Single Amino Acid Preferences for Specific Locations at Type-I α-Turns in Globular Proteins

Mitsuaki Narita, Koji Sode,* and Shokichi Ohuchi#

Department of Biotechnology, Faculty of Technology, Tokyo University of Agriculture and Technology,
Koganei, Tokyo 184-0012

(Received September 17, 1998)

Using both dihedral angles and hydrogen-bond definitions, 225 type-I α-turns were extracted from 125 analyzed proteins. The extracted α-turns were built up from 5% (1125) of the total amino acid residues (23286) found in the proteins. The type-I α-turn is an independent kind of secondary structural unit, although it has been frequently classified as $(i, i+1)$ double β-turns. Both the nucleation of a helix with the type-I β-turn and its propagation with the repetitive addition of the type-I β-turn suggest that the origin of the helix in a globular protein might be the type-I β-turn. Single amino acid preferences for specific locations at type-I α-turns were compared to those at type-I β-turns and the ends of helices. The result indicates that the type-I α-turn, the type-I β-turn, and the helix are of similar kin. Similarities in the structure and amino acid preferences for specific locations among them suggest that the amino acid residues on defined positions along both turns could be properly assigned to the corresponding helix elements.

A variety of secondary structural units, such as helices, β-strands, and β-turns, are assembled in various ways to form a protein tertiary structure. On the other hand, they are built up from secondary structure elements assigned to amino acid residues in protein sequences.[1] Therefore, the protein tertiary structure can be analyzed on the amino acid level. Here, the term "secondary structure element" has been frequently used for "secondary structural unit" in this study.[2,3] Helices consisting of repetitive turns[4] are major secondary structural units, and appear to be folding units. Thus, it is important to investigate the origin of the helix in a globular protein so as to elucidate the mechanism of protein folding.

In general, although helices consist of mainly 5→1 hydrogen bonds (type-I α-turn in this study), 4→1 hydrogen bonds (type-I β-turn[5]) are commonly observed at their N- and C-termini. As shown by model-building technics, a part (type-III β-turn classified by Venkatachalam[6]) of the type-I β-turns are the basic repetitive turn (one winding) of a 3_{10}-helix. However, allowed combinations of various types of β-turns form 5→1 hydrogen bonds between the main chain C=O (i) and the N–H ($i+4$),[7] and they are classified into some types of α-turns.[8] The most frequently occurring α-turn is the basic repetitive turn of a right-handed α-helix.[4,9] In this paper we classify this α-turn as the "type-I α-turn". The 6→1 hydrogen bonds (π-turns[10]) are rarely observed in helices,[7] although a different type of 6→1 hydrogen bond often occurs at the C-termini of helices.[11,12]

To analyze helices,[1] 8000 (20^3) possible kinds of amino acid residues in the middle of triplets have been introduced for the statistical characterization of 11 kinds of helix elements (a—k), which are the building blocks of helices. In this paper, however, we improve the classification of the helix elements. A helix identified as a dodecapeptide sequence, a helix of average length, is represented by using 9 kinds of helix elements (a—i) as follows:

9 kinds of defined positions	(N′)	N_0	N_1	N_2	N_3	M	M	M	M	C_3	C_2	C_1	C_0	(C′)
amino acid sequence		1	2	3	4	5	6	7	8	9	10	11	12	
9 kinds of helix elements a—i	—	a	b	c	d	e	e	e	e	f	g	h	i	—

Using the α-carbon definition, the first (1) or last amino acid residue (12) with its α-carbon within the cylinder defined by the helix is termed an N cap (N_0) or C cap (C_0) residue of the helix, respectively, by Richardson and Richardson.[13] In this study, the end residues of both type-I α- and β-turns as well as those of helices are also called N cap (N_0) and C cap (C_0). The 9 kinds of defined positions in this study are labeled on the basis of the helix end residues (N_0 and C_0). These residues are defined by both the dihedral angles and the hydrogen-bond definitions; amino acid residues (1—12) located on the defined positions (N_0—C_0) are assigned to secondary structure elements (a—i), as shown above. In helix boundary regions, the N′ and C′ residues flank the N_0 and C_0 residues, respectively.

In this paper we demonstrate the similarity in the position-specific preferences of single amino acid residues among type-I β-turns, type-I α-turns, and the ends of helices. The structural similarity among both turns and short helices is illustrated when viewing the nucleation of a helix with the type-I β-turn followed by its propagation with a repetitive addition of the type-I β-turn.

Present Address: Department of Biochemical Engineering & Science, Faculty of Computer Science & Systems Engineering, Kyushu Institute of Technology, Iizuka, Fukuoka 820-0067.

Materials and Methods

Proteins Examined in This Study. The proteins examined in this study can be identified by the Protein Data Bank (PDB) code.[14] A capital letter or a number (the chain identifier) following a hyphen in codes is based on DSSP.[7] Proteins, except for 32 (1cgl), 34 (2dnj), 37 (1ctf), 49 (1fd2), 53 (1fha), 60 (1tab), 74 (1cpc), 80 (1msb), 81 (1mat), 92 (1f3g), 106 (1lpe), 113 (1lts), 119 (3fis), and 125 (1lba), are those used in a previous series of papers.[1,15] The numbers of proteins are also those used previously.

