Bull. Chem. Soc. Jpn., 70, 1639—1648 (1997)

A New Type of ϕ, ψ Representation of the Protein Tertiary Structure and the Analysis of the Amino Acid Preferences for Specific Locations at Type-II β-Turn by Using 8000 Possible Kinds of Amino Acid Residues

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

Department of Biotechnology, Tokyo University of Agriculture and Technology, 2-24-16, Naka-machi, Koganei, Tokyo 184

(Received December 16, 1996)

ϕ, ψ Representations of three-dimensional structures of 125 globular proteins were depicted for analyzing conformational details in their secondary and tertiary structures. They can be drawn in the form of a two-dimensional diagram. The new type of representation of the protein tertiary structure can be used as a stereotyped finger printing for 125 protein molecules. It is powerful for a precise comparison of the relationship of the primary, secondary, and tertiary structures of homologous proteins. The precise or broad similarity between the patterns of proteins classified into the same family is obvious. This representation is also useful in readily recognizing all of the secondary structures as one progresses along the chain. Complicated three-dimensional structures of 125 globular proteins are easily recognized to be built up from only four secondary structures: helix, β-strand, β-turn, and unordered structure. A combination of intersegments hydrogen bonds among β-strands was also shown in the diagram. In order to demonstrate the usefulness of the diagram, 195 type-II β-turns were extracted from 125 proteins using the new type of two-dimensional ϕ, ψ diagrams. Their amino acid sequences were surveyed so as better to understand amino acid preferences previously observed. For their analysis, a normalized preference (NP) value of an amino acid residue at a particular position of the secondary structure is defined as the ratio of the average percentage composition at the position based on average percentage composition at large. The statistical significance of the NP-value used in this study is more simple and clearer than that of the d-test based on a normal distribution, which was used in the previous study. It is clearly shown that a high NP-value (5.0) of a single P residue for the position $i+1$ of the type-II β-turn is an average preference of a variety of P residues in the middle of triplets consisting of consecutive amino acid residues. The NP-values of P residues, in the middle of 25 particular triplets are outstandingly high. Out of 400 possible P residues 322 kinds of P residues were found in 125 proteins. A remarkable high NP-value (8.8) of a single G residue for the position $i+2$ is also an average preference of G residues in the middle of various triplets.

The atomic coordinates of a large number of globular proteins can be utilized from the Brookhaven Protein Data Bank (PDB).[1] Although they are convenient for protein graphics, it is difficult to directly assign the relationship of the protein secondary structure with the amino acid sequence from the data. The ϕ, ψ backbone dihedral angles of each amino acid residue in a sequence calculated based on the atomic coordinates completely specify the backbone conformation of the protein, since the peptide unit can be taken to be planar for all practical purposes.[2] Although a linked ϕ, ψ chain plot and an unlinked ϕ, ψ chain plot have already been proposed for a visual comparison of the backbone conformation of peptides and proteins[3] and for mapping the protein folding,[4] they are rather complicated. Here, we proposed a new type of ϕ, ψ representation of the protein tertiary structure in the form of a two-dimensional diagram. The new graphic representation is distinct and useful for visually recognizing the relationship among primary, secondary, and tertiary structures. Practically, sequences of 195 type-II β-turns are extracted.

Here, we further survey amino acid sequences of 195 type-II β-turns and analyze amino acid preferences for specific locations at the type-II β-turn by using 8000 possible kinds of amino acid residues in the middle of triplets consisting of consecutive amino acid residues.

Materials and Methods

Proteins Examined in This Study. In Table 1, 125 different globular proteins examined in this study are listed and identified by the PDB code.[1] Proteins **1**—**30** in Table 1 were used to tabulate amino acid preferences for specific locations at the ends of helices.[5] Proteins **31**—**125** were used to predict the protein secondary structure.[6,7] In order to examine their homologies, they were classified into families according to Chothia.[8] Homologous proteins in this data set, (for example, **8** and **81**; **4** and **125**; **15** and **31**; **16**, **106**, and **113**) did not bias the results through analyses by using 8000 possible kinds of amino acid residues.

Amino Acid Residues Used in This Study and Chain Breaks. The amino acid residues in the data set for which there are coordinates[1] were used in this study. Chain breaks in a protein were assumed if the peptide bond length (distance C′–N) exceeded 2.5 Å according to Kabsch and Sander.[9] Seven chain breaks could be observed in proteins **50**, **64**, **64**, **85**, **106**, **111**, and **118**.

The ϕ, ψ Representation of Protein Tertiary Structure. Amino acid residues in sequence, which are represented by one-

Table 1. Proteins Examined[a] and Type II β-Turn Sequences[b]

