© 1998 The Chemical Society of Japan

Statistical Characterization of Eleven Kinds of Helix Elements with Amino Acid Residues in the Middle of Triplets

Mitsuaki Narita, Koji Sode,* Shokichi Ohuchi,# Yuka Murakawa, and Mitsuo Hitomi

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

(Received April 28, 1997)

Based on the definition of a helix region in terms of dihedral angles ϕ and ψ, 681 helices could be easily extracted from 125 kinds of analyzed proteins by using two-dimensional ϕ and ψ diagrams of their three-dimensional structures. The 8000 (20^3) possible kinds of amino acid residues in the middle of triplets consisting of consecutive residues were used for a precise analysis of helical segments in globular proteins. They correspond to the 8000 kinds of amino acid residues in a sequence of proteins. The 681 helices were built up from 8836 amino acid residues, which could be classified into 11 kinds of helix elements (a—k). Each of the amino acid residues in a sequence of helical segments could be allotted to one of 11 helix elements (a—k). Glycine residues in the proteins were used for its precise assignment as internal standards. The 11 kinds of helix elements could be statistically characterized with amino acid residues in the middle of triplets. Amino acid residues, observed as a specific helix element 3-times or more often in the total helical segments, were evaluated regarding their propensities for a specific helix element by using their inherent preference (IP)-values for the specific helix element, which had been defined in a previous paper. Amino acid residues found 13-times or more often in the data set are also listed to show their propensities for specific helix elements.

The study of helix formation in globular proteins is expected to provide an essential feature for the mechanism of protein folding.[1] Richardson and Richardson[2] documented the preference for certain amino acid residues at defined positions along helical segments of the peptide chain in globular proteins, especially near to the ends of those helices. This preference can be related to the observation that the first four NH groups of an α-helix in a globular protein do not form an intrahelical hydrogen bond, and neither do the four CO groups at the C terminus. The helix hypothesis introduced by Presta and Rose[3] at the same time states that a necessary condition for helix formation is the presence of residues flanking the helix termini, whose side chains can form hydrogen bonds with the initial four-helix NH groups and the final four-helix CO groups. Attention is now being directed to specific helix endpoints.[2—8] In studying helix formation, Harper and Rose[6] later reported that at the first residues in helices, termed "N-cap", a reciprocal backbone-side-chain hydrogen-bonding interaction was frequently observed, and was dubbed the "capping box". They also reported on the preferences of pairing residues in the capping boxes. The field of helix formation is now at an exciting, but speculative, stage.

Principally, each amino acid residue in a protein sequence should be allotted to a building block of secondary structure. Here, we call the building block a secondary structure element. Thus, the secondary structures of a protein can be analyzed on an amino acid level. However, the definitions of helix and β-sheet endpoints are so far ambiguous, and secondary structure elements cannot always be resolved, even for X-ray structures known at high resolution. However, in our preceding paper[9] it was emphasized that particular amino acid residues out of the 8000 kinds of residues in a protein sequence favored specific locations at the $i+1$ and $i+2$ positions of the type-II β-turn. Similarly, particular residues favored specific locations at helix endpoints. These observations lead us to investigate the amino acid preferences for specific locations at defined positions along helical segments of the peptide chain in globular proteins by using the 8000 amino acid residues. In this paper we define the helix endpoints in terms of the dihedral angles (ϕ and ψ) and characterize the 11 kinds of helix elements with amino acid residues in the middle of triplets.

Materials and Methods

Proteins Examined in This Study. In Table 5 (in Appendix), the 125 different globular proteins examined in this study are listed and identified based on the Protein Data Bank (PDB) code.[10] Proteins 1—30 in Table 5 were used to tabulate the amino acid preferences for specific locations at the ends of helices.[7] Proteins 31—125 were used to predict the protein secondary structure.[11,12] They were classified into families according to Chothia et al.[13] in order to understand their homologies. Since homologous proteins were present in this data set, all of the helices were carefully cross-referenced in order to prevent the use of identical information that could bias the results. Homologous proteins in this data set did not bias the results through analyses by using the 8000 possible kinds of

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

amino acid residues.

Amino Acid Residues Used in This Study and Chain Breaks. The amino acid residues (23132) are those used in a preceding study;[9] the sequence numbers of the residues in the helix regions in Table 5 are those of PDB.[10] Chain breaks in a protein are assumed if the peptide bond length (distance C′–N) exceeds 2.5 Å, according to Kabsch and Sander.[14] Seven chain breaks can be observed in proteins 50, 64, 64, 85, 106, 111, and 118.

Definition of a Helix Region and Extraction of Helices from the 125 Proteins. In the definition of a helix region, the endpoints of helices in proteins were defined in terms of the dihedral angles (ϕ and ψ). According to Dasgupta, et al.,[7] the helix region, defined as intrahelical segments in this study, comprises ϕ angles of between $-133°$ and $-17°$ and ψ angles of between $-112°$ and $+18°$ (here, $-105°$ and $+6°$ in Ref. 7 are changed into $-112°$ and $+18°$, respectively). The dihedral angles (ϕ and ψ) of the amino acid residues in proteins 1—125 were based on the Dictionary of Secondary Structure of Protein (DSSP).[15] In this study, helix endpoints are defined by N-cap (No) and C-cap (Co) termed by Richardson and Richardson.[2] Each N-cap and C-cap residue makes one additional intrahelical hydrogen bond, but departs from helical values of the ϕ and ψ angles. Helices are identified as sequences of at least 6 consecutive residues.

Using two-dimensional ϕ and ψ diagrams of the three-dimensional structures of the 125 proteins,[9] 681 helices could be easily extracted from the 125 proteins, which are listed in Table 5. The assignments of these helices are not identical to Kabsch and Sander's own DSSP in detail.[15] Especially, most of the C-cap residues are different from those of Kabsch and Sander's assignment.

The Statistical Characterization of Eleven Kinds of Helix Elements with Amino Acid Residues in the Middle of Triplets. For the statistical characterization of 11 kinds of helix elements with amino acid residues in the middle of triplets, they were obtained by sliding one amino acid residue step-by-step on amino acid sequences of the 125 proteins from *N*- to *C*-termini in a previous paper.[9] The absence of a peptide bond and missing density in the crystallography map give rise to chain breaks, which reduce the number of triplets in the proteins. The protein chains given in Table 5 consist of 23132 single amino acid residues, and are regarded as being constructed of 22868 ($23132-2\times125-2\times7$) amino acid residues in the middle of triplets. Among the 8000 possible triplets, 6352 independent triplets were found in 125 proteins; among them, 1631 independent triplets were found only once and 1287 ones, twice. Thus, 18663 (82%) total triplets were found among 3434 independent triplets, which were found 3-times or more often in the data set.[9]

Triplets in helical segments were similarly obtained by sliding one amino acid residue step-by-step in the helix regions. In a helix boundary region, N′ and C′ amino acid residues flank N-cap and C-cap ones, respectively.[2] The N-cap residue is situated in the middle of the N-terminal triplet of a helix and the N′ residue is at the left of the triplet. The *C*-terminal triplet contains the C-cap residue in the middle and the C′ residue at the right. In the helix regions assembled in Table 5, 8836 triplets, namely amino acid residues in the middle of triplets, are found. Each of the amino acid residues in a sequence of the helical segments was allotted to one of the 11 helix elements (a—k) (see text), and were statistically characterized with these amino acid residues (Tables 1, 2, and 3). In the case of linked helices, a linking amino acid residue functions as both helix elements a and k (see text).

Results

Based on the definition of a helix region in terms of the dihedral angles (ϕ and ψ), the 681 helices (Table 5 in Appendix) could be easily extracted from the 125 analyzed proteins by using their two-dimensional ϕ and ψ diagrams, which had been devised in a previous paper.[9] Regions of repetitive secondary structures (helices and β-strands) can be recognized as repetitive linking patterns in the diagrams, and standard helices are seen as parallel lines with their ϕ and ψ values 10°—20° apart. Each N-cap and C-cap residue departs from the helical values of the ϕ and ψ angles, and intrahelical segments have backbone dihedral angles of ϕ and ψ near to the observed mean values of $-64°\pm7°$ and $-41°\pm7°$, respectively.[3] The relationship between the primary and secondary structures is also readily recognized on the diagrams.

Except for the *N*- and *C*-termini of the protein chains, the amino acid residues in a sequence of proteins are found among the 8000 possible kinds of residues by taking into account the difference in their N- and C-sides residues.[9] Namely, the amino acid residues in a sequence of proteins can be regarded as being those in the middle of triplets consisting of consecutive amino acid residues; also, each of 20 single common amino acid residues in a sequence of proteins comprises the 400 (20^2) kinds of residues in the middle of triplets by taking into account the difference in its N- and C-sides residues. The 125 analyzed proteins consist of the 22868 total amino acid residues in the middle of triplets, and comprise 6352 independent amino acid resides among the 8000 ones.[9] The 681 helices are built up from 8836 amino acid residues in the middle of triplets.

Since the previously observed, 20 single common amino acid preferences for specific locations at defined positions along helical segments have been observed; especially near to the ends of helices,[2,7] the 11 kinds of helix elements a—k are allotted to amino acid residues on the 11 kinds of defined positions along helical segments, as follows:

11 kinds of defined positions (N')$N_0N_1N_2N_3N_4$MM- - -
MM$C_4C_3C_2C_1C_0$(C')
11 kinds of helix elements - - -abcdeff- - -ffghijk- - -.

Here, we coin the terms "helix elements a—k" for amino acid residues on the 11 defined positions along helical segments. The 11 defined positions are labeled on the basis of the helix endpoints N_0 and C_0, as shown above. Thus, each of the amino acid residues which build up helical segments of the peptide chain in a protein is a building block of helical segments, and can be allotted to one of the 11 helix elements. They are represented by one of the small letters (a—k), as illustrated above. The helical segments identified as sequences of 6, 8, 10, 12, and 20 residues can be expressed using the 11 helix elements (a—k), as follows:

- - -abcijk- - -abcdhijk- - -abcdeghijk- - -abcdeffghijk- - -
- - -abcdefffffffffghijk- - -.

Table 1. Amino Acid Residues Characteristic of the 10 Helix Elements Except for f

Amino acid residue[a]	Total[b]	Helix			Other helix elements[f]	Amino acid residue[a]	Total[b]	Helix			Other helix elements[f]
		Element[c]	Number[d]	IP-value[e]				Element[c]	Number[d]	IP-value[e]	
ADP	6	a	4	67	k	ADF	4	g	3	75	d
AGP	6	"	4	67	k	LAA	13	"	3	23	g)
CGS	5	"	3	60	—	LLR	14	"	3	21	g)
DNA	6	"	3	50	b, e	QEL	11	"	4	36	d, f, i
FGD	8	"	3	38	f	AAA	27	h	3	11	g)
GDA	7	"	3	43	b, f	AIA	12	"	3	25	f, g, i
GDP	8	"	3	38	k	IIE	7	"	3	43	f
GDV	14	"	3	21	b, i	IRE	7	"	3	43	f, i
GGS	13	"	4	31	g)	LEK	13	"	3	23	g)
GSA	13	"	4	31	g)	LEN	13	"	3	23	g)
ISA	9	"	3	33	k	LKA	19	"	4	21	g)
KDL	11	"	3	27	e, f, i	LKE	13	"	5	38	g)
KSA	10	"	3	30	b, f, h, i, k	LKK	13	"	3	23	g)
KSP	8	"	3	38	k	ALL	17	i	4	24	g)
LPA	7	"	5	71	b	EEF	10	"	3	30	d, f, g
LSA	10	"	3	30	f, h	IAL	14	"	3	21	g)
LSE	9	"	3	33	c, f	IAN	8	"	3	38	e, j
LSK	10	"	3	30	g, i, j	KDL	11	"	3	27	a, e, f
RDP	5	"	3	60	—	LDS	9	"	3	33	d
SGK	12	"	3	25	f, k	LIA	12	"	3	25	f
TDA	10	"	4	40	b, c	LKK	13	"	4	31	g)
TSP	6	"	3	50	—	LLE	9	"	3	33	c, e, h
DAN	6	b	3	50	c	LQQ	6	"	3	50	b, f
DLK	10	"	3	30	d, j, k	REA	12	"	3	25	a, c, d, f,
DPA	8	"	4	50	—						g, h, j
DPE	7	"	4	57	—	AAG	21	j	4	19	g)
DPV	3	"	3	100	—	ALG	11	"	5	45	f, k
DVA	7	"	3	43	c, e	ASG	15	"	3	20	g)
GDV	14	"	3	21	g)	ATG	15	"	4	27	g)
KPE	5	"	3	60	i	EAG	6	"	3	50	c, d
NDA	8	"	3	38	a	EIG	6	"	3	50	c, e
RPE	4	"	3	75	—	EKG	11	"	4	36	f, h, k
SAA	13	"	6	46	g)	ELG	11	"	6	55	e, i
SAD	7	"	4	57	f	IAG	10	"	3	30	f
SEA	7	"	3	43	d, f	IAN	8	"	3	38	e, i
SEQ	6	"	4	67	c	KAA	17	"	4	24	g)
SKE	5	"	3	60	c, g	KAG	11	"	3	27	f, i, k
SLA	9	"	3	33	f, g, k	KEG	7	"	4	57	f, k
SPE	6	"	6	100	—	KYG	10	"	4	40	a
SVD	11	"	3	27	d	LEN	13	"	5	38	g)
TLE	9	"	3	33	d, h	LKN	11	j	3	27	c, e, f, g, i
TLS	8	"	3	38	—	LLG	13	"	5	38	g)
TPA	7	"	3	43	—	LSG	16	"	5	31	g)
TPD	8	"	3	38	—	NLA	13	"	3	23	g)
AAE	13	c	4	31	g)	QQK	6	"	3	50	—
AAQ	11	"	3	27	d, h, i, j	QYG	5	"	3	60	—
ADA	16	"	3	19	f)	RAG	7	"	3	43	f
ADV	8	"	3	38	a, d, e, k	SKG	5	"	3	60	—
KEE	8	"	3	38	f	SSG	11	"	4	36	b
LEE	10	"	3	30	a, f, g, i	TSG	10	"	3	30	e, h
PAA	6	"	4	67	f	AAP	7	k	3	43	—
PEA	8	"	3	38	b	AGA	17	"	6	35	g)
PEF	6	"	3	50	b, d	AGK	11	"	3	27	a, f, g
PEQ	3	"	3	100	—	AGL	13	"	4	31	g)
PET	11	"	3	27	a	AGV	13	"	5	38	g)
PKL	9	"	3	33	f	ANI	6	"	3	50	a, f
QQS	4	"	3	75	j	EGI	8	"	3	38	f
						EGK	9	k	3	33	c
AAK	12	d	3	25	c, e, f, h, i	IGA	6	"	3	50	a
AEA	8	"	3	38	c, e, f	IGR	11	"	3	27	a, d
AEL	12	"	3	25	e, f, i	KGV	7	"	4	57	a
DEA	9	"	6	67	a, c, f	KHP	6	"	3	50	—
DEL	10	"	4	40	f	KNL	13	"	3	23	g)
EEF	10	"	3	30	f, g, i	LGA	10	"	4	40	f
EEL	12	"	3	25	c, f, g, h, i	LGI	11	"	3	27	—
ETV	8	"	5	63	—	LGL	12	"	4	33	e, i
GDL	10	"	3	30	a, f	LGS	9	"	3	33	e, f, j
TLK	14	"	3	21	g)	LGT	11	"	4	36	f
DLL	16	e	3	19	g)	MGA	8	"	3	38	f, g
DVI	8	"	3	38	b, f	QKR	5	"	3	60	i
EAK	10	"	3	30	c, j	SGL	11	"	5	45	d, g, i
EAL	17	"	3	18	g)	SGR	8	"	3	38	—
EFE	7	"	3	43	f, j	SGS	14	"	3	21	g)
ELA	7	"	3	43	f	VAA	14	"	3	21	g)
ELK	15	"	5	33	g)						

a) Amino acid residues in the middle of triplets correspond to those in a protein sequence and are characterized by the helix element observed 3 times or more often in the total helical segments and other helix elements. b) The total observed occurrence number of the amino acid residue in the data set. c) The helix element observed for the residue 3 times or more often in the total helical segments. d) The observed occurrence number of the amino acid residue as the corresponding helix element in the total helical segments. e) See text. f) Other helix elements observed for the residue in the total helical segments. g) See Table 3.

Table 2. Amino Acid Residues characteristic of the Helix Element f

Amino acid residue[a]	Total[b]	Helix Number[c]	Helix IP-value[d]	Other helix elements[e]	Amino acid residue[a]	Total[b]	Helix Number[c]	Helix IP-value[d]	Other helix elements[e]
AAA	27	11	41	f)	KGI	11	4	36	k
AAG	21	4	19	f)	KLF	7	4	57	—
AAI	9	6	67	e, g	KRL	11	5	45	i, j
AAL	16	7	44	f)	KTL	10	4	40	a, e
AAV	17	8	47	f)	KVL	13	5	38	f)
AEL	12	4	33	d, e, i	LAA	13	6	46	f)
AKA	11	4	36	c, e, h, i	LAE	8	4	50	c, e, h
AKK	15	4	27	f)	LAT	7	4	57	g
ALA	19	8	42	f)	LIE	9	5	56	h, i
ALE	11	5	45	b, g, h	LIR	8	4	50	e, g, i
ALI	9	4	44	e	LKA	19	6	32	f)
ALK	15	5	33	f)	LKL	12	4	33	g
ALL	17	5	29	f)	LLK	14	6	43	f)
AQA	10	6	60	c, d,	LLL	16	5	31	f)
ARR	7	5	71	e	LNQ	8	4	50	a, k
ATI	9	5	56	d, h	LRA	9	4	44	d, e, h
AVA	15	5	33	f)	LRD	10	4	40	d, g, h
AVD	9	4	44	—	LRR	8	4	50	c, i
DAL	14	5	36	f)	LSV	9	4	44	a, g
DLI	6	4	67	e, j	LVA	13	4	31	f)
DLL	16	4	25	f)	NLA	13	4	31	f)
EAF	8	4	50	d, e, k	QAA	13	4	31	f)
EAI	9	4	44	c, d, e	QEL	11	4	36	d, g, i
EAL	17	7	41	f)	RAA	11	4	36	b, g
ELL	12	4	33	d, g, h	RIQ	6	4	67	c, h
FAD	7	4	57	i	RLK	7	4	57	g
FEK	8	4	50	e, k	RQL	5	4	80	i
IAA	10	4	40	h	SAI	9	6	67	b
IAK	8	4	50	h	VAA	14	4	29	f)
IAL	14	4	29	f)	VAK	12	6	50	c, h, i, j
IDD	6	4	67	—	VDL	12	4	33	h
IDL	9	4	44	a	VIA	18	4	22	f)
IKE	7	4	57	g, h	VLH	6	4	67	g, j
IRS	6	4	67	e	VLT	13	5	38	f)
KAA	17	5	29	f)	VSA	9	4	44	j
KFL	7	4	57	d, h					

a) Amino acid residues in the middle of triplets correspond to those in a protein sequence and are characterized by the helix element f observed 4-times or more often in the total helical segments and other helix elements. b) The total observed occurrence number of the amino acid residue in the data set. c) The observed occurrence number of the amino acid residue as the helix element f in the total helical segments. d) See text. e) Other helix elements observed for the residue in the total helical segments. f) See Table 3.

As illustrated above, each of the 8836 amino acid residues in the helical segments can be allotted to one of the 11 helix elements (a—k), and the growth of a helical segment longer than 10 residues gives rise to increase the helix element f in the helical segment. In the case of linked helices, a linking amino acid residue functions as both helix elements a and k. In this data set, 62 linking amino acid residues are present. Therefore, the 681 helices are built up from the 8898 helix elements. A large number of G residues in proteins have unique dihedral angles (ϕ and ψ) which are rarely allowed for other amino acid residues. Taking advantage of the definition of a helix region in terms of the dihedral angles (ϕ and ψ), they are used as internal standards for a precise assignment of each of the amino acid residues in a sequence of helical segments to one of the 11 helix elements.

A precise assignment makes it possible to analyze the 11 kinds of helix elements with amino acid residues in the middle of triplets, and statistically to characterize the 11 helix elements with them. Practically, the 11 helix elements (a—k) are allotted to amino acid residues on the 11 defined positions along the 681 helices, as illustrated above. In the case of linked helices, both helix elements a and k are allotted to a linking amino acid residue. Those amino acid residues observed 3-times or more often as a specific helix element, except for f, are summarized in Table 1 together with both their observed occurrence number as the specific helix element and their IP-values for the specific element.[9] Other helix elements which were observed fewer times for them are also listed. The IP-value is the overall percentage of the observed occurrence number of a certain amino acid residue as a specific helix element. For example, an IP-value of ADP for helix element a is 67. Since 6 D residues in the middle of ADP are found in the data set, the IP-value then means that 4 out of 6 ADP are observed as helix element a in

Table 3. the observed Occurrence Numbers of Each Helix Element for Amino Acid Residues Found out 13 Times or More Often in the Data Set

Amino acid		The observed occurrence number of helix element												Amino acid		The observed occurrence number of helix element											
residue[a]	Total[b]	a	b	c	d	e	f	g	h	i	j	k	Number[c]	residue[a]	Total[b]	a	b	c	d	e	f	g	h	i	j	k	Number[c]
AAA	27			1	2	1	11	1	3	1	2		5	KLV	15						3		2		1	1	8
AAE	13			4			3		2	2			2	KNL	13						1	1			1	3	7
AAG	21					1	4				4		12	KVL	13						5	2			1		5
AAL	16			1		1	7	1	2		1		3	LAA	13	1		1			6	3	1				1
AAV	17						8	2					7	LAK	13						3	1	1	2	2	1	3
ADA	16	2	1	3	1		1			1			7	LEK	13			1			2	2	3	1			4
ADG	16										1		15	LEN	13						3		3	2	5		0
AFS	13			1		1	3	1					7	LGG	14	2					2						10
AGA	17		1		1		1	1				6	7	LKA	19			1		1	6		4	2	2		3
AGL	13	1			2		2					4	4	LKE	13						2		5	1			5
AGV	13	1					1					5	6	LKK	13	1				2	1		3	4	1		1
AKG	15		1	1	1		1			1	1	1	8	LLG	13			1			1				5		6
AKK	15				1		4		1	2		2	5	LLK	14						6	1		1	1	1	4
AKL	13			1			2	1		1		1	7	LLL	16						5	1	2	1			7
ALA	19					1	8	2		2	1		5	LLN	13							1	2	2			8
ALK	15				2	1	5	2	1	1	2		1	LLR	14			1	1		2	3	2	2	1		2
ALL	17			1		1	5			4			6	LSG	16						1		1		5		9
ASG	15						1				3		11	LTK	14	1		1			1		1				10
ATA	15	1		1	1		2						10	LVA	13						4	2	1				6
ATG	15				1		2				4		8	LVK	15						3	1	1	1			9
AVA	15						5	1	2				7	NLA	13						4		1	1	3		4
DAA	14		2			2	3				1		6	PGS	14		1	1									12
DAL	14		1	1	2		5	1	1				3	QAA	13				1	1	4	1		1			5
DGS	13			1									12	SAA	13		6	1			2	1					3
DLL	16	1		2	1	3	4	1	1				3	SAL	16		1	1		2	1		1	2			8
DTV	15					1	1			1			12	SGS	14											3	11
EAL	17		1		2	3	7		1				3	SGT	14	1							1		2	10
ELK	15					5	1	1	1	1	1		5	SGV	15	1	1		1		1				2	9
GAT	16			1			1						14	SSS	15	1	2						1			11
GDG	20		1										19	TAG	13						2			2		9
GDV	14	3	3							1			7	TGS	19						1					18
GET	14		1		1				1			1	10	TLK	14		2		3		2				1	6
GGA	14		2	1			1					1	9	TLN	14						3			1		10
GGI	15	2											13	TVG	14		1			1			1			11
GGS	13	4											9	TVL	13		1			2	2		1			7
GID	14		2				1						11	TVS	14					1							13
GKL	13	1	1				1						10	VAA	14						4		1		3	6
GKT	16						1					1	14	VAL	14			1	1	1	3		1				7
GLA	15					2	1		1				11	VAS	17					1	2	1	2	2			9
GSA	13	4					2						7	VGD	13			1								1	11
GSG	16	1	1										14	VGG	16	2	1				1					2	10
GSL	17	1		1							1	1	13	VGL	14				1							2	11
GTG	13		2								1	1	9	VIA	18	1					4			1			12
GVD	18		2	1	1		2						12	VIG	13								1				12
GVT	14		1				3						10	VLG	13						2	1	1	1			8
IAL	14					1	4			3			6	VLN	14						1				1	2	10
IGG	14						1					1	12	VLT	13						5						8
IGV	15	1		1	1							1	11	VSV	13	2											11
KAA	17			1			5			2	4		5	VTV	14	1					1						12
KAD	13	1		2			1				2	1	6	VVV	13					1							12
KEL	13	1			1		1	1	2	2			5	YTG	13										1		12

a) Amino acid residues in the middle of triplets correspond to those in sequence of a protein and are characterized by the observed occurrence numbers of specific helix element for them. b) The total observed occurrence number of the amino acid residue in the data set. c) The observed occurrence number of the amino acid residue as other secondary structure elements.

the total helix regions. ADP functions as helix element a at high probability. Furthermore, it is also observed twice (not shown) as much as helix element k. Its IP-value for the helix element k in this data set is 33. One of the small letters (a—

k) in Table 1 denotes helix elements observed for amino acid residues in the sequence of 681 helices. In Table 1, the amino acid residues are also denoted by one capital symbol.[9] All of the amino acid residues in the middle of triplets in Table 1 correspond to those in a protein sequence. Table 1 indicates that a variety of amino acid residues function as specific helix elements with high frequency. This fact is important for understanding the amino acid propensities for none to some of specific helix elements. As a typical example, the 10 S residues in the middle of KSA are found in the data set, and are observed as helix element a 3-times, and as other helix elements (b, f, h, i, and k). Amino acid residues observed 4-times or more often as helix element f are also summarized in Table 2 together with both their observed occurrence number and their IP-values for element f. As assembled in Tables 1 and 2, SEQ, KEE, DEA, FEK, QEL, LEK, REA, and KEG function as helix elements b, c, d, f, g, h, i, and j, respectively, at high probability. DAN, AAE, AAK, EAK, AAI, LAA, AAA, IAN, AAG, and VAA are also observed as helix elements b, c, d, e, f, g, h, i, j, and k, respectively, at high probability. These results emphasize that amino acid residues X in the middle of triplets UXZ have strong propensities for specific helix elements (a—k), depending on their N- and C-sides residues, U and Z. The results given in Tables 1 and 2 also indicate that a certain amino acid residue functions as one to some specific helix elements and, conversely, each helix element favors a variety of particular amino acid residues out of the 8000 possible ones. As a result, the 11 kinds of helix elements are statistically characterized by amino acid residues in the middle of triplets. An IP-value of a certain amino acid residue for a specific helix element is more useful for evaluating the amino acid propensities for a specific helix element than the observed occurrence number of the amino acid residue as the specific helix element, since the latter depends entirely on the size of a data set.

In Table 3, amino acid residues found 13-times or more often in the data set are assembled. The observed occurrence numbers of each helix element for these amino acid residues are listed according to a precise assignment of each of the 8836 amino acid residues in a sequence of the 681 helices to one of the 11 helix elements. Although their IP-values for a specific helix element are not shown, they are obtained as the overall percentage of the observed occurrence number (listed in Table 3) of a certain amino acid residue as a specific helix element. For example, the IP-value of GGS for the helix element a is 31, meaning that 4 out of 13 GGS occur as helix element a. As another example, 27 A residues in the middle of AAA are found in the data set and the 1, 2, 1, 11, 1, 3, 1, and 2 middle A residues are observed as the helix elements c, d, e, f, g, h, i, and j, respectively. Thus, the IP-values of the middle A residue for the helix elements c, d, e, f, g, h, i, and j are 3.7 (100/27), 7.4 (200/27), 3.7, 41 (1100/27), 3.7, 11 (300/27), 3.7 and 7.4, respectively.

Secondary structures of the 125 analyzed proteins are built up from their building blocks, which are classified into a variety of secondary structure elements. For example, long helices are built up from the 11 kinds of helix elements, and β-strands are constructed by the 5 kinds of β-strand elements. An additional 14 loop elements are further classified.[16] Generally, each of the 8000 kinds of amino acid residues functions as some specific secondary structure element in proteins and has its IP-values for them.[16] The A residue in the middle of AAA has large IP-values for helix elements f and h, showing strong propensities for these helix elements. On a whole, 22 out of the 27 A residues are observed as helix elements, and the residual 5 A residues occur as other secondary structure elements in this data set. As another example, the 13 A residues in the middle of AAE are found in the data set and the 4, 3, 2, and 2 A residues in the middle of AAE are observed as the helix elements c, f, h, and i, respectively. Thus, the IP-values of the middle A residue for helix elements c, f, h, and i are 31, 23, 15, and 15, respectively. The residual 2 A residues function as other secondary structure elements in the data set. The 21 A residues in the middle of AAG are found in the data set and the 1, 4, and 4 A residues are observed as helix elements e, f, and j, respectively. The other 12 A residues are observed as other secondary structure elements. The 16 D residues in the middle of ADG are found in the data set, and only one of them is observed as helix element j, and the D residue does not have strong propensities for the helix element. The 17 G residues in the middle of AGA are found in the data set and the 1, 1, 1, 1, and 6 G residues out of them are observed as the helix elements b, d, f, g, and k, respectively. The G residue is evaluated to have strong propensities for helix element k.

Since the normalized preference (NP)-value[9] is statistically more distinct than the corresponding IP-value for evaluating the amino acid propensities for a specific helix element, the numbers of helices having a variety of chain lengths and

Table 4. The Numbers of Helices Having a Variety of Chain Lenhths, Those of Each Helix Element in the 681 Helices and the *S*-Factors of Each Helix Element

Chain length of helices	6	7	8	9	10 or greater		Total
The number of helices	65	43[a)]	65	37[a)]	471		681
Helix element	a	b, c, i, j	d, h	e, g	f	k	Total
The number of helix element (n)	670	681	573	471	2683	653	8818 (8898)[b)]
S-Factor[c)]	2.9	3.0	2.5	2.1	12	2.9	

a) Amino acid residues in the center of helical segments are not assigned to any of helix elements. b) The total number of helix elements in the 681 helices is 8898 (8818+43[a)]+37[a)]). c) An *S*-factor can be obtained as the overall percentage of the number of each helix element ($100n$/22868) to the number of the total amino acid residues (22868).

those of helix elements in the 681 helices were examined, and are listed in Table 4 in order to obtain both structure (*S*)-factors of helix elements[9] and NP-values of amino acid residues for a specific helix element. As noticed in the footnote of Table 5, the 11 and 28 helices existing in the *N*- and *C*-termini of protein chains do not comprise N′ and C′ residues, respectively. Thus, the total observed occurrence numbers of helix elements a and k allotted to amino acid residues in the middle of triplets are 670 and 653, respectively. Another representative example of the total observed occurrence number of a helix element is that of helix element f to be 2683. Amino acid residues (43+37) in the center of helical segments identified as sequences of 7 and 9 residues are not assigned to any helix elements. Therefore, in the helix regions, the 8898 (8818+43+37) total helix elements are found as listed in Table 4. Namely, the 681 helices are built up from the 8898 total helix elements, while 62 particular amino acid residues linking two helices in the data set function as both helix elements a and k. Thus, the 681 helices are built up from 8836 (8898-62) amino acid residues. Using the total number of amino acid residues (22868) in the midle of triplets, an *S*-factor of a helix element is obtained as the percentage of the number of each helix element to that of the total residues, as noticed in the footnote of Table 4. For example, the *S*-factors for helix elements a, d, and h are 2.9, 2.5, and 12, respectively. These values then mean that 2.9% (670) of the total residues (22868) are helix element a, 2.5% (573) helix element d, 12% (2683) helix element f, and so on. The implication of NP-values of a certain amino acid residue for specific helix elements is discussed later.

Discussion

A two-dimensional ϕ and ψ diagram of the protein three-dimensional structure was powerful for extracting helices from a protein and a precise assignment of each of the amino acid residue in a sequence of helical segments to one of the 11 helix elements. As a result, each of the 8836 amino acid residues could be precisely assigned to one or two of the 11 kinds of helix elements (a—k). At 62 particular residues linking two helices, they function as both helix elements a and k. Generally, each of the amino acid residues in a protein sequence can be allotted to one or two of the secondary structure elements of the protein. An analysis of the previously observed single amino acid preferences for specific locations at the 11 defined positions along helical segments with the 8000 possible kinds of amino acid residues could add a more detailed understanding of helical segments in globular proteins so as to clarify the statistical characteristics of the 11 kinds of helix elements (a—k). As assembled in Table 3, the amino acid residue X in the middle of a triplet UXZ has strong propensities for none to some of the specific helix elements. The propensities remarkably depend on its N- and C-sides residues. This fact reflects that each of 20 single common amino acid residues in a sequence of proteins comprises 400 kinds of residues in the middle of triplets, whose properties are related to their conformations, reflecting the difference in their N- and C-sides residues. Practically, the A residues in the middle of triplets UAA, which comprise another A residue at the right of a triplet has strong propensities for a variety of helix elements, and are remarkably dependent on their N-side residues (U). AAA, DAA, KAA, LAA, QAA, SAA, and VAA (Table 3) have strong propensities for different helix elements, depending on the amino acid residues at the left of triplets, A, D, K, L, Q, S, and V residues, respectively. Concretely, QAA and SAA have strong propensities for helix elements f and b, respectively. Similarly, the G residues in the middle of triplets UGZ show strong propensities for specific helix elements, depending on their N- and C-sides residues. The middle G residues of AGA, AGL, AGV, SGV, and VGL and those of GGI, GGS, LGG, and VGG (Table 3) have strong propensities for helix elements k and a, respectively. However, those of DGS, IGG, IGV, and TGS (Table 3) do not have strong propensities for any of the helix elements. Most of the amino acid residues listed in Table 3 indicate that each of them locates preferentially at a few specific positions along helical segments. As a typical example, the middle A residue of AAV and the middle L residue of VLT predominantly locate at the defined positions M and C_4 and the defined position M, respectively. Alternatively, particular residues, such as AAA and AKG, locate widely at a variety of defined positions. The amino acid residues given in Tables 1, 2, and 3 cleary show that the 11 kinds of helix elements are statistically characterized by amino acid residues in the middle of triplets. These results appear to imply that the 8000 kinds of amino acid residues in the middle of triplets correspond to the 8000 (20^3) words consisting of three letters, since 20 single common amino acid residues in a protein sequence are denoted by 20 single letters, and that they mean amino acid propensities for specific helix elements. Since the secondary structures of a protein can be analyzed on an amino acid level, and each amino acid residue in a protein sequence can be allotted to one or two of the secondary structure elements of a protein, it is expected that they can be elucidated by deciphering the message in the amino acid sequence of a protein with the 8000 words, meaning the amino acid propensities for some of specific secondary structure elements.

For the purpose of evaluating the amino acid propensities for a specific helix element, the NP-values of a certain amino acid residue for specific helix elements can be obtained by dividing its IP-values for the corresponding helix elements with their *S*-factors. The NP-values of the A residue in the middle of AAA for the helix elements c, d, e, f, g, h, i, and j are 1.2 (3.7/3.0), 3.0 (7.4/2.5), 1.8 (3.7/2.1), 3.5 (41/12), 1.8 (3.7/2.1), 4.4 (11/2.5), 1.2 (3.7/3.0), and 2.5 (7.4/3.0), respectively. Its NP-value of 3.5 for the helix element f means that it is observed as the helix element f 3.5 times as often as at large. Similarly, its NP-value of 4.4 for helix element h means that it is observed as helix element h 4.4 times as often as at large. In contrast to the fact that the IP-values of the middle A residue of AAA for helix elements f and h are 41 and 11, respectively, its NP-values for helix elements f and h are 3.5 and 4.4, respectively. As another example of the NP-values of an amino acid residue in Table 3,

the IP-values of AAG for helix elements f and j are equal to 19, but its NP-values for helix elements f and j are remarkably different, 1.6 and 7.6, respectively. The strong propensities of LEN for specific helix elements are typical. The 13 middle E residues of LEN are found in the data set and the observed occurrence numbers of the middle E residue as elements f, h, i, and j are 3, 3, 2, and 5, respectively. No LEN is observed as other secondary structure elements in the data set. Since the *S*-factors of the elements f, h, i, and j are 12, 2.5, 3.0, and 3.0 (Table 4), respectively, the middle E residue of LEN locates at the defined positions M, C_3, C_2, and C_1 along helical segments 2.0, 9.2, 5.1, and 13 times, respectively, as preferentially as at large. Statistically, although the NP-value is more distinct than the corresponding IP-value, the NP-value is obtained by using *S*-factors which are entirely dependent on the data sets. However, the IP-value is weakly dependent on the data sets and is a characteristic of each amino acid residue.

In conclusion, the 8000 kinds of amino acid residues in the middle of triplets correspond to those in a sequence of proteins by taking into account the difference in their N- and C-sides residues. The previously observed, 20 single common amino acid preferences for specific locations at the 11 defined positions along helical segments were analyzed with the 8000 amino acid residues in a sequence of proteins to add a more detailed understanding of helical segments of the peptide chain in globular proteins. The helices in proteins are built up from the 11 kinds of helix elements, and are statistically characterized by amino acid residues in the middle of triplets. The amino acid residue X in the middle of a triplet UXZ has the strong propensities for none to some of specific helix elements. The propensities remarkably depend on its N- and C-sides residues. Conversely, each helix element favors a variety of particular amino acid residues out of the 8000 ones. The triplets appear to correspond to the 8000 words of protein language consisting of three letters and meaning amino acid propensities for specific helix elements. Our results also emphasize that, generally, each of the amino acid residues in a protein sequence has the strong propensities for some of specific secondary structure elements. Our results strongly suggest that the genetic information for helices in a protein sequence should be able to be deciphered by the words.

Appendix

The 125 proteins examined in this study are listed in Table 5 and identified by PDB code.[10] The 681 helices extracted from the 125 proteins are also listed in Table 5.

Table 5. Proteins Examined and Helices Identified

No.	PDB code	Helices[a]
1	1eca	2-19, 19-31, 31-38, 45-52, 52-73, 75-91, 93-112, 113-136[b]
2	1ppt	13-33
3	1rhd	11-22, 42-50, 76-87, 106-118, 129-137, 163-174, 183-189, 211-216, 224-235, 250-264, 274-282
4	2act	6-11, 24-43, 49-57, 69-81, 99-105, 120-131, 141-147
5	2aza	52-67, 116-121
6	2cab	12-20, 20-25, 130-137, 154-167, 180-185, 219-229
7	2cdv	29-34, 64-71, 78-88, 90-99
8	2cro	1-14, 16-25, 27-37, 44-53, 55-62
9	2cyp	15-33, 41-57, 73-79, 84-99, 103-119, 150-162, 164-177, 181-186, 200-209, 232-241, 241-254, 254-272, 288-293
10	2hmz-A	18-39, 41-67, 68-87, 91-111
11	2lhb	12-29, 29-45, 45-52, 60-67, 67-90, 90-110, 112-128, 131-149[b]
12	2sn3	22-31
13	2sns	54-60, 60-69, 98-107, 121-136
14	2stv	12-24,[c] 46-51, 101-106, 116-123
15	3adk	1-8,[c] 20-33, 38-50, 51-64, 68-87, 98-110, 121-137, 143-168, 178-193
16	3app	47-52, 58-63, 109-115, 139-148, 222-233, 268-273, 299-306
17	3b5c	8-15, 31-39, 42-50, 53-61, 64-75, 80-87[b]
18	3bp2	1-12,[c] 16-22, 38-57, 88-108
19	3grs	29-43, 56-61, 62-86, 95-122, 176-183, 196-210, 227-242, 299-304, 338-356, 383-392, 405-411, 439-454, 456-463, 469-476
20	3lzm	2-12, 38-51, 59-81, 81-91, 92-107, 107-114, 114-124, 125-135, 136-142, 142-156, 157-164[b]
21	4fxn	10-27, 62-75, 93-107, 123-137
22	4mbn	3-20, 20-36, 36-43, 43-49, 51-58, 58-79, 81-98, 100-119, 124-150
23	5cpa	14-29, 72-90, 93-103, 112-123, 142-147, 173-187, 215-235, 242-248, 253-262, 282-307[b]
24	5cpv	7-18, 25-34, 39-51, 59-71, 78-90, 98-108[b]
25	5cyt	2-18, 49-56, 60-70, 70-75, 87-103[b]
26	5rsa	3-13, 22-34, 50-60
27	7lyz	4-15, 24-37, 59-64, 79-85, 87-102, 108-116

Table 5. (Continued)

No.	PDB code	Helices[a]
28	8adh	46-55, 100-105, 165-174, 174-188, 201-215, 225-236, 249-259, 271-282, 304-311, 323-338, 341-346, 354-364
29	8cat	5-18, 53-67, 96-101, 156-168, 177-188, 188-200, 245-256, 256-271, 284-291, 323-331, 347-366, 437-450, 451-469, 470-485, 485-500[b]
30	8tln	64-89, 136-154, 159-181, 207-212, 229-247, 259-275, 280-297, 300-314
31	1acx	
32	1ak3-A	17-29, 35-45, 48-61, 65-79, 93-104, 115-126, 163-189, 198-213
33	1bbp-A	139-152, 166-171
34	1bds	
35	1bmv-1	54-63
36	1bmv-2	46-52, 57-64, 137-143, 230-239, 275-280
37	1cbh	
38	1cc5	7-22, 31-42, 47-55, 69-82
39	1cdt-A	
40	1crn	6-20, 22-31
41	1cse-I	17-29
42	1fdl-H	60-67
43	1fdx	12-18, 39-45
44	1fkf	38-43, 56-66
45	1fxi-A	23-32
46	1gd1-O	10-23, 36-46, 78-83, 83-88, 101-112, 150-168, 193-198, 209-219, 252-269, 280-285, 313-333
47	1gdj	4-21, 21-37, 37-44, 57-82, 87-101, 103-123, 127-153[b]
48	1hip	11-18, 27-32
49	1il8-A	55-72[b]
50	1lap	21-31, 33-43, 83-104, 115-128, 150-173, 179-195, 204-212, 213-223, 266-291, 333-350, 361-369, 379-394, 403-411, 427-440, 471-484[b]
51	1lmb-3	8-30, 32-41, 43-53, 58-70, 72-77, 77-92[b]
52	1mcp-L	127-132, 188-193
53	1mrt	
54	1ovo-A	33-45
55	1paz	96-105, 108-120[b]
56	1pyp	179-187, 187-194, 210-231, 253-260
57	1r09-1	66-72, 92-97, 109-119, 165-170, 214-219
58	1rbp	16-21, 145-160
59	1s01	5-12, 12-20, 63-73, 103-118, 132-146, 219-238, 242-253, 259-264, 269-275[b]
60	1sh1	
61	1tgs-I	33-43
62	1tnf-A	
63	1ubq	22-35, 55-60
64	1wsy-A	1-14,[c] 29-44, 61-75, 77-92, 102-110, 110-122, 135-146, 160-170, 192-204,[c] 216-227, 235-244, 247-265[b]
65	1wsy-B	18-38, 38-54, 85-101, 113-127, 145-153, 165-172, 172-179, 196-221, 234-247, 310-320, 328-344, 348-365, 380-393[b]
66	256b-A	2-20, 22-43, 45-51, 55-82, 83-106[b]
67	2aat	29-36, 60-71, 84-96, 115-128, 142-152, 170-181, 202-217, 237-247, 276-294, 299-311, 312-344, 349-355, 367-375, 393-408[b]
68	2alp	197-206
69	2ccy-A	4-32, 39-58, 71-78, 78-103, 103-126
70	2fnr	130-139, 171-187, 212-223, 246-254, 254-264, 274-293, 295-307
71	2fxb	15-21, 47-61, 76-81[b]
72	2gbp	14-31, 44-59, 69-84, 95-102, 111-130, 152-170, 184-198, 198-204, 211-226, 238-248, 257-275, 299-307
73	2gcr	152–157
74	2gls-A	1-13,[c] 41-46, 103-119, 156-162, 188-203, 227-250, 291-304, 304-312, 315-322, 366-383, 413-425, 425-430, 435-456, 458-467
75	2gn5	
76	2i1b	33-39
77	2ltn-A	
78	2ltn-B	11-16
79	2mev-4	26-31

Table 5. (Continued)

No.	PDB code	Helices[a)]
80	2mhu	
81	2or1-L	1-14,[c)] 16-25, 27-36, 42-53, 55-62
82	2pab-A	74-83
83	2pcy	51-56, 84-91
84	2phh	11-25, 35-42, 49-60, 61-69, 87-93, 101-117, 163-169, 235-247, 248-255, 296-320, 321-351, 357-374, 374-387
85	2rsp-A	110-118
86	2sod-B	129-135
87	2tgp-I	2-7, 47-56
88	2tmv-P	8-14, 19-32, 37-52, 73-85, 110-135, 140-149
89	2tsc-A	1-15,[c)] 51-65, 68-76, 78-84, 93-100, 110-122, 134-141, 169-193, 211-222
90	2utg-A	3-16, 17-29, 31-48, 49-66
91	2wrp	8-33, 33-43, 44-64, 67-76, 78-92, 93-105
92	3ait	
93	3blm	32-41, 68-83, 84-90, 103-110, 115-127, 128-140, 141-153, 163-168, 179-193, 197-211, 211-217, 217-222, 272-287[b)]
94	3cd4	49-55, 58-65
95	3cla	16-29, 41-50, 54-70, 112-130, 198-215
96	3cln	5-20,[c)] 28-40, 44-56, 64-93, 101-113, 117-129, 137-147[b)]
97	3ebx	
98	3gap-A	8-17, 98-109, 109-135, 138-151, 168-177, 179-193
99	3hmg-A	65-72, 73-80, 104-116, 187-196
100	3hmg-B	37-57, 75-127, 145-155, 158-171
101	3icb	2-17, 24-36, 36-41, 45-54, 62-75[b)]
102	3pgm	28-46, 57-72, 98-105, 106-112, 150-172, 181-191
103	3rnt	12-30
104	3sdh	2-10,[c)] 11-28, 28-44, 44-51, 63-86, 86-105, 107-126, 128-146[b)]
105	3tim-A	17-31, 44-53, 79-87, 95-103, 105-120, 130-137, 138-154, 155-162, 179-199, 199-206, 215-225, 233-239, 239-249
106	4cms	47-53, 109-115, 135-144, 223-235, 247-254, 269-274, 297-304
107	4pfk	15-31, 40-47, 53-58, 73-78, 78-93, 102-115, 138-161, 173-185, 197-212, 226-239, 248-253, 257-278, 295-302, 307-319[b)]
108	4rhv-3	42-49, 63-68, 90-95, 95-104, 141-148
109	4rxn	
110	4sgb-I	
111	4ts1-A	1-11,[c)] 18-29, 45-62, 70-77, 90-109, 122-130, 131-143, 144-151, 151-161, 163-185, 192-211,[b)] 247-258, 259-271, 274-288, 292-308, 308-319[b)]
112	4xia-A	13-19, 34-46, 53-58, 63-82, 107-129, 149-172, 194-205, 216-224, 226-238, 263-277, 300-328, 328-339, 339-345, 353-360, 368-375, 378-392
113	5er2-E	110-116, 140-149, 225-236, 271-276, 293-298, 303-310
114	5hvp-A	86-94
115	5ldh	29-43, 54-69, 83-88, 110-128, 139-152, 163-179, 211-217, 226-241, 244-266, 311-331
116	6acn	17-33, 37-46, 72-87, 108-135, 145-155, 165-174, 182-192, 216-229, 229-234, 242-248, 249-264, 273-283, 285-300, 333-345, 362-379, 393-404, 404-413, 465-476, 566-571, 574-579, 582-590, 615-627, 644-654, 665-675, 684-691, 732-742, 743-754[b)]
117	6cpp	37-47, 67-77, 89-96, 105-120, 120-146, 149-154, 154-168, 170-186, 192-214, 218-226, 234-267, 267-277, 277-292, 321-328, 359-378
118	6dfr	24-36, 43-51, 77-86, 96-107
119	6hir	
120	7icd	36-57, 71-79, 85-96, 113-123, 169-184, 202-220, 234-254, 281-292, 302-318, 352-368, 369-387, 390-397, 404-416[b)]
121	8abp	14-32, 42-57, 67-82, 109-129, 145-163, 178-193, 205-220, 233-242, 255-274, 290-302
122	9api-A	21-45, 53-67, 70-80, 88-104, 127-138, 149-166, 259-267, 268-278, 298-307
123	9api-B	
124	9ins-B	8-20
125	9pap	6-11, 23-43, 49-57, 67-79, 95-101, 117-128, 138-144

a) N-cap and C-cap residues are included. b) These 28 helices do not comprise C′ residues. The helix (111, 192–211) give rise to a chain break at the *C*-terminal of the helix by missing density in the crystallography map. c) These 11 helices do not comprise N′ residues. The helix (64, 192–204) give rise to a chain break at the *N*-terminal of the helix by missing density in the crystallography map.

We thank C. Sander's working group for the use of their data bases. We also thank all those crystallographers who have deposied their hard-earned coordinate sets in the PDB.

References

1) J. M. Scholtz and R. L. Baldwin, *Annu. Rev. Biophys. Biomol. Struct.*, **21**, 95 (1992).

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

3) L. G. Presta and G. D. Rose, *Science*, **240**, 1632 (1988).

4) P. C. Lyu, H. X. Zhou, N. Jelveh, D. E. Wemmer, and N. R. Kallenbach, *J. Am. Chem. Soc.*, **114**, 6560 (1992).

5) P. C. Lyu, D. E. Wemmer, H. X. Zhou, R. J. Pinker, and N. R. Kallenbach, *Biochemistry*, **32**, 421 (1993).

6) E. T. Harper and G. D. Rose, *Biochemistry*, **32**, 7605 (1993).

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

8) D. L. Minor, Jr., and P. S. Kim, *Nature*, **380**, 730 (1996).

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

10) F. C. Benstein, 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.*, **112**, 535 (1977); PDB is generally available on the internet (http address://www.pdb.bnl/).

11) B. Rost and C. Sander, *J. Mol. Biol.*, **232**, 584 (1993).

12) V. V. Solovyev and A. A. Salamov, *Comput. Appl. Biosci.*, **10**, 661 (1994).

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

14) W. Kabsch and C. Sander, *Biopolymers*, **22**, 2577 (1983).

15) We used DSSP data base of C. Saner's working group (http address://www.embl-heidelberg.de/).

16) M. Narita and K. Sode, unpublished data.