11.8	0.13	1.25	0.980	C	86
2LH1	S127-A152	6.2 ± 2.3	10.3	0.17	0.72	0.001	C	67
2ZTA	A. R 1-V 30	4.4 ± 1.8	7.0	0.11	0.51	0.992	C	141
2ZTA	B. R 1-V 30	4.2 ± 2.5	9.4	0.08	0.62	0.935	C	107
3GRS	G 62-M 79	5.9 ± 2.4	9.7	0.15	0.48	0.988	C	54
3GRS	W 96-H122	6.2 ± 4.3	18.1	0.32	0.18	0.999	L	148
3LAD	A. C 53-H 70	8.9 ± 5.1	16.5	0.09	0.38	0.969	C	58
3LAD	A. D 86-N112	7.3 ± 3.7	15.2	0.31	0.26	0.825	L	58
3LAD	B. G 52-E 71	6.3 ± 2.2	9.8	0.14	1.21	0.973	C	65
3LAD	B. D 86-N112	7.0 ± 2.9	13.2	0.27	0.84	0.993	C	135
3MDS	A. D 21-E 45	7.1 ± 3.1	12.3	1.78	0.29	0.955		22
3MDS	B. D 21-E 45	7.7 ± 3.0	12.9	1.03	2.65	0.920		38
4MDH	A. N305-L330	6.5 ± 2.4	11.1	0.09	0.74	0.994	C	151
4MDH	B. N305-F327	4.9 ± 1.6	6.9	0.09	0.46	0.993	C	98
4XIA	A. Y301-A327	8.3 ± 4.2	17.1	0.49	0.81	0.756	C	55
4XIA	B. Y301-A327	8.3 ± 4.2	17.4	0.49	0.82	0.794	C	55
4XIS	D295-F320	6.7 ± 5.0	16.6	0.30	2.10	0.952	C	55
6XIA	F295-A321	6.9 ± 4.5	16.8	0.40	2.00	0.959	C	61
7AAT	A. P313-E344	5.2 ± 2.6	11.2	0.33	0.93	0.972	C	84
7AAT	B. P313-E344	5.3 ± 2.6	10.3	0.37	0.83	0.952	C	87
MCC	1-28	1.7 ± 0.3	2.0	0.004	0.62	0.996	C	170

Mean B.A., Mean bending angle with standard deviation; Max. B.A., maximum bending angle. r.m.s.d. (c); root mean square deviation from best fit circle; r.m.s.d. (l), root mean square deviation from best fit line; r^2, square of linear correlation coefficient; C, L, and K, curved, linear, and kinked; Rad. curve, radius of curvature; MCC, model coiled coil for paramyosin from fiber diffraction data (Parry and Suzuki, 1969).

*r.m.s.d. (c) is only marginally better than r.m.s.d. (l), but r^2 is low (<0.4).

The number of helices with unique sequences and structures is 64.

kinked), to which helix axes have been fitted using our program. Fig. 2 (*A*, *B*, and *C*) shows the local bending angle at each residue, as well as main-chain hydrogen bond lengths for the helices in Fig. 1. Because the C-O···N-H distance is sensitive to the inclination of the peptide units with respect to the helix axis, a linear or a curved helix with small values of bending angles shows only a weak correlation between the magnitude of local bending angle and the hydrogen bond length (Fig. 2, *A* and *B*). However, in the case of kinked helices, the correlation is strong and it is found that at least two hydrogen bonds are broken (C-O···N-H distance > 3.5 Å) at or near the site of the kink (Fig. 2 *C*), which is clearly identified by the large bending angles. This finding can be utilized to identify bends and kinks in helices in a general and precise manner, especially when nonproline, nonglycine residues are present in the region of a kink, as shown in Fig. 2 *C*, where residues His 61, Trp 62, and Ala 63 are located in the region of a kink, for the helix Thr 44–Asp 74 of the fertilization protein lysin (1LIS).

Kinked helices

The 15 long α helices, classified as kinked, contain a total of 18 kinks, with three helices being kinked twice. In the majority of cases (12 cases), either Pro (seven cases) or Gly (five cases) is located in the region of kink. Angles of kink vary between 21° and 54°, with most (14 of 18) being in the range of 21–40°. In the seven cases, where a Pro is located at the kink, angles of kink vary in the range of 21–45°, with a mean angle of 31 ± 8°. In the five cases, where a Gly residue is located at the kink, the angles of kink vary in the range between 26° and 39°, with a mean angle of 32 ± 7°.

In six out of 18 cases, the kink is caused by residues other than Pro and Gly. In these cases, it is mostly the aromatic

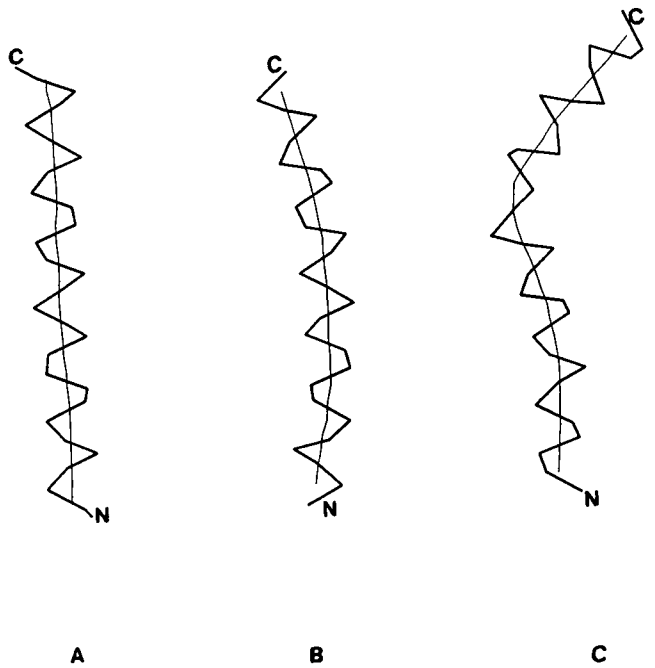

FIGURE 1 Representative examples of the three types of long α helices characterized by our method. A thick line connects C^{α} atoms in a helix and a thin line indicates the path traced by its local helix origins. (*A*) Linear: residues Q12–H40 in 1BGC. (*B*) Curved: residues A144–R170 in 1BGC. (*C*) Kinked: residues T44–D74 in 1LIS.