No.	PDB code	Type II β-Turn sequences[b]	No.	PDB code	Type II β-Turn sequences[b]	No.	PDB code	Type II β-Turn sequences[b]	No.	PDB code	Type II β-Turn sequences[b]
1	1eca	FAGK	**28**	8adh	EEKK	**59**	1s01	GSNV	**99**	3hmg-A	FQNE
		VPGA			KAHE	**60**	1shl	NAGW			VPDY
		LNGG			GEGV	**62**	1tfn-A	AEGQ			WTGV
		IRGS			RPGD			ANGV			KPGD
4	2act	PYNN			TQGS			EKGD	**100**	3hmg-B	AIAG
5	2aza	KAGD	**29**	8cat	HIGK	**63**	1ubq	QKES	**103**	3rnt	HTGA
		GGGE			PPGI	**64**	1wsy-A	TLGD	**105**	3tim-A	FTGE
		TPGE			GPMC	**65**	1wsy-B	LHGG	**106**	4cms	DIQQ
		FPGH	**30**	8tln	ISGD			DTSV			SEYS
6	2cab	TKGK	**32**	1ak3-A	APGS	**68**	2alp	VGGI			CEGG
		VEGD			MLRG			TVNA	**107**	4pfk	SPGM
7	2cdv	TAGC	**33**	1bbp-A	EKYG			AVGA			VGDI
9	2cyp	EKGR	**35**	1bmv-1	QQTV			GRGD			IPGT
		NAGL	**36**	1bmv-2	AKSA			DSGG			TIGF
10	2hmq	YRGK	**37**	1cbb	SHYG	**69**	2ccy-A	AKGT	**108**	4rhv-3	LPGS
11	2lhb	FKGL			GIGY			LPNG			GSGQ
		AAGD			ASGT	**70**	2fnr	EEGI			LPNY
13	2sns	EKYG	**38**	1cc5	KVGD			REGQ			NANR
14	2stv	GPGA	**39**	1cdt-A	PEGK			KPGA			PPGA
15	3adk	GPGS			TDRC			NEKG			PPET
16	3app	QSGH	**41**	1cse-1	VVGK	**72**	2gbp	EPGH	**109**	4rxn	NPGT
		LSGY			PEGS			ADGT	**110**	4ssgb-1	YKGC
		AHGQ	**42**	1fdl-H	APSQ	**73**	2gcr	QPNF	**111**	4tsl-A	PSGK
		VSGA			PPGK	**74**	2gls-A	GYDR			KFGK
17	3b5c	QAGG			SSGV			KAGG	**112**	4xia-A	VFGH
19	3grs	KLGG	**44**	1fkf	KRGQ	**76**	2ilb	LKEK	**113**	5er2-E	LSGA
		IPGA			MLGK			FPNW			VSGA
		LKFS			HPGI			AENM			STGS
		VPGR			PPHA	**77**	2ltn-A	QNGE	**114**	5hvp-A	GIGG
		VKGI	**46**	1gd1-0	YNGS	**83**	2pcy	SPGE	**116**	6acn	RPGS
		YNNI	**47**	1gdj	LKGT			PSGV			YPGV
		GLGC	**48**	1hip	AAGA			AKGE			GTGA
21	4fxn	STKI			FPGK	**84**	2phh	VYQA			DPGC
		ISGK	**50**	1lap	KAGK	**86**	2sod-B	IIGR			KKGE
23	5cpa	NYGQ			AANM	**87**	2tgp-1	CGGA			FTGR
		GFGK			KPGD	**92**	3ait	APGQ			NYGE
		YANS			RSAG	**93**	3blm	YVGK			APGK
24	5cpv	AADS	**52**	1mcp-L	KPGQ			KKDT	**117**	6cpp	KKGD
25	5cyt	ENGG			ESGV			PKGQ			APGA
		LWGL	**55**	1paz	NPGD	**94**	3cd4	KKGD	**118**	6dfr	AVDR
		LFGR			IKDM			LQGQ			TLDK
		KTGQ			PEGA	**95**	3cla	FTDY			LPGR
		AEGY			KINE	**97**	3edx	SPGE	**119**	6hir	ESGQ
		IPGT	**56**	1pyp	FPQT			KPGI			GQGN
27	7lyz	LDNY	**57**	1r09-1	ATGI			SEVC	**120**	7icd	YQGT
		CKGT			PPGA	**98**	3gap-A	PSKS			KADS
		IRGC			KVGD			HQGE			IDGG
					SIGR			NQGD			YAGQ
			58	1rbp	KENF			ELGL	**122**	9api-A	DEGK
					LARQ						

a) The type II turn is not included in the following proteins: **2**, 1ppt; **8**, 2cro; **12**, 2sn3; **18**, 3bp2; **20**, 3lzm; **22**, 4mbn; **26**, 5rsa; **31**, 1acx; **34**, 1bds; **40**, 1cc5; **43**, 1cse-I; **45**, 1fdx; **49**, 1gdj; **51**, lil8-A; **53**, 1lmb-3; **54**, 1mcp-L; **61**, 1s01; **66**, 1wsy-A; **67**, 1wsy-B; **71**, 2fxb; **75**, 2gn5; **78**, 2ltn-B; **79**, 2mev-4; **80**, 2mhu; **81**, 2orl-L; **82**, 2pab-A; **85**, 2rsp-A; **88**, 2tmv-P; **89**, 2tsc-A; **90**, 2utg-A; **91**, 2wrp-R; **96**, 3cln; **101**, 3icb; **102**, 3pgm; **104**, 3sdh; **115**, 5ldh; **121**, 8adp; **123**, 9api-B; **124**, 9ins-B; **125**, 9pap. b) Amino acid residues in sequences are represented by one letter symbols.

letter symbols, are sequentially taken along the *X*-axis. The corresponding ϕ and ψ values are plotted together on the *Y*-axis above the amino acid residues in sequence. The symbols for each residue are connected by different lines, — for ϕ and --- for ψ. The dihedral angles (ϕ and ψ) for the 125 proteins used in this study were from DSSP.[10] The amino acid residues in sequences are represented by one letter symbols: A, alanine; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine;

I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine.

β-Turns Extraction. A β-turn forms a hydrogen bond between the main chain C=O (i) and the N–H ($i+3$), and is defined as 3-turn in the automatic Kabsch and Sander assignment.[10] The central $i+1$ and $i+2$ residues of the β-turns used in this study are not in a helical conformation in their automatic assignment. β-Turns were extracted from 125 proteins. All of the turns used in this study were assigned as "T" in the nomenclature of Kabsch and Sander.[10] They do not include bends to be assigned as "S".

Classification of β-Turns. β-Turn types were assigned to each of the extracted turns by using the two-dimensional ϕ, ψ diagrams. β-Turns were classified into the four common types (I, I′, II and II′) and a miscellaneous one (IV).[11] Ideal ϕ, ψ angles for the turn types were allowed to vary by about $\pm 50°$. No cross among turn types occurs by this allowance. The sequences of 195 type-II β-turns classified by using ϕ, ψ angles[11] are listed in Table 1.

Results

The new type of ϕ, ψ representations of tertiary structures of cro 434 (**8**, 2cro)[12] and 434 C1 repressor (**81**, 2orl-L)[13] are depicted in Fig. 1. The amino acid residues of 2cro (Fig. 1a) and 2orl-L (Fig. 1b) are sequentially taken along the X-axis. The corresponding ϕ, ψ values are plotted together on the Y-axis above the amino acid residues in sequence. In the amino acid alignment of homologous proteins, 2cro and 2orl-L, identical residues between them are shown by symbols in Fig. 1c. The homology is 52%.

The new type of ϕ, ψ representation of the protein tertiary structure in the form of a two-dimensional diagram can be used as a stereotyped finger printing for the 125 proteins. The precise (**8** and **81**; **4** and **125**; **16** and **113**) or broad (**15** and **31**; **16** and **106**; **106** and **113**) similarities between the patterns of proteins classified into the same family can be readily clarified by comparing their diagrams. The new type of two-dimensional ϕ, ψ diagram is powerful for a precise comparison of the relationship between the primary, secondary, and tertiary structures of these homologous proteins.