In general, the alignment of two random sequences can produce a 25% or less residue identity.[16] Homologous proteins (32, 81, 106, 113, and 125) in the previous data set have a residue identity of higher than 25% with the corresponding proteins in the data set. We excluded these proteins from the 125 analyzed proteins, since we compared position-specific preferences of not 8000 kinds of amino acids in the middle of triplets, but single amino acids. Thus, we newly enter the following 5 proteins: 1cgl as 32; 1mat as 81; 1lpe as 106; 1lts as 113 and 1lba as 125. Furthermore, proteins (34, 37, 49, 53, 60, 74, 80, 92, and 119) in the previous papers[1,15] were excluded, since their structures were determined by NMR measurements, or since data to more than 3.1 Å resolution were used in refinements of their coordinates. Thus, we enter the following 9 independent proteins: 2dnj as 34; 1ctf as 37; 1fd2 as 49; 1fha as 53; 1tab as 60; 1cpc as 74; 1msb as 80; 1f3g as 92; 3fis as 119. We name this new data set Subset A.

The 125 proteins of Subset A were classified into families according to Murzin et al.[17] in order to understand their homologies. Homologous proteins of Subset A (for example, globins: 1eca, 1; 2lhb, 11; 4mbn, 22; 1gdj, 47; 3sdh, 104) have residue identity of less than 25% and do not bias the results of this study.

Amino Acid Residues Used in This Study and Chain Breaks. The amino acid residues (23286) in the 125 proteins for which there are coordinates[14] were used in this study. The amino acid residues in sequences are represented by one-letter symbols.[15] Chain breaks in the proteins are assumed if the peptide bond length (distance C′–N) exceeds 2.5 Å, according to Kabsch and Sander.[7] Seven chain breaks can be observed in proteins 34 (2dnj), 50 (1lap), 64 (1wsy-A), 64, 85 (2rsp-A), 111 (4ts1-A), and 118 (6fdr). Some of the amino acid residues in these proteins have no coordinates.

Type-I α-Turns and Type-I β-Turns Extraction. The 435 type-I β-turns and the 225 type-I α-turns (Table 1) were easily extracted from the 125 analyzed proteins of Subset A using both ϕ and ψ representations of their 3-dimensional structure[15] and Kabsch and Sander's automatic assignments.[7] At first, both the secondary structural units in the proteins were extracted using the dihedral angles definition, irrespective of the presence or absence of a hydrogen bond. Although most of the N cap and C cap residues of the secondary structural units make an intrahydrogen bond, they depart from the helical values of the ϕ and ψ angles as well as those of the helices. Intrasegments of type-I α-turns and type-I β-turns as well as those of helices comprise ϕ angles of between $-130°$ and $-20°$ and ψ angles of between $-110°$ and $+20°$. Namely, the central $i+1$ and $i+2$ residues of the β-turns and the central $i+2$ and $i+3$ residues of the α-turns possess helical values of the ϕ and ψ angles.

Next, according to both Kabsch and Sander's automatic assignments and the dihedral angles definition of Richardson,[5] 3-turns on DSSP[7] were extracted as type-I β-turns, even though they comprise a dihedral ϕ or ψ angle (their residue $i+1$ or $i+2$) slightly deviated from the helical values. Similar to the type-I β-turns, a small fraction of the type-I α-turns have a dihedral ϕ or ψ angle (their residue $i+2$ or $i+3$) slightly deviated from the helical values.

Results

The dihedral angles and hydrogen-bond definition of the type-I β-turn in this study is essentially identical with the widely used definition of Richardson, although, in his definition, a hydrogen bond is defined ambiguously.[5] The 435 type-I β-turns were easily extracted from the 125 analyzed proteins using both ϕ and ψ representations of their 3-dimensional structure[15] and Kabsch and Sander's automatic assignments of 3-turn based on a hydrogen bond.[7] At first, the type-I β-turns were extracted using the backbone dihedral ϕ and ψ angles near to the helical values, irrespective of the presence or absence of a hydrogen bond. A fraction of the extracted β-turns were actually free from the hydrogen bond defined by Kabsch and Sander.[7]

Next, due to Kabsch and Sander's automatic assignments,[7] their small fraction, which comprised a dihedral ϕ or ψ angle (their residue $i+1$ or $i+2$) slightly deviated from the helical values, were assigned to type-I β-turns, since they formed a hydrogen bond between the main chain C=O (i) and the N–H ($i+3$). Thus, distorted type-I β-turns make the ϕ and ψ regions of intrasegments of the type-I β-turns ambiguous. The extracted β-turns were classified into the 3 groups based on both dihedral (ϕ, ψ) angles and a hydrogen bond, as follows:

a) typical type-I β-turns formed by a 4→1 hydrogen bond,
b) type-I β-turns free from a hydrogen bond,
c) type-I β-turns formed by a 4→1 hydrogen bond, but comprising a dihedral ϕ or ψ angle (their $i+1$ or $i+2$ residue) slightly deviated from helical values.