amino acids Tyr, Phe, His, or β-branched residue Ile that occur in the region of a kink. In fact, the largest kink (local bending angle 54.4°) found in our data set spans the residues Thr 60, His 61, Trp 62, and Ala 63 in the helix Thr 44–Asp 74 of the fertilization protein lysin (1LIS in PDB) from the mollusk red abalone (Shaw et al., 1993) and is shown in Figs. 1 *C* and 2 *C* as an example of a kinked α helix. An examination of the side-chain conformation of His 61 shows that it has a side-chain dihedral angle χ^1(N, C^{α}, C^{β}, C) of *gauche*$^-$ (−63.7°) and χ^2(C^{α}, C^{β}, C, $N^{\delta 1}$) of *gauche*$^+$ (+73.8°). In this conformation, its side-chain atom $N^{\delta 1}$ comes within hydrogen bonding distance of the carbonyl oxygen of Tyr 57 ($N^{\delta 1}{}_{61}$–O_{57} = 3.04 Å), which is also hydrogen bonded to the backbone NH (N_{61}–O_{57} = 2.73 Å). A survey of side-chain conformations of His residues in the kinked and highly curved helices (radius of curvature <80 Å) shows that the formation of such hydrogen bonds involving the His side chain $N^{\delta 1}$ is mostly facilitated by χ^1 and χ^2 being *gauche*$^-$ and *gauche*$^+$, respectively.

Curved helices

Radii of curvature of 40 long α helices, classified as curved, lie in the range of 30–200 Å. The example shown in Figs. 1 *B* and 2 *B* is the helix from 1BGC, spanning residues Ala 144 to Arg 170 and corresponding to a radius of curvature of 59 Å. Fig. 3 shows a histogram of radii of curvature versus number of long helices. The peak of the histogram lies in the range 80–100 Å. The mean radius of curvature

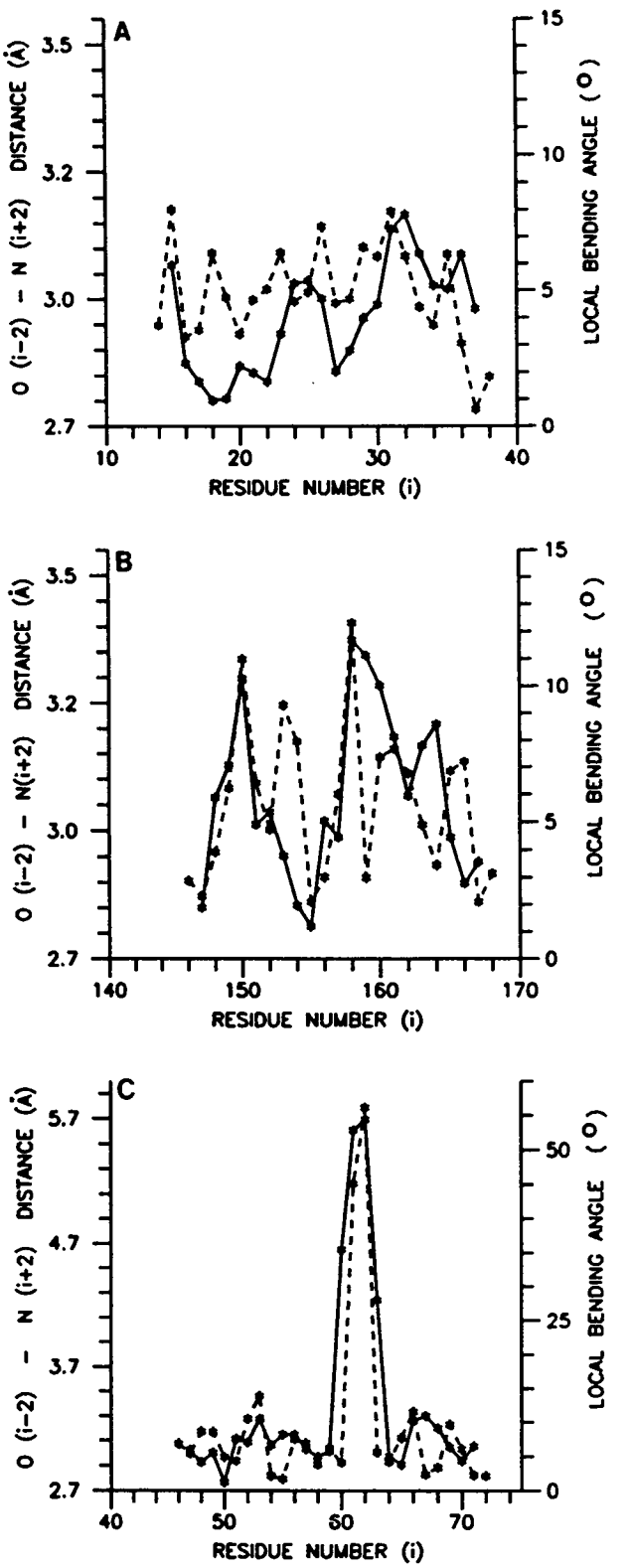

FIGURE 2 Plots of the local bending angles at each C_i^{α}, i.e., angles between the local helix axes fitted to C^{α} atoms of residues (i − 3 to i) and (i to i + 3) are shown by solid lines, along with the hydrogen bond distances between atoms O_{i-2}···N_{i+2} (indicated by dashed lines). The representative examples of (*A*) linear, (*B*) curved, and (*C*) kinked helices are the same as shown in Fig. 1.

