Assingments of Tri- and Tetrapeptide Sequences in Globular Proteins to the 18 Kinds of Local Structures along Helices and Their Propensities for Specific Local Structures

Mitsuaki Narita,* Akihiko Mochizuki, and Shoukichi Ohuchi#

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

(Received August 9, 1999)

The genetic information for local structures along helices, encoded in local sequences of protein chains, was statistically investigated for tri- and tetrapeptide sequences (3- and 4-letter words) using the 411 analyzed protein chains. The 18 kinds of local structures adopted by tri- and tetrapeptide sequences were represented 1-dimensionally by using a single helix element and pairs of helix elements in parentheses, respectively. The local sequences have propensities for none to some specific local structures along helices. A local structure (LS)-value is introduced for the evaluation of the normalized preference (NP)-value of a local sequence for a particular local structure. It should be emphasized that N-capping tetrapeptide sequences of helices do not necessarily prevent helix elongation. Local structures adopted by tetrapeptide sequences are plastic, and a particular local structure along helices is determined by the sequence context of helices.

One-dimensional amino acid sequences encode the information required to specify 3-dimensional final structures of proteins,[1,2] but how the amino acid sequences determine protein folding is a major unsolved problem in structural biology and chemistry.[3—10] The folding is likely to be initiated through local structure formation in many regions of the polypeptide chain during its earliest stages.[3—7] Local structures are formed that depend only on local sequences, and local structure formation is independent of long-range interactions.[7] Therefore, decoding the genetic information for 3-dimensional structures of proteins appears to require that the genetic information for local structures, encoded in local sequences, should be decoded. However, so far, even correlating local sequences with local structures[11—18] remains an important area of research, although the relationship between tripeptide sequences and local structures along helices has been compiled using 125 unrelated proteins.[11]

Protein structures are arranged in a structural hierarchy: the primary, secondary, supersecondary, and tertiary structures. The primary structure is an amino acid sequence, and each amino acid in the sequence is allotted to one of the secondary structural elements, which are building blocks of protein structures.[11] The secondary structure is local structures which include the basic units, such as helices and β-strands as well as α-, β-, and π-turns. These secondary structural units are built up from a variety of secondary structural elements. Here, the term "secondary structure element"[6,7] is frequently used for a "secondary structural unit" in this study. While, the supersecondary structure is a recurring motif of a combination of secondary structural units[18] and includes β-hairpins, β-α-β units, β-sheet topologies, and so forth. These secondary and supersecondary structures are assembled in various ways to form the tertiary structure.

Recently, Thornton and collaborators have carried out important work in characterizing the secondary structural units and their recurring motifs.[18] On the other hand, in order to analyze helices in proteins precisely, we have introduced both the 11 kinds of helix elements and the 8000 kinds of amino acid residues in the middle of tripeptide sequences.[11] As a result, we could compile their preferences for none to some of specific helix elements using the 125 analyzed proteins. For the purpose of compilation, it was quite important that the secondary structural elements constructing proteins were classified as simply as possible in order to characterize the amino acid residues by using secondary structural elements without complexity. Furthermore, structural similarity among type-I α-turns, type-I β-turns, and short helices was illustrated by both the nucleation of a helix with a type-I β-turn and its propagation by the repetitive addition of type-I β-turns.[19] The similarity of amino acid preferences for specific locations at helices, type-I α-, and β-turns also suggested that each amino acid on defined positions along both turns should be properly assigned to one of the helix elements.[19] The end residues of both turns as well as those of helices[20] were termed to be N-cap (N_0) and C-cap (C_0), and the definition of N-cap and C-cap residues of type-I α- and β-turns was identical to that of helices. It also suggests the propriety of the classification of secondary structural elements

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

from both turns as helix elements. In this study a statistical investigation of the relationship between local sequences and local structures was performed under the classification of secondary structural elements from type-I α- and β-turns as helix elements using the 411 analyzed protein chains. The genetic information for local structures along helices, encoded in tri- and tetrapeptide sequences, is illustrated as 3- and 4-letter words, respectively, meaning the propensities of local sequences for none to some of the specific local structures along helices.

Results and Discussion

Extraction of Helices, Type-I α-, and β-Turns from the 411 Analyzed Protein Chains. For the purpose of choosing the 411 analyzed protein chains, sequence alignments of sequentially identical hexapeptides were performed so as to exclude homologous proteins. Distantly related homologous proteins included in the 411 protein chains, which are classified into the same families according to Murzin,[21] do not bias the results in this study. The 3191 helices, the 975 type-I α-turns, and the 1890 type-I β-turns were easily extracted from the 411 protein chains by using both 2-dimensional ϕ and ψ representations of their 3-dimensional structures[12,19] and Kabsch and Sander's automatic assignments based on a hydrogen bond.[22] Upon the extraction of these secondary structural units, at first, their N-cap and C-cap residues were defined in terms of the backbone dihedral ϕ and ψ angles near to the helical values, irrespective of the presence or absence of a hydrogen bond. Intrasegments of type-I α- and β-turns as well as those of helices[23] comprise ϕ angles of between $-130°$ and $-20°$ and ψ angles of between $-110°$ and $+20°$.[19] However, the N-cap and C-cap residues of both turns as well as those of helices depart from the helical values of the ϕ and ψ angles. A small fraction of the extracted α- and β-turns are actually free from a hydrogen bond defined by Kabsch and Sander.[22] A variety of G residues in a protein have unique dihedral angles, ϕ and ψ, which are rarely allowed for other amino acid residues. Thus, they are useful for assigning the N-cap and C-cap residues as an internal standard.