The (i, $i+1$) double β-turns[18,19] formed by the allowed combinations of various types (I, I′, II, and II′) of β-turns occur frequently in globular proteins, and consist of 5 residues, such that residues 1-4 and residues 2-5 form β-turns. Although a part of combinations are identified as the minimal 3_{10}-helix of 2 successive 3-turns on DSSP,[7] only their fraction corresponds to the minimal 3_{10}-helix of 2 successive type-I β-turns. On the other hand, a part of double β-turns form a hydrogen bond between the main-chain C=O (i) and the N–H ($i+4$), and are identical with α-turns.[8] The most frequently occurring α-turn is the basic repetitive turn of a right-handed α-helix.[4,9]

As classified later, typical type-I α-turns in this study include both the minimal 3_{10}-helix of 2 successive type-I β-turns and the basic repetitive turn of an α-helix, since it is difficult to distinguish the former from the latter. We can recognize a gradual transformation from 2 successive 4→1 hydrogen bonds (the minimal 3_{10}-helix, one of typical type-I α-turns) to an isolated 5→1 hydrogen bond (the other typical type-I α-turn) by using the ϕ and ψ representations of 3-dimensional structures of the analyzed proteins.[15]

However, a fraction of the non-typical type-I α-turns were free from a hydrogen bond defined by Kabsch and Sander,[7] and the residual α-turns included type-I α-turns forming a single 4→1 hydrogen bond and distorted type-I α-turns. Similar to the type-I β-turn, the type-I α-turn was funda-

Table 1. Proteins Examined[a)] and Type-I α-Turn Sequences[b)]

No.	PDB Code	Sequence		No.	PDB Code	Sequence	
1	1eca	KFTQFAG		48	1hip	QDATKSE	LPPEEQH
3	1rhd	FDIEECR	LNRSLLK			QHCADCQ	
		APPETWV		49	1fd2	VTDNCIK	IHPDECI
4	2act	NTEENYP	CGIATMP			FSEDEVP	
5	2aza	MAKSAMG	AGLAQDY	50	1lap	IDEQENW	EYDDLKQ
		FDVSKLT				TPANEMT	
6	2cab	YDDKNGP	IKTSETK	52	1mcp-L	VQAEDLA	WKIDGSE
		YNPATAK	WNSAKYS			QDSKDST	
		SKADGLA		53	1fha	DRDDVAL	
7	2cdv	FNHSTHK		54	1ovo-A	VDCSEYP	
9	2cyp	WDKHDNT	GGSYGGT	55	1paz	PHYAMGM	
		GGTYRFK	NDPSNAG	56	1pyp	PLAPKLN	DVEKYFP
		TPEDTTP				IDKSIDK	
10	2hmz-A	WDISFRT		57	1r09-2	TANETGA	NGSETDV
11	2lhb	FFPKFKG				INLSSLV	
12	2sn3	VKKSDGC	KAKNQGG	58	1rbp	CRVSSFR	
14	2stv	AVAASNG		59	1s01	GYPAKYP	
16	3app	AFSSINT	IDSSKYT	60	1tab	SAAHCYK	YNSNTLN
		VDNSQGF	QDSNAGG	62	1tnf-A	NRPDYLL	
18	3bp2	KLDSCKV	DNPYTNN	63	1ubq	IPPDQQR	
		YNKEHKN	LDKKNC-	65	1wsy-B	KREDLLH	GSYETAH
19	3grs	PHESQIP	PGASLGI			HGIETGE	ISAGLDF
		VGDVCGK	YGIENVK	67	2aat	GLEEDAE	VNVAGMT
21	4fxn	INVSDVN		68	2alp	TAGHCGT	SGRTTGY
22	4mbn	RHPGDFG				GNNCGIP	IPASQRS
23	5cpa	RSTNTFN	QHPELVS	70	1fnd	VGKEMLM	PTSSSLL
24	5cpv	VGLTSKS				KAPDNFR	AVSREQT
27	7lyz	NGMNAWV		71	2fxb	VDKETCI	AAPDIYD
28	8adh	VDEISVA	VNPQDYK			YVTLDDN	
29	8cat	DITRYSK	RNPQTHL	72	2gbp	ANQGWDL	
		FSDRGIP	DGHRHMD	73	2gcr	NLQPYFS	DYQQWMG
		WPHGDYP	QIPVNCP			SSLQDRF	
		HQPSALE	FNSANDD	76	2i1b	VDPKNYP	MEKRFVF
30	8tln-E	IPLSGGI	IGEDVYT			IEINNKL	
		DHYSKRY		77	2ltn-A	WDRETGN	YDKTTQT
31	1acx	VDCATDA				YNAAWDP	FNAATNV
33	1bbp-A	FDWSNYH	PNSVEKY	82	2pab-A	LDAVRGS	ELHGLTT
		YDEDKKG	VDSQKLV	83	2pcy	FDEDSIP	
34	2dnj	DDPNTYH	FRPNKVS	84	2phh	IPAERLK	EKVEDWS
35	1bmv-1	VDLLGGG	VKRSDWA			AGDAAHI	VPPTGAK
		CGPNNGF	DNPKQST	85	2rsp-A	ISEEDWP	
36	1bmv-2	TNLFKLS	TICSQDC	86	2sod-B	GGPKDEE	
		WNPACTK	FSQEEFL	87	2tgp-I	YNAKAGL	
		TTLLADG		88	2tmv-P	LGAFDTR	ANPTTAE
40	1crn	CPGDYAN				TAETLDA	
41	1cse-I	SFPEVVG		89	2tsc-A	FNLQDGF	ESIFDYR
42	1fdl-H	FSLTGYG	KDNSKSQ			YRFEDFE	
		LHTDDTA	VPSSPRP	90	2utg-A	KSPLCM-	
		AHPASST		92	1f3g	VNIEDVP	
43	1fdx	IDADSCI		93	3blm	LDTKSGK	INKDDIV
44	1fkf	ISPDYAY		94	3cd4	LKIEDSD	LELQDSG
45	1fxi-A	ACSTCAG	DQIQAGY	95	3cla	FDVKNWV	QFDELRM
		ILTCVAY				FHQETET	VHHAVCD
46	1gd1-O	YDSVHGR	VNQDKYD	102	3pgm	HYGDLQG	
		YDPKAHH	RAAAESI	104	3sdh	YFKRLGN	GNVSQGM
		VLPELKG	IDALSTM	105	3tim-A	EPVWAIG	
47	1gdj	LFSFLKG		107	4pfk	GDIIHRG	VAEGVGS

