PROTEINS: Structure, Function, and Genetics 38:165–175 (2000)

How Good Is Prediction of Protein Structural Class by the Component-Coupled Method?

Zhi-Xin Wang* and Zheng Yuan
National Laboratory of Biomacromolecules, Institute of Biophysics, Academia Sinica, Peoples Republic of China

ABSTRACT **Proteins of known structures are usually classified into four structural classes: all-α, all-β, α+β, and α/β type of proteins. A number of methods to predicting the structural class of a protein based on its amino acid composition have been developed during the past few years. Recently, a component-coupled method was developed for predicting protein structural class according to amino acid composition. This method is based on the least Mahalanobis distance principle, and yields much better predicted results in comparison with the previous methods. However, the success rates reported for structural class prediction by different investigators are contradictory. The highest reported accuracies by this method are near 100%, but the lowest one is only about 60%. The goal of this study is to resolve this paradox and to determine the possible upper limit of prediction rate for structural classes. In this paper, based on the normality assumption and the Bayes decision rule for minimum error, a new method is proposed for predicting the structural class of a protein according to its amino acid composition. The detailed theoretical analysis indicates that if the four protein folding classes are governed by the normal distributions, the present method will yield the optimum predictive result in a statistical sense. A non-redundant data set of 1,189 protein domains is used to evaluate the performance of the new method. Our results demonstrate that 60% correctness is the upper limit for a 4-type class prediction from amino acid composition alone for an unknown query protein. The apparent relatively high accuracy level (more than 90%) attained in the previous studies was due to the preselection of test sets, which may not be adequately representative of all unrelated proteins. Proteins 2000;38:165–175.**

Key words: SCOP database; Bayes decision rule; jack-knife analysis; amino acid composition; α domains; β domains; α + β domains; α/β domains.

INTRODUCTION

The protein structure prediction problem remains one of the most important problems in molecular biology. At present time, available methods have unable to meet in a satisfactory way. Protein can be considered as a hierarchy of structure: amino acid sequence → secondary structure → supersecondary structure → domain → three-dimensional structure. Many approaches to the protein-folding problem have reflected this hierarchical scheme, and in these secondary structure prediction is the first and most critical step in the achievement of correct prediction.

The concept of protein structural classes was originally introduced by Levitt and Chothia based on a visual inspection of polypeptide chain topologies in a data set of 31 globular proteins.[1] According to the contents of secondary structures, proteins of known structures can be classified into one of the following four structural classes: all α, all β, α + β, and α/β type of proteins.[1,2] These class definitions have been generally accepted, and are still in common use, although slight changes have been made.[3,4] It is well known that the knowledge of protein structural class can help in the determination of the three-dimensional structure of a protein, particularly in improving the prediction of secondary structure.[5,6] Therefore, it would be very useful if a rapid and reliable method could be developed to predict the structural class of a protein. During the past decade, various efforts have been made to reach such a goal by many investigators. On the basis of different criteria, such as the discriminant analysis,[7,8] the least Minkowski distance,[9,10] the least Euclidean distance,[11] the optimization approach principle,[12] or the maximum projection principle,[13] various prediction methods have been developed in these studies. Recently, a statistical method was developed for predicting protein structural class according to amino acid composition.[14] This method is based on the least Mahalanobis distance principle, and yields much better predicted results in comparison with the previous methods. However, the success rates reported for secondary structural class prediction by different investigators are contradictory. The highest reported accuracies obtained by this method are near 100%, but the lowest one is only about 60%.[15–20]

These studies have been difficult to compare to each other for several reasons. First, the protein sets used vary widely in both size and extent of sequence homology. For example, some of the sets of proteins on which these algorithms were tested contained high levels of sequence homology (more than 90% identity in some cases) with

Grant sponsor: The 863 High-Technology Foundation and Pandeng project of the Ministry of Science and Technology; Grant number: 103–13–03–02.

*Correspondence to: Zhi-Xin Wang, National Laboratory of Biomacromolecules, Institute of Biophysics, Academia Sinica, Beijing 100101, Peoples Republic of China. E-mail: zxwang@sun5.ibp.ac.cn

Received 23 April 1999; Accepted 11 August 1999

each other and with the proteins used in determining the numeric parameters of the algorithm. Second, the number structural classes in which to categorize the proteins has varied from study to study, ranging from three to seven structural classes. Third, criteria for dividing proteins into their structural classes has differed. For example, Nakashima et al. classified a protein as α type if the protein is known to contain greater than 15% α helices and less than 10% β strands.[11] Chou, on the other hand, requires their α class proteins to contain at least 40% α helices and not more than 5% β strands.[15] Fourth, cross-validation methods have varied widely, ranging from leave-one-out methods to single-test-set methods.

In the first part of this paper, according to the normality assumption and the Bayes decision rule for minimum error, we propose a new approach for predicting the structural class of a protein from its amino acid composition. A detailed theoretical analysis indicates that the least Mahalanobis distance method developed previously can be viewed as the approximation to the present method if the four protein structural classes are governed by the normal distributions. The question of why different conclusions were obtained from different studies is addressed in the next section by constructing an unbiased data set based on all unrelated protein domains. In particular, structures were taken from the Protein Data Bank and definitions of domains and structural classes were taken from the Structural Classification of Proteins (SCOP) database.[3] One of the most important aspects of our analysis is that we carefully tested the present method against this new data set. This testing allowed us to decide unambiguously whether a given comparison resulted in a true or false-positive and to decide objectively between different statistical schemes. Our results demonstrate that 60% correctness is the upper limit for a 4-type class prediction by the existing component-coupled method for an unknown non-homologous domain (sequence identity $<$ 30%). Therefore, further improvement of prediction accuracy will mainly depend on the introduction of other features (e.g., average hydrophobicity, net charge, amino acid sequence, and so on).