As typical examples, the ϕ, ψ representations of the tertiary structures of pancreatic polypeptide (**2**, 1 ppt)[14] and carbonic anhydrase (**6**, 2cab)[15] are also depicted in Figs. 2 and 3, respectively, in order to easily recognize the relationship between the primary and secondary structures of a protein. All of the secondary structures can be readily recognized as partial linking patterns in the ϕ, ψ diagrams. Regions of repetitive secondary structure can be recognized as repetitive linking patterns. Standard helices are seen as parallel lines will their ϕ and ψ values 10—20° apart. Standard β-strands appear as parallel lines with their ϕ and ψ values 80—130° apart. In the definition of a helix region, the end points of helices in proteins were defined in terms of dihedral angles ϕ and ψ. According to Dasgupta, et al.,[5] the helix region, defined as intrahelical segments in this study, comprise ϕ angles between $-133°$ and $-17°$, and ψ angles between $-105°$ and $+18°$ (here, $+6°$ in Ref. 5 is changed into $+18°$). Helix ends are defined by N-cap and C-cap, termed by Richardson and Richardson.[16] Each N-cap and C-cap residue makes one additional intrahelical hydrogen bond,

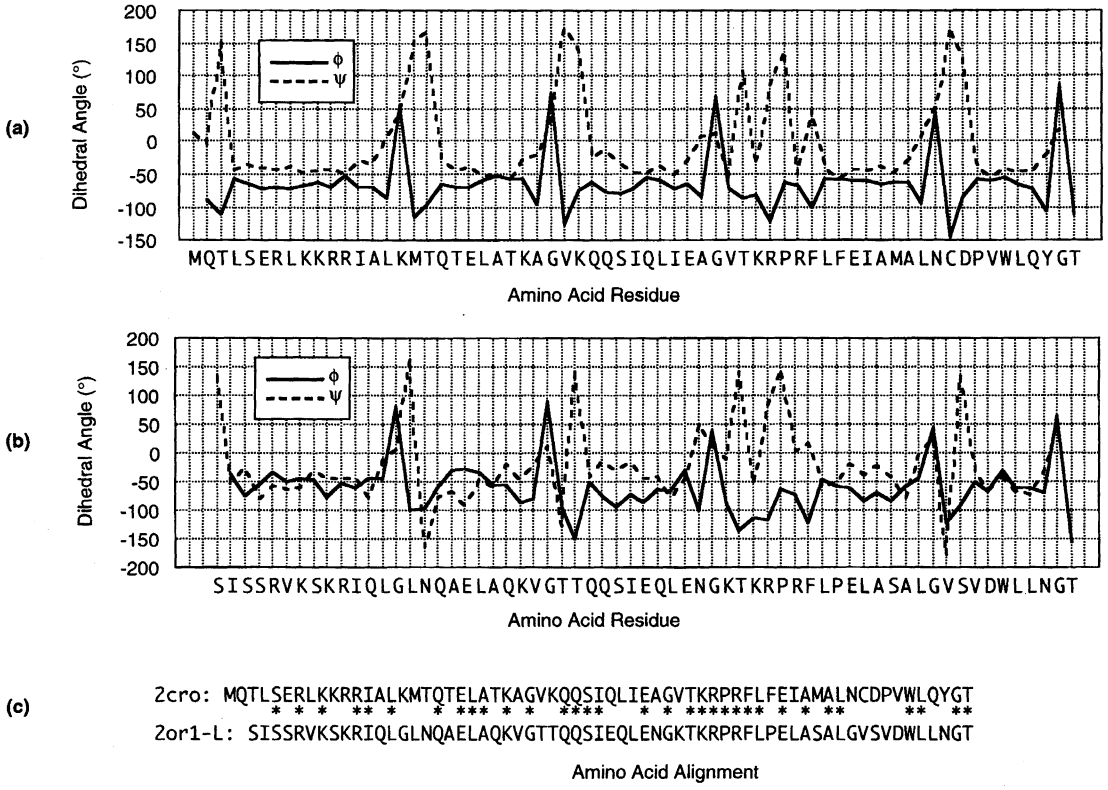

Fig. 1. The ϕ, ψ representation of tertiary structure of homologous proteins. (a) cro 434 (**8**, 2cro); (b) 434 C1 repressor (**81**, 2orl-L); (c) their amino acid alignments.

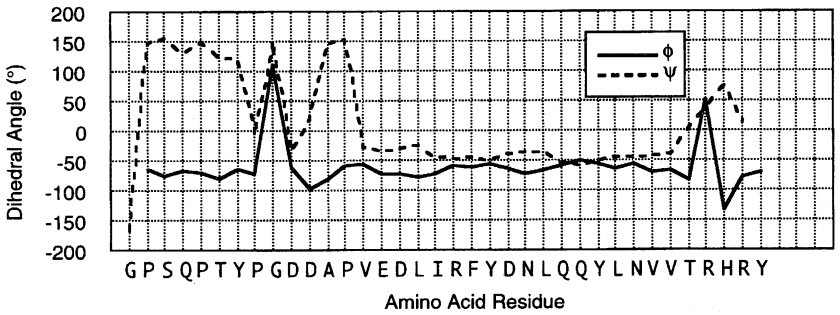

Fig. 2. The ϕ, ψ representation of tertiary structure of pancreatic polypeptide (**2**, 1ppt) in the form of two-dimensional diagram.