Next, according to Kabsch and Sander's automatic assignments based on a hydrogen bond, the 3 kinds of secondary structural units were extracted even though their intrasegments comprise dihedral ϕ or ψ angle that are slightly deviated from the helical values. Thus, distorted secondary structural units make their end residues ambiguous. The definition of a type-I β-turn in this study is essentially identical to the widely used definition of type-I β-turn by Lewis et al.[24] Type-I α-turns extracted include an isolated $5 \rightarrow 1$ hydrogen bond, 2 successive $4 \rightarrow 1$ hydrogen bonds or an isolated $4 \rightarrow 1$ hydrogen bond, although a part of the α-turns are free from a hydrogen bond.[19]

The Importance of 1-Dimensional Representation of 3-Dimensional Helices by Using Helix Elements. A one-dimensional representation of 3-dimensional secondary structural units such as helices, type-I α-, and β-turns is essential to decoding the genetic information for 3-dimensional local structures encoded in 1-dimensional local sequences. The representation makes it possible for sequences of amino acid residues in proteins to be translated into sequences of secondary structural elements constructing protein structures, since each amino acid in a sequence is assigned to one of the secondary structural elements.[11,12] In this study, a sequence of amino acid residues along a helix identified as a dodecapeptide is 1-dimensionally translated into a sequence of the 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	0	1	2	3	4	5	6	7	8	9	10	11	12 13
helix element sequence	—	a	b	c	d	e	e	e	e	f	g	h	i —

The 9 kinds of defined positions (N_0–C_0) are labeled on the basis of the N_0 and C_0 positions, and the amino acid residues (1—12) located on the defined positions (N_0–C_0) are assigned to helix elements a—i, as shown above. The N-cap and C-cap residues of both turns as well as the helices are assigned to elements a and i, respectively. Each amino acid residue along type-I α- and β-turns is assigned to one of the helix elements (a, b, h, i, and j) as follows:

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

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

The 1-dimensional sequence of helix elements comprising the 12-residue α-helix, for example, encodes the information required to specify the 3-dimensional 12-residue α-helix on an amino acid level. Amino acid residues (1—12) assigned to helix elements a—i have 3-dimensional characteristics according to the defined positions (N_0–C_0). Similarly, 3-dimensional helices consisting of a variety of chain lengths can be specified by using 1-dimensional sequences of helix elements, and amino acid residues assigned to helix elements have 3-dimensional characteristics. Consequently, 1-dimensional local sequences can be assigned to 3-dimensional local structures on an amino acid level.

For hydrogen bonds, $5 \rightarrow 1$, $6 \rightarrow 2$, $7 \rightarrow 3$, $8 \rightarrow 4$, $9 \rightarrow 5$, $10 \rightarrow 6$, $11 \rightarrow 7$, and $12 \rightarrow 8$ hydrogen bonds along the 12-residue α-helix are formed successively. The initial 4 NH groups (residues 1—4) do not form intrahelical hydrogen bonds; neither do the final 4 CO groups (residues 9—12) at the C-terminus. In the middle (residues 5—8) of the helix, intrahelical hydrogen bonds are formed through both 4 NH and 4 CO groups successively. For dihedral ϕ and ψ angles, each N-cap (1) and C-cap (12) residue always departs from the helical values of the ϕ and ψ angles, and the residues (2—11) of an intrahelical segment always have dihedral ϕ and ψ angles within the helical values.

The Genetic Information for Local Structures Illustrated as 3-Letter Words Meaning Propensities of Tripep-

tide Sequences for Some of Specific Local Structures. New approaches to correlating local sequences with local structures along helices are indispensable in order to decode the information required to specify helix units in proteins. As mentioned above, a 1-dimensional representation of helices is useful for this purpose. Since each amino acid residue constructing helices in proteins is regarded as being one of the 8000 kinds of amino acid residues in the middle of tripeptides, and is assigned to one of the 9 kinds of helix elements, the 9 kinds of local structures adopted by tripeptide sequences can be represented by single helix elements in parentheses as local structures (a)—(i).

As an example, the amino acid sequence shown below is used to correlate tripeptide sequences with local structures along helices. It corresponds to the sequence from residues 24 to 36 in the protein (3icb)[25] and forms a 13-residue helix in 3icb. The correlation of tripeptide sequences with the 9 kinds of local structures along helices is represented using local sequences and single helix elements in parentheses, as follows: LSK (a), SKE (b), KEE (c), EEL (d), ELK (e), LKL (e), KLL (e), LLL (e), LLQ (e), LQT (f), QTE (g), TEF (h), and EFP (i).

defined positions	N'	N_0	N_1	N_2	N_3	M	M	M	M	M	C_3	C_2	C_1	C_0	C'
amino acid sequence	L	S	K	E	E	L	K	L	L	L	Q	T	E	F	P
helix element sequence	—	a	b	c	d	e	e	e	e	e	f	g	h	i	—

The 411 analyzed protein chains consist of 101122 total single amino acid residues. Since they comprise 110 chain breaks, they consist of 100102 total tripeptide sequences, and comprise 7769 independent sequences among the 8000 kinds of ones. They also consist of 99593 total tetrapeptide sequences. In any case, an inspection of the amino acid sequences always allows an unambiguous identification of the tri- (100102) and tetrapeptide (99593) sequences based on DSSP,[22] as mentioned in Methods. Conversely, protein chains can be assembled by successive hybridization of adjacent local sequences.

As previously described,[11] by compiling the relationship between tripeptide sequences and the local structures in the 411 protein chains, the 8000 kinds of tripeptide sequences can be characterized by their propensities for none to some of specific local structures along helices. Protein chains can be assembled by the successive hybridization of adjacent tripeptide sequences, and the assembly of local structures adopted by successive tripeptide sequences specifies helices in proteins depending on the sequence context of helices, since the folding of many proteins in vitro is an indication that the process of folding is dictated only by the amino acid sequence. Therefore, the relationship between the local sequences and the local structures is identical to the genetic information required to specify helices. The genetic information for local structures, encoded in tripeptide sequences, can be illustrated as 3-letter words meaning the propensities of the tripeptide sequences for none to some of the specific local structures along helices, since each of 20 single common amino acid residues in protein sequences is denoted by a single-letter code.