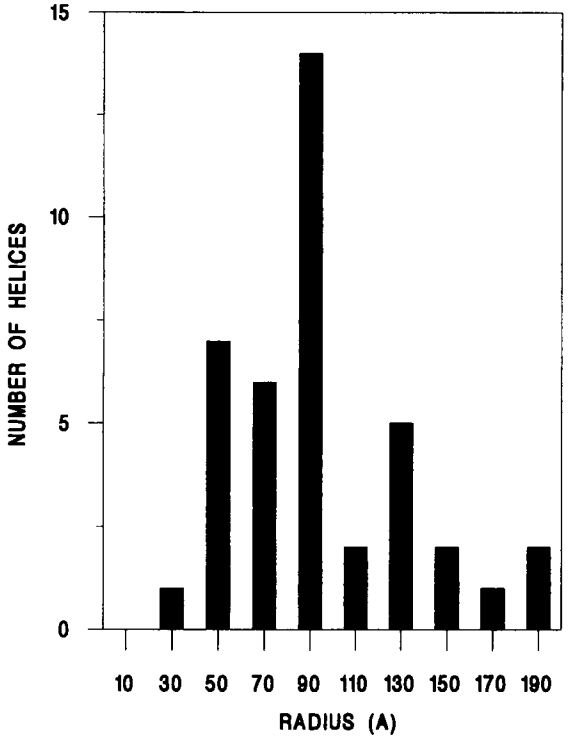

FIGURE 3 Histogram showing distribution of the 40 curved long α helices, in the normalized data set, with respect to radii of curvature. Each vertical bar corresponds to a range of ± 10 Å about the value indicated along the horizontal axis.

for long α helices is 94 Å. A simple calculation shows that for a smoothly curved α helix of 27 residues (mean residue length in our data set) with a radius of curvature of 94 Å (mean radius of curvature in our data set), the ratio of end-to-end distance to the trace length will be 0.992 and the decrease in end-to-end distance due to curvature is only 0.31 Å. However, if two helices of 27 residues, one being straight and the other being smoothly curved, with a radius of curvature 94 Å, start from the same point, their end points will diverge by $\approx$8.7 Å. Thus, a curved helix can interact with the rest of the protein core more extensively than a straight helix of similar length. This perhaps explains why 62.5% of long α helices are curved.

We have thus been able to correlate the local features of long α helices such as bending angles and local helix origins with their global geometries and categorize them in a simple and efficient manner into three classes, viz. linear, curved, and kinked. However, it may be mentioned that in our classification, the axis of a "linear" long α helix may not be a perfect straight line. Rather, it represents a set of local helix origins to which a line fits better than a circle, and hence, overall they correlate well with a line. Hence this category includes the cases where the curvature of a helix is not uniform or the helix axis curves locally in a random manner. For example, in the case of the helix spanning the residues Pro 56 to Asn 80 in subunit A of cytochrome *b* 562 (256B in PDB) from *Escherichia coli* (Lederer et al., 1981), the circle fit to the path traced by local helix origins gives a radius of curvature of 26 Å, and the r.m.s.d. from the circle is 1.68 Å. A line fit to the origins gives a r.m.s.d of 0.34 Å, and the origins correlate well to the line, with r^2 (square of linear correlation coefficient) being 0.99. We have classified this helix as linear, although it contains a small region that appears curved (Fig. 4 *A*). Similarly, a "curved" helix implies that the local helix origins fit better to a circle than a line and includes the cases where the curvature is limited largely to only one end or region of the helix rather than being distributed throughout, as expected by a rigorous definition. Furthermore, in the case of helices with large radii of curvature, mere visual examination may lead them to be erroneously described as linear. For example, consider

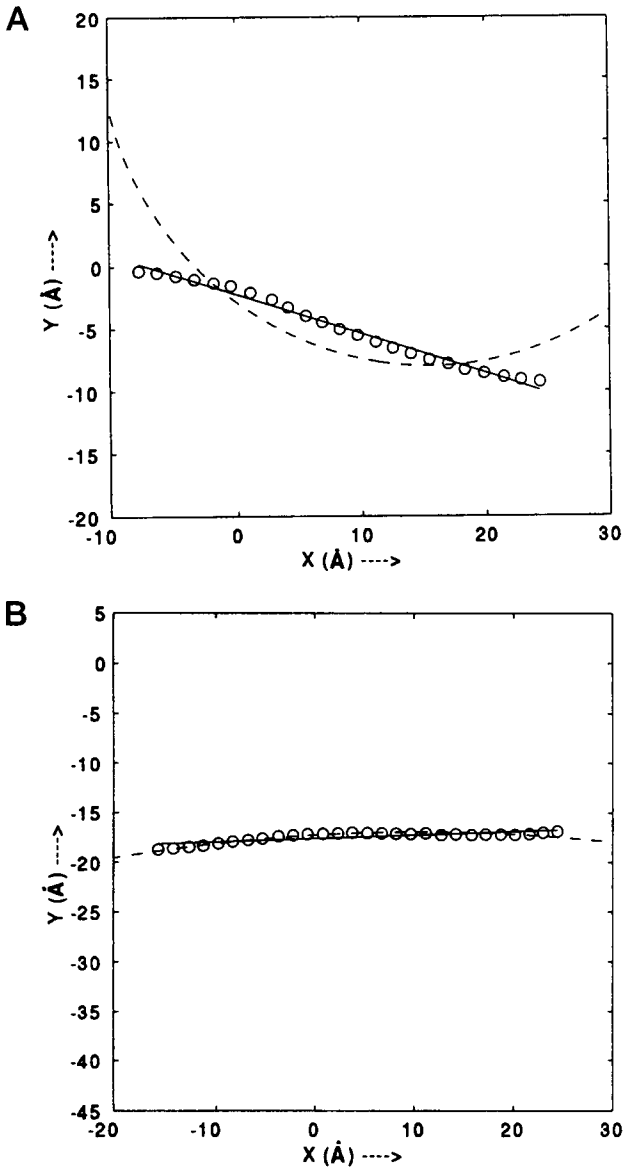

FIGURE 4 Comparison of the best line and circle fitted to the local helix origins, reoriented in the *X-Y* plane. (*A*) Residues: A P56–N80 in 256B fit better to a line. (*B*) Residues: L6–Y35 in 1HUW give a better circle fit with a large radius of curvature (182 Å).

the helix spanning the residues Leu 6 to Tyr 35 in the human growth hormone mutant (1HUW in PDB) (Ultsch et al., 1994). A circle fitted to the local helix origins for this helix gives a radius of curvature of 182 Å and a r.m.s.d. of 0.18 Å. A visual examination could lead to the helix being described as linear (Fig. 4 *B*), but we classify this helix as curved because the circle fit to the local helix origins is much better than the line fit, with a r.m.s.d of 0.31 Å and r^2 of 0.65. Because helices with a large radius of curvature are also important constituents of coiled coil motifs, an understanding of their precise geometries is of great interest.

