PROTEINS: Structure, Function, and Genetics 31:460–476 (1998)

Dissecting α-Helices: Position-Specific Analysis of α-Helices in Globular Proteins

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

ABSTRACT **An analysis of the amino acid distributions at 15 positions, viz., N″, N′, Ncap, N1, N2, N3, N4, Mid, C4, C3, C2, C1, Ccap, C′, and C″ in 1,131 α-helices reveals that each position has its own unique characteristics. In general, natural helix sequences optimize by identifying the residues to be avoided at a given position and minimizing the occurrence of these avoided residues rather than by maximizing the preferred residues at various positions. Ncap is most selective in its choice of residues, with six amino acids (S, D, T, N, G, and P) being preferred at this position and another 11 (V, I-,F, A, K, L, Y, R, E, M, and Q) being strongly avoided. Ser, Asp, and Thr are all more preferred at Ncap position than Asn, whose role at helix N-terminus has been highlighted by earlier analyses. Furthermore, Asn is also found to be almost equally preferred at helix C-terminus and a novel structural motif is identified, involving a hydrogen bond formed by $N^{\delta 2}$ of Asn at Ccap or C1 position, with the backbone carbonyl oxygen four residues inside the helix. His also forms a similar motif at the C-terminus. Pro is the most avoided residue in the main body (N4 to C4 positions) and at C-terminus, including Ccap of an α-helix. In 1,131 α-helices, no helix contains Pro at C3 or C2 positions. However, Pro is highly favoured at N1 and C′. The doublet X-Pro, with Pro at C′ position and extended backbone conformation for the X residue at Ccap, appears to be a common structural motif for termination of α-helices, in addition to the Schellman motif. Main body of the helix shows a high preference for aliphatic residues Ala, Leu, Val, and Ile, while these are avoided at helix termini. A propensity scale for amino acids to occur in the middle of helices has been obtained. Comparison of this scale with several previously reported scales shows that this scale correlates best with the experimentally determined values. Proteins 31:460–476, 1998.**

Key words: ***α*-helix; sequence; structure; database; amino acid; secondary structure**

INTRODUCTION

One of the characteristic features of α-helices is the intrachain $5 \rightarrow 1 > \text{N–H} \ldots \text{O} = \text{C} <$ hydrogen bonds. A typical α-helix contains four residues at either end, which lack hydrogen bonding partners.[1] Besides, every α-helix has a net dipole moment, with N-terminus being positively charged and C-terminus negatively charged.[2] Charged residues are found to occur preferentially at either end of the helices and this preference has been explained on the basis of helix stabilizing, charge-helix dipole interaction.[2–8] These characteristics make various positions within and around an α-helix nonequivalent with respect to the amino acid distribution.[9,10] Similar observations have been made by substitution experiments in helices in proteins[11–14] as well as in peptide helices.[15–21] Two types of motifs that play important structural roles at the helix termini have been described in recent literature. These motifs are the capping box or reciprocal hydrogen bond motif,[1,17,22–25] a hydrogen-bonding-based motif found at the N-terminus of helices, and the hydrophobic staple motif, also at the N-terminus of α-helices.[26–28] Recently, specific rotamer preferences for sidechains of the residues favored at the N-termini of α-helices have been observed.[29]

Here, we report a comprehensive and systematic positionwise dissection of α-helix sequences found in globular proteins. We obtained a database of 1,131 α-helices with lengths of nine or more residues and nonidentical sequences, from 205 nonhomologous globular protein chains, in protein crystal structures available from Brookhaven Protein Data Bank (PDB).[30] This database has been used to analyze sequence–structural characteristics at 15 positions within and around an α-helix, namely, N″, N′, Ncap, N1, N2, N3, N4, Mid, C4, C3, C2, C1, Ccap, C′, and C″. We asked the following questions: Do each of these 15 positions show unique features? Alternatively, can they be grouped into a few classes like helix-neighboring, helix termini, and helix main body positions that may suffice for complete understanding of α-helices? Which amino acids are preferred, avoided, or neutral at each positions? Do

*Correspondence to: Manju Bansal, Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India.

Received 11 August 1997; 23 December 1997

some positions have more stringent sequence preferences than others, perhaps reflecting their important role in helix formation or stability? Do the residues preferred at the helix termini position form structural motifs to stabilize the helix? If yes, what are the backbone conformation, sidechain rotamer, and sequence preferences of the residues involved in the structural motifs at the helix termini? Our analysis is able to provide answers to most of these questions. In addition, selection of a large database with statistically sound criteria, comprehensive, self-consistent analysis, and use of several parameters to study a given property helps to avoid many pitfalls. Our results are largely free of small database biases as well as parameter choices. The lessons learnt from this study could have important implications both for protein secondary-structure prediction and de novo protein design.

MATERIALS AND METHODS

Composition of Database

We have used the June 1995 list of nonhomologous (sequence identity ≤25%) protein chains compiled by Hobohm and Sander[31,32] to select a subset of 205 nonhomologous globular protein chains whose crystal structures have been solved to resolution of 2.5 Å or better. These protein chains, listed in Table I by their PDB[30] codes, contain a total of 53,238 amino acid residues, with 15,677 amino acids in 1,131 α-helices, of lengths of nine or more amino acid residues and nonidentical sequence.

Helix Position Nomenclature

For each α-helix we have considered 15 positions, within and around it. They are N″, N′, Ncap, N1, N2, N3, N4, Mid, C4, C3, C2, C1, Ccap, C′, and C″. Nine positions, from N1 to C1, constitute an α-helix. Ncap and Ccap are interface positions between helical and nonhelical regions at N and C terminus of the helix, respectively. N″, N′ and C′, C″ are the positions in the nonhelical regions preceding and succeeding the helix. Unlike other positions, Mid contains one or more residues per helix because it represents the whole middle region of an α-helix. Thus, Mid position corresponds to (N − 8) residues, where N is the helix length (number of residues in the helix).

Definition of Helix Boundaries

Initially, protein chain segments of length nine residues or more and defined as helices in the Dictionary of Protein Secondary Structure (DSSP)[33] for each of the 205 protein chains were taken. Boundaries of these helices were checked and, if necessary, reassigned using the following two criteria: one, distance $|O_i \ldots N_{i+4}| \leq 3.5$ Å at the helix termini; and two, angles between successive local helix axes at the helix termini are $<20°$.[34]

TABLE I. List of PDB Filenames of 205 Nonhomologous Globular Protein Chains Whose Structures Have Been Solved to a Resolution of 2.5 Å or Better*

119L	1GHSA	1PGB	2CTC
153L	1GKY	1PHP	2CTS
1AAK	1GLT	1PII	2DNJA
1ABK	1GMFA	1PLQ	2EBN
1ADD	1GOX	1PMY	2END
1ADS	1GP1A	1PNT	2GSTA
1ALKA	1GTRA	1POA	2HBG
1AMP	1HDCA	1POC	2HHMA
1AORA	1HDGO	1POXA	2HMZA
1APME	1HEX	1PPN	2HSC
1ARS	1HFC	1PPT	2IHL
1ASH	1HJRA	1PYAB	2LIV
1AYAA	1HLB	1RBLM	2MGE
1BABB	1HLEA	1RCB	2MNR
1BBPA	1HMY	1REC	2MTAC
1BGEB	1HSLA	1RIBA	2OHXA
1BMDA	1HTP	1ROPA	2PGD
1BPB	1HUCB	1RTP1	2PIA
1BSAA	1HUW	1RVAA	2REB
1CBN	1HVD	1SO1	2RN2
1CCR	1IAE	1SBP	2RSLB
1CDE	1IAG	1SCUA	2SAS
1CDG	1IGP	1SCUB	2SCPA
1CEWI	1ISCA	1SESA	2TGI
1CHMA	1KAB	1SMRA	2TMDA
1CMBA	1LBA	1TADA	2TPRA
1CPCA	1LDM	1TCA	2TS1
1CPCB	1LGAA	1TGSI	3AAHB
1CRL	1LIS	1THV	3CHY
1CSEI	1LKI	1TML	3CLA
1CTN	1LPBB	1TPH1	3COX
1CTT	1LPE	1TPLA	3DFR
1CUS	1MAT	1TRB	3GAPA
1DHR	1MINA	1TRKA	3GLY
1DSBA	1MINB	1TYS	3IL8
1DTS	1MMOB	1WHTA	3MDDA
1DYNA	1MMOG	1WHTB	3SDHA
1ECA	1MRG	1WSYA	3SGBI
1EDE	1MUP	1WSYB	4BLMA
1EFT	1MYLB	1XNB	4ENL
1ENH	1NAR	1YPTA	4GPB
1EPAB	1NBAA	1YTBA	4XIAA
1ERB	1NHKL	1ZAAC	5P21
1FBAA	1OLBA	2AK3A	7CCP
1FHA	1OMP	2AZAA	7RSA
1FIAB	1OVB	2CCYA	8ABP
1FNC	1OXY	2CDV	8ACN
1FRPA	1OYB	2CHSA	8ATCA
1GAL	1PBE	2CP4	8CATA
1GCA	1PBP	2CPL	8TLNE
1GDHA	1PFKA	2CRO	9RNT
1GDM			

*First four characters indicate PDB filename and the fifth character is the protein chain.

Statistical Methods

Distribution and frequency of occurrence (%) of individual amino acids were computed for the 205 nonhomologous globular protein chains, 1,131

TABLE II. Amino Acid Distribution at 15 Positions* Within and Around the 1,131 α-Helices From 205 Nonhomologous Protein Chains

Amino																Full helix
acid	N″	N′	Ncap	N1	N2	N3	N4	Mid	C4	C3	C2	C1	Ccap	C′	C″	(N1–C1)
A	82	82	45	116	109	114	145	936	134	175	149	158	113	68	60	2036
C	18	13	14	13	8	8	8	77	16	13	11	6	15	14	7	160
D	79	74	178	62	116	97	28	299	45	38	47	49	58	49	92	781
E	56	59	43	101	174	145	33	477	69	71	104	105	60	51	65	1279
F	43	41	16	38	32	46	67	253	58	46	39	42	41	33	47	621
G	141	104	119	80	79	66	46	225	31	24	19	26	226	214	97	596
H	27	24	23	15	26	37	21	115	19	29	28	47	46	29	32	337
I	50	63	21	51	35	40	106	431	85	62	70	41	31	30	57	921
K	58	59	31	68	57	50	61	458	76	92	127	107	77	85	87	1096
L	61	118	49	91	59	77	162	761	156	199	132	118	79	81	82	1755
M	17	28	15	16	12	17	42	201	44	45	21	19	23	9	21	417
N	57	60	99	33	49	34	22	234	35	39	31	60	90	67	62	537
P	68	53	63	130	54	33	1	28	1	0	0	2	2	113	76	249
Q	33	27	31	28	49	85	56	339	47	56	63	51	47	43	48	798
R	34	31	28	51	46	33	78	408	70	56	79	68	43	48	51	889
S	79	75	171	51	73	63	37	276	53	41	42	68	57	59	50	704
T	63	77	137	55	55	72	37	318	41	42	53	58	41	45	52	731
V	70	67	17	56	49	71	116	472	78	51	67	45	30	40	62	1005
W	15	18	7	19	17	11	17	90	28	16	18	13	7	13	10	229
Y	49	37	22	33	32	32	48	231	45	36	31	48	44	22	35	536
Total	1100	1110	1129	1131	1131	1131	1131	6629	1131	1131	1131	1131	1130	1106	1083	15677

*The total number of amino acids at the nonhelical positions before the start and after the end of the helices, i.e., at N″, N′, Ncap, Ccap, C′, and C″, is <1,131 because some helices start or end very near to the beginning or end of the protein chain. In several cases, the electron densities for the residues at these positions were disordered in the protein crystal structures and no residues were assigned to these positions. The total number of amino acids at the mid position in helices is >1,131 because the mid position can contain more than one amino acid residue per helix. The number of amino acids in the mid position of a helix is N − 8, where N is the number of residues in the helices (N1–C1).

