Bull. Inst. Chem. Res., Kyoto Univ., Vol. 66, No. 4, 1989

Correlation of Amino Acid Sequence to Higher-ordered Structure of Globular Proteins

Hiroshi NAKASHIMA

Received September 30, 1988

The profile of location measure (N_{14}) obtained from the X-ray structure of a protein gives information on a relative location of a residue in the molecule from the center of mass in the protein. We analyzed protein conformations statistically in terms of segments, using the location profiles of 87 proteins of known three-dimensional structures. The protein chain was divided into segments along the sequence at minima of the smoothed N_{14} profile, then the character of a segment was expressed by corresponding original location profile. Cutting regions into segments are designated as c-segments and analyzed in connection with the segments. A protein conformation was described by an arrangement of characterized segments and c-segments. The feature of segments was analyzed in terms of secondary structures and occurrence of segment pairs. The sites in a polypeptide chain to divide into segments could be predicted as minima of a location profile calculated from its amino acid sequence with an accuracy of 77%. The position of the first minimum of a smoothed profile showed a good correspondence to the cleavage site for a signal peptide of a secretory precursor protein. The significance of this approach to describe protein conformation by characterized segments is discussed.

KEY WORDS: Protein conformation/ Segment/ Location measure profile/

INTRODUCTION

The information on the native tertiary structure of a protein is encoded in its amino acid sequence[1]. In order to decipher the information in the amino acid sequence, the sequence has to be converted into a series of numerals replacing each amino acid residue to such a quantity as hydrophobicity. It is important to find a quantity that shows a high correlation between the tertiary structure and its sequence. Nishikawa and Ooi[2] found that the following quantity (N_{14}) has a high correlation between the 3D structure of a protein and its amino acid sequence; the quantity N_{14} is given for each residue along the sequence as a number of C^{α} atoms within a sphere of radius 14Å. Thus, the number N_{14} is small for a residue at the surface and large in the interior of a globular protein. The quantity N_{14} (designated as a location measure) has a good correlation with the distance from the center of mass in a protein. Another advantage of the use of the location measure is that a profile can be calculated from the sequence with a good accuracy[2] by allotting an optimized value to each amino acid. Therefore, the present analysis on protein sequence and conformation has been performed using the location measure profiles.

Many workers have used a segment as a structure unit according to their definition to examine the chain folding[3,4] or turn conformation[5] in X-ray structures of proteins, but segments have not been analyzed systematically so far. The segmentation may be done after smoothing along the polypeptide chain using a location measure

中島広志 : School of Allied Medical Professions, Kanazawa University, Kanazawa, 920, Japan

profile of a zig-zag curve. A polypeptide chain is divided at minima of the smoothed profile obtained by a similar method to those by other people[3,4]. The site of every minimum of the smoothed profile corresponds to a turn conformation[5], so that segments may be regarded as a part of the polypeptide chain from a turn to the next turn. Regions at cutting positions in the chain are defined as c-segments, which are analyzed in connection with segments. Recently, we have reported[6] the characterization and classification of both segments and c-segments from various proteins, and described a protein conformation in terms of the characterized segments and c-segments. The further implication of the segmentation will be described in this report, and the correlation between amino acid sequences and characterized segments is reconfirmed.

METHODS

Protein data

The 87 proteins consisted of a single polypeptide chain more than 50 residues were selected from the Protein Data Bank[7] (PDB) for the present analysis. The 87 proteins shown in Table I in the PDB code are the same as in the previous work[6]. They differ with each other by at least 50% in their sequences. The location of secondary structures (α-helix, β-structure, and coil) described in the PDB was used, and the turn was included in the coil conformation.

For the analysis on cleavage sites of signal peptides, signal peptides of 201 known proteins were chosen from the Protein Identification Resources (PIR) database[8] of release 13. Proteins were selected on the basis that they differ mutually by at least

Table I. The average (AV) and standard deviation (SD) of components of segments and c-segments. They were determined by using 1240 segments and 1158 c-segments from 87 proteins, respectively. The 87 proteins are indicated by PDB codes. Five-letter codes are used as follows: 2MHBA and 2MHBB for α- and β-chains of horse hemoglobin, 3FABH and 3FABL for heavy and light chains of immunoglobulin Fab' (NEW), 2ATCR for R-chain of aspartate carbamoyl transferase, 1TGSI for trypsin inhibitor from pig pancreas, respectively.

	component of segments					component of c-segments		
	1	2	3	4		1	2	3
AV	44.92	21.23	1.00	12.02	AV	47.33	19.21	0.93
SD	22.30	9.48	0.22	4.75	SD	26.70	11.28	0.23

PDB codes of 87 proteins

1CPV	1ICB	1CTS	156B	3CYT	2C2C	2CDV	1CC5	155C	351C
1CCY	1CYP	1ECD	1HMQ	2MHBA	2MHBB	1LHB	1LH1	2MBN	1MHR
2APE	1APR	1ACX	1AZA	2STV	4SBV	3CNA	1GCR	1GN5	1EST
3RP2	1HIP	1REI	3FABH	3FABL	1FC1	1CTX	2ALP	1NXB	2APP
1PCY	2PAB	3SGB	3RXN	1SN3	2SOD	2ADK	4ADH	1ABP	1AAT
2ATCR	5CPA	8CAT	3DFR	4DFR	3FXN	1FX1	1GP1	2GRS	2GPD
4LDH	3PGK	3PGM	1RHD	1SBT	2TAA	1SRX	1TIM	2ACT	2CAB
2B5C	1FDX	3FXC	2FD1	1PYP	2LYZ	1LZM	2MT2	1OVO	1TGSI
8PAP	1P2P	5RSA	2SNS	2SSI	3TLN	4PTI			