Does local environment affect helix geometry?

Do differences in local environments, e.g., crystal packing and/or hydration around a long α helix, have some effect on its conformation? This question can be addressed by examining the conformation of long α helices with identical sequences that are present on different subunits of a multimeric protein. We have utilized the redundancy arising due to the presence of 17 multimeric proteins (14 dimers and three trimers) in our data set to answer this question. If a protein is a multimer of subunits with identical sequences (i.e., a homodimer or homotrimer), then as expected, long α helices present in the different subunits generally adopt similar structures (13 out of 17, 76%), even though the subunits are not crystallographically related. For example, alkaline phosphatase (1Alk in PDB) from *E. coli* solved at a resolution of 2.0 Å, with an *R* factor of 0.177 (Kim and Wycoff, 1991), is a homodimer. This enzyme has two long α helices, one on each subunit, that are identical in sequence, and the helix C^α atoms superpose with an r.m.s.d. of 0.1 Å. Both of these long α helices are curved, with radii of curvature of 80 Å and 76 Å. On the other hand, a few helices with identical sequence have different geometries (four out of 17, 24%). For example, human granulocyte colony-stimulating factor (R-HU-CSF; 1RHG in PDB), solved at a resolution of 2.2 Å, with an *R* factor of 0.215 (Hill et al., 1993), is a homotrimer. Each subunit of this protein contains two long α helices. The helices spanning the residues Gln 11–Tyr 39 in each subunit do not fall in the same class. The helix in subunit A is linear, whereas those in subunits B and C are curved according to our criteria. Their different classification is supported by the observation that the C^α superpositions of the helix in subunit A on the corresponding helices in subunits B and C give r.m.s.d. of 0.48 Å and 0.54 Å, respectively, whereas the C^α superposition of the helices in subunits B and C gives a much smaller r.m.s.d. of 0.27 Å. Thus long helices with the same sequence but in different subunits of a protein, and presumably differing local environments, tend to adopt similar geometries, except in a few cases.

The above section described a simple and efficient method of characterizing the geometries of α helices, and the results indicate that most long α helices have simple geometries that can be characterized as linear, curved, or kinked. One can further ask whether the sequence requirements of long α helices are different from short α helices and whether these sequence requirements can be correlated with their geometries. To answer these questions at a preliminary level, we have analyzed the sequences of long α helices and compared them with the results of earlier studies on short α helices. We have also analyzed the sequence compositions of long α helices in various structural classes and, in the case of curved helices, we have further analyzed their sequence compositions by grouping them according to their radii of curvature.

Amino acid composition of long α helices

Both long and short α helices are expected to have similar N- and C- capping regions, and differences, if any, in sequence compositions should be observed in their middle regions. The distribution of amino acids in middle regions of long α helices was compared with that in the middle regions of short helices (number of residues 8–24) taken from the data set used by Richardson and Richardson (1988), after removing the helices with more than 25 residues and the helices in common with our data set. As mentioned in the Methods section, the criteria adopted by us to define helix boundaries are similar to their implementation, and hence the data sets in the two studies can be directly compared. Table 2 shows the distribution of amino acid residues in the middle regions of short helices taken from a data set analyzed by Richardson and Richardson (1988) and the middle regions of long helices in our data set. χ^2 analysis shows that the differences between overall distributions of amino acids in the middle regions of short helices and middle regions of long helices are highly significant at the 5% level. This result indicates that sequences of long α helices are not mere extensions of sequences of short helices but have characteristics that are unique to them. To determine exactly what characteristics are different between long and short α helices, we looked at frequencies of occurrence of various residue types, viz. apolar (A, F, I, L, M, V, Y), basic (K, R), acidic (D, E), and others (C, G, H, N, P, Q, S, T, W) in the middle regions of the two classes of helices. Both short and long α helices have approximately similar amounts of apolar (49.0% in short helices and 46.5% in long helices) and other (29.7% and 27.7% in short and long helices, respectively) residues. However, in the case of polar amino acids (acidic + basic), long helices have a significantly higher proportion (25.8%), as compared to short helices (21.3%). Middle regions of short α helix sequences contain a slightly higher amount of basic residues (11.7%) than acidic residues (9.6%), indicating that short α helices in general have a net positive charge in their middle that can potentially interfere with the helix dipole and destabilize the helix. In contrast, middle regions of long α helices have almost equal proportions of basic (12.8%) and acidic (13.0%) residues, showing that long α helices do not have any net residual charge in the middle.