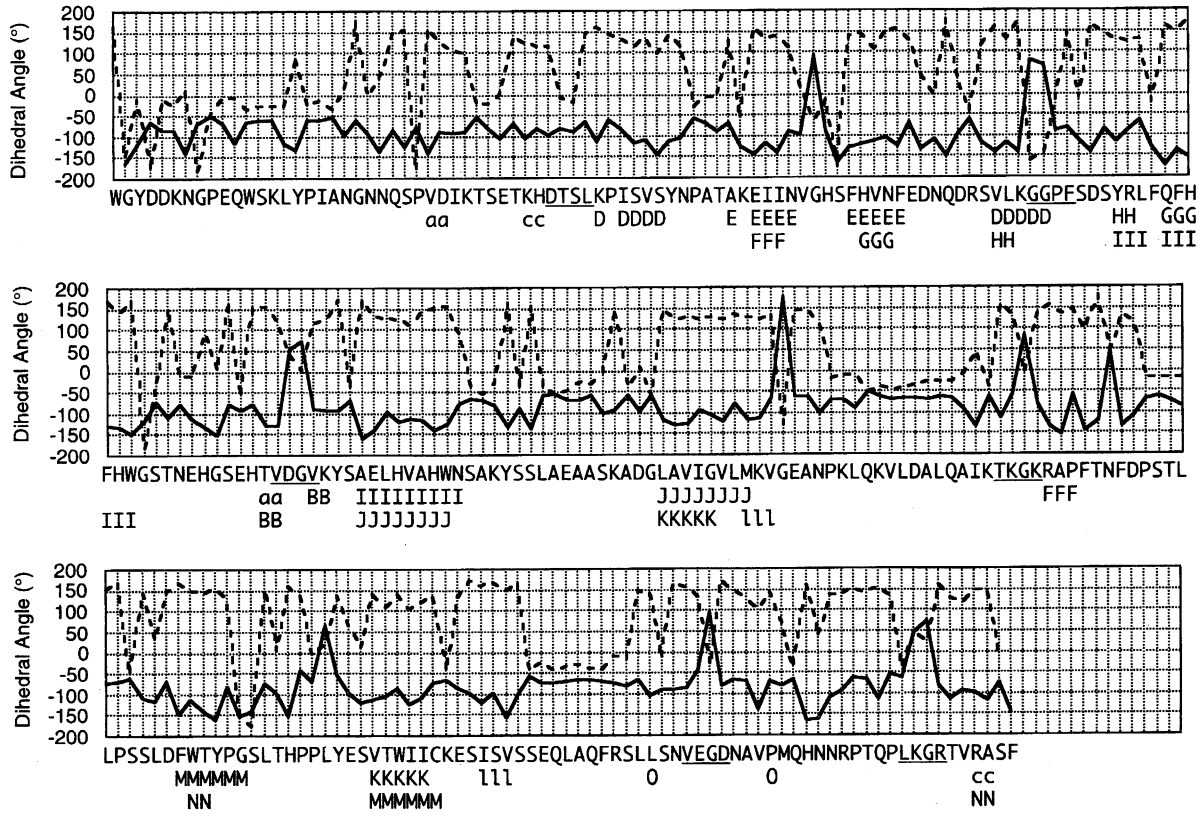

Fig. 3. The ϕ, ψ representation of tertiary structure of carbonic anhydrase (**6**, 2cab) in the form of two-dimensional diagram. The four types I, I′, II, and II′ of β-turns are indicated by underlines. DTSL (I), VDGK (I′), LKGR (I′), TKGK (II), VEGD (II), and GGPF (II′). A combination of intersegment hydrogen bonds among β-strands is shown in Fig. 3 a, B, etc.: see text.

but departs from helical values of the ϕ, ψ angles, as depicted in Figs. 1, 2, and 3. Helices are identified as sequences of at least six consecutive residues. Using ϕ, ψ representations of 125 protein tertiary structures in the form of a two-dimensional diagram, 681 helices were extracted from 125 proteins. As well as the end points of helices, it is convenient for the assignment of a β-strand that those of β-strands in proteins are also defined in terms of the ϕ, ψ dihedral angles. β-Strands defined as intrastrand segments, for example, comprise a ϕ angle of between $-180°$——$-45°$ and between $+145°$—$+180°$ and a ψ angle of between $-180°$——$-150°$ and between $+90°$—$+180°$. β-Strand ends as well as helix ends can also be defined by N-cap and C-cap residues which depart from the ϕ, ψ values in the β-strand region. Furthermore, intrastrand segments consist of at least four consecutive residues in the ϕ, ψ β-strand region. By this definition, the β-strand shown in Fig. 2 (GPSQPTYP) is built up from one continuous segment, which is extended and free from an intersegment hydrogen bond. A combination of intersegment hydrogen bonds among β-strands can be shown in the two-dimensional diagram. Along the sequence, the first sheet is named "a", the second "B", etc. according to the Kabsch and Sander assignment.[10] In Fig. 3, a, c, k, and l are for parallel β-sheets, B, D, E, F, G, H, I, J, M, N, and O for antiparallel ones.

This ϕ, ψ representation is useful in readily recognizing all of the secondary structures as one progresses along the sequence. Regions of nonrepetitive secondary structure can be recognized as partial linking patterns. β-Turns extracted from 125 proteins using the automatic Kabsch and Sander assignment[10] can be easily recognized and classified into the four most common conventional types (I, I′, II, and II′) and a

miscellaneous one (IV).[11] Each of them can be characterized by the $i+1$ and $i+2$ residues of a β-turn. For example,the typical ϕ, ψ patterns of four types of β-turns are easily recognized in Fig. 3. Typical patterns of these β-turn types shown in Fig. 3 are as follows: type-I, DTSL; type-I′, VDGV and LKGR; type-II, TKGK and VEGD; type-II′, GGPF.

In order to analyzed amino acid preferences for specific locations at the type-II β-turn, 195 of type-II β-turns from 125 proteins were listed in Table 1 and ascertained position-specific preferences of single 20 common amino acid residues at type-II β-turns (Table 2). In Table 2, the upper entry is the observed occurrence number of an amino acid residue out of the total number (195) at each of the four i, $i+1$, $i+2$, and $i+3$ positions of type-II β-turns. The middle entry is average percentage composition at each of the four positions of the type-II β-turn. The lower entry is a normalized preference (NP) value defined in Discussion. It is the ratio of the average percentage composition at each of the four positions of the type-II β-turn based on the average percentage composition at large, listed at the top in Table 2. NP-Values of single 20 common amino acid residues are tabulated for each of the four positions of the type-II β-turn, as given in Table 2. Type-II turns favor A, K, and P at i; E, K, and P at $i+1$; G and N at $i+2$; C, K, and Q at $i+3$. As previously observed, the most notable of these are for P at the $i+1$ position and for G at the $i+2$ position.

Discussion

The new type of two-dimensional ϕ, ψ diagrams of tertiary structure of homologous proteins make it possible to clarify the differences in their secondary and tertiary structures in detail. Figure 1 clearly shows that the three-dimensional structures of 2cro and 2orl-L are nearly equal. The new type of two-dimensional ϕ, ψ diagram seems to have many advantages over many other types of representations reported in the literature.[3] For example, the amino acid sequence of a protein is represented by one-letter symbols on the X-axis, and the relationship among primary, secondary, and tertiary structures of homologous proteins can be at a glance recognized on it. Based on the assignment of helices and β-turns, 681 helices were easily extracted from 125 proteins, and β-turns were readily classified. The end points of β-strands in proteins are distinctly assigned by their definition in terms of the ϕ, ψ angles regardless of their intersegment hydrogen bonds (Figs. 2 and 3). The β-sheet is built up from a combination of several β-strands of the polypeptide chain. Alternatively, a β-strand as well as helix is built up from one continuous segment. Therefore, it appears to be rational that the β-strand is classified into a secondary structure instead of a β-sheet. Usually β-strands are from 6 to 10 residues long and aligned adjacent to each other, such that hydrogen bonds can form between the C=O groups of one strand and NH groups of an adjacent β-strand, and vice versa.[17] By the definition of the helix and β-strand ends in the terms of the ϕ, ψ dihedral angles and β-turns in terms of hydrogen bond, the three secondary structures make up more than 80 percent of the conformational structure of most of 125 proteins. 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 as internal standards for assigning of the relationship between primary and secondary structures in a protein. As a typical example, the 195 type-II β-turns are specified by the ϕ, ψ dihedral angles of two residues which occupy the $i+1$ and $i+2$ positions of the type-II β-turn using G residues in a protein as internal standards.

In order to analyze amino acid preferences for specific locations at the $i+1$ and $i+2$ positions of type-II β-turns more precisely, we used 8000 (20^3) kinds of amino acid residues for the analysis. Except for the N- and C-terminal amino acid residues of a protein, we can regard that amino acid residues in sequence of proteins comprise 8000 (20^3) kinds of residues by taking into account the difference in their N- and C-sides residues. Namely, amino acid residues in sequence can be regarded as those in the middle of triplets

Table 2. Position-Specific Amino Acid Preference at the Type II β-Turn[a]

Amino acid residue	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
%[b]	8.2	1.9	5.9	5.9	3.9	8.2	2.2	5.5	6.5	8.0	1.9	4.7	4.6	3.8	4.0	6.7	6.2	7.0	1.4	3.5
	23	3	5	11	10	15	4	10	22	16	2	9	14	6	4	10	9	12	1	9
i	12	1.5	2.6	5.6	5.1	7.7	2.0	5.1	11	8.2	1.0	4.6	7.1	3.1	2.0	5.1	4.6	6.1	0.51	4.6
	1.5	0.79	0.44	0.95	1.3	0.94	0.91	0.93	1.7	1.0	0.53	0.98	1.5	0.82	0.50	0.75	0.74	0.87	0.39	1.3
	18	0	4	17	4	4	3	9	22	7	0	6	46	8	5	18	11	7	1	5
$i+1$	9.2	0	2.0	8.7	2.0	2.0	1.5	4.6	11	3.6	0	3.1	23	4.1	2.6	9.2	5.6	3.6	0.51	2.6
	1.1	0	0.34	1.5	0.51	0.24	0.68	0.84	1.7	0.45	0	0.66	5.0	1.1	0.65	1.4	0.90	0.51	0.39	0.74
	2	0	10	3	1	141	2	0	4	0	1	16	0	3	3	3	1	1	0	4
$i+2$	1.0	0	5.1	1.5	0.51	72	1.0	0	2.0	0	0.51	8.2	0	1.5	1.5	1.5	0.51	0.51	0	2.0
	0.12	0	0.86	0.25	0.13	8.8	0.45	0	0.31	0	0.27	1.7	0	0.39	0.38	0.22	0.082	0.073	0	0.57
	20	8	16	13	3	18	4	10	20	4	4	2	0	16	10	16	12	10	2	7
$i+3$	10	4.1	8.2	6.7	1.5	9.2	2.0	5.1	10	2.0	2.0	1.0	0	8.2	5.1	8.2	6.1	5.1	1.0	3.6
	1.2	2.2	1.4	1.1	0.38	1.1	0.91	0.93	1.5	0.25	1.1	0.21	0	2.2	1.3	1.2	0.98	0.73	0.77	1.0

a) The upper, the middle, and the lower entries: see in text. b) Average percentage composition at large.

consisting of consecutive amino acid residues and each of single 20 common amino acid residues in sequence comprise 400 (20^2) kinds of residues in the middle of triplets. A protein consisting of n single amino acid residues is generally regarded to be constructed by n-2 amino acid residues in the middle of triplets, and the absence of a chemical peptide bond and missing density in the crystallography map reduce the numbers of triplets in the protein. Triplets in 125 proteins were obtained by sliding one amino acid residue step-by-step from N- to C-termini. The proteins listed in Table 1 consist of 23132 amino acid residues and are regarded to be constructed by 22868 ($23132-2\times125-2\times7$) triplets. Among 8000 possible triplets, 6352 independent triplets were found in 125 proteins and among them 1631 independent triplets were found only once and 1287 independent triplets, twice. Thus, 18664 (82%) total triplets were among 3434 independent triplets which were found three times or more often in the data set.

Many kinds of triplets were preferentially observed at the $i+1$ and $i+2$ positions of the type-II β-turn. Their specific locations at these positions were clearly shown. Here, the specific locations of a triplet at the $i+1$ position of the type-II β-turn means that of the amino acid residue in the middle of the triplet. For a more detailed analysis, we also introduce the inherent preference (IP) value of a triplet for a specific location at a particular position of secondary structure as Eq. 1. In Eq. 1, an IP-value of a certain triplet is obtained for a helix region. The IP-value of a triplet has the same meaning as that of the amino acid residue in the middle of the triplet. The IP-value is the overall percentage of the observed occurrence number of a certain triplet at a particular position of the secondary structure. An IP-value of 100, for example, will then mean that all the certain triplets in the data set occur at a particular position of the secondary structure. Each of the 8000 triplets has IP-values which are inherent for each particular position of the secondary structure. However, a normalized preference (NP) value for single amino acid residues (Table 2) is to the ratio of average percentage composition at the particular position of secondary structure based on the average percentage composition at large. Statistically, this also has the same meaning as the normalized frequency of occurrence, an f-value for doublets[18] and to Pα of Chou and Fasman[19] for single amino acid residues in a helix region. In Eq. 2, the NPα-value of a triplet in a helix region is defined. When we define a structure (S) factor of the data set for a helix region as the following Eq. 3, the NPα-value is related to both the IPα-value and the Sα-factor through Eq. 4. Pα of Chou and Fasman had been defined by Eq. 4 for each of 20 single common amino acid residues. Although an NP-value and an S-factor are dependent on data sets, the IP-value of each of the 8000 triplets defined by Eq. 1 is independent of data sets and inherent for each particular position of the secondary structure.

