The folding type of a protein is relevant to the amino acid composition

The Folding Type of a Protein Is Relevant to the Amino Acid Composition

Hiroshi NAKASHIMA,[*] Ken NISHIKAWA,[**] and
Tatsuo OOI[**]

[*]School of Allied Medical Professions, Kanazawa University,
Kanazawa, Ishikawa 920, and [**]Institute for Chemical Research,
Kyoto University, Gokasho, Uji, Kyoto 611

Received for publication, July 20, 1985

The folding types of 135 proteins, the three-dimensional structures of which are known, were analyzed in terms of the amino acid composition. The amino acid composition of a protein was expressed as a point in a multidimensional space spanned with 20 axes, on which the corresponding contents of 20 amino acids in the protein were represented. The distribution pattern of proteins in this composition space was examined in relation to five folding types, α, β, α/β, $\alpha+\beta$, and irregular type. The results show that amino acid compositions of the α, β, and α/β types are located in different regions in the composition space, thus allowing distinct separation of proteins depending on the folding types. The points representing proteins of the $\alpha+\beta$ and irregular types, however, are widely scattered in the space, and the existing regions overlap with those of the other folding types. A simple method of utilizing the "distance" in the space was found to be convenient for classification of proteins into the five folding types. The assignment of the folding type with this method gave an accuracy of 70% in the coincidence with the experimental data.

Previous analyses of amino acid composition data have shown that the amino acid composition of a protein contains information about protein character: *i.e.*, the composition has significant correlations to its biological characters: its location (inside or outside the cell), its function (enzyme or nonenzymes), it structural features, and the presence of disulfide bonds (*1–3*). In the present study, we focus our attention on the folding types of proteins and analyze the relation of folding type to amino acid composition. We use the same method as before of the "composition space" to represent an amino acid composition; *i.e.*, the contents of 20 amino acids are plotted on the corresponding orthogonal axes. The dimensionality of the composition space is twenty (in the previous studies an 18-dimensional space was used). We gathered as many proteins of known sequence and three-dimensional structure as possible by surveying the literature. The previous finding that the amino acid composition has a fairly strong correlation to the folding type (*1*, *3*) was reascertained with this extended data set. We will show here that the folding type of a protein is "predictable" in a statistical sense from its amino acid composition.

METHODS

Definition of the Composition Space—The method to express the composition of a protein as one point in the 20-dimensional orthogonal coordinate system is to take every amino acid content on the corresponding axis. We introduce a normalized scale in order to adjust the variance of amino acid contents along an axis and make the contribution of each amino acid equivalent. A normalized composition of a protein j is calculated as follows;

$$a_{ij}=(c_{ij}-\bar{c}_i)/\sigma_i \qquad (1)$$

where a_{ij} is the normalized content, c_{ij} is the real content (%) of amino acid i in the protein, and $\bar{c}_i$ and σ_i are the average composition and standard deviation (S.D.) for amino acid i, respectively. The average composition of all the proteins was taken as the origin of the coordinate system, as described in the next section. The distance d_{jk} between two points in the space for two proteins, j and k, is defined in terms of their normalized compositions in a form

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

and d_{jk} calculated from Eq. 2 is expressed in d.u. (distance units). A distribution of proteins in the composition space was analyzed in terms of the distance between two points (*i.e.*, the difference in the respective amino acid compositions) and radius (a root mean squares of distances) from the center of a given cluster.

Protein Data—The definition of our composition space depends on the average composition and the standard deviations as given in Eq. 1. In order to construct it on a solid basis, $\bar{c}_i$ and σ_i was calculated using data on 569 protein sequences which were selected by eliminating incomplete, short, and closely related sequences from our databank containing 3,010 proteins sequences. The σ_i values of some residues have been adjusted by omitting proteins which have contents that deviate greatly from the mean (*i.e.*, more than ±3.5 S.D. unit). Since we could determine $\bar{c}_i$ and σ_i in this way, the composition of any protein is represented in the space.

We collected 135 proteins of known three-dimensional structure for the analysis of the folding types and composition data. To avoid an artifact arising from inclusion of a number of homologous proteins, proteins were selected on the basis that they have more than 50 residues and differ by at least 50% in their sequences. The names of the selected proteins are listed in Table I. (In the table, proteins of the same family, *e.g.*, cytochrome, are found, but the sequences differ by more than 50% as described above.)