THEORY

A pattern is the description of an object. Almost anything that is within the reach of our five senses can be chosen as a pattern—a character, a photograph, speech pattern, odors, tastes, etc. A pattern class is a group of patterns with certain properties. Pattern recognition is that of classifying a pattern into one of the pattern classes on the grounds of certain measurement or properties. Depending on the problem of interest, the variations of the member of a pattern class can be deterministic (nonrandom) or random in nature. If the variations of the pattern from the stored reference, which is the ideal or average pattern, are random, the statistical pattern-recognition approach should be used. In this case, it is necessary then to describe such variations with a probabilistic quantity. The design of a statistical pattern recognition system is generally based on the Bayes classification rule and its variations. This rule yields an optimum classifier (decision procedure) when the probability density function of each pattern population and the probability of occurrence of each pattern class are known.[21–23]

Suppose there are M possible pattern classes, ω_1, ω_2, . . . , ω_M, and an arbitrary pattern belongs to class ω_l with a priori probability, $P(\omega_l)$, $l = 1, 2, \ldots, M$. Pattern or feature-vector, $\boldsymbol{x}$, are n-components, random vectors taking value in n-dimensional feature-space, $\boldsymbol{x}$, and governed by a multivariate conditional probability density function, $P(\boldsymbol{x} \mid \omega_l)$, when pattern $\boldsymbol{x}$ is known to belong to class ω_l. The recognition problem can now be viewed as that of generating the decision boundaries which separate the M pattern classes on the basis of the observed measurement vectors. Let the decision boundaries be defined by decision functions $d_1(\boldsymbol{x})$, $d_2(\boldsymbol{x})$, . . . , $d_M(\boldsymbol{x})$. These functions, which are also called discriminant functions, are scalar and single-valued functions of the pattern $\boldsymbol{x}$. If $d_l(\boldsymbol{x}) > d_u(\boldsymbol{x})$ for $u = 1, 2, \ldots, M$, and $l \neq u$, the pattern $\boldsymbol{x}$ belongs to ω_l. In other words, if the lth decision function, $d_l(\boldsymbol{x})$, has the largest value for a pattern $\boldsymbol{x}$, then $\boldsymbol{x} \in \omega_l$. A decision rule based simply on probabilities is to assigns a particular pattern $\boldsymbol{x}$ to class ω_l if

$$P(\omega_l \mid \boldsymbol{x}) > P(\omega_u \mid \boldsymbol{x}), \qquad u = 1, 2, 3, \ldots, M; \quad l \neq u \tag{1}$$

that is, the pattern class ω_l with the highest *posteriori* probability is chosen as the assignment for $\boldsymbol{x}$. The *a posteriori* probabilities $P(\omega_l \mid \boldsymbol{x})$ may be calculated from the a priori probabilities $P(\omega_l)$ and the conditional density functions $P(\boldsymbol{x} \mid \omega_l)$, using Bayes' theorem, that is

$$P(\omega_l \mid \boldsymbol{x}) = P(\omega_l)P(\boldsymbol{x} \mid \omega_l)/P(\boldsymbol{x}) \tag{2}$$

It can be verified that the average of classification error probability is minimized if the decision rule of Eq.(1) is used, and therefore it is also called the Bayes decision rule for minimum error. From the discussion on decision function given above, it is note that the decision functions for the Bayes decision rule for minimum error can be written as

$$d_l(\boldsymbol{x}) = P(\omega_l \mid \boldsymbol{x}) \tag{3}$$

or

$$d_l(\boldsymbol{x}) = P(\omega_l)P(\boldsymbol{x} \mid \omega_l) \tag{4}$$

where $P(\boldsymbol{x})$ has been eliminated since it does not depend on l. If all a priori probabilities are equal: $P(\omega_l) = 1/M$, for $l = 1, 2, \ldots, M$, the decision function can then be written as:

$$d_l(\boldsymbol{x}) = P(\boldsymbol{x} \mid \omega_l) \tag{5}$$

Equation (5) is called the (conditional) maximum-likelihood decision, and it can be regarded as the Bayes decision rule for minimum error with equal a priori probabilities, i.e., $P(\omega_l) = 1/M$, for $l = 1, 2, \ldots, M$.

Prediction of protein structural class from amino acid composition can be considered as a pattern recognition

problem. The amino acid composition of a protein molecule serves as the properties of the recognition system, which identifies the protein structural class by analysis of these properties. According to its amino acid composition, a protein molecule can be represented by a point or a vector in a 20-dimensional space, the so-called composition space. However, as pointed by Chou and Zhang,[16] of the 20 amino acid composition components only 19 are independent, since the amino acid composition of a protein must be constrained by

$$\sum_{i=1}^{20} x_i = 1 \tag{6}$$

where x_i is the composition component of the ith amino acid in a protein. Therefore, by leaving out any one of its 20 components, one can still uniquely represent a protein by a point in a 19-D space. Suppose the 20 amino acids are alphabetically ordered according to their single-letter code. If the last amino acid component is left out, then the 19-D space will be defined by the bases corresponding to the components of A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, and W, respectively. Once the 19-D space is established, the kth protein in a given set of N proteins can be expresses by

$$\boldsymbol{x}_k = \begin{bmatrix} x_{k,1} \\ x_{k,2} \\ \\ x_{k,19} \end{bmatrix} \qquad (k = 1, 2, \ldots, N) \tag{7}$$

where $x_{k,1}$, $x_{k,2,}$ . . . , $x_{k,19}$ are, respectively, the 19 amino acid composition of the kth protein $\boldsymbol{x}_k$, and N is the total number of proteins in the set.