TABLE 2 Distribution of individual amino acids in α helices

Amino acid	Middle regions of short helices[a] (#)	Middle regions of long helices[b] (#)	Linear long helices[c] (#)	Kinked long helices[d] (#)	Curved long helices[e] (#)	Curved long helices with $R < 80$ Å[f] (#)	Curved long helices with $R < 80$ Å[g] (#)
ALA (A)	89	101	24	38	124	50	74
CYS (C)	10	13	3	6	19	7	12
ASP (D)	31	54	16	22	59	18	41
GLU (E)	23	74	20	27	105	23	82
PHE (F)	31	27	12	21	41	14	27
GLY (G)	22	41	5	19	42	23	19
HIS (H)	12	28	8	13	22	9	13
ILE (I)	33	55	14	25	37	13	24
LYS (K)	39	64	20	33	82	18	64
LEU (L)	47	136	31	51	139	43	96
MET (M)	14	38	7	12	34	11	23
ASN (N)	23	34	11	15	39	13	26
PRO (P)	7	6	2	9	8	3	5
GLN (Q)	23	54	19	18	52	12	40
ARG (R)	27	63	19	22	74	28	46
SER (S)	23	47	7	26	52	20	32
THR (T)	36	37	6	24	33	10	23
VAL (V)	49	63	9	24	61	21	40
TRP (W)	11	13	3	8	14	9	5
TYR (Y)	13	39	7	22	40	13	27
TOTAL	563	987	243	435	1077	358	719
χ^2*			$34.0^{c,d}$	$47.3^{d,e}$			
		$153.7^{a,b}$	$88.1^{c,f}$	$38.2^{d,f}$		$73.4^{f,g}$	
			$52.0^{c,g}$	$83.7^{d,g}$			

*Only χ^2 values significant at the 5% level for a 19-parameter data set, i.e., values greater than 30.1, are shown. The pairs of superscripts for χ^2 values denote the helical classes corresponding to the superscripts indicated in the column headings. Only middle regions of helices are included in the χ^2 comparison between short and long helices, whereas all of the residues in the long helices have been included in the analysis of linear, curved, and kinked classes. Short helices have been taken from Richardson's data set (Richardson and Richardson, 1988).

A comparison of the frequencies of occurrence of individual amino acids in the middle regions of short and long helices is shown in Fig. 5. Frequencies of Ala, Phe, Thr, and Val decrease by at least 1%, whereas frequencies of Glu, Leu, Met, Gln, Arg, and Tyr increase by more than 1%. χ^2 analysis of the occurrence of individual amino acids in long and short α helices shows that differences in the frequencies of Ala, Glu, Phe, Leu, and Thr are highly significant at the 5% level. The frequency of occurrence of Ala decreases in long α helices by as much as 5.6%, which is almost entirely compensated for by the increase in frequency of occurrence of Leu by 5.5%. In general, the proportion of residues with longer side chains and/or a greater number of potential interacting functional groups increases in the middle regions of long α helices, at the expense of those with shorter side chains. Hence it may be postulated that because long α helices, by virtue of their greater length, can interact more extensively with the other elements of secondary structures in globular proteins, their sequences may be biased in such a way as to maximize not only the number of possible interactions but also the extent (area) of the helix surface available for interaction. In this regard, especially when the interaction is hydrophobic, Leu, with its bulky aliphatic side chain, is a much better choice than Ala, as well as the β branched residues Ile, Val, and Thr (all decrease in the middle regions of long helices). Frequency of occurrence of Glu increases in long helices by 3.4% and almost wholly accounts for the increase in the proportion of acidic residues (13.0% in long helices as compared to 9.6% in short helices, as mentioned above). The fact that the proportion of Asp remains more or less constant while that of Glu increases considerably is consistent with the above hypothesis, because Glu has a longer side chain. Statistical preferences (propensities) calculated for individual amino acids to occur in the middle regions of long α helices, using the methods of Chou and Fasman (1978) and Williams et al. (1987), are markedly different from those in the middle regions of short α helices and show trends that are very similar to those observed from the variations in the frequencies of occurrence (shown in Table 3).

All of these results taken together indicate that in globular proteins, long α helices have sequence characteristics that are different from short α helices, and they tend to favor amino acids with longer side chains and a greater number of functional groups, so as to maximize the number as well as the extent of stabilizing interactions. The observation that sequences of secondary structure elements in globular proteins have length-dependent features could be an important input for protein secondary structure prediction algorithms and, if incorporated into such algorithms, should improve their accuracy, which is currently around 70% at the maximum (Rost and Sander, 1994; Rost et al., 1994; Eisenhaber

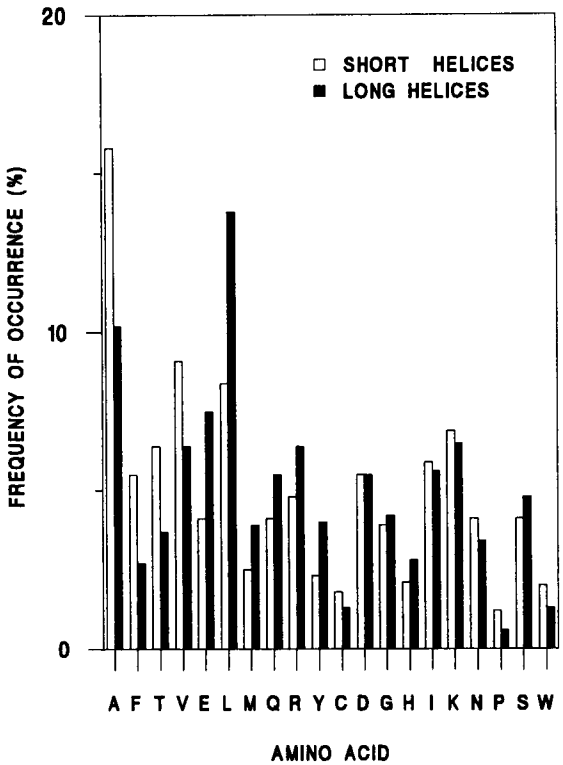

FIGURE 5 Bar diagram showing a comparison of frequencies of occurrence of individual amino acids in the middle regions of short and long α helices. Amino acids are written in single-letter code. A, E, F, L, and T show significant differences between the two classes at the 5% level.

et al., 1995). However, it must be stressed that the exact nature of length dependence of amino acid distribution in α helices remains to be determined. In addition, although our efforts have been concentrated on an analysis of α helices, it is obvious that sequences of β sheets may also show some length-dependent features.