α-helices, and each of the 15 positions within and around these α-helices.

Propensity

Propensity (P_{ij}) of amino acid i to occur at position j and its standard deviation (sd_{ij}) was calculated using the formula[35]

$$P_{ij} = \frac{n_{ij}/n_i}{N_j/N} \qquad (1)$$

$$sd_{ij} = \frac{\mathrm{sqrt}(f_{ij}(1 - f_{ij})/n_i)}{N_j/N} \qquad (2)$$

where n_{ij} = number of ith amino acid at jth position; n_i = total number of ith amino acids in the 205 protein chains; N_j = total number of amino acids at the jth position; N = total number of amino acids in 205 protein chains; $f_{ij} = n_{ij}/n_i$; i = 1, 2 . . . 20 (20 amino acids); and j = 1, 2 . . . 15 (15 positions N″ to C″).

Amino acid distribution in the 1,131 α helices was taken as reference distribution for calculating preference, χ^2 values and the change in proportions of individual amino acids at the 15 positions.

TABLE III. χ^2 and Euclidean and Hamming Distances for Amino Acid Compositions at 15 Positions*

		Euclidean	Hamming
Position	χ^2 value	distance	distance
N″	535.9	14.5	46.9
N′	280.8	10.9	35.0
Ncap	1344.5	24.7	90.7
N1	774.1	11.7	30.0
N2	348.4	13.3	44.6
N3	187.6	10.1	36.2
N4	170.5	9.4	31.5
Mid	102.6	2.4	8.2
C4	69.0	5.2	19.6
C3	106.7	8.3	24.1
C2	77.1	5.6	17.1
C1	105.2	6.2	23.5
Ccap	983.6	18.9	48.7
C′	1388.7	20.7	58.5
C″	405.2	12.5	38.0

*For a distribution with 19 degrees of freedom, null hypothesis is rejected at 95% level of confidence ($P < 0.05$) if $\chi^2 > 30.14$.

Preference

Preference for the ith amino acid to occur at the jth position was calculated using the formula[10]

$$\mathrm{Pref}_{ij} = n_{ij}/\mathrm{nexpec}_{ij} \qquad (3)$$

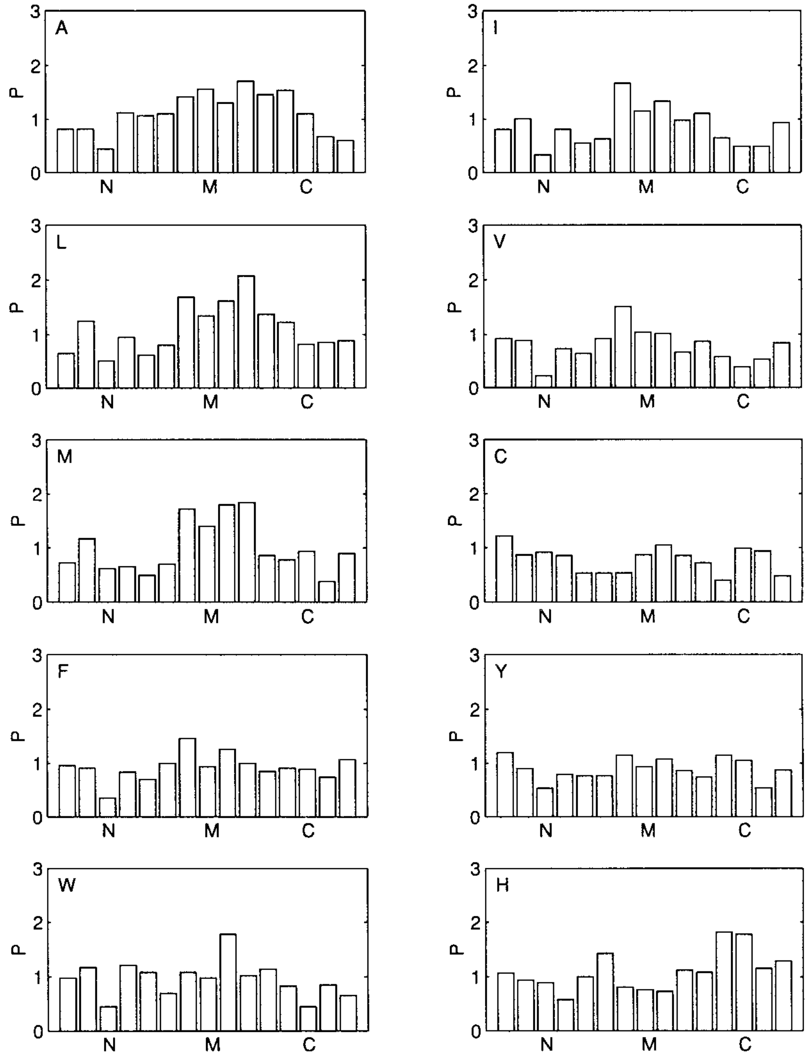


Fig. 1. Position-dependent variation of amino acid propensities P at each of the 15 positions within and around α-helices, viz., N″, N′, Ncap, N1, N2, N3, N4, Mid, C4, C3, C2, C1, Ccap, C′, and C″ in α-helices. Only the Ncap, Mid, and Ccap positions in the helix have been indicated along the X-axis by symbols N, M, and C, respectively. Amino acids are denoted by single-letter code in the upper left corner of each box.

where n_{ij} = observed number of ith amino acid at jth position; and $nexpec_{ij}$ = expected number of ith amino acid at the jth position = Y_iN_j/R. So

$$\mathrm{Pref}_{ij} = \frac{n_{ij}/Y_i}{N_j/R} \qquad (4)$$

where r_i = number of ith amino acids in the reference distribution; N_j = total number of amino acids at the jth position; and R = total number of amino acids in the reference distribution.

χ² Values

χ^2 values at the jth position were calculated as follows:

$$\chi_j^2 = \Sigma_{i=1,20}\,(n_{ij} - nexpec_{ij})^2/nexpec_{ij} \qquad (5)$$

where $nexpec_{ij}$ is the expected number calculated as given above.

For a 19-parameter system such as amino acid distribution in different classes, χ^2 value at 95% level of confidence (probability of accepting the null hypothesis, $P < 0.05$) should be greater than 30.14 to reject the null hypothesis. Null hypothesis is rejected at 99.99% level of confidence, if the value of χ^2 is greater than 43.82.

Differences in amino acid compositions (frequencies of occurrence as %) of individual amino acids at various positions were geometrically interpreted by calculating Euclidean and Hamming distances in 20-dimensional amino acid composition space. The amino acid composition of 1,131 α-helices was taken as origin. Similar calculations were performed to compute these distances among the various positions.

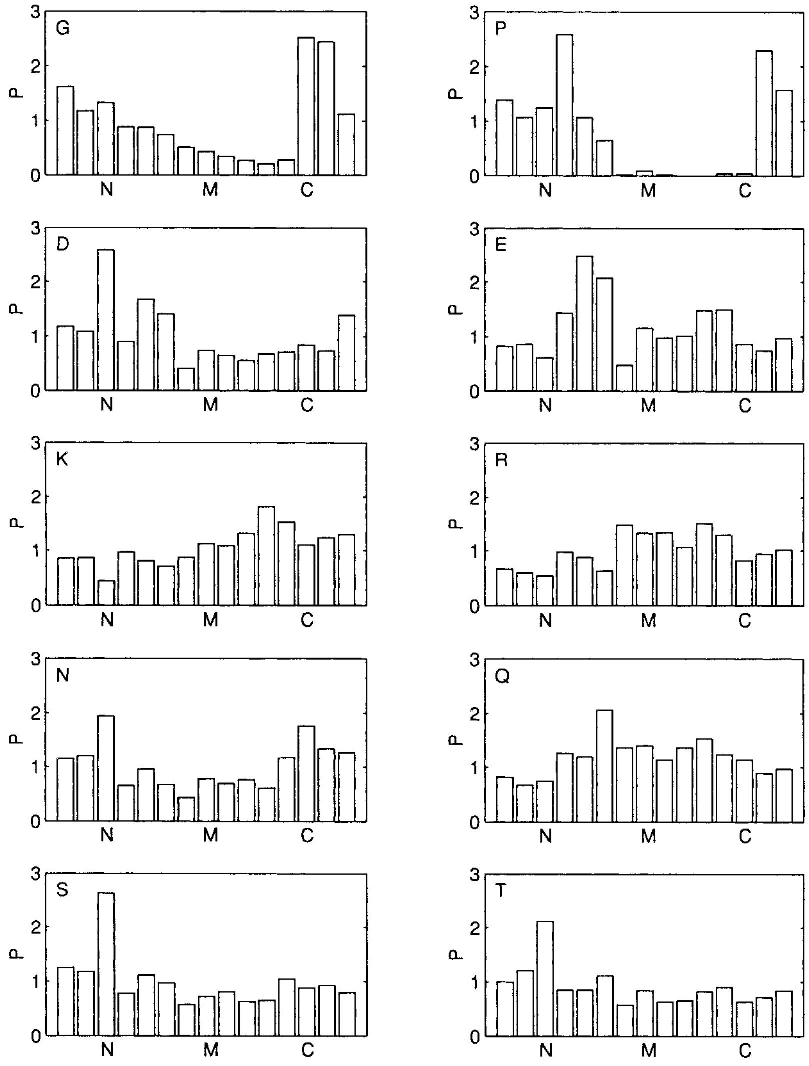


Figure 1. (Continued)

Hamming distance

Hamming distance of the amino acid composition at the jth position was computed by the formula[36]

$$D_j^H = \Sigma_{t=1,20} |x_{ij} - xr_i| \qquad (6)$$

Euclidean distance

Euclidean distance of the amino acid composition at the jth position was computed by the formula[36]

$$D_j^E = (\Sigma_{i=1,20} (x_{ij} - xr_i)^2)^{1/2} \qquad (7)$$

where x_{ij} = frequency of occurrence (%) of the ith amino acid at the jth position; and xr_i = frequency of occurrence (%) of the ith amino acid in the reference distribution (origin).