According to statistical theory,[24] when the number of the proteins in database is sufficiently large, the distribution density functions for α, β, α + β, and α/β proteins can be assumed to be the 19-dimensional normal density functions[25]

$$P(\boldsymbol{x}\,|\,\omega_l) = \frac{1}{\sqrt{(2\pi)^{19}\,|\,\Sigma_l\,|}}\ \exp\{-\frac{1}{2}(x - \boldsymbol{\mu}_l)^T\Sigma_l^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_l)\}$$
$$l \in \alpha, \beta, \alpha + \beta, \text{ or } \alpha/\beta \tag{8}$$

where $\boldsymbol{\mu}_l$ and Σ_l are mean vector and 19 × 19 covariance matrix for the l type protein, $l \in$ α, β, α + β, or α/β, respectively. According to the maximum likelihood estimation method, the sample average of the protein set concerned can be estimated by

$$\boldsymbol{\mu}_l = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \\ \bar{x}_{19} \end{bmatrix} \tag{9}$$

where

$$\bar{x}_i = \frac{1}{N}\sum_{k=1}^{N} x_{k,i} \qquad (i = 1, 2, \ldots, 19)$$

and the covariance matrix is given by

$$\Sigma_l = \begin{bmatrix} s_{1,1} & s_{1,2} & \ldots & \ldots & s_{1,19} \\ s_{2,1} & s_{2,2} & \ldots & \ldots & s_{2,19} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{19,1} & s_{19,2} & \ldots & \ldots & s_{19,19} \end{bmatrix} \tag{10}$$

where

$$s_{i,j} = \frac{1}{N}\sum_{k=1}^{N} [x_{k,i} - \bar{x}_i][x_{k,j} - \bar{x}_j] \qquad (i,j = 1, 2, \ldots, 19)$$

It has been shown that if $N > 19$ and the sample are drawn from a normal population, the estimate of Σ_l given by Eq.(10) possesses an inverse Σ_l^{-1} with probability 1.[26] Note that the maximum likelihood estimator for $\boldsymbol{\mu}_l$ is unbiased but Σ_l is not unbiased. The desired unbiased estimator for Σ_l is defined by

$$s_{i,j} = \frac{1}{N-1}\sum_{k=1}^{N} [x_{k,i} - \bar{x}_k][x_{k,j} - \bar{x}_j]$$
$$(i,j = 1, 2, \ldots, 19) \tag{11}$$

which is very nearly the maximum likelihood estimator for large values of N (say $N > 30$). Because of this, some statisticians choose to define the sample variance by Eq.(11) rather than Eq.(10). In this case, it can be verified that the present method is mathematically identical to the method used by Chou and Maggiora.[20] When the N proteins in Eqs.(9), (10), and (11) are all α proteins, $\boldsymbol{\mu}_l$ and Σ_l thus defined would become the mean and the covariance matrix of α protein set, denoted by $\boldsymbol{\mu}_\alpha$ and Σ_α. Likewise, when the N proteins in equations (9)–(11) are all β, or α + β, or α/β protein sets, denoted by $\boldsymbol{\mu}_\beta$, $\boldsymbol{\mu}_{\alpha+\beta}$, $\boldsymbol{\mu}_{\alpha/\beta}$, and $\boldsymbol{\Sigma}_\beta$, $\Sigma_{\alpha+\beta}$, $\Sigma_{\alpha/\beta}$, respectively.

Because of the exponential form of the normal density function, it is sometime more convenient to work with natural logarithm of this decision function

$$d_l(\boldsymbol{x}) = \ln[P(\omega_l)P(\boldsymbol{x}\,|\,\omega)] = \ln P(\omega_l) + \ln P(\boldsymbol{x}\,|\,\omega_l) \tag{12}$$

which is totally equivalent to Eq.(4) in terms of classification performance since ln is a monotonically increasing function. Substituting Eq.(8) into Eq.(12) yields

$$d_l(\boldsymbol{x}) = \ln P(\omega_l) - (19/2)\ln 2\pi$$
$$- (1/2)\ln\,|\Sigma_l| - (1/2)[(\boldsymbol{x} - \boldsymbol{\mu}_l)^T\Sigma_l^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_l)] \tag{13}$$

Since the term (19/2)ln2π does not depend on l, it can be eliminated from the expression; $d_l(\boldsymbol{x})$ then becomes

$$d_l(\boldsymbol{x}) = \ln P(\omega_i) - (1/2)\ln|\Sigma_l| - (1/2)[(\boldsymbol{x} - \boldsymbol{\mu}_l)^T\Sigma_l^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_l)] \quad (14)$$

Equations (13) and (14) also represent the Bayes decision functions for normal patterns. If all a priori probabilities are equal: $P(\omega_1) = 1/M$, for $l = 1, 2, \ldots, M$, it can easily be shown that by dropping the terms independent of index l, Eq.(14) becomes

$$d_l(\boldsymbol{x}) = -(1/2)\{\ln|\Sigma_l| + [(\boldsymbol{x} - \boldsymbol{\mu}_l)^T\Sigma_l^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_l)]\} \quad (15)$$

Note that the second term in the right side of Eq.(15) is the Mahalanobis distance from $\boldsymbol{x}$ to $\boldsymbol{\mu}_l$. Thus, it is evident from the discussion given above that if the M pattern classes are governed by the multivariate normal density functions, the decision rules for the least Mahalanobis distance can be viewed as the approximation the Bayes decision rule.

