An optimization approach to predicting protein structural class from amino acid composition

CHUN-TING ZHANG AND KUO-CHEN CHOU

Computational Chemistry, Upjohn Laboratories, Kalamazoo, Michigan 49001

(RECEIVED September 24, 1991; ACCEPTED November 1, 1991)

Abstract

Proteins are generally classified into four structural classes: all- α proteins, all- β proteins, $\alpha+\beta$ proteins, and α/β proteins. In this article, a protein is expressed as a vector of 20-dimensional space, in which its 20 components are defined by the composition of its 20 amino acids. Based on this, a new method, the so-called maximum component coefficient method, is proposed for predicting the structural class of a protein according to its amino acid composition. In comparison with the existing methods, the new method yields a higher general accuracy of prediction. Especially for the all- α proteins, the rate of correct prediction obtained by the new method is much higher than that by any of the existing methods. For instance, for the 19 all- α proteins investigated previously by P.Y. Chou, the rate of correct prediction by means of his method was 84.2%, but the correct rate when predicted with the new method would be 100%! Furthermore, the new method is characterized by an explicable physical picture. This is reflected by the process in which the vector representing a protein to be predicted is decomposed into four component vectors, each of which corresponds to one of the norms of the four protein structural classes.

Keywords: all α ; $\alpha+\beta$; α/β ; all β ; component coefficients; contradictory equations; 20-dimensional space; vector decomposition

Current computational approaches to predicting the structure of a protein from its amino acid sequence may be classified into two categories. The first one is the free energy minimization method, which is based on the empirical atomic potentials. By means of this method, one can calculate the detailed atomic coordinates (Némethy & Scheraga, 1977; Chou, 1992) and reveal the origin of the structural handedness in proteins (Chou & Scheraga, 1982; Chou et al., 1983, 1989, 1990; Chou & Carlacci, 1991a). The second one is the statistical method, which was developed based on various statistical data in an attempt to predict the "state" of a residue (Chou & Fasman, 1974; Lim, 1974; Garnier et al., 1978) or the "class" of a protein (Chou, 1980, 1989; Klein, 1986; Klein & Delisi, 1986; Nakashima et al., 1986).

Because of the notorious local minimum problem, the free energy minimization method has trouble determining the real global minimum for even a very small protein molecule (Scheraga, 1987). Although the simulated annealing approach is a powerful tool in overcoming the

local minimum problem (Kawai et al., 1989; Wilson & Cui, 1990; Chou & 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 approaches (Carlacci et al., 1991). However, there are not many proteins from which one can get enough heuristic inputs to result in a successful prediction.

On the other hand, the statistical method has the problem of poor accuracy, as reflected by large statistical fluctuations. Nevertheless, the statistical method has the merit of simplicity and convenience in application and hence has been widely adopted by biochemists. Besides, the results predicted by the statistical method, although less accurate, might reduce the scope of searching conformational space or provide some useful information for heuristic approach.

Therefore, in determining the three-dimensional structure of proteins, the statistical method can certainly play a complementary role. In this article, we propose a new method for predicting the structural class of a protein according to its amino acid composition.

It is well known that the structurally known proteins

Reprint requests to: Kuo-Chen Chou, Computational Chemistry, Upjohn Laboratories, Kalamazoo, Michigan 49001.

are generally categorized into four structural classes, i.e., all- α proteins, all- β proteins, $\alpha+\beta$ proteins, and α/β proteins (Levitt & Chothia, 1976; Richardson & Richardson, 1989). Which of these structural classes a protein belongs to was also found to correlate to its amino acid composition (Chou, 1980, 1989; Nakashima et al., 1986). Therefore, various methods have been proposed to predict the structural class of a protein according to the occurrence frequencies of its amino acids. Based on the maximum likelihood rule, the multidimensional statistical technique of discriminant analysis was used in some methods (Klein, 1986; Klein & Delisi, 1986). But in the others, the minimum distance rule was adopted (Chou, 1980, 1989; Nakashima et al., 1986). In this article, we propose a different approach to predicting the structural class of a protein. The new method is based on the optimization principle or the maximum component coefficient principle, as will be described in the next section.