Composition and structure correlation in long α helices

Sequence compositions of long α helices correlate, in general, with broad structural families of long α helices. For example, differences in the distribution of amino acid are significant at the 5% level between linear and kinked helices and between kinked and curved helices, but not between linear and curved helices (Table 2). Fig. 6 *A* shows the differences in frequencies of occurrence of individual amino acids in linear and kinked helices relative to the curved helices. Only Gln is significantly different at the 5% level in linear helices as compared to the kinked helices, increasing by 3.7% in linear helices. Differences between amino acid distributions in kinked and curved helices are more pronounced, with Glu, Pro, and Thr showing significant differences at the 5% level. The frequency of Pro increases by 1.4% in kinked helices as compared to the curved helices, and this is expected because Pro is known to cause kinks in α helices. The frequency of Glu decreases by 3.5%, whereas that of Thr increases by 2.4% in kinked helices. On other hand, linear helices show much smaller differences from curved helices, and none of the amino acid residue shows a significant difference at the 5% level. These observations are supported by the fact that in the four cases where long helices of identical sequences had different conformations on different subunits of multimeric proteins, the structures of helices changed between linear and curved only, not between linear and kinked or curved and kinked.

Structural studies on the 64 long α helices with unique sequences have shown that a majority (40) of them are smoothly curved (as discussed above), with radii of curvature varying between 30 and 200 Å. The sequences of long curved helices were initially grouped into nine classes according to their radii of curvature, in intervals of 20 Å (Fig. 3). χ^2 analysis of the distributions of amino acid residues in these classes showed that they can be merged into two classes that are significantly different at the 5% level in their amino acid composition, as shown in Table 2. These classes are "highly curved" (radius of curvature < 80 Å) and "less curved" (radius of curvature > 80 Å). The majority of long α helices belong to the less curved class (26 of 40, or 65%), whereas only 14 of 40 (35%) long α helices are highly curved. Fig. 6 *B* shows the differences in frequencies of occurrence in the distribution of individual amino acids in the highly curved and less curved helices, with respect to the distribution in all of the curved helices. Individual amino acids that show large changes are Ala, Gly, Lys, and Glu, with the differences for Gly, Glu, and Lys being significant

TABLE 3 Statistical preferences for individual amino acids to occur in the middle regions of α helices

Amino acid	Middle regions of short helices*	Middle regions of long helices
ALA (A)	1.76 ± 0.01	1.08 ± 0.00
CYS (C)	0.89 ± 0.02	1.09 ± 0.02
ASP (D)	0.98 ± 0.01	0.89 ± 0.00
GLU (E)	0.87 ± 0.01	1.08 ± 0.00
PHE (F)	1.38 ± 0.01	0.69 ± 0.01
GLY (G)	0.44 ± 0.00	0.55 ± 0.00
HIS (H)	0.82 ± 0.02	1.12 ± 0.01
ILE (I)	1.17 ± 0.01	1.06 ± 0.01
LYS (K)	1.01 ± 0.01	1.10 ± 0.01
LEU (L)	1.19 ± 0.01	1.41 ± 0.00
MET (M)	1.38 ± 0.03	1.54 ± 0.01
ASN (N)	0.85 ± 0.01	0.92 ± 0.01
PRO (P)	0.28 ± 0.01	0.14 ± 0.00
GLN (Q)	1.20 ± 0.02	1.31 ± 0.01
ARG (R)	1.36 ± 0.02	1.25 ± 0.01
SER (S)	0.52 ± 0.00	0.90 ± 0.01
THR (T)	1.01 ± 0.01	0.74 ± 0.01
VAL (V)	1.24 ± 0.01	1.01 ± 0.02
TRP (W)	1.37 ± 0.04	0.98 ± 0.02
TYR (Y)	0.61 ± 0.01	1.26 ± 0.01

*Helices from Richardson's data set (Richardson and Richardson, 1988). Statistical preferences have been calculated using the method of Williams et al. (1987).