METHODS

Data Set

It is impossible to accurately know in advance the accuracy of a prediction method when applied to a new protein. In order to use the data bank of known structures to estimate the performance on new proteins, two requirements are essential to derive a reasonable assessment of the method's generalization ability: (1) the pairwise identity of the protein chains or domains used for developing the prediction method and those for testing should be lower than the value sufficient for modeling tertiary structure by homology, and (2) a multiple cross-validation test (ideally jack-knife) has to be performed to exclude a potential dependency of the evaluated accuracy on the particular test set chosen.

It is now well established that protein domains having more than 30% of their sequence in common adopt the same fold structures.[27–30] Therefore, a tool not using the homology to a protein of known structure has to be tested on those cases for which it will be used, i.e., protein domains without significant pairwise homology to those used for developing the method. In the present study, the classification method of Murzin et al.[3] was used for the distinction between different structural classes (SCOP, version 1.38). This database provides a detailed and comprehensive description of the structural and evolutionary relationships of proteins whose three-dimensional structures have been determined. The basic unit for classification in SCOP is the protein domain. 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 classification in SCOP is entirely manual and does not incorporate any of the "hard and fast" rules, such as those described by Nakashima et al.,[11] Klein and Delisi,[7] and Chou.[15] Rather, it focuses on what structural elements are within the "core" of the protein. A protein with 7 strands and 1 helix may be "α + β" if the helix is integral to the core, while it would be "all β" if the helix were a non-conserved elaboration. The distinction of α/β and α + β is made on the interactions between the α and β sections of the structure. α/β proteins typically have interspersed α and β units, while α + β proteins typically have separate regions of mostly α and mostly β structure. 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. In the SCOP database, protein domains are classified into the following 10 categories: (1) all α; (2) all β; (3) α/β; (4) α + β; (5) multidomain; (6) membrane and cell surface protein; (7) small protein; (8) peptides; (9) designed protein; and (10) non-protein. In this study, only categories (1)–(4) will be considered. The creators of SCOP have clustered the domains in the Protein Data Bank on the basis of sequence identity.[31] At a sequence identity level of 40%, 1,189 unique sequences corresponding to the known structural domains were found in the PDB40D_1.37 database of SCOP. These 1,189 sequences are used as both the training and test sets in the present study.

Self-Consistency and Jack-Knife Tests

The prediction quality was examined by two approaches. One is based upon the re-substitution test and the other upon the jack-knife test. The former is for testing the self-consistency of the algorithm, whereas the latter is for testing the results by cross-validation. When the self-consistency test is performed for the current study, the structural class for each of the domain in a given data set is predicted using the rules derived from the same data set. Testing predictive accuracy on the training set could lead to unrealistically high accuracies. An objective test of a structural class prediction method will predict the structure of a test set of proteins that are not in the training set and show no detectable sequence similarity with the training set. Since the number of proteins of known structure is limited, it is normal to develop structural class prediction methods by cross-validation techniques, or jack-knife test. In a full jack-knife test, each protein or domain in the data set is in turn moved from the set, the parameters are developed on the remaining domains, the structure of the removed domain is predicted and its accuracy measured. In other words, the structural class of each domain is predicted by the rules derived using all other domains except the one that is being predicted. During the process of jack-knife analysis, both the training data set and testing data set are actually open, and a domain will in turn move from one to the other.

Prediction Procedure

Suppose $\boldsymbol{x}$ is a protein whose structural class is to be predicted. Using the Bayes decision rule for minimum error, the prediction can be performed according to the following procedure:

TABLE I. Comparison of Prediction Results for the Data Sets Used by Nakashima et al.[11] and Chou[15]

		Prediction method		
		Chou and Zhang[14]	Wang and Yuan (this study)	
Test set	Protein type	Self-consistency	Self-consistency	Jack-knife
131 proteins of Nakashima et al.[11]	All α	31/31 = 100%	31/31 = 100%	17/31 = 54.9%
	All β	34/34 = 100%	33/34 = 97.1%	17/34 = 50.0%
	α + β	24/27 = 88.9%	27/27 = 100%	4/27 = 14.8%
	α/β	35/39 = 89.7%	39/39 = 100%	18/39 = 46.2%
	Average	124/131 = 94.7%	130/131 = 99.2%	56/131 = 42.7%
4 × 30 = 120 proteins of K.C. Chou[15]	All α	30/30 = 100%	30/30 = 100%	20/30 = 66.7%
	All β	30/30 = 100%	30/30 = 100%	17/30 = 56.7%
	α + β	30/30 = 100%	30/30 = 100%	14/30 = 46.7%
	α/β	29/30 = 96.7%	30/30 = 100%	13/30 = 43.3%
	Average	119/120 = 99.2%	120/120 = 100%	64/120 = 53.3%

(1) Knowing the amino acid compositions of the database proteins, normalize their amino acid components by dividing the number of each component amino acid by the total number of amino acids in the protein.