60% in their precursor sequences of 50 residues at the N-terminal portion. The proteins consist of 116 superfamilies from various sources; 164 from eukaryotic (142 from vertebrate, 12 from invertebrate, and 10 from plant), 22 from prokaryotic, and 15 from virus and phage proteins. The length of signal peptides was in the range from 12 to 34 residues, and its average was 21.8. The zein precursor from maize has the shortest signal peptide of 12 residues long, and α-amylase precursor from *bacillus stearothermophilus* and human lymphotoxin precursor have the longest signal peptides of 34 residues long.

Calculation of N_r profiles and smoothing

The experimental N_r was calculated for each residue along the sequence as a number of C^α atoms within a sphere of radius rÅ using the coordinate data of C^α atoms in PDB. In the calculation of N_r, several neighboring residues (two for N_8, three for N_{14}, five for N_{20}, and six for N_{26}) on both sides of a given residue were omitted from counting to avoid the end effects since these residues are always within a sphere of radius of rÅ because the distance of $C^\alpha_i - C^\alpha_{i+1}$ is a fixed length of 3.8Å. A value of N_{14} for a given residue i along the sequence was devided into two components; one is the number of C^α atoms within a sphere of radius 14Å from residue 1 to residue $i-1$ as the contribution from the N-terminal part (designated as N_{term14}), and the other is the number of C^α atoms from residue $i+1$ to C-terminal residue as the contribution from the C-terminal part (C_{term14}). The sum of two components, N_{term14} and C_{term14}, is equal to the original location measure, N_{14}. The degree of coincidence of minima of the smoothed profile (see below) of N_{term14} and those of C_{term14} was examined as follow: when a site of minimum in a location profile is coincident with those from both curves of N_{term14} and C_{term14} within ± 3 residues, respectively, the site of minimum is assumed to be coincident. The number of coincidence versus the total cutting positions of the location profile was used as the measure of the coincidence. The predicted location profiles were claculated from the protein sequence using the optimized parameter set[2].

The experimental and predicted profiles were smoothed using a Fourier transformation method, *i.e.*, a profile was transformed into Fourier components of angular frequencies from 0 to 180°, then the Fourier synthesis was performed by summing up components of lower angular frequencies. The Fourier synthesis using the frequencies of 0–180° gave an original profile, and the degree of smoothing was dependent on the range of angular frequencies used for the Fourier synthesis. In this work, angular frequencies over 0–45° were used to smooth the profile according to the previous method[6]. The sites of minima of the smoothed N_{14} profiles were assigned as positions for cutting the chain into segments. The minima within five residues of both the N- and C-terminus of the chain were omitted to avoid an artifact of the smoothing.

Characterization of segments and c-segments

The character of a segment was expressed numerically by the following four components using the location measure profile; (1) an average of angular moment

at 40°, $M(40°)$, (2) an average at 100°, $M(100°)$, (3) an average of location measures within a segment, (4) a length of a segment. The angular moment, $M(\theta)$, for the location measure profile was calculated in a frame of 11 residues using relative values of N_{14} from the average, X_i, as follows:

$$M(\theta) = (\{\sum X_i \cos(\theta i)\}^2 + \{\sum X_i \sin(\theta i)\}^2)^{1/2}$$
$$X_i = N_{14,i} - \overline{N_{14}}$$

where $N_{14,i}$ is the value of N_{14} at i-th residue in the frame of 11 residues and $\overline{N_{14}}$ is the mean value of N_{14} in the frame, and the moment was allocated to the central residue. Here we used 40° and 100° for the calculation of angular moments, as they were suitable angles to discriminate α-helix and β-structure[6]. Angular moments were calculated by moving the frame from one residue to another along the sequence. The average of angular moments within a segment was used to characterize the segment. The c-segment was defined as a region of 11 residues long at a junction of two adjacent segments, *i.e.*, five residues on each side of the central residue which is a minimum of the smoothed N_{14} profile. The character of c-segment was expressed by three components; (1) $M(40°)$, (2) $M(100°)$, (3) an average of location measures.

Classification of segments and c-segments into groups

Four components of a segment were normalized using mean values and standard deviations to make the contribution of each component equivalent[6,9,10]; *i.e.*

$$a_{jk} = (c_{jk} - \overline{c_j})/\sigma_j$$

where a_{jk} and c_{jk} are the normalized and original value of component j of segment k, respectively, and $\overline{c_j}$ and σ_j are the average and standard deviation for component j, respectively. A segment in terms of normalized components was then represented as a point in the four-dimensional space spanned with four components as axes. The distance d_{kl} between two segments, k and l, in the space was calculated according to the following equation;

$$d_{kl} = \{\sum (a_{jk} - a_{jl})^2\}^{1/2}$$

where a_{jk} and a_{jl} are the normalized components of the segment k and l, respectively. The distribution of segments in the component space was analyzed in terms of a distance to classify them into several groups. The region of the highest density was searched by examining the number of segments existing within a certain distance r from a given segment. The segment that included the largest number of segments in the space was chosen as the center of the first group. Then, another group of the next highest density was searched using the rest of the segments. This procedure was repeated until the number of remaining segments became less than a cutoff value of 40. Each group has a representative segment at its center. The radius r was chosen in such a way that the largest group included about one-fifth of the total segments. The assignment of segments to a group was done as follows:

a distance from a segment to each one of the 10 representative segments was calculated, and the segment was assigned to the group which had a shortest distance.

Classification of c-segments into groups was done similarly, and a three-dimensional component space was used for the analysis of distribution of c-segments in this case.

Assignment of cleavage sites for signal peptides

The residue which corresponded to the first minimum located after the N-terminal peak in a smoothed N_{14} profile was assigned as the cleavage site for a signal peptide. Any minima within five residues from the N-terminus of the chain were not taken into consideration to avoid an artifact of the smoothing. The degree of coincidence between experimental and calculated data was estimated as follows: when the experimental cleavage site (between -1 and $+1$) is coincident with a calculated one within $\pm i$ residues, we assume that the assignment is correct, and the number of sites correctly assigned and the percentage of the correct assignment are used as the measure of the accuracy. In the calculation, $i=3$, and 4 were used to determine the degree of coincidence.

RESULTS

Segmentation

The smoothed location profile N_{14} along a sequence of a protein generally shows a low-high-low pattern, and the profile reflects the path of a chain in the tertiary structure running from a point at the surface to another through the inside of a molecule. This structural unit, designated as segment, may be a fundamental feature of the chain folding of a globular protein. In order to identify this pattern clearly, the original zig-zag profile was smoothed and the polypeptide chain was divided into segments at minima. Smoothing was done by Fourier transformation using only angular frequencies of 0–45°, since this was an optimum condition to assign the turn conformation at minima[6]. The minima within five residues from the N- and C-terminus were not taken into account to avoid an artifact of smoothing. The 1158 dividing positions into segments were determined from the 87 experimental profiles. Similarly, the positions were assigned from the 87 predicted profiles, and the agreement of the dividing positions within three residues was 77% for 87 proteins[6].

The segmentation was examined for various smoothed profiles of N_r such as N_8, N_{14}, N_{20}, and N_{26}. In Fig. 1, the smoothed N_r profiles of bovine ribonuclease A are shown as function of r, where r increases from 8Å to 26Å with an interval of 6Å. A radius 8Å was used to examine the contact number of a residue in a protein[11]. As shown in Fig. 1, the smoothed N_{14} profile indicates nine minima, and the number and locations of minima of other N_r profiles are almost the same except for the absence of minima at residues 39 and 90 in the N_8 profile, respectively. This result indicates that the choice of a radius from 14Å to 24Å does not affect the result of the segmentation; *i.e.*, the segmentation does not depend on a radius of the sphere

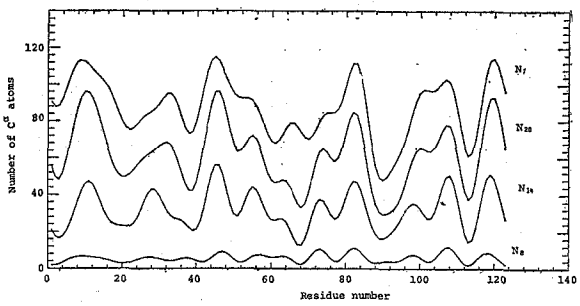

Fig. 1. The smoothed experimental profiles of N_r (r=8, 14, 20, 26) of bovine ribonulease A are plotted along the residue number.

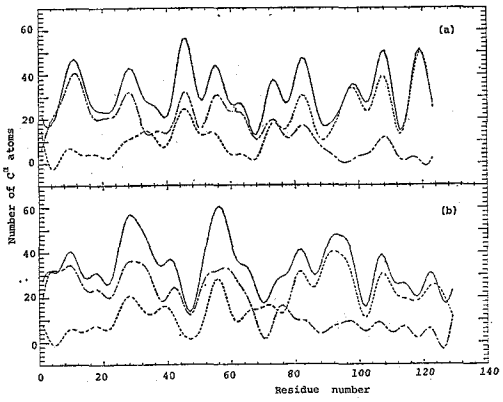

Fig. 2. The smoothed location profile N_{14} (solid line) and its two components; one from the N-terminal part, N_{term14} (broken line) and the other from the C-terminal part, C_{term14} (dot-and-dash line) are plotted along the residue number. Fig. 2a and 2b indicate bovine ribonuclease A, and hen egg lysozyme, respectively.

in this range.

A quantity of N_{14} for a given residue i was divided into two components; one is the contribution from the N-terminal part (N_{term14}) and another is from the C-terminal part (C_{term14}). The smoothed N_{14} profile (solid line), smoothed N_{term14} profile (broken line), and smoothed C_{term14} profile (dot-and-dash line) of bovine ribonuclease A (a) and hen egg lysozyme (b) are shown in Fig. 2. Nine cutting positions into segments were obtained from the smoothed N_{14} profile in both examples. In Fig. 2, the differences of minima between N_{term14} and C_{term14} profiles were found at residues 90 and 102 and at residues 38, 70 and 102 for ribonuclease A (a) and lysozyme (b), respectively. The degrees of coincidence of minima for two proteins were 78% (7/9) and 67% (6/9) for ribonuclease A and lysozyme, respectively. The average conicidence of 87 proteins was 69%.

Angular moment

Angular moments of $M(40°)$ and $M(100°)$ were useful to characterize each segment using its corresponding N_{14} profile. The experimental profiles of location measure (top), and angular moment $M(40°)$ in solid line and $M(100°)$ in broken

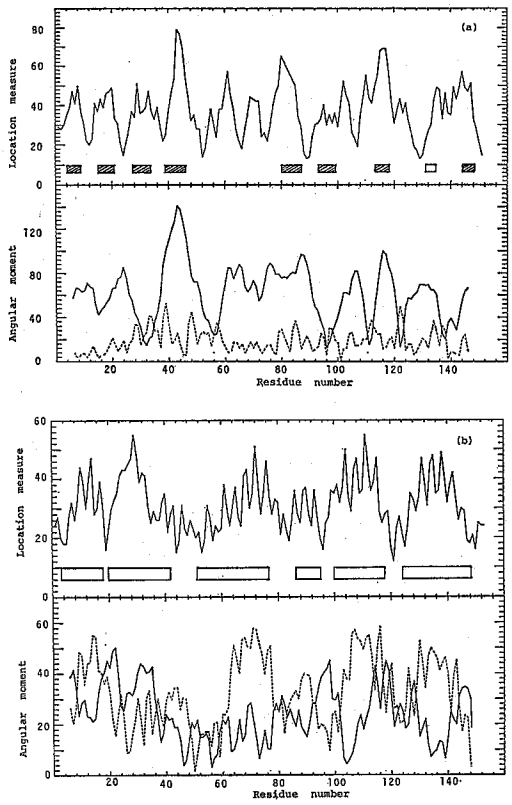

Fig. 3. The experimental location measure profile (top) and the angular moment $M(40°)$ in solid line and $M(100°)$ in broken line (bottom) are plotted along the residue number. Fig. 3a and 3b indicate superoxide dismutase of bovine, and myoglobin of sperm whale, respectively. Open and hatched bars indicate α-helical and β-structural regions, respectively.

line are plotted along the sequence (bottom) in Fig. 3(a) and 3(b). Profiles for superoxide dismutase of bovine and myoglobin of sperm whale are illustrated in Fig. 3a and 3b, respectively. Open and hatched bar represent α-helical and β-structural regions, respectively. As shown in Fig. 3a and 3b, an α-helical region corresponds to a zig-zag pattern of the 3.6 residues repeat and a β-structural region corresponds to a sharp bell-shaped pattern[2], and those patterns can be distinguishable by profiles of $M(40°)$ and $M(100°)$, respectively. The average of angular moments within a segment was used to characterize the segments.

Grouping and characterization of segments

Four components ((1) the average of angular moment at 40°, $M(40°)$, (2) $M(100°)$, (3) the average of location measures within the segment, (4) the length of a segment), were normalized using averages and standard deviations given in Table I, then each segment was represented as a point in the normalized component space. The 1240 segments in total were classified into 10 groups using the radius of $r=1.3$ for grouping. Table II shows the number of segments, the components of the

Table II. The normalized 10 representative segments for classification. The percent of secondary structures and the number of assigned segments for the 10 groups. Component 1, 2, 3, and 4 are $M(40°)$, $M(100°)$, average of location measure, and length, respectively. They were determined by using 1240 segments from 87 proteins.

No.	Group	Component				Conformation		No. of segment
		1	2	3	4	helix	sheet	
1	A1	−0.98	1.93	−0.72	−0.85	81.9	0.3	85
2	A2	−0.63	1.49	0.31	−0.00	79.1	1.2	99
3	A3	−0.42	1.40	−0.09	2.10	68.8	4.3	70
4	B1	2.47	−0.20	1.09	−0.00	2.8	52.1	104
5	B2	0.82	−0.17	0.63	0.21	8.3	45.7	162
6	B3	0.52	−0.87	−0.50	0.21	4.0	38.2	150
7	B4	−0.00	−0.57	1.59	−0.64	29.4	33.2	74
8	C1	−0.89	−0.73	−1.61	−1.06	13.0	18.5	131
9	C2	−0.48	−0.26	−0.38	−0.43	19.8	23.5	246
10	C3	−0.01	0.10	0.55	1.26	39.6	19.7	119

representative segment in the normalized scale, and average content of α-helix and β-structure for each group. As listed in Table II, group 1 had 81.9% α-helix and 0.3% β-structure on the average over 85 segments, and so on. Since the averages of secondary structures in the available proteins were 31.4% and 21.9% for α-helix and β-structure, respectively, three groups in Table II are considered to be favorable for α-helix and four groups to be favorable for β-structure. Therefore, groups favorable for α-helix were designated as A1, A2, A3, and those for β-structure as B1, B2, B3, B4 and those for coil as C1, C2, C3. Even if contents of the secondary structures are similar in the magnitude, segments in group A1 are distinguishable from those in A2 by this grouping. A segment in A1 locates at the more exposed side than that in A2 and the length of A1 group is shorter than A2 group, as is shown by the representative segments in Table II.

Grouping and characterization of c-segments

The 1158 c-segments from the 87 proteins were characterized by three components; an average of angular moment at 40°, $M(40°)$, a similar moment $M(100°)$, and an average of location measures of the c-segment. The three components were normalized using average values and standard deviations given in Table I, and then the 1158 c-segments were classified into eight groups using the radius of $r=1.0$. In Table III, the number of c-segments and components of the representative c-segment in the normalized scale are listed for each group. The contents of α-helix and β-structure for every c-segment group averaged over 11 residues including five residues on each side of a given cutting position are shown also in Table III. According to the contents of secondary structures, c-segments in group 7 and 3 seem to be favorable for α-helix and those in group 2, 8, and 6 favorable for β-structure. The dividing positions into segments were analyzed in terms of N_{term14} and C_{term14} profiles as mentioned in segmentation. The degree of coincidence of minima from N_{term14} profile with those from C_{term14} profile was examined for every group of

Table III. The normalized 8 representative c-segments for classification. Component 1, 2, and 3 are $M(40°)$, $M(100)°$, and average of location measure, respectively. The percent of secondary structures of 11 residues and the number of assigned c-segments for the 8 groups. The degree of coincidence between a component of location number from the N-terminal part and that from the C-terminal part is shown in percentage at the column of coincidence. They were determined by using 1158 c-segments from 87 proteins.

	Component			Conformation		No. of	Percent of
Group	1	2	3	helix	sheet	c-segment	coincidence
1	−0.09	−0.27	−0.20	25.3	17.8	210	71
2	1.20	−0.56	0.28	10.2	37.1	208	86
3	−0.80	0.89	−0.15	42.0	11.9	132	52
4	−0.06	−0.79	−1.26	19.1	14.5	183	67
5	0.61	1.24	0.16	32.8	17.8	99	89
6	−0.81	−0.61	0.58	24.4	31.2	136	59
7	−0.95	1.88	0.71	68.7	7.5	91	41
8	0.27	0.25	1.10	26.4	32.0	99	70

c-segments. As shown in Table III, the coincidence in group 5 was 89% (88/99), and that in group 2 was 86% (179/208), and so on. The degree of coincidence is correlated with the representative value of component 1 or $M(40°)$ in Table III. This result may be explained as follows: when a minimum of the N_{term14} profile overlaps with the corresponding one of the C_{term14} profile, a sharp trough will appear in a location profile, since the location number is a sum of N_{term14} and C_{term14}. A sharp trough in the location profile yields a large value of $M(40°)$ as shown in Fig. 3a.

Occurrence of segment pairs and triplets

Frequencies of occurrence for 1153 segment-segment pairs and 1066 triplets of successive three segments were examined here. We classified the 1240 segments into three types; 254 α-helix favorable segments (in group A1, A2 and A3), 490 β-structure favorable segments (in group B1, B2, B3 and B4), and 496 coil favorable segments (in group C1, C2 and C3) according to the contents of secondary structures. Since segments are classified into the three types, there are nine possible pairs and 27 possible triplets. The observed number of segment pairs are shown with the ratios of observed occurrence versus expected one in Table IV. The expected number was computed from the number of each group assuming that the pairs (or triplets) are random. An occurrence is favorable if its ratio is greater than 1.0, and unfavorable if less than 1.0. Five most favorable and unfavorable triplets are shown in Table IV with the observed numbers and ratios. As shown in Table IV, the segment pairs such as helix-helix and sheet-sheet are more favorable than those of helix-sheet and sheet-helix. Interestingly, the triplet of helix-sheet-helix is found to be favorable, although pairs of helix-sheet and sheet-helix are less favorable.

Frequencies of occurrence for the segment-c-segment-segment were examined for 1153 combinations. Since segments are classified into three types of those favorable

Table IV. Occurrence frequencies of pairs of segments and five most favorable and unfavorable triplets of segments. Segment that is favorable for α-helix, β-structure, and coil is represented as helix, sheet and coil, respectively. The observed numbers are indicated with their ratios (ratio=obs/cal). Occurrence frequencies are analyzed using 1153 pairs and 1066 triplets, respectively.

Segment pair	Ratio	Obs
helix–helix	1.63	70
sheet–sheet	1.16	217
coil–coil	1.06	199
helix–coil	0.99	89
coil–helix	0.98	88
coil–sheet	0.95	179
sheet–coil	0.90	168
sheet–helix	0.88	79
helix–sheet	0.71	64

Favorable triplet	Ratio	Obs	Unfavorable triplet	Ratio	Obs
helix–helix–helix	3.44	25	helix–sheet–coil	0.51	17
helix–coil–helix	2.11	33	helix–sheet–sheet	0.54	18
coil–helix–helix	1.85	29	helix–coil–sheet	0.57	19
helix–sheet–helix	1.67	26	sheet–coil–helix	0.60	20
sheet–sheet–sheet	1.46	102	coil–helix–sheet	0.60	20

Table V. Occurrence frequencies of c-segments at a junction of two adjacent segments. The observed number and three most favorable and unfavorable combinations are shown in percentage of occurrence for each c-segment. Segment that is favorable for α-helix, β-structure, and coil is represented as helix, sheet and coil, respectively.

Group	No.	Favorable					
1	209	coil–coil	25	sheet–coil	18	coil–sheet	17
2	208	sheet–sheet	51	sheet–coil	19	coil–sheet	14
3	132	coil–coil	24	coil–helix	21	helix–coil	19
4	181	coil–coil	32	coil–sheet	19	sheet–coil	12
5	99	helix–sheet	24	coil–sheet	18	sheet–helix	16
6	135	coil–sheet	27	coil–coil	23	sheet–coil	16
7	90	helix–helix	32	coil–helix	17	sheet–helix	12
8	99	sheet–sheet	43	sheet–coil	16	coil–sheet	13
Group		Unfavorable					
1		helix–sheet	3	sheet–helix	4	helix–coil	6
2		coil–helix	0	helix–helix	1	helix–coil	2
3		sheet–sheet	1	helix–sheet	3	sheet–helix	5
4		helix–sheet	4	helix–helix	4	sheet–sheet	7
5		helix–helix	2	coil–coil	3	coil–helix	4
6		helix–sheet	1	helix–helix	1	helix–coil	5
7		sheet–sheet	3	coil–sheet	4	coil–coil	6
8		coil–helix	0	helix–helix	1	helix–coil	5

for α-helix, β-structure, and coil, there are nine combinations for each c-segments. Three most favorable and unfavorable combinations are shown in Table V with the percentage of occurrence. A c-segment in group 1 appears at the junction of two coil segments with a frequency of 25% (53/209), and a c-segment in group 2 at the junction of two β-structural segments with that of 51% (106/208), and so on. A c-segment favorable for α-helix such as that in group 7 is frequently observed (32%) at the junction of two α-helical segments, and a c-segment favorable for β-structure

such as that in group 2 is frequently observed (51%) at the junction of two β-structural segments.

Comparison of homologous sequence and segmentation

Two sets of homologous proteins are shown in Fig. 4. One example pair is sperm whale myoglobin (2MBN) and yellow lupin leghemoglobin (1LH1) from globin family. Another pair is pig elastase (1EST) and rat group-specific protease (3RP2) from serine protease family. Since these two pairs of proteins satisfy four out of five criteria necessary for the detection of the weak sequence similarity[12], the similarity in the tertiary structure is expected for the two pairs of proteins. Myoglobin and leghemoglobin have 17% (27/159) identical amino acid residues in their aligned sequences after the tertiary structure comparison[13]. Another pair of elastase and group-specific protease has 32% (77/241) identical residues in their sequences after

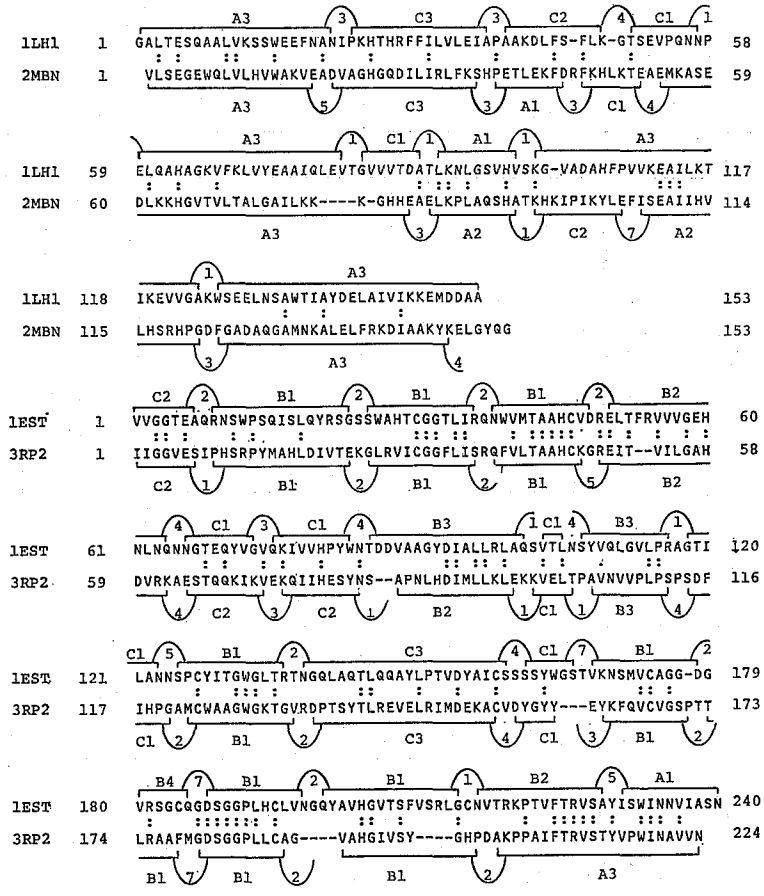

Fig. 4. Two aligned protein sequences are described with the classified segments and c-segment (in brackets). The codes 1LH1, 2MBN, 1EST, and 3RP2 represent yellow lupin leghemoglobin, sperm whale myoglobin, pig elastase and rat group-specific protease, respectively. Residues are numbered according to the order of the C^{α} atom in Protein Data Bank.

homology alignment[14]. The locations and assigned groups of segments are shown along the aligned sequence in Fig. 4. The structural similarities and differences could be identified by comparing the assigned segments. The coincidence of the corresponding segments was 37% (4/9) for the pair of myoglobin and leghemoglobin, and 70% (14/20) for elastase and group-specific protease. Therefore, description of a protein structure in terms of classified segments is useful to detect the structure similarities in their classified segments by comparing the corresponding segments.

Correlation between segmentation and proteolytic processing sites

The 201 proteins with known signal peptides were analyzed using their predicted location profiles. The predicted curve and its smoothed one from bovine proglucagon precursor and human vasoactive intestinal peptide (VIP) precursor are shown in Fig. 5(a) and 5(b), respectively. The assigned cleavage sites for signal peptides are at residues -2 and $+3$, as indicated by the filled arrows in Fig. 5(a) and 5(b). Both of them are known to have a signal peptide of 20 residues long at the N-terminus, so that differences in the number between experimental and predicted sites are -2 and $+3$ for glucagon and VIP (the experimental cleavage site is located between -1 and $+1$), respectively. In this way, the coincidence was calculated on all the proteins, and the results were 71% (142/201) and 85% (171/201) of agreement within ± 3 and ± 4 residues, respectively. The site of the first minimum in the

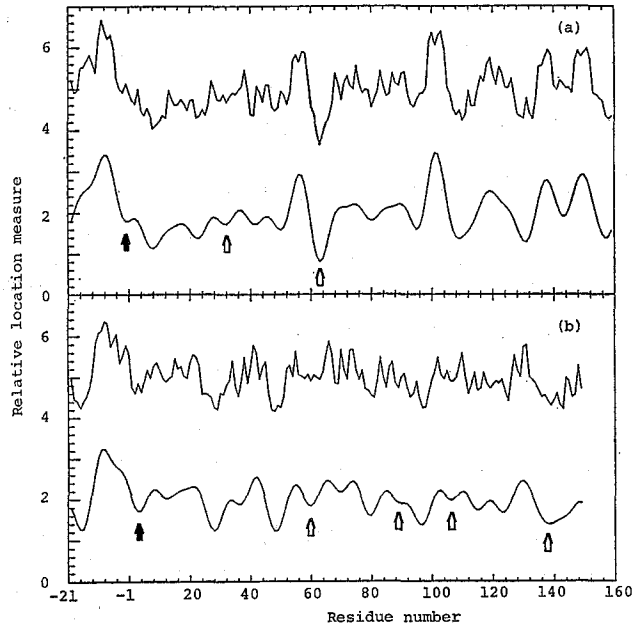

Fig. 5. The computed location measure profiles and their smoothed ones along the residue number of bovine gulcagon precursor (a), and human vasoactive intestinal peptide precursor (b). Filled arrows indicate the assigned processing sites as the cleavage of signal peptides. Open arrows indicate the expected processing sites to yield mature peptide chains. Residues are numbered so as the cleavage site of signal peptide is between -1 and 1.

smoothed profile was adopted as a cleavage site for a signal peptide, while the position of the second minimum corresponded to the experimental cleavage sites in six proteins. This method is quite different from those of other people[15,16], where empirically-derived rules have been used to predict the cleavage site for a signal peptide. The proteins which have a signal peptide show generally a high peak near the N-terminal portion as shown in Fig. 5a and 5b. However, the following seven proteins did not show a high peak in the N-terminal portion; winter flounder antifreeze peptide, *E. coli* lambda receptor, halobacterium halobium bacteriorhodopsin, tobacco ribulose bisphosphate carboxylase large chain, trypanosoma brucei variant S glycoprotein 117 precursor, *E. coli* penicillin-binding protein 5 precursor, and bacteriophage T4 internal protein I precursor.

The residues 33–61 in Fig. 5a correspond to the mature peptide glucagon, and the residues 61–87 and 105–132 in Fig. 5b correspond to the mature peptide PHM-27 and VIP, respectively. The minima are expected to be the processing sites by a proteolytic enzyme, as indicated by open arrows such as the residues 32 and 63 in Fig. 5a, and the residues 60, 90, 106, and 138 in Fig. 5b. This result indicates that the predicted N_{14} location profile has information on the proteolytic processing sites to yield mature peptides.

DISCUSSION

There are several approaches to predict the relative location of a residue (inside or outside) in a globular protein by means of hydrophobicity plot, N_8 contact profile, N_{14} location profile, and so on. As mentioned at the beginning, the quantity N_{14} of a residue has a good correlation with its relative location in a globular protein, and it is predictable from the amino acid sequence with a good accuracy[2]. Therefore, it seems to be fruitful to use a location profile for the prediction of a native protein conformation from its amino acid sequence. The experimental location profile shows wave-like patterns along the amino acid sequence, and this pattern could be identified by smoothing the profile. The Fourier transformation technique used to smooth the profile is different from those usually employed[5,17,18], where smoothing is done by taking the moving average of n residues. The advantage of our smoothing is that it can take the effect of local sequences into account whereas the simple average method cannot. The polypeptide chain was divided into segments at minima of the smoothed location profile. The sites of minima correspond to turn conformation, so that a segment can be regarded as a turn-to-turn element in a globular protein.

The segmentation was analyzed using N_r profiles such as N_8, N_{14}, N_{20}, and N_{26}, where N_8 and N_{14} are the contact and location number, respectively. Both the number and sites of minima are almost independent of the radius of sphere as shown in Fig. 1, implying that the segmentation cannot be arbitrary. The feature of dividing positions into segments was examined by using the N_{term14} profile and C_{term14}. The degrees of coincidence of minima between two profiles are in the range from 41% to 89% according to the type of c-segments, and the average coincidence of 87 proteins is 69%. The coincidence occurring by chance was evaluated to be 43%

from random sequences[6]. Therefore, conformations of the N- and C-terminal parts are correlated with each other through a cutting point; for example some c-segments such as group 5 and 2 show a high coincidence. Interestingly, group 7 is the group of the lowest coincidence, and c-segments in this group are frequently found at helix-helix junctions.

In order to characterize the segments from a corresponding location profile, angular moments of $M(40°)$ and $M(100°)$, relative location and length were used. Angular moments of $M(40°)$ and $M(100°)$ are correlated with the preference of β-structure and α-helix, respectively. The 1240 segments were classified into 10 groups according to their distribution. The classification method used here is dependent on the choice of a distance and a cut-off value for grouping. The criterion to determine the distance was that the largest group includes about one-fifth of the total segments. When we searched the center of the local high density by varying the distance, most of the centers were invariant. The 10 groups can be described by their representative segments or the average contents of secondary structures. As shown in Table II, however, the character of a segment cannot be expressed simply by secondary structures. In other words, even if the secondary structures are similar, segments are distinguished in other aspects like group A1 and A2 segments. In order to obtain a clear grouping for occurrence frequencies, 10 groups of segments were classified into three types such as α-helical, β-structual and coil segments according to the contents of secondary structures.

The 10 groups have characteristic amino acid compositions, especially the compositions of group A1 were different from those of group B1. Large differences were found in the following five amino acid residues in their occurrences shown in group A1 and B1 in parenthesis, respectively; these are expressed in %: Glu(7.0, 2.7), Gly(3.8, 12.2), Lys(8.8, 3.3), Ser(4.5, 9.4) and Val(6.8, 10.8). This is consistent with the occurrence frequencies of amino acid residues in α-helix and β-structure[19].

In this work, the experimental location profiles from various 87 proteins were used to analyze the protein conformation in terms of segments. As a result, protein structure can be described by characterized segments and c-segments at the junction of adjacent segments. The present study shows that one-dimensional experimental location profile has sufficient information on segments characterization. Our final aim is to obtain a similar representation of a protein in terms of characterized segments from an amino acid sequence alone. As the assignment of the cutting positions seems to be quite satisfactory, the next step is to assign a sequence into a characterized group.

An attempt to assign segments into characterized groups was made by comparing sequences using a homology search method[14]. First, 11 triplets of B3-C1-B3 and segment pairs of B3-C1 and C1-B3 observed frequently[6] were chosen as target sequences of three and two successive segments in the 87 proteins, respectively. Then, one of the sequences of triplets of B3-C1-B3 was taken as a reference and this was examined against all the sequences in the 87 proteins, but we could not find any homologous triplets of B3-C1-B3 by this method. Therefore, this method would not be sufficient enough to assign a sequence into characterized segments.

A cleavage site of a signal peptide corresponds to the first minimum of the computed location profile. The proteins which have a signal peptide generally showed a high peak at the N-terminal region as shown in Fig. 5. However, cytoplasmic proteins or the mature proteins (after the removal of signal peplide) except membrane proteins do not show such a high peak at the N-terminal region. When the hydropathy plot[18] was used to assign a cleavage site of signal peptide for 201 proteins similarly, the degrees of coincidence within ± 3 and ± 4 residues were 42% and 54%, respectively. Thus, location profiles have a higher correlation with a protein structure than other profiles.

ACKNOWLEDGEMENT

The author thanks to Professor T. Ooi of Kyoto University for reading this manuscript and valuable comment.

REFERENCES

(1) C.B. Anfinsen and H.A. Sheraga, *Adv. Protein Chem.*, **29**, 205 (1975).
(2) K. Nishikawa and T. Ooi, *J. Biochem.*, **100**, 1043 (1986).
(3) G.D. Rose and J.P. Seltzer, *J. Mol. Biol.*, **113**, 153 (1977).
(4) G.M. Crippen, *J. Mol. Biol.*, **126**, 315 (1978).
(5) G.D. Rose, *Nature*, **272**, 586 (1978).
(6) H. Nakashima, K. Nishikawa and T. Ooi, *J. Protein Chem.*, **7**, 509 (1988).
(7) F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, Jr., M.D. Brice, J. Rodgers, O. Kennard, T. Simanouchi and M. Tasumi, *J. Mol. Biol.*, **112**, 535 (1977).
(8) W.C. Baker, L.T. Hunt, D.G. George, L.S. Yeh, H.R. Chen, M.C. Blomquist., E.I. Seibel-Ross, A. Elzanowski, J.K. Bair, D.A. Ferrick, M.K. Hong, R.S. Lekley, *Protein Identification Resources* (PIR), National Biomedical Research Foundation, Washington, D.C. (1987).
(9) K. Nishikawa, Y. Kubota and T. Ooi, *J. Biochem.*, **94**, 981 (1983).
(10) H. Nakashima, K. Nishikawa and T. Ooi, *J. Biochem.*, **99**, 153 (1986).
(11) K. Nishikawa and T. Ooi, *Int. J. Peptide Protein Res.* **16**, 19 (1980).
(12) K. Nishikawa, H. Nakashima, M. Kanehisa and T. Ooi, *Protein Seq. Data Anal.*, **1**, 107 (1987).
(13) D. Bashford, C. Chothia and A.M. Lesk, *J. Mol. Biol.*, **196**, 199 (1987).
(14) W.B. Goad and M. Kanehisa, *Nucleic Acids Res.*, **10**, 247 (1982).
(15) G. von Heijne, *Nucleic Acids Res.*, **14**, 4683 (1986).
(16) R.J. Folz and J.I. Gordon, *Biochem. Biophys. Res. Commun.*, **146**, 870 (1987).
(17) T.P. Hopp and K.R. Woods, *Proc. Natl. Acad. Sci. USA*, **78**, 3824 (1981).
(18) J. Kyte and R.F. Doolittle, *J. Mol. Biol.*, **157**, 105 (1982).
(19) P.Y. Chou and G.D. Fasman, *Annu. Rev. Biochem.*, **47**, 251 (1978).