Local Structures Adopted by Tripeptide Sequences (3-Letter Words) in the 411 Analyzed Protein Chains. By using a correlation of 100102 tripeptide sequences with the 9 kinds of local structures, (a)—(i), and other local structure (–), 5987 and 6030 out of 100102 tripeptide sequences, for example, are assigned to the local structures (a) and (b), respectively. The total observed occurrence numbers of each local structure along helices are according to the following: (a), 5987, 6.0%; (b), 6030, 6.0%; (c), 3166, 3.2%; (d), 2662, 2.7%; (e), 19038, 19.0%; (f), 2615, 2.6%; (g), 3119, 3.1%; (h), 5982, 6.0%; and (i), 5938, 5.9%. For example, 5987 of local structures (a) adopted by tripeptide sequences is 6.0% of the total observed occurrence number (100102) of the local structures adopted by the total tripeptide sequences. The 4 kinds of local structures ((a), (b), (h), and (i)) include those along the type-I α- and β-turns. The total occurrence number (54537) of the 9 kinds of local structures is 54% of that (100102) of those adopted by the total tripeptide sequences. Frequently, tripeptide sequences adopt both the local structures (a) and (i) at the same time. Practically, 939 out of 5987 tripeptide sequences are assigned to both the local structures (a) and (i).

A local sequence-local structure dictionary for 3-letter words could be compiled using the relationship between the 100102 tripeptide sequences and the local structures. The dictionary is expected to be used for decoding the genetic information for helices, encoded in a sequence of unrelated proteins.

Definition of a Local Structure (LS)-Value. A local structure (LS)-value is introduced to evaluate local sequence preferences for a specific local structure. To evaluate the local sequence preferences for a specific local structure, at first, the N-cap structure adopted by N-capping tripeptide sequences of helices is represented by using the single helix element a in parentheses as the N-cap structure (a). The LS-value of the N-cap structure (a) is defined as the average percentage of its occurrence number (5987) in the data set based on the total occurrence number (100102) of local structures adopted by the tripeptide sequeneces in the data set. It is defined by Eq. 1 :

$$\text{LS-value of the N-cap structure (a)} = \frac{\text{The occurrence number of the N-cap structure (a) adopted by N-capping tripeptides in the data set.}}{\text{The occurrence number of local structures adopted by total tripeptides in the data set.}} \times 100. \quad (1)$$

Therefore, the LS-values of the N-cap structure (a) and the C-cap structure (i) of this data set, for example, are 6.0 and 5.9, respectively. The LS-values are entirely dependent on the data sets.

Evaluation of Local Sequence Preferences for Specific Local Structures. As previously described,[12] each of the 8000 kinds of amino acid residues in the middle of tripeptide sequences has its inherent preference (IP)-values for some specific secondary structural elements. The IP-value of an

amino acid residue for a particular secondary structural element was the average percentage of its observed occurrence number at the defined position assigned to the particular secondary structural element on the basis of its total observed occurrence number at large.

Similarly, we can define inherent preference (IP)-values of a local sequence for particular local structures. The IP-value of an N-capping tripeptide sequence of helices for the N-cap structure (a), for example, is also the average percentage of its observed occurrence number at the N-cap structure (a) on the basis of its total observed occurrence number at large. Table 1 lists those N-capping tripeptide sequences having strong propensities for the N-cap structure (a) and their IP-values for the N-cap structure. It also lists their observed occurrence numbers both at the N-cap structure and at large. The IP-values of the tripeptide sequences in Table 1 for the N-cap structure are 40 or greater, indicating their strong propensities for the N-cap structure. They are observed at the N-cap structure 4-times or more often.

In Table 1, 556 (9.3%) out of 5987 tripeptide sequences which adopt the N-cap structure (a) are listed as a typical example. Table 1 indicates that 70 (0.9%) out of 8000 kinds of tripeptide sequences occupy 9.3% of the N-cap structure (a) in the 411 protein chains. Tripeptide sequences such as KHP, TDP, VHP, HDP, WDP, NDP, CDP, DDP, and LSP in protein sequences preferentially occupy the N-cap structure. Other tripeptide sequences listed in Table 1 also adopt preferentially the N-cap structure among the 9 kinds of local structures. Distinctly, the genetic information for local structures along helices, encoded in tripeptide sequences, is valuable to decode the information for helices, encoded in protein sequences.

A normalized preference (NP)-value of a tripeptide sequence for a corresponding particular local structure is de-

Table 1. N-Capping Tripeptide Sequences for the N-Cap Structure (a) of Helices, Type-I α-, and β-Turns