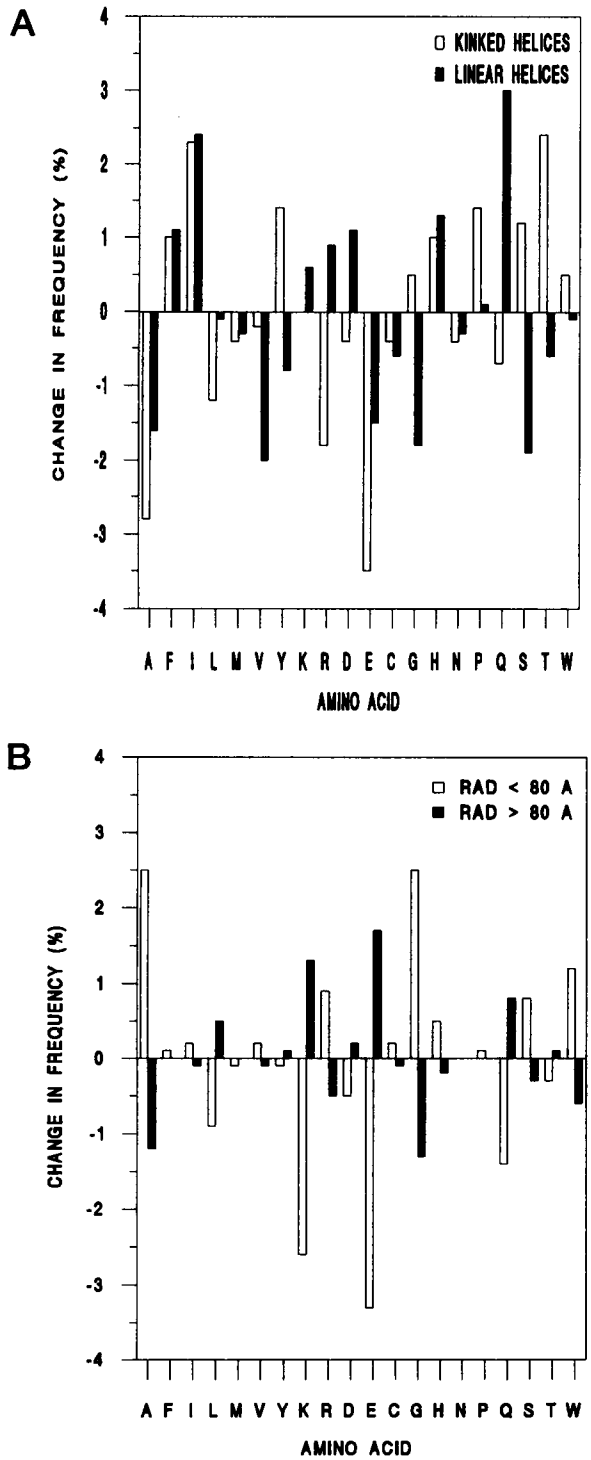

FIGURE 6 Bar diagrams showing the changes in frequencies of occurrence of individual amino acids in various structural classes of long α helices with respect to the amino acid composition of curved long α helices. Amino acids are written in single-letter code. (*A*) Linear and kinked helices. Q shows significant differences at the 5% level between linear and kinked helices, and E, P, and T show significant differences at the 5% level between curved and kinked helices. No amino acid residue shows significant difference at the 5% level between linear and curved helices. (*B*) Curved helices with radii of curvatures less than 80 Å ("highly curved" helices) and with radii of curvature greater than 80 Å ("less curved" helices). E, G, K, and W show significant differences between the two classes at the 5% level.

at the 5% level. The increase in frequency of occurrence of Gly in the case of highly curved helices is consistent with its high conformational flexibility due to the absence of a side chain. It is pertinent to mention here that Gly occurs frequently as a C-cap residue in α helices and is found in the kink regions of α helices, as described above. Given this conformational versatility of Gly, it is possible that the presence of Gly residues can also lead to a number of local bends (less severe than kinks) occurring in phase in the inner regions of long helices, thus causing them to be significantly curved in a smooth manner. Helical wheel plots of the 14 highly curved helices show that in eight of the nine helices that contain more than one Gly, the Gly residues lie on one face of the helices, whereas Ala, which also increases in the highly curved helices, shows no such preference but generally occurs close to Gly in the sequences.

A more surprising observation is that proportions of Glu and Lys increase in less curved helices and decrease in highly curved helices. The differences in absolute terms between highly curved and less curved helices are as large as 3.9% for Lys and 5% for Glu and, as mentioned earlier, these differences are highly significant at the 5% level. It may also be mentioned that the proportion of Glu also decreases by 3.5% in the kinked helices as compared to the curved helices. Increased proportions of Lys and Glu in the case of less curved helices can be rationalized because Glu and Lys often form intrahelical (i, $i \pm 3/4$) ion pairs and contribute to the stability of α helices. However, it is less clear why proportions of Lys and Glu should decrease in the highly curved helices. We hypothesize that when a helix is highly curved and if Lys and Glu lie on the inside (concave) face of the helix, some atoms in the long side-chains of Lys and Glu may come too close, leading to steric hindrance, for certain combinations of side-chain torsion angles. It may be mentioned that only four of the 14 highly curved helices contain a total of five Lys and Glu pairs separated by (i, $i \pm 3/4$) spacing in the sequence. In four such pairs, the side chains of Glu and Lys project out from the convex (outside) face of the helix, and in only one Lys and Glu pair do the side chains orient toward the concave face of the helix and form a salt bridge.

The above results indicate that the amino acid sequence in long α helices can be correlated broadly with their curvature and hence imply the existence of sequence-dependent bending of α helices. These results may be useful in the de novo design of α helices with defined curvature.

CONCLUSIONS

The overall geometry of an α helix can, in general, be characterized as kinked, linear, or curved, using its local structural features, viz. local helix origins and angles between successive local helix axes, in conjunction with well-established statistical methods. By plotting the local bending angle at residue i along with backbone

$N_{i+2}\cdots O_{i-2}$ distance, one can identify and localize the kinks in α helices more precisely. A majority of long α helices show varying degrees of smooth curvature, probably because of the fact that a curved α helix can interact more extensively with the rest of the protein core than an equivalent straight helix can. The long α helices have unique sequence features that are different from short α helices in globular proteins. The distribution and statistical propensities of individual amino acids to occur in long α helices are different from those found in short α helices, with amino acids having longer side chains and/or a greater number of functional groups occurring more frequently in long α helices. Hence, the distribution and propensity of an amino acid to occur in a particular secondary structure apparently depend not only on the conformation but also on the length of the secondary structure. A significant correlation is also found between the sequence of a long helix and the structural family into which it falls. Furthermore, in the case of long α helices that have been classified as curved, the sequence can be correlated with their radius of curvature. The correlations between sequence composition and length of α helices as well as the structural features of long α helices should lead to a better understanding of the sequence-structure relationship in globular proteins.

The authors are grateful to Prof. N. V. Joshi for advice on statistical analysis. Dr. D. Mohanty and M. Ravikiran are thanked for many useful discussions. We thank the Interactive Graphics Facility at IISc for assistance in data retrieval.

SK acknowledges CSIR, India, for a fellowship.

REFERENCES

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. J., 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 M. Tasumi. 1977. The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* 112:535–542.

Blundell, T., D. Barlow, N. Borkakoti, and J. M. Thornton. 1983. Solvent induced distortions and the curvature of α helices. *Nature.* 306:281–283.

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

Chou, P. Y., and G. D. Fasman. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol.* 47:45–148.

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

DeGrado, W. F., Z. R. Wasserman, and J. D. Lear. 1989. Protein design, a minimalist approach. *Science.* 243:622–628.

Eisenberg, D., W. Wilcox, S. M. Eshita, P. M. Prycak, S. P. Ho, and W. F. DeGrado. 1986. The design, synthesis and crystallization of an α-helical peptide. *Proteins Struct. Funct. Genet.* 1:16–22.