a) Type-I α-turn is not included in the following proteins:2,1ppt; 8,2cro; 13,2sns; 15,3adk; 17,3b5c; 20,3lzm; 25,5cyt; 26,5rsa; 32,1cgl-D; 37, 1ctf; 38,1cc5; 39,1cdt-A; 51,1lmb-3; 61,1tgs-I; 64,1wsy-A; 66,256b-A; 69,2ccy-A; 74,1cpc; 75,2gn5; 78,2ltn-B; 79,2mev-4; 80,1msb; 81,1mat; 91,2wrp; 96,3cln; 97,3ebx; 98,3gap-A; 99,3hmg-A; 100,3hmg-B; 101,3icb; 103,3rnt; 106,1lpe; 113,1lts-D; 114,5hvp-A; 119,3fis; 124,9ins-B.

b) Type-I α-turn seuquences from N′ to C′ are represented by one-letter symbols.

Table 1. (Continued)

No.	PDB Code	Sequence		No.	PDB Code	Sequence	
108	4rhv-3	PDTYTSA				EDYRHFS	FIPTSMD
109	4rxn	TCTVCGY	YDPEDGD			LDERENA	WDPATTK
		GDPDDGV	TDFKDIP	118	6dfr	YEPDDWE	
		VCPLCGV	VGKDEFE	120	7icd	KHPELTD	EDIYAGI
110	4sgb-I	TNCCAGY				HKGNIMK	KNPNTGK
111	4ts1-A	LDKEKTS	EAPEKRA			LRPAEYD	GGIGIAP
112	4xia-A	PTPADHF	DGGFTSN			TAPKYAG	
		EYDGSKD	NDSASFA	121	8abp	WDVKESA	FPEKQIY
115	5ldh	LPKHRVI	AVWSGVN			FKAADII	
116	6acn	DDPANQE	HCDHLIE	122	9api-A	FNLTEIP	FEVKDTE
		INLSELK	CGPCIGQ			HCKKLSS	DLSGVTE
		TGRNDAN	FNPETDF	123	9api-B	IEQNTKS	
		INIENRK	RNAVTQE	125	1lba	KGYNHNS	
117	6cpp	VPEHLVF	SNLSAGV				

mentally defined in terms of the backbone dihedral ϕ and ψ angles; however, distorted α-turns also make the ϕ and ψ regions of intrasegments of the type-I α-turns ambiguous. The sequences from N′ to C′ of the type-I α-turns extracted in this study are listed in Table 1. They are classified into the 5 groups based on both the dihedral ϕ and ψ angles and a hydrogen bond, as follows:

a) typical type-I α-turns formed by a 5→1 hydrogen bond,
b) typical type-I α-turns formed by 2 successive type-I β-turns, corresponding to the minimal 3_{10}-helix on DSSP,[7]
c) type-I α-turns forming a single 4→1 hydrogen bond,
d) type-I α-turns free from a hydrogen bond,
e) type-I α-turns formed by one or two hydrogen bonds, but comprising a dihedral ϕ or ψ angle (their $i+1$, $i+2$ or $i+3$ residue) slightly deviated from helical values.

Although the N cap and C cap residues of helices are fundamentally defined in terms of the dihedral ϕ and ψ angles,[1] those of a small fraction of helices are also ambiguous, because of the presence of a hydrogen bond at their distorted C-termini.