N-Capping tripeptide sequence[a)]	Observed occurrence number		IP-value[b)]	N-Capping tripeptide sequence[a)]	Observed occurrence number		IP-value[b)]
	at N-cap	at large			at N-cap	at large	
KHP	7	8	88	RDP	7	13	54
TDP	15	18	83	EDP	8	15	53
VHP	5	6	83	TNP	9	17	53
HDP	8	10	80	YNP	9	17	53
WDP	4	5	80	GNP	10	19	53
NDP	14	18	78	CGP	4	8	50
CDP	6	8	75	DNP	9	18	50
DDP	9	12	75	EHP	4	8	50
LSP	14	19	74	FDE	7	14	50
ADP	12	17	71	FDP	7	14	50
KDP	14	20	70	PMP	4	8	50
IDP	11	16	69	RSP	5	10	50
LHP	8	12	67	VPQ	6	12	50
MPK	4	6	67	SDP	9	19	47
QHP	4	6	67	SSP	9	19	47
RNP	8	12	67	ENP	7	15	47
WNP	4	6	67	TDA	14	30	47
YDP	8	12	67	APP	5	11	45
AHP	5	8	63	QNE	5	11	45
FNP	10	16	63	NDT	4	9	44
MNP	5	8	63	CSD	7	16	44
LDP	16	26	62	IDQ	7	16	44
GDP	18	30	60	GSP	9	21	43
SHP	6	10	60	LPP	12	28	43
YSP	9	15	60	TSP	9	21	43
ATP	13	22	59	VSD	8	19	42
FNH	4	7	57	YTY	5	12	42
GHP	8	14	57	INP	7	17	41
QDP	4	7	57	KNP	8	20	40
ANP	14	25	56	KTP	4	10	40
YTP	5	9	56	NDE	4	10	40
VNP	11	20	55	NSP	4	10	40
AGP	12	22	55	TDI	6	15	40
PDP	6	11	55	VPP	6	15	40
VDP	13	24	54	YND	4	10	40

a) Amino acid redidues in the middle of tripeptide sequences are assigned to the helix element a. b) The IP-value is defined in text.

fined as the ratio of the average percentage of the tripeptide sequence at the particular local structure to its average percentage at large. Thus, an NP-value of unity means that the frequency of occurrence of a certain tripeptide sequence at a particular local structure is the same as the frequency of its occurrence at large. The NP-values of N-capping tripeptide sequences in Table 1 for the N-cap structure were obtained by dividing the individual IP-values by the LS-value (6.0). As an example, the NP-value of TDP for the N-cap structure was obtained to be 14 by dividing the IP-value (83) of TDP for the N-cap structure by the LS-value (6.0). This means that the tripeptide sequence of TDP is observed at the N-cap structure 14-times as often as at large.

Although the statistical significance of the NP-value is distinct, it is entirely dependent on the data sets. However, the IP-value is inherent to each local sequence. Therefore, the propensities of the tripeptide sequences for a particular local structure can be evaluated more conveniently by using their IP-values rather than their NP-values. The N-capping tripeptide sequences in Table 1 would be valuable to identify the precise location of the N-cap structure (a) of helices, type-I α- and β-turns in protein sequences. Similarly, C-capping tripeptide sequences would be valuable to identify the precise location of their C-cap structure (i).

One-Dimensional Representation of the 9 Kinds of 3-Dimensional Local Structures Adopted by Tetrapeptide Sequences by Using Pairs of Helix Elements. The 160000 (20^4) possible kinds of tetrapeptide sequences can be obtained by hybridization of a pair of adjacent tripeptide sequences. For example, the tetrapeptide sequence of AAGA is obtained by the hybridization of a pair of corresponding tripeptide sequences of AAG and AGA. Similarly, the tetrapeptide sequence of VAAG is obtained from VAA and AAG. When we analyzed helices with the 8000 kinds of amino acid residues in the middle of tripeptide sequences, the middle amino acid residue could be assigned to one of the helix elements.[11] Therefore, successive amino acid residues in the middle of tetrapeptide sequences can be assigned to a pair of helix elements, and local structures adopted by tetrapeptide sequences can be represented by pairs of helix elements in parentheses.

Taking the sequence from residues 24 to 36 in 3icb[25] as an example, the correlation of tetrapeptide sequences with the 9 kinds of local structures along helices is represented by using local sequences and pairs of helix elements in parentheses, as follows: LSKE (ab), SKEE (bc), KEEL (cd), EELK (de), ELKL (ee), LKLL (ee), KLLL (ee), LLLQ (ee), LLQT (ef), LQTE (fg), QTEF (gh), and TEFP (hi). One-dimensional tetrapeptide sequences can be assigned to 3-dimensional local structures on an amino acid level. Amino acid residues assigned to helix elements a—i have 3-dimensional characteristics according to the defined positions (N_0–C_0).

Local Structures Adopted by Tetrapeptide Sequences (4-Letter Words) in the 411 Analyzed Protein Chains. Local structures adopted by tetrapeptide sequences (99593) in the 411 protein chains can be represented by using the 9 kinds of local structures ((ab)—(hi)) and other local structure (—). The local structures (ab) and (bc), for example, are adopted by 5987 and 3166 out of 99593 tetrapeptide sequences (4-letter words), respectively. Both local structures ((ab) and (hi)) include those along type-I α- and β-turns. The observed occurrence numbers of the 9 kinds of local structures along helices and their LS-values are as follows: (ab), 5987, 6.0; (bc), 3166, 3.2; (cd), 2662, 2.7; (de), 2410, 2.4; (ee), 16602, 16.7; (ef), 2362, 2.4; (fg), 2615, 2.6; (gh), 3119, 3.1, and (hi), 5932, 5.9. The total occurrence number (44855) of the 9 kinds of local structures is 45% of that (99593) of those adopted by the total tetrapeptide sequences in the data set. Both the IP- and NP-values of a tetrapeptide sequence for a particular local structure can be evaluated as well as those of a tripeptide sequence. For example, the IP-values of the tetrapeptide sequence of LSKE in Table 2 for the local structures (ab) and (ee) are 63 and 38, respectively, since 5 and 3 out of 8 LSKE are observed at the local structures (ab) and (ee), respectively. Its NP-values for the local structures (ab) and (ee) are 10 and 2.2, respectively, since their LS-values are 6.0 and 16.7.

A local sequence-local structure dictionary for 4-letter words could be compiled using the relationship between 99593 tetrapeptide sequences and the local structures. We can distinctly identify the locations of the tetrapeptide sequences (99593) and the local structures adopted by them in the 411 protein chains. The relationship is identical to the genetic information for the local structures along helices, type-I α- and β-turns, encoded in tetrapeptide sequences of protein chains. The dictionary is expected to be used for decoding the genetic information for helices, encoded in a sequence of unrelated proteins.

