doi:10.1016/S0022-2836(02)01338-4 J. Mol. Biol. (2003) 326, 273–291

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Sequence and Structure Patterns in Proteins from an Analysis of the Shortest Helices: Implications for Helix Nucleation

Lipika Pal[1], Pinak Chakrabarti[1]* and Gautam Basu[2,3]*

[1]*Department of Biochemistry Bose Institute P-1/12 CIT Scheme VIIM Calcutta 700 054, India*

[2]*Department of Biophysics Bose Institute P-1/12 CIT Scheme VIIM Calcutta 700 054, India*

[3]*Graduate School of Information Science Nara Institute of Science and Technology 8916-5 Takayama, Ikoma Nara 630-01, Japan*

**Corresponding authors*

The shortest helices (three-length 3_{10} and four-length α), most abundant among helices of different lengths, have been analyzed from a database of protein structures. A characteristic feature of three-length 3_{10}-helices is the shifted backbone conformation for the C-terminal residue (φ,ψ angles: −95°,0°), compared to the rest of the helix (−62°, −24°). The deviation can be attributed to the release of electrostatic repulsion between the carbonyl oxygen atoms at the two C-terminal residues and further stabilization (due to a more linear geometry) of an intrahelical hydrogen bond. A consequence of this non-canonical C-terminal backbone conformation can be a potential origin of helix kinks when a 3_{10}-helix is sequence-contiguous at the α-helix N-terminal. An analysis of hydrogen bonding, as well as hydrophobic interactions in the shortest helices shows that capping interactions, some of them not observed for longer helices, dominate at the N termini. Further, consideration of the distribution of amino acid residues indicates that the shortest helices resemble the N-terminal end of α-helices rather than the C terminus, implying that the folding of helices may be initiated at the N-terminal end, which does not get propagated in the case of the shortest helices. Finally, pairwise comparison of β-turns and the shortest helices, based on correlation matrices of site-specific amino acid composition, and the relative abundance of these short secondary structural elements, leads to a helix nucleation scheme that considers the formation of an isolated β-turn (and not an α-turn) as the helix nucleation step, with shortest 3_{10}-helices as intermediates between the shortest α-helix and the β-turn. Our results ascribe an important role played by shortest 3_{10}-helices in proteins with important structural and folding implications.

Keywords: 3_{10}-helix; α-helix; β-turn; protein folding; protein modeling

Introduction

The shortest α-helix consists of four contiguous core residues and two flanking residues that participate in two consecutive hydrogen bonds of type $5 \rightarrow 1$ (Figure 1). Similarly the shortest 3_{10}-helix contains three core and two flanking residues and the hydrogen bonds are of the type $4 \rightarrow 1$. Of all the helix lengths the two shortest categories are the most abundant (Figure 2). These helices can be considered as extensions of an α-turn[1] and a β-turn,[2–4] respectively. β-Turns are the most common form for the reversal of polypeptide chain direction. However, instead of occurring in isolation, in 58% of cases they occur as multiple turns, with two overlapping type I turns being the most common.[5] While type III turns share identical backbone dihedral angles (−60°, −30°) with 3_{10}-helices, they are also conformationally very similar to type I turns.[6] The shortest 3_{10}-helix, therefore, can be considered to be a special case of multiple turns, stabilized by a pair of backbone hydrogen bonds, where the chain trajectory follows a regular helical pattern. This implies that the occurrence of a large number of multiple β-turns and the shortest helices may be related phenomena. In this context, the shortest helices may have features of both turns as well as helices and it would be of interest to compare the amino

Abbreviations used: DSSP, Define Secondary Structure of Proteins; PDB, Protein Data Bank.

E-mail addresses of the corresponding authors: pinak@boseinst.ernet.in; gautam@is.aist-nara.ac.jp

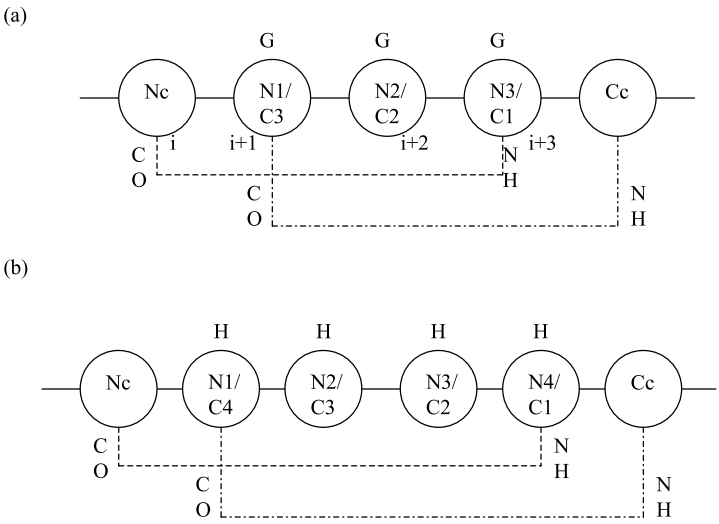

Figure 1. Schematic representations of (a) three-length (marked GGG) 3_{10}-helix with two consecutive $4 \rightarrow 1$ hydrogen bonds and (b) four-length (marked HHHH) α-helix with two consecutive $5 \rightarrow 1$ hydrogen bonds. Helical residues are flanked by two capping residues Nc and Cc. In (a) the removal of the second hydrogen bond produces a β-turn structure involving four residues (i to $i + 3$).

acid distribution in these helices with both the categories of the secondary structures.

To improve our understanding of protein folding and design, it is necessary to establish rules relating sequence and structure. An important signature of α-helix in the polypeptide sequence is the repeat *PPHHPP* of hydrophobic (*H*) and polar (*P*) residues[7–10] in the main body of the helix. However, more specific patterns of the use of polar, hydrophobic and conformationally unique residues (Gly and Pro) are found at the two terminal turns of the helix and the immediately surrounding positions.[11–16] This is because the backbone N–H and C═O groups at the initial and final turns of the helix cannot partake in normal intrahelical hydrogen bonding and need to be "capped" by alternative hydrogen bond patterns involving residues beyond the helix, which is also normally accompanied by a hydrophobic interaction between apolar residues in the helix and its flanking turn.[17] Like α-helices, the 3_{10}-helices also exhibit similar features of residue distribution.[18] Given this background, the shortest helix offers a unique structure to address a few interesting questions. As a single turn of a helix the choice of residues and their interactions with the flanking groups may either resemble the N-terminal or the C-terminal turn of a long helix or may even be different from either. Here we analyze the shortest helices and local determinants of the stabilization of their backbone conformation, their capping motifs and location in the tertiary structure. The results provide an insight into how the code for a particular local structure is distributed along the sequence.

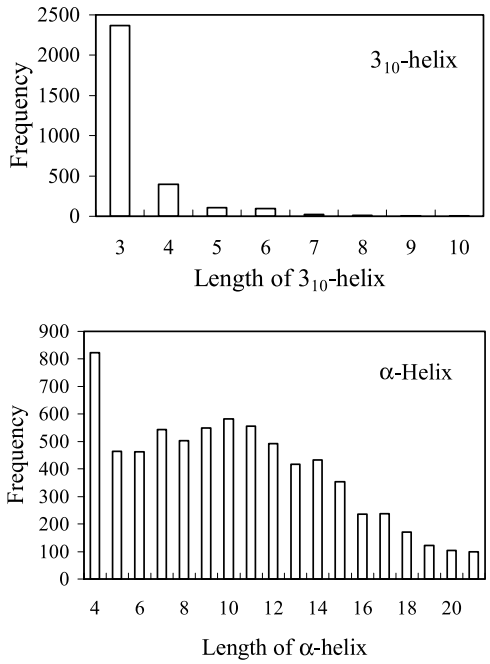

Figure 2. Length distribution of 3_{10}-helices and α-helices in proteins.

Results

In our database of 1085 polypeptide chains, there are 7548 α-helices of which 822 (~11%) are of length four residues (Figure 2). The figures for 3_{10}-helices are 3004 overall, 2366 of length three (~79%) and 396 of length four (~13%). Here, we have considered only those short 3_{10}-helices (2140 in number) for which the backbone conformation of all the residues lie in the typical right-handed helical region of the Ramachandran plot;[19] the remaining are variants of 3_{10}-helices as they contain at least one residue with non-standard ϕ,ψ angles.[20] There are 6500 cases of isolated type I β-turns.

Backbone conformation at different positions in helices

The distributions of ϕ,ψ angles in the two types of helices (Figure 3(a) and (b)) show that compared to α-helices the angles are extended to more negative ϕ and positive ψ values in 3_{10}-helices. A closer

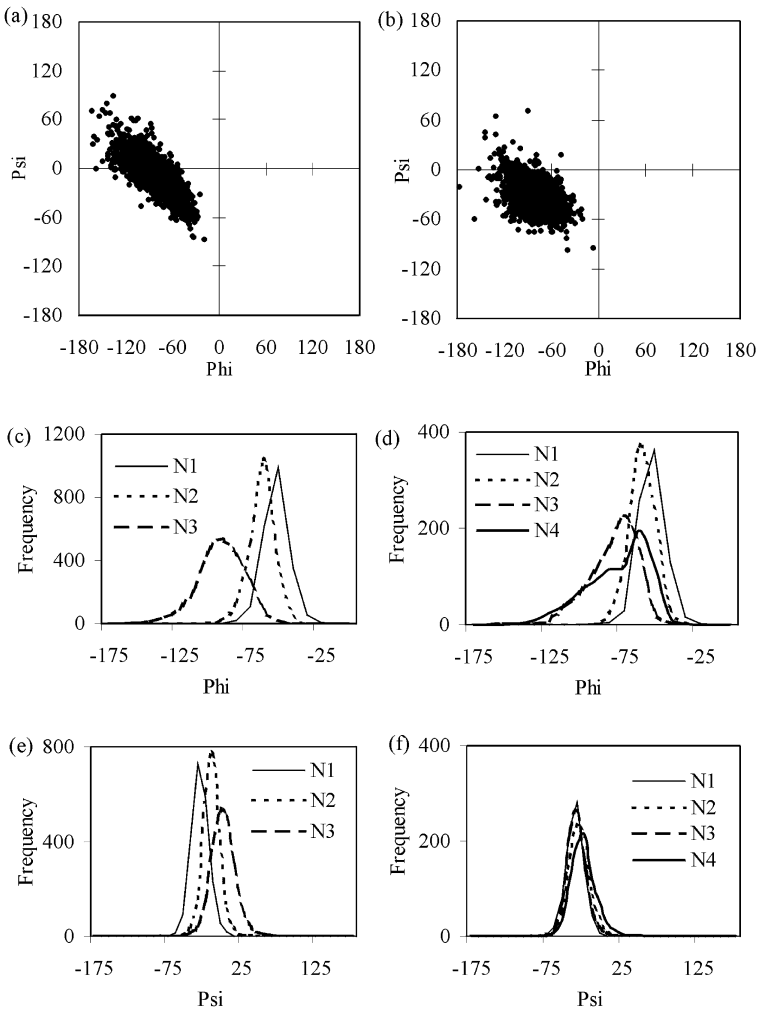

Figure 3. The distribution of ϕ,ψ angles (degrees) for residues in (a) 3-length 3_{10}-helices and (b) 4-length α-helices. Frequency of occurrence along the ϕ direction at (c) the three positions of 3-length 3_{10}-helices and (d) the four positions of 4-length α-helices; (e) and (f) are the equivalent plots along the ψ axis.

look at the distribution revealed the presence of position-specific shift in the backbone dihedral angles (Figure 3(c) and (e)). As one moves from N1 to N2 to N3 there is a clear shift in the ϕ and ψ values, the change being the most prominent for the angle ϕ between N2 and N3 positions. The average ϕ,ψ angles are at N1: −57(9)°, −32(12)°; N2: −67(9)°, −16(12)°; N3: −95(16)°,0(6)°. The average of the angles at N1 and N2, −62(10)°, −24(14)°, can be considered as the representative backbone torsion angles in a 3_{10}-helix. The ϕ,ψ angles in α-helices (Figure 3(b), (d) and (f)) is more evenly distributed around the average value, especially in ψ. Although there is some shift in the peak of the ϕ angles towards a more negative value as one moves from N1 to N3, the trend is reversed for N4. The average ϕ,ψ angles at the four positions are: N1: −57(9)°, −38(12)°; N2: −63(8)°, −32(14)°; N3: −82(17)°, −35(13)°; N4: −79(21)°, −26(17)°. While all the distributions of ϕ and ψ angles are symmetric, for N3 and N4 positions, these are asymmetric (with a negative skewness).