The method

The quantitative criterion for the four structural classes

Based on the classification made by Chou (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 structural classes can be defined as follows:

1. α class: Proteins of this class contain more than 45% α -helices and less than 5% β -strands, e.g., cytochrome c, hemoglobin (human), myoglobin (sperm whale), etc., belong to this class.
2. β class: Proteins of this class contain less than 5% α -helices and more than 45% β -strands, e.g., α -chymotrypsin, concanavalin A, elastase, etc., belong to this type.
3. $\alpha+\beta$ class: Proteins of this class contain more than 30% α -helices and more than 20% β -strands with dominantly antiparallel β -strands, e.g., actinidin, lysozyme (chicken), papain, etc., belong to this class.
4. α/β class: Proteins of this class contain more than 30% α -helices and more than 20% β -strands with dominantly parallel β -strands, e.g., adenylate kinase, carbonic anhydrase B, flavodoxin, etc., belong to this class. Most of the α/β -class proteins are enzymes.

The maximum component coefficient principle

In order to predict the structural class of a protein based on its amino acid composition, Nakashima et al. (1986)

introduced the concept of composition space. This is a 20-dimensional (20-D) Euclidean space, in which each protein corresponds to a point whose 20 coordinates are given by the occurrence frequencies of the 20 constituent amino acids. According to such a concept, a protein can be represented by a 20-D unit vector through the following formulation. Suppose x is a protein to be predicted, and $f_i(x)$ ($i = 1, 2, \dots, 20$) represents the occurrence frequencies of its 20 constituent amino acids. The compositions of the 20 amino acids in the protein x are given by the following normalization equation:

$$v_k(x) = \frac{f_k(x)}{\sum_{i=1}^{20} f_i(x)} \quad (k = 1, 2, \dots, 20). \quad (1)$$

Thus, in the composition space, the protein x can be expressed by the vector

$$\mathbf{v}(x) = [v_1(x), v_2(x), \dots, v_{20}(x)]. \quad (2)$$

Because $v_1(x), v_2(x), \dots, v_{20}(x)$ are normalized as defined by Equation 1, the vector $\mathbf{v}(x)$ is a 20-D unit vector. Similarly, we can also use the following four standard unit vectors to represent the norms of the four protein structural classes:

$$\begin{cases} \mathbf{v}(\alpha) &= [v_1(\alpha), v_2(\alpha), \dots, v_{20}(\alpha)] \\ \mathbf{v}(\beta) &= [v_1(\beta), v_2(\beta), \dots, v_{20}(\beta)] \\ \mathbf{v}(\alpha+\beta) &= [v_1(\alpha+\beta), v_2(\alpha+\beta), \dots, v_{20}(\alpha+\beta)] \\ \mathbf{v}(\alpha/\beta) &= [v_1(\alpha/\beta), v_2(\alpha/\beta), \dots, v_{20}(\alpha/\beta)] \end{cases} \quad (3)$$

where $v_1(\alpha), v_2(\alpha), \dots, v_{20}(\alpha)$ are the average compositions of the 20 amino acids occurring in the 19 known α -type proteins, $v_1(\beta), v_2(\beta), \dots, v_{20}(\beta)$ the average compositions of the 20 amino acids occurring in the 15 known β -type proteins, and so forth (Chou, 1989). Thus, each of such standard vectors is uniquely defined by its 20 components as given in Table 1.

The Euclidean distance between the end of $\mathbf{v}(x)$ and the ends of $\mathbf{v}(k)$ ($k = \alpha, \beta, \alpha+\beta, \alpha/\beta$) are given by

$$d_k^E = |\mathbf{v}(x) - \mathbf{v}(k)| = \left\{ \sum_{i=1}^{20} [v_i(x) - v_i(k)]^2 \right\}^{1/2} \quad (4)$$

$(k = \alpha, \beta, \alpha+\beta, \alpha/\beta).$

Thus, according to the method proposed by Nakashima et al. (1986), the protein x is assigned to be the k -class if d_k^E is the smallest one among the four distances given by Equation 4. Instead of Euclidean distance, Chou (1980, 1989) chose the following Minkowski's distance,

$$d_k^M = \left\{ \sum_{i=1}^{20} |v_i(x) - v_i(k)|^p \right\}^{1/p} \quad (5)$$

$$(k = \alpha, \beta, \alpha + \beta, \alpha/\beta; p = 1),$$

as a criterion to assign the predicted protein x .