Some of Specific Local Structures Adopted by N-Capping Tetrapeptide Sequences of Helices, Type-I α-, and β-Turns. Generally, tetrapeptide sequences are assigned to some of specific local structures represented by pairs of secondary structural elements in parentheses. As an example, some of the specific local structures adopted by each N-capping tetrapeptide sequence of helices, type-I α-, and β-turns are illustrated in Table 2. At the top of Table 2, the 9 kinds of local structures ((ab)—(hi)) and another local structure (—) are arranged. The sequences are observed at the N-cap structure (ab) 3-times or more often in the 411 protein chains, and frequently determine their N-cap structure (ab). Table 2 lists their observed occurrence numbers at the N-cap structure (ab) at a variety of specific local structures and at large. For example, it indicates that N-capping tetrapeptide sequences such as GDPN, TDAT, YDAT, APPE, CGAC, DDPE, GDTE, IDPE, IPRE, and ISEE in protein sequences predominantly occupy the N-cap structure (ab).

In Table 2, 236 (3.9%) out of 5987 N-capping tetrapeptide sequences are listed as a typical example. This indicates that 74 (0.05%) out of 160000 kinds of tetrapeptide sequences occupy 3.9% of the N-cap structure (ab) in the data set. However, 5 and 3 out of 8 LSKE adopt the local structures (ab) and (ee), and 3, 2, 1, and 2 out of 8 ADAA adopt the local structures (ab), (cd), (ee), and (—), respectively. It should be emphasized that N-capping tetrapeptide sequences do not

Table 2. N-Capping Tetrapeptide Sequences for the N-Cap Structure (ab) of Helices, Type-I α-, and β-Turns

N-Capping tetrapeptide sequence[a]	Total occurrence number[b]	Observed occurrence number of tetrapeptide sequences at local structures adopted by them[c]									
		(ab)	(bc)	(cd)	(de)	(ee)	(ef)	(fg)	(gh)	(hi)	(–)
LPPE	7	5									2
LSKE	8	5				3					
LTAD	8	5				1					2
ADPS	5	4									1
ATPA	6	4		1							1
DSLD	6	4								2	
GDPN	4	4									
GSAL	7	4							1		2
INPE	5	4				1					
TDAT	4	4									
YDAT	4	4									
ADAA	8	3		2		1					2
AGPS	6	3									3
APPE	3	3									
ATPR	4	3									1
CGAC	3	3									
CSSE	3	3									
DDPE	3	3									
EGAG	5	3									2
EHPE	4	3									1
GDKV	9	3									6
GDTE	3	3									
GGLT	6	3		1	1						1
GHPE	4	3									1
GTPE	6	3									3
IDPE	3	3									
IDSK	4	3									1
IPAE	5	3									2
IPRE	3	3									
ISAA	7	3		1		2					1
ISEE	3	3									
KDPS	3	3									
KSPE	3	3									
LDFE	5	3				1					1
LDPA	4	3									1
LDPD	4	3									1
LLPD	4	3									1
LNDD	4	3									1
LNLD	4	3									1
LNLQ	4	3									1
LPAE	5	3									2
LSPD	4	3									1
LSPE	3	3									
LSPS	3	3									
LSSQ	6	3				1					2
LSVE	6	3				1	1				1
LTED	4	3					1				
LTPD	4	3									1
LTRA	3	3									
NDPR	3	3									
NDPS	3	3									
NNPN	3	3									
PGIE	4	3									1
PGLG	4	3									1
RDPE	3	3									
SDSK	4	3									1

a) Successive amino acid residues in the middle of tetrapeptide sequences are assigned to pairs of successive helix elements. b) The total observed occurrence number of tetrapeptide sequences found in the data set. c) Local structures along helices are illustrated in text and a symbol (—) represents other local structure except for the 9 kinds of ones.

Table 2. (Continued)

N-Capping tetrapeptide sequence[a]	Total occurrence number[b]	Observed occurrence number of tetrapeptide sequences at local structures adopted by them[c]									
		(ab)	(bc)	(cd)	(de)	(ee)	(ef)	(fg)	(gh)	(hi)	(–)
SGPN	3	3									
SNPA	4	3									1
TDAN	5	3	1		1						
TDQT	3	3									
TKPE	3	3									
TSPE	3	3									
VDAS	6	3									3
VDLS	3	3									
VGKE	3	3									
VGLD	7	3				1					3
VNPS	3	3									
VPAE	3	3									
VSDD	3	3									
VSDE	3	3									
YLPS	3	3									
YNAA	4	3			1						
YNPS	5	3									2
YNSN	3	3									

necessarily prevent helix elongation[26] although N-capping sequences present in long peptide fragments are suggested to prevent helix fraying and to function as helix stop signals.[7]

Plasticity of Local Structures Adopted by Tetrapeptide Sequences. So far, a 1-dimensional representation of local structures has never been developed, and remains an important area of study for decoding the genetic information required to specify 3-dimensional protein structures. In this study we developed a 1-dimensional representation of 3-dimensional local structures along helices by using helix elements. Now, we can demonstrate the plasticity of the local structures adopted by tetrapeptide sequences in details.