(2) Eliminate one of the 20 normalized amino acid components, thereby defining a 19-D space, and express the proteins as points in the 19-D space.

(3) Calculate the mean vectors and 19 × 19 covariance matrixes for the α, β, α + β, and α/β proteins from the database proteins.

(4) Calculate the normalized frequencies of 20 amino acids of the unknown protein $\boldsymbol{x}$ as follows:

$$x_i = \frac{v_i}{\sum_{j=1}^{20} v_j} \qquad (i = 1, 2, \ldots, 20)$$

where v_i is number of ith amino acid in the protein $\boldsymbol{x}$ (i = 1, 2 , . . . , 20). Thus, the protein $\boldsymbol{x}$ also corresponds to a point ($x_1, x_2, \ldots, x_{19}$) in the 19-D space.

(5) For the point of the unknown protein $\boldsymbol{x}$, calculate the conditional probabilities of occurrence of $\boldsymbol{x}$ for the above four types of protein, $P(\boldsymbol{x} \mid \omega_l)$, $l \in$ α, β, α + β, α/β [cf., Equation (8)].

(6) The unknown protein is predicted to have the same structural class as the one which the value of decision function, $d_l(\boldsymbol{x}) = P(\omega_l)P(\boldsymbol{x} \mid \omega_l)$ or $d_l(\boldsymbol{x}) = P(\boldsymbol{x} \mid \omega_l)$ if all a priori probabilities are equal, is the largest [cf., Equations (3) and (4)].

RESULTS AND DISCUSSION

Comparison to the Previous Component-Coupled Methods

In order to compare class prediction quality by the present and the other component-coupled methods, two protein data sets used by Nakashima et al.[11] and Chou[15] were tested first. In their original paper Nakashima et al. examined 135 proteins of which 31 are α proteins, 34 β, 27 α + β, 39 α/β, and 4 irregular. The criteria of classification are given by

α proteins	⇒ α > 15%, β < 10%
β proteins	⇒ α < 15%, β > 10%
α + β proteins	⇒ α> 15%, β > 10%
	with dominantly antiparallel β-sheets
α/β proteins	⇒ α > 15%, β > 10%
	with dominantly parallel β-sheets
Irregular proteins	otherwise

The irregular proteins have been left out in this study because their number is only four, too small to have any statistical significance. Therefore, the prediction and comparison will be made based on the remaining 131 proteins. Assuming that four a priori probabilities are equal, $P(\omega_\alpha) = P(\omega_\beta) = P(\omega_{\alpha+\beta}) = P(\omega_{\alpha/\beta}) = 1/4$, the rates of correct prediction for 31 α, 34 β, 27 α + β, and 39 α/β proteins are 100%, 97.1%, 100%, and 100%, respectively. If the average accuracy is defined by the percentage of the number of correct prediction events for all classes divided by the number of total prediction events, i.e.,

$$Q = \text{average accuracy} = \frac{\text{total number of correct prediction events}}{\text{total number of prediction events}}$$

we have the average accuracy of 99.2% for predicting the 131 proteins by the current method. This result shows that the average prediction accuracy of our method is 4.5% higher than that obtained by the least Mahalanobis distance method. However, when predicted with jack-knife test, the average accuracy was only 42.7%. The results are summarized in Table I.

Although proteins of known structure are generally classified into one of the four structural classes, there is no unified quantitative measure for making such a classifica-

TABLE II. Comparison of Prediction Results for the Data Sets Used by Eisenhaber et al.[17]

		Prediction method			
		Eisenhaber et al[17]		Wang and Yuan (this study)	
Test set	Protein type	Self-consistency	Jackknife	Self-consistency	Jackknife
260 proteins of Eisenhaber et al.[17]	All α (55)	54.5%	47.3%	78.2%	45.5%
(set ≥2.0 Å)	All β (78)	82.1%	76.9%	88.5%	69.2%
	Mixed (127)	50.4%	50.4%	89.0%	66.1%
	Average	60.8%	57.7%	86.5%	62.7%
471 proteins of Eisenhaber et al[17]	All α (99)	57.6%	54.5%	69.7%	56.6%
(Set 3.0 Å)	All β (140)	77.9%	76.4%	82.9%	65.0%
	Mixed (232)	46.6%	47.0%	81.9%	72.0%
	Average	58.2%	57.3%	79.6%	66.7%

tion. The relevant percentages set by Nakashima et al. for α proteins ($\alpha > 15\%$) and β proteins ($\beta > 10\%$) do not seem large enough to reflect the real features of the two structural classes. In view of these, Chou[15] proposed a new method that classifies protein according to the following quantitative criterion:

α proteins $\Rightarrow \alpha \geq 40\%, \beta \leq 5\%$

β proteins $\Rightarrow \alpha \leq 5\%, \beta \geq 40\%$

α + β proteins $\Rightarrow \alpha \geq 15\%, \beta \geq 15\%$

with more than 60% parallel β-sheets

α/β proteins $\Rightarrow \alpha \geq 15\%, \beta \leq 15\%$

with more than 60% parallel β-sheets

ζ proteins $\Rightarrow \alpha \leq 10\%, \beta \leq 10\%$

where the contents of protein secondary structures were computed based on the dictionary by Kabsch and Sander.[32] According to the new criteria and selection principle, 120 structure-known proteins were thus selected and classified into 30 α, 30 β, 30 α + β and 30 α/β proteins. Based on such a training database, all of the 120 proteins were correctly predicted according to the Bayes decision rule. It can be seen from Table I that for this carefully selected database, although the prediction accuracy for the self-consistency test was 100% by our method, only about half proteins were correctly predicted for the jack-knife test.