Eisenhaber, F., B. Persson, and P. Argos. 1995. Protein structure prediction: recognition of primary, secondary and tertiary structural features from amino acid sequence. *Crit. Rev. Biochem. Mol. Biol.* 30:1–94.

Finzel, B. C., P. C. Weber, K. D. Hardman, and F. R. Salemme. 1985. Structure of ferricytochrome c′ from *Rhodospirillum molischianum* at 1.67 Å resolution. *J. Mol. Biol.* 186:627–643.

Hecht, M. M., J. S. Richardson, D. C. Richardson, and R. C. Ogden. 1990. De novo design, expression and characterization of felix: a four-helix bundle protein of native-like sequence. *Science.* 249:884–891.

Hill, C. P., T. D. Oslund, and D. Eisenberg. 1993. The structure of granulocyte-colony-stimulating factor and its relationship to other growth factors. *Proc. Natl. Acad. Sci. USA.* 90:5167–5171.

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

Hodges, R. S., P. D. Semchuk, A. K. Taneja, C. M. Kay, J. M. R. Parker, and C. T. Mant. 1988. Protein design using model synthetic peptides. *Pept. Res.* 1:19–30.

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

Karplus, P. A., and G. E. Schulz. 1987. Refined structure of glutathione reductase at 1.54 Å resolution. *J. Mol. Biol.* 195:701–729.

Kim, E. E., and H. W. Wyckoff. 1991. Reaction mechanism of alkaline phosphatase base on crystal structures: two metal ion catalysis. *J. Mol. Biol.* 218:449–464.

Lawson, D. M., P. J. Artymuik, S. J. Yewdall, J. M. A. Smith, J. C. Livingstone, A. Treffery, A. Luzzago, S. Levi, P. Arosio, G. Cesareni, C. D. Thomas, W. V. Shaw, and P. M. Harrison. 1991. Solving the structure of human H ferritin by genetically engineering intermolecular crystal contacts. *Nature.* 349:541–544.

Lederer, F., A. Glatigny, P. H. Bethge, H. D. Bellamy, and F. S. Mathews. 1981. Improvement of 2.5 Å resolution model of cytochrome b562 by redetermining the primary structure and using molecular graphics. *J. Mol. Biol.* 148:427–448.

Lovejoy, B., S. Choe, D. Cascio, D. K. McRorie, W. F. DeGrado, and D. Eisenberg. 1993. Crystal structure of a synthetic triple stranded α-helical bundle. *Science.* 259:1288–1292.

Ludwig, M. L., A. L. Metzer, K. A. Pattridge, and W. C. Stallings. 1991. Manganese superoxide dismutase from *Thermus thermophilus*. A structural model refined at 1.8 Å resolution. *J. Mol. Biol.* 219:335–358.

McCaldon, P., and P. Argos. 1988. Oligopeptide biases in protein sequences and their use in predicting protein coding regions in nucleotide sequences. *Proteins Struct. Funct. Genet.* 4:99–122.

Myszka, D. G., and I. M. Chaiken. 1994. Design and characterization of an intramolecular antiparallel coiled coil peptide. *Biochemistry.* 33: 2363–2372.

Oldfield, T. J., and R. E. Hubbard. 1994. Analysis of C^{α} geometry in protein structures. *Proteins Struct. Funct. Genet.* 18:324–337.

O'Shea, E. K., J. D. Klemm, P. S. Kim, and T. Alber. 1991. X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science.* 254:539–544.

Parry, D. A. D., and E. Suzuki. 1969. Intrachain potential energy of the α-helix and a coiled coil strand. *Biopolymers.* 7:189–197.

Pauling, L., and R. B. Corey. 1953. Compound configurations of polypeptide chains: structure of proteins of the α-keratin type. *Nature.* 171: 59–61.

Poulos, T. L., B. C. Finzel, I. C. Gunsalus, G. C. Wagner, and J. Kraut. 1985. The 2.6 Å crystal structure of *Pseudomonas putida* cytochrome P-450. *J. Biol. Chem.* 260:16122–16130.

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

Rost, B., and C. Sander. 1994. Structure prediction of proteins—where are we now? *Curr. Opin. Struct. Biol.* 5:372–380.

Rost, B., C. Sander, and R. Schneider. 1994. Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.* 235:13–26.

Satyshur, K. A., T. R. Sambhorao, D. Pyzalska, W. Drendel, M. Greaser, and M. Sundaralingam. 1988. Refined structure of chicken skeletal muscle troponin C in the two-calcium state at 2-Å resolution. *J. Biol. Chem.* 263:1628–1647.

Shaw, A., D. E. McRee, V. D. Vacquir, and C. D. Stout. 1993. The crystal structure of lysin, a fertilization protein. *Science.* 262:1864–1867.

Srinivasan, R., R. Balasubramanian, and S. S. Rajan. 1975. Some new methods and general results of analysis of protein crystallographic data. *J. Mol. Biol.* 98:739–747.

Sugeta, H., and T. Miyazawa. 1967. General method for calculating helical parameters of polymer chains from bond lengths, bond angles and internal-rotation angles. *Biopolymers.* 5:673–679.

Sundaralingam, M., W. Drendel, and M. Greaser. 1985. Stabilization of the long central helix of troponin C by intra helical salt bridges between charged amino side chains. *Proc. Natl. Acad. Sci. USA.* 82:7944–7947.

Ultsch, M. H., W. Somers, A. A. Kossiakoff, and A. M. de Vos. 1994. Crystal structure of affinity-matured human growth hormone at 2 Å resolution. *J. Mol. Biol.* 236:286–299.

Wiegand, G., S. Remington, J. Deisenhofer, and R. Huber. 1984. Crystal structure analysis and molecular model of a complex of citrate synthase with oxaloacetate and S-acetonyl-coenzyme A. *J. Mol. Biol.* 174: 205–219.

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

Wilson, C., M. R. Wardell, K. H. Weisgraber, R. W. Mahley, and D. A. Agard. 1991. Three-dimensional structure of the LDL receptor-binding domain of human apolipoprotein E. *Science.* 252:1817–1822.