Practically, Table 3 assembles the local structures adopted by tetrapeptide sequences, which are found 9-times or more often in the 411 analyzed protein chains. The 9 kinds of local structures are arranged at the top of Table 3. A symbol (—) in Table 3 also represents other local structures. The tetrapeptide sequences listed in Table 3 have propensities for none to 6 out of 9 kinds of local structures. For example, 6 out of 9 kinds of local structures can be adopted by AAAA (20) and LEKA (9), and 5 out of 9 kinds of local structures are observed for AAAE (13), AVAA (12), QAAA (12), and AAEG (10). However, none of the 9 AGES is observed at helices in the 411 protein chains, and 9 out of 11 AAGV, 7 out of 9 EELG, 5 out of 9 ELGL are predominantly observed at local structures (hi), (gh), and (hi), respectively. Similarly, 9 (ee) and 2 (fg) are preferred for 11 EALK, 5 (ee) and 4 (fg) for 10 ELKK, and 2 (de) and 6 (hi) for 9 ELGA. The IP-values of the tetrapeptide sequence of ELGA for the local structures (de) and (hi) are 22 and 67, respectively. Its NP-values for the local structures (de) and (hi) are 9.2 and 11, respectively, and the tetrapeptide sequence of ELGA is observed at the local structures (de) and (hi) 9.2- and 11-times as often as at large, respectively.

Tables 2 and 3 suggest that tetrapeptide sequences greatly limit the number of local structures available to each tetrapeptide sequence of a polypeptide chain. Thus, the genetic information for local structures along helices, encoded in tetrapeptide sequences, would be more valuable for decoding the information for helices encoded in protein sequences than that encoded in tripeptide sequences. The plasticity of the local structures leads us to the concept of context-dependent secondary structure formation.[27,28] As well as the sequence context-dependent helix formation, the assembly of local structures adopted by each tetrapeptide sequence of protein chains would specify their secondary structural units dependently on their sequence context alone, since the folding of many proteins in vitro is an indication that the process of folding is dictated only by the amino acid sequence.

Materials and Methods

Proteins Examined in This Study. Proteins whose atomic coordinates have been entered to the Protein Data Bank (PDB)[29] were used in this study. The protein structures were excluded if data to more than 3.1 Å resolution were used in refinements of the coordinates, or if they were determined by NMR measurements. For the purpose of choosing the nonhomorogous data set of 411 analyzed protein chains (Data Set A), sequence alignments of sequentially identical hexapeptides have been performed so as to exclude homologous proteins. The 411 analyzed protein chains, which are identified by the PDB code,[29] do not bias the results in this study. They are as follows: 153l ; 1aak ; 1ack ; 1acx ; 1add ; 1ads ; 1adt ; 1aep ; 1alk-A ; 1alo ; 1amp ; 1aor-A ; 1aoz-A ; 1apy-A ; 1apy-B ; 1ast ; 1atn-A ; 1avh-A ; 1bam ; 1bbp-A ; 1bco ; 1bcx ; 1bet ; 1bgs-A ; 1bia ; 1bmc ; 1bmt-A ; 1bmv-1 ; 1bmv-2 ; 1bnc-A ; 1bov-A ; 1brl-A ; 1bro-A ; 1btc ; 1bvp-1 ; 1byh ; 1cbg ; 1cby ; 1cc5 ; 1cde ; 1cdt-A ; 1cel-A ; 1cew-I ; 1cgl-A ; 1cgt ; 1chd ; 1chk-A ; 1chm-A ; 1cks-A ; 1clc ; 1cma-A ; 1col-A ; 1cpc-A ; 1crn ; 1cse-I ; 1csm-A ; 1csp ; 1ctf ; 1ctm ; 1ctn ; 1ctt ; 1cus ; 1cyw ; 1d66-A ; 1daa-A ; 1dch-A ; 1ddt ; 1dea-A ; 1dhr ; 1dhy ; 1dih ; 1dik ; 1div ; 1dka ; 1dkz-A ; 1dlc ; 1dnp-A ; 1dog ; 1dpg-A ; 1dsb-A ; 1dup-A ; 1eca ; 1ecl ; 1ecm-A ; 1eft ; 1eri-A ; 1esc ; 1f3g ; 1fbp-A ; 1fd2 ; 1fdl-H ; 1fdx ; 1fha ; 1fjm-A ; 1fkf ; 1fle-E ; 1fmp ; 1fnd ; 1fod-1 ; 1fps ; 1fxi-A ; 1gal ; 1gd1-O ; 1gdh-A ; 1gdj ; 1ggt-A ; 1ghr ; 1gla-G ;

Table 3. Plasticity of Local Structures Adopted by Tetrapeptide Sequences

Tetrapeptide sequence[a]	Total occurrence number[b]	Observed occurrence number of tetrapeptide sequences at local structures adopted by them[c]									
		(ab)	(bc)	(cd)	(de)	(ee)	(ef)	(fg)	(gh)	(hi)	(–)
AAAA	20				2	7	1	1	1	1	7
AAAE	13	1				5	2	2		1	2
AAGA	13					2				4	7
LAAA	13			1		8	2				2
AVAA	12	1			1	6	1	1			2
QAAA	12			1	1	6	1	1			2
VAAA	12			1		5		1	1		4
VAAG	12					2			6		4
AAGV	11									9	2
AALE	11				2	6	1				2
ALLA	11			1		7	2				1
AQAA	11			1		10					
EALK	11					9		2			
QAAL	11			1		8	1				1
AADL	10		1	1		2		1			5
AAEG	10	1		1		2			2	2	2
AALK	10			1		5		1	1		2
ALAA	10				1	5	1			1	2
ALEA	10		1			6	1	2			
ALLD	10			1		3	2				4
EEAL	10			1	1	4	2				2
EELG	10								7		3
ELKK	10					5		4			1
LLEK	10					5		2			3
VIAG	10					1			1	2	6
VIGG	10									1	9
AAAG	9				1	1			2		5
AAGY	9					2	1			3	3
AALR	9					8	1				
AEKL	9					4		2	1		2
AGAA	9				1	4		1			3
AGES	9										9
ALAE	9	1				5	3				
ATAG	9					1			1	1	6
AVRG	9	1				4			1		3
EGSS	9	2								1	6
ELGA	9				2					6	1
ELGL	9									6	3
GDKV	9	3									6
GSAA	9	1	1							1	6
KALA	9				1	4	1		1		2
LEKA	9	1		1		3	1	1	1		1
LKAA	9					4		3	1		1
LKAL	9			1		3		2	1		2
LVGG	9	1		1		1					6
RLSA	9		1			2		2			4
VAGG	9				1						8
VASG	9								3		6
VEAL	9			2		5	1		1		

a) Successive amino acid residues in the middle of tetrapeptide sequences are assigned to pairs of successive helix elements. b) The total observed occurrence number of tetrapeptide sequences in the data set. c) Local structures along helices are illustrated in text and a symbol (—) represents other local structure except for the 9 kinds of ones.

1gln ; 1glu-A ; 1glv ; 1gof ; 1gox ; 1gpb ; 1gpc ; 1gph-1 ; 1gpm-A ; 1grj ; 1grl ; 1gss-A ; 1gtp-A ; 1gtq-A ; 1hav-A ; 1hcn-A ; 1hcn-B ; 1hcr-A ; 1hip ; 1hjr-A ; 1hlc-A ; 1hmy ; 1hpl-A ; 1hst-A ; 1htd-A ; 1hxp-A ; 1hyp ; 1ice-A ; 1ice-B ; 1ice-T ; 1ifb ; 1ign-A ; 1inp ; 1itg ; 1jud ; 1kbp-A ; 1knb ; 1kpt-A ; 1lap ; 1lba ; 1lbd ; 1lbu ; 1leh-A ; 1lfa-A ; 1lis ; 1lmb-3 ; 1lpe ; 1lts-D ; 1lxa ; 1mas-A ; 1mat ; 1mcp-L ; 1mda-H ; 1mda-L ; 1mdy-A ; 1mla ; 1mmd ; 1mmo-B ; 1mrr-A ; 1msa-A ; 1msb-A ; 1msp-A ; 1myp-A ; 1myp-C ; 1nar ; 1nba-A ; 1nci-A ; 1nfp ; 1nip-A ; 1nn2 ; 1occ-E ; 1oen ; 1ord-A ; 1otf-A ; 1otg-A ; 1ovo-A ; 1paz ; 1pda ; 1pdg-A ; 1pdn-C ; 1pdo ; 1pil

; 1pk4 ; 1pkp ; 1plq ; 1png ; 1pnk-A ; 1pnk-B ; 1poc ; 1ppt ; 1pre-A ; 1prt-A ; 1psd-A ; 1pta ; 1ptd ; 1ptq ; 1pvi-A ; 1pxt-A ; 1pya-A ; 1pya-B ; 1pyd-A ; 1pyp ; 1qrd-A ; 1r09-1 ; 1rbp ; 1reg-X ; 1rgs ; 1rhd ; 1rhg-A ; 1rie ; 1ris ; 1rlr ; 1rnl ; 1rop-A ; 1rpa ; 1rsy ; 1rvv-A ; 1s01 ; 1sac-A ; 1scu-A ; 1scu-B ; 1sfe ; 1shf-A ; 1sly ; 1smn-A ; 1spb-P ; 1sra ; 1sry-A ; 1std ; 1sto ; 1stp ; 1tab-E ; 1taf-A ; 1taf- B ; 1tah-B ; 1tbp-A ; 1ten ; 1tfd ; 1tfe ; 1tfg ; 1tgs-I ; 1thj-A ; 1tht-A ; 1tie ; 1tif ; 1tlk ; 1tnd-A ; 1tnf-A ; 1tnr-A ; 1tpl-A ; 1trk-A ; 1tsp ; 1tss-A ; 1tul ; 1tup-A ; 1ubq ; 1ulb ; 1vcc ; 1vhh ; 1vin ; 1vmo-A ; 1vnc ; 1vsg-A ; 1wap-A ; 1was ; 1wsy-A ; 1wsy-B ; 1xrc ; 1xva-A ; 1xxa-A ; 1ypt-A ; 1zym-A ; 256b-A ; 2aai-B ; 2aat ; 2abk ; 2act ; 2alp ; 2atc-A ; 2atc-B ; 2aza-A ; 2baa ; 2bbv-A ; 2bbv-D ; 2bgu ; 2bop-A ; 2bpa-1 ; 2cab ; 2ccy-A ; 2cdv ; 2chs-A ; 2cpk-E ; 2cpl ; 2cro ; 2cts ; 2cyp ; 2dln ; 2dnj-A ; 2end ; 2eng ; 2fgf ; 2fxb ; 2gbp ; 2gcr ; 2gn5 ; 2had ; 2hhm-A ; 2hmz-A ; 2hpr ; 2hts ; 2i1b ; 2kau-A ; 2kau-B ; 2kau-C ; 2lhb ; 2ltn-A ; 2ltn-B ; 2mev-4 ; 2mnr ; 2pab-A ; 2pcd-A ; 2pcd-M ; 2pcy ; 2pgd ; 2phh ; 2phy ; 2pia ; 2pol-A ; 2por ; 2rsp-A ; 2rve-A ; 2sbl-B ; 2sn3 ; 2sns ; 2snv ; 2sod-B ; 2spc-A ; 2stv ; 2taa-A ; 2tgp-I ; 2tmd-A ; 2tmv-P ; 2trt ; 2tsc-A ; 2utg-A ; 2wrp-R ; 3adk ; 3app ; 3b5c ; 3blm ; 3bp2 ; 3cd4 ; 3cla ; 3cln ; 3cox ; 3ebx ; 3eca-A ; 3fis-A ; 3gap-A ; 3grs ; 3hmg-A ; 3hmg-B ; 3icb ; 3lzm ; 3mdd-A ; 3mon-A ; 3pgm ; 3pmg-A ; 3rnt ; 3rub-L ; 3rub-S ; 3sc2-A ; 3sc2-B ; 3sdh-A ; 3sdp-A ; 3tim-A ; 4enl ; 4fxn ; 4mbn ; 4pfk ; 4rhv-3 ; 4rxn ; 4sgb-I ; 4ts1-A ; 4xia-A ; 5cpa ; 5cpv ; 5cyt-R ; 5hvp-A ; 5ldh ; 5rsa ; 6acn ; 6cpp ; 6dfr ; 7icd ; 7lyz ; 8abp ; 8adh ; 8cat-A ; 8tln-E ; 9api-A ; 9api-B, and 9ins-B.

Amino Acid Residues Used in This Study. The amino acid residues in the sequences are represented by one-letter symbols, and the sequence numbers of the residues in the proteins are based on DSSP.[22] Chain breaks in a protein are assumed if the peptide bond length (distance C′–N) exceeds 2.5 Å, according to Kabsch and Sander.[22] They are labeled "!" and counted as a break residue. The 411 analyzed protein chains comprise 110 chain breaks. The residues for which there are coordinates in PDB are numbered sequentially, including break residues. The amino acid residues (Nos. 1—3 of 1mdy-A and Nos. 1—9 of 1rsy) are excluded, although their coordinates are entered. The amino acid residues (101122) in the 411 protein chains of Data Set A are used in this study. In any case, an inspection of the amino acid sequences always allows an unambiguous identification of tri- (100102) and tetrapeptide (99593) sequences. Therefore, we can identify the locations and meanings of 3- and 4-letter words in the 411 protein chains.

Extraction of Helices, Type-I α-, and β-Turns from the 411 Protein Chains. The 3191 helices, the 975 type-I α-turns and the 1890 type-I β-turns were easily extracted from the 411 analyzed protein chains by using both 2-dimensional ϕ and ψ representations of their 3-dimensional structures[12,19] and Kabsch and Sander's automatic assignments.[22]

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 of the crystallographers who have deposited their hard-earned coordinate sets in the PDB.

References

1 C. B. Anfinsen, E. Haber, M. Sela, and F. H. White, Jr., *Proc. Natl. Acad. Sci. U.S.A.*, **47**, 1309 (1961).
2 C. B. Anfinsen, *Science*, **181**, 223 (1973).
3 S. C. Harrison and R. Durbin, *Proc. Natl. Acad. Sci. U.S.A.*, **82**, 4028 (1985).
4 K. A. Dill, K. M. Fiebig, and H. S. Chan, *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 1942 (1993).
5 L. S. Itzhaki, D. E. Otzen, and A. R. Fersht, *J. Mol. Biol.*, **254**, 260 (1995).
6 J. C. Martinez, M. T. Pisabarro, and L. Serrano, *Nature Struct. Biol.*, **5**, 721 (1998).
7 M. T. Reymond, S. Huo, B. Duggan, P. E. Wright, and H. J. Dyson, *Biochemistry*, **36**, 5234 (1997).
8 S. Honda, N. Kobayashi, E. Munekata, and H. Uedaira, *Biochemistry*, **38**, 1203 (1999).
9 K. A. Dill and H. Sun Chan, *Nature Struct. Biol.*, **4**, 10 (1997).
10 B. Kuhlman, J. A. Boice, R. Fairman, and D. P. Raleigh, *Biochemistry*, **37**, 1025 (1998).
11 M. Narita, K. Sode, S. Ohuchi, Y. Murakawa, and M. Hitomi, *Bull. Chem. Soc. Jpn.*, **71**, 385 (1998).
12 M. Narita, K. Sode, S. Ohuchi, M. Hitomi, and Y. Murakawa, *Bull. Chem. Soc. Jpn.*, **70**, 1639 (1997).
13 W. Kabsch and C. Sander, *Proc. Natl. Acad. Sci. U.S.A.*, **81**, 1075 (1984).
14 P. Argos, *J. Mol. Biol.*, **197**, 331 (1987).
15 W. R. Taylor, *Protein Eng.*, **2**, 77 (1988).
16 B. I. Cohen, S. R. Presnell, and F. E. Cohen, *Protein Sci.*, **2**, 2134 (1993).
17 C. Bystroff, K. T. Simons, K. F. Han, and D. Baker, *Curr. Opin. Biotechnol.*, **7**, 417 (1996).
18 E. G. Hutchinson and J. M. Thornton, *Protein Sci.*, **5**, 212 (1996).
19 M. Narita, K. Sode, and S. Ohuchi, *Bull. Chem. Soc. Jpn.*, **72**, 385 (1999).
20 J. S. Richardson and D. C. Richardson, *Science*, **240**, 1648 (1988).
21 A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, *J. Mol. Biol.*, **247**, 536 (1995).
22 W. Kabsch and C. Sander, *Biopolymers*, **22**, 2577 (1983). We used Secondary Structure Definition Program (DSSP) data base of C. Sander's research group.
23 S. Dasgupta and J. A. Bell, *Int. J. Peptide Protein Res.*, **41**, 499 (1993).
24 P. N. Lewis, F. A. Momany, and H. A. Scheraga, *Biochim. Biophys. Acta*, **303**, 211 (1973).
25 D. M. E. Szebenyi and K. Moffat, *J. Biol. Chem.*, **261**, 8761 (1986).
26 M. A. Jimenez, V. Munoz, M. Rico, and L. Serrano, *J. Mol. Biol.*, **242**, 487 (1994).
27 D. L. Minor, Jr., and P. S. Kim, *Nature*, **371**, 264 (1994).
28 D. L. Minor, Jr., and P. S. Kim, *Nature*, **380**, 730 (1996).
29 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).