$$\mathrm{IP}\alpha = \frac{\text{(no. of the certain triplets in the helix regions)}}{\text{(no. of the certain triplets in the data set)}} \times 100 \quad (1)$$

$$\mathrm{NP}\alpha = \frac{\dfrac{\text{(no. of the certain triplets in the helix regions)}}{\text{(no. of all triplets in the helix regions)}}}{\dfrac{\text{(no. of the certain triplet in the data set)}}{\text{(no. of all triplets in the data set)}}} \quad (2)$$

$$\mathrm{S}\alpha = \frac{\text{(no. of all triplets in the helix regions)}}{\text{(no. of all triplets in the data set)}} \times 100 \quad (3)$$

$$\mathrm{NP}\alpha = \frac{\mathrm{IP}\alpha}{\mathrm{S}\alpha} \quad (4)$$

The statistical significance of the NP-value used in this study is more simple and clearer than that of the d-test based on a normal distribution, which was used in a previous study.[20] Practically, the position-specific preferences of single 20 common amino acid residues at type-II β-turns can be clearly exhibited by using their NP-values. An NP-value of a unity in Table 2 for a single amino acid residue Y, for example, at the $i+3$ position will then mean that the single Y residue is observed at a position that is same as often as at large. A single P residue is the most strongly preferred amino acid residue at position $i+1$ of the type-II turn, and the NP-value of the P residue at the position $i+1$ is 5.0. This preference value would mean that the single P residue occurs at this position 5.0 times as often as at large. This is due to the fact that the inherent distinction on ϕ to about $-60°$ corresponds to the ψ ($i+1$) requirement for the turn type. The NP-value of a single G residue at $i+2$ shows an extremely high value of 8.8. The position $i+2$ of the type-II turn is dominated by a G residue and to a lesser extent by D and N residues. These three amino acid residues readily adopt the +/+ ϕ, ψ conformation and account for 86% of the $i+2$ position of the type-II turns found. However, the NP-value of single I, L, and P residues at the $i+2$ position is zero. This value means that the I, L, and P residues do not occur at this position. Low NP-values of A, C, F, T, V, and W at the position $i+2$ mean that these residues rarely adopt the +/+ ϕ, ψ conformation. Since the total number of each of the four positions of type-II β-turns in 125 proteins is 195 and the number of amino acid residues used in this study is 23132, this data set S-factor for each of the four positions is 0.84 ($195\times100/23132$). This data set S-factor (0.84) for each of the four positions of the type-II β-turn means that the $i+2$ positions in 125 proteins, for example, comprise 0.84% of the total positions in 125 proteins. A single G residue occurs at this position 8.8-times as often as at large. While, a single P residue occurs at the $i+1$ position 5.0 times as often as at large.

As mentioned above, each of 20 common amino acid residues in sequence of proteins comprises 400 kinds of amino acid residues by taking into account the difference of its N- and C-sides residues. An analysis of the observed single amino acid preferences listed in Table 2 using 8000 (20×400) kinds of amino acid residues is expected to add to a more detailed understanding of the type-II β-turn. In Tables 3 and 4, the triplets found at the $i+1$ and $i+2$ positions in 195 β-turns are assembled. Triplets found out at the $i+1$ and $i+2$ positions mean the amino acid residues in the middle of the triplets. The observed occurrence numbers of triplets at these positions and at large and their IP-values are

Table 3. The Observed Occurrence Number of Triplets and Their IP-Values at the $i+1$ Positions of Type II Turns

Triplet	The observed occurrence number		IP-value	Triplet	The observed occurrence number		IP-value
	at $i+1$	at large			at $i+1$	at large	
AAD	1	10	10	KVG	2	10	20
AAG	2	21	10	LAR	1	5	20
AAN	1	7	14	LDN	1	4	25
ADG	1	16	6	LFG	1	5	20
AEG	2	11	18	LHG	1	3	33
AEN	1	5	20	LKE	1	13	8
AHG	1	5	20	LKF	1	5	20
AIA	1	12	8	LKG	1	12	8
AKG	2	15	13	LNG	1	5	20
AKS	1	12	8	LPG	2	9	22
ANG	1	6	17	LPN	2	4	50
APG	4	10	40	LQG	1	10	10
APS	1	4	25	LSG	2	16	13
ASG	1	15	7	LWG	1	2	50
ATG	1	15	7	MLG	1	4	25
AVD	1	9	11	MLR	1	4	25
AVG	1	9	11	NAG	2	8	25
CEG	1	6	17	NAN	1	4	25
CGG	1	3	33	NEK	1	3	33
CKG	1	2	50	NPG	2	5	40
DEG	1	6	17	NQG	1	9	11
DIQ	1	1	100	NYG	2	5	40
DPG	1	3	33	PEG	3	7	43
DSG	1	8	13	PKG	1	3	33
DTS	1	5	20	PPE	1	7	14
EEG	1	9	11	PPG	4	6	67
EEK	1	4	25	PPH	1	2	50
EKG	2	11	18	PSG	2	7	29
EKY	2	6	33	PSK	1	4	25
ELG	1	11	9	PYN	1	5	20
ENG	1	8	13	QAG	1	4	25
EPG	1	4	25	QKE	1	2	50
ESG	2	12	17	QNG	1	6	17
FAG	1	7	14	QPN	1	1	100
FKG	1	3	33	QQT	1	4	25
FPG	2	4	50	QSG	1	3	33
FPN	1	2	50	REG	1	5	20
FPQ	1	3	33	RPG	2	6	33
FQN	1	4	25	RSA	1	10	10
FTD	1	6	17	SEV	1	7	14
FTG	2	6	33	SEY	1	4	25
GEG	1	3	33	SHY	1	1	100
GFG	1	3	33	SIG	1	7	14
GGG	1	11	9	SPG	3	6	50
GIG	2	11	18	SSG	1	11	9
GLG	1	7	14	STG	1	8	13
GPG	2	6	33	STK	1	8	13
GPM	1	1	100	TAG	1	13	8
GQG	1	4	25	TDR	1	1	100
GRG	1	8	13	TIG	1	4	25
GSG	1	16	6	TKG	1	8	13
GSN	1	8	13	TLD	1	10	10
GTG	1	13	8	TLG	1	4	25
GYD	1	6	17	TPG	1	5	20
HIG	1	4	25	TQG	1	4	25
HPG	1	6	17	TVN	1	5	20
HQG	1	2	50	VEG	1	5	20
HTG	1	1	100	VFG	1	7	14
IDG	1	10	10	VGD	1	13	8
IIG	1	8	13	VGG	1	16	6
IKD	1	7	14	VKG	1	8	13
IPG	3	8	38	VPD	1	8	13
IRG	2	6	33	VPG	2	5	40
ISG	2	8	25	VSG	2	9	22
KAD	1	13	8	VVG	1	9	11
KAG	3	11	27	VYQ	1	1	100
KAH	1	5	20	WTG	1	3	33
KEN	1	6	17	YAG	1	8	13
KFG	1	2	50	YAN	1	7	14
KIN	1	5	20	YKG	1	4	25
KKD	1	5	20	YNG	1	3	33
KKG	3	12	25	YNN	1	4	25
KLG	1	8	13	YPG	1	3	33
KPG	5	7	71	YQG	1	4	25
KRG	1	8	13	YRG	1	5	20
KTG	1	4	25	YVG	1	7	14