Criteria of Assigning the Folding Type to a Protein—Proteins of known structure are usually classified into one of five folding types, α type, β type, α/β type, $\alpha+\beta$ type, and irregular type, as done by Levitt and Chothia (*4*). We, however, encountered difficulty and ambiguity in assigning a unique type for some proteins. Thus, we introduced a quantitative measure to define a folding types by contents of α-helix (α) and β-structure (β) as follows: α type proteins, $\alpha>15\%$ and $\beta<10\%$; β type proteins, $\alpha<15\%$ and $\beta>10\%$; α/β type proteins, $\alpha>15\%$ and $\beta>10\%$ with dominantly parallel β-sheets; $\alpha+\beta$ type proteins, $\alpha>15\%$ and $\beta>10\%$ with dominantly antiparallel β-sheets; and irregular type proteins, $\alpha<15\%$ and $\beta<10\%$. In this way, the 135 proteins divided into 31 α type, 34 β type, 39 α/β type, 27 $\alpha+\beta$ type, and 4 irregular type proteins. The results classified by this method are shown in Table I. The major analyses on these 135 proteins were determined irrespective of domains which are found experimentally: *i.e.*, the folding types were determined automatically by the experimental contents of secondary structures described in the data of the Protein Data Bank (*5*), on the basis of the criteria described above. When the domains, the folding types of which are known, are treated as independent data, we have 202 data for the analysis. This data set was used for the test of our results.

RESULTS

Characterization of the Five Folding Groups in the Composition Space—The average compositions obtained by classification of the proteins into five folding groups are listed in Table II. There are marked differences in amino acid contents between the compositions of α type and β type proteins. In order to demonstrate the differences between

TABLE I. A list of the 135 structurally known proteins used in this work. The number of residues of a protein is given in the column " No. aa." The next column (A) gives the folding type of proteins predicted from the amino acid composition, where the irregular type is represented with " R." The column " PDB" indicates the identification code of the Protein Data Bank entries. Reference numbers headed by @ refer to the entry number in the table of 356 proteins in Ref. *2*.

No.	Protein name	Source	No. aa	A	PDB	Ref.
α proteins						
1	Apoferritin	Horse spleen	175	α	0AF1	@26
2	Bacteriorhodopsin	*Halobacterium halobium*	248	α		12
3	Calcium-binding parvalbumin B	Carp	108	α	1CPV	@55
4	Calcium-binding protein	Bovine	75	α	1ICB	
5	Citrate synthase	Pig heart	437	α/β	1CTS	
6	Coat protein	Tobacco mosaic virus	158	β	0TMV	@77
7	Complement C3a anaphylatoxin	Human serum	77	$\alpha+\beta$	0C3A	@83
8	Cytochrome b_{562}	*Escherichia coli*	103	α	156B	@92
9	Cytochrome c	Albacore tuna	103	α	3CYT	@93
10	Cytochrome c_2	*Rhodospirillum rubrum*	112	α	2C2C	@94
11	Cytochrome c_3	*Desulfovibrio desulfuricans*	116	α	0CY3	@95
12	Cytochrome c_3	*Desulfovibrio vulgaris*	107	α	2CDV	
13	Cytochrome c_5	*Azotobacter vinelandii*	83	α	1CC5	
14	Cytochrome c_{550}	*Paracoccus denitrificans*	121	α	155C	@96
15	Cytochrome c_{551}	*Pseudomonas aeruginosa*	82	α	351C	@97
16	Cytochrome c_{555}	*Chlorobium thiosulfatophilum*	86	α	05C1	
17	Cytochrome c'	*Rhodospirillum molischianum*	128	α	1CCY	
18	Cytochrome c peroxidase	Baker's yeast	294	α/β	1CYP	@99
19	Erythrocruorin	*Chironomus thummi thummi*	136	α	1ECD	@114
20	Globin	*Glycera dibranchiata*	147	α	0HBG	@136
21	Hemerythrin	*Themiste dyscritum*	113	α	1HMQ	@156
22	Hemoglobin α chain	Horse erythrocyte	141	α	2MHB	@158
23	Hemoglobin β chain	Horse erythrocyte	146	α	2MHB	@159
24	Hemoglobin	Sea lamprey	148	α	1LHB	@160
25	Leghemoglobin	Yellow lupin	153	α	1LH1	@203
26	Myoglobin	Sperm whale	153	α	2MBN	@227
27	Myoglobin	*Aplysia limacina*	145	α	0MBA	
28	Myohemerythrin	*Themiste zostericola*	118	α	1MHR	@228
29	λ Repressor, N-terminal domain	Bacteriophage λ	92	α	0LRP	@284
30	Tropomyosin, α	Rabbit skeletal muscle	284	α		@335
31	Uteroglobin	Rabbit uterus	70	α	0UTG	@354
β proteins						
32	Acid proteinase	*Endothia parasitica*	318	β	2APE	
33	Acid proteinase	*Rhizopus chinensis*	324	β	1APR	@285
34	Actinoxanthin	*Actinomyces globisporus*	108	β	1ACX	
35	Azurin	*Pseudomonas fluorescens*	128	α	1AZU	@43
36	Chymotrypsin A	Bovine pancreas	245	β	2CHA	@71
37	Coat protein	Satellite tobacco necrosis virus	195	β	2STV	
38	Coat protein	Southern bean mosaic virus	260	β	3SBV	
39	Coat protein	Tomato bushy stunt virus	387	β	2TBV	
40	Concanavalin A	Jack bean	237	α/β	3CNA	@84
41	γ Crystallin II	Bovine lens	174	$\alpha+\beta$	0GCR	@88
42	DNA unwinding protein	Bacteriophage fd gene 5	88	β	0GN5	@105
43	Elastase	Pig pancreas	240	β	1EST	@107
44	Group-specific protease	Rat small intestine	224	α/β	3RP2	@152
45	Hemagglutinin, HA1 chain	Influenza virus A Hong Kong 68	293	β	0HMG	