Change in proportion

Change in proportion of ith amino acid at the jth position was considered to be significant at 95% confidence level if it is greater than twice the estimated standard deviation[37]

$$(prop_{ij} - propr_i) > 2\ \sigma_{ij} \qquad (8)$$

where $prop_{ij}$ = proportion of ith amino acid at jth position; and $propr_i$ = proportion of the ith amino acid in reference distribution.

σ_{ij}, the estimated standard deviation for the ith amino acid at the jth position, is defined as

$$\sigma_{ij} = sqrt(propav_i(1 - propav_i)(1/N_j + 1/R))$$

where $propav_i$ is the average proportion of i^{th} amino acid and is defined as $propav_i = (n_{ij} + r_i/N_j + R)$.

TABLE IV. Residues That Show a Significant Change in Proportion at the 15 Positions Within and Around α-Helices*

N″	N′	Ncap	N1	N2	N3	N4	Mid	C4	C3	C2	C1	Ccap	C′	C″
Proportion increases														
G (1.62)	**T** (1.21)	**S** (2.63)	**P** (2.58)	**E** (2.48)	**E** (2.07)	**M** (1.72)	A (1.56)	M (1.80)	**L** (2.07)	**K** (1.82)	**H** (1.82)	**G** (2.52)	**G** (2.44)	**P** (1.57)
P (1.39)	**N** (1.20)	**D** (2.58)		**D** (1.68)	**Q** (2.06)	**L** (1.69)		W (1.78)	M (1.84)		**K** (1.53)	**H** (1.78)	**P** (2.29)	**D** (1.39)
S (1.25)	**G** (1.18)	**T** (2.12)		S (1.12)	H (1.42)	**I** (1.68)		I (1.34)	A (1.71)		**N** (1.17)	**N** (1.76)	**N** (1.34)	**N** (1.27)
D (1.18)	**S** (1.18)	**N** (1.94)		**P** (1.07)	**D** (1.41)	**V** (1.51)					S (1.05)			**G** (1.13)
N (1.15)	**D** (1.09)	**G** (1.33)			T (1.11)	**F** (1.46)								
	P (1.07)	**P** (1.25)			**G** (0.74)									
					P (0.62)									
Proportion decreases														
K (0.85)	K (0.86)	**Q** (0.75)	A (1.13)	**A** (1.07)	**L** (0.80)	K (0.87)	E (1.16)	E (0.98)	E (1.01)	**G** (0.21)	**I** (0.65)	A (1.11)	**Q** (0.89)	**E** (0.97)
Q (0.82)	**E** (0.86)	M (0.62)	**L** (0.95)	R (0.88)	**K** (0.72)	**T** (0.57)	**P** (0.10)	**P** (0.02)	V (0.66)	**P** (0.00)	**V** (0.58)	**E** (0.86)	**L** (0.86)	**Q** (0.96)
A (0.82)	**A** (0.82)	**E** (0.61)	**M** (0.66)	K (0.82)	M (0.69)	**E** (0.47)			D (0.55)		G (0.29)	R (0.82)	**E** (0.74)	**L** (0.89)
E (0.82)	**Q** (0.67)	**R** (0.54)		V (0.64)	**I** (0.63)	N (0.43)			G (0.27)		**P** (0.04)	I (0.49)	**A** (0.68)	**A** (0.61)
M (0.72)	**R** (0.60)	Y (0.53)		**L** (0.61)	**R** (0.63)	**D** (0.41)			**P** (0.00)			W (0.45)	Y (0.54)	
R (0.67)		**L** (0.51)		**I** (0.55)		**P** (0.02)						**V** (0.39)	**V** (0.53)	
L (0.65)		**K** (0.44)		**M** (0.49)								**P** (0.04)	**I** (0.49)	
		A (0.44)											**M** (0.38)	
		F (0.35)												
		I (0.33)												
		V (0.22)												

*For each position, residues whose proportions change significantly (>2σ) with respect to the overall amino acid distribution in 1,131 α-helices are shown. Change in proportion of a residue shown in bold letters is greater than 3 σ. σ is the estimated standard deviation for the residue at a given position. Number within a parenthesis indicates propensity of a residue to occur at a particular position.

Hydrogen Bonding Motifs at the Helix Termini

The values of hydrogen bond distances and angles for a regular poly (L-Alanine) α-helix, generated using the coordinates of L-Alanine in α-helical conformation and the α-helical parameters given by Arnott and Dover,[38] have been used as benchmarks to check for occurrence of hydrogen bonds. Since positions of hydrogen atoms are not defined in most protein crystal structures, the distances and angles were calculated from the heavy atom positions. At each of the 14 positions at the helix termini, first and last turn of the helix (N″–N4, C4–C″), i.e., hydrogen bonds that involve both donor and acceptor atoms from the helix, were computed and divided into two types: MM, both donor and acceptor atoms belonging to helix backbone; and M/S, either donor or acceptor atom belonging to the residue sidechain.

Residues involved in hydrogen bond type M/S were used for identifying structural motifs at helix termini. Structural motifs were characterized by the backbone torsion angles (ϕ, ψ) of all the residues and the sidechain rotamers (χ^1s) for the residues containing donor/acceptor atoms. The sequences involving these structural motifs were analyzed to identify sequence preferences, if any.

TABLE V. Propensitywise Rank Orders for Amino Acid Residues at the 15 Positions

N″	N′	Ncap	N1	N2	N3	N4	Mid[a]	C4	C3	C2	C1	Ccap	C′	C″
G	L	S	P	E	E	M	A (1.56)	M	L	K	H	G	G	P
P	T	D	E	D	Q	L	M (1.41)	W	M	Q	A	H	P	D
S	N	T	Q	Q	H	I	Q (1.40)	L	A	R	K	N	N	K
C	G	N	W	S	D	V	L (1.35)	I	Q	E	E	Q	K	H
Y	S	G	A	W	T	R	R (1.33)	R	K	A	R	A	H	N
D	M	P	R	P	A	F	I (1.16)	A	H	L	Q	K	C	G
N	W	C	K	A	F	A	E (1.16)	F	R	W	L	Y	R	F
H	D	H	L	H	S	Q	K (1.12)	Q	W	I	N	C	S	R
T	P	Q	D	N	V	Y	V (1.05)	K	E	H	Y	M	Q	E
W	I	M	G	G	L	W	W (0.98)	Y	F	V	S	F	L	Q
F	H	E	C	R	Y	K	Y (0.94)	C	I	M	F	S	W	I
V	F	R	T	T	G	H	F (0.94)	V	Y	F	T	E	E	M
K	Y	Y	F	K	K	T	C (0.87)	E	C	T	W	D	F	L
Q	V	L	I	Y	W	S	T (0.84)	S	N	Y	M	R	D	Y
A	C	W	Y	F	M	C	N (0.78)	H	V	C	D	L	T	T
E	K	K	S	V	N	G	H (0.76)	N	T	D	I	T	A	V
I	E	A	V	L	P	E	D (0.74)	D	S	S	V	I	Y	S
M	A	F	M	I	I	N	S (0.72)	T	D	N	C	W	V	W
R	Q	I	N	C	R	D	G (0.43)	G	G	G	G	V	I	A
L	R	V	H	M	C	P	P (0.10)	P	P	P	P	P	M	C

[a] Numbers in parenthesis indicate propensity of a residue to occur in middle region of α-helices.

RESULTS AND DISCUSSION

Every Position in α-Helix Has Its Own Characteristic Amino Acid Distribution

The distribution of individual amino acids at each of the 15 positions (defined in Materials and Methods) in the 1,131 α-helices (positions N″ to C″) is listed in Table II. The values of χ^2, Euclidean and Hamming distances at each of the 15 positions computed with respect to the reference distribution, are shown in Table III. Entries in this table show that differences in the distributions of amino acids at each of the 15 positions are highly significant, especially at Ncap, Ccap, and C′ positions. These observations indicate that each position within and around an α-helix has its own characteristic amino acid requirements. Large values of Euclidean and Hamming distances[36] for each of the 15 positions in the 20-dimensional amino acid composition space with respect to the amino acid composition of the 1,131 α-helices (taken as origin) reflect large differences among them and support our hypothesis that each of these 15 positions has its own characteristic sequence requirements and, thus, significantly distinct amino acid distribution. Relatively smaller values of Euclidean and Hamming distances for the Mid position are due to the fact that amino acid distribution in the middle of helices is closer to the reference distribution, being the difference between the reference distribution and the amino acid distributions in the first and last turns of the 1,131 α-helices. However, the differences between the reference amino acid distribution and that at Mid position are also significant, as shown by the χ^2 test.

Preferred and Avoided Residues at Termini, First Turn, Last Turn, and Middle Region of α-Helices

Since χ^2 tests indicate that each of the 15 positions has unique residue requirements, it is implied that propensities of individual amino acids to occur in α-helices may vary in a position-dependent manner. Figure 1 shows the variation of amino acid propensities with respect to the 15 positions within and around α-helices. We have used the change in proportion test to identify residues whose proportions change significantly at a given helix position with respect to the reference distribution. At each of the 15 positions, a residue whose proportion increases significantly ($>2\sigma$) and whose propensity to occur at the position is greater than one is taken to be "preferred" at that position. As a corollary, a residue whose proportion decreases significantly and whose propensity to occur at the position is less than one is said to be "avoided" at that position. Table IV lists amino acid residues whose proportions change significantly at each of the 15 positions and Table V shows propensitywise rank orders for amino acids to occur at each of the 15 positions. Salient features observed for preferred and avoided residues at various positions are described below.

Ncap position prefers residues with small polar sidechains, viz., Ser, Thr, Asp, and Asn. Sidechain hydroxyl and carbonyl groups of these residues often form hydrogen bonds with the free mainchain amide NH groups in the first turn of α-helices.[1,10,29,39,40] In their analysis, Richardson and Richardson[10] had found that Asn has extraordinarily large preference

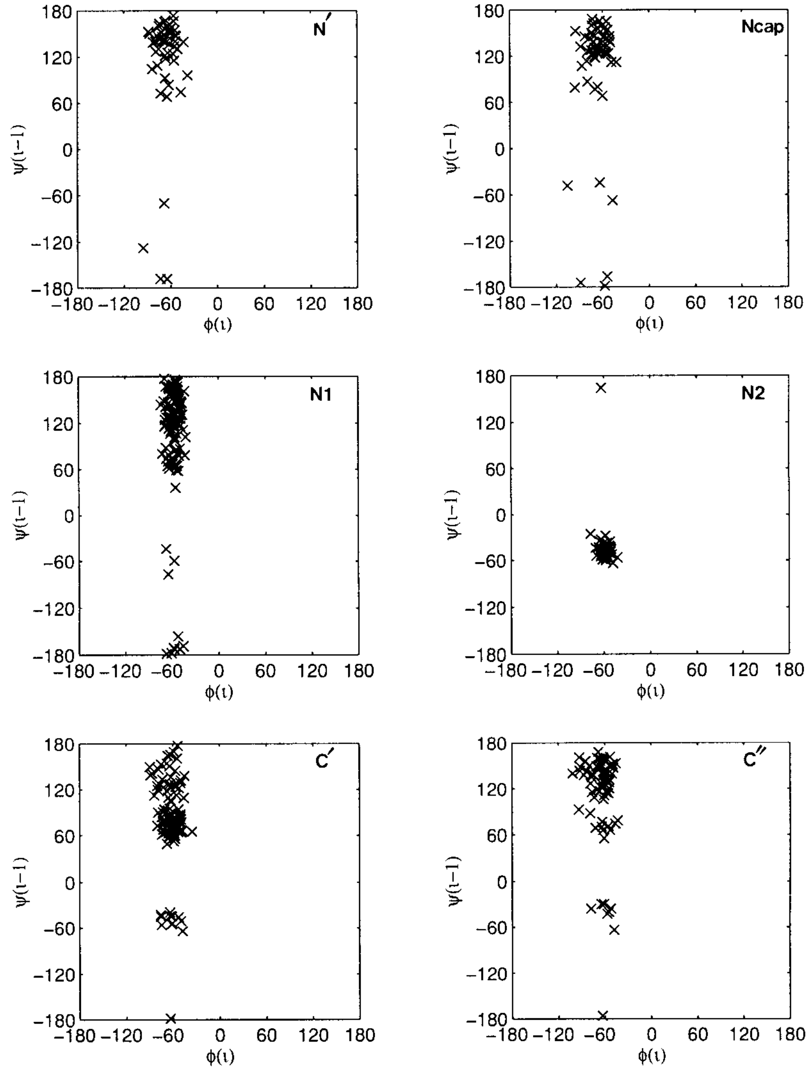


Fig. 2. ψ values of the residues preceding a proline residue are plotted against ϕ of the proline residue at its preferred helical positions. In each box, the helix position for proline is indicated in upper right corner.

(3.5:1) to occur at Ncap position, while, in our database, Asn shows only a marginally higher propensity at Ncap (1.94) than that at Ccap position (1.76). We also find that Ser (2.63), Thr (2.12), and Asp (2.58) have higher propensity to occur at Ncap position than Asn. Besides Ncap, Asn has high propensities, and is thus favoured at N″, N′, Ccap, C′, and C″ positions also. Since the Asn sidechain contains both amide and carbonyl groups, it has been suggested that it is an ambidextrous residue and can form hydrogen bonds with the free mainchain amide groups at N-terminus of α-helix as well as with the mainchain carbonyl groups at the C-terminus of α-helix. Hydrogen bond formation by Asn sidechain $O^{\delta 1}$ atom with mainchain NH groups in the first turn has been observed earlier.[10,29,40] The present analysis reveals that an (i, i − 4) hydrogen bonding motif is also formed by Asn in the C-terminal region of α-helices, and the conformational characteristics of this motif will be described later. Pro is also found to be a favored residue at Ncap and more so at N1. The following residue-pairs involving Pro at N1 have high propensities : SP (4.02), DP (3.36), TP (2.41), HP (2.21), GP (1.81), and NP (1.78).

Consistent with the early studies[2–8] on helix dipole and its interaction with polar residues near the helix termini, N2 and N3 positions prefer Glu and Gln while Lys and His are preferred in the last turn of the helix. His also has a high propensity to occur at Ccap position, and, like Asn, it can form a hydrogen-bonding motif when it occurs at Ccap position. Structural characteristics of this motif are very similar to that formed by Asn and will be described later.

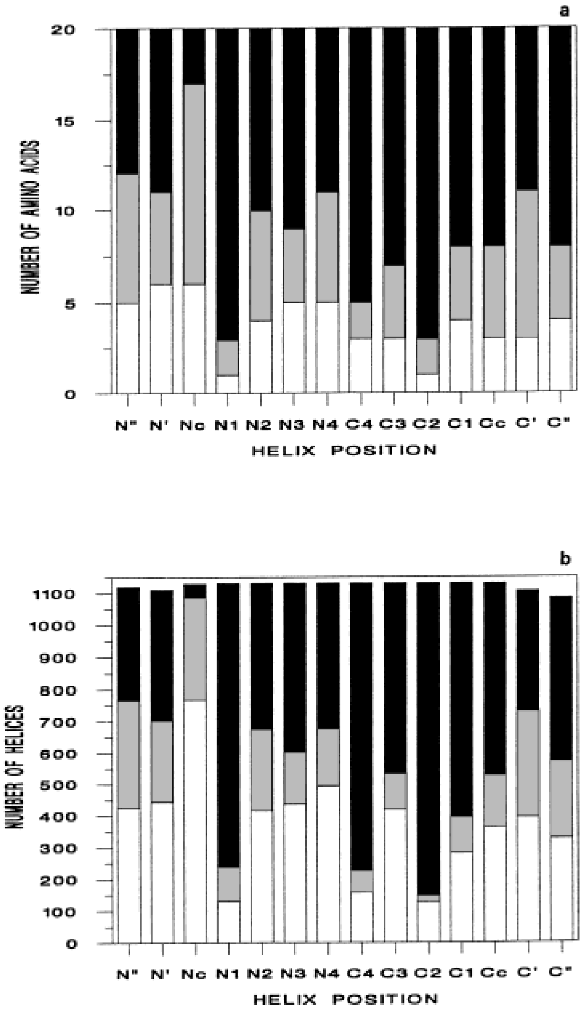


Fig. 3. Relative importance of 14 positions (N″, N′, Ncap, N1, N2, N3, N4, C4, C3, C2, C1, Ccap, C′, C″) within and around α-helices with respect to helix structure. **a:** Stacked bar plot for the number of amino acids (Y-axis) versus each helix position (X-axis). A white bar represents the number of amino acids preferred at a helix position, a gray bar represents the number of amino acids avoided at a helix position, and a black bar represents the number of amino acids neither preferred nor avoided at a helix position. **b:** Stacked bar plot for number of helices (Y-axis) versus each helix position (X-axis). A white bar represents the number of helices containing a preferred amino acid at a helix position, a gray bar represents the number helices containing an avoided amino acid at a helix position, and a black bar represents the number of helices containing a residue that is neither preferred nor avoided at a helix position.

From N4 to Ccap position in helices, Pro is the most avoided residue, with its propensity being nearly zero at these positions. Proline is known to cause kinks if it occurs in the middle of helices,[34,41] thus it can be rationalized that Pro should be disfavored in the main body of an α-helix. Among 1,131 α-helices, we could not find any helix that contains a Pro residue at C3 or C2 positions (Table II). Gly is the second most avoided residue at these positions. Absence of the helix-breaking residues Pro and Gly at these positions, which may cause premature termination of the helix due to absence/breakage of intrachain hydrogen bonds, suggests that the hydrogen bonds in the last turn of the helix are very essential.

Interestingly, Pro is highly favored at C′ and C″ positions, indicating that Pro causes a helix break one or two residues before its actual occurrence. Figure 2 shows a (ψ_{i-1}, ϕ_i(Pro)) plot in which ϕ of the Pro residue has been plotted against ψ of the preceding residue for all the positions that favor Pro. It can be seen that except for Pro at N2 position, the presence of Pro tends to restrict the previous residue to the extended conformation ($\psi > 60°$), as expected from simple energy considerations.[42] However, if Gly precedes Pro in the sequence, then this restriction can be relieved due to absence of a bulky sidechain.[42] In our database, only in six out of the 189 cases, where Pro occurs at C′ and C″ positions, does a Gly residue precede Pro; but even in these six cases, Gly residues also have extended conformations. The amino acids with high propensity to occur at Ccap position along with Pro at the C′ position are C(2.70), F(2.70), H(2.59), Y(2.23), N(1.64), and Q(1.64), of which only His and Asn have high propensity to occur at Ccap. Thus, X-Proline with the X-residue adopting an extended conformation is a recurring structural motif that causes helix termination, in addition to the motifs involving Gly and described by previous analyses, e.g., the Schellman motif,[24,43–46] which terminates 333 (29%) out of 1,131 α-helices in our database.

Identification of the doublets involving Pro at both the helix termini has obvious importance in prediction of helix N- and C-termini, since location of these doublets by a prediction program can facilitate identification of the helix ends.

Hydrophobic contribution to helix stability comes largely from the middle region of an α-helix, with apolar aliphatic residues Ala, Leu, Val, and Ile being highly preferred at N4, Mid, C4, and C3 positions. Since α-helices usually form the core of protein molecules, which is hydrophobic in nature, it explains the occurrence of hydrophobic amino acids at the mid and nearby helix positions. It is interesting to note that these apolar residues, along with the residues that contain large aliphatic sidechains, have small propensities and are thus avoided at several positions at helix termini and just outside the helix. These positions are often part of reverse turns or loops connecting two secondary-structural elements that are generally found on the protein surface, exposed to aqueous environment, and are therefore avoided by the hydrophobic residues.

Designing a sequence that will form an α-helix can be thought of as an optimization problem between helix-favoring and -disfavoring tendencies of each of the 20 amino acids at various positions within and around the helix. Intuitively, a good design should

TABLE VI. Square of Linear Correlation Coefficient (r^2) Between Various Propensity* Scales

	Our study	Williams	Levitt	Chou and Fasman	Scheraga	AK/AQ†
Our study		0.72	0.59	0.47	0.75	0.61
Williams			0.57	0.64	0.65	0.39
Levitt				0.75	0.45	0.30
Chou and Fasman					0.33	0.25
Scheraga						0.29

*Our study includes 53,238 residues in 205 nonhomologous protein chains. Williams = Williams et al. (1987),[35] with 39,707 residues in 212 proteins; Levitt = Levitt (1978),[49] with 11,569 residues in 66 proteins; Chou and Fasman = Chou and Fasman (1974),[48] with 2,473 residues in 15 proteins; and Scheraga = Scheraga et al. (1990),[50,51] with propensities determined by using host guest technique in peptides.

†(AAKAA)n and (AAQAA)n peptides studied by Baldwin et al.[47]

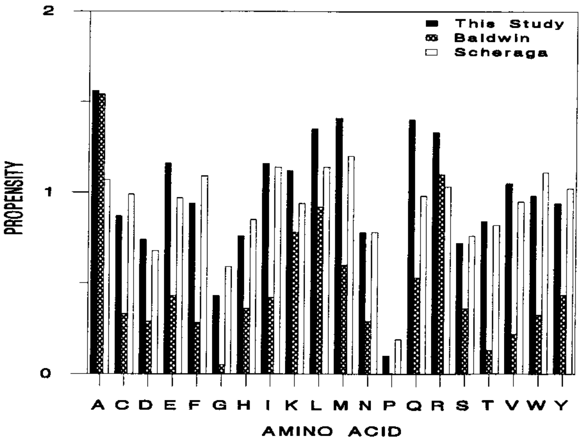


Fig. 4. Comparison of the amino acid propensities calculated for the middle region of 1,131 helices in this study (black bars) with those determined experimentally by Scheraga et al.[50,51] (blank bars) and by Baldwin et al.[47] (hatched bars). The amino acid propensities determined in this study correlate well with these two sets of propensity values, as shown in Table VI.

simultaneously maximize the number of positions containing preferred residues and minimize those containing avoided residues. Figure 3a shows the number of residues preferred, avoided, and neither preferred nor avoided at each of the 14 positions (N″–N4, C4–C″). Figure 3b shows the number of helices containing preferred, avoided, and neither preferred nor avoided residues at each of the 14 positions. At almost every position, the total number of amino acid residues avoided is more than the number of amino acid residues preferred at that position. Inspite of this, at each position the number of helices containing any of the large number of avoided residues is much smaller than the number of helices containing one of the few preferred residues, as seen from Figure 3b. Thus, natural helix sequences optimize by identifying the residues that should be avoided and stressing upon minimum occurrence of these avoided residues in the sequences. For example, Ncap position, which immediately precedes the helix-initiating position (N1), prefers only six residues, Ser, Thr, Asp, Asn, Gly, and Pro, while it avoids 11 (Ala, Glu, Phe, His, Ile, Lys, Leu, Met, Gln, Arg, and Val) residues. However, 767 (68%) helices contain one of the six preferred residues at the Ncap position while the 11 avoided residues are present only in 318 (28%) helices (Fig. 3b). Thus, Ncap position has very stringent sequence requirements, as it either prefers or strongly avoids as many as 17 out of the 20 amino acid residues (Table IV). This probably reflects its importance to helix structure initiation, a hypothesis supported by the fact that four of the six preferred residues at the Ncap position contain a sidechain hydroxyl or carbonyl groups that can form hydrogen bonds and stabilize helix formation. In contrast, N1, Mid, C2, and C4 positions are most neutral, with more than 15 residues being neither preferred nor avoided. Thus, almost any residue can be accommodated at these positions in the helical structure, although Pro is preferentially accommodated at N1 and Lys at C2. In conclusion, all the positions at the N-terminus of the helix (except N1) are more selective than the middle region or last turn of the helix. Surprisingly, the positions Ccap, C′, and C″ are also more selective than Mid to C1 positions.

Propensities of Amino Acids to Occur in Middle of α-Helices

In our analysis, Ala is the only preferred residue and Pro is the only avoided residue at the mid position. A large body of experimental and theoretical data exists on determination of propensities of amino acid residues to occur in α-helices, and work in this field has been recently reviewed.[47] In Table V, we have listed propensities of amino acids to occur at the mid position of the helices in our database (53,238 residues in 205 nonhomologous protein chains found in high-resolution protein crystal structure). We have also compared these values with the propensities determined by Chou and Fasman[48] using 2,473 residues in 15 protein crystal structures, by Levitt[49] using 11,569 residues in 66 protein crystal structures, by Williams et al.[35] using 39,707 residues in 212 proteins, by Scheraga et al.,[50,51] and in AK/AQ peptides from Baldwin et al.[47] We have not compared our propensity data with those determined using coiled coils,[52] since these are biased by

TABLE VII. Average Parameters for Mainchain–Mainchain (i, i + 3) and (i, i + 4) Hydrogen Bonds Found at or Near Helix Termini in Globular Proteins*

	Donor NH at (i + 4) position				Donor NH at (i + 3) position			
Acceptor O position (i)	d (Å)	∠O (°)	∠N (°)	Number of hydrogen bonds	d (Å)	∠O (°)	∠N (°)	Number of hydrogen bonds
N″	2.95 ± 0.15	156 ± 18	116 ± 6	22	2.98 ± 0.18	124 ± 16	106 ± 7	71
N′	2.98 ± 0.17	152 ± 9	114 ± 5	12	3.01 ± 0.13	128 ± 8	107 ± 5	34
Ncap	2.99 ± 0.12	159 ± 8	118 ± 7	805	3.01 ± 0.15	122 ± 10	101 ± 5	565
N1	2.95 ± 0.12	156 ± 8	117 ± 6	991	3.06 ± 0.12	114 ± 8	97 ± 4	349
N2	2.98 ± 0.12	152 ± 8	115 ± 6	909	3.07 ± 0.13	111 ± 8	97 ± 4	192
N3	2.97 ± 0.12	155 ± 7	117 ± 5	986	3.07 ± 0.13	111 ± 7	97 ± 4	312
N4	2.96 ± 0.12	156 ± 7	116 ± 5	983	3.08 ± 0.11	111 ± 6	98 ± 4	335
C4	2.94 ± 0.15	152 ± 12	109 ± 11	851	3.05 ± 0.13	113 ± 7	100 ± 5	320
C3	3.03 ± 0.14	147 ± 12	102 ± 12	177	3.02 ± 0.15	114 ± 8	101 ± 6	271
C2	3.00 ± 0.15	153 ± 15	111 ± 13	41	2.98 ± 0.14	116 ± 8	103 ± 6	345
C1					2.99 ± 0.15	122 ± 14	106 ± 7	112
Fiber model of α-helix	2.86	160.47	119.00		3.13	108.49	94.72	

*d = distance between the mainchain carbonyl oxygen O and the mainchain amide nitrogen N (Cutoff D ≤ 3.2 Å); ∠O = angle between vectors C═O and O . . . N; ∠N = angle between vectors O . . . N and N–C^{α}; fiber model of α-helix = a poly(L-alanine) structure generated using atomic coordinates for L-alanine in α-helical conformation and α-helix parameters given by Arnott and Dover (1967).[38]

the problem of aggregation[53] or with Barnase[54] and T4 lysozyme[55] because absolute measurement of helix propensity is not possible in the protein systems.[47]

Agreement among various propensity scales can be assessed from the values of square of linear correlation coefficient given in Table VI. There is no significant correlation between amino acid propensities in our analysis and those proposed by Chou and Fasman[48] (square of linear correlation coefficient, $r^2 = 0.47$) from analyses of much smaller data sets, while there is a good correlation between our propensity set and that determined by Williams et al.[35] ($r^2 = 0.72$) from a large protein data set. However, our set contains a larger number of residues and consists of nonhomologous protein chains. The best correlation ($r^2 = 0.75$) is obtained with the experimentally determined propensities by Scheraga et al. from studies on peptides.[50,51] Propensity set determined using alanine-based peptides[47] also correlates well with our data ($r^2 = 0.61$), though this set appears to underestimate the propensities of most residues relative to that of alanine and does not correlate with Scheraga's set ($r^2 = 0.29$). It can be seen from Table VI that correlation between experimental and theoretical scales is improving as the size of data set analyzed becomes larger. Figure 4 shows a comparison of our propensity data with the experimentally determined scales proposed by Scheraga's and Baldwin's groups. It is seen that the values of propensities for several helix-formers, particularly Ala, are consistently higher in our protein data set than that in the scale by Scheraga et al., which has a much smaller range of propensity values, while, as mentioned above, in the AK/AQ set, Ala has a significantly higher propensity than all other residues.

In summary, these studies indicate that the propensity scales have improved with the size of the database analyzed. Hence, it is expected that use of our propensity set, which has been determined using a truly nonhomologous set of protein chains and is the largest database of α-helices analyzed to date, should result in improved prediction for α-helices in proteins.

Hydrogen Bond Motifs at Helix Termini

Mainchain–mainchain hydrogen bonds

As mentioned in the Introduction, current literature describes motifs at the helix termini where sidechain groups participate in hydrogen bonding to stabilize the helices. However, a large number of additional hydrogen bonds involving mainchain carbonyl and amide groups at the helix termini are also seen. The first characteristic intrachain 5 → 1 hydrogen bond in an α-helix is formed by the mainchain amino group of the residue at N5 position, while the last such hydrogen bond involves the amino group of the residue at C1. However, it is clear from Table VII that the amino groups of a majority of the residues at the neighboring N4 and Ccap positions are also involved in 5 → 1 hydrogen bonds, viz., N4 (NH) to Ncap (CO) and Ccap (NH) to C4(CO). In addition, at the C-terminus, a significant number (177) of (i, i − 4) hydrogen bonds are observed between C′ (NH) and C3 (CO). Many of these terminal groups also form bifurcated hydrogen bonds as implied from the large number of (i, i + 3) interactions seen for carbonyl oxygens of residues at Ncap, N1, and C4 to C1 positions.

TABLE VIII. Various Recurring Mainchain–Sidechain Hydrogen Bond Motifs at the Helix Termini*

Position (i), residue and side chain atom involved in H-bond			d (Å)	∠O (°)	∠N (°)	Conformation of the residue involved in side chain H-bond (ψ)	Sidechain rotamer region (χ^1)	Number of examples in the database
(i, i + 1) hydrogen bond motif with mainchain NH group at (i + 1) position								
N″	S	O^{γ}	3.01 ± 0.17	83 ± 7	136 ± 8	α_R, E		21
N″	D	$O^{\delta1}$	2.93 ± 0.22	101 ± 10	130 ± 7	α_R, E	*g*−, *t*	11
N′	T	$O^{\gamma1}$	2.88 ± 0.19	85 ± 3	140 ± 10	E	*t*	10
N′	S	O^{γ}	3.08 ± 0.10	83 ± 3	133 ± 8	α_R, E	*g*−	12
N′	D	$O^{\delta1}$	2.97 ± 0.17	108 ± 7	126 ± 9	α_R, E	*g*−, *t*	11
Ncap	S	O^{γ}	3.06 ± 0.12	76 ± 10	76 ± 10	E	*t*	22
Ncap	D	$O^{\delta1}$	3.08 ± 0.12	105 ± 8	119 ± 4	E	*t*	22
Ncap	N	$O^{\delta1}$	3.07 ± 0.12	107 ± 7	116 ± 3	E	*t*	12
N2	D	$O^{\delta1}$	2.98 ± 0.12	93 ± 12	131 ± 10	α_R	*g*−	15
N2	S	O^{γ}	3.03 ± 0.12	85 ± 4	138 ± 4	α_R	*g*−	22
C1	S	O^{γ}	3.04 ± 0.10	86 ± 4	131 ± 5	α_R	*g*−	17
(i, i + 2) hydrogen bond motif with mainchain NH group at (i + 2) position†								
Ncap	D	$O^{\delta1}$	2.99 ± 0.18	136 ± 11	103 ± 7	E	*g*−, *t*	45
Ncap	N	$O^{\delta1}$	3.01 ± 0.14	140 ± 13	102 ± 5	E	*g*−, *t*	39
(i, i + 3) hydrogen bond motif with mainchain NH group at (i + 3) position								
Ncap	S	O^{γ}	3.05 ± 0.13	134 ± 11	112 ± 7	E	*g*−, *t*	65
Ncap	D	$O^{\delta1}$	2.98 ± 0.15	141 ± 25	115 ± 10	E	*g*−, *t*	61
Ncap	N	$O^{\delta1}$	2.98 ± 0.13	149 ± 20	119 ± 7	E	*g*−, *t*	39
Ncap	T	$O^{\gamma1}$	3.08 ± 0.09	142 ± 11	111 ± 6	E	*g*−	53
(i, i − 3) hydrogen bond motif with mainchain NH group at (i − 3) position								
N3	Q	$O^{\epsilon1}$	2.95 ± 0.14	134 ± 11	126 ± 6	α_R	*g*+	31
N3	E	$O^{\epsilon1}$	2.98 ± 0.13	136 ± 10	124 ± 7	α_R	*g*+	25
(i, i − 4) hydrogen bond motif with mainchain NH/CO group at (i − 4) position								
N3	E	$O^{\epsilon1}$	2.85 ± 0.14	134 ± 17	112 ± 9	α_R	*g*+	15
C1	N	$N^{\delta2}$	2.94 ± 0.14	130 ± 7	107 ± 8	α_R	*g*+	21
Ccap	N	$N^{\delta2}$	2.90 ± 0.15	134 ± 9	110 ± 15	—	*g*+	24
Ccap	H	$N^{\delta1}$	2.94 ± 0.14	140 ± 12	114 ± 15	—	*g*+	15
Reciprocal hydrogen bond motif containing both (i, i + 3) and (i, i − 3) hydrogen bonds								
Ncap	T/S/D/N	$O^{\gamma1/\gamma/\delta2}$	2.95 ± 0.13	133 ± 14	116 ± 6	E	*g*−	32
N3	E/Q/D	$O^{\epsilon1/\delta2}$	3.06 ± 0.11	144 ± 11	125 ± 7	α_R	*g*+	

*d = distance between two heavy atoms involved in a hydrogen bond (cutoff: d ≤ 3.2 Å); ∠O = angle between vectors C═O and O . . . N; ∠N = angle between vectors O . . . N and N–C^{α}. In case of Asn at C1 and Ccap, ∠N = angle between vectors O . . . $N^{\delta2}$ and $N^{\delta2}$–C^{γ}. In case of His at Ccap, ∠N = angle between vectors O . . $N^{\delta1}$ and $N^{\delta1}$–C^{γ}. α_R = right-handed α-helical region in the Ramachandran (ϕ, ψ) map; E = extended region in the Ramachandran Map; *g*− = *gauche*− (∼ +60°), *g*+ = *gauche*+ (∼ −60°), and *t* = *trans* (∼180°).

†Most (i, i + 2) motifs involve side chains which also form (i, i + 1) or (i, i + 3) hydrogen bonds.

Mainchain–sidechain hydrogen bonds

Additional stabilization to helical structure comes from several structural motifs that contain mainchain–sidechain (M/S) hydrogen bonds formed by the preferred residues, especially at the N-terminal region of α-helices. Fewer hydrogen bonds are observed between the free mainchain carbonyl oxygen atoms and sidechain NH groups of the preferred residues at the C-terminus.

Geometrical characteristics for all M/S hydrogen bond motifs present at helix termini in our database of 1,131 α-helices are summarized in table VIII. Sidechains of residues at Ncap are involved in maximum number of hydrogen bonds, with (i, i + 3) type of hydrogen bonds between residues at Ncap and N3 positions being more frequent than (i, i + 1) and (i, i + 2) types. Two hundred eighty-two (48.2%) out of a total of 585 residues with sidechain oxygen atoms (Asp + Asn + Ser + Thr) at the Ncap position form 56 (i, i + 1), 84 (i, i + 2) and 218 (i, i + 3) hydrogen bonds with free mainchain NH groups of residues at N1, N2, and N3 positions, respectively. Ser and Thr mostly form (i, i + 3) hydrogen bonds, while Asn and Asp form (i, i + 1) and (i, i + 2) types of hydrogen bonds along with the (i, i + 3) hydrogen bonds. At the Ncap position, 53 Thr out of 137 (38%) form (i, i + 3) hydrogen bonds and 80 Ser out of 171 (47%) form 22 (i, i + 1) and 65 (i, i + 3) hydrogen bonds. Compared to this, 93 Asp out of 178 (52%) at Ncap position form 22 (i, i + 1), 45 (i, i + 2), and 61

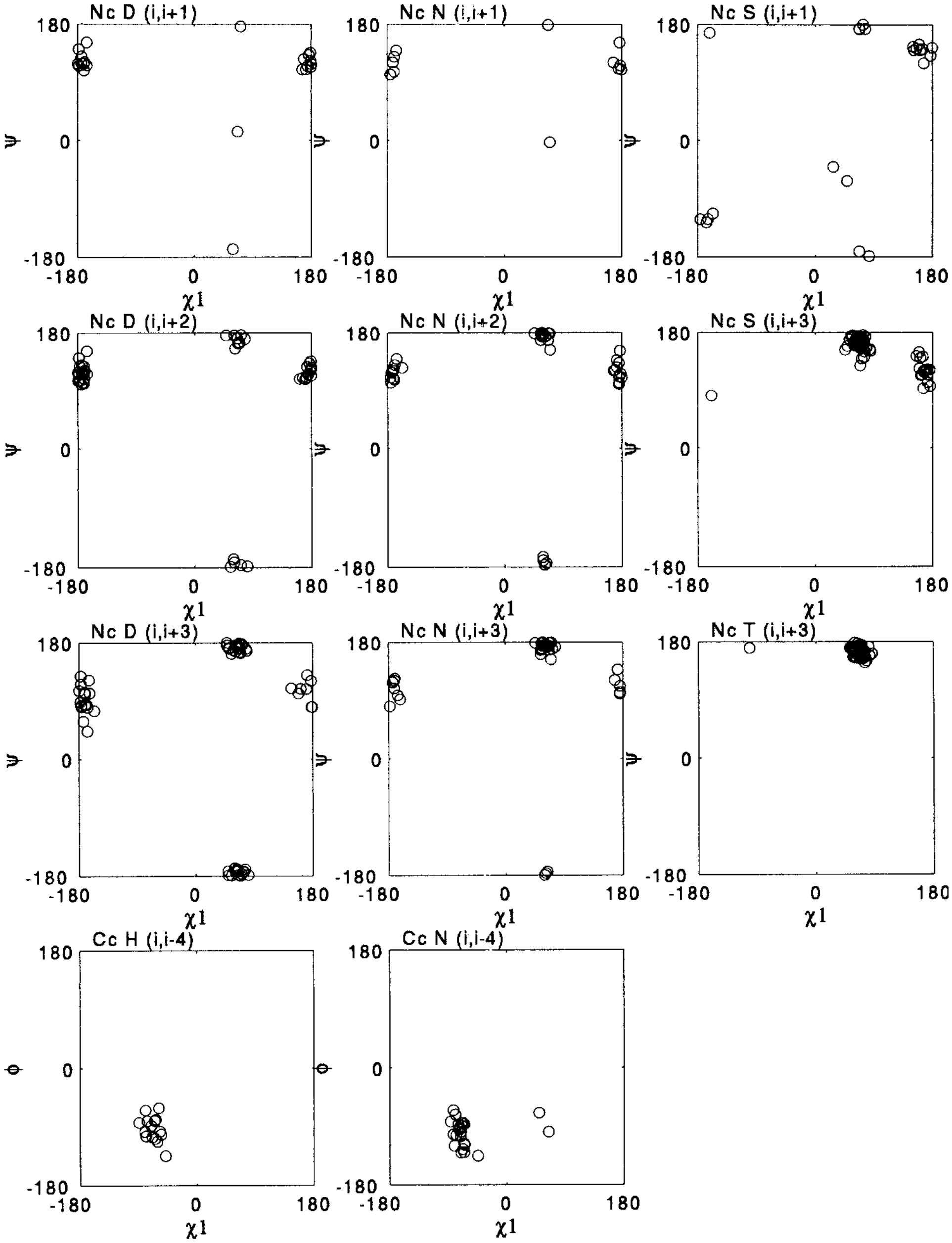
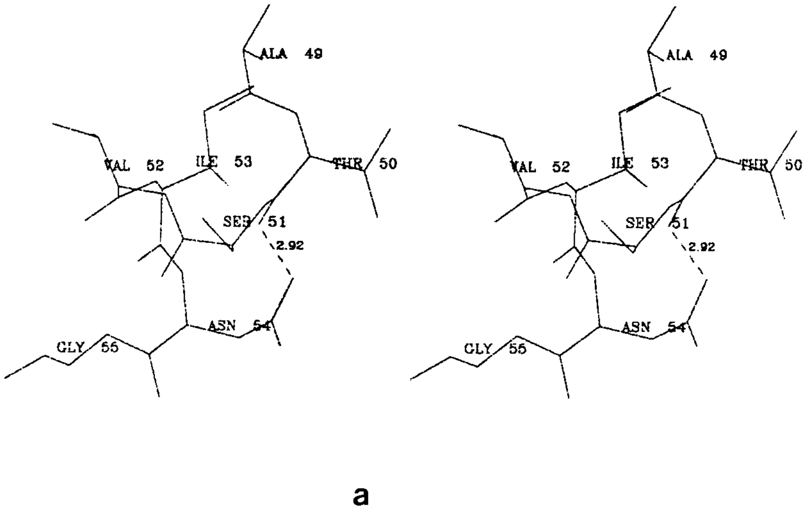


Fig. 5. ψ–χ^1 plots for residues forming M/S hydrogen bonds at Ncap and ϕ $-\chi^1$ plots for Ccap positions. In each plot, X-axis denotes the sidechain torsion angle χ^1 in degrees and Y-axis denotes the backbone torsion angle for the residue whose sidechain is involved in an M/S hydrogen bond. The position, the name of residue whose sidechain is involved in M/S hydrogen bond, and the type of hydrogen bond (denoted by the positions of the residues involved in the hydrogen bond) are indicated at top left of each box. Nc stands for Ncap, Cc for Ccap. Amino acids are denoted by their single-letter code. (i, i + 1) indicates that sidechain carbonyl of residue at position i (Ncap) forms an M/S hydrogen bond with the mainchain NH group at position (i + 1) (N1); similarly for (i, i + 2) and (i, i + 3). In case of (i, i − 4) M/S hydrogen bonds, the sidechain NH groups in Asn and His at position i (Ccap) form hydrogen bonds with the free mainchain carbonyl groups at position i − 4 (C4).

(i, i + 3) hydrogen bonds and 56 Asn out of 99 (56%) form 12 (i, i + 1), 39 (i, i + 2) and 39 (i, i + 3) hydrogen bonds. Thus, Asn and Asp are more efficient hydrogen-bond-formers at the helix N-terminus than Ser and Thr, indicating that the longer sidechains make their carbonyl groups more accessible as acceptors of hydrogen bonds, from more than one backbone NH group.

Backbone conformation and sidechain rotamer preferences for residues involved in various types of hydrogen bonding motifs at the helix termini are shown in the (ψ, χ^1) plots in Figure 5. At Ncap position, the residues forming (i, i + 1) hydrogen bonds motifs have ψ values close to 120° and the preferred sidechain rotamer (χ^1) is *trans* (~180°),



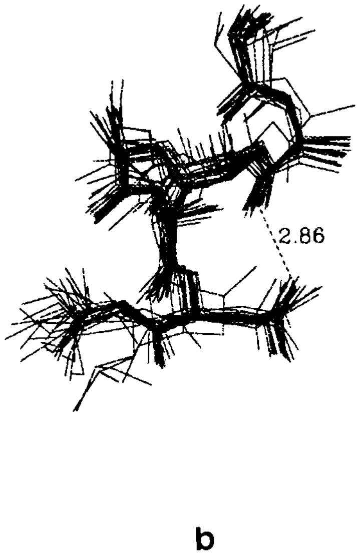


Fig. 6. A representative example of (i, i − 4) sidechain-mainchain hydrogen bond motif, formed by an Asn residues at Ccap position of an α-helix. There are 24 such examples in our database. Sidechain $N^{\delta2}$ atom of Ccap Asn forms a hydrogen bond with the free mainchain carbonyl oxygen of the residue at C4 position of the helix. **a:** Stereo diagram showing an example of this motif (helix S45 - I53 of pdb1sbp). **b:** Superimposed plots of all the 24 examples of this motif. The sidechain atoms beyond C^{β} are not shown for the residues at C1, C2, C3 and C4 positions.

while the same motif is seen inside the α-helices at N2 and C1 positions, with the sidechain rotamer being *gauche*- ($\chi^1 \sim +60°$). The β-branched Thr residues at Ncap, which form only (i, i + 3) motif, have χ^1 in *gauche*- conformation, with ψ being 180°, while Ser, Asp, and Asn show two clusters with (ψ, χ^1) being (120°, 180°) and (180°, 60°) for (i, i + 3) motifs.

While (i, i + 3) mainchain-sidechain hydrogen bonds formed by residues at Ncap positions have been studied earlier,[29] (i, i + 2) hydrogen bond motif formed by Asn and Asp at Ncap position has not been reported. It is interesting to note that such hydrogen bonds are formed only by Asn/Asp and not by Ser/Thr residues. Most of the (i, i + 2) hydrogen bonds are formed by Asn and Asp residues which are also involved in (i, i + 1) or (i, i + 3) hydrogen bonds, hence their characteristics are similar to these motifs.

In addition to these motifs, our analysis reveals the presence of novel (i, i − 3) and (i, i − 4) hydrogen bond motifs formed by sidechain $O^{\epsilon1}$ of Gln or Glu at N3 position within α-helix, with the mainchain NH of the residues at Ncap or N′ positions, respectively. The values of sidechain rotamer (χ^1) for the residues involved in these motifs cluster in *gauche*+ ($\sim -60°$) region.

In all, 218 (i, i + 3) and 56 (i, i − 3) hydrogen bonds are formed between the residues at Ncap and N3 positions. However, we find only 32 examples of the reciprocal hydrogen bonding motif[22] that contains both (i, i + 3) and (i, i − 3) hydrogen bonds between the residues at Ncap and N3 positions. This clearly indicates that the reciprocal hydrogen bond motif is not a very common structural motif at the helix N-terminus, but constitutes a subset of a more general motif that contains only one of the two M/S hydrogen bonds.

Yet another new hydrogen bond motif is found at the C-terminus of the helix, viz., at C1 and Ccap positions, $N^{\delta2}$ of Asn residues and $N^{\delta1}$ of His residues

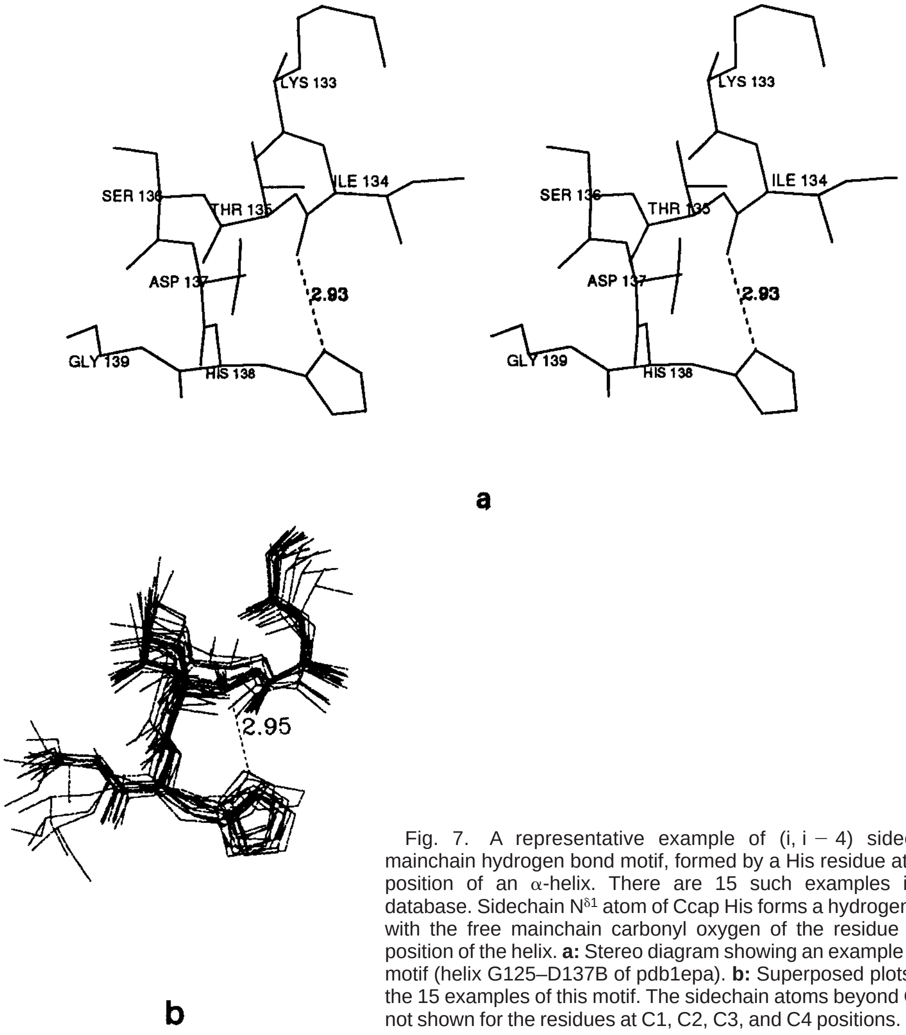


Fig. 7. A representative example of (i, i − 4) sidechain-mainchain hydrogen bond motif, formed by a His residue at Ccap position of an α-helix. There are 15 such examples in our database. Sidechain $N^{\delta1}$ atom of Ccap His forms a hydrogen bond with the free mainchain carbonyl oxygen of the residue at C4 position of the helix. **a:** Stereo diagram showing an example of this motif (helix G125–D137B of pdb1epa). **b:** Superposed plots of all the 15 examples of this motif. The sidechain atoms beyond C^{β} are not shown for the residues at C1, C2, C3, and C4 positions.

TABLE IX. Sequence Preferences for M/S Hydrogen Bond Motifs at the Helix N-Termini

Type of M/S motif (number of examples)	Residues that occur with more than expected frequencies in the motif	Observed frequency of the residues	Expected frequency for the residues	Significance level
(i, i + 3) with Thr at Ncap position (53)	Glu/Gln at N3 position	28	11	95% level of confidence
(i, i − 3) with Glu at N3 position (31)	Ser/Thr at Ncap position	23	8	95% level of confidence
(i, i − 3) with Gln at N3 position (25)	Ser/Thr at Ncap position	16	7	95% level of confidence
Reciprocal motif with (i, i + 3), (i, i − 3) hydrogen bonds between Ncap and N3 positions (32)	Ser/Thr at Ncap position Glu/Gln at N3 position	28 22	9 7	95% level of confidence

form the (i, i − 4) hydrogen bonds with the free mainchain carbonyl oxygen of residues inside the helix. Twenty-one Asn residues at C1 and 24 at Ccap positions form such hydrogen bonds with carbonyl oxygens at C5 and C4 positions in the helices, while 15 out of 46 (~33%) His residues at Ccap position form a similar motif. The sidechain rotamers seen for the residues involved in these motifs cluster in *gauche+* region, with ϕ of Ccap being between −130° to −80°. Figures 6 and 7 show examples of the (i, i − 4) hydrogen bonding motifs formed by His and Asn at the Ccap position, while ($\phi - \chi^1$) plots for these motifs are shown in Figure 5.

Sequence preferences in hydrogen bonded motifs

All motifs listed in Table VIII involve sidechain atoms of a preferred residue, forming hydrogen bond with either a free mainchain amino group or a carbonyl group, of a residue spaced 1, 2, or 3 residues apart in the helix. It is, therefore, expected that there may not be a significant sequence bias for the residue contributing the mainchain amino nitrogen or carbonyl oxygen, over and above the preference of the residue for the position involved. This is indeed true for most of the motifs, except for the (i, i + 3) hydrogen bonding motif involving Thr at Ncap, the (i, i − 3) motifs involving Glu and Gln at N3 position, and the reciprocal (i, i + 3) and (i, i − 3) hydrogen bonding motif.[1,17,22,26,40] The sequence preferences for these motifs are described in Table IX. It is interesting to note that though (i, i + 1) and (i, i + 2) types of hydrogen bond motifs involving residues at Ncap and the (i, i + 3) motif formed by Ser at Ncap have similar structural characteristics as the (i, i + 3) motif formed by Thr, they do not show any sequence preferences. In case of the (i, i + 3) motif involving Thr at Ncap, Glu/Gln are observed with more than their expected frequency at N3. In case of (i, i − 3) motif involving Glu/Gln at N3 position, Ser/Thr occur at Ncap position more frequently than their expected frequency. Similarly, in the reciprocal M/S motif, Ser/Thr at Ncap and Glu/Gln at N3 position occur more frequently than their expected frequencies and are much more dominant than previously reported,[22] suggesting that the consensus sequence for this motif is Thr/Ser–X–X–Glu/Gln.

CONCLUSIONS

Most of the questions raised in the Introduction can be answered on the basis of this analysis. Each of the 15 positions within and around α-helices has its own unique sequence characteristics and hence they cannot be grouped into a few classes such as terminal and middle regions. As a consequence, propensities of amino acids to occur in α-helices, especially at the termini, show a position-dependent variation. The various positions have different choices for preferred and avoided amino acid residues, reflecting their differing roles in helix structure and stability. Positions at the N-terminus are more selective than those in the main body of the helix. Ncap position appears to be most important for helix structure. Ncap favors six residues (D, G, N, P, S, T) and avoids another eleven residues (A, E, F, H, I, K, L, M, Q, R, V). The rigid sequence requirement at Ncap position indicates a predominant role for this position in helix initiation and stability.

In contrast to the previous studies that showed Asn to be the most preferred residue at helix N-terminus, this analysis shows that Ser, Thr, and Asp are more preferred than Asn at Ncap. Also, Asn is almost equally favored at both the helix termini, with its propensity to be present at Ncap position being only marginally higher than that at the Ccap position.

Natural helix sequences seem to be better adapted at identifying the residues to be avoided at any position and minimizing the occurrence of these avoided residues. Residues with sidechains that can form hydrogen bonds are preferred at the helix termini and they are often involved in hydrogen bond formation with free NH and CO groups in the helix backbone. These capping residues generally have extended conformation and well-defined sidechain rotamer preferences for *gauche*- and *trans* at Ncap and *gauche*+ at Ccap. A large number of (i, i + 3) and (i, i − 3) types of mainchain-sidechain hydrogen bonds are formed by the residues at Ncap and N3 positions. However, the reciprocal motif containing both these hydrogen bonds[22,26] forms only a small subset of these motifs. Furthermore, both (i, i + 3) and i, i − 3) motifs show characteristic sequence preferences, with Ser/Thr (rather than Asn) being preferred at Ncap and Glu/Gln at N3. This is in contrast to (i, i + 2) motifs formed by Asn and Asp residues at Ncap position, which do not show any strong preference for residues at N2 position.

It appears that α-helices are terminated in several different ways. In addition to the well-documented Schellman motif involving Gly in α_L conformation at Ccap position, two other motifs are also seen. These are characterized as an X-Pro motif, where residues at X (often Cys, Phe, His, Asn, Gln, or Tyr) adopt an extended backbone conformation, and a sidechain–mainchain hydrogen bonded motif, involving Asn or His in a non helical backbone conformation and *gauche*+ sidechain rotamer.

Interestingly, an X-Pro motif with Pro at N1 and X residue at Ncap position adopting an extended conformation is also seen at the N-terminus, where it facilitates the formation of hydrogen bonds involving the sidechains of residues at Ncap.

Propensity scale determined in this study for the middle region of α-helices correlates best with the experimental propensity scales. Hence, use of the scale presented here can result in improved secondary-structure prediction for α-helices. Besides, identification of several doublets involving Pro, at both the helix termini, can be a good input for locating the helix termini.

ACKNOWLEDGMENTS

We thank Professor N.V. Joshi for advice on statistical analysis, and Dr. M. RaviKiran for several useful discussions and suggestions during the course of this investigation. SK acknowledges CSIR, India for a fellowship.

REFERENCES

1. Presta, L.G., Rose, G.D. Helix signals in proteins. Science 240:1632–1641, 1988.
2. Hol, W.G.J., Van Duijnen, P.T., Brendsen, H.J.C. The α-helix dipole and properties of proteins. Nature 273:443–446, 1978.
3. Hol, W.G.J., Hail, L.M., Sander, C. Dipoles of the α-helix and β-sheet: Their role in protein folding. Nature 294:532–536, 1981.
4. Hol, W.G. Effect of the α-helix dipole upon the functioning and structure of proteins and peptides. Adv. Biophys. 19:133–165, 1985.

5. Hol, W.G. The role of the α-helix dipole in protein function and structure. Prog. Biophys. Mol. Biol. 45:149–195, 1985.
6. Blagdon, D.E., Goodman, M. Mechanism of protein and polypeptide helix initiation. Biopolymers 14:241–245, 1975.
7. Wada, A., Nakamura, H. Nature of charge distribution in proteins. Nature, 293:757–758, 1981.
8. Warwicker, J., Watson, H.C. Calculation of the electric potential in the active-site cleft due to α-helix dipoles. J. Mol. Biol. 157:671–679, 1982.
9. Argos, P., Palau, J. Amino acid distribution in protein secondary structures. Int. J. Prot. Pept. Res. 19:380–393, 1982.
10. Richardson, J.S., Richardson, D.C. Amino acid preferences for specific locations at the ends of α-helices. Science 240:1648–1652, 1988.
11. Serrano, L., Fersht, A.R. Capping and α-helix stability. Nature 342:296–299, 1989.
12. Serrano, L., Neira, J.L., Sancho, J., Fersht, A.R. Effect of alanine versus glycine in α-helices on protein stability. Nature 256:453–456, 1992.
13. Lecomte, J.T.J., Moore, C.D. Helix formation in Apocytochrome-B5: The role of neutral histidine at the N-cap position. J. Am. Chem. Soc. 113:9663–9665, 1991.
14. Bell, J.A., Becktel, W.J., Sauer, C., Baase, W.A., Matthews, B.W. Dissection of helix capping in T4 lysozyme by structural and thermodynamic analysis of 6 amino acid substitutions at Thr 59. Biochemistry 31:3590–3596, 1992.
15. Chakrabartty, A., Doig, A.J., Baldwin, R.L. Helix capping propensities in peptides parallel those in proteins. Proc. Natl. Acad. Sci. U.S.A. 90:11332–11336, 1993
16. Farood, B., Feliciano, E.J., Nambiar, K.P. Stabilisation of α-helical structures in short peptides via end capping. Pro. Natl. Acad. Sci. U.S.A. 90:838–842, 1993.
17. Lyu, P.C., Wemmer, D.E., Zhou, H.X., Pinker, R.J., Kallenbach, N.R. Capping interactions in isolated α-helices: Position-dependent substitution effects and structure of a serine-capped helix. Biochemistry 32:421–425, 1993.
18. Yumoto, N., Murase, S., Hattori, T., Yamamoto, H., Tatsu, Y., Yoshikawa, S. Stabilisation of α-helix in C-terminal fragments of neuropeptide-Y. Biochem. Biophys. Res. Commun. 196:1490–1495, 1993.
19. Doig, A.J., Chakrabartty, A., Kingler, T.M., Baldwin, R.L. Determination of free energies of N-capping in α-helices by modification of Lifson-Roig helix–coil theory to include N- and C-capping. Biochemistry 33:3396–3403, 1994.
20. Doig, A.J., Baldwin, R.L. N- and C-capping preferences for all 20 amino acids in α helical peptides. Protein Sci. 4:2247–2251, 1995.
21. Petukhov, M., Yumoto, N., Murase, S., Onmura, R., Yashikawa, S. Factors that affect the stabilisation of α-helices in short peptides by a capping box. Biochemistry, 35:387–397, 1996.
22. Harper, E.T., Rose, G.D. Helix stop signals in proteins and peptides: The capping box. Biochemistry 32:7605–7609, 1993.
23. Dasgupta, S., Bell, J.A. Design of helix ends: Amino acid preferences, hydrogen bonding and electrostatic interactions. Int. J. Pept. Prot. Res. 41:499–511, 1993.
24. Aurora, R., Srinivasan, R., Rose, G.D. Rules for alpha-helix termination by glycine. Science 264:1126–1130, 1994.
25. Zhou, H.X., Lyu, P.C., Wemmer, D.E., Kallenbach, N.R. α-helix capping in synthetic model peptides by reciprocal sidechain–mainchain interactions : Evidence for an N-terminal "Capping Box". Proteins 18:1–7, 1994.
26. Seale, J.W., Srinivasan, R., Rose, G.D. Sequence determinants of the capping box, a stabilising motif at the N-termini of alpha-helices. Protein Sci. 3:1741–1745, 1994.
27. Munoz, V., Serrano, L. Analysis of i, i + 5 and i, i + 8 hydrophobic interactions in α-helical model peptide bearing hydrophobic staple motif. Biochemistry 34:15301–15306, 1995.
28. Munoz, V., Blanco, F.J., Serrano, L. The hydrophobic staple motif and a role for loop-residues in alpha-helix stability and protein folding. Nat. Struct. Biol. 2:380–385, 1995.
29. Doig, A.J., MacArthur, M.W., Stapely, B.J., Thornton, J.M. Structures of N-termini of helices in proteins. Protein Sci. 6:147–155, 1997.
30. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., et al. The Protein Data Bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
31. Hobohm, U., Scharf, M., Schneider, R., Sander, C. Selection of representative protein data sets. Protein Sci. 1:409–417, 1992.
32. Hobohm, U., Sander, C. Enlarged representative set of protein structures. Protein Sci. 3:522–524, 1994.
33. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
34. Kumar, S., Bansal, M. Structural and sequence characteristics of long alpha-helices in globular proteins. Biophys. J. 71:1574–1586, 1996.
35. Williams, R.W., Chang, A., Juretic, D., Loughran, S. Secondary-structure prediction and medium-range interactions. Biochim. Biophys. Acta 916:200–204, 1987
36. Chou, K.C., Zhang, C.T. Prediction of protein structural classes. Crit. Rev. Bioch. Mol. Biol. 30:275–349, 1995.
37. Medhi, J. "Statistical Methods: An Introductory Text." New Delhi: Wiley Eastern Limited, 1992.
38. Arnott, S., Dover, S.D. Refinement of bond angles of an α-helix. J. Mol. Biol. 30:209–212, 1967.
39. Gray, T.M., Mathews, B.W. Intrahelical hydrogen bonding of serine, threonine and cysteine residues within α-helices and its relevance to membrane-bound proteins. J. Mol. Biol. 175:75–81, 1984.
40. Bordo, D., Argos, P. The role of sidechain hydrogen bonds in the formation and stabilisation of secondary structures in soluble proteins. J. Mol. Biol. 243:504–519, 1994.
41. Woolfson, D.N., Williams, D.H. The influence of proline residues on alpha-helical structure. FEBS Lett. 277:185–188, 1990.
42. Schimmel, P.R., Flory, P.J. Conformational energies and configurational statistics of copolypeptides containing L-Proline. J. Mol. Biol. 34:105–120, 1968.
43. Schellman, C. The α_L conformation at the ends of helices. In: "Protein Folding." Jaenicke, R. (ed.). Amsterdam: Elsvier, 1980:53–61.
44. Preißner, R., Bork, P. On α-helices terminated by glycine. 1. Identification of common structural features. Biochem. Biophys. Res. Comm. 180:660–665, 1991.
45. Bork, P., Preißner, R. On α-helices terminated by glycine. 2. Recognition by sequence patterns. Biochem. Biophys. Res. Comm. 180:666–672, 1991.
46. Atschular, E.L., Lades, M. Possible exceptions to rules of α-helix termination by Glycine. Science 269:1451–1452, 1995.
47. Chakrabartty, A., Baldwin, R.L. Stability of α-helices. Adv. Prot. Chem. 46:141–176, 1995
48. Chou, P.Y., Fasman, G.D. Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. Biochemistry 13:211–221, 1974.
49. Levitt, M. Conformational preferences of amino acids in globular proteins. Biochemistry 17:4277–4285, 1978.
50. Altmann, K.-H., Wojcik, J., Vasuez, M., Scheraga, H.A. Helix-coil stability for the naturally occurring amino acids in water. XXIII. Proline parameters from random poly(hydroxybutylgulatame-co-L-proline). Biopolymers 30:107–120, 1990.
51. Wojcik, J., Altmann, K.-H., Scheraga, H.A. Helix-coil stability for naturally occurring amino acids in water. XXIV. Half-cystine parameters from random poly(hydroxybutylglutamate-co-S-methylthio-L-cysteine). Biopolymers 30: 121–134, 1990.
52. O'Neil, K.T., DeGrado, W.F. A thermodynamic scale for helix-forming tendencies of the commonly occurring amino acids. Science 250:646–650, 1990.
53. Betz, S., Fairman, R., O'Neil, K., Lear, J., DeGrado, W.F. Design of two-stranded and three-stranded coiled coil peptides. Phil. Trans. R. Soc. London. B348:81–88, 1995.
54. Horovitz, A., Matthews, J., Fersht, A.R. α-helix stability in proteins. II. Factors that influence stability at an internal position. J. Mol. Biol. 227:560–568, 1992
55. Blaber, M., Zhang, X., Matthews, B. Structural basis of amino acid helix propensity. Science 260:1637–1640, 1993.