Table 4. The Observed Occurrence Number of Triplets and Their IP-Values at the $i+2$ Positions of Type II Turns

Triplet	The observed occurrence number		IP-value	Triplet	The observed occurrence number		IP-value
	at $i+2$	at large			at $i+2$	at large	
ADS	2	8	25	NGS	1	7	14
AGA	1	17	6	NGV	1	6	17
AGC	1	3	33	NNI	1	2	50
AGD	2	8	25	PDY	1	7	14
AGG	2	5	40	PET	1	11	9
AGK	2	11	18	PGA	7	12	58
AGL	1	13	8	PGC	1	1	100
AGQ	1	6	17	PGD	4	12	33
AGW	1	1	100	PGE	3	5	60
AHE	1	4	25	PGH	2	2	100
ANM	1	2	50	PGI	3	7	43
ANR	1	4	25	PGK	3	5	60
ANS	1	5	20	PGM	1	1	100
ARQ	1	1	100	PGQ	2	4	50
DGG	1	4	25	PGR	2	5	40
DGT	1	5	20	PGS	4	14	29
DNY	1	2	50	PGT	3	12	25
DRC	1	3	33	PGV	1	5	20
EGA	1	9	11	PHA	1	1	100
EGD	1	6	17	PMC	1	1	100
EGG	1	3	33	PNF	1	1	100
EGI	1	8	13	PNG	1	8	13
EGK	2	9	22	PNW	1	2	50
EGQ	2	7	29	PNY	1	5	20
EGS	1	5	20	PQT	1	6	17
EGV	1	8	13	PSQ	1	4	25
EGY	1	6	17	QGD	1	5	20
EKG	1	11	9	QGE	1	5	20
EKK	1	4	25	QGN	1	3	33
ENF	1	3	33	QGQ	1	4	25
ENM	1	5	20	QGS	1	4	25
EVC	1	2	50	QGT	1	4	25
EYS	1	4	25	QNE	1	5	20
FGH	1	4	25	QTV	1	4	25
FGK	2	9	22	RGC	1	4	25
FGR	1	5	20	RGD	1	6	17
GDI	1	6	17	RGK	1	6	17
GGA	1	14	7	RGQ	1	5	20
GGE	1	4	25	RGS	1	5	20
GGI	1	15	7	SAG	1	11	9
HGG	1	2	50	SGA	3	12	25
HGQ	1	3	33	SGD	1	12	8
HYG	1	4	25	SGG	1	10	10
IAG	1	10	10	SGH	1	2	50
IGF	1	3	33	SGK	2	12	17
IGG	1	14	7	SGQ	2	7	29
IGK	1	4	25	SGT	1	14	7
IGR	2	11	18	SGV	3	15	20
IGY	1	4	25	SGY	1	4	25
INE	1	5	20	SKS	1	7	14
IQQ	1	3	33	SNV	1	4	25
KDM	1	1	100	TDY	1	6	17
KDT	1	5	20	TGA	2	11	18
KEK	1	8	13	TGE	1	7	14
KES	1	6	17	TGI	1	7	14
KFS	1	4	25	TGQ	1	6	17
KGC	1	3	33	TGR	1	5	20
KGD	3	8	38	TGS	1	19	5
KGE	2	8	25	TGV	1	6	17
KGI	1	11	9	TKI	1	5	20
KGK	1	8	13	TSV	1	6	17
KGL	1	7	14	VDR	1	3	33
KGQ	1	6	17	VGA	1	8	13
KGR	1	3	33	VGD	2	13	15
KGT	3	11	27	VGK	2	10	20
KSA	1	10	9	VNA	1	7	14
KYG	2	10	20	WGL	1	1	100
LDK	1	8	13	YDR	1	1	100
LGC	1	3	33	YGE	1	3	33
LGD	1	8	13	YGQ	1	5	20
LGG	1	14	7	YNN	1	4	25
LGK	1	8	13	YQA	1	2	50
LGL	1	12	8				
LRG	1	6	17				
NGE	1	2	50				
NGG	2	9	22				

Table 5. The Observed Number of Kinds of Triplets in 125 Proteins

Triplet[a)]	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Kinds[b)]	366	222	345	346	302	376	259	342	354	361	219	329	322	313	330	358	349	364	199	300

a) An amino acid residue in the middle of a triplet. b) The observed number of kinds of triplets.

tabulated in Tables 3 and 4. An IP-value of 71 for KPG at the $i+1$ position of the type-II β-turn was obtained by the observation that 7 P residues in the middle of KPG are found at large and 5 P residues out of them, at the $i+1$ position (Table 3).

