Biochemical and Biophysical Research Communications **264,** 216–224 (1999)
Article ID bbrc.1999.1325, available online at http://www.idealibrary.com on IDEAL®

A Key Driving Force in Determination of Protein Structural Classes

Kuo-Chen Chou
Computer-Aided Drug Discovery, Pharmacia and Upjohn, Kalamazoo, Michigan 49007-4940

Received August 13, 1999

The three-dimensional structure of a protein is uniquely dictated by its primary sequence. However, owing to the very high degenerative nature of the sequence–structure relationship, proteins are generally folded into one of only a few structural classes that are closely correlated with the amino-acid composition. This suggests that the interaction among the components of amino acid composition may play a considerable role in determining the structural class of a protein. To quantitatively test such a hypothesis at a deeper level, three potential functions, $U^{(0)}$, $U^{(1)}$, and $U^{(2)}$, were formulated that respectively represent the 0th-order, 1st-order, and 2nd-order approximations for the interaction among the components of the amino acid composition in a protein. It was observed that the correct rates in recognizing protein structural classes by $U^{(2)}$ are significantly higher than those by $U^{(0)}$ and $U^{(1)}$, indicating that an algorithm that can more completely incorporate the interaction contributions will yield better recognition quality, and hence further demonstrate that the interaction among the components of amino acid composition is an important driving force in determining the structural class of a protein during the sequence folding process.

Key Words: **folding patterns; 0th-order potential; 1st-order potential; 2nd-order potential; amino acid components; cluster-tolerant capacity.**

The 3-D (dimensional) structure of a protein is uniquely dictated by its amino acid sequence, the so-called primary structure. However, owing to the degenerate nature of the sequence–structure relationship, although the number of protein sequences is extremely large, the number of their folding patterns is quite limited. Actually, according to their chain folding topologies, proteins are usually folded into one of the following four structural classes: all-α, all-β, α/β, and $\alpha + \beta$ (1). The all-α and all-β proteins are essentially formed by α-helices (Fig. 1a) and β-sheets (Fig. 1b), respectively. The α/β class represents those proteins in which α-helices and β-strands are largely interspersed with the main sheet consisting mainly of parallel strands (Fig. 1c), while the $\alpha + \beta$ class represents those in which α-helices and β-strands are largely segregated with the β-sheets almost always built up from antiparallel strands (Fig. 1d). Moreover, some proteins, the so-called ζ proteins (2), are highly irregular that contain very little or no α-helices and β-sheets at all.

These class definitions clearly describe the underlying architecture of a protein's structure, and hence have been generally accepted and are still in common use today. This represents that the degree of degeneracy between protein sequences and structural classes is extremely high. On the other hand, a high degeneracy also exists between protein sequences and amino acid compositions because a same amino acid composition can be derived from many different amino acid sequences. Thus, the following questions are naturally raised. (1) Is there a correlation between protein structural classes and amino acid compositions so that the prediction of protein structural classes can be significantly simplified? (2) If yes, what is the physical mechanism underlying this?

For the first question, many encouraging results about the structural class prediction based on amino acid composition alone have been reported during the past two decades (3–17). All these results have demonstrated that some correlation between the protein structural class and amino acid composition does exist, and the former can be recognized from the latter at least to some approximate degree. The second question is more essential, but the answer is not clear yet, and the test not so straightforward either. However, according to the logic in physics, the existence of the above correlation would imply that the interaction among the components of amino acid composition might play a dominant role in determining the structural class of a protein during the sequence-folding process. In other words, although the detailed 3-D structure (in the level of atomic coordinates) of an

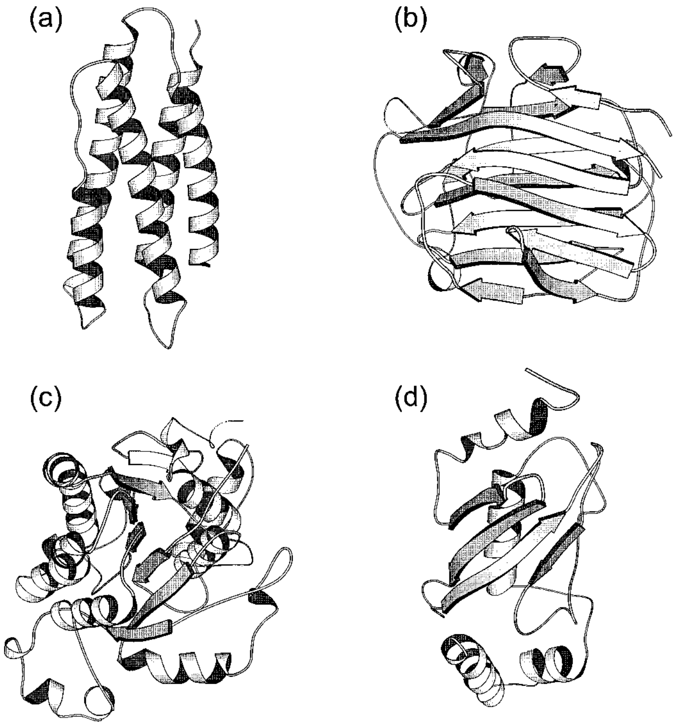

FIG. 1. Ribbon drawings to show proteins representative of the four structural classes: (a) all-α, (b) all-β, (c) α/β, and (d) $\alpha + \beta$. The PDB codes of the four proteins are 1aep, 1gbg, 1enp, and 1aak, respectively.

isolated protein is dictated by the amino acid interactions over the entire sequence chain, its structural class might be determined to a considerable degree by the interactions among 20 constituent amino acid components.

If the above rationale is correct, a potential function that can more completely cover the interaction of amino acid components will yield a better recognition quality for structural classes; in contrast, a potential function covering such an interaction in a less complete way will yield a poorer recognition quality.

The present study is devoted to test such a rationale, so as to gain some insight into the physical mechanism about the correlation between the protein structural class and amino acid composition.

A NEW PARADIGMATIC DATASET

The dataset originally studied by Levitt and Chothia (1) was the first structural dataset consisting of only 31 proteins that were classified completely based on a visual inspection. In order to develop a statistical method for studying protein structural classes, a dataset of much more than 31 proteins must be constructed. Thus, various quantitative classification rules were proposed based on the percentages of α-helices and β-sheets in a protein (see, e.g., Refs. 3–5, 7, 12, and 18). The introduction of these quantitative rules has stimulated the development of protein structural class prediction. But on the other hand, the structural classification solely based on the percentages of α-helices and β-sheets could hardly be without arbitrariness and hence would go short of objectivity; i.e., according to different rules, a same protein could be assigned to completely different classes. This would often cause confusion or lead to a wrong conclusion, especially when using these rules to classify more and more proteins in the Brookhaven Protein Databank (19), as recently elaborated by Chou *et al.* (15) and Zhou (16).

To avoid this kind of arbitrariness, instead of using the percentages of secondary structure as a sole criterion to classify the classes of proteins, Murzin *et al.* (20) proposed a method which is based upon the evolutionary relationships of proteins and on the principles that govern their 3-D structure. Small proteins, and most of those with medium size, have a single domain and are, therefore, treated as a whole. The domains in large proteins are usually classified individually. The database thus constructed is called SCOP (structural classification of proteins). In addition to the information of structural classes, SCOP (20) also provides a detailed and comprehensive description of the structural and evolutionary relationships of proteins whose 3-D structures have been determined. Therefore, in comparison with the other classifications only based on the percentages of secondary structures, the classification in SCOP is more natural, better reflects the objective reality, and provides a more reliable database for the study of protein structural class prediction. However, since so far no powerful method is available to predict the domain region(s) for a given protein sequence, the application of the SCOP database for many practical problems is often of limit; e.g., when using the knowledge of structural class of a protein to improve the prediction of its secondary structure contents (see, e.g., (21, 22)).

In view of this, rather than a protein domain, here let us still use the entire protein chain as a basic unit for the structural class classification. However, rather than an arbitrary criterion, the classification is made according to a rule derived from a learning process, as illustrated below. From some very large datasets in SCOP database, e.g., Tables 10 and 11 of (17), we can extract 2,540 PDB codes that all represent a whole protein chain and are well defined in SCOP as one of the all-α, all-β, α/β, and $\alpha + \beta$ structural classes. The ζ proteins are left out here for further consideration because their number is too small to have any statistical significance. Besides, the purpose of this study is not to provide a complete dataset of structural classes since it is apparently too premature to do so now. The current

TABLE 1

The Average (Mean ± Standard Deviation) Length of Protein Chains and Their Percentages of α-Helices, β-Strands, Parallel β-Sheets, and Antiparallel β-Sheets Derived from the SCOP Database for Each of the Four Structural Classes

Class	Length[a]	α-Helices	β-Strands	Parallel β-sheets	Antiparallel β-sheets
All-α	150 ± 92	58 ± 15%	2 ± 4%	—	—
All-β	201 ± 93	7 ± 6%	39 ± 11%	—	—
α/β	302 ± 135	36 ± 8%	18 ± 6%	66 ± 23%	34 ± 23%
$\alpha + \beta$	143 ± 54	36 ± 20%	18 ± 11%	9 ± 19%	91 ± 20%

[a] The length of a protein chain is defined by the total number of residues it contains.

study was devoted to explore the physical mechanism through a new paradigmatic working dataset that shall satisfy the following conditions: (1) Any protein in the new dataset must, as a whole, clearly and unambiguously belong to one of the four structural classes. (2) Each subset in the dataset must contain a statistically significant number of proteins that belong to a same structural class. To establish such a dataset, as a first step, the average (mean ± standard deviation) length and secondary structure content percentages for each of the four classes were derive from the 2,540 protein chains and their DSSP files (23). These results are listed in Table 1. Secondly, within the length and percentage scopes as prescribed in Table 1 for each class, 204 non-homologous proteins were extracted from the Brookhaven Protein Databank (19) as the representatives of the four structural classes. Of the 204 proteins, 52 are all-α proteins, 61 all-β, 45 α/β, and 46 $\alpha + \beta$ (Table 2). The 204 proteins of Table 2 form a new working dataset which will be used to study the physical mechanism about the correlation between the structural class and amino acid composition.

AN APPROACH TO THE PHYSICAL MECHANISM

Through what kind of concrete methods can we observe the correlation between the structural class and amino acid composition, so as to gain some useful insight into the physical mechanism therein? Actually, the methods have already existed, and what we need to do here is just to systematically conceptualize them according to the rationale described in the introduction, as illustrated below.

Suppose there are N proteins forming a set S, which is the union of four subsets;

$$S = S_{\alpha} \cup S_{\beta} \cup S_{\alpha/\beta} \cup S_{\alpha+\beta}. \qquad [1]$$

Each subset is composed of proteins with a same structural class. Its size is given by N_{ξ} ($\xi = \alpha, \beta, \alpha/\beta, \alpha + \beta$), where N_{ξ} represents the number of proteins in the subset S_{ξ}. Obviously, $N = N_{\alpha} + N_{\beta} + N_{\alpha/\beta} + N_{\alpha+\beta}$. Thus, any protein in the set S will correspond to a vector or a point in the 20-D amino acid composition space; i.e., it can be described by

$$\mathbf{X}_k^{\xi} = \begin{bmatrix} x_{k,1}^{\xi} \\ x_{k,2}^{\xi} \\ \cdot \\ \cdot \\ \cdot \\ x_{k,20}^{\xi} \end{bmatrix} \quad (k = 1, 2, \ldots, N_{\xi};$$

$$\xi = \alpha, \beta, \alpha/\beta, \alpha + \beta), \qquad [2]$$

where the components $x_{k,1}^{\xi}, x_{k,2}^{\xi}, \ldots, x_{k,20}^{\xi}$ are the normalized occurrence frequencies of the 20 amino acids for the kth protein $\mathbf{X}_k^{\xi}$ in the subset S_{ξ}. Without loss of generality, all the components in this paper follow the alphabetical order of the single-letter codes for the 20 amino acids: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y. The *standard vector* for the subset S_{ξ} is defined by

$$\bar{\mathbf{X}}^{\xi} = \begin{bmatrix} \bar{x}_1^{\xi} \\ \bar{x}_2^{\xi} \\ \cdot \\ \cdot \\ \cdot \\ \bar{x}_{20}^{\xi} \end{bmatrix} \quad (\xi = \alpha, \beta, \alpha/\beta, \alpha + \beta), \qquad [3]$$

where

$$\bar{x}_i^{\xi} = \frac{1}{N_{\xi}} \sum_{k=1}^{N_{\xi}} x_{k,i}^{\xi} \quad (i = 1, 2, \ldots, 20). \qquad [4]$$

Suppose $\mathbf{X}$ is a protein whose structural class is to be recognized. It can be either one of the N proteins in the set S, or a protein outside it. It also corresponds to a point $(x_1, x_2, \ldots, x_{20})$ in the 20-D space, where x_i has the same meaning as $x_{k,i}^{\xi}$ but is associated with protein $\mathbf{X}$ instead of $\mathbf{X}_k^{\xi}$. Now the essential problem is how to effectively define the potential of a query protein $\mathbf{X}$ in the 20-D composition space with respect to the four

TABLE 2

The PDB Codes of the 204 Protein Chains Classified According to the Criteria of Table 1

(1) 52 α-proteins									
1aep_	1ash_	1bcfA	1cnt1	1gdy_	1hlb_	1ilk_	1maz_	1mls_	1rhgA
1spgB	1sra_	1vls_	2fal_	2hbg_	3sdhA	1allA	1flp_	1ibeA	1ithA
2gdm_	2lhb_	1hdsB	1myt_	1osa_	1sctA	1spgA	1fslA	1hlm_	1lht_
1outA	1outB	1pbxA	1pbxB	1sctB	1babB	2asr_	1babA	1bgc_	1bgeA
1emy_	1hdaB	1hdsA	1ibeB	1mbs_	2mm1_	2pghA	2pghB	1hdaA	1hrm_
1mygA	1vlk_								
(2) 61 β-proteins									
1bbt2	1cfb_	1edhA	1gen_	1sacA	1tcrA	2ayh_	3hhrC	6fabL	8fabB
1pex_	1vcaA	1mfbL	1gnhA	1yna_	8fabA	1flrH	1ggiH	1indH	1JELH
2cgrH	7fabH	1bbdH	1eapA	1gafL	1gbg_	1ggiL	1ghfH	1hilB	1ncbL
1nldH	1opgL	1ospL	1vgeL	2fbjL	2mcg1	7fabL	1acyL	1bafL	1bjmA
1bqlH	1bqlL	1dfbL	1forL	1ghfL	1iaiL	1iaiM	1igcL	1ikfL	1indL
1macA	1mamL	1mreH	1ngqH	1nsnH	1plgH	1plgL	1tetH	1xnd_	1yuhA
3hfmH									
(3) 45 α/β proteins									
1amp_	1ceo_	1cvl_	1dorA	1gca_	1ghr_	1gym_	1lbiA	1lucA	1masA
1nar_	1pbn_	1pfkA	1sbp_	1scuA	1thtA	1vdc_	1vpt_	1xel_	1xyzA
2bgu_	2ctc_	2ebn_	3pga1	8abp_	1enp_	1gdhA	1lucB	1obr_	1cnv_
1exp_	1trb_	1ghsA	1hdg0	1lwiA	1wsaA	2alr_	3ecaA	4pfk_	1agx_
1cer0	1gia_	2lip_	1ula_	2gbp_					
(4) 46 $\alpha + \beta$ proteins									
1aak_	1afb1	1bplA	1cof_	1cyw_	1def_	1doi_	1epaB	1fil_	1grj_
1gtqA	1hjrA	1htp_	1ino_	1itg_	1lit_	1mkaA	1msc_	1nhkL	1pkp_
1poc_	1rbu_	1seiA	1sfe_	1snc_	1std_	1tfe_	1vhh_	1vhiA	1vsd_
1whtB	1ytbA	2tbd_	8atcB	1apyB	1div_	1pvuA	1npk_	2uce_	1ril_
2prd_	1hup_	1nueA	1cdwA	1pne_	2kmb1				

Note. The fifth character in the PDB code indicates a specific chain of the protein; if it is _, the corresponding protein has only one chain.

structural classes. Actually, potential functions with different approximate orders are hidden in some existing algorithms, as can be formulated according to the following three categories.

1. The 0th-Order Component-Coupled Algorithm

According to the 0th-order component-coupled algorithm, the potential function of a query protein in the 20-D composition space is given by

$$U^{(0)}(\mathbf{X}, \bar{\mathbf{X}}^{\xi}) = k^{(0)} \sum_{i=1}^{20} (x_i - \bar{x}_i^{\xi})^2 \qquad (\xi = \alpha,\ \beta,\ \alpha/\beta,\ \alpha + \beta), \quad [5]$$

where $k^{(0)}$ is a force constant, which is trivial here and can be left out in calculation because it is the same for all the structural classes. As we know from a basic law in physics, a system will become the most stable when it is in a state with the lowest potential, or strictly speaking, the lowest free energy. Accordingly, if $U^{(0)}(\mathbf{X}, \bar{\mathbf{X}}^{\alpha})$ is the smallest among $U^{(0)}(\mathbf{X}, \bar{\mathbf{X}}^{\xi})$ ($\xi = \alpha$, β, α/β, and $\alpha + \beta$), the protein $\mathbf{X}$ will fold into the all-α structural class; if $U^{(0)}(\mathbf{X}, \bar{\mathbf{X}}^{\beta})$ the smallest, then it will fold into the all-β class; and so forth. Therefore, the recognition rule should be formulated as

$$U^{0}(\mathbf{X}, \bar{\mathbf{X}}^{\ell}) = \mathbf{Min}\{U^{0}(\mathbf{X}, \bar{\mathbf{X}}^{\alpha}),\ U^{0}(\mathbf{X}, \bar{\mathbf{X}}^{\beta}),\ U^{0}(\mathbf{X}, \bar{\mathbf{X}}^{\alpha/\beta}),\ U^{0}(\mathbf{X}, \bar{\mathbf{X}}^{\alpha+\beta})\}, \quad [6]$$

where ℓ can be α, β, α/β, or $\alpha + \beta$, and the operator **Min** means taking the least one among those in the parentheses, and the superscript ℓ represents the very structural class which the protein $\mathbf{X}$ belongs to. If there is a tie case, ℓ is not uniquely determined, but that rarely occurs. As we can see from Eq. [5], after leaving out the constant $k^{(0)}$, $U^{0}(\mathbf{X}, \bar{\mathbf{X}}^{\ell})$ actually represents the squared Euclidean distance between $\mathbf{X}$ and $\bar{\mathbf{X}}^{\xi}$ as used by Nakashima *et al.* (5) for recognizing the protein structural class. The 20 amino acid components in Eq.

[5] are independent of one another, and hence the formulation thus obtained represents the 0th-order component-coupled algorithm.

2. The 1st-Order Component-Coupled Algorithm

Instead of Eq. [5], the potential function in the 1st-order component-coupled algorithm is given by

$$U^{(1)}(\mathbf{X}, \bar{\mathbf{X}}^{\xi}) = k^{(1)}(\mathbf{X} - \bar{\mathbf{X}}^{\xi})^{\mathbf{T}}\mathbf{C}_{\xi}^{-1}(\mathbf{X} - \bar{\mathbf{X}}^{\xi}) \quad (\xi = \alpha, \beta, \alpha/\beta, \alpha+\beta), \quad [7]$$

where $k^{(1)}$ is the force constant for the 1st-order component-coupled potential, $\mathbf{C}_{\xi}$ is the covariance matrix for subset S_{ξ} as defined by

$$\mathbf{C}_{\xi} = \begin{bmatrix} c^{\xi}_{1,1} & c^{\xi}_{1,2} & \cdots & c^{\xi}_{1,20} \\ c^{\xi}_{2,1} & c^{\xi}_{2,2} & \cdots & c^{\xi}_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ c^{\xi}_{20,1} & c^{\xi}_{20,2} & \cdots & c^{\xi}_{20,20} \end{bmatrix}, \quad [8]$$

the superscript $\mathbf{T}$ is the transposition operator, and $\mathbf{C}_{\xi}^{-1}$ is the inverse matrix of $\mathbf{C}_{\xi}$. The matrix elements $c^{\xi}_{i,j}$ in Eq. [8] are given by

$$c^{\xi}_{i,j} = \frac{1}{N_{\xi} - 1}\sum_{k=1}^{N_{\xi}} [x^{\xi}_{k,i} - x^{\xi}_{i}][x^{\xi}_{k,j} - x^{\xi}_{j}] \quad (i, j = 1, 2, \ldots, 20). \quad [9]$$

And the recognition rule is formulated by

$$U^{(1)}(\mathbf{X}, \bar{\mathbf{X}}^{\ell}) = \mathbf{Min}\{U^{(1)}(\mathbf{X}, \bar{\mathbf{X}}^{\alpha}), U^{(1)}(\mathbf{X}, \bar{\mathbf{X}}^{\beta}), U^{(1)}(\mathbf{X}, \bar{\mathbf{X}}^{\alpha/\beta}), U^{(1)}(\mathbf{X}, \bar{\mathbf{X}}^{\alpha+\beta})\}. \quad [10]$$

For the same reason as discussed for Eq. [5], the force constant $k^{(1)}$ can also be left out during calculation. After doing so, the function of $U^{1}(\mathbf{X}, \bar{\mathbf{X}}^{\xi})$ in Eq. [7] is actually reduced to the squared Mahalanobis distance between $\mathbf{X}$ and $\bar{\mathbf{X}}^{\xi}$ as introduced by Chou and Zhang (10) for structural class prediction. Three years later, it was also used by Bahar *et al.* (14) as a key parameter in performing the singular value decomposition (SVD) analysis for recognizing the protein structural class. Note that the Mahalanobis distance is unit-independent, i.e., its value will not be changed by using different units of coordinates. As we can see from Eqs. [7]–[9], the 20 amino acid components in the function $U^{(1)}(\mathbf{X}, \bar{\mathbf{X}}^{\xi})$ are coupled with one another, and hence the recognition rule based on Eq. [7] represents the 1st-order component-coupled algorithm.

Note that because the amino acid composition must be normalized, i.e., constrained by

$$\sum_{i=1}^{20} x^{\xi}_{k,i} = 1 \quad (k = 1, 2, \ldots, N_{\xi}; \quad \xi = \alpha, \beta, \alpha/\beta, \alpha+\beta), \quad [11]$$

$\mathbf{C}_{\xi}$ defined by Eq. [9] is a singular matrix, and its inverse matrix $\mathbf{C}_{\xi}^{-1}$ must be of divergence and meaninglessness. In order to avoid the divergence difficulty of $\mathbf{C}_{\xi}^{-1}$, the function $U^{(1)}$ was originally defined in a 19-D space (10, 12) rather than 20-D space. However, such a difficulty can also be overcome through the following eigenvalue–eigenvector approach. Suppose

$$\mathbf{C}_{\xi}\mathbf{\Psi}^{\xi}_{i} = \lambda^{\xi}_{i}\mathbf{\Psi}^{\xi}_{i} = \lambda^{\xi}_{i}\begin{bmatrix} \psi^{\xi}_{i,1} \\ \psi^{\xi}_{i,2} \\ \vdots \\ \psi^{\xi}_{i,20} \end{bmatrix} \quad (i = 1, 2, \ldots, 20; \quad \xi = \alpha, \beta, \alpha/\beta, \alpha+\beta), \quad [12]$$

where λ^{ξ}_{i} is the ith eigenvalue of $\mathbf{C}_{\xi}$, and $\psi^{\xi}_{i,j}$ the jth component of the corresponding eigenvector $\mathbf{\Psi}^{\xi}_{i}$. It can be proved that for the covariance matrix $\mathbf{C}_{\xi}$ as defined by Eq. [9], there is no negative eigenvalue. Actually, because of Eq. [11], $\mathbf{C}_{\xi}$ must have one eigenvalue, denoted by λ^{ξ}_{1}, equal to zero (24). Besides, since proteins in a practical training dataset are always formed by those without high sequence similarity to one another, all the other 19 eigenvalues $\lambda^{\xi}_{2}, \lambda^{\xi}_{3}, \ldots, \lambda^{\xi}_{20}$ are generally greater than zero, and hence Eq. [7] can be converted to (24)

$$U^{(1)}(\mathbf{X}, \bar{\mathbf{X}}^{\xi}) = k^{(1)}\sum_{i=2}^{20}\frac{1}{\lambda^{\xi}_{i}}\Big[\sum_{j=1}^{20}(x_j - \bar{x}^{\xi}_{j})\psi^{\xi}_{i,j}\Big]^2, \quad [13]$$

for which the divergence difficulty no longer exists.

3. The 2nd-Order Component-Coupled Algorithm

In the above algorithm, although the component-coupled effects are incorporated through a set of covariant matrices (Eq. [8]), it omits an important term in further reflecting the difference among these covariant matrices for different classes. Therefore, it is only a 1st-order approximation. When the interactions among different amino acid components are extended to cover the 2nd-order component-coupled effect, instead of Eq. [7] or [13] we should have

TABLE 3

Rates of Correct Recognition in Structural Classes by the 0th-Order, 1st-Order, and 2nd-Order Component-Coupled Algorithms for the 204 Proteins of Table 2

Algorithm	Rate of correct recognition for each class				Overall rate of correct recognition
	α	β	α/β	$\alpha + \beta$	
0th-order component-coupled (Eq. [6])	39/52 = 75%	52/61 = 85%	25/45 = 56%	29/46 = 63%	145/204 = 71%
1st-order component-coupled (Eq. [10])	52/52 = 100%	60/61 = 98%	26/45 = 58%	45/46 = 98%	183/204 = 90%
2nd-order component-coupled (Eq. [15])	52/52 = 100%	61/61 = 100%	45/45 = 100%	46/46 = 100%	204/204 = 100%

$$U^{(2)}(\mathbf{X}, \bar{\mathbf{X}}^{\xi}) = k^{(2)}\left\{\sum_{i=2}^{20}\frac{1}{\lambda_i^{\xi}}\Big[\sum_{j=1}^{20}(x_j-\bar{x}_j^{\xi})\psi_{i,j}^{\xi}\Big]^2 + \ln(\lambda_2^{\xi}\lambda_3^{\xi}\lambda_4^{\xi}\cdots\lambda_{20}^{\xi})\right\} + \upsilon_0, \quad [14]$$

where $k^{(2)}$ is the force constant for the 2nd-order component-coupled potential, and υ_0 a potential constant. Accordingly, the recognition rule should be formulated as

$$U^{(2)}(\mathbf{X}, \bar{\mathbf{X}}^{\ell}) = \mathbf{Min}\{U^{(2)}(\mathbf{X}, \bar{\mathbf{X}}^{\alpha}),\ U^{(2)}(\mathbf{X}, \bar{\mathbf{X}}^{\beta}),\ U^{(2)}(\mathbf{X}, \bar{\mathbf{X}}^{\alpha/\beta}),\ U^{(2)}(\mathbf{X}, \bar{\mathbf{X}}^{\alpha+\beta})\}. \quad [15]$$

It can be easily proved that using different units of coordinates will not change the size order among $U^{(2)}(\mathbf{X}, \bar{\mathbf{X}}^{\xi})$ ($\xi = \alpha, \beta, \alpha/\beta, \alpha + \beta$), and hence the results recognized by Eq. [15] is unit-independent. Again, during calculation, the constants $k^{(2)}$ and υ_0 can be left out according to the same reason as addressed for Eq. [5]. After that, Eq. [14] will be reduced to the covariant-discriminant function (15, 16, 18).

As we can see, when all the covariant matrices $\mathbf{C}_{\xi}$ are an unitary matrix (i.e., all the diagonal elements equal to one, and all the non-diagonal elements equal to zero), both Eq. [7] and Eq. [14] will be reduced to Eq. [5], implying that no coupling effect exists among different amino acid components and that both the 1st- and 2nd-order component-coupled algorithms can be reduced to the 0th-order one. If the matrices $\mathbf{C}_{\xi}$ are not unitary but are all identical, the term $\ln(\lambda_2^{\xi}\lambda_3^{\xi}\lambda_4^{\xi} \ldots \lambda_{20}^{\xi})$ in Eq. [14] must be the same for different subsets ($\xi = \alpha, \beta, \alpha/\beta, \alpha + \beta$), then the results obtained by Eq. [15] will be the same as those by Eq. [10], and hence the

TABLE 4

The Standard Vector and Eigenvalues Derived from Table 2 for Each of the Four Protein Structural Classes

Amino acid code	Standard vector Components (normalized to 1)				Order i	Eigenvalues ($\times 10^5$)			
	$\bar{\mathbf{X}}^{\alpha}$	$\bar{\mathbf{X}}^{\beta}$	$\bar{\mathbf{X}}^{\alpha/\beta}$	$\bar{\mathbf{X}}^{\alpha+\beta}$		λ_i^{α}	λ_i^{β}	$\lambda_i^{\alpha/\beta}$	$\lambda_i^{\alpha+\beta}$
A	0.114	0.057	0.093	0.084	1	0	0	0	0
C	0.007	0.019	0.010	0.016	2	0.4	0.3	0.2	0.6
D	0.055	0.045	0.062	0.056	3	2.5	0.8	1.4	3.8
E	0.059	0.044	0.054	0.069	4	3.0	1.8	2.0	4.9
F	0.048	0.036	0.039	0.039	5	5.6	2.3	2.7	9.6
G	0.069	0.074	0.078	0.076	6	8.0	2.6	4.7	10.6
H	0.039	0.015	0.022	0.024	7	9.8	3.2	4.9	14.1
I	0.040	0.038	0.061	0.060	8	11.3	4.9	8.0	14.7
K	0.077	0.054	0.059	0.073	9	17.0	5.7	10.9	17.9
L	0.113	0.066	0.083	0.082	10	20.7	6.9	11.4	23.4
M	0.024	0.012	0.022	0.021	11	25.0	7.7	14.0	27.5
N	0.038	0.043	0.054	0.043	12	26.4	10.0	18.7	29.6
P	0.034	0.056	0.041	0.044	13	37.9	14.0	20.1	36.1
Q	0.041	0.041	0.036	0.032	14	45.5	16.3	25.4	51.3
R	0.034	0.030	0.039	0.048	15	59.9	21.6	32.7	59.4
S	0.058	0.128	0.060	0.056	16	71.9	30.7	42.4	69.5
T	0.046	0.096	0.061	0.053	17	90.9	34.9	51.7	94.7
V	0.069	0.073	0.075	0.076	18	148.5	71.2	78.8	100.2
W	0.013	0.022	0.011	0.014	19	180.6	96.0	85.5	113.9
Y	0.022	0.049	0.038	0.034	20	263.9	170.5	161.1	162.2

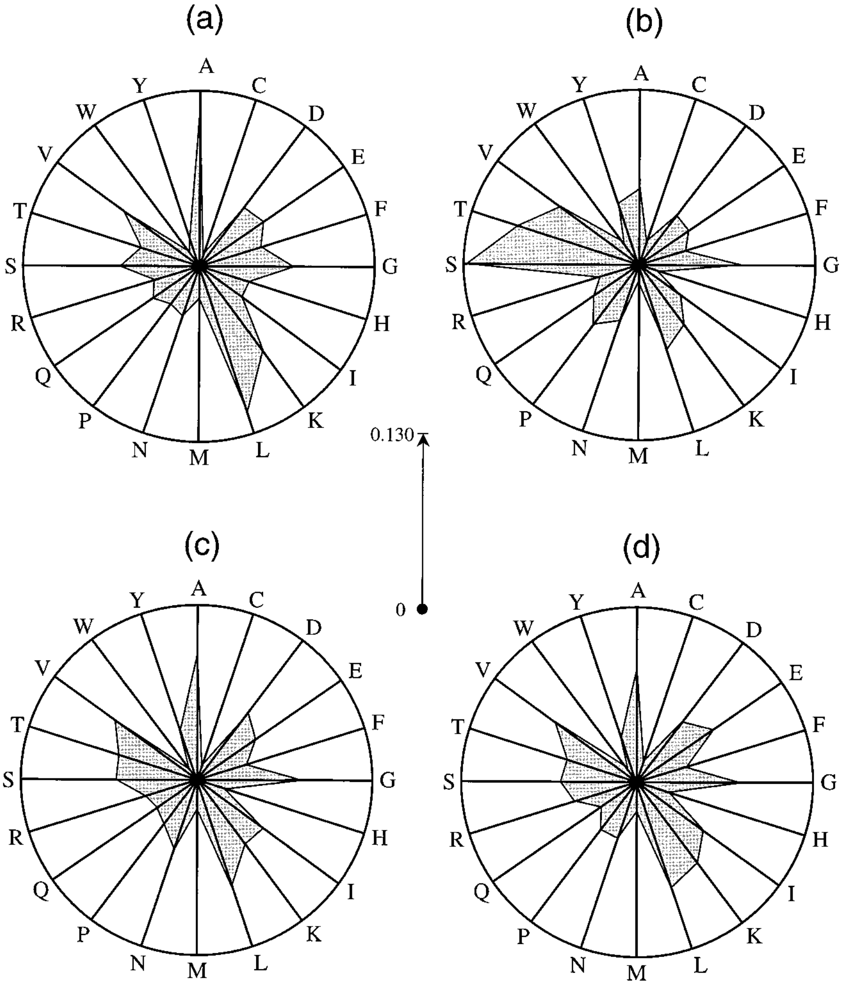

FIG. 2. Radar diagrams to show the distinction of the 20-D standard vectors, i.e., the average amino acid compositions for the proteins in the following structural class subsets: (1) all-α, (2) all-β, (3) α/β, and (4) $\alpha + \beta$. Amino acids are denoted by their single-letter codes (see Table 4).

2nd-order component-coupled algorithm can be reduced to the 1st-order component-coupled algorithm.

RESULTS AND DISCUSSION

The rates of correct recognition for the 204 proteins in Table 2 by the 0th-order, 1st-order, and 2nd-order component-coupled algorithms are given in Table 3, from which the following can be observed.

1. The overall rates of correct recognition by the 0th-order, 1st-order, and 2nd-order component-coupled algorithms are in the range from 71 to 100%. According to the probability theory, if the samples of proteins are completely randomly assigned among four possible subsets, the rate of correct assignment would generally be $\frac{1}{4} = 25.0\%$. If, however, the random assignment is weighted according to the sizes of subsets, then the rate of correct assignment would be $p_\alpha^2 + p_\beta^2 + p_{\alpha/\beta}^2 + p_{\alpha+\beta}^2$, where $p_\alpha = N_\alpha/N$, $p_\beta = N_\beta/N$, and so forth. Substituting the subset sizes (see Table 2) into the above equation, we obtain the rate of correct assignment by the weighted random assignment is $(52/204)^2 + (61/204)^2 + (45/204)^2 + (46/204)^2 \approx 25.4\%$. Therefore, the rate of correct recognition by any of the three algorithms is significantly higher than the corresponding completely randomized rate and weighted randomized rate, implying that the structural class of a protein is considerably correlated with its amino acid composition.

2. The overall rate of correct recognition by the 2nd-order component-coupled algorithm is 100%, which is 10% higher than that by the 1st-order component-coupled algorithm and about 30% higher than that by the 0st-order component-coupled algorithm. This indi-

cates that the interaction among the components of amino acid composition does play an important role in determining the structural class of a protein during its folding process. Therefore, if an algorithm can more accurately encompass such an interaction, its power in recognizing the structural class of a protein will be higher; and vice versa. To show the difference in amino acid compositions that distinguish the structural classes of proteins, the 20-D standard vector (see Eq. [3]) derived from the proteins in Table 2 for each of the four structural classes is given in Table 4. Meanwhile, to provide an intuitive picture, each such 20-D standard vector is projected onto a 2-D radar diagram as given in Fig. 2. Furthermore, to show the difference of the covariant matrices for different classes, the eigenvalue set (see Eq. [12]) for each of these matrices are also given in Table 4. From these eigenvalues, the 2nd-order component-coupled term of Eq. [14] can be calculated. Only when such a term has the same value for all the structural classes, will the results recognized by the 2nd-order component-coupled algorithm (Eq. [15]) be the same as those by the 1st-order component-coupled algorithm (Eq. [10]).

3. It is instructive to introduce a new concept called the *cluster-tolerant capacity* for the dataset studied here. If such a capacity is high for a dataset, then the removal of any entry from the dataset will not significantly change the original clustered picture (such as the distribution of the standard vectors for each subset, the spacial gaps between the boundaries of any two subsets, and the original attribution of the removed entry to a subset); conversely, if the clustered tolerant capacity is low for a dataset, the removal of some entry from it will have a significant impact on the original clustered picture. A quantity that is associated with the cluster-tolerant capacity is τ, the overall rate of correct recognition obtained by a jackknife analysis. During the analysis process, each of the N proteins in Table 2 was singled out in turn as a "query protein" for recognition and all the rule-parameters were determined from the remaining $N - 1$ proteins. Therefore, although τ is a scale used mainly for measuring the cluster-tolerant capacity of a dataset, it is also, to some degree, depend on different recognition algorithms. In other words, for a same dataset distorted by jackknifing, some algorithms may give higher τ values than the others. For the dataset of Table 2, the value of τ by the 2nd-order component-coupled algorithm is 77%, about 11 and 8% higher than those by the 1st-order and 0th-order component-coupled algorithms. This implies that the 2nd-order component-coupled algorithm is more powerful in recognizing the protein structural classes not only for an original dataset but also for a jackknife-distorted dataset, fully consistent with our hypothesis and rationale.

CONCLUSION

It is demonstrated that the interactions among the components of amino acid composition is an important driving force during the sequence folding process for finally determining the structural class of a protein.

The paradigmatic dataset mined according to a new rule derived from a learning process is featured by 100% rate of correct recognition, and hence can become a better base for understanding the recognition of protein structural classes by amino acid composition from other approaches as well, such as the one recently performed by Bahar *et al.* (14). The remarkable dataset can also be used as a powerful vehicle for the effort of improving the prediction of proten secondary structure contents by means of the knowledge of structural class, as pursued by Zhang *et al.* (21, 22).

ACKNOWLEDGMENTS

The author thanks Raymond B. Moeller, Cynthia A. Ludlow, and Wendy Vanderheide for their help with drawing Figs. 2 and 3.

REFERENCES

1. Levitt, M., and Chothia, C. (1976) *Nature* **261,** 552–557.
2. Chou, J. J., and Zhang, C. T. (1993) *J. Theor. Biol.* **161,** 251–262.
3. Chou, P. Y. (1980) *in* Abstracts of Papers, Part I, Second Chemical Congress of the North American Continent, Las Vegas.
4. Chou, P. Y. (1989) *in* Prediction of Protein Structure and the Principles of Protein Conformation (Fasman, G. D., Ed.), pp. 549–586, Plenum, New York.
5. Nakashima, H., Nishikawa, K., and Ooi, T. (1986) *J. Biochem.* **99,** 152–162.
6. Klein, P. (1986) *Biochim. Biophys. Acta* **874,** 205–215.
7. Klein, P., and Delisi, C. (1986) *Biopolymers* **25,** 1659–1672.
8. Metfessel, B. A., Saurugger, P. N., Connelly, D. P., and Rich, S. T. (1993) *Protein Sci.* **2,** 1171–1182.
9. Dubchak, I., Holbrook, S. R., and Kim, S-H. (1993) *Proteins: Struct. Funct. Genet.* **16,** 79–91.
10. Chou, K. C., and Zhang, C. T. (1994) *J. Biol. Chem.* **269,** 22014–22020.
11. Mao, B., Chou, K. C., and Zhang, C. T. (1994) *Protein Eng.* **7,** 319–330.
12. Chou, K. C. (1995) *Proteins Struct. Funct. Genet.* **21,** 319–344.
13. Chandonia, J. M., and Karplus, M. (1995) *Protein Sci.* **4,** 275–285.
14. Bahar, I., Atilgan, A. R., Jernigan, R. L., and Erman, B. (1997) *Proteins* **29,** 172–185.
15. Chou, K. C., Liu, W., Maggiora, G. M., and Zhang, C. T. (1998) *Proteins Struct. Funct. Genet.* **31,** 97–103.
16. Zhou, G. P. (1998) *J. Protein Chem.* **17,** 729–738.
17. Chou, K. C., and Maggiora, G. M. (1998) *Proteins Eng.* **11,** 523–538.
18. Liu, W., and Chou, K. C. (1998) *J. Protein Chem.* **17,** 209–217.
19. Abola, E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F., and Weng, J. (1987) *in* Crystallographic Databases—Information Contents, Software Systems, Scientific Applications (Allen, F. H., Bergerhoff, G., and Sierers, R., Eds.), pp. 107–132, Com-

mission of International Union of Crystallography, Bonn, Cambridge, Chester.

20. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) *JMB* **247,** 536–540.

21. Zhang, C. T., Zhang, Z., and He, Z. (1996) *J. Protein Chem.* **15,** 775–786.

22. Zhang, C. T., Zhang, Z., and He, Z. (1998) *J. Protein Chem.* **17,** 261–272.

23. Kabsch, W., and Sander, C. (1983) *Biopolymers* **22,** 2577–2637.

24. Chou, K. C., and Zhang, C. T. (1995) *Crit. Rev. Biochem. Mol. Biol.* **30,** 275–349.