TABLE I (Continued)

No.	Protein name	Source	No. aa	A	PDB	Ref.
46	High potential iron-sulfur protein	*Chromatium vinosum*	85	α	1HIP	@162
47	Immunoglobulin κ Bence-Jones REI	Human patient REI	107	β	1REI	@52
48	Immunoglobulin λ Fab. L-chain	Human patient NEW	207	β	3FAB	@179
49	Immunoglobulin λ Fab. H-chain	Human patient NEW	219	β	3FAB	@180
50	Immunoglobulin G Fc fragment	Human serum	224	β	1FC1	
51	Kallikrein A	Pig pancreas	232	α/β	2PKA	
52	Long neurotoxin (α cobratoxin)	Cobra (*naja naja siamlnsis*)	71	α+β	1CTX	@237
53	α-Lytic protease	Myxobacter 495	198	β	1ALP	@231
54	Neuraminidase	Influlenza virus, Tokyo 3 67	469	β		13
55	Neurotoxin B (Erabutoxin B)	Broad-banded sea snake	62	β	1NXB	@238
56	Penicillopepsin	*Penicillium janthinellum*	323	β	2APP	@250
57	Pepsin	Pig stomach mucosa	326	β	1PEP	@251
58	Plastocyanin	Poplar leaves	99	α	1PCY	@266
59	Prealbumin	Human plasma	127	α/β	2PAB	@267
60	Proteinase B (SGPB)	*Streptomyces griseus*	185	β	3SGB	@277
61	Rubredoxin	*Desulfovibrio vulgaris*	52	α+β	3RXN	
62	Scorpion venom	*Centrupoides sculpturatus*	65	α+β	1SN3	
63	Superoxide dismutase, Cu-Zn	Bovine erythrocyte	151	β	2SOD	@311
64	Trypsin β	Bovine pancreas	223	β	3PTP	@339
65	Trypsin inhibitor (Kunitz)	Soy bean	181	α/β		14
α/β proteins						
66	Adenylate kinase	Pig muscle	194	α/β	2ADK	@8
67	D-Alanyl-D-alanine peptidase	*Streptomyces albus*	212	β	0ZGP	
68	Alcohol dehydrogenase	Horse liver	374	α/β	4ADH	@15
69	Aldolase	*Pseudomonas putida*	225	α/β	1KGA	@262
70	Alkaline phosphatase	*Escherichia coli*	449	α/β		15
71	L-Arabinose binding protein	*Escherichia coli*	306	α/β	1ABP	@29
72	Aspartate aminotransferase	Chicken heart	411	α/β	1AAT	
73	Aspartate carbamoyltransferase C-chain	*Escherichia coli*	310	α/β	2ATC	@36
74	Aspartate carbamoyltransferase R-chain	*Escherichia coli*	152	α/β	2ATC	@37
75	Carboxypeptidase A	Bovine pancreas	307	α/β	5CPA	@60
76	Carboxypeptidase B	Bovine pancreas	306	α/β	1CPB	@61
77	Catalase	Bovine liver	506	α/β	3CAT	@65
78	Dihydrofolate reductase	*Lactobacillus casei*	162	α/β	3DFR	@321
79	Dihydrofolate reductase	*Escherichia coli*	159	α/β	4DFR	
80	Elongation factor Tu	*Escherichia coli*	393	α/β	0ETU	@108
81	Flavodoxin	Clostridium MP	138	α	3FXN	@124
82	Flavodoxin	*Desulfovibrio vulgaris*	148	α+β	0FX1	
83	D-Galactose binding protein	*Salmonella typhimurium*	309	α/β	1GBP	
84	Glutathione peroxidase	Bovine erythrocyte	178	β	0GP1	@145
85	Glutathione reductase	Human erythrocyte	478	α/β	2GRS	@146
86	Glyceraldehyde-phosphate dehydrogenase	Lobster	333	α/β	2GPD	@148
87	Glycogen phosphorylase A	Rabbit skeletal muscle	841	α/β	0PPA	@150
88	*p*-Hydroxybenzoate hydroxylase	*Pseudomonas fluorescens*	394	α/β	0PHH	
89	Lactate dehydrogenase C4	Mouse testis	329	α/β	1LDX	@196
90	Lactate dehydrogenase M4	Dogfish muscle	329	α/β	4LDH	@197
91	Malate dehydrogenase (cytoplasmic)	Pig heart	314	α/β	2MDH	@218
92	Methionyl-tRNA synthetase	*Escherichia coli*	581	α/β	0MTS	
93	Phosphofructokinase	*Bacillus stearothermophilus*	316	α/β	0PFK	@255
94	6-Phosphogluconate dehydrogenase	Sheep liver	466	α/β		16
95	Phosphoglycerate kinase	Horse muscle	416	α/β		17