The 125 analyzed proteins are consisted of 23286 total single amino acid residues; each amino acid residue was assigned to one of the secondary structure elements constructing the protein structures. They also contained the 435 β-turns (their secondary structure elements a, b, h, and i: 435×4 = 1740, 8%), the 225 α-turns (their secondary structure elements a, b, h, i, and j: 225×5 = 1125, 5%), and the 706 helices (their secondary structure elements: 9487, 41%). Amino acid residues in the center ($i+2$) of the type-I α-turns were assigned to the newly classified secondary structure element j, as illustrated in Discussion. These secondary structural units are built up from 12352 secondary structure elements. In the 125 analyzed proteins, we found 220 particular residues linking two secondary structural units out of α-turns, β-turns, and helices. Thus, the total secondary structural units are built up from 52% (12352 − 220 = 12132) of the amino acid residues out of those (23286) found in the 125 analyzed proteins.

In order to examine the similarity of the amino acid locations among the type-I α-turns, the type-I β-turns, and the ends of helices, single amino acid preferences for specific locations were compared among them. The position-specific preferences of each single common amino acid at these N- and C-termini are assembled in Table 2. The values for the rarer amino acids, such as C, H, M, and W, are statistically insignificant because the data base is not large enough. Position-specific preferences of each amino acid are tabulated for the caps (N_0 and C_0) at each end, one position out (N′ and C′) from the ends, and one position in (N_1 and C_1) from the ends. Each position of the type-I α- and β-turns is illustrated in Discussion. For each position the left-hand column is data for type-I β-turns, the central column for type-I α-turns, and the right-hand column for the ends of helices (adapted from Ref. 20).

The upper entry is the observed occurrence number of each amino acid at the 6 positions (N′, N_0, N_1, C_1, C_0, and C′) for each secondary structural unit. The middle entry is the average percentage composition of each amino acid at the 6 positions. The lower entry is a normalized preference (NP) value defined by the ratio of the average percentage composition at the 6 positions to the average percentage composition at large.

Table 3 lists the average percentage composition of each amino acid found in the 125 analyzed proteins of Subset A. The upper entry is the number of observed occurrences of each amino acid out of the total number (23286). The lower entry is the average percentage composition at large.

Discussion

The α-turns defined by Toniolo[21] are essentially identical with the 4-turns.[7] Corresponding to the classification of both β-turns[5] and α-turns,[8] the α-turns are classified into the following 5 types: type-I (double type-I β-turns), type-I′ (double type-I′ β-turns), type-II (successive type-II and -I′ β-turns), type-II′ (successive type-II′ and -I β-turns), and type-III (miscellaneous) α-turns. The type-I′ and -II′ α-turns are main-chain mirror images of the type-I and -II α-turns, respectively, and would be disfavored due to steric consideration. Similar to type-I′ and -II′ β-turns, they would be minor secondary structural units in globular proteins. According to Kabsch and Sander's automatic assignments,[7] which are most widely used for secondary structure assignments, all of

Table 2. Positions-Specific Amino Acid Preferences at Type-I α- and β-Turns and the Ends of Helices[a)]