Here, we approach this problem from a completely different avenue. The new method can be formulated as follows. First, we decompose the vector $v(x)$ into the four component vectors, each of which corresponds to one of the four standard vectors $v(\alpha)$, $v(\beta)$, $v(\alpha + \beta)$, and $v(\alpha/\beta)$ as defined by Equation 3; i.e.,

$$v(x) = a_\alpha v(\alpha) + a_\beta v(\beta) + a_{\alpha+\beta} v(\alpha + \beta) + a_{\alpha/\beta} v(\alpha/\beta), \quad (6a)$$

or

$$v_i(x) = a_\alpha v_i(\alpha) + a_\beta v_i(\beta) + a_{\alpha+\beta} v_i(\alpha + \beta) + a_{\alpha/\beta} v_i(\alpha/\beta) = \sum_k a_k v_i(k) \quad (6b)$$

$$(i = 1, 2, \dots, 20),$$

where a_k ($k = \alpha, \beta, \alpha + \beta, \alpha/\beta$) are the four component coefficients, and they should be subject to the following constraints:

$$\begin{cases} \sum_k a_k = 1 \\ 0 \leq a_k \leq 1 \end{cases} \quad (k = \alpha, \beta, \alpha + \beta, \alpha/\beta). \quad (7)$$

Our goal is to determine the four component coefficients a_k ($k = \alpha, \beta, \alpha + \beta, \alpha/\beta$). However, Equation 6b contains 20 equations. Therefore, we have to deal with a problem mathematically termed a set of contradictory equations, in which the number of equations is greater than that of the unknown variables. To solve these kinds of contradictory equations we adopt the least-squares method, as illustrated below.

Define an objective function Q by

$$Q(a_\alpha, a_\beta, a_{\alpha+\beta}, a_{\alpha/\beta}) = \sum_{i=1}^{20} [v_i(x) - \sum_k a_k v_i(k)]^2, \quad (8)$$

such that

$$Q(a_\alpha, a_\beta, a_{\alpha+\beta}, a_{\alpha/\beta}) = \text{minimum} \quad (9)$$

with the constraint imposed by Equation 7. It can be proved that the above optimization problem possesses a unique solution (Rao, 1984). The solution can be easily obtained by means of any existing numerical minimizer, such as the one provided in the IMSL LIBRARY (For-

Table 1. The four standard vectors representing the norms of four structural classes in the 20-D composition space^a

Amino acid		Standard vector			
Order ^b	Name	$v(\alpha)$	$v(\beta)$	$v(\alpha + \beta)$	$v(\alpha/\beta)$
1	Arg	0.022	0.024	0.041	0.034
2	Leu	0.090	0.064	0.058	0.078
3	Ser	0.050	0.123	0.067	0.075
4	Thr	0.049	0.091	0.062	0.055
5	Pro	0.034	0.046	0.038	0.043
6	Ala	0.116	0.073	0.093	0.083
7	Gly	0.081	0.107	0.091	0.087
8	Val	0.068	0.082	0.065	0.087
9	Lys	0.102	0.041	0.059	0.074
10	Asn	0.040	0.050	0.064	0.042
11	Gln	0.027	0.044	0.039	0.026
12	His	0.045	0.018	0.017	0.025
13	Glu	0.055	0.031	0.046	0.059
14	Asp	0.067	0.044	0.059	0.056
15	Tyr	0.026	0.040	0.057	0.030
16	Cys	0.009	0.027	0.039	0.015
17	Phe	0.050	0.031	0.028	0.036
18	Ile	0.037	0.043	0.049	0.055
19	Met	0.020	0.006	0.013	0.021
20	Trp	0.013	0.016	0.016	0.017

^a The four standard vectors represent all- α -, all- β -, $\alpha + \beta$ -, and α/β -type proteins, respectively. Each such vector has 20 components, which are defined by the average amino acid compositions of the corresponding folding type, and whose values are derived from the 64 structurally known proteins (Chou, 1989).

^b The order of amino acids, each of which corresponds to a component of the 20-dimensional composition space, is numbered according to the codon usage table compiled by Wada et al. (1990); i.e., the order of an amino acid increases with a decrease in the degree of degeneracy in its genetic code. If two amino acids are the same in such a degeneracy, then they are arranged in alphabetical order.