TABLE I (Continued)

No.	Protein name	Source	No. aa	A	PDB	Ref.
96	Phosphoglycerate kinase	Baker's yeast	415	α/β	3PGK	
97	Phosphoglyceromutase	Baker's yeast	241	α/β	3PGM	@261
98	Rhodanese	Bovine liver	293	α/β	1RHD	@324
99	Subtilisin BPN′	*Bacillus amyloliquefaciens*	275	β	1SBT	@310
100	Taka-amylase	*Aspergillus oryzae*	478	α/β	2TAA	@316
101	Thioredoxin	*Escherichia coli*	108	α/β	1SRX	@322
102	Thioredoxin	Bacteriophage T4	87	α	0TT4	@323
103	Triose phosphate isomerase	Chicken muscle	247	α/β	1T1M	@334
104	Tyrosyl-tRNA synthetase	*Bacillus stearothermophilus*	419	α/β	1TS1	@348
$\alpha+\beta$ proteins						
105	Actinidin	Chinese gooseberry (kiwifruit)	220	β	2ACT	@5
106	α-1-Antitrypsin	Human plasma	383	α/β	6API	
107	Bacteriochlorophyll A protein	*Chlorobium limicola*	358	α/β	2BCL	@44
108	Carbonic anhydrase B	Human erythrocyte	260	α/β	2CAB	@59
109	Catabolite gene activator protein	*Escherichia coli*	209	α/β	0GAP	
110	Cytochrome b_5	Bovine liver	93	α	2B5C	@91
111	Deoxyribonuclease I	Bovine pancreas	257	α/β		18
112	DNA binding protein II	*Bacillus stearothermophilus*	90	α		19
113	Hemagglutinin, HA2 chain	Influenza virus A Hong Kong 68	293	α/β	0HMG	
114	Inorganic pyrophosphatase	Baker's yeast	285	α/β	1PYP	@182
115	Lysozyme	Chicken egg white	129	$\alpha+\beta$	2LYZ	@212
116	Lysozyme	Bacteriophage T4	164	α/β	1LZM	@213
117	Ovomucoid, third domain	Japanese quail	56	$\alpha+\beta$	1OVO	@244
118	Pancreatic secretory trypsin inhibitor	Pig pancreas	56	$\alpha+\beta$	1TGS	
119	Papain	Papaya	212	$\alpha+\beta$	8PAP	@247
120	Phospholipase A_2	Porcine pancreas	124	$\alpha+\beta$	1P2P	@273
121	*cro* Repressor	Bacteriophage λ	66	α	0CRO	
122	Ribonuclease A	Bovine pancreas	124	$\alpha+\beta$	4RSA	@287
123	Ribonuclease (Barnase)	*Bacillus amyloliquefaciens*	110	β	0RNB	
124	Ribonuclease ST	*Streptomyces erythreus*	101	$\alpha+\beta$	0RST	
125	Ribonuclease T_1	*Aspergillus oryzae*	104	$\alpha+\beta$	0RNT	@288
126	Ribosomal protein L7/L12	*Escherichia coli*	120	α	0CTF	@293
127	Staphylococcal nuclease	*Staphylococcus aureus*	149	α/β	2SNS	@221
128	Streptomyces subtilisin inhibitor	*Streptomyces albogriseolus*	113	β	2SSI	@309
129	Superoxide dismutase, Mn	*Escherichia coli*	205	α/β	0SDE	@312
130	Thermolysin	*Bacillus thermoproteolyticus*	316	$\alpha+\beta$	3TLN	@54
131	Trypsin inhibitor, BPTI	Bovine pancreas	58	$\alpha+\beta$	4PTI	@341
Irregular type proteins						
132	Agglutinin isolectin 2	Wheat germ	171	R	2WGA	20
133	Ferredoxin, bacterial-type	*Peptococcus aerogenes*	54	R	1FDX	@117
134	Ferredoxin, chloroplast-type	*Spirulina platensis*	98	$\alpha+\beta$	3FXC	@118
135	Ferredoxin	*Azotobacter vinelandii*	106	$\alpha+\beta$	2FD1	