One possible explanation for this remarkable difference between the self-consistency and jack-knife tests is that the data base used is too small so that the information loss due to jack-knife will have a greater impact on the prediction results, because these kind of methods need more training data to make their prediction mechanism work properly. Eisenhaber et al.[17] have tested the least Mahalanobis distance method with a larger database of non-homologous crystal structures (residue identity among all aligned pairs sequences ≤ 35%, minimal sequence length 80 residues). They concluded that a jury decision among four structural classes based only on the amino acid composition of the query protein is at best 50–60% correct, and consideration of the coupling effect among different amino acid components would not improve the class prediction quality. However, since the original least Mahalanobis distance method developed by Chou and Zhang is not valid for their training set, in which the subset sizes are very much different, the conclusion obtained by Eisenhaber et al. could still be questionable.[19,20] As mentioned earlier, the component-coupled method proposed in the present study can be used to deal with the cases where the training subset sizes are different, therefore, it is interested in testing our method with the same database used by Eisenhaber et al. According to their paper, the following rule was used to classify the structural classes of proteins:

all α proteins $\Rightarrow \alpha > 15\%, \beta < 10\%$

all β proteins $\Rightarrow \alpha < 15\%, \beta > 10\%$

mixed class proteins $\Rightarrow \alpha > 15\%, \beta > 10\%$

irregular proteins $\Rightarrow$ otherwise

Their rule does not distinguish between the α/β class and the α + β class and places them in one class, the so-called mixed class. As the subsets of irregular proteins is too small to be statistically significant, they have been removed from consideration. Table II gives a comparison of results obtained by our method and by the vector decomposition method of Eisenhaber et al. with two different databases. It can be seen from this table that the self-consistency test of the current method for 260 proteins (the second training set in Table I of their paper) is 86.5%, which is about 26% higher than the result obtained by their component-coupled method. Similarly, when the jack-knife test was performed for the same data set, the success rate drops to the value of 60%. Note that for the larger data set containing 475 proteins, the overall rate of correct prediction by the current method for the self-consistency test decreases but that for jack-knife test increases slightly. This result suggests that the limited size of data set may not be the main reason for the remarkable difference between the self-consistency and jack-knife tests.

TABLE III. Prediction Results for the 1,189 Domains in the PDB40D_1.37 Database by the Present Method

Test set	Protein type	Number of domains	Prediction accuracy	
			Self-consistency	Jack-knife
1,189 protein domains obtained from the SCOP database (sequence identity <40%)	Allα	263	167/263 = 63.5%	144/263 = 54.8%
	Allβ	317	204/317 = 64.4%	181/317 = 57.1%
	$\alpha + \beta$	270	107/270 = 39.6%	60/270 = 22.2%
	α/β	339	281/339 = 82.9%	255/339 = 75.2%
	Average		759/1189 = 63.8%	640/1189 = 53.8%

Effect of the Size of Data Set on Prediction Accuracy

The validation of all prediction methods is, of course, dependent on the classification scheme. The surprisingly poor results (with the data set in Table II) of the composition-coupled method may be due partly to the classification principles. As pointed out by Chou et al.,[19,20] protein structural classification should be based on the domain structure, while the classification of the three data sets tested above are based on the whole protein chain. If a data set is constructed according to an arbitrary or incorrect classification rule, it certainly cannot objectively reflect the relationship between the structural class of a protein and its amino acid composition. All the calculated results based on such a data set would be meaningless. To avoid this, we also test our method with a more reasonable classification scheme and the largest data set at present. As mentioned above, SCOP database is now a more reasonable data base and available in the network. At a sequence identity level of 40%, 1,189 unique sequence corresponding to the known structural domains were found in the PDB40D_1.37 database of SCOP. There are 263 all α, 317 all β, 270 $\alpha + \beta$, and 339 α/β protein domains. The results predicted by the current method for the 1,189 domains are summarized in Table III. The overall rate of correct prediction by the current method is 63.8% for the self-consistency test, and 53.8% for the jack-knife test. Thus we think that the poor results obtained cannot be explained by differences in protein structural classification.

The goal of testing the prediction tool is to assess the accuracy to be expected for any new protein sequence. Since different test sets yield different results, it is not sufficient to use only one set. In order to exclude the artifact owing to selecting different sets of protein and analyze the impact of the learning set, the new method was tested with learning sets of varying representativity and size. N (N = 40, 80,120, 160, 200, 240) representative domains were selected randomly from each of the four structural classes in the PDB40D database. For a database of such $4 \times N$ domains, both the self-consistency and jack-knife tests were performed. This procedure was repeated for 200 times. These 200 sets were tested to determine the average results and the degree of variation that can occur. The average over all 200 tests gives a reasonable estimate of the prediction accuracy. The individual and overall rates of class prediction are summarized in Table IV. We also computed the average difference for the current method on the 200 different data sets, which is the deviation of prediction accuracy and reflects the statistical fluctuation of the data sets. The larger the N, the smaller the deviation. The performance of the method turns out to be strongly dependent on the size of the data set. For the self-consistency test, more than 90% accuracy is obtained with a representative learning set of size = 160 while the performance came in about 20% lower with a data set of size = 960. This suggests that considerable care has to be used in evaluating the results of a single test set. By enlarging the data set, the self-consistency performance showed a steady decrease, yet the jack-knife performance improved slightly. This decrease in the "generalization gap"—the difference in performance between the learning and test sets—illustrates that a decrease in some of the noisy sequence-specific (or example-specific) information occurs with this simplification but without a corresponding decrease in structural classes, this underscores the importance of cross-validation in studies of this type.[33]