tran subroutines for mathematics and statistics). However, the solution for Equations 8 and 9 can also be obtained via the analytical method as given in the Appendix. Once the solution for a_k ($k = \alpha, \beta, \alpha + \beta, \alpha/\beta$) is found, the predicted structural class for protein x is given by the subscript j in the following equation:

$$a_j = \max(a_\alpha, a_\beta, a_{\alpha+\beta}, a_{\alpha/\beta}), \quad (10)$$

where j can be α , β , $\alpha + \beta$, or α/β , and the operator **max** means taking the largest one among those in the parentheses. The physical implication of Equation 10 can be further illustrated as follows. When, e.g., $j = \alpha$, i.e., $a_\alpha = \max(a_\alpha, a_\beta, a_{\alpha+\beta}, a_{\alpha/\beta})$, it is obvious according to Equation 6a that the weight of the standard vector $v(\alpha)$ in $v(x)$ is the largest, and so is the "ingredient" of the α protein in the protein x . Therefore, the predicted protein x should be assigned to be the α structural class.

Results and discussion

Predicted results

In order to facilitate comparison, let us also consider the 64 proteins investigated by Chou (1989). It is known that in the 64 proteins there are 19 α proteins, 15 β proteins, 14 $\alpha+\beta$ proteins, and 16 α/β proteins. Below, let us see what predicted results are obtained if the new method is employed.

1. α proteins: The four component coefficients a_j ($j = \alpha, \beta, \alpha+\beta, \alpha/\beta$) for the 19 α proteins are given in Table 2, from which we can see that a_α has the largest values for all the proteins listed there. Therefore, according to the maximum component coefficient principle (cf. Equation 10), all of the 19 proteins are correctly predicted to be the α structural class. These results indicate that the rate of correct prediction for the α proteins is 100%.
2. β proteins: The component coefficients for the 15 β proteins are listed in Table 3. As we can see, of the 15 proteins, 12 are correctly predicted as β proteins according to the maximum component coefficient principle. The rate of correct prediction for the β proteins is 80.0%.
3. $\alpha+\beta$ proteins: The component coefficients for the

14 $\alpha+\beta$ proteins are listed in Table 4. Of the 14 proteins, 10 are correctly predicted as $\alpha+\beta$ proteins according to the maximum component coefficient principle. The rate of correct prediction for the β proteins is 71.4%.

4. α/β proteins: The component coefficients for the 16 α/β proteins are listed in Table 5. Of the 16 proteins, 12 are correctly predicted as α/β proteins according to the maximum component coefficient principle. The rate of correct prediction for the α/β proteins is 75.0%.

Comparison with the other methods

If the method proposed in this paper is termed as the maximum component coefficient method, then, according to Equations 4 and 5, the method proposed by Nakashima et al. (1986) can be termed the least Euclid's distance method, whereas the method proposed by Chou (1980, 1989) can be termed the least Minkowski's distance method.

Listed in Table 6 are the rates of correct prediction by means of the maximum component coefficient method, the least Minkowski's distance method (Chou, 1980, 1989), and the least Euclid's distance method (Nakashima

Table 2. The component coefficients^a obtained for 19 α proteins by minimizing the objective function of Equation 8

	Component coefficient				Predicted structural class
	a_α	a_β	$a_{\alpha+\beta}$	$a_{\alpha/\beta}$	
Calcium-binding parvalbumin (carp)	1.00*	0.00	0.00	0.00	α
Cytochrome b_{562} (<i>E. coli</i>)	1.00*	0.00	0.00	0.00	α
Cytochrome c (tuna)	0.61*	0.00	0.39	0.00	α
Cytochrome c_2 (<i>R. rubrum</i>)	1.00*	0.00	0.00	0.00	α
Cytochrome c_{550} (<i>P. denitrificans</i>)	0.73*	0.00	0.27	0.00	α
Cytochrome c_{555} (<i>C. thiosulfatophilum</i>)	0.70*	0.00	0.30	0.00	α
Hemerythrin B (<i>G. gouldii</i>)	0.74*	0.00	0.26	0.00	α
Methemerythrin (<i>T. dyscritum</i>)	0.63*	0.00	0.37	0.00	α
Methemerythrin (<i>T. pyroides</i>)	0.70*	0.00	0.00	0.30	α
α -methemoglobin (horse)	0.81*	0.19	0.00	0.00	α
β -methemoglobin (horse)	0.89*	0.00	0.00	0.11	α
α -deoxyhemoglobin (human)	0.95*	0.05	0.00	0.00	α
β -deoxyhemoglobin (human)	0.83*	0.00	0.00	0.17	α
γ -deoxyhemoglobin (human fetal)	0.58*	0.21	0.00	0.21	α
Hemoglobin (glycera)	0.88*	0.01	0.11	0.00	α
Hemoglobin (lamprey)	0.68*	0.14	0.00	0.19	α
Hemoglobin (midge larva)	0.77*	0.23	0.00	0.00	α
Myoglobin (scal)	1.00*	0.00	0.00	0.00	α
Myoglobin (sperm whale)	1.00*	0.00	0.00	0.00	α

Rate of correct prediction = 100%

^a See Equations 6 and 7 for the definitions of the four component coefficients, of which the one with the largest value (marked by *) is assumed to correspond to the structural class for the predicted protein.