them, deviations from the corresponding average content for every amino acid normalized with the S.D. values were calculated for α type and β type proteins. As shown in Fig. 1, deviations obtained on the α type proteins are in the opposite direction to those on the β type proteins for most of amino acids; exceptional amino acids are Arg, Ile, and Cys. Positive deviations of Lys, Ala, His, Asp, Met, Leu, Phe, Glu, and Cys are observed in the α type proteins, while Ser, Cys, Thr, Gly, Pro, Trp, Asn, Val, Tyr, and Gln are amino acids of positive deviations for the β type proteins. A

TABLE II. The average amino acid composition (AV) and standard deviation (S.D.) of the total proteins, and the average composition of the five folding types, α, β, α/β, $\alpha+\beta$, and the irregular, R (all the data are in molar percent, %). In the two sets of values given for each folding type (except for R), the former one is the average over the proteins included in a given type, and the latter is the corresponding one obtained by optimization so as to yield the best prediction of the folding type (see the text). The total average and S.D. were determined by using data for 569 proteins (see "METHODS").

Amino acid	AV	S.D.	α		β		α/β		$\alpha+\beta$		R
Ala	8.74	3.67	11.63	11.67	7.54	8.29	8.83	9.19	8.89	9.12	8.90
Cys	1.62	1.53	1.71	1.81	3.48	3.51	1.43	1.77	2.94	3.97	12.04
Asp	5.72	2.20	6.52	6.47	5.37	4.98	6.12	6.59	5.76	4.05	8.85
Glu	6.39	2.88	6.52	6.56	3.75	3.81	6.12	6.61	6.18	5.12	6.85
Phe	3.87	1.85	4.22	4.30	3.57	3.67	3.88	3.38	3.60	3.01	1.73
Gly	7.82	2.98	7.66	7.05	9.87	9.65	8.71	8.91	8.00	7.99	10.49
His	2.15	1.32	2.79	2.59	1.64	2.04	2.19	2.27	2.00	1.25	1.02
Ile	5.15	2.29	3.72	3.72	4.76	4.63	5.82	5.71	4.74	4.85	6.99
Lys	6.78	3.34	10.10	10.00	4.66	4.82	6.55	6.75	7.18	5.95	3.27
Leu	8.20	3.16	8.89	9.18	6.69	6.36	8.54	8.09	6.37	5.21	4.02
Met	2.08	1.26	2.42	2.15	1.24	1.47	2.14	2.07	1.40	0.62	0.53
Asn	4.39	1.99	3.79	3.99	4.90	4.65	4.13	4.35	5.60	7.31	4.16
Pro	4.49	2.04	3.81	3.79	5.23	6.03	4.36	4.21	4.29	4.88	5.82
Gln	3.91	1.74	3.33	3.48	4.12	3.99	3.44	3.46	3.17	3.33	4.03
Arg	4.81	2.53	2.79	3.12	3.22	3.98	4.35	4.08	4.05	3.14	1.08
Ser	6.56	2.73	5.44	6.17	9.50	9.14	5.89	5.34	7.05	8.04	6.42
Thr	5.84	2.30	4.91	4.41	7.83	7.42	5.50	4.52	6.41	8.26	4.35
Val	7.01	2.48	6.02	5.59	7.48	6.70	7.62	7.77	6.50	6.54	4.89
Trp	1.17	0.98	1.17	1.30	1.48	1.35	1.38	1.37	1.28	1.22	0.62
Tyr	3.33	1.87	2.55	2.64	3.67	3.41	3.02	2.66	4.59	6.14	3.95

similar tendency is seen for amino acids with negative deviations (Fig. 1). Apparently, an amino acid composition of an α type protein is different from that of a β type protein.

The distribution of points representing proteins of one folding type in the composition space may be characterized by the location of a center (*i.e.*, an average composition) and the root-mean squares (rms) radius of the distribution for the folding type from the center. The distance to the origin and the rms radius (both in d.u.) are (2.04, 5.26), (2.62, 4.95), (0.77, 3.38), (1.65, 5.08), and (9.69, 6.47) for proteins of α type, β type, α/β type, $\alpha+\beta$ type, and irregular type, respectively. The distribution of the α/β type proteins is narrowest, having its center near the origin as indicated by a short distance to the origin (0.77 d.u) and a small rms radius (3.38 d.u.). Compositions of irregular-type proteins, though the number of data is small, locate far away from the origin (9.69 d.u.).

In order to illustrate the distribution, points in the composition space were projected onto a plane which is defined by two vectors, one pointing from the average composition of the $\alpha+\beta$ type proteins to that of the α type proteins (vector A) and the other from the $\alpha+\beta$ type to the β type (vector B). The projection on the plane was done in a way similar to the previous method (*1*, *3*). As the angle between the A and B vectors was 107.6°, the X axis was taken in the direction of the A vector from the origin and the Y axis was chosen so as to be perpendicular to the X axis in the plane defined with the vectors A and B both pass-

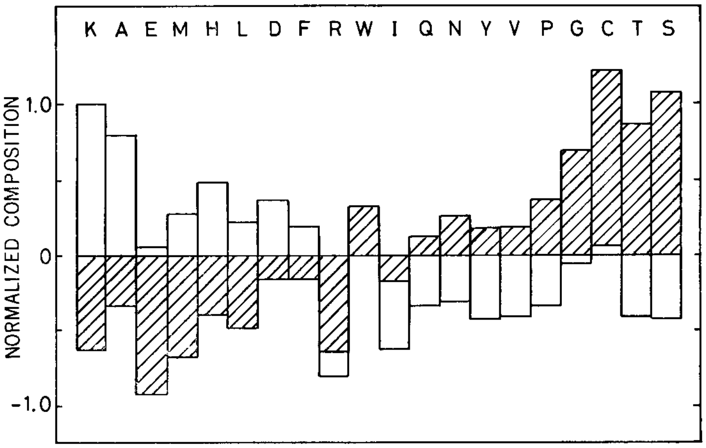

Fig. 1. Comparison of the average composition between the α and β type proteins. The normalized composition is used in accordance with Eq. 1, so that the value plotted indicates the deviation from the total average. Open columns are used for the α type and shaded columns for the β type. Amino acids are denoted by the one-letter code.

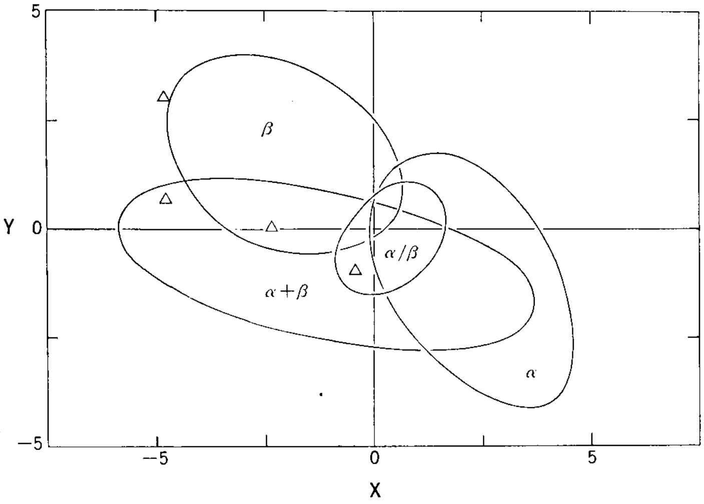

Fig. 2. Schematic drawing of the distribution pattern of the four (α, β, α/β, $\alpha+\beta$) types of proteins projected onto a plane. The X axis represents the α-helical character and Y approximately represents β-sheet character (see the text). The origin was taken at the total average composition. The four irregular-type proteins are plotted with the symbol △. There are some exceptional proteins which fall outside the individual regions indicated. The number of such proteins are: 2 out of 31 proteins for the type α, 3 out of 34 for β, 7 out of 39 for α/β, and 3 out of 27 for $\alpha+\beta$.

ing through the origin. Thus, the X axis represents the α-helical character and the Y axis approximately the β-sheet character (17.6° from the B vector).

Figure 2 shows the resulting distribution of proteins schematically, indicating that the amino acid compositions of proteins which belong to the same folding type locate in a specified region of the space. Proteins of the irregular type seem to scatter in the extended direction of the $\alpha+\beta$ and β regions (the points are shown in Fig. 2). The regions of α and β types appear almost completely separated from each other, while the α/β and $\alpha+\beta$ regions seem to overlap with the other regions. The actual overlap in the composition space is, however, not so serious as expected from the appearance, since Fig. 2 is just a two-dimensional representation of the distribution. The degree of overlap (or separation) quantitatively shown below demonstrates that distinction among α, β, and α/β is good in practice.

Classification of Proteins into Folding Groups by Their Compositions—Since points representing proteins of the same folding type cluster in the composition space as shown in Fig. 2, it is possible to classify proteins into five folding types according to their amino acid compositions. There are two ways to assign the folding type of a protein according to its amino acid composition. One method is as follows: proteins whose composition is located beyond a distance greater than 11 d.u. from the origin are assigned as the irregular type. A protein is assigned as an α/β type one when its composition locates smaller than 3.91 d.u. from the center of α/β type proteins. The folding type of a protein which does not meet the above conditions is determined by the criterion of the shortest distance to the centers of the α, β, and $\alpha+\beta$ types, respectively. In this way, folding types of proteins are automatically assigned in terms of distances measured in the composition space. The theoretical folding types are shown in column A in Table I and the comparison of this classification with that based on secondary structures determined experimentally from X-ray analysis is summarized in Table III(a). The numbers on the diagonal represent proteins for which the theoretical analysis coincides with experimental data. The score of the agreement is 70% (94/135), and this is equivalent to a measure of separation of the different types in the composition space. The worst distinction is seen between the $\alpha+\beta$ and α/β types. When these two types were unified and the classification was made into the three types of α, β, and "$\alpha\beta$ complex" (excluding irregular type), the score of coincidence increased to 80% (105/131).

The other method is to use the criterion of the shortest distance to a group center; *i.e.*, the distance from the origin is first computed, and then the distances to the four centers of the folding types are compared. When the distance from the origin is greater than 11 d.u., the protein is assigned as the irregular type as done before. If not, the closest folding type is selected as that for the protein. Since only distances are used for the criterion, the latter method is simpler than the former method. Since we don't have so much data for proteins whose folding types is known, we have to optimize the location of the "centers" of corresponding folding types. Numbers shown in parentheses of Table III(a) are the results obtained by the latter method using optimized aver-

TABLE III. Comparison of the two classifications of folding types, one based on the amino acid composition (column) and the other from X-ray observation (row). (a) Comparison for 135 proteins. The numbers in parentheses show the result when the optimized composition (Table II) is used as the "center" of each protein group. (b) Comparison for 202 domains. The optimized composition (Table II) is used as the "center" of each protein group.

	(a) Observed type					Total		(b) Observed type					Total
	α	β	α/β	$\alpha+\beta$	R			α	β	α/β	$\alpha+\beta$	R	
α	27 (28)	3 (2)	2 (0)	4 (7)	0 (0)	36 (37)	α	30	2	0	8	0	40
β	1 (1)	22 (26)	3 (4)	3 (3)	0 (0)	29 (34)	β	3	51	6	3	1	64
α/β	2 (2)	5 (4)	33 (33)	10 (5)	0 (1)	50 (45)	α/β	6	7	49	10	1	73
$\alpha+\beta$	1 (0)	4 (2)	1 (2)	10 (12)	2 (1)	18 (17)	$\alpha+\beta$	2	5	2	13	1	23
R	0 (0)	0 (0)	0 (0)	0 (0)	2 (2)	2 (2)	R	0	0	0	0	2	2
Total	31	34	39	27	4	135	Total	41	65	57	34	5	202

age compositions of every folding type, which are listed in the second column of Table II. The score is (101/135), or 75%. This value is not surprising because we optimized the average compositions of each folding type to get a better score. The test of this method should be done for an expanded data set; *e.g.*, we have data on domains of several proteins listed in Table I, which are available for the test of the method. Table III(b) shows that the score is (145/202) or 71%. Therefore, we have a similar prediction score of about 70% by the use of both methods.

When we optimized the average compositions again for the expanded data set, we did not obtain any improvement on the score, indicating that the values in Table II are already at the optimum.

DISCUSSION

The present analyses indicate that proteins of the same folding type have amino acid compositions which locate within a specified region in the composition space. The four regions are well separated from each other, except for that of the $\alpha+\beta$ type (see above). In other words, proteins of every folding type are characterized by their amino acid compositions; the differences in amino acid contents of α type and β type proteins shown in Fig. 1 are examples. At a glance, one might feel that the relative abundance or scarcity of amino acids plotted in Fig. 1 is not consistent with the well-known propensity of amino acids to form α-helix or β-structure (*6–8*), particularly so for β-proteins. However, what is shown in Fig. 1 is the difference in the "whole" compositions between α type and β type proteins. If we look at the difference, *i.e.*, the total length of a column of each amino acid, and take account of the fact that β type proteins usually contain many more turns and irregular conformations than α proteins, the data shown in Fig. 1 would be attributed to the reflection of the properties of single amino acids.

This result has an important meaning: properties of single amino acid residues are, as the first approximation, additive regardless of the sequence. The nature of a protein as a whole is, therefore, approximately determined by its amino acid composition. Moreover, analysis using the composition space should be a useful tool for general analyses of proteins.

As described before, we introduced a quantitative criterion using the amount of α-helix and β-structure for classifying proteins from the X-ray data. The present assignment of folding types (Table I) is more objective, and therefore better than those in the previous study (the Table in the appendix of Ref. *2*), where a few proteins were assigned differently (*i.e.*, 18, 46, 126, 128, and 134 by the serial number in Table I).

Busetta and Barrans (*9*) have recently reported a prediction study of protein types, A, B, and AB, corresponding to our α, β and the unified type of α/β and $\alpha+\beta$. They employed the sequence data and made the prediction using a single quantity, which was derived from a predictive function developed for secondary structures of a protein (*10*). If the present assignment based on the X-ray data (Table I) is replaced by theirs and the comparison is made with their prediction, the score of coincidence becomes 79% (*i.e.*, 41/52), comparable with our result in the case of using the three folding types, α, β, and $\alpha\beta$ complex. This result again suggests that a considerable amount of information is obtained from the composition data.

Generally, it is difficult to make an unequivocal assignment of the folding type to a protein. The best way would be to divide a protein molecule into folding units, namely domains, as has been done by Richardson (*11*). When we know the domains in the three-dimensional structure of a protein, then we can examine the correlation between sequence and folding type, as shown in Table III(b). On the other hand, it is not possible to assign domains according only to the amino acid sequence at the present stage. Therefore, the problem of how to determine the domain arises.

Since the regions corresponding to each folding type are separated from each other in the composition space, it might be possible to get better separation if we use amino acid pairs instead of a single residue; *i.e.*, the 400-dimensional space might give better clustering of points which represent proteins of the same folding type. As far as we tried, we could not find any tendency toward clustering in the 400-dimensional space. Presumably, the meaning of a "sequence" appears already in doublet, triplet, and so on.

One of the authors (H.N.) expresses his thanks for the fellowship given by the Ministry of Education, Science and Culture of Japan for a leave from Kanazawa University to stay at the Institute for Chemical Research, Kyoto University, in 1984.

REFERENCES

1. Nishikawa, K. & Ooi, T. (1982) *J. Biochem.* **91**, 1821–1824
2. Nishikawa, K., Kubota, Y., & Ooi, T. (1983) *J. Biochem.* **94**, 981–995
3. Nishikawa, K., Kubota, Y., & Ooi, T. (1983) *J. Biochem.* **94**, 997–1007
4. Levitt, M. & Chothia, C. (1976) *Nature* **261**, 552–558
5. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977) *J. Mol. Biol.* **112**, 535–542
6. Chou, P.Y. & Fasman, G.D. (1978) *Annu. Rev. Biochem.* **47**, 251–276
7. Levitt, M. (1978) *Biochemistry* **17**, 4277–4285
8. Lifson, S. & Sander, C. (1979) *Nature* **282**, 109–111
9. Busetta, B. & Barrans, Y. (1984) *Biochim. Biophys. Acta* **790**, 117–124
10. Busetta, B. & Hospital, M. (1982) *Biochim. Biophys. Acta* **701**, 111–118
11. Richardson, J.S. (1981) *Adv. Protein Chem.* **35**, 167–339
12. Leifer, D. & Henderson, R. (1983) *J. Mol. Biol.* **163**, 451–466
13. Varghese, J.N., Laver, W.G., & Colman, P.M. (1983) *Nature* **303**, 35–44
14. McLachlan, A.D. (1979) *J. Mol. Biol.* **133**, 557–563
15. Sowadski, J.M., Foster, B.A., & Wyckoff, H.W. (1981) *J. Mol. Biol.* **150**, 245–272
16. Adams, M.J., Archibald, I.G., Helliwell, J.R., & Jenkins, S.E. (1978) *Acta Crystallogr.* **34**, S64
17. Banks, R.D., Blake, C.C.F., Evans, P.R., Haser, R., Rice, D.W., Hardy, G.W., Merrett, M., & Phillips, A.W. (1979) *Nature* **279**, 773–777
18. Suck, D., Oefner, Ch., & Kabsch, W. (1984) *EMBO J.* **3**, 2423–2430
19. Tanaka, I., Appelt, K., Dijk, J., White, S.W., & Wilson, K.S. (1984) *Nature* **310**, 376–381
20. Wright, C.S., Gavilanes, F., & Peterson, D.L. (1984) *Biochemistry* **23**, 280–287