Figure 1 shows the dependence of the prediction accuracy achieved by the current method on the size of data set. Since the limited size of the current protein data bank, the predicted accuracy is, to some extent, dependent on the set of proteins selected by the predictor. The similar problem also exists when prediction is performed for a set of testing proteins. Only when the number of proteins considered is sufficiently large can the bias due to the selection of different protein sets be eliminated. Therefore, to make a fair comparison of different prediction methods, one should adopt the objective accuracy as a criterion. The objective accuracy is actually an asymptotical limit for the rate of correct prediction computed for a sufficiently large number of proteins. It can be seen from Figure 1 that as expected, the jack-knife test rate will be close to the self-consistency test rate and the bias due to selection of different sets can be eliminated when the data set becomes sufficiently large. The common horizontal asymptotical limit suggests that the objective accuracy of our prediction method should be about 60%.

For a small data set, a large generalization gap may be considered as implying that the algorithm is "memorizing" the information in a rigid fashion rather than learning the underlying informational concepts behind a classification scheme. Both the current method and the least Mahalanobis distance method show the largest generalization gaps, with the performance on the learning sets at or near 100% accuracy while the performance on the test sets came in

TABLE IV. Prediction Results for the Different Size of Data Sets by the Present Method

Test set	Protein type	Prediction accuracy: Self-consistency (%)	Prediction accuracy: Jack-knife (%)
4 × 40 = 160 protein domains randomly selected from the SCOP database containing 1,189 domains (200 times)	All α	88.1 ± 4.7	50.9 ± 8.8
	All β	93.9 ± 3.8	50.1 ± 7.9
	α + β	90.7 ± 4.1	42.9 ± 9.3
	α/β	98.3 ± 2.1	36.0 ± 6.4
	Average	93.0 ± 2.0	45.0 ± 5.0
4 × 80 = 320 protein domains randomly selected from the SCOP database containing 1,189 domains (200 times)	All α	74.6 ± 4.2	53.4 ± 5.2
	All β	82.5 ± 4.1	52.9 ± 5.7
	α + β	73.8 ± 4.2	42.1 ± 5.8
	α/β	94.5 ± 2.4	56.6 ± 4.4
	Average	81.0 ± 2.0	51.0 ± 3.0
4 × 120 = 480 protein domains randomly selected from the SCOP database containing 1,189 domains (200 times)	All α	69.2 ± 3.4	54.5 ± 4.4
	All β	76.7 ± 3.9	53.8 ± 4.8
	α + β	64.8 ± 3.6	40.6 ± 4.2
	α/β	90.9 ± 2.1	64.7 ± 3.4
	Average	75.0 ± 3.0	53.0 ± 2.0
4 × 160 = 640 protein domains randomly selected from the SCOP database containing 1,189 domains (200 times)	All α	66.4 ± 2.3	55.0 ± 2.8
	All β	73.1 ± 3.1	54.7 ± 3.4
	α + β	59.4 ± 3.0	39.1 ± 3.4
	α/β	89.1 ± 2.0	69.5 ± 2.9
	Average	72.0 ± 1.0	55.0 ± 2.0
4 × 200 = 800 protein domains randomly selected from the SCOP database containing 1,189 domains (200 times)	All α	64.3 ± 2.1	55.3 ± 2.2
	All β	70.9 ± 2.6	54.9 ± 3.0
	α + β	55.8 ± 2.3	38.5 ± 2.6
	α/β	87.9 ± 1.7	72.2 ± 2.2
	Average	70.0 ± 1.0	55.0 ± 1.0
4 × 240 = 960 protein domains randomly selected from the SCOP database containing 1,189 domains (200 times)	All α	63.3 ± 1.3	55.4 ± 1.6
	All β	68.9 ± 1.8	55.8 ± 2.2
	α + β	52.8 ± 2.0	37.4 ± 2.2
	α/β	86.5 ± 1.5	73.9 ± 1.7
	Average	67.9 ± 0.7	56.0 ± 1.0

50–55% lower. This memorization is likely due to the large number of parameters with respect to the size of each learning set, which tends to cause the current method to readily extract and retain sequence-specific information. The best solution to the memorization problem appears to be to increase the size of the training set. An increased training set size will also likely decrease some of the "wobble" associated with small sets; that is, larger sets will contain more structural class information, decreasing the variation in prediction accuracy due to the random compositions of the training and test sets.[33]

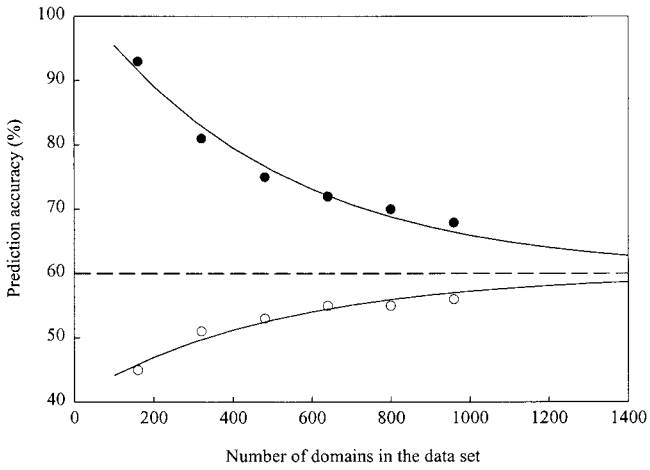

Fig. 1. Dependence of structural class prediction accuracy on the size of the protein domain data set. (·) Self-consistency test; (○) Jack-knife test.