We wanted to see if the shift in ϕ,ψ angles observed at the N3 position of three-length 3_{10}-helices is restricted only to the short helices or if it is a general phenomenon to be found at the last position of longer 3_{10}-helices[21] also. The terminal position of longer helices has average ϕ,ψ values of −93(23)°,0(19)°, suggesting that when the ϕ,ψ angles of a residue in a 3_{10}-helix are shifted by about −35° and 25° from the average helical values of −62° and −24°, the helix cannot continue any further.

Origin of the irregularity in 3_{10}-helices

To understand the reason why the N3 position of the shortest (or the C-terminal position, in general) 3_{10}-helices adopts a large deviation in the ϕ,ψ angles, we constructed two models. The first is the ideal 3_{10}-helix, with all helical positions with ϕ,ψ angles of (−60°, −30°) and the second represents the real (the terminology "real" and "ideal" are invoked in the spirit of ideal and real gases) 3_{10}-helix, with ϕ,ψ values set to observed position specific average values, as shown in Figure 4. The distance between the carbonyl oxygen atoms at N2 and N3 increases from 3.6 Å in the ideal helix to 4.5 Å in the real helix, with a simultaneous increase in the N−H···O angle (connecting groups at Cc and N1 positions) from 139° to 151°. The

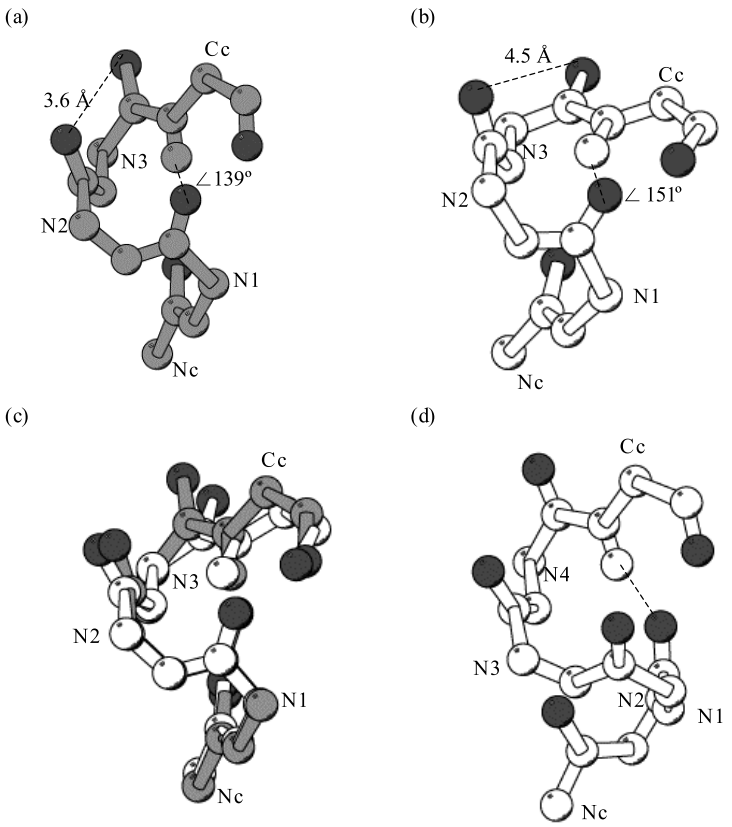

Figure 4. Molecular plots of (a) an ideal and (b) a real 3_{10}-helix. The ideal helix is built using the average (ϕ,ψ) values of (−60°, −30°) in all the three helical positions. The real helix has the average values of the (ϕ,ψ) angles as observed in these positions: N1 (−57°, −32°), N2 (−67°, −16°) and N3 (−95°,0°). An extended conformation has been used for Nc and Cc positions. The distance between the carbonyl oxygen atoms at N2 and N3 and the N−H···O angle involving the N−H group at Cc (the proton is shown only for this residue) and the oxygen of N1 are shown. The superimposition (using N1, N2 and N3 positions) of the two helices (the ideal in gray) is shown in (c). (d) A real α-helix with average ϕ,ψ angles at different position: N1 (−57°, −38°), N2 (−63°, −32°), N3 (−82°, −35°), and N4 (−79°, −26°); the intrahelical hydrogen bond between N1 and Cc is shown.

superimposition of the two helices shows that the difference in the ϕ,ψ angles at N3 has the effect of rotating the peptide plane linking N3 and Cc. Thus a change from the ideal to the real structure has two stabilizing consequences: the carbonyl group at N3 is moved away from the helix axis, resulting in lesser electrostatic repulsion between oxygen atoms at N2 and N3, and the linearity of an intrahelical hydrogen bond improves, giving a more favorable hydrogen bond energy. Various parameters for the real 3_{10}-helix match quite well with those obtained by averaging the values over all the structures (Table 1). It can be seen (Figure 5) that the O···O distance is linearly related to both the ϕ,ψ angles over the range of the backbone angles observed in different 3_{10}-helices.

A similar exercise was done to examine why the last position in the equivalent α-helices does not show any shift in backbone dihedral angles. Towards this goal a real α-helix (with average position-specific ϕ,ψ angles) was generated (Figure 4(d)), along with a chimeric helix, which had the ϕ,ψ angles at the N4 position replaced by those at the N3 position of the real 3_{10}-helix. Table 1 shows that though the O···O distance between N3 and N4 residues increases in the chimeric structure, relative to the real α-helix, the N1−Cc hydrogen bond is essentially broken (N−H···O angle ≈100°), while a new one between N2 and Cc is also not formed (N···O distance ≈3.8 Å). Thus the structure does not gain any extra stability by mimicking the changes at the C-terminal residue

Table 1. Selected values of distance (O···O, N···O) and angles (N−H···O) in shortest helices

	O···O (Å)		N···O (Å), N−H···O (°)		
(A) 3_{10}-Helix	N1−N2	N2−N3	N3−Nc	Cc−N1	
Ideal	3.6	3.6	3.0, 139	3.0, 139	
Real	3.9	4.5	2.8, 144	3.0, 151	
Average	3.9 (2)	4.4 (3)	3.0 (7), 157 (11)	3.2 (3), 160 (11)	
(B) α-Helix	N2−N3	N3−N4	N4−Nc	Cc−N1	Cc−N2
Real	3.9	3.5	2.9, 151	3.0, 156	3.9, 118
Chimeric	3.9	4.4	2.9, 151	2.5, 101	3.8, 143

Ideal and real 3_{10}-helices are displayed in Figure 4(a) and (b) and real α-helix in Figure 4(d). The chimeric α-helix has all the ϕ,ψ angles of the real α-helix, but the ϕ,ψ angles of N3 of 3_{10}-helix grafted at N4. Distance and angles in the row corresponding to average 3_{10}-helix were obtained by averaging (with standard deviations in parenthesis) over all three-length 3_{10}-helices in the database.

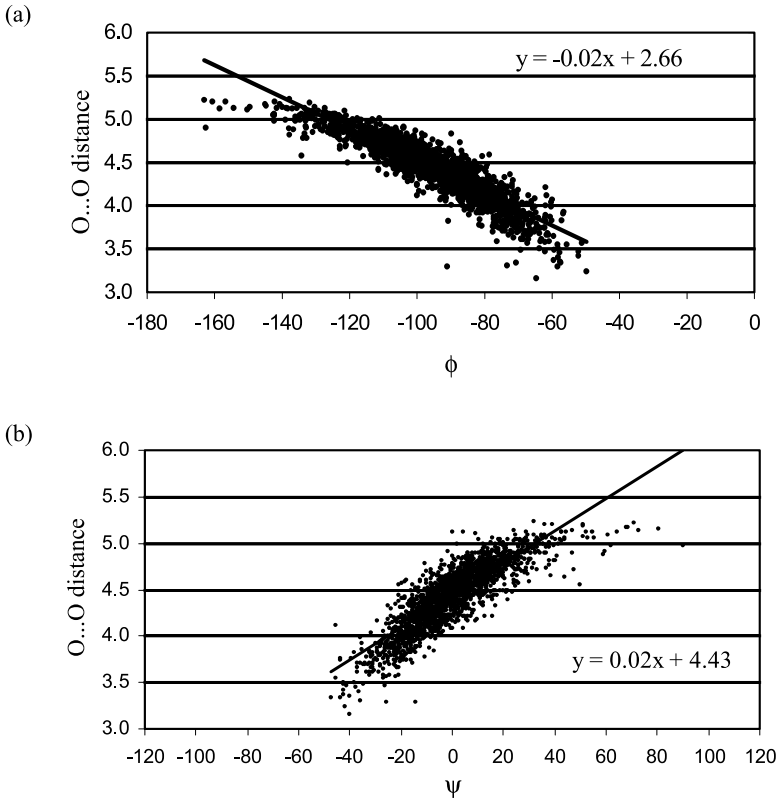

Figure 5. (a) ϕ (°) and (b) ψ (°) of N3 plotted against the O···O distance (Å) between the carbonyl oxygen atoms at N2 and N3. The correlation coefficients between the two sets of parameters are −0.9 and 0.9 respectively; the equations of the least-squares lines are also shown.

observed in the 3_{10}-helix. The slight shift in ϕ angle that was observed at the N3 position relative to N2 (Figure 3(d)) results in a small increase (by 0.3 Å) in the O···O distance between N2 and N3 positions as compared to the normal value of ~3.6 Å between the oxygen atoms of neighboring residues in an α-helix.

Positional potentials in shortest helices

The propensity and *z*-value of 20 amino acid residues to occupy each position in three-length 3_{10}-helices and four-length α-helices were calculated (Table 2). While *Karpen et al.*[18] calculated position-specific amino acid distribution based on 77 three-residue 3_{10}-helices only, our values using a database of 2140 helices are much more robust. Compared to the earlier values there have been a few changes in the individual values. Among all the residues, Asp, Glu, His, Ile, Asn, Pro and Val stand out, being either over or under-represented in four out of the five (Nc to Cc) positions, the representation of Arg being the most neutral. Likewise, for the shortest α-helices, Gly, Asn, Pro and Glu show strong positive or negative preferences in five out of six positions (Nc to Cc). For both the helices Pro has very high propensities at Nc and N1, but is totally absent at the last two positions. At the Nc position, the propensities decrease in the order Asp > Pro > Asn > His > Ser in 3_{10}-helix and Asp > Ser=Pro > Thr=Asn in α-helix. It should be noted that unlike the shortest helices, when all helices are considered with no regard to the length, Gly is also found among residues with high propensities at Nc.[22]

Comparison of percentage occurrences of residues in different helices to β-turn and terminal turns of long helices

We wanted to see if the shortest helices resemble one or the other of the terminal turns of long helices and also find out the similarity to type I β-turns. For this the correlation coefficients between the percentage compositions of different residues at each position in the shortest helices and the equivalent numbers at a given position in another secondary structure were calculated. The largest values in Table 3(B) occur when the position Nc in four-length α-helix is compared to Nc in long α-helix, N1 (i.e. N1/C4) with N1, N2 to N2 and so on. When compared to the other end of long α-helices it is found that there is no systematic trend, with only the N4 position in the four-length helix showing rather high values with three C-terminal positions of long helices. Similarly, when three-length 3_{10}-helices are compared to long α-helices (Table 3(A)) the correlation is much better at the N-terminal end of the latter (Nc matching with Nc, N1 with N1, N2 with N2). Thus both types of shortest helices correlate better with the N-terminal rather than the C-terminal turn of long α-helix. Between the two types of α-helices the match is also good at Cc.

Correlation between the three-length 3_{10}-helix and the type I β-turn is shown in Table 3(C). A

Table 2. Amino acid propensity and *z*-values of three-length 3_{10}-helical and four-length α-helical sequences

	Nc			N1			N2			N3			Cc		
Res	*N*	*P*	*Z*	*N*	*P*	*Z*	*N*	*P*	*Z*	*N*	*P*	*Z*	*N*	*P*	*Z*
(A) *Three-length* 3_{10}*-helix*															
Ala	128	0.7	−4.1	241	**1.3**	**4.7**	210	1.2	**2.2**	138	0.8	−3.3	134	0.7	−3.7
Cys	29	1.0	−0.3	28	0.9	−0.5	19	0.6	−2.1	40	**1.3**	1.7	47	**1.5**	**3.0**
Asp	316	**2.5**	**17.1**	90	0.7	−3.5	234	**1.8**	**9.6**	240	**1.9**	**10.2**	68	0.5	−5.5
Glu	74	0.6	−5.4	149	1.1	1.2	305	**2.3**	**15.1**	169	**1.3**	**3.0**	58	0.4	−6.9
Phe	73	0.9	−1.2	67	0.8	−1.8	42	0.5	−4.6	110	**1.3**	**3.0**	141	**1.7**	**6.4**
Gly	156	0.9	−0.8	62	0.4	−8.4	100	0.6	−5.3	70	0.4	−7.8	189	1.1	1.9
His	82	**1.7**	**4.7**	27	0.6	−3.2	66	**1.3**	**2.4**	72	**1.5**	**3.3**	36	0.7	−1.9
Ile	51	0.4	−6.5	110	0.9	−0.9	28	0.2	−8.6	57	0.5	−5.9	218	**1.8**	**9.3**
Lys	90	0.7	−3.3	111	0.9	−1.3	148	1.2	**2.1**	161	**1.3**	**3.3**	100	0.8	−2.4
Leu	132	0.7	−3.9	187	1.0	0.3	72	0.4	−8.6	196	1.1	1.0	302	**1.7**	**9.2**
Met	28	0.6	−2.8	49	1.0	0.3	23	0.5	−3.6	61	**1.3**	**2.0**	55	1.2	1.2
Asn	173	**1.8**	**7.9**	38	0.4	−6.1	129	**1.3**	**3.3**	177	**1.8**	**8.3**	86	0.9	−1.2
Pro	235	**2.4**	**13.9**	391	**3.9**	**29.9**	102	1.0	0.2	0	0.0	−10.2	0	0.0	−10.2
Gln	55	0.7	−2.9	53	0.7	−3.2	82	1.0	0.1	121	**1.5**	**4.6**	70	0.9	−1.2
Arg	92	0.9	−1.1	93	0.9	−1.0	95	0.9	−0.8	107	1.0	0.5	97	1.0	−0.6
Ser	164	**1.3**	**3.2**	134	1.0	0.5	262	**2.0**	**12.1**	121	0.9	−0.7	107	0.8	−2.0
Thr	118	1.0	−0.4	74	0.6	−4.5	91	0.7	−3.0	90	0.7	−3.0	94	0.8	−2.7
Val	70	0.5	−6.7	133	0.9	−1.4	40	0.3	−9.3	56	0.4	−7.9	204	**1.4**	**4.6**
Trp	23	0.8	−1.4	36	1.2	1.0	43	**1.4**	**2.2**	36	1.2	1.0	36	1.2	1.0
Tyr	49	0.6	−3.2	64	0.8	−1.5	48	0.6	−3.4	114	**1.5**	**4.3**	96	**1.3**	**2.2**

	Nc			N1			N2			N3			N4			Cc		
Res	*N*	*P*	*Z*	*N*	*P*	*Z*	*N*	*P*	*Z*	*N*	*P*	*Z*	*N*	*P*	*Z*	*N*	*P*	*Z*
(B) *Four-length* α*-helix*																		
Ala	52	0.8	−2.1	74	1.1	0.7	76	1.1	1.0	63	0.9	−0.7	94	**1.4**	**3.3**	76	1.1	1.0
Cys	16	**1.4**	1.3	11	1.0	−0.2	8	0.7	−1.0	15	**1.3**	1.0	25	**2.2**	**4.0**	11	1.0	−0.2
Asp	115	**2.4**	**9.9**	39	0.8	−1.4	73	**1.5**	**3.7**	63	**1.3**	**2.2**	45	0.9	−0.5	32	0.7	−2.4
Glu	37	0.7	−2.0	32	0.6	−2.7	130	**2.6**	**11.5**	91	**1.8**	**5.8**	30	0.6	−3.0	46	0.9	−0.7
Phe	19	0.6	−2.3	35	1.1	0.7	23	0.7	−1.5	43	**1.4**	**2.1**	35	1.1	0.7	50	**1.6**	**3.4**
Gly	62	1.0	−0.1	34	0.5	−3.8	39	0.6	−3.1	24	0.4	−5.1	16	0.3	−6.1	110	**1.8**	**6.3**
His	22	1.2	0.8	12	0.7	−1.5	20	1.1	0.3	13	0.7	−1.3	18	1.0	−0.1	25	**1.4**	1.5
Ile	19	0.4	−4.0	49	1.1	0.6	27	0.6	−2.8	53	1.2	1.2	56	**1.3**	1.7	29	0.6	−2.5
Lys	24	0.5	−3.5	39	0.8	−1.2	48	1.0	0.1	55	1.2	1.2	31	0.7	−2.4	54	1.1	1.0
Leu	49	0.7	−2.5	100	**1.5**	**3.9**	53	0.8	−2.0	66	1.0	−0.4	128	**1.9**	**7.5**	71	1.0	0.3
Met	10	0.6	−1.9	9	0.5	−2.1	7	0.4	−2.6	15	0.8	−0.7	21	1.2	0.8	15	0.8	−0.7
Asn	59	**1.6**	**3.8**	12	0.3	−4.2	41	1.1	0.8	21	0.6	−2.6	18	0.5	−3.1	52	**1.4**	**2.6**
Pro	72	**1.9**	**5.8**	103	**2.7**	**10.9**	38	1.0	0.1	11	0.3	−4.4	0	0.0	−6.3	0	−0.0	−6.3
Gln	18	0.6	−2.3	23	0.8	−1.4	31	1.0	0.1	41	**1.4**	**2.0**	31	1.0	0.1	28	0.9	−0.5
Arg	21	0.5	−2.9	40	1.0	0.2	34	0.9	−0.8	42	1.1	0.6	41	1.1	0.4	42	1.1	0.6
Ser	90	**1.9**	**6.1**	56	1.2	1.1	60	1.2	1.7	36	0.7	−1.9	36	0.7	−1.9	66	**1.4**	**2.6**
Thr	73	**1.6**	**4.1**	32	0.7	−2.2	32	0.7	−2.2	46	1.0	0.0	61	**1.3**	**2.3**	21	0.5	−3.8
Val	22	0.4	−4.7	44	0.8	−1.7	33	0.6	−3.2	55	1.0	−0.2	59	1.1	0.4	37	0.7	−2.7
Trp	9	0.8	−0.8	24	**2.1**	**3.7**	14	1.2	0.7	13	1.1	0.4	16	**1.4**	1.3	9	0.8	−0.8
Tyr	16	0.6	−2.5	37	**1.3**	1.5	18	0.6	−2.1	39	**1.4**	1.9	43	**1.5**	**2.7**	31	1.1	0.4

N stands for the number of occurrences, *P* for propensity and *Z* for *z*-values of amino acid residues; $P \geq 1.3$ and $Z \geq 1.96$ are given in bold; $P < 0.7$ and $Z < -1.96$ are underlined.

three-length 3_{10}-helix can be considered to be two overlapping β-turns, where the two turns correspond to Nc–N3 and N1–Cc segments of the helix. Helix–turn correlation coefficients, corresponding to the two individual turns ((Nc/*i*, N1/$i+1$, N2/$i+2$, N3/$i+3$) and (N1/*i*, N2/$i+1$, N3/$i+2$, Cc/$i+3$)) indicate that Nc–N3 is more strongly correlated with a single β-turn than N1–Cc, especially for the two N-terminal residues. Whether or not the Nc–N3 β-turn can sustain a second β-turn (N1–Cc) strongly depends on how correlated is N3 with $i+2$ position of a β-turn, similar to what was found for variant 3_{10}-helices as well.[20] In other words, a contiguous stretch of four residues, with strong β-turn forming potential, will add a second overlapping β-turn at its C-terminal only if the turn potential of position $i+3$ is low when compared to the canonical $i+3$ position of isolated β-turns but high compared to the canonical $i+2$ position of isolated β-turns. Thus helix propagation from an isolated β-turn is determined by the nature of the $i+3$ residue. Similarly a three-length 3_{10}-helix thus formed can propagate into a four-length 3_{10}-helix by adding a β-turn at the C terminus. Table 3(F) shows positional correlation between a three-length and a four-length 3_{10}-helix. There is strong correlation between the two across the sequence, when aligned at the N termini. Correlation between a β-turn and the C-terminal end of four-length 3_{10}-helix (N2/*i*, N3/$i+1$, N4/$i+2$, Cc/$i+3$) shows Cc and $i+3$ to be strongly correlated

Table 3. Correlation coefficients of percentage composition of amino acids at different positions of helices and type I β-turn

	Three-length 3_{10}-helix					
(A) Long α-helix	Nc	N1/C3	N2/C2	N3/C1	Cc	
Nc	**0.77**	0.06	0.59	0.39	−0.03	
N1	0.44	**0.92**	0.41	0.15	0.29	
N2	0.41	0.43	**0.89**	0.57	0.01	
N3	0.29	0.34	0.72	0.60	0.23	
C3	0.00	0.33	0.31	0.56	0.60	
C2	−0.05	0.27	0.37	0.57	0.55	
C1	0.06	0.28	0.46	0.66	0.51	
Cc	0.28	0.00	0.35	0.38	0.50	
	Four-length α-helix					
(B) Long α-helix	Nc	N1/C4	N2/C3	N3/C2	N4/C1	Cc
Nc	**0.92**	0.10	0.36	0.07	−0.03	0.33
N1	0.38	**0.87**	0.60	0.52	0.42	0.27
N2	0.47	0.31	**0.98**	**0.76**	0.22	0.39
N3	0.37	0.31	**0.90**	**0.89**	0.48	0.37
N4	−0.15	0.50	0.18	0.60	**0.90**	0.40
C4	−0.01	0.54	0.47	0.75	**0.90**	0.49
C3	0.02	0.50	0.52	0.73	**0.85**	0.51
C2	−0.04	0.40	0.58	**0.82**	0.74	0.46
C1	0.13	0.43	0.60	0.74	**0.80**	0.55
Cc	0.34	0.15	0.36	0.24	0.27	**0.95**
	Three-length 3_{10}-helix					
(C) Type I β-turn	Nc	N1	N2	N3	Cc	
i	**0.89**	0.11	0.58	0.62	0.12	
$i+1$	0.61	**0.92**	0.55	0.05	−0.15	
$i+2$	0.69	−0.11	0.68	**0.81**	−0.01	
$i+3$	0.29	−0.08	0.20	0.21	0.54	
	Four-length 3_{10}-helix					
(D) Type I β-turn	Nc	N1	N2	N3	N4	Cc
i	**0.92**	0.24	0.58	0.61	0.12	0.39
$i+1$	0.47	0.65	0.62	0.20	−0.11	0.03
$i+2$	0.70	−0.09	0.51	0.67	0.06	0.35
$i+3$	0.49	0.44	0.30	0.22	0.30	**0.86**
	Four-length α-helix					
(E) Type I β-turn	Nc	N1	N2	N3	N4	Cc
i	**0.92**	0.20	0.43	0.20	0.16	0.45
$i+1$	0.55	**0.76**	0.51	0.20	0.03	0.05
$i+2$	**0.76**	−0.05	0.55	0.39	0.15	0.34
$i+3$	0.36	0.09	0.23	0.16	0.19	**0.85**
	Three-length 3_{10}-helix					
(F) Four-length 3_{10}-helix	Nc	N1	N2	N3	Cc	
Nc	**0.93**	0.37	0.47	0.48	0.23	
N1	0.41	**0.79**	0.22	−0.05	0.39	
N2	0.59	0.50	**0.84**	0.47	0.12	
N3	0.43	0.12	0.63	**0.81**	0.32	
N4	−0.03	0.08	−0.12	0.43	**0.76**	
Cc	0.19	0.01	0.41	0.47	0.57	
	Four-length α-helix					
(G) Four-length 3_{10}-helix	Nc	N1	N2	N3	N4	Cc
Nc	**0.89**	0.46	0.42	0.17	0.18	0.44
N1	0.36	**0.84**	0.35	0.22	0.38	0.45
N2	0.67	0.54	**0.75**	0.51	0.27	0.49
N3	0.50	0.28	0.56	0.60	0.60	0.48
N4	−0.02	0.44	0.06	0.39	**0.79**	0.53
Cc	0.22	0.15	0.51	0.53	0.33	**0.93**

Values >0.75 are in bold. In (C), the positions in 3_{10}-helix corresponding to the first β-turn (i.e. Nc to N3) are singly underlined, while those for the next turn (N1 to Cc) are doubly underlined.

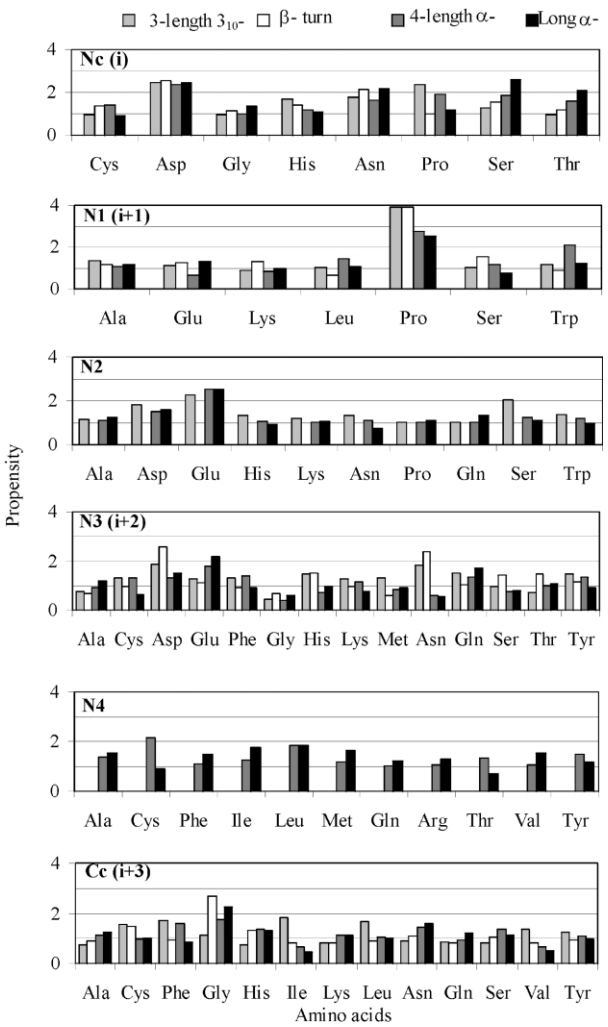

Figure 6. Histogram of residue propensities in 3-length 3_{10}-helix, type I β-turn, 4-length α-helix and long α-helix as a function of positions in helix (or β-turn); the positions in β-turn equivalent to those in helices are obtained from data in Table 4. Only those residues are shown that are over-represented (based on *z*-values) in at least one of the structures.

(Table 3(D)). Thus the nature of C1′ position of a three-length 3_{10}-helix (Cc position of the equivalent four-length 3_{10}-helix), essentially determines helix propagation from three to four-length 3_{10}-helix.

Unlike β-turns, isolated α-turns are rare in proteins; a survey found only 50 in 107 protein chains.[1] This prevented us from calculating correlation between two consecutive α-turns and the four-length α-helix, like we did for two consecutive β-turns and the three-length 3_{10}-helix. However, surprisingly, the correlation between four-length α and four-length 3_{10}-helices was found to be very strong (Table 3(G)). This suggests that four contiguous and two flanking residues show similar local propensity to form an α or a 3_{10}-helix. The helical form it ultimately adopts is probably determined by the global tertiary structure context. In other words, four-length α and four-length 3_{10}-helices are easily inter-convertible with similar amino acid preference. The implications of these results, in the context of helix nucleation will be discussed later.

Positional potentials in shortest helices *vis-à-vis* longer α-helix and β-turn

For comparison, the propensities of residues that are over-represented ($z \geq 1.96$) in three-length 3_{10} and four-length α-helices (Table 2) are plotted in Figure 6 along with the values observed in the N-terminal turn of long α-helices and isolated type I β-turns. The positions in helices of different lengths are compared in sequence, so that compared to the three-length helix, the four-length helix has an extra position, N4. The equivalence between positions in β-turn and helices is as established in Table 3. A comparison between three-length 3_{10}-helices and longer α-helices has also been made by Karpen *et al.*[18] and that between the different positions in the first turn of the α-helix by Penel *et al.*[15]

Asp and Asn have high propensities for the Nc (or position *i* in type I β-turn, Figure 1) in all the secondary structural elements. Propensities of Ser and Thr decrease in the order long α, four-length α-helix, β-turn and three-length 3_{10}-helix, while for His it is reversed. Doig *et al.*[22] explained why Ser and Thr are poorer Nc residues in 3_{10} than α-helices; in the latter they primarily hydrogen bond with the N3 NH, which is not available as a hydrogen bond partner to the Nc side-chain in a 3_{10}-helix. The preference for Pro is much more pronounced in the shortest helices and Gly has a slightly lesser preference as compared to long α-helices and β-turn. Pro is highly preferred at the N1 position in all the secondary structures, the preference being very striking for the 3_{10}-helices and β-turn. Polar residues (Glu, Lys and Ser) are slightly more preferred in β-turns, whereas non-polar (Leu and Trp) are found more in four-length α-helices.

The N2 position is occupied preferentially by polar residues, Glu in particular, when all three helix types are compared. The position N3 of helices and $i + 2$ of β-turn are compared together. As has also been noted by Hutchinson and Thornton,[5] Asp, Asn, Ser and Thr are the preferred residues at this turn position. Of all the helices, the three-length 3_{10}-helix has the best match to turn, especially in the high propensity values of Asp and Asn. Thus N3 in 3_{10}-helix and the position $i + 2$ in β-turn, which have similar φ,ψ angles, also show similar residue preferences.

At the N4 position, the comparison is really between the four-length and longer α-helices. The hydrophobic residues, Phe, Ile, Met and Val are preferred more in the longer helices and Cys in the four-length helices. Cc of helices and the position $i + 3$ of β-turn have been put together. The three-length helices strongly prefer hydrophobic residues, such as Cys, Phe, Ile, Leu, Val and Tyr. The four-length and longer α-helices have

quite similar preferences, except that Gly is found more in the longer helices. Gly is also the most preferred residue in the equivalent $i + 3$ position of β-turn.

Hydrophobic interaction and capping motifs involving the shortest helices

A hydrophobic interaction straddling the helix terminus is associated with hydrogen-bonded capping of helices.[17] As the helices under consideration here are short, the two ends can have hydrophobic contacts with each other. Thus both Nc and N1 can interact with Cc in 3_{10}-helices, whereas in α-helices it is the latter interaction that is normally observed (Figure 7). In α-helices the interactions of Nc are more with residues N-terminal to it and for Cc the residues are towards the C-terminal. In contrast to α-helices, the predominant interaction of Cc is with N1′ and N2 with Nc in 3_{10}-helices. The typical hydrophobic interactions in the two types of helices are illustrated in Figure 8.

The capping motifs that can be identified considering interactions (hydrogen bonding as well as hydrophobic) between helical residues and those outside the helix in the shortest helices are given in Figure 9. The nomenclature in the first three patterns in Figure 9(a) and first four patterns in Figure 9(b) are following Aurora & Rose (1998);[17] the rest were found primarily involving 3_{10}-helices and have been named in conformity with the already known patterns.

The N-terminal capping patterns in long α-helices, such as the capping box[23,24] and big box[25] are not frequent in the shortest α-helices and not observed at all in the 3_{10}-helices. In capping box, Nc and N3 are linked to each other with a pair of backbone amide to side-chain hydrogen bonds. In 3_{10}-helix, as the N3 backbone amide is already engaged in $4 \rightarrow 1$ hydrogen bonding the capping box cannot be formed and instead there can be a cap′-box where Nc side-chain is hydrogen bonded with N2 backbone amide. β-Box (hydrogen bond between NH of Nc and CO of N3′)[17] and α-box (between NH of Nc and CO of N4′) are quite common in the shortest helices. α′-Box is mainly found in those shortest 3_{10}-helices where an α-helix has the 3_{10}-helix at its C-terminal end. When NH of Nc in the shortest helix is hydrogen bonded to both the carbonyl groups at N3′ and N4′, an αβ-box is formed. β′-Box is generated when Nc has its NH hydrogen bonded to CO of N3′ and the side-chain hydrogen bonded to NH group of N2. Thus the standard N-terminal capping patterns in long α-helices are not common in the shortest helices; instead, there are new motifs. This trend is also observed in the C-terminal capping patterns. Schellman and Pseudo-Schellman motifs[26] are totally absent for 3_{10}-helices, though it occurs to some extent in the shortest α-helices. The Pro-box motif (with a Pro at C1′)[27] occurs almost equally in the two types of shortest helices. Additionally however, there are instances in these helices, where the required three-center hydrogen bonding is present without any Pro at the C1′ position, resulting in a pseudo-Pro box motif. The residues at C1′ position in this motif have ϕ,ψ values as in the Pro-box motif.

The capping patterns considered so far are of classical kind, meaning that they involve interactions between helical residues and residues extraneous to the helix at one of the two termini. In the shortest helices, however, as Nc and Cc are only three or four residues away from each other, there can also be hydrogen bonding across the helix, resulting in the motifs shown in Figure 9(c). In the three-length 3_{10}-helices, several cases were found in which NH of Nc is hydrogen bonded to the CO group at C1′ and, reciprocally, NH of C1′ is hydrogen bonded to CO of Nc. The residues outside the helix actually define this motif, which in reality is a 4:4 β-hairpin motif,[28] with the 3_{10}-helix constituting the turn region (Figure 10(a)). This motif is unique only for the shortest 3_{10}-helices. Across these helices, there can also be a $6 \rightarrow 1$ hydrogen bond involving NH of C1′ and CO of Nc (π-box motif). A few additional cases are found with the reciprocal pattern of hydrogen bonds (π′-box motif).

About 25% of all the three types of helices exhibit the N-terminal capping patterns given in Figure 9(a). However, while about 20% of C termini of long helices participate in the motifs in Figure 9(b), this Figure reduces to 10% when the shortest helices are considered. Thus compared to the longer α-helices, lesser number of C-terminal capping motifs are observed in three-length 3_{10}- and four-length α-helices.

Secondary structures on either side of the shortest helices

We also investigated the structural context (helix or strand as sequence neighbors) of the shortest helices. Results, given in Table 4, indicate that the shortest helices can be present in two distinct structural neighborhoods, one, in which there are no immediate secondary structural elements, making the shortest helix a connector between two secondary structures and another, in which the helix is contiguous to another secondary structure or sandwiched between two of them with no gap in between. The four-length α-helix has a greater probability (79%) of occurring in isolation than a three-length 3_{10}-helix (61%).

Short bits of 3_{10}-helix are known to occur at the ends of α-helices,[29] especially at the C-terminal end.[6] Table 4(category (a)) indicates that the 3_{10}-helix is found more at the N-terminal end of an α-helix, when there is usually a kink at the junction (Figure 10(b)) and interestingly, there are examples where such a helix is sandwiched between two regular α-helices. A 3_{10}-helix leading into a β-strand is found more than a four-length α-helix in a similar situation. A few of the cases of β-strand

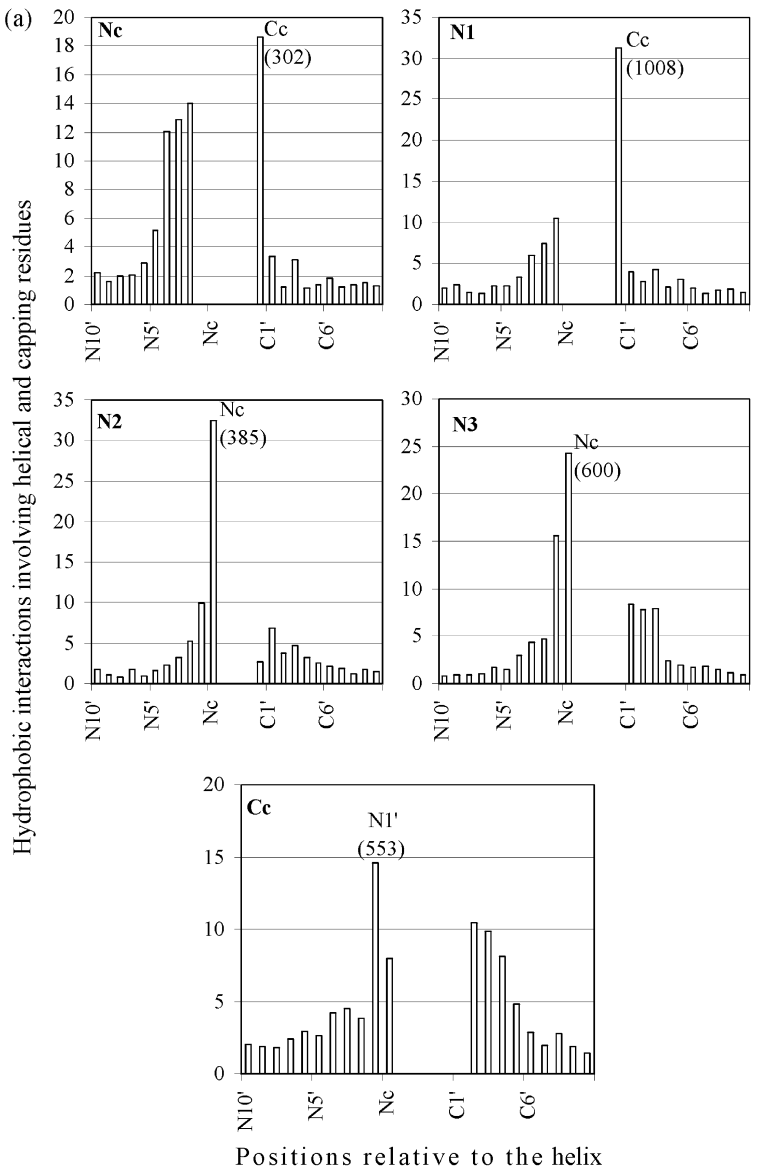

Figure 7 (*legend opposite*)

leading into a 3_{10}-helix are part of 4:4 β-hairpin structures (Figure 10(a)) mentioned in the section entitled Hydrophobic interaction and capping motifs involving the shortest helices.

Discussion

The shortest possible helices, three-length 3_{10} and four-length α (Figure 1), are very abundant in nature (Figure 2). Here we report an analysis of such helices. Specifically we focus on the extent to which (in terms of conformation and residue distribution) they resemble β-turns and longer helices, the origin of their stability, the tertiary structural context in which they occur (and a typical signature of their occurrence, especially when a 3_{10}-helix occurs immediately before an α-helix) and the perspectives the results provide on the folding process.

Conformational irregularity of 3_{10}-helices

The φ,ψ angles of residues in four-length α-helices are distributed within a compact region (Figure 3), whereas the distribution of angles in 3_{10}-helices is elongated along the direction of more negative φ and positive ψ towards what is normally called the bridging region between helical and extended conformations.[19] Interestingly, this spread occurs because there is a systematic shift of the backbone torsion angles at positions N1 to N3. This is due to a shift in the φ,ψ angles of N3 in three-length 3_{10}-helices (or the C-terminal position of 3_{10}-helices in general) by about −35° and 25°, respectively, from the average values in the

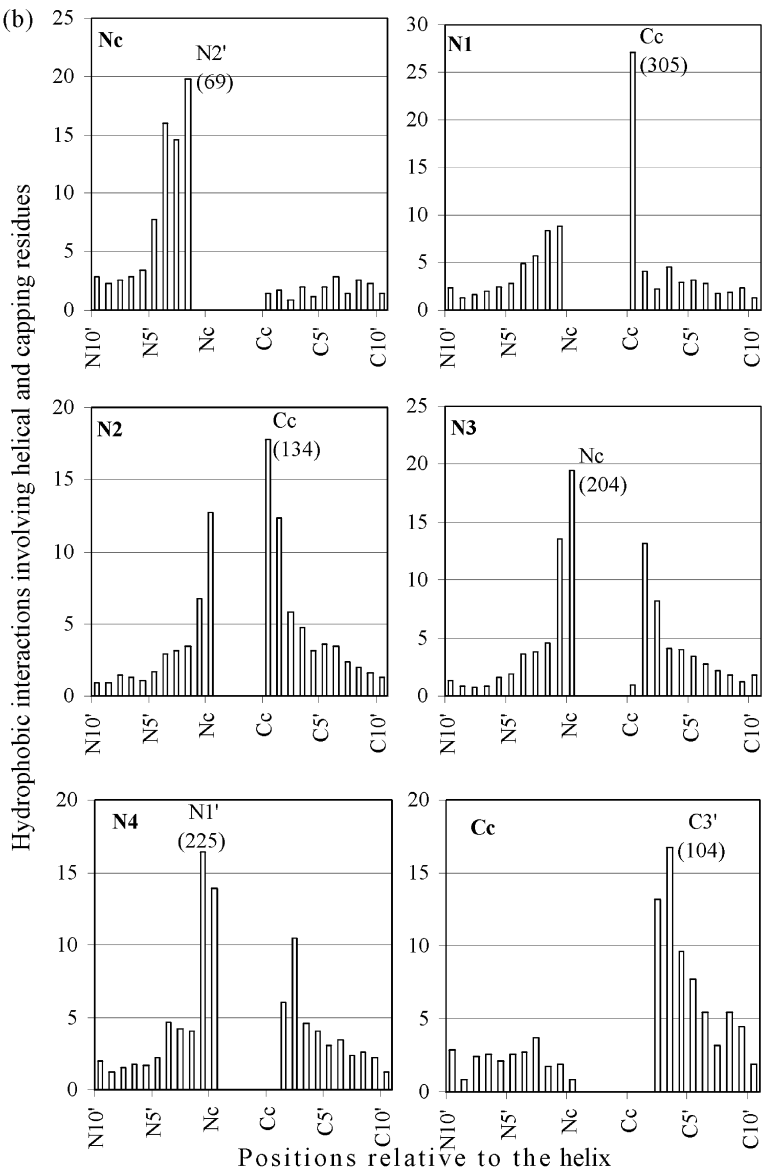

Figure 7. Position-specific hydrophobic interactions in shortest (a) 3_{10}- and (b) α-helices. The interactions are between helical residues (Nc to Cc, indicated on top of each histogram) and the flanking residues (N10′ to Nc and Cc to C10′). The ordinate is normalized (by the total number of interactions) and for the highest peak in each panel the number of hydrophobic interactions is given in parentheses.

remaining positions, which brings about two stabilizing effects in the structure (Figure 4 and Table 1). The O···O separation between two adjacent carbonyl groups at N2 and N3 is widened to reduce the electrostatic repulsion and the N–H···O angle between the NH group at Cc and the CO at N1 becomes more linear to improve the hydrogen bond stability. A similar deviation in ϕ,ψ angles is not observed at the last position (N4) of a four-length α-helix, as this would perturb the normal hydrogen bond geometry linking N1 carbonyl to NH of Cc and the longer distance across the α-helix (with a radius of 2.3 Å, *vis-à-vis* 1.9 Å in a 3_{10}-helix[30] would not allow the formation of an additional hydrogen bond (typical of 3_{10}-helices) involving positions N2 and Cc.

Similarity to type I β-turn

The four-length α-helix can be considered as an extension of α-turn. Though in an analogous manner the three-length 3_{10}-helix can also be considered to be two consecutive type III turns, this is untenable as the observed ϕ,ψ angles at N3 are atypical. In fact, the backbone angles at N2 and N3 are very similar to (−64°, −27°) and (−90°, −7°) observed at the two central residues of type I β-turn.[5] Hence a better description of a three-length 3_{10}-helix is a type III turn followed by a type I turn.

Proposed originally by Venkatachalam,[2] type III β-turn was included by Lewis *et al.*[3] in an enlarged list of β-turns and also retained by Rose *et al.*[4] as a class of turns. However, Richardson[6] suggested eliminating type III as a distinct category as it occupies contiguous regions of ϕ,ψ space as type I β-turns. Our results indicate that a single, isolated type III turn is not likely to exist. This would require the carbonyl groups of two adjacent peptide groups to align in a parallel fashion, which is electrostatically less stable and can exist only when they are embedded within a 3_{10}-helix where

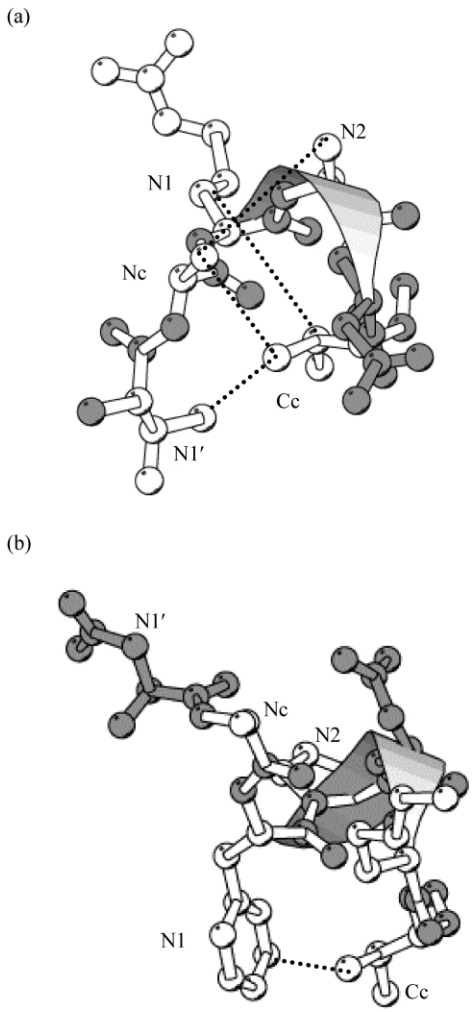

Figure 8. Hydrophobic interactions (shown by dashed lines) in shortest (a) 3_{10}- and (b) α-helices. The former is represented by residues 18–23 in 1a15A and the latter by residues 327–333 in 1a8i_. The atoms (N, C and O) of main-chain and polar residues are shown in gray shade.

the groups are held in position by intrahelical hydrogen bonds. In the three-length 3_{10}-helix two peptide carbonyl groups (at positions N2 and N3, Figure 1(a)) are not involved in such hydrogen bonds and the last carbonyl group is tilted outwards. As such, a type III turn can only be considered as part of a 3_{10}-helix and is not likely to exist in the isolated state, where it is likely to get converted to a type I β-turn.

The conformational similarity between 3_{10}-helices and type I β-turn made us find out if they have compositional similarity also. For this, correlation coefficients were calculated between the percentage composition at each position of a three-length 3_{10}-helix and the four positions of type I β-turn. As illustrated in Table 3(C), strong correlation was observed for Nc and N1 with the first turn and N3 with the second turn. This clearly spells out the sequence requirement for the formation of a three-length 3_{10}-helix originating from a pair of overlapping β-turns.

Helix kink

Helices need not necessarily be linear; in fact, a large fraction of them is curved or kinked.[31–34] There is no clear understanding of the origin of the distortion in helix, except that the occurrence of a Pro inside the helix is known to introduce a kink in the structure.[32,35] Our results suggest a possible reason for helix distortion. Because of the local relaxation from typical helical values in the ϕ,ψ angles at the last position of a 3_{10}-helix, if there is an α-helix immediately following a 3_{10}-helix, the two helix axes may not be aligned. Indeed, when we look at the composite helices, a 3_{10}-helix leading into an α-helix (Table 4 and also the cases with longer 3_{10}-helices), there is a bend at the last position of the 3_{10}-helix. As an example, Figure 10(b) shows a 16-residue helix where the first seven residues adopt the 3_{10}-helical conformation and the next nine adopt the α-helical conformation. The C-terminal 3_{10}-helical residue has ϕ,ψ angles of $(-103°, -7°)$ with a Δϕ of about $-40°$ and Δψ of 15° from the average 3_{10}-helical conformation. The two halves of the helix deviate from linearity by an angle of 42°, as measured following the procedure given by Chakrabarti & Chakrabarti (1998).[35] Thus the kink angle would depend on Δϕ and Δψ (the former, in particular) of the C-terminal 3_{10}-helical residue, preceding an α-helix in a contiguous manner.

Stabilizing interactions in the shortest helices

β-Turns may or may not be accompanied by a hydrogen bond between the CO group of residue i and NH of residue $i+3$ in the turn.[3] However, a three-length 3_{10}-helix has two interlocked tight turns held together by two hydrogen bonds (Figure 1(a)), contributing to its stability. In addition, characteristic hydrophobic interactions also exist, which may not necessarily be of type i, $i+3$. As the three-residue helix is short, the most prominent interaction is across it (Figures 7 and 8). Thus Cc can interact with both Nc (i, $i+4$ interaction) and N1 (i, $i+3$ interaction) and consequently this position has an over-representation of hydrophobic residues (Table 2). In the α-helical counterpart, though N1 and Cc interact (i, $i+4$ interaction), the other interactions for Nc and Cc are with a residue on the same side of the helix. For example, most of the interactions of Nc are with N2′ and Cc with C3′. In contrast, a distinguishing feature of 3_{10}-helices is the interaction of Cc with N1′ (i, $i+5$ interaction), which is almost absent in α-helices. Also the interaction between N2 and Nc (i, $i+2$ interaction) is quite prominent in 3_{10}-helices. The various interaction patterns exhibited clearly show that the shortest helices can be stabilized by hydrophobic interactions similar to single turns occurring at the two termini of longer helices.[17]

The common patterns in capping (hydrogen bonding and hydrophobic) interactions involving residues, N-terminal as well as C-terminal, in the shortest helices and regular α-helices are shown in Figure 9. Although the degree of occurrence of a particular motif varies among different helix types (for example, the capping box in Figure 9(a) is not found in the shortest 3_{10}-helix at all, whereas all the motifs in Figure 9(c) are exclusive to the shortest 3_{10}-helix), 25% N termini of all individual helix types adopt one of the several typical N-terminal motifs. However, when the C termini of the helices are considered, participation in a distinct motif (identified in Figure 9(b)) by the shortest helices, compared to the longer α-helices, is reduced by about 50%. Thus the capping of the N-terminal end of the shortest helices is much more extensive than in the other direction.

Resemblance of shortest helices to N-terminal end of long helices and implications in protein folding

The potentials of different residues to occur in α-helix and β-turns have long been established.[5,36] Here we have shown that the residue preference in type I β-turn is very similar to the N-terminal end of α-helices (Table 3(E), Figure 6). Protein folding is thought to be a hierarchic process in which folding begins with structures that are local in sequence, which interact to produce intermediates of increasing complexity, growing ultimately into the native conformation.[37] With a similar distribution of amino acid residues, interactions important for the formation of turns and helix N terminus during folding are likely to be similar.

Being equivalent to a single helical turn, the shortest helices may have residue preferences similar to either the N or the C terminus of longer helices. In the former case, the shortest helices may be thought to have originated like a normal helix, which could not propagate along the sequence in the C-terminal direction, whereas the latter would indicate that long helices could also be initiated at the C-terminal end and then propagate in the N-terminal direction. Table 3 shows that the position-specific preference in the shortest helices compares better with the N-terminal rather than the C-terminal end of regular helices. In the previous section it has been shown that typical N-terminal motifs are found more than the C-terminal motifs in these helices. Our observations suggest that there may exist a pronounced directionality (N to C terminus of the polypeptide chain) in the way the local structures are formed during folding. Computer simulation studies of helix unfolding[38–41] or helix-helix transition[42] show preferential unfolding from the C-terminal, although some recent folding simulations[43] do not support directional propagation of helix, as suggested here. However, it should be noted that typical peptide sequences for which helix folding simulations exist are mostly homopolymers (often polyalanine) with no inherent directional bias in the sequence. Our study shows that protein helices have clear directional bias in their sequence and that may be responsible for directional preference of helix growth.

β-Turns as helix nucleation sites and short 3_{10}-helices as early helix propagation intermediates

Helix–coil transition theories[44,45] consider helix formation as a series of elementary steps. The first step is helix nucleation; this is then followed by a series of helix propagation steps. Helix nucleation involves restricting the backbone ϕ,ψ angles of three sequence contiguous residues ($i+1$ through $i+3$) in the helical region with concomitant formation of a single i, $i+4$ hydrogen bond (an α-turn). The first helix propagation step, on the other hand, involves restricting the backbone ϕ,ψ angles of one additional residue ($i+4$) in the helical region and the addition of another hydrogen bond ($i+1$, $i+5$). In other words, helix propagation steps represent addition of additional overlapping α-turns to the growing helix. The α-turn is assumed to be the nucleation conformation because its elementary hydrogen bonding pattern is identical to long α-helices. Similarly, the nucleation conformation for 3_{10}-helices is considered to be the β-turn.[46–48]

However, as can be observed from protein structures, the occurrence of the α-turn is rare. This is because there is a large loss in conformational entropy in restricting three residues in a helical conformation that cannot be compensated by the formation of only one hydrogen bond or any other associated favorable non-bonded interactions. β-turns, on the other hand, restrict the conformational space of only two residues and as a result they occur frequently in proteins. If the formation of an α-turn is energetically much less favorable than the β-turn, can β-turns be the effective nucleating conformations for α-helices? Correlation coefficients between β-turns and the four-length α-helix (Table 3(E)) showed strong correlation at the N-terminal (between Nc, N1 positions in the helix and i, $i+1$ positions in the turn), as was seen between β-turns and the three-length 3_{10}-helix. In fact this N-terminal correlation was seen for all helices, independent of type or length, indicating that β-turns can indeed be the nucleation conformation of all helices.

The propagation of a β-turn into 3_{10}-helices, by addition of overlapping β-turns, was inferred from the fact that a three-length 3_{10}-helix is correlated with two overlapping β-turns and a four-length 3_{10}-helix is correlated with the three-length 3_{10}-helix and an overlapping β-turn. In an analogous way can a β-turn propagate into a four-length α-helix? Four-length α and four-length 3_{10}-helices were found to be highly correlated (Table 3(G)), indicating that the two helical forms are interconvertible (by reshuffling hydrogen bonds and a slight shift in ϕ,ψ angles). Clearly, a four-length

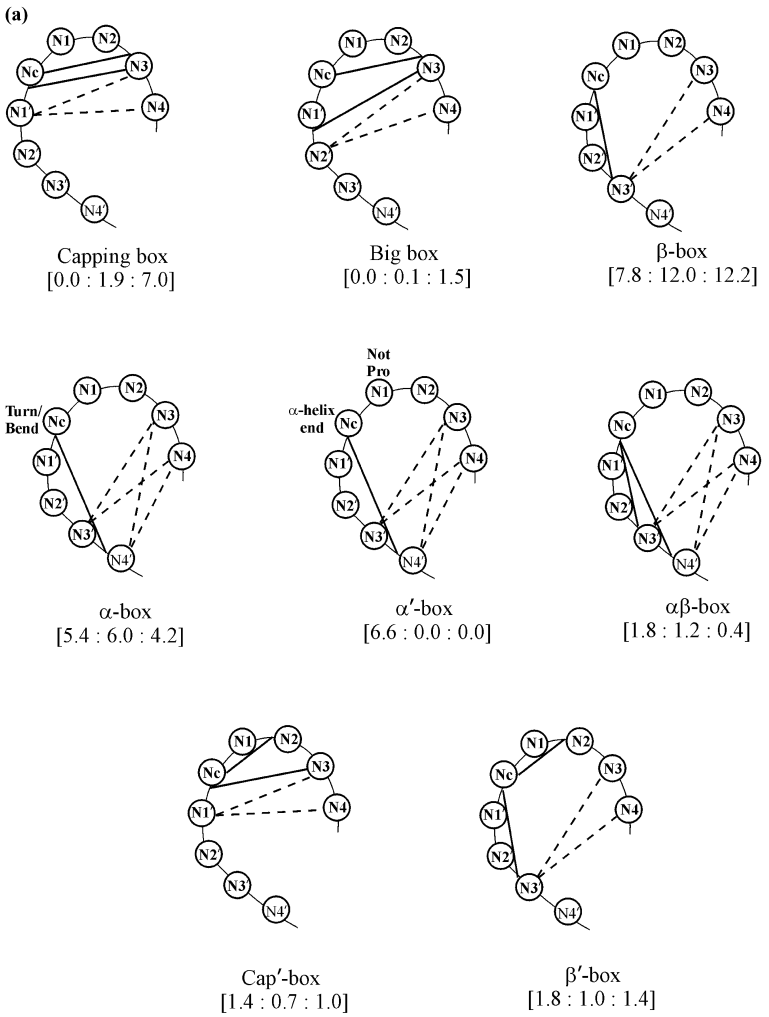

Figure 9 (*legend opposite*)

3_{10}-helix is therefore the potential precursor to the four-length α-helix, linking a β-turn to α-helices.

The observed inter-correlation between β-turns, three-length 3_{10}-helices and four-length α and 3_{10}-helices lead to a comprehensive model of α-helix nucleation and propagation, shown in Figure 11. The nucleation step for the formation of both α and 3_{10}-helices involves the formation of an isolated β-turn. This then can propagate into a longer 3_{10}-helix by adding successive β-turns. However, intrinsic instability of longer 3_{10}-helices makes helix propagation difficult beyond $n = 4$. In addition, the four-length 3_{10}-helix can rearrange its backbone hydrogen bonds and dihedral angles into a four-length α-helix. Thus the formation of four-length α-helix occurs in a concerted manner from three-length 3_{10}-helix. The α-helix can then propagate by addition of successive α-turns. From an analysis of hydrated α-helical segments in proteins, Sundaralingam and Sekharudu[49] had earlier proposed that reverse turns could act as intermediates in the folding–unfolding of helices. In addition, based on double-labeled ESR data (along with CD and NMR data) on short peptides, Millhauser (1995)[50] put forward a proposal of positioning the 3_{10}-helix along the thermodynamic folding pathway of helices (random coil ↔ nascent helix ↔ 3_{10}-helix ↔ α-helix). The picture that emerges from our work encompasses both these views. It is worth noting that α-aminoisobutyric acid (Aib), a non-coded amino acid that strongly favors the helical state,[51] also exhibits a length-dependent 3_{10}- to α-helical transition.[52] A recent review[53] summarizes the 3_{10}-α-helical equilibrium in short peptides.

The helix nucleation parameter in the Zimm–Bragg model,[44] σ, is the equilibrium constant connecting a α-turn and the statistical coil. Analysis of experimental unfolding data of peptides, when analyzed by the Zimm–Bragg model, typically yields a rather low (~0.003) value for σ.[54] However, if indeed helix nucleation bypasses the α-turn, as suggested in Figure 11, this low number may not necessarily reflect the equilibrium constant of isolated α-turn formation. Isolated β-turns are often observed in short peptide fragments or denatured proteins,[55] and therefore in all likelihood, the bottleneck in helix formation is the

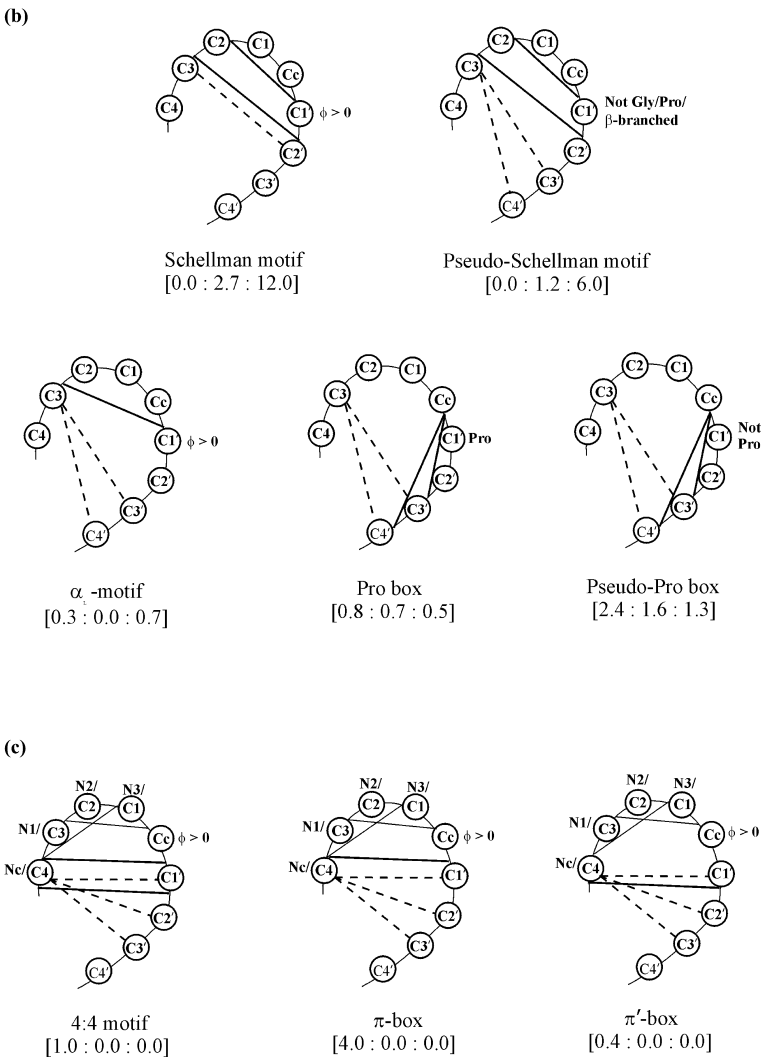

Figure 9. Schematic representations of capping patterns in (a) N-terminal, (b) C-terminal and (c) involving both N- and C-termini in three types of helices—3-length 3_{10}, 4-length α- and longer α- helices. Each motif has a name (found in the literature, or assigned here based on standard conventions) and the three numbers in square brackets correspond to the percentage occurrence of the three types of helices mentioned above. Hydrogen bonds and hydrophobic interactions typical of the motif are shown in solid and broken lines, respectively. Motifs in (c) are found only in 3_{10}-helices and the two intrahelical hydrogen bonds are also shown in thin solid lines. For 3_{10}-helices, N4 is Cc in (a) and C4 is Nc in (b). In some motifs, restrictions on the type of residue or the secondary structure or backbone torsion angle at Nc, Cc and C1′ are indicated. Depending on whether the hydrogen bond involves the side chain, the main-chain NH or CO, the corresponding line starts from the middle of the circle representing the helix position, or before it or following it.

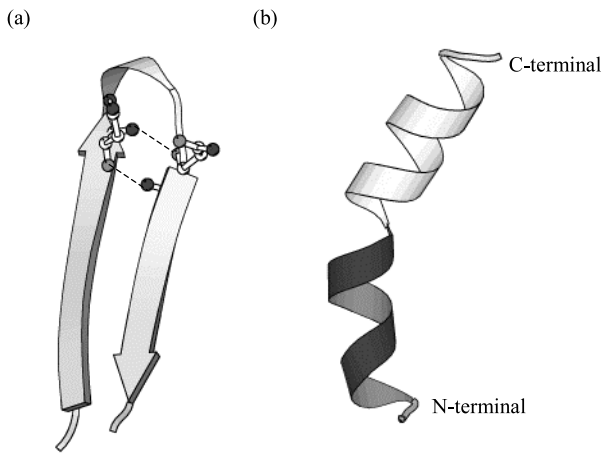

Figure 10. (a) Cartoon representation of 4:4 β-hairpin motif in the structure 1a2yB. Residues 73–75 constitute the 3_{10}-helix; a pair of intra-strand hydrogen bonds links residues at Nc and C1′, which are shown in ball-and-stick; (b) A nine-length α-helix (residues 325–333) immediately following a seven-length 3_{10}-helix (318–324) (shown in darker shade) in 1aoxA, the composite helix having a kink at position 324 with ϕ,ψ angles of (−103.2°, − 6.8°).

shortest 3_{10}-helical intermediates in Figure 11, giving rise to low values of σ. Independent support for our model comes from computer simulations of helix formation.[43] A long time simulation of reaction pathway for coil-to-helix transition in polyalanine showed formation of isolated β-turns (not α-turns) at early stages of transition. In addition, short (and not long) 3_{10}-helices were observed as intermediates. Interestingly, all the early equilibrium intermediates to α-helix, as proposed here (β-turn, three and four-length 3_{10}-helices), are populated substantially in proteins, as if these equilibrium intermediates have been frozen along the helix folding pathway due to tertiary structural demands of the global structure.

Tertiary structural context of the shortest helices

An earlier study on structural characterization of protein helices[32] had reported that a large number of short 3_{10}-helices occur mostly at α-helix termini, which has since become the generally accepted view. Contrary to this notion, among all right-handed three-length 3_{10}-helices (for 3_{10}-helix

Table 4. Occurrences of different types of secondary structures around shortest helices

	Three-length 3_{10}-helix (total 2140)			Four-length α-helix (total 805)		
	Type	Number	%	Type	Number	%
Category (a)	Connected 3_{10}-helix	826	38.6	Connected α-helix	173	21.5
	$(3_{10}) + \alpha$	207	25.1	$(\alpha) + 3_{10}$	46	26.6
	$\alpha + (3_{10})$	138	16.7	$3_{10} + (\alpha)$	39	22.5
	$(3_{10}) + \beta$	160	19.4	$(\alpha) + \beta$	17	9.8
	$\beta + (3_{10})$	131	15.9	$\beta + (\alpha)$	34	19.7
	$(3_{10}) + B$	58	7.0	$(\alpha) + B$	16	9.3
	$B + (3_{10})$	37	4.5	$B + (\alpha)$	15	8.7
	$\alpha + (3_{10}) + \alpha$	13	1.6	$3_{10} + (\alpha) + 3_{10}$	1	0.6
	$\beta + (3_{10}) + \beta$	19	2.3	$\beta + (\alpha) + \beta$	2	1.2
	$\alpha + (3_{10}) + \beta$	11	1.3	$3_{10} + (\alpha) + \beta$	0	0.0
	$\beta + (3_{10}) + \alpha$	23	2.8	$\beta + (\alpha) + 3_{10}$	3	1.7
	$\alpha + (3_{10}) + B$	7	0.9	$3_{10} + (\alpha) + B$	0	0.0
	$B + (3_{10}) + \alpha$	8	1.0	$B + (\alpha) + 3_{10}$	0	0.0
	$\beta + (3_{10}) + B$	6	0.7	$\beta + (\alpha) + B$	0	0.0
	$B + (3_{10}) + \beta$	6	0.7	$B + (\alpha) + \beta$	0	0.0
	$B + (3_{10}) + B$	2	0.2	$B + (\alpha) + B$	0	0.0
Category (b)	Isolated 3_{10}-helix	1314	61.4	Isolated α-helix	632	78.5
	α–α connector	88	6.7	α–α connector	33	5.2
	β–β connector	103	7.8	β–β connector	27	4.3
	α–β connector	87	6.6	α–β connector	27	4.3
	β–α connector	78	5.9	β–α connector	37	5.9
	α–3_{10} connector	21	1.6	α–3_{10} connector	14	2.2
	3_{10}–α connector	16	1.2	3_{10}–α connector	13	2.1
	β–3_{10} connector	22	1.7	β–3_{10} connector	10	1.6
	3_{10}–β connector	24	1.8	3_{10}–β connector	11	1.7
	3_{10}–3_{10} connector	1	0.1	3_{10}–3_{10} connector	5	0.8

The overall percentage values in two categories (a) and (b) (for example, 38.6 and 61.4 for three-length 3_{10}-helix) are with respect to the total number (2140), whereas the individual values within the two categories are relative to the total numbers (826 and 1314) in them. The two categories are distinguished as follows: (a), when any (or both) of Nc and Cc positions are part of neighboring helix or β-strand (in which a β-bridge with DSSP notation "B" is also included); (b) contains the rest; under (b), a few sub-categories have been identified as connectors when two secondary structural elements occur within four flanking positions (N1′ → N4′ and C1′ → C4′) of the helix.

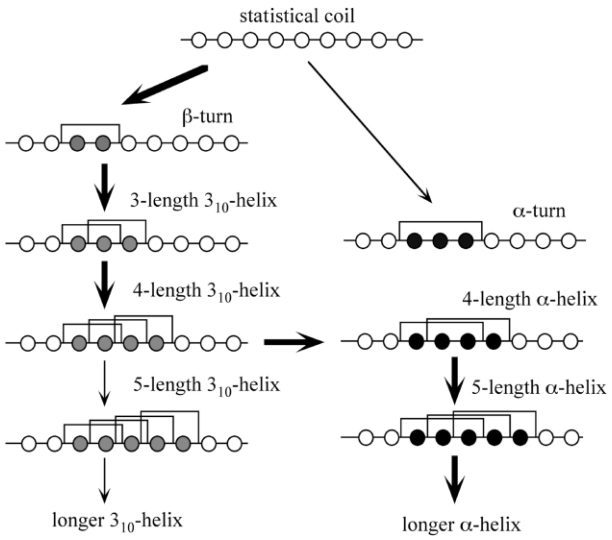

Figure 11. A schematic description of proposed helix nucleation and propagation scheme in proteins. Shaded circles indicate helical backbone dihedral angles (dark: α- and light: 3_{10}-) and lines connecting non-contiguous residues indicates backbone hydrogen bonds. Arrows indicate the proposed pathways to longer helices (thicker the arrow, stronger the pathway).

variants, see Pal *et al.*[20] only ~20% occur as α-helix termini (Table 4), which is ~50% of all three-length 3_{10}-helices connected contiguously to another secondary structure; for the rest, the contiguous secondary structure is a β-strand or a β-bridge. More importantly, more than 60% of total three-length 3_{10}-helices occur independent of any other contiguous secondary structure. Occurring as an α-helix terminus, a three-length 3_{10}-helix can provide a unique geometry for the terminal helix turn. Occurring in isolation, often three-length 3_{10}--helices participate in reversal of peptide chain direction, including participation in established motifs that induce such chain reversal (4:4 β-hairpin motif in Figure 9). Collective experimental evidence, from several proteins, also point out that short 3_{10}-helices may play an important role in the overall folding and stability[56–58] or function[59–61] of proteins.

Conclusion

In conclusion, an analysis of the shortest helices in protein structures has provided several useful insights into protein structure and folding. Far from being a regular structure, the DNA helix is known to be capable of accommodating a

large number of distortions, which have been parameterized.[62] Here we show that one of the counterparts in protein structures, viz., the 3_{10}-helix, can have a significant deviation in the backbone torsion angle at the last position (or the first position, in some 3_{10}-helix variants,[20] this being an interesting manifestation of the interplay of two types of interactions, electrostatics and hydrogen bonding. The distortion can explain the kink that is observed when a α-helix follows a 3_{10}-helix and the information should be useful in protein modeling. The similarity in the residue composition of the shortest helices with type I β-turn and the N-terminal end of regular helices suggests that turns and helices may be nucleated in an analogous fashion during the folding, and depending on the residues that follow, a turn can exist as such or extend to a helix.

Finally, using clues from site–site correlation coefficients of amino acid composition and the frequency of occurrence in the database, we propose an important role of three and four-length 3_{10}-helices in α-helix nucleation. At variance from the general notion that α-turns nucleate α-helices, our model (Figure 11) stipulates formation of a β-turn to be the helix nucleation event; this is then followed by sequential formation of a three and four-length 3_{10}-helix. The shortest α-helix (four-length) is formed from the four-length 3_{10}-helix, which then can propagate by adding α-turns.

Materials and Methods

The culledpdb data set (March, 1999 version),[63,64] containing 1085 protein chains with less than 30% sequence identity and ≤2.5 Å resolution, was used in this analysis. Secondary structure assignments were made using the Define Secondary Structure of Proteins (DSSP) program.[65] A Protein Data Bank (PDB) file is mentioned in the text as the four-letter PDB code (in lowercase letters) followed by the chain identifier (in uppercase letters); when there is no chain ID an underscore (_) is used.

The nomenclature for the three-length 3_{10}-helices and their flanking residues is as follows:

$$\cdots N2'-N1'-Nc-N1/C3-N2/C2-N3/C1-Cc-C1'-C2'\cdots$$

where N1/C3 through N3/C1 belong to helix proper, Nc and Cc represent helix capping positions, while the primed residues represent residues preceding and succeeding the helix. A position within the shortest helix can be labeled counting from the N-terminal or the C-terminal end. Thus the first position in the 3_{10}-helix is N1 or C3 and the label N1/C3 is used to designate it. However, for brevity, N1 is also used synonymously in the text. The labels in α-helix are given in a similar way and shown in Figure 1.

As in Ref. 5, turns were characterized as four consecutive residues (i to $i+3$) with the distance between C^{α}_{i} and $C^{\alpha}_{i+3} < 7$ Å, and where the central residues were not helical. To constitute a type I β-turn the ϕ,ψ angles of residues $i+1$ and $i+2$ had to be within 30° of the ideal values ((−60°, −30°) and (−90°,0°)), with one angle being allowed to deviate by 45°. In addition to hydrogen-bonded β-turns (for which the two central residues would normally appear with designation TT in DSSP output) this definition would also include non-hydrogen bonded turns. Two consecutive turns were considered to be part of multiple turns when they had a common central residue; such turns were excluded to retain only the isolated β-turns.

The propensity of a given residue to occur at a specific position (say, helix Nc) of a secondary structure was calculated as the ratio of the actual number of observations to the expected number of observations, where the expected number of observations is given by $(n_{\mathrm{xaa}}/n_{\mathrm{total}}) \times n^{*}$, where n_{xaa} is the total number of residue type Xaa in the data set, n_{total} is total number of all residues in the data set, and n^{*} is the total number of residues in the data set at that particular position. In calculating amino acid composition, over-representation was estimated as described.[18] For each amino acid at a specific helix position, a z-value was calculated. If $|z| \geq 1.96$ (5% significance level), the observed number of occurrences was considered to deviate significantly from its expected value. Negative values of z indicate under-representation and positive values indicate over-representation.

Two hydrophobic residues were considered to be making a contact when the distance between any two atoms of either residue was less than or equal to the sum of their van der Waals radii plus the diameter of a water molecule, 2.8 Å.[17] In addition to the typical hydrophobic residues (Ala, Val, Ile, Leu, Met, Phe, Trp, Cys, Tyr, Pro and His) Lys and Arg are also considered hydrophobic as most of their contacts within the helix involve non-polar atoms. Hydrogen bonded partners were assigned using the program HBPLUS.[66] Molecular diagrams were made using MOLSCRIPT.[67]

Acknowledgements

The financial support for the work was provided by the Council of Scientific and Industrial Research, India. Additional support was provided by the special coordination fund promoting science and technology from ministry of education, culture, sports, science and technology, Japan.

References

1. Nataraj, D. V., Srinivasan, N., Sowdhamini, R. & Ramakrishnan, C. (1995). α-Turns in protein structures. *Curr. Sci.* **69**, 434–447.
2. Venkatachalam, C. M. (1968). Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers*, **6**, 1425–1436.
3. Lewis, P. N., Momany, F. A. & Scheraga, H. A. (1973). Chain reversals in proteins. *Biochim. Biophys. Acta*, **303**, 211–229.
4. Rose, G. D., Gierasch, L. M. & Smith, J. A. (1985). Turns in peptides and proteins. *Advan. Protein Chem.* **37**, 1–109.
5. Hutchinson, E. G. & Thornton, J. M. (1994). A revised set of potentials for β-turn formation in proteins. *Protein Sci.* **3**, 2207–2216.

6. Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Advan. Protein Chem.* **34**, 167–339.
7. Lim, V. I. (1974). Algorithms for prediction of α-helical and β-structural regions of globular proteins. *J. Mol. Biol.* **88**, 873–894.
8. Huang, E. S., Subbiah, S. & Levitt, M. (1995). Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.* **252**, 709–720.
9. Sun, S. J., Thomas, P. D. & Dill, K. A. (1995). A simple protein-folding algorithm using a binary code and secondary structure constraints. *Protein Eng.* **8**, 769–778.
10. West, M. W. & Hecht, M. H. (1995). Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. *Protein Sci.* **4**, 2032–2039.
11. Richardson, J. S. & Richardson, D. C. (1988). Amino acid preferences for specific locations at the ends of α-helices. *Science*, **240**, 1648–1652.
12. Presta, L. G. & Rose, G. D. (1988). Helix signals in proteins. *Science*, **240**, 1632–1641.
13. Kumar, S. & Bansal, M. (1998). Dissecting α-helices: position-specific analysis of α-helices in globular proteins. *Proteins*, **31**, 460–476.
14. Petukhov, M., Muñoz, V., Yumoto, N., Yoshikawa, S. & Serrano, L. (1998). Position dependence of nonpolar amino acid intrinsic helical propensities. *J. Mol. Biol.* **278**, 279–289.
15. Penel, S., Hughes, E. & Doig, A. J. (1999). Side-chain structures in the first turn of the α-helix. *J. Mol. Biol.* **287**, 127–143.
16. Penel, S., Morrison, R. G., Mortishire-Smith, R. J. & Doig, A. J. (1999). Periodicity in α-helix lengths and C-capping preferences. *J. Mol. Biol.* **293**, 1211–1219.
17. Aurora, R. & Rose, G. D. (1998). Helix capping. *Protein Sci.* **7**, 21–38.
18. Karpen, M. E., De Haseth, P. L. & Neet, K. E. (1992). Differences in the amino acid distributions of 3_{10}-helices and α-helices. *Protein Sci.* **1**, 1333–1342.
19. Chakrabarti, P. & Pal, D. (2001). The interrelationships of side-chain and main-chain conformations in proteins. *Prog. Biophys. Mol. Biol.* **76**, 1–102.
20. Pal, L., Basu, G. & Chakrabarti, P. (2002). Variants of 3_{10}-helices in proteins. *Proteins*, **48**, 571–579.
21. Pal, L. & Basu, G. (1999). Novel protein structural motifs containing two-turn and longer 3_{10}-helices. *Protein Eng.* **12**, 811–814.
22. Doig, A. J., MacArthur, M. W., Stapley, B. J. & Thornton, J. M. (1997). Structures of N-termini of helices in proteins. *Protein Sci.* **6**, 147–155.
23. Dasgupta, S. & Bell, J. A. (1993). Design of helix ends. Amino acid preferences, hydrogen bonding and electrostatic interactions. *Int. J. Peptide Res.* **41**, 499–511.
24. Harper, E. T. & Rose, G. D. (1993). Helix stop signals in proteins and peptides: the capping box. *Biochemistry*, **32**, 7605–7609.
25. Seale, J. W., Srinivasan, R. & Rose, G. D. (1994). Sequence determinants of the capping box, a stabilizing motif at the N-termini of alpha-helices. *Protein Sci.* **3**, 1741–1745.
26. Schellman, C. (1980). The alpha-L conformation at the ends of helices. In *Protein Folding* (Jaenicke, R., ed.), pp. 53–61, Elsevier/North Holland, New York.
27. Prieto, J. & Serrano, L. (1997). C-capping and helix stability: the Pro C-capping motif. *J. Mol. Biol.* **274**, 276–278.
28. Sibanda, B. L., Blundell, T. L. & Thornton, J. M. (1989). Conformation of β-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. *J. Mol. Biol.* **206**, 759–777.
29. Baker, E. N. & Hubbard, R. E. (1984). Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **44**, 97–179.
30. Schulz, G. E. & Schirmer, R. H. (1984). *Principles of Protein Structure*, Springer-Verlag, New York.
31. Blundell, T., Barlow, D., Borkakoti, N. & Thornton, J. (1983). Solvent-induced distortions and the curvature of α-helices. *Nature*, **306**, 281–283.
32. Barlow, D. J. & Thornton, J. M. (1988). Helix geometry in proteins. *J. Mol. Biol.* **201**, 601–619.
33. Woolfson, D. N. & Williams, D. H. (1990). The influence of proline residues on α-helical structures. *FEBS Letters*, **277**, 185–188.
34. Kumar, S. & Bansal, M. (1996). Structural and sequence characteristics of long α-helices in globular proteins. *Biophys. J.* **71**, 1574–1586.
35. Chakrabarti, P. & Chakrabarti, S. (1998). C–H···O hydrogen bond involving proline residues in α-helices. *J. Mol. Biol.* **284**, 867–873.
36. Chou, P. Y. & Fasman, G. D. (1974). Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry*, **13**, 211–222.
37. Baldwin, R. L. & Rose, G. D. (1999). Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem. Sci.* **24**, 26–33.
38. Soman, K. V., Karimi, A. & Case, D. (1991). Unfolding of an α-helix in water. *Biopolymers*, **31**, 1351–1361.
39. Tirado-Rives, J. & Jorgensen, W. L. (1991). Molecular dynamics simulation of the unfolding of an α helical analog of ribonuclease A S-peptide in water. *Biochemistry*, **30**, 3864–3871.
40. Daggett, V. & Levitt, M. (1992). Molecular-dynamics simulations of helix denaturation. *J. Mol. Biol.* **223**, 1121–1138.
41. Hummer, G., Garcia, A. E. & Garde, S. (2001). Helix nucleation kinetics from molecular simulations in explicit solvent. *Proteins: Struct. Funct. Genet.* **42**, 77–84.
42. Basu, G., Kitao, A., Hirata, F. & Go, N. (1994). A collective motion description of the 3_{10}-/α-helix transition: implications for a natural reaction coordinate. *J. Am. Chem. Soc.* **116**, 6307–6315.
43. Huo, S. & Straub, J. E. (1999). Direct computation of long time processes in peptides and proteins: reaction path study of the coil-to-helix transition in polyalanine. *Proteins: Struct. Funct. Genet.* **36**, 249–261.
44. Zimm, B. H. & Bragg, J. K. (1959). Theory of phase transition between helix and random coil in polypeptide chains. *J. Chem. Phys.* **31**, 526–535.
45. Lifson, S. & Roig, A. (1961). On the theory of helix-coil transition in polypeptides. *J. Chem. Phys.* **34**, 1963–1974.
46. Basu, G. & Kuki, A. (1992). Conformational preferences of oligopeptides rich in alpha-aminoisobutyric acid. II. A model for the 3_{10}-/alpha-helix transition with composition and sequence sensitivity. *Biopolymers*, **32**, 61–71.
47. Rohl, C. A. & Doig, A. J. (1996). Models for the 3_{10}-helix/coil, pi-helix/coil, and alpha-helix/3_{10}-helix/coil transitions in isolated peptides. *Protein Sci.* **5**, 1687–1696.

48. Sheinerman, F. B. & Brooks, C. L., III (1995). 3_{10}-helices in peptides and proteins as studied by modified Zimm-Bragg theory. *J. Am. Chem. Soc.* **117**, 10098–10103.
49. Sundaralingam, M. & Sekharudu, Y. C. (1989). Water-inserted α-helical segments implicate reverse turns as folding intermediates. *Science*, **244**, 1333–1337.
50. Millhauser, G. L. (1995). Views of helical peptides: a proposal for the position of 3_{10}-helix along the thermodynamic folding pathway. *Biochemistry*, **34**, 3873–3877.
51. Karle, I. L. & Balaram, P. (1990). Structural characteristics of alpha-helical molecules containing Aib residues. *Biochemistry*, **29**, 6747–6756.
52. Pavone, V., Benedetti, E., Di Blasio, B., Pedone, C., Santini, A., Bavoso, A. *et al.* (1990). Critical main-chain length for conformational conversion from 3(10)-helix to alpha-helix in polypeptides. *J. Biomol. Struct. Dyn.* **7**, 1321–1331.
53. Bolin, K. A. & Millhauser, G. L. (1999). α and 3_{10}: the split personality of polypeptide helices. *Acc. Chem. Res.* **32**, 1027–1033.
54. Scholtz, J. M., Qian, H., York, E. J., Stewart, J. M. & Baldwin, R. L. (1991). Parameters of helix-coil transition theory for alanine-based peptides of varying chain lengths in water. *Biopolymers*, **31**, 1463–1470.
55. Dyson, H. J. & Wright, P. E. (1998). Equilibrium NMR studies of unfolded and partially folded proteins. *Nature Struct. Biol.* **5**, 499–503.
56. Worthylake, D. K., Wang, H., Yoo, S., Sundquist, W. I. & Hill, C. P. (1999). Structures of the HIV-1 capsid protein dimerization domain at 2.6 Å resolution. *Acta Crystallog. sect. D*, **55**, 85–92.
57. Martinez, J. C. & Serrano, L. (1999). The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nature Struct. Biol.* **6**, 1010–1016.
58. Demarest, S. J., Boice, J. A., Fairman, R. & Raleigh, D. P. (1999). Defining the core structure of the alpha-lactalbumin molten globule state. *J. Mol. Biol.* **294**, 213–221.
59. Xu, R. M., Jokhan, L., Cheng, X., Mayeda, A. & Krainer, A. R. (1997). Crystal structure of human UP1, the domain of hnRNP A1 that contains two RNA-recognition motifs. *Structure*, **5**, 559–570.
60. Ventura, S., Villegas, V., Sterner, J., Larson, J., Vendrell, J., Hershberger, C. L. & Aviles, F. X. (1999). Mapping the pro-region of carboxypeptidase B by protein engineering. Cloning, overexpression, and mutagenesis of the porcine proenzyme. *J. Biol. Chem.* **274**, 19925–19933.
61. Hashimoto, Y., Kohri, K., Kaneko, Y., Morisaki, H., Kato, T., Ikeda, K. & Nakanishi, M. (1998). Critical role for the 3_{10} helix region of p57(Kip2) in cyclin-dependent kinase 2 inhibition and growth suppression. *J. Biol. Chem.* **273**, 16544–16550.
62. Dickerson, R. E. (1983). The DNA helix and how it is read. *Sci. Am.* **249**, 94–111.
63. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
64. Hobohm, U., Scharf, M. & Schneider, R. (1993). Selection of representative protein data sets. *Protein Sci.* **1**, 409–417.
65. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
66. McDonald, I. K. & Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777–793.
67. Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946–950.

Edited by F. E. Cohen

(Received 24 June 2002; received in revised form 30 September 2002; accepted 12 November 2002)