	N' position			N_0 position			N_1 position			C_1 position			C_0 position			C' position		
	β-turn	α-turn	helix	β-turn	α-turn	helix	β-turn	α-turn	helix	β-turn	α-turn	helix	β-turn	α-turn	helix	β-turn	α-turn	helix
A	22	12		17	12		41	21		25	13		35	15		23	13	
	5.1	5.3		3.9	5.3		9.4	9.3		5.7	5.8		8	6.7		5.3	5.8	
	0.6	0.6	0.8	0.5	0.6	0.2	1.1	1.1	1.3	0.7	0.7	1.6	1	0.8	0.8	0.6	0.7	0.9
C	11	5		15	5		7	5		7	10		8	8		5	2	
	2.5	2.2		3.4	2.2		1.6	2.2		1.6	4.4		1.8	3.6		1.2	0.9	
	1.4	1.2	0.4	1.9	1.2	0.8	0.9	1.2	1.5	0.9	2.5	2.6	1	2	0.4	0.6	0.5	0.8
D	17	13		82	48		29	11		78	33		29	13		38	11	
	3.9	5.8		19	21		6.7	4.9		18	15		6.7	5.8		8.8	4.9	
	0.7	1	0.8	3.1	3.6	2.7	1.1	0.8	0.8	3	2.4	1.0	1.1	1	1.1	1.5	0.8	1.1
E	20	9		10	6		32	18		36	18		17	2		19	15	
	4.6	4		2.3	2.7		7.4	8		8.3	8		3.9	0.9		4.4	6.7	
	0.8	0.7	0.6	0.4	0.4	0.7	1.2	1.3	2.0	1.4	1.3	1.1	0.6	0.1	0.2	0.7	1.1	0.3
F	31	22		11	7		9	2		14	4		15	12		11	9	
	7.2	9.8		2.5	3.1		2.1	0.9		3.2	1.8		3.4	5.3		2.5	4	
	1.8	2.5	1.3	0.6	0.8	0.3	0.5	0.2	0.8	0.8	0.5	0.5	0.9	1.4	0.9	0.7	1	0.5
G	34	11		48	25		14	5		28	11		104	35		23	22	
	7.9	4.9		11	11		3.2	2.2		6.4	4.9		24	16		5.3	9.9	
	1	0.6	1.5	1.4	1.4	2.0	0.4	0.3	0.2	0.8	0.6	0.3	3	1.9	3.7	0.7	1.2	1.5
H	6	6		20	12		5	5		8	6		7	5		9	6	
	1.4	2.7		4.6	5.3		1.1	2.2		1.8	2.7		1.6	2.2		2.1	2.7	
	0.6	1.2	0.9	2	2.3	1.4	0.5	1	0.6	0.8	1.2	1.7	0.7	1	2.3	0.9	1.2	0.9
I	29	23		6	4		19	15		5	6		11	12		20	10	
	6.7	10		1.4	1.8		4.4	6.7		1.1	2.7		2.5	5.3		4.6	4.5	
	1.2	1.8	1.1	0.2	0.3	0.1	0.8	1.2	0.8	0.2	0.5	0.4	0.5	1	0.6	0.8	0.8	1.0
K	32	10		13	9		31	17		24	15		16	12		40	21	
	7.4	4.4		3	4		7.1	7.6		5.5	6.7		3.7	5.3		9.2	9.4	
	1.1	0.7	1.5	0.5	0.6	0.5	1.1	1.1	0.7	0.8	1	1.4	0.6	0.8	1.2	1.4	1.4	1.5
L	39	14		12	8		23	14		13	17		29	16		14	8	
	9	6.2		2.8	3.6		5.3	6.2		3	7.6		6.7	7.1		3.2	3.6	
	1.1	0.8	1.4	0.3	0.4	0.2	0.7	0.8	0.8	0.4	0.9	0.4	0.8	0.9	1.0	0.4	0.4	0.7
M	9	2		3	0		3	1		4	2		5	9		1	5	
	2.1	0.9		0.7	0		0.7	0.4		0.9	0.9		1.1	4		0.2	2.2	
	1	0.4	1.8	0.3	0	0.6	0.3	0.2	0.0	0.5	0.4	0.6	0.6	2	1.2	0.1	1.1	0.3
N	16	7		46	36		16	4		45	15		24	8		29	13	
	3.7	3.1		11	16		3.7	1.8		10	6.7		5.5	3.6		6.7	5.8	
	0.8	0.7	0.6	2.2	3.4	3.3	0.8	0.4	0.4	2.2	1.4	0.8	1.2	0.8	1.5	1.4	1.2	1.0
P	16	8		24	16		91	48		2	1		1	0		40	22	
	3.7	3.6		5.5	7.1		21	21		0.5	0.4		0.2	0		9.2	9.9	
	0.8	0.8	1.0	1.2	1.6	0.7	4.6	4.7	2.6	0.1	0.1	0.0	0.1	0	0.0	2.1	2.2	2.7
Q	16	7		9	3		14	7		16	10		15	7		27	4	
	3.7	3.1		2.1	1.3		3.2	3.1		3.7	4.4		3.4	3.1		6.2	1.8	
	1	0.8	0.8	0.5	0.4	0.2	0.8	0.8	0.8	1	1.2	2.2	0.9	0.8	0.5	1.6	0.5	1.5
R	18	5		5	7		15	5		16	4		10	9		20	7	
	4.2	2.2		1.1	3.1		3.4	2.2		3.7	1.8		2.3	4		4.6	3.1	
	1	0.6	0.8	0.3	0.8	0.4	0.9	0.6	0.5	0.9	0.4	1.0	0.6	1	0.9	1.2	0.8	1.3
S	22	6		51	12		40	19		43	14		38	14		32	15	
	5.1	2.7		12	5.3		9.2	8.4		9.9	6.2		8.7	6.2		7.4	6.7	
	0.8	0.4	0.8	1.8	0.8	2.3	1.4	1.3	0.8	1.5	0.9	1.9	1.3	0.9	0.6	1.1	1	1.0
T	27	13		37	5		21	11		41	29		29	12		39	17	
	6.2	5.8		8.5	2.2		4.8	4.9		9.4	13		6.7	5.3		9	7.6	
	1	1	1.1	1.4	0.4	2.0	0.8	0.8	0.8	1.6	2.1	0.7	1.1	0.9	0.7	1.5	1.3	0.6
V	45	27		11	5		14	10		12	7		23	12		21	12	
	10	12		2.5	2.2		3.2	4.4		2.8	3.1		5.3	5.3		4.8	5.4	
	1.5	1.7	0.8	0.4	0.3	0.1	0.5	0.6	1.4	0.4	0.4	0.7	0.8	0.8	0.3	0.7	0.8	0.7
W	7	9		3	0		3	2		5	3		8	7		7	1	
	1.6	4		0.7	0		0.7	0.9		1.1	1.3		1.8	3.1		1.6	0.4	
	1.2	3.1	0.5	0.5	0	0.5	0.5	0.7	2.6	0.9	1	0.0	1.4	2.4	0.5	1.2	0.3	0.5
Y	16	16		12	5		8	5		13	7		11	17		15	10	
	3.7	7.1		2.8	2.2		1.8	2.2		3	3.1		2.5	7.6		3.5	4.5	
	1.1	2.1	0.7	0.8	0.7	0.5	0.5	0.7	1.7	0.9	0.9	1.8	0.7	2.2	0.7	1	1.3	0.4

a) For each position the left-hand column is data for type-I β-turns, the central column for type-I α-turns, and the right-hand column for the ends of helices (adapted from Ref. 20). The upper entry is the observed occurrence number of each amino acid at the 6 positions N′, N_0, N_1, C_1, C_0, and C′ for each secondary structural unit. The middle entry is average percentage composition of each amino acid at the 6 positions. The lower entry is a normalized preference (NP) value defined by the ratio of average percentage composition at the 6 positins to average composition at large.