Table 3. The component coefficients^a obtained for 15 β proteins by minimizing the objective function of Equation 8

	Component coefficient				Predicted structural class
	a_α	a_β	$a_{\alpha+\beta}$	$a_{\alpha/\beta}$	
α -chymotrypsin (bovine)	0.14	0.86*	0.00	0.00	β
Concanavalin A (jack bean)	0.16	0.67*	0.00	0.17	β
Elastase (porcine)	0.00	0.79*	0.21	0.00	β
Erabutoxin B (sea snake)	0.00	0.76*	0.24	0.00	β
Immunoglobulin F _{ab} (V _H and C _H) (human)	0.00	1.00*	0.00	0.00	β
Immunoglobulin F _{ab} (V _L and C _L) (human)	0.09	0.91*	0.00	0.00	β
Immunoglobulin B-J MCG (human)	0.00	1.00*	0.00	0.00	β
Immunoglobulin B-J REI (human)	0.00	1.00*	0.00	0.00	β
Penicillopepsin (<i>P. janthinellum</i>)	0.00	1.00*	0.00	0.00	β
Prealbumin (human)	0.01	0.28	0.00	0.71*	α/β
Protease A (<i>S. griseus</i>)	0.00	1.00*	0.00	0.00	β
Protease B (<i>S. griseus</i>)	0.00	1.00*	0.00	0.00	β
Rubredoxin (<i>C. pasteurianum</i>)	0.00	0.00	0.48	0.52*	α/β
Superoxide dismutase (bovine)	0.00	0.21	0.00	0.79*	α/β
Trypsin (bovine)	0.00	0.97*	0.23	0.00	β

Rate of correct prediction = 80.0%

^a See Equations 6 and 7 for the definitions of the four component coefficients, of which the one with the largest value (marked by *) is assumed to correspond to the structural class for the predicted protein.

Table 4. The component coefficients^a obtained for 14 $\alpha+\beta$ proteins by minimizing the objective function of Equation 8

	Component coefficient				Predicted structural class
	a_α	a_β	$a_{\alpha+\beta}$	$a_{\alpha/\beta}$	
Actindin (kiwi fruit)	0.00	0.22	0.78*	0.00	$\alpha+\beta$
Cytochrome <i>b</i> ₅ (bovine)	0.15	0.00	0.00	0.85*	α/β
Ferredoxin (<i>P. aerogenes</i>)	0.00	0.00	1.00*	0.00	$\alpha+\beta$
High-potential iron protein (chromatium)	0.61*	0.00	0.39	0.00	α
Insulin (A and B chains) (porcine)	0.00	0.06	0.91*	0.03	$\alpha+\beta$
Lysozyme (bacteriophage T ₄)	0.27	0.00	0.63*	0.10	$\alpha+\beta$
Lysozyme (chicken)	0.00	0.00	1.00*	0.00	$\alpha+\beta$
Papain (papaya)	0.00	0.12	0.88*	0.00	$\alpha+\beta$
Phospholipase A ₂ (bovine)	0.00	0.00	1.00*	0.00	$\alpha+\beta$
Ribonuclease S (bovine)	0.00	0.30	0.70*	0.00	$\alpha+\beta$
Staphylococcal nuclease (<i>Staphylococcus aureus</i>)	0.93*	0.00	0.07	0.00	α
Subtilisin inhibitor (streptomyces)	0.44	0.52*	0.04	0.00	β
Thermolysin (<i>B. thermoproteolyticus</i>)	0.00	0.36	0.64*	0.00	$\alpha+\beta$
Trypsin inhibitor (bovine)	0.00	0.00	1.00*	0.00	$\alpha+\beta$

Rate of correct prediction = 71.4%

^a See Equations 6 and 7 for the definitions of the four component coefficients, of which the one with the largest value (marked by *) is assumed to correspond to the structural class for the predicted protein.

et al., 1986), respectively. As shown in Table 6, a remarkable improvement by using the new method is in predicting the α proteins, for which the rate of correct prediction by the maximum component coefficient method is 16.2% higher than that by the least Minkowski's distance

method, and 13.3% higher than that by the least Euclid's distance method. Defining the general accuracy rate as the percentage of the number of correct prediction events for all types divided by the number of prediction events for all types, namely,

Table 5. The component coefficients^a obtained for 16 α/β proteins by minimizing the objective function of Equation 8

	Component coefficient				Predicted structural class
	a_α	a_β	$a_{\alpha+\beta}$	$a_{\alpha/\beta}$	
Adenylate kinase (porcine)	0.00	0.00	0.00	1.00*	α/β
Alcohol dehydrogenase (horse)	0.00	0.05	0.00	0.95*	α/β
Carbonic anhydrase B (human)	0.13	0.27	0.11	0.49*	α/β
Carbonic anhydrase C (human)	0.36	0.03	0.00	0.61*	α/β
Carboxypeptidase A (bovine)	0.05	0.55*	0.20	0.20	β
Carboxypeptidase B (bovine)	0.00	0.23	0.58*	0.19	$\alpha+\beta$
Dihydrofolate reductase (<i>E. coli</i>)	0.00	0.00	0.14	0.86*	α/β
Flavodoxin (<i>Clostridium</i> MP)	0.00	0.00	0.00	1.00*	α/β
Glyceraldehyde 3-P dehydrogenase (lobster)	0.04	0.00	0.00	0.96*	α/β
Glyceraldehyde 3-P dehydrogenase (<i>B. stearothermophilus</i>)	0.16	0.00	0.00	0.84*	α/β
Lactate dehydrogenase (dogfish)	0.00	0.00	0.00	1.00*	α/β
Phosphoglycerate kinase (horse)	0.44	0.00	0.00	0.56*	α/β
Rhodanese (bovine)	0.00	0.00	0.00	1.00*	α/β
Subtilisin BPN' (<i>B. amyloliquefaciens</i>)	0.14	0.86*	0.00	0.00	β
Thioredoxin (<i>E. coli</i>)	0.89*	0.00	0.11	0.00	α
Triose phosphate isomerase (chicken)	0.38	0.00	0.00	0.62*	α/β

Rate of correct prediction = 75.0%

^a See Equations 6 and 7 for the definitions of the four component coefficients, of which the one with the largest value (marked by *) is assumed to correspond to the structural class for the predicted protein.

Table 6. Comparison of various prediction methods

Method	Rate of correct prediction				General accuracy ^a
	α -type	β -type	$\alpha+\beta$ -type	α/β -type	
This paper ^b	$\frac{19}{19} = 100.0\%$	$\frac{12}{15} = 80.0\%$	$\frac{10}{14} = 71.4\%$	$\frac{12}{16} = 75.0\%$	$\frac{53}{64} = 82.8\%$
Chou (1989) ^c	$\frac{16}{19} = 84.2\%$	$\frac{12}{15} = 80.0\%$	$\frac{11}{14} = 78.6\%$	$\frac{12}{16} = 75.0\%$	$\frac{51}{64} = 79.7\%$
Nakashima et al. (1986) ^d	$\frac{27}{31} = 87.1\%$	$\frac{22}{34} = 64.7\%$	$\frac{10}{29} = 37.0\%$	$\frac{33}{39} = 84.6\%$	$\frac{92}{131} = 70.2\%$

^a See Equation 11 for the definition of the general accuracy. The denominator in this table represents the number of prediction events and the numerator represents that of the corresponding correct prediction events.

^b Based on the maximum component coefficient principle (cf. Equation 10).

^c Based on the least Minkowski's distance principle (cf. Equation 5).

^d Based on the least Euclid's distance principle (cf. Equation 4).

the general accuracy rate

$$= \left(\frac{\text{total number of correct prediction events}}{\text{total number of prediction events}} \right) \%, \quad (11)$$

then we find from Table 6 that the general accuracy rate by the maximum component coefficient method would reach 82.8%, which is 3.1% higher than that by the least Minkowski's distance method, and about 12.6% higher than that by the least Euclid's distance method. However, it should be pointed out that the data given in Table 6 for

the least Euclid's distance method are the predicted results for five rather than four folding types, i.e., α , β , $\alpha+\beta$, α/β , and ζ (irregular) types. Therefore, in order to make these three methods more comparable, a 5% credit should be added to the data reported by Nakashima et al. (1986), even though, however, the maximum component coefficient method still has the highest rate of correct prediction as reflected by the highest general accuracy rate.

It is instructive to note that some of the same proteins were always missed in their predictions regardless of which of the three methods was used. Generally speaking, these missed proteins are the most deviant in the data set, such as subtilisin BPN' (*Bacillus amyloliquefaciens*)

and thioredoxin (*Escherichia coli*) (Table 5). However, it would be more essential to realize that, although the structural class of a protein is correlated to its amino acid composition, the former cannot be uniquely defined by the latter. In other words, the amino acid composition of a protein is not yet sufficiently complete to form a set of complete parameters to define its structural class uniquely. 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 three prediction methods. Obviously, these hidden variables will 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 a Monte Carlo simulation approach will be presented elsewhere (Zhang & Chou, manuscript in preparation).

Conclusion

The structural class of a protein is correlated with its amino acid composition. The present study was initiated in an attempt to formulate such a correlation in an appropriate way so as to develop a more effective method for predicting the structural class of a protein according to its amino acid composition.

A protein can be expressed as a vector of a 20-D space, in which its 20 components are defined by the compositions of its 20 amino acids, respectively. Likewise, the norms of the four structural classes can also be represented by four different vectors, the so-called four standard vectors. Thus, the vector representing a protein to be predicted can be decomposed into four component vectors, each of which corresponds to one of the four standard vectors. Because such a decomposition is carried out in a 20-D space, the number of equations is greater than that of unknown variables. Therefore, we have to deal with a set of contradiction equations. For this, the operation of minimizing an objective function was adopted to determine the component coefficients. The component coefficients thus obtained reflect the weights of various standard vectors on the decomposed vector. In other words, they represent some sort of scale to measure the ingredients of the four standard structural classes in the protein to be predicted. Therefore, the predicted protein should be assigned to the structural class corresponding to the largest component coefficient. This is the essence of the maximum component coefficient principle.

The predicted results for 64 known proteins by the new method indicate that the rates of correct prediction for the α -, β -, $\alpha+\beta$ -, and α/β -type folding proteins were 100%, 80.0%, 71.4%, and 75.0%, respectively. The general accuracy rate was 82.8%. In comparison with the ex-

isting methods, the new method has the highest accuracy rate of prediction.

Acknowledgments

Sincere gratitude is extended to Dr. Ken Nishikawa, Dr. Hiroshi Nakashima, and Professor Tasuo Ooi for kindly providing relevant data to the authors. C.-T.Z. is on sabbatical leave from the Department of Physics, Tianjin University, Tianjin, China.

References

- Carlacci, L., Chou, K.C., & Maggiora, G.M. (1991). A heuristic approach to predicting the tertiary structure of bovine somatotropin. *Biochemistry* 30, 4389–4398.
- Chou, K.C. (1992). Energy-optimized structure of antifreeze protein and its binding mechanism. *J. Mol. Biol.* 223, 509–517.
- Chou, K.C. & Carlacci, L. (1991a). Energetics approach to the folding of α/β barrels. *Proteins Struct. Funct. Genet.* 9, 280–295.
- Chou, K.C. & Carlacci, L. (1991b). Simulated annealing approach to the study of protein structures. *Protein Eng.* 4, 661–667.
- Chou, K.C., Némethy, G., & Scheraga, H.A. (1983). Effect of amino acid composition on the twist and the relative stability of parallel and antiparallel β -sheets. *Biochemistry* 22, 6213–6221.
- Chou, K.C., Némethy, G., & Scheraga, H.A. (1989). Energy of stabilization of the right-handed $\beta\alpha\beta$ crossover in proteins. *J. Mol. Biol.* 205, 241–249.
- Chou, K.C., Némethy, G., & Scheraga, H.A. (1990). Energetics of interactions of regular structural elements in proteins. *Accounts Chem. Res.* 23, 134–141.
- Chou, K.C. & Scheraga, H.A. (1982). Origin of the right-handed twist of β -sheets of poly(L-Val) chains. *Proc. Natl. Acad. Sci. USA* 79, 7047–7051.
- Chou, P.Y. (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, P.Y. (1989). Prediction of protein structural classes from amino acid composition. In *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G.D., Ed.), pp. 549–586. Plenum Press, New York.
- Chou, P.Y. & Fasman, G.D. (1974). Conformational parameters for amino acids in helical, β -sheet, and random coil regions calculated from proteins. *Biochemistry* 13, 222–245.
- Garner, J., Osguthorpe, D.J., & 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, 1659–1672.
- Levitt, M. & Chothia, C. (1976). Structural patterns in globular proteins. *Nature* 261, 552–557.
- Lim, V.I. (1974). Structural principles of globular protein secondary structure. *J. Mol. Biol.* 88, 857–872.
- 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.
- Némethy, G. & Scheraga, H.A. (1977). Protein folding. *Q. Rev. Biophys.* 10, 239–352.
- Rao, S.S. (1984). *Optimization, Theory and Applications*, 2nd Ed., Chap. 6. Wiley Eastern Limited, New Delhi.
- Richardson, J.S. & Richardson, D.C. (1989). Principles and patterns of protein conformation. In *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G.D., Ed.), pp. 1–98. Plenum Press, New York.
- Scheraga, H.A. (1987). Conformational analysis of polypeptides and

proteins for the study of protein folding, molecular recognition, and molecular design. *J. Protein Chem.* 6, 61–80.

Wada, K., Aota, S., Tsuchiya, R., Ishibashi, F., Gojobori, T., & Ikemura, T. (1990). Codon usage tabulated from the GenBank genetic sequence data. *Nucleic Acids Res.* 18, 2367–2411.

Wilson, S.R. & Cui, W. (1990). Applications of simulated annealing to peptides. *Biopolymers* 29, 225–235.

Appendix: An analytic method to optimize the objective function $Q(a_\alpha, a_\beta, a_{\alpha+\beta}, a_{\alpha/\beta})$

The analytical solutions for Equations 8 and 9 with the constraints as determined by Equation 7 can be derived as follows.

By introducing the slack variables μ_k and ν_k defined in the real number field, we can transform the inequality constraint of Equation 7 into two equality constraints; i.e., the constraints on the component coefficients a_k ($k = \alpha, \beta, \alpha+\beta, \alpha/\beta$) as imposed by Equation 7 are now completely equivalent to the following equality constraints:

$$\begin{cases} \Sigma_k a_k - 1 = 0 \\ -a_k + \mu_k^2 = 0 \\ a_k - 1 + \nu_k^2 = 0 \end{cases} \quad (k = \alpha, \beta, \alpha+\beta, \alpha/\beta). \quad (\text{A.1})$$

Using the Lagrange multiplier method, we obtain a new objective function as given by

$$\begin{aligned} \tilde{Q} = Q + \lambda(\Sigma_k a_k - 1) + \Sigma_k \lambda_k(-a_k + \mu_k^2) \\ + \Sigma_k \lambda_k^*(a_k - 1 + \nu_k^2), \end{aligned} \quad (\text{A.2})$$

where λ , λ_k , and λ_k^* are the Lagrange multipliers. Now the optimization problem has been further transformed

to that without any constraints, and its solution can be obtained by solving the following set of equations:

$$\begin{cases} \frac{\partial \tilde{Q}}{\partial a_k} = \frac{\partial Q}{\partial a_k} + \lambda - \lambda_k + \lambda_k^* = 0 \\ \frac{\partial \tilde{Q}}{\partial \lambda} = \Sigma_k a_k - 1 = 0 \\ \frac{\partial \tilde{Q}}{\partial \lambda_k} = -a_k + \mu_k^2 = 0 \\ \frac{\partial \tilde{Q}}{\partial \lambda_k^*} = a_k - 1 + \nu_k^2 = 0 \\ \frac{\partial \tilde{Q}}{\partial \mu_k} = 2\lambda_k \mu_k = 0 \\ \frac{\partial \tilde{Q}}{\partial \nu_k} = 2\lambda_k^* \nu_k = 0. \end{cases} \quad (\text{A.3})$$

In the above set of equations, the first two are linear equations, and the last four are nonlinear equations. However, the nonlinear equations bear such simple forms that their solutions can be immediately written out. Substitution of their solutions into the first two equations of Equation A.3 will reduce the problem to a set of linear equations, whose solutions can be obtained by routine procedure. It should be pointed out, however, that in solving the nonlinear equations, various possible cases should be taken into account that would result in different sets of solutions. Therefore, the real set of solutions for the component coefficients should be the one in which Q has the smallest value.