Effect of Sequence Homology on Results

Whether knowledge of the fractions of the 20 amino acids is sufficient alone for predicting the structural class of a given protein is steeped in controversy. The level of prediction accuracy is still available ranging from 60% to near 100% and clearly depends on the definitions of structural classes and the set of database structures considered for performing the analysis. Analyses of known three-dimensional protein structures and amino acid sequences revealed that proteins are clustered into families whose members have evolved from a common ancestor, share a characteristic fold, and, sometimes, have a similar function. The number of families of related protein structures critically depends on the value of the homology threshold applied in the protein structure comparison routines. According to the classification of protein structures in SCOP, all protein domains with more than 30% identity belong to the same of protein family and must have the same structural class. Therefore, the structural class assignment of a new protein domain with homologous > 30% to a protein of known structure can be easily performed by sequence alignment, and any prediction method for the protein structural classes should only address those proteins for which no homologous proteins (at a sequence identity level of 30%) are found in the Protein Data Bank. Since smaller test sets always open the theoretical possibility that just proteins which would be badly predicted have been missed and that the prediction rates are overestimated, it is extremely important that success rates are calculated from as large as possible representative subsets of PDB. The most stringent test set of tertiary structures available would consist of all unrelated proteins, the structure of which is known. In the PDB40D_1.37 database, the 1,189 domains belong to 675

TABLE V. The 675 Protein Domains (Sequence Identity <30%) Extracted From the PDB40D_1.37 Database of SCOP for Self-Consistency and Jack-Knife Tests

1. 155 α domains

3sdha_	1lea__	1erc__	2hmqa_	1pnra1	1bmta1	1ezm_1	2abk__	1aora1	1csma_
1cpca_	1aoy__	1aca__	1vtmp_	2tct_1	1tpt_1	2ts1_1	1gln_1	1vnc__	1pprm1
1grj_1	1cgpa1	1lre__	1buca1	1mi_1	1c5a__	1bmfa1	1rlr_1	1sig__	1poc__
1hdj__	1opc__	1vii__	1fapb_	1coo__	1occh_	1abv__	1dnpa1	1bvp11	1poa__
1seta1	1hsta_	2end__	1ryt_1	1an2a_	1hyp__	1ak4c_	1lla_1	1rgp__	1beo__
1mnga1	1etd__	1lis__	1mmob_	4icb__	1bip__	2hmx__	2pgd_1	2bct__	1rtm11
1dvh__	2hts__	1bmfg_	1bgc__	1sra__	1pnb.1	1jvr__	1yvei1	1lrv__	2ztaa_
1aofa1	1dpra1	1lbu_1	1csga_	1rro__	1olga_	1vin_1	1utg__	1sly_1	1ifj__
1etpa1	1xgsa1	1hme__	1ilk__	2scpa_	1adt_1	1vola1	1glm__	1ocee_	1vdfa_
1enh__	1fow__	1bfma_	1acp__	1djxa1	1ihfb_	1cuk_2	1cem__	2sblb1	1dkga1
1hcra_	1cuk_1	1tafa_	1imq__	1cpo_1	1alo_1	1dpra2	5eas_1	2tct_2	1dipa_
1msec1	1tns__	1mmog_	1rpo__	1pax_1	1ab3__	1ngr__	1csh__	1lbd__	
1pdnc_	2spca_	1lpe__	1ytfb1	1myka_	1aep__	1crka1	1phb__	1fps__	
1igna1	1fc2c_	2liga_	1ecia_	1cmba_	1nkl__	1agre_	1fipa_	5eas_2	
1sfe_1	1gab__	256ba_	1octc2	1dsba1	1hvd__	1aru__	1prcc_	1grl_1	
1bia_1	1bbl__	2ccya_	1llia_	2gsta1	1tada1	1mhl.1	2wrpr_	1ecma_	

2. 156 β domains

1neu__	1nbca_	2bpa1_	1hgea_	1ckaa_	1mjc__	1arb__	1wpoa_	2cba__	1rgs_1
1cd1a1	1qba_2	1stma_	1aol__	1mmd_1	1ckma1	1bty__	1pkya1	3bcl__	2arca_
1vcaa1	1tupa_	2bbva_	1knb__	1vie__	1rip__	2snv__	1hbp__	1ospo_	1wapa_
1vcaa2	1ctm_1	1bbt1_	1aly__	1pse__	1yhb__	1hava_	1hms__	1vmoa_	1bdo__
1gof_1	1cdg_2	4gcr_1	1thw__	1ihwa_	1pyp__	1bco_1	1sria_	1dlc_2	1bnca1
2hft_1	1eta1_	1prