Amino Acids (1994) 6: 231–246

Amino
Acids

Prediction of protein folding types from amino acid composition by correlation angles*

Review Article

K.-C. Chou

Computational Chemistry, Upjohn Research Laboratories, Kalamazoo, Michigan, U.S.A.

Accepted September 6, 1993

Summary. A protein is usually classified into one of the following four folding types: all-α, all-β, $\alpha + \beta$, and α/β. On the other hand, a protein can also be expressed as a vector of a 20-D space, in which its 20 components are defined by the composition of its 20 amino acids, respectively. Thus, the similarity between any two proteins in their amino acid composition can be quantitatively described in terms of their mutual projection in the 20-D space. The larger the projection value between the two vectors is, the higher their similar extent would be. Based on such a physical picture, a new method, the maximum projection method, has been proposed for predicting the folding type of a protein according to its amino acid composition. In comparison with the existing methods, the new method has the merits of yielding a higher rate of correct prediction, displaying a more intuitive physical picture, and being convenient in application.

Keywords: Amino acids – All α – All β – $\alpha + \beta$ – α/β – Amino acid composition – 20-D vector

Introduction

The prediction of the three-dimensional structure of a protein from its primary sequence is one of the most important, but also one of the most difficult problems in molecular biology. Despite years of both experimental and theoretical study, this problem is far from being solved yet. While empirical force fields have been developed in an attempt to predict the conformation of a protein by energy minimization (e.g., Scheraga, 1987), it is a formidable task to find the real global

* Part content of this paper was presented in the Third International Congress on Amino Acids held in Vienna.

minimum. The difficulty of this approach lies in the fact that the number of local minima often increases exponentially with the problem size, making a practical solution for the global minimum problem appear to be impossible. Therefore, the method based on energy calculations can so far only be successfully used to deal with regular structural elements in proteins (Chou et al., 1990), such as in explaining the trend of their handedness (Ooi et al., 1967; Chou and Scheraga, 1982), their packing arrangement (Chou et al., 1985), as well as the relevant folding features (Chou and Carlacci, 1991a). Although the simulated annealing approach is a powerful tool in overcoming the local minimum problem (Kawai et al., 1989; Wilson and Cui, 1990; Chou and Carlacci, 1991b), using it to predict the conformation of a protein merely based on its primary sequence is still impracticable. A feasible approach is a combination of the energy minimization and heuristic approach (Carlacci et al., 1991). However, there are not many proteins from which one can get enough heuristic inputs to result in a successful prediction.

A completely different approach to this problem is based on the statistical concept. The statistical method was established in an attempt to predict what secondary structure an amino acid segment would assume (Chou and Fasman, 1974; Garnier et al., 1978), and which structural type a protein would fold into (Chou, 1980; Nakashima et al., 1986; Klein, 1986; Klein and Delici, 1986; Chou, 1989; Zhang and Chou, 1992; Chou and Zhang, 1992). The prediction of protein folding type can help improve the prediction of protein secondary structure (Chou, 1989; Deléage and Roux, 1989), may reduce the scope of searching conformational space during energy optimization (e.g., see Chou, 1992), and provide useful information for a heuristic approach (Carlacci et al., 1991). In this review article, a new method, the "maximum projection method" or "least correlation angle method", is proposed to predict the folding type of a protein according to its amino acid composition.

Method

A. The four folding types

Proteins of known structures are usually classified into one of the following four folding types (Levitt and Chothia, 1976; Richardson and Richardson, 1989): all α-proteins (α type), all β-proteins (β type), $\alpha + \beta$ proteins ($\alpha + \beta$ type), and α/β proteins (α/β type). According to the classification made by Chou, P. Y. (1989) for the 64 known proteins of which 19 proteins are assigned as α type, 15 proteins as β type, 14 proteins as $\alpha + \beta$ type, and 16 proteins as α/β type, the quantitative criteria for the four folding types can be defined as follows:

(1) α type. Proteins of this type contain more than 45% α-helices, and less than 5% β-strands. Listed in Table 1 are the 19 α proteins and their amino acid composition.

(2) β type. Proteins of this type contain less than 5% α-helices, and more than 45% β-strands. Listed in Table 2 are the 15 β proteins and their amino acid composition.

(3) $\alpha + \beta$ type. Proteins of this type contain more than 30% α-helices and more than 20% β-strands with dominantly antiparallel β-strands. Listed in Table 3 are the 14 $\alpha + \beta$ proteins and their amino acid composition.

(4) α/β type. Proteins of this type contain more than 30% α-helices and more than 20% β-strands with dominantly parallel β-strands. Most of α/β type proteins are enzymes. Listed in Table 4 are the 16 α/β proteins and their amino acid composition.

Table 1. Amino acid composition in 19 α proteins (from P. Y. Chou, 1989)

α Proteins	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	Total
Calcium-binding parvalbumin (carp)	20	1	3	14	1	2	6	8	1	5	9	13	0	10	0	5	5	0	0	5	108
Cytochrome b_{562} (*E. coli*)	17	4	9	10	0	8	7	3	2	3	9	12	3	2	3	1	4	0	2	4	103
Cytochrome *c* (tuna)	7	2	6	4	2	4	5	13	2	4	6	16	2	3	3	4	7	2	5	6	103
Cytochrome c_2 (*R. rubrum*)	15	0	7	6	2	1	9	8	2	2	8	17	2	5	3	5	8	1	5	6	112
Cytochrome c_{550} (*P. denitrificans*)	15	1	8	10	2	6	10	17	1	5	6	17	4	4	6	3	8	1	3	7	134
Cytochrome c_{555} (*C. thiosulfatophilum*)	16	0	4	4	2	2	0	12	2	2	1	11	8	0	4	3	4	1	4	6	86
Hemerythrin B (*G. gouldii*)	6	3	7	12	1	3	6	7	6	9	8	11	1	9	4	3	4	4	5	4	113
Methemerythrin (*T. dyscritum*)	7	4	6	12	2	5	4	6	7	9	10	9	1	7	3	3	7	3	6	2	113
Myohemerythrin (*T. pyroides*)	7	2	5	9	2	1	11	6	6	6	7	15	3	7	5	4	5	3	5	9	118
α-Methemoglobin (horse)	16	3	4	9	1	1	3	10	10	0	21	11	1	7	6	13	9	1	3	12	141
β-Methemoglobin (horse)	15	4	6	8	1	4	10	14	9	0	19	11	1	8	5	6	3	2	3	17	146
α-Deoxyhemoglobin (human)	21	3	4	8	1	1	4	7	10	0	18	11	2	7	7	11	9	1	3	13	141
β-Deoxyhemoglobin (human)	15	3	6	7	2	3	8	13	9	0	18	11	1	8	7	5	7	2	3	18	146
γ-Deoxyhemoglobin (human fetal)	11	3	5	8	1	4	8	13	7	4	17	12	2	8	4	11	10	3	2	13	146
Hemoglobin (*glycera*)	28	3	4	8	1	6	4	20	6	8	11	11	5	3	3	10	1	2	3	10	147
Hemoglobin (lamprey)	21	5	4	10	1	4	6	6	2	8	10	13	4	8	6	13	9	2	4	12	148
Hemoglobin (midge larva)	17	3	5	9	0	4	5	11	4	9	6	10	4	14	5	9	9	1	2	9	136
Myoglobin (seal)	14	5	3	8	0	3	14	12	13	8	19	19	2	7	4	7	5	2	2	6	153
Myoglobin (sperm whale)	17	4	1	7	0	5	14	11	12	9	18	19	2	6	4	6	5	2	3	8	153
Total	285	53	97	163	22	67	134	197	111	91	221	249	48	123	82	122	119	33	63	167	2447

Table 2. Amino acid composition in 15 β proteins (from P. Y. Chou, 1989)

β Proteins	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	Total
α-Chymotrypsin (bovine)	22	3	13	9	10	10	5	23	2	10	19	14	2	6	9	27	22	8	4	23	241
Concanavalin A (jack bean)	19	6	12	20	0	5	7	16	6	15	18	12	2	11	11	31	19	4	7	16	237
Elastase (porcine)	17	12	18	6	8	15	4	25	6	10	18	3	2	3	7	22	19	7	11	27	240
Erabutoxin B (sea snake)	0	3	3	1	8	4	4	5	2	4	1	4	0	2	4	8	5	1	1	2	62
Immunoglobulin Fab′ (V_H and C_H) (human)	11	7	7	8	6	8	5	18	3	4	19	9	1	6	14	34	23	4	9	24	220
Immunoglobulin Fab′ (V_L and C_L) (human)	19	5	7	5	5	11	10	14	4	5	15	13	0	5	14	30	19	3	8	16	208
Immunoglobulin B-J MCG (human)	17	3	9	7	5	9	12	18	3	4	11	14	0	5	14	32	20	3	11	19	216
Immunoglobulin B-J REI (human)	6	3	2	5	2	13	2	8	0	8	8	4	1	3	6	14	11	1	8	3	108
Penicillopepsin (*P. janthinellum*)	24	0	18	19	2	25	4	40	3	13	21	5	0	20	12	47	29	3	14	24	323
Prealbumin (human)	12	4	3	5	1	2	10	10	4	5	7	8	1	5	8	11	12	2	5	12	127
Protease A (*S. griseus*)	19	7	13	3	5	6	2	31	3	9	10	0	1	5	4	21	21	1	8	12	181
Protease B (*S. griseus*)	14	8	10	7	4	2	2	33	2	7	7	1	2	5	5	22	28	2	10	14	185
Rubredoxin (*C. pasteurianum*)	0	0	1	10	4	0	6	6	0	2	1	4	1	2	5	0	3	1	3	5	54
Superoxide dismutase (bovine)	9	4	6	11	3	2	9	25	8	9	8	10	1	4	6	8	12	0	1	15	151
Trypsin (bovine)	14	2	17	5	12	10	4	25	3	15	14	14	2	3	8	34	10	4	10	17	223
Total	203	67	139	121	75	122	86	297	49	120	177	115	16	85	127	341	253	44	110	229	2776

Table 3. Amino acid composition in 14 $\alpha + \beta$ proteins (from P. Y. Chou, 1989)

$\alpha + \beta$ Proteins	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	Total
Actinidin (kiwi fruit)	18	5	11	16	7	10	10	28	1	17	8	6	2	5	7	12	18	6	14	17	218
Cytochrome b_5 (bovine)	4	3	2	7	0	2	12	6	5	5	8	9	0	3	3	8	7	1	4	4	93
Ferredoxin (*P. aerogenes*)	7	0	3	5	8	2	2	4	0	6	0	1	0	0	5	5	0	0	2	4	54
High-potential iron protein (chromatium)	19	2	5	5	4	5	4	6	1	2	5	5	1	2	5	3	4	3	1	3	85
Insulin (A and B chains) (porcine)	2	1	3	0	6	3	4	4	2	2	6	1	0	3	1	3	2	0	4	4	51
Lysozyme (bacteriophage T4)	15	13	12	10	2	5	8	11	1	10	16	13	5	5	3	6	11	3	6	9	164
Lysozyme (chicken)	12	11	13	8	8	3	2	12	1	6	8	6	2	3	2	10	7	6	3	6	129
Papain (papaya)	14	12	13	6	7	13	7	28	2	12	11	10	0	4	10	13	8	5	19	18	212
Phospholipase A_2 (bovine)	6	2	16	9	14	3	5	6	2	5	8	11	1	4	5	10	4	1	7	4	123
Ribonuclease S (bovine)	12	4	10	5	8	7	5	3	4	3	2	10	4	3	4	15	10	0	6	9	124
Staphylococcal nuclease (*S. aureus*)	14	5	7	7	0	6	12	10	4	5	11	23	4	3	6	5	10	1	7	9	149
Subtilisin inhibitor (streptomyces)	18	4	3	6	4	1	5	11	2	0	9	2	3	3	8	9	8	1	3	13	113
Thermolysin (*B. thermoproteolyticus*)	28	10	19	25	0	13	8	36	8	18	16	11	2	10	8	26	25	3	28	22	316
Trypsin inhibitor (bovine)	6	6	3	2	6	1	2	6	0	2	2	4	1	4	4	1	3	0	4	1	58
Total	175	78	120	111	74	74	86	171	33	93	110	112	25	52	71	126	117	30	108	123	1889

Table 4. Amino acid composition in 16 α/β proteins (from P. Y. Chou, 1989)

α/β Proteins	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	Total
Adenylate kinase (porcine)	8	11	2	11	2	6	19	19	2	9	18	21	6	5	6	11	14	0	7	17	194
Alcohol dehydrogenase (horse)	28	12	8	17	14	8	21	38	7	24	25	30	9	18	20	26	24	2	4	39	374
Carbonic anhydrase B (human)	19	7	17	14	1	9	13	16	11	10	20	18	2	11	17	30	14	6	8	17	260
Carbonic anhydrase C (human)	13	7	8	21	1	10	14	22	12	9	26	24	1	12	17	18	12	7	8	17	259
Carboxypeptidase A (bovine)	21	11	17	12	2	11	14	23	8	21	23	15	3	16	10	32	26	7	19	16	307
Carboxypeptidase B (bovine)	22	13	15	13	7	6	19	22	7	16	21	17	6	12	12	27	27	8	22	14	306
Dihydrofolate reductase (*E. coli*)	13	9	6	13	2	5	11	10	5	12	11	6	5	6	10	9	6	5	4	11	159
Flavodoxin (*Clostridium* MP)	6	2	8	9	3	3	18	14	0	15	8	10	5	5	3	8	5	3	3	10	138
Glyceraldehyde 3-P dehydrogenase (lobster)	32	9	10	22	5	7	17	30	5	18	18	28	10	15	12	25	20	3	9	38	333
Glyceraldehyde 3-P dehydrogenase (*B. stearotherm.*)	39	14	22	18	2	3	23	25	9	19	26	23	7	5	11	17	18	2	8	43	334
Lactate dehydrogenase (dogfish)	19	9	13	22	7	10	14	24	12	22	35	29	11	7	10	26	12	7	7	33	329
Phosphoglycerate kinase (horse)	41	11	22	23	7	7	26	40	6	18	38	42	13	16	16	24	18	4	4	40	416
Rhodanese (bovine)	23	20	8	14	4	6	22	25	8	7	25	15	5	15	18	21	13	8	11	25	293
Subtilisin BPN' (*B. amyloliquefaciens*)	37	2	17	11	0	11	4	33	6	13	15	11	5	3	14	37	13	3	10	30	275
Thioredoxin (*E. coli*)	12	1	4	11	2	3	5	9	1	9	13	10	1	4	5	3	6	2	2	5	108
Triose phosphate isomerase (chicken)	28	8	6	13	4	9	17	27	8	17	17	22	2	8	7	13	10	5	4	23	248
Total	361	146	183	244	63	114	257	377	107	239	339	321	91	158	188	327	238	72	130	378	4333

B. The maximum projection principle

Suppose x is a protein to be predicted. It corresponds to a 20-D space unit vector as defined by

$$v(x) = \{v_1(x), v_2(x), \ldots, v_{20}(x)\} \tag{1}$$

where

$$v_j(x) = \frac{n_j(x)}{\sqrt{\sum_{i=1}^{20} n_i^2(x)}} \quad (j = 1, 2, \ldots, 20) \tag{2}$$

where $n_j(x)$ is the frequency of amino acid j occurring in protein x. Similarly, the norm for each of the four protein folding types can be expressed as

$$\begin{cases} \mathbf{v}(\alpha) = \{v_1(\alpha), v_2(\alpha), \ldots, v_{20}(\alpha)\} \\ \mathbf{v}(\beta) = \{v_1(\beta), v_2(\beta), \ldots, v_{20}(\beta)\} \\ \mathbf{v}(\alpha + \beta) = \{v_1(\alpha + \beta), v_2(\alpha + \beta), \ldots, v_{20}(\alpha + \beta)\} \\ \mathbf{v}(\alpha/\beta) = \{v_1(\alpha/\beta), v_2(\alpha/\beta), \ldots, v_{20}(\alpha/\beta)\} \end{cases} \tag{3}$$

where $v_j(\alpha)$, $v_j(\beta)$, $v_j(\alpha + \beta)$, and $v_j(\alpha/\beta)(j = 1, 2, \ldots, 20)$ can be calculated by means of eq. 2 and the data of Table 5, which are derived from the training set data in Tables 1, 2, 3, and 4, respectively.

Table 5. The standard amino acid composition of the four protein folding types derived from Tables 1–4

Amino acid	$n_j(\alpha)^a$	$n_j(\beta)^b$	$n_j(\alpha + \beta)^c$	$n_j(\alpha/\beta)^d$
Ala	285	203	175	361
Arg	53	67	78	146
Asn	97	139	120	183
Asp	163	121	111	244
Cys	22	75	74	63
Gln	67	122	74	114
Glu	134	86	86	257
Gly	197	297	171	377
His	111	49	33	107
Ile	91	120	93	239
Leu	221	177	110	339
Lys	249	115	112	321
Met	48	16	25	91
Phe	123	85	52	158
Pro	82	127	71	188
Ser	122	341	126	327
Thr	119	253	117	238
Trp	33	44	30	72
Tyr	63	110	108	130
Val	167	229	123	378
$\sum n_j$	2447	2776	1889	4333

[a] Derived from the 19 α proteins in Table 1. [b] Derived from the 15 β proteins in Table 2. [c] Derived from the 14 $\alpha + \beta$ proteins in Table 3. [d] Derived from the 16 α/β proteins in Table 4.

The projections of the protein x with the norms of the α, β, $\alpha + \beta$, α/β proteins are defined by

$$\begin{cases} p_\alpha^x = \mathbf{v}(x) \cdot \mathbf{v}(\alpha) = |\mathbf{v}(x)||\mathbf{v}(\alpha)| \cos(\theta_\alpha^x) \\ p_\beta^z = \mathbf{v}(x) \cdot \mathbf{v}(\beta) = |\mathbf{v}(x)||\mathbf{v}(\beta)| \cos(\theta_\beta^x) \\ p_{\alpha+\beta}^x = \mathbf{v}(x) \cdot \mathbf{v}(\alpha + \beta) = |\mathbf{v}(x)||\mathbf{v}(\alpha + \beta)| \cos(\theta_{\alpha+\beta}^x) \\ p_{\alpha/\beta}^x = \mathbf{v}(x) \cdot \mathbf{v}(\alpha/\beta) = |\mathbf{v}(x)||\mathbf{v}(\alpha/\beta)| \cos(\theta_{\alpha/\beta}^x) \end{cases} \tag{4}$$

where θ_α^x (Fig. 1) is the correlation angle of the vector $\mathbf{v}(x)$ with the standard vector $\mathbf{v}(\alpha)$, θ_β^x the correlation angle of the vector $\mathbf{v}(x)$ with the vector $\mathbf{v}(\beta)$, and so forth, and they are given by:

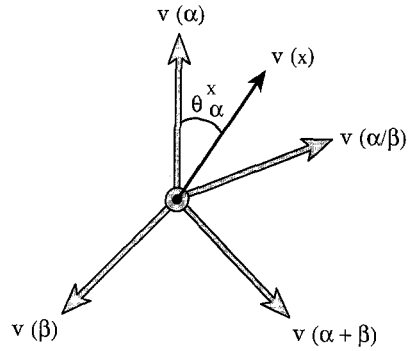

Fig. 1. Schematic illustration to show the correlation angle between two vectors in a 20-D space. The black single-line vector $\mathbf{v}(x)$ representing the protein x to be predicted, while the four shaded double-line vectors $\mathbf{v}(\alpha)$, $\mathbf{v}(\beta)$, $\mathbf{v}(\alpha + \beta)$, and $\mathbf{v}(\alpha/\beta)$ represent the norms of the four protein folding types, i.e., α, β, $\alpha + \beta$, and α/β types, respectively. θ_α^x is the correlation angle between $\mathbf{v}(\alpha)$ and $\mathbf{v}(x)$. The correlation angles between $\mathbf{v}(x)$ and the other three standard vectors can be likewise illustrated although they are not explicitly marked here

$$\begin{cases} \theta_\alpha^x = \arccos\left\{ \dfrac{\sum_{i=1}^{20} v_i(x) v_i(\alpha)}{\{[\sum_{i=1}^{20} v_i(x)^2][\sum_{i=1}^{20} v_i(\alpha)^2]\}^{1/2}} \right\} \\ \theta_\beta^x = \arccos\left\{ \dfrac{\sum_{i=1}^{20} v_i(x) v_i(\beta)}{\{[\sum_{i=1}^{20} v_i(x)^2][\sum_{i=1}^{20} v_i(\beta)^2]\}^{1/2}} \right\} \\ \theta_{\alpha+\beta}^x = \arccos\left\{ \dfrac{\sum_{i=1}^{20} v_i(x) v_i(\alpha + \beta)}{\{[\sum_{i=1}^{20} v_i(x)^2][\sum_{i=1}^{20} v_i(\alpha + \beta)^2]\}^{1/2}} \right\} \\ \theta_{\alpha/\beta}^x = \arccos\left\{ \dfrac{\sum_{i=1}^{20} v_i(x) v_i(\alpha/\beta)}{\{[\sum_{i=1}^{20} v_i(x)^2][\sum_{i=1}^{20} v_i(\alpha/\beta)^2]\}^{1/2}} \right\} \end{cases} \tag{5}$$

The protein x is predicted to belong to the folding type for which the projection is the largest, or according to trigonometry, the correlation angle is the smallest. Therefore, the method is based on the "maximum projection" or "least correlation angle" principle. The rationale of the current method is self-evident, as can be seen by the fact that the correlation angle of two identical vectors must be zero, and this is also the case in which their projection reaches the maximum. Generally speaking, the larger the projection between two vectors, the smaller their correlation angle, and the more similar the two vectors.

Results and discussion

Two sets of data, the training set of proteins and the test set of proteins are used to demonstrate the predicted results according to the maximum projection principle.

A. Predicted results for the training set of proteins

To test our method we used the same training set used by Chou (1989). The predicted results for the 19 α, 15 β, 14 $\alpha + \beta$, and 16 α/β proteins are given in Tables 6–9, respectively. The rates of correct prediction by means of the current method as well as the other four existing methods are listed in Table 10. The new method gives a $\geq 80\%$ rate of correct prediction for three of the four folding types, but the other four methods can do so only for one or two classes. None of the four methods could predict $\alpha + \beta$ proteins with an accuracy higher than 80%. Even for the Klein & Delisi method (1986) in which the prediction was made by grouping the $\alpha + \beta$ and α/β classes into one "mixed class", the predicted result was poor, with an accuracy of only 68.2%. Similar grouping with the current method would increase the correct prediction to $\frac{26}{30} = 86.7\%$. Table 10

Table 6. The correlation angles of the 19 α type proteins with the four standard folding type proteins

Name of the 19 α proteins[a]	Correlation angles[b] (deg)				The predicted folding type
	θ_α^x	θ_β^x	$\theta_{\alpha+\beta}^x$	$\theta_{\alpha/\beta}^x$	
Calcium-binding parvalbumin (carp)	22*	43	35	32	α
Cytochrome b_{562} (*E. coli*)	26*	46	33	34	α
Cytochrome *c* (tuna)	23*	33	25	25	α
Cytochrome c_2 (*R. rubrum*)	18*	37	28	26	α
Cytochrome c_{550} (*P. denitrificans*)	21*	35	25	25	α
Cytochrome c_{555} (*C. thiosulfatophilum*)	32*	42	35	37	α
α-Deoxyhemoglobin (human)	20*	33	31	26	α
β-Deoxyhemoglobin (human)	19*	33	29	22	α
γ-Deoxyhemoglobin (human fetal)	15*	24	24	15*	α or α/β
Hemerythrin B (*G. gouldii*)	25*	40	30	28	α
Hemoglobin (*glycera*)	24*	35	29	28	α
Hemoglobin (lamprey)	18*	29	23	19	α
Hemoglobin (midge larva)	20*	29	25	22	α
Methemerythrin (*T. dyscritum*)	26*	38	29	30	α
Methemerythrin (*T. pyroides*)	21*	39	30	24	α
α-Methemoglobin (horse)	21*	31	31	25	α
β-Methemoglobin (horse)	19*	35	30	23	α
Myoglobin (scal)	19*	41	34	26	α
Myoglobin (sperm whale)	17*	41	33	25	α
Rate of correct prediction = $\frac{18.5}{19} = 97.4\%$					

[a] In alphabetical order. [b] See eq. 4 for the definitions of the four correlation angles, of which the one with the smallest value (marked by *) is assumed to correspond to the folding type for the predicted protein.

Table 7. The correlation angles of the 15 β type proteins with the four standard folding type proteins

Name of the 15 β proteins[a]	Correlation angles[b] (deg)				The predicted folding type
	θ_α^x	θ_β^x	$\theta_{\alpha+\beta}^x$	$\theta_{\alpha/\beta}^x$	
α-Chymotrypsin (bovine)	27	12*	20	19	β
Concanavalin A (jack bean)	27	17*	23	19	β
Elastase (porcine)	35	17*	22	26	β
Erabutoxin B (sea snake)	50	35*	38	42	β
Immunoglobulin F_{ab} (V_H and C_H) (human)	37	15*	29	27	β
Immunoglobulin F_{ab} (V_L and C_L) (human)	31	15*	25	23	β
Immunoglobulin B-J MCG (human)	32	14*	24	24	β
Immunoglobulin B-J REI (human)	42	25*	32	35	β
Penicillopepsin (*P. janthinellum*)	37	15*	28	29	β
Prealbumin (human)	23	20	21	15*	$\alpha + \beta$
Protease A (*S. griseus*)	40	20*	27	32	β
Protease B (*S. griseus*)	43	22*	31	35	β
Rubredoxin (*C. pasteurianum*)	46	50	43*	43*	$\alpha + \beta$ or α/β
Superoxide dismutase (bovine)	28	26	26	22	α/β
Trypsin (bovine)	36	18*	25	27	β
Rate of correct prediction = $\frac{12}{15}$ = 80.0%					

[a] In alphabetical order. [b] See eq. 4 for the definitions of the four correlation angles, of which the one with the smallest value (marked by *) is assumed to correspond to the folding type for the predicted protein.

Table 8. The correlation angles of the 14 $\alpha + \beta$ type proteins with the four standard folding type proteins

Name of the 14 $\alpha + \beta$ proteins[a]	Correlation angles[b] (deg)				The predicted folding type
	θ_α^x	θ_β^x	$\theta_{\alpha+\beta}^x$	$\theta_{\alpha/\beta}^x$	
Actindin (kiwi fruit)	32	22	17*	24	$\alpha + \beta$
Cytochrome b_5 (bovine)	28	34	31	25*	α/β
Ferredoxin (*P. aerogenes*)	48	42	37*	43	$\alpha + \beta$
High-potential iron protein (chromatium)	30*	41	31	36	α
Insulin (A and B chains) (porcine)	39	36	33*	35	$\alpha + \beta$
Lysozyme (bacteriophage T_4)	24	32	21*	22	$\alpha + \beta$
Lysozyme (chicken)	33	28	21*	28	$\alpha + \beta$
Papain (papaya)	34	25	19*	26	$\alpha + \beta$
Phospholipase A_2 (bovine)	39	38	30*	37	$\alpha + \beta$
Ribonuclease S (bovine)	35	28	26*	31	$\alpha + \beta$
Staphylococcal nuclease (*S. aureus*)	19*	37	27	24	α
Subtilisin inhibitor (streptomyces)	30	29	29	28*	α/β
Thermolysin (*B. thermoproteolyticus*)	29	19	16*	23	$\alpha + \beta$
Trypsin inhibitor (bovine)	39	42	30*	38	$\alpha + \beta$
Rate of correct prediction = $\frac{10}{14}$ = 71.4%					

[a] In alphabetical order. [b] See eq. 4 for the definitions of the four correlation angles, of which the one with the smallest value (marked by *) is assumed to correspond to the folding type for the predicted protein.

Table 9. The correlation angles of the 16 α/β type proteins with the four standard folding type proteins

Name of the 16 α/β proteins[a]	Correlation angles[b] (deg)				The predicted folding type
	θ_α^x	θ_β^x	$\theta_{\alpha+\beta}^x$	$\theta_{\alpha/\beta}^x$	
Adenylate kinase (procine)	25	31	27	20*	α/β
Alcohol dehydrogenase (horse)	21	22	21	12*	α/β
Carbonic anhydrase B (human)	23	19	21	16*	α/β
Carbonic anhydrase C (human)	19	27	25	17*	α/β
Carboxypeptidase A (bovine)	28	24	58	12*	α/β
Carboxypeptidase B (bovine)	26	19	16*	18	$\alpha + \beta$
Dihydrofolate reductase (*E. coli*)	24	29	22	17*	α/β
Flavodoxin (*Clostridium* MP)	33	37	31	26*	α/β
Glyceraldehyde 3-P dehydrogenase (lobster)	18	22	19	11*	α/β
Glyceraldehyde 3-P dehydrogenase (*B.* stearotherm.)	20	29	22	16*	α/β
Lactate dehydrogenase (dogfish)	21	27	24	14*	α/β
Phosphoglycerate kinase (horse)	13	28	21	12*	α/β
Rhodanese (bovine)	22	25	21	25*	α/β
Subtilisin BPN′ (*B. amyloliquefaciens*)	31	19*	24	24	β
Thioredoxin (*E. coli*)	19*	35	36	23	α
Triose phosphate isomerase (chicken)	15	28	20	13*	α/β
Rate of correct prediction = $\frac{13}{16}$ = 81.3%					

[a] In alphabetical order. [b] See eq. 4 for the definitions of the four correlation angles, of which the one with the smallest value (marked by *) is assumed to correspond to the folding type for the predicted protein.

Table 10. Comparison of various prediction methods in self-consistency

Method[a]	Rate of correct prediction				Average accuracy
	α type	β type	$\alpha + \beta$ type	α/β type	
This paper	$\frac{18.5}{19}$ = 97.4%	$\frac{12}{15}$ = 80.0%	$\frac{10}{14}$ = 71.4%	$\frac{13}{16}$ = 81.3%	$\frac{53.5}{64}$ = 83.6%
P. Y. Chou (1989)	$\frac{16}{19}$ = 84.2%	$\frac{12}{15}$ = 80.0%	$\frac{11}{14}$ = 78.6%	$\frac{12}{16}$ = 75.0%	$\frac{51}{64}$ = 79.7%
Nakashima et al. (1986)	$\frac{27}{31}$ = 87.1%	$\frac{22}{34}$ = 64.7%	$\frac{10}{27}$ = 37.0%	$\frac{33}{39}$ = 84.6%	$\frac{92}{131}$ = 70.2%
Klein (1986)	$\frac{20}{29}$ = 68.9%	$\frac{25}{27}$ = 92.6%	$\frac{10}{16}$ = 62.5%	$\frac{20}{26}$ = 76.9%	$\frac{75}{98}$ = 76.5%
Klein & Delisi (1986)	$\frac{17}{20}$ = 85.0%	$\frac{14}{17}$ = 82.4%	$\frac{15}{22}$ = 68.2%		$\frac{46}{59}$ = 78.0%

[a] The methods listed here are all based on the amino acid composition except the Klein and Delisi method, which is based on the hydrophobic values of the constituent amino acids. Also, in that method, the $\alpha + \beta$ and α/β proteins were treated as one class called the "mixed" class.

also indicates that the average accuracy of the new method is about 4% higher than that by the Chou's method, about 13% higher than that by the Nakashima

Table 11. Amino acid composition in the 35 testing proteins, of which 9 are α-type, 8 β-type, 8 α + β-type, and 10 α/β-type

9 α proteins of known structure																					
Name[a]	Ala	Arg	Asn	Cys	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Ser	Ser	Thr	Trp	Tyr	Val	Total
1CC5	13	2	3	8	4	1	1	14	1	1	9	7	2	0	3	6	3	1	1	3	83
0AF1	15	11	6	12	2	12	15	11	6	4	28	9	3	8	2	10	6	1	6	8	175
2CDV	12	0	4	8	8	2	3	10	9	0	3	20	3	2	4	5	6	0	3	5	107
0CY3	13	1	2	10	9	3	4	9	8	3	4	16	1	3	8	4	10	0	1	6	116
351C	13	1	3	5	2	5	5	7	1	3	4	8	2	2	6	3	2	2	1	7	82
1CCY	28	2	2	5	2	8	10	10	1	12	12	15	3	4	7	5	6	3	0	3	128
1ECD	17	3	5	9	0	4	5	11	4	9	6	10	4	14	5	9	9	1	2	9	136
1LHB	21	5	0	14	1	0	10	6	2	8	10	13	4	8	6	13	9	2	4	12	148
1LH1	21	1	6	6	0	4	14	7	5	9	14	14	1	7	5	9	8	3	2	17	153
8 β proteins of known structure																					
Name[a]	Ala	Arg	Asn	Cys	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Ser	Ser	Thr	Trp	Tyr	Val	Total
1ACX	20	1	4	5	4	4	1	15	1	1	4	1	0	5	6	15	10	0	3	8	108
—[b]	17	27	32	28	22	16	21	38	10	39	21	17	8	14	20	46	34	12	13	34	469
1PEP	16	2	13	29	6	13	13	35	1	25	26	1	4	14	15	44	26	5	16	22	326
1ALP	24	12	13	2	6	9	4	32	1	8	10	2	2	6	4	20	18	2	4	19	198
2APE	46	1	16	17	2	14	0	44	3	14	25	4	1	14	7	41	30	3	13	23	318
3SBV	28	17	13	6	4	10	7	15	2	13	22	12	8	4	17	17	29	4	9	23	260
1NXB	0	3	3	1	8	4	4	5	2	4	1	4	0	2	4	8	5	1	1	2	62
1REI	6	2	2	5	2	13	2	8	0	8	8	4	1	3	6	14	11	1	8	3	107

Table 11. *(continued)*

8 $\alpha + \beta$ proteins of known structure																					
Name[a]	Ala	Arg	Asn	Cys	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Ser	Ser	Thr	Trp	Tyr	Val	Total
1OVO	4	1	6	3	6	0	2	3	1	0	3	5	0	2	4	5	3	0	3	5	56
1P2P	8	4	13	9	14	1	6	6	3	5	7	9	2	5	5	10	6	1	8	2	124
ORNB	8	6	6	9	0	4	3	10	2	8	7	8	0	4	3	9	9	3	7	4	110
ORST	5	6	2	9	2	1	13	12	2	3	5	2	0	5	8	3	8	0	9	6	101
ORNT	5	1	9	6	4	3	6	12	3	2	3	1	0	4	4	15	6	1	9	8	104
0HMG	17	13	25	16	6	15	24	27	5	23	19	25	6	12	7	9	17	7	11	12	296
0SDE	26	6	12	12	0	6	13	13	8	6	21	17	2	11	9	11	9	6	7	10	205
1GF1	6	6	1	4	6	2	4	7	0	1	6	3	1	4	5	5	3	0	3	3	70
10 α/β proteins of known structure																					
Name[a]	Ala	Arg	Asn	Cys	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Ser	Ser	Thr	Trp	Tyr	Val	Total
2ATC	10	8	10	10	4	3	13	6	4	12	15	10	2	5	7	11	7	0	3	12	152
0PHH	36	38	4	19	5	15	34	34	9	18	46	12	6	12	17	19	20	5	17	28	394
0PB1	63	63	45	51	9	31	64	48	22	49	79	48	21	38	36	29	35	12	36	62	841
0ETU	27	23	7	24	3	8	37	41	11	29	28	23	10	14	20	10	30	1	10	37	393
3CAT	35	31	30	39	4	21	25	34	21	18	35	27	10	31	38	25	22	6	20	34	506
1KGA	31	14	6	11	5	5	15	21	1	19	21	7	7	7	14	10	11	3	3	14	225
2MDH	32	8	11	13	8	8	16	28	5	23	27	25	6	11	21	18	21	0	5	28	314
0TT4	3	3	3	7	2	4	5	7	2	6	7	9	3	7	4	2	4	0	4	5	87
1ABP	31	8	10	21	1	11	21	29	3	16	21	30	10	12	15	14	16	5	6	26	306
8DFR	10	8	8	11	1	6	14	10	4	12	16	18	5	8	11	12	7	3	8	4	176

[a] Here only PDB (Protein Data Bank (Bernstein et al., 1977) codes are given owing to the limitation of space; the full names of these proteins can be found in Table 12. [b] *Tokyo influenza virus* whose PDB code is not available.

Table 12. Application of the new method to the independent 35 testing proteins listed in Table 11

Name of protein	Correlation angles[a] (deg)				X-ray	Predicted	PDB[b]
	θ_α^x	θ_β^x	$\theta_{\alpha+\beta}^x$	$\theta_{\alpha/\beta}^x$			
Cytochrome c_5	27*	34	29	31	α	α	1CC5
Apoferritin	25*	35	30	29	α	α	0AF1
Cytochrome c_3 (*D. vulgaris*)	30*	44	37	37	α	α	2CDV
Cytochrome c_3 (*D. desulfuricans*)	27*	39	32	32	α	α	0CY3
Cytochrome c_{551}	21*	36	26	25	α	α	351C
Cytochrom c'	26*	44	36	35	α	α	1CCY
Erythrocruorin	20*	29	25	22	α	α	1ECD
Hemoglobin, (Sea lamprey)	21*	33	28	22	α	α	1LHB
Leghemoglobin A (Soybean)	18	34	28	19*	β	β	1LH1
Actinoxanthin	37	25*	30	34	β	β	1ACX
Neuraminidase (Tokyo influenza virus)	35	20*	21	23	β	β	—
Pepsin	37	18*	27	27	β	β	1PEP
α-Lytic protease	37	21*	26	29	β	β	1ALP
Acid proteinase	33	19*	26	29	β	β	2APE
Coat protein	29	24	22*	23	β	$\alpha + \beta$	3SBV
Neurotoxin A (Sea snake)	50	35*	38	42	β	β	1NXB
Immunoglobulin RHE (Human)	42	25*	32	35	β	β	1REI
Ovomucoid, 3rd domain	36	33	28*	32	$\alpha + \beta$	$\alpha + \beta$	1OVO
Phospholipase A_2	37	36	27*	35	$\alpha + \beta$	$\alpha + \beta$	1P2P
Ribonuclease	26	23	17*	21	$\alpha + \beta$	$\alpha + \beta$	0RNB
Ribonuclease ST	38	37	31*	33	$\alpha + \beta$	$\alpha + \beta$	0RST
Ribonuclease T_1	40	21*	26	31	$\alpha + \beta$	β	0RNT
Hemagglutinin, HA2 chain	26	32	21*	23	$\alpha + \beta$	$\alpha + \beta$	0HMG
Superoxide dismutase, Mn	11*	31	22	19	$\alpha + \beta$	α	0SDE
Insulin-like growth factor (Human)	32	31	25*	28	β	β	1GF1
Aspartate transcarbamoylase (*E. Coli*)	19	28	23	18*	α/β	α/β	2ATC
p-Hydroxybenzoate hydroxylase	28	32	26	23*	α/β	α/β	0PHH
Glycogen phosphorylase A	22	32	21	18*	α/β	α/β	0PPA
Elongation factor T_u	25	31	25	19*	α/β	α/β	0ETU
Catalase	22	28	20	19*	α/β	α/β	3CAT
Aldolase	26	32	24	22*	α/β	α/β	1KGA
Malate dehydrogenase	18	24	19	11*	α/β	α/β	2MDH
Thioredoxin (Bacteriophage T4)	25	36	28	24*	α/β	α/β	0TT4
L-Arabinose binding protein (*E. Coli*)	13	28	20	12*	α/β	α/β	1ABP
Dihydrfolate reductase (Mouse L1210)	21	30	23	15*	α/β	α/β	8DFR
q = Average accuracy = $\frac{32}{35}$ = 91.4%							

[a] See eq. 5 for the definitions of the four correlation angles, of which the one with the smallest value (marked by *) is assumed to correspond to the folding type for the predicted protein.
[b] The identification code of the Protein Data Bank entries (Bernstein et al., 1977).

et al.' method, and about 7% higher than that by the Klein's method. Therefore, the new method has the highest rate of correct prediction for the training set of proteins, indicating a better self-consistency than any other existing methods.

B. Predicted results for the test set of proteins

The prediction based on an independent test set is instructive because it will indicate the extrapolating-effectiveness of a method. Klein (1986) has made a prediction for an independent set of 27 proteins, and he found the accuracy is $\frac{17}{27} = 63.0\%$. Chou (1989) has also made a prediction for an independent set of 12 proteins, and he found the accuracy is $\frac{10}{12} = 83.3\%$. However, in the paper by Nakashima et al. (1986), no calculated results for an independent set of proteins are reported. Listed in Table 11 are 35 structure-known proteins and their amino acid composition which did not enter into the initial parametrization. The predicted results by the new method for these 35 independent proteins are listed in Table 12, from which we can see the average accuracy is $\frac{32}{35} = 91.4\%$. This value is much higher than 63.0%, the accuracy reported by Klein (1986), and also higher than 83.3%, the accuracy reported in Table 16 by Chou, P. Y. (1989) even though he used a test set less than half the size of ours. Furthermore, if the same set of 35 proteins was predicted by means of the P. Y. Chou's method, the average accuracy would drop to $\frac{26}{35} = 74.3\%$, i.e., 17% lower than the result by the new method.

Consequently, the new method can lead to a considerably higher average accuracy than any of the previous methods for both training base and independent test sets of proteins, indicating that a significant improvement has been achieved by the new method in both the self-consistency and extrapolating-effectiveness.

Finally, it should be pointed out that, although protein structural class is correlated with amino acid composition, the former cannot be uniquely defined by the latter. The effect of some "hidden variables", such as the amino acid order along the sequence of a protein, was not taken into account in any of the aforementioned prediction methods. These variables certainly somehow affect the folding of a protein. Therefore, the accuracy of prediction based on such a set of incomplete parameters must have an upper limit with fluctuation. It is merely within such a limit that the new method has been improved, yielding a better rate of correct prediction than the previous methods.

A detailed analysis and discussion about the upper limits by means of Monte Carlo simulation approach will be presented elsewhere.

Acknowledgements

The authors are greatly indebted to Dr. Ken Nishikawa, Dr. Hiroshi Nakashima, and Professor Tasuo Ooi for their kindly providing the relevant data.

References

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) The protein data bank: a computer-based archival file for macromolecular structures. J Mol Biol 112: 535–542

Carlacci L, Chou KC, Maggiora GM (1991) A heuristic approach to predicting the tertiary structure of bovine somatotropin. Biochemistry 30: 4389–4398
Chou KC (1992) Energy-optimized structure of antifreeze protein and its binding mechanism. J Mol Biol 223: 509–517
Chou KC, Carlacci L (1991a) Energetics approach to the folding of α/β barrels. Proteins: Structure, Function and Genetics 9: 280–295
Chou KC, Carlacci L (1991b) Simulated annealing approach to the study of protein structures. Protein Eng 4: 661–667
Chou KC, Scheraga HA (1982) Origin of the right-handed twist of β-sheets of poly(L-Val) chains. Proc Natl Acad Sci USA 79: 7047–7051
Chou KC, Zhang CT (1992) A correlation-coefficient method to predicting protein-structural classes from amino-acid compositions. Eur J Biochem 207: 429–433
Chou KC, Némethy G, Scheraga HA (1990) Energetics of interactions of regular structural elements in proteins. Acc Chem Res 23: 134–141
Chou KC, Némethy G, Rumsey S, Tuttle RW, Scheraga HA, (1985) Interactions between an α-helix and a β-sheet: energetics of α/β packing in proteins J Mol Biol 186: 591–609
Chou KC, Némethy G, Scheraga HA (1990) Energetics of interactions of regular structural elements in proteins. Accounts Chem Res 23: 134–141
Chou PY (1980) Amino acid composition of four classes of proteins. In: Abstracts of papers, part I. Second Chemical Congress of the North American Continent, Las Vegas
Chou PY (1989) Prediction of protein structural classes from amino acid composition. In: Fasman GD (ed) Prediction of protein structure and the principles of protein conformation. Plenum Press, New York, pp 519–586
Chou PY, Fasman GD (1974) Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. Biochemistry 13: 222–245
Deléage G, Roux B (1989) Use of class prediction to improve protein secondary structure prediction: joint prediction with methods based on sequence homology. In: Fasman GD (ed) Prediction of protein structure and the principles of protein conformation. Plenum Press, New York, pp 587–597
Garnier J, Osguthorpe DJ, Robson B (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 120: 97–120
Kawai H, Kikuchi T, Okamoto Y (1989) A prediction of tertiary structures of peptide by the Monte Carlo simulated annealing method. Protein Eng 3: 85–94
Klein P (1986) Prediction of protein structural class by discriminant analysis. Biochim Biophys Acta 874: 205–215
Klein P, Delisi C (1986) Prediction of protein structural class from amino acid sequence. Biopolymers 25: 1569–1672
Levitt M, Chothia C (1976) Structural patterns in globular proteins. Nature 261: 552–557
Nakashima H, Nishikawa K, Ooi T (1986) The folding type of a protein is relevant to the amino acid composition. J Biochem 99: 152–162
Richardson JS, Richardson DC (1989) Principles and patterns of protein conformation. In: Fasman GD (ed) Prediction of protein structure and the principles of protein conformation. Plenum Press, New York, pp 1–98
Scheraga HA (1987) Conformational analysis of polypeptides and proteins for the study of protein folding, molecular recognition, and molecular design. J Prot Chem 6: 61–80
Wilson SR, Cui W (1990) Applications of simulated annealing to peptides. Biopolymers 29: 225–235
Zhang CT, Chou KC (1992) An optimization approach to predicting protein structural class from amino acid composition. Protein Science 1: 401–408

Author's address: K.-C. Chou, Ph.D. Dr. Sci., Computational Chemistry, Upjohn Research Laboratories, Kalamazoo, MI 49001-4940, U.S.A.

Received August 24, 1993