Ten APG triplets are observed at large and their 4 middle P residues are observed at the position $i+1$. Therefore, the IP-value of APG for the position $i+1$ is 40. Since the S-factor for this position is 0.84, an NP-value of APG for this position is 48 (40/0.84). This value means that APG occurs at this position 48 times as often as at large. It is outstandingly greater than the NP-value of the single P residue. Similarly, NP-values of P residues in the following triplets: APS, IPG, EPG, GPG, FPN, FPQ, GPG, GPM, IPG, KPG, LPG, LPN, NPG, PPG, PPH, QPN, RPG, SPG, TPG, VPD, VPG, and YPG, are 29 or greater and an NP-value of P residues in HPG and PPE is 17. That of the P residues in the middle of other triplets is zero. The high NP-value of the single P residue, 5.0 in Table 2, is the average preference of P residues in the middle of triplets occurring in 125 proteins. As tabulated in Table 5, out of 400 possible P residues, 322 kinds of P residues were practically found in 125 proteins. The NP-values of a single P residue in Table 2 are the average preference of the 322 kinds of P residues.

The NP-value of a single K residue, 1.7 in Table 2, is also the average preference of K residues in the middle of triplets occurring in 125 proteins. K residues in the middle of the limited triplets out of 400 possible kinds of triplets can be observed at the $i+1$ position of the type-II β-turn. For example, for 1 out of 2 CKG, 2 out of 6 EKY, 3 out of 12 KKG and 1 out of 2 QKE, their middle K residues occur at the position $i+1$ of the type-II β-turn. These K residues contribute to the high NP-value of the single K residue.

Although the NP-values at the $i+1$ position of single amino acid residues A, E, I, R, S, T, V, and Y are 1.1, 1.5, 0.84, 0.65, 1.4, 0.90, 0.51, and 0.74, respectively, and are not so large, these values are the average preferences of amino acid residues in the middle of various triplets. These residues can be preferentially observed at high frequency of occurrences in the middle of particular triplets. Practically, the A, E, I, S, T, and Y residues could be observed at the $i+1$ position of type-II β-turns twice or more often in the middle of the following triplets: AEG, ESG, FTG, GIG, GSG, IRG, KAG, KVG, LSG, NAG, NYG, PEG, PSG, and VSG. As tabulated in Table 3, these triplets have high IP-values at the $i+1$ position of the type-II β-turn. Therefore, their NP-values are also high and in the range between 7 and 50.

In Table 4, triplets having strong preferences for specific locations at the position $i+2$ are also assembled. The observed occurrence numbers of triplets at that position and at large and their IP-values are listed in Table 4. A remarkable high NP-value (8.8) of the single G residue at the $i+2$ position is also the average preference of the G residues in the middle of various triplets. Out of 400 possible G residues, 376 kinds of G residues can be found in 125 proteins and 97 out of 376 G residues have high NP-values. Although the NP-values at the $i+2$ position of single amino acid residues for A, E, F, H, K, M, Q, R, S, T, V, and Y are 0.12, 0.25, 0.13, 0.45, 0.31, 0.26, 0.39, 0.38, 0.22, 0.08, 0.07, and 0.59, respectively, and quite low, these values are the average preferences of amino acid residues in the middle of a variety of triplets. These residues in the middle of particular triplets can be observed frequently. Triplets having high IP-values are in the following: AHE, ARQ, DRC, EKG, EKK, EVC, EYS, HYG, IAG, IQQ, KEK, KES, KFS, KSA, KYG, LRG, PET, PHA, PMC, PQT, PSQ, QTV, SAG, SKS, TKI, TSV, and YQA. Although the conformational preference of each of single amino acid residues is weak, limited residues out of 8000 kinds of residues have extremely strong amino acid preferences for specific locations at the $i+1$ or $i+2$ position of the type-II β-turn.

In conclusion, the new type of ϕ, ψ representation of protein tertiary structure makes it possible to draw a precise protein tertiary structure in the from of a two-dimensional diagram. The tertiary structure of homologous proteins can be analyzed in detail using this diagram. These results lead us to elucidate pathways to protein folding. In order to demonstrate amino acid preferences for specific locations at particular positions of the secondary structure, the type-II β-turn was used as an example. At 0.84% (195 positions) of the total positions, particular amino acid residues out of 8000 showed extremely strong preferences for specific locations at the $i+1$ or $i+2$ position. The preferences reflect the stereochemistry of amino acid residues in the middle of triplets.

We thank C. Sander's research group for the use of their data bases, and also thank the crystallographers for the use of the atomic coordinates of the 125 proteins.

References

1) 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/).

2) IUPAC-IUB Commission of Biochemical Nomenclature, *Biochemistry*, **9**, 3471 (1970).

3) R. D. McClain and B. W. Erickson, *Int. J. Peptide Protein Res.*, **45**, 272 (1995).

4) R. Balasubramanian, *Nature*, **266**, 856 (1977).

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

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

7) V. V. Solovyev and A. A. Salamov, *Comput. Appl. Biosci.*, **10**, 661 (1994).
8) A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, *J. Mol. Biol.*, **247**, 536 (1995).
9) W. Kabsch and C. Sander, *Biopolymers*, **22**, 2577 (1983).
10) We used the Definition of Secondary Structure of Protein (DSSP) data base of C. Saner's research group (http address://www.embl-heidelberg.de/).
11) J. S. Richardson, *Adv. Protein Chem.*, **34**, 167 (1981).
12) A. K. Aggarwal, D. W. Rodgers, M. Drottar, and M. Prashne, *Science*, **242**, 899 (1988).
13) A. Mondragon, C. Wolberger, and S. C. Harrison, *J. Mol. Biol.*, **205**, 179 (1989).
14) T. L. Blundell, J. E. Pitts, I. J. Tickles, S. P. Wood, and C. -W. Wu, *Proc. Natl. Acad. Sci. U.S.A.*, **78**, 4175 (1981).
15) K. K. Kannan, M. Ramanadham, and T. A. Jones, *Ann. N. Y. Acad. Sci.*, **429**, 49 (1984).
16) J. S. Richardson and D. C. Richardson, *Science*, **240**, 1648 (1988).
17) C. Branden and J. Tooze, "Introduction to Protein Structure," Garl and Publishing (1991).
18) E. T. Harper and G. D. Rose, *Biochemistry*, **32**, 7605 (1993).
19) P. Y. Chou and G. D. Fasman, *Biochemistry*, **13**, 222 (1974).
20) C. M. Wilmot and J. M. Thornton, *J. Mol. Biol.*, **203**, 221 (1988).