Table 3. Average Percentage Composition of Each Amino Acid Found in the 125 Analyzed Proteins of Subset A

Amino acid	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
The observed number[a)]	1931	418	1392	1412	902	1863	534	1302	1531	1887	462	1103	1046	876	935	1545	1398	1646	314	789
Average percentage (%)	8.3	1.8	6.0	6.1	3.9	8.0	2.3	5.6	6.6	8.1	2.0	4.7	4.5	3.8	4.0	6.6	6.0	7.1	1.3	3.4

a) The total number of amino acids found in the 125 analyzed proteins of Subset A is 23286.

the α-turns are classified only as the 4-turn. It should be emphasized that the type-I α-turn is an independent kind of secondary structural units, although it has been frequently classified as $(i, i+1)$ double β-turns.[18,19]

The typical type-I α-turns extracted in this study were formed by an isolated 5→1 hydrogen bond or 2 successive 4→1 hydrogen bonds (the minimal 3_{10}-helix on DSSP[7]), and consisted of 5 residues. A structural similarity between the type-I α-turn formed by a 5→1 hydrogen bond and the minimal 3_{10}-helix on DSSP of 2 successive type-I β-turns can be recognized by using ϕ and ψ representations of the 3-dimensional structures of the analyzed proteins. Local patterns of their intrasegments could be readily superimposed to recognize any structural similarity between them. As a result, both secondary structural units were extracted as typical type-I α-turns in this study.

The type-I β-turn is a nucleus of the type-I α-turn and a helix, and the double type-I β-turns form the type-I α-turn. By a repetitive addition of the type-I β-turn to the nucleus, short helical segments in protein chains propagate step-by-step to hexa- and heptapeptide sequences and so forth. Short α-helices, identified as hexa- and heptapeptide sequences, include secondary structural units of 2 and 3 successive type-I α-turns, respectively. Frequently, hexa- and heptapeptide helices are also formed by 3 and 4 successive type-I β-turns, respectively. The $(i, i+1, i+2, i+3)$ quadruple β-turns, for example, consist of 7 residues, such that residues 1-4, residues 2-5, residues 3-6, and residues 4-7 form 4 successive type-I β-turns, and the $(i, i+1, i+2)$ triple type-I α-turns consist of 7 residues, such that residues 1-5, residues 2-6, and residues 3-7 form 3 successive type-I α-turns. However, the difference between 4 successive 4→1 hydrogen bonds and 3 successive 5→1 hydrogen bonds is ambiguous due to the gradual transformation from the former to the latter. Here, 4→1 and 5→1 hydrogen bonds are the basic repetitive turns of a 3_{10}-helix and of a right-handed α-helix, respectively.

Both the nucleation of a helix with the type-I β-turn and its propagation with a repetitive addition of the type-I β-turn suggest that the origin of the helix in a globular protein be the type-I β-turn. Type-I β-turns, type-I α-turns, and helices are classified as repetitive secondary structural units, and are of similar kin because the backbone dihedral angles of their intrasegments have repeating values near to the canonical value of (−60°, −40°). The N cap and C cap residues of the type-I α-turn and the type-I β-turn as well as those of helices depart from the helical values of the ϕ and ψ angles.

Single amino acid preferences for specific locations at helices have been reported to be characteristic, especially at their N- and C-terminal ends.[13,20] In order to examine the similarity of amino acid locations among the type-I α-turns, type-I β-turns, and the ends of helices, the position-specific preferences of each single common amino acid at the type-I β-turns and the type-I α-turns are compared with those at the ends of the helices in Table 2. The position-specific preferences of each of 20 single common amino acid residues can be distinctly compared by using their NP-values at each position of the secondary structural units. The NP-value in Table 2 is a better measure of amino acid selection at each position of the secondary structural units. An NP-value of unity for a single amino acid residue (Q), for example, at the N′ position of the β-turn, will then mean that the single Q residue is observed at the position the same as often as at large.

There are highly position-specific preferences of single amino acid residues that are in common among both the turns and the ends of helices. A single P residue is the most strongly preferred amino acid residue at the N_1 positions of the 3 kinds of secondary structural units, and the NP-value of a P residue, for example, at the N_1 position of type-I α-turn is 4.6. This NP-value would mean that the single P residue occurs at this position 4.6-times as often as at large. This is due to the fact that the inherent distinction on ϕ to about −60° corresponds to the $\phi(N_1)$ requirement for this unit. The NP-values of single D and N residues at the N_0 positions of the 3 kinds of structural units are also found to be high. These N_0 positions are dominated by the D and N residues, and to a lesser extent by the G residue. These 3 amino acid residues account for 48% of the N_0 position of the α-turns found. The A, E, F, I, K, L, Q, R, V, and Y residues appear to be disfavored at these N_0 positions. The D, F, G, I, L, N, Q, R, and T residues also seem to be disfavored at the N_1 positions. Position N′ prefers hydrophobics. All 3 kinds of secondary structural units favor F and I at position N′. The A, D, E, N, and S residues are found to be unfavorable at the N′ position.

Some strong and highly position-specific preferences were also observed at the C-terminal ends of the 3 kinds of secondary structural units. The C_1 position of each unit shows a preference for D, E, (N), Q, (S), and (T), a weak preference for K and Y, and a bias against F, G, I, P, and V. For example, at the C_1 position of the β-turn, P occurs 0.1-times as frequently as does at large. At the C_0 position a very strong preference for G is observed. In contrast, an NP-value of a single P residue at the C_0 positions of the 3 kinds of structural units is zero. This value means that a P residue does not nearly occur at their C_0 positions. A weak preference for D is found at the C_0 positions. The E, P, Q, and V are selected against at the C_0 positions. No significant selection for A has been seen at any position of the C-terminal ends. The C′ position of each unit is significantly enriched for P, with a lesser preference for K, N, and S at this site.

The similarity of the position-specific preferences of single amino acid residues among both turns and the ends of helices suggests that the type-I α-turn, the type-I β-turn, and the helix are of similar kin. Similarities in the structure and amino acid preferences among them suggest that each of amino acid residues on the defined positions along both turns could be properly assigned to one of the helix elements (a, b, h, and i) as follows:

4 kinds of defined positions along the β-turn	(N′)	N_0	N_1	C_1	C_0	(C′)
type-I β-turn, amino acid sequence	$i-1$	i	$i+1$	$i+2$	$i+3$	$i+4$
4 kinds of helix elements	—	a	b	h	i	—

5 kinds of defined positions along the α-turn	(N′)	N_0	N_1	$i+1$	C_1	C_0	(C′)
type-I α-turn, amino acid sequence	$i-1$	i	$i+1$	$i+2$	$i+3$	$i+4$	$i+5$
5 kinds of helix elements	—	a	b	j	h	i	—

The amino acid residue in the center ($i+2$) of type-I α-turns is assigned to the newly classified secondary structure element j as shown above.

We thank C. Sander's research group and C. A. Chothia's research group for the use of their data bases, and also thank all those crystallographers who have deposited their hard-earned coordinate sets in the PDB.

References

1 M. Narita, K. Sode, S. Ohuchi, Y. Murakawa, and M. Hitomi, *Bull. Chem. Soc. Jpn.*, **71**, 385 (1998).

2 L. S. Itzhaki, D. E. Otzen, and A. R. Fersht, *J. Mol. Biol.*, **254**, 260 (1995).

3 A. R. Viguera, F. J. Blanco, and L. Serrano, *J. Mol. Biol.*, **247**, 670 (1995).

4 L. Pauling, R. B. Corey, and H. R. Branson, *Proc. Natl. Acad. Sci. U.S.A.*, **37**, 205 (1951).

5 J. S. Richardson, *Adv. Protein Chem.*, **34**, 167 (1981).

6 C. M. Venkatachalam, *Biopolymers*, **6**, 1425 (1968).

7 W. Kabsch and C. Sander, *Biopolymers*, **22**, 2577 (1983). We used Secondary Structure Definition Program (DSSP) data base of C. Sander's research group.

8 V. Pavone, G. Gaeta, A. Lombardi, F. Nastri, O. Maglio, C. Isernia, and M. Saviano, *Biopolymers*, **38**, 705 (1996).

9 D. J. Barlow and J. M. Thornton, *J. Mol. Biol.*, **201**, 601 (1988).

10 K. R. Rajashankar and S. Ramakumar, *Protein Sci.*, **5**, 932 (1996).

11 C. Schellman, in "Protein Folding," ed by R. Jaenicke, Elsevier, Amsterdam (1980), pp. 53—61.

12 H. A. Nagarajaram, R. Sowdhamini, C. Ramakrishnan, and P. Balaram, *FEBS Lett.*, **321**, 79 (1993).

13 J. S. Richardson and D. C. Richardson, *Science*, **240**, 1648 (1988).

14 F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.*, **122**, 535 (1977).

15 M. Narita, K. Sode, S. Ohuchi, M. Hitomi, and Y. Murakawa, *Bull. Chem. Soc. Jpn.*, **70**, 1639 (1997).

16 W. R. Taylor, *Protein Eng.*, **2**, 77 (1988).

17 A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, *J. Mol. Biol.*, **247**, 536 (1995).

18 Y. Isogai, S. J. Leach, and H. A. Scheraga, *Biopolymers*, **19**, 1183 (1980).

19 E. G. Hutchinson and J. M. Thornton, *Protein Sci.*, **3**, 2207 (1994).

20 S. Dasgupta and A. B. Bell, *Int. J. Peptide Protein Res.*, **41**, 499 (1993).

21 C. Toniolo, *CRC Crit. Rev. Biochem.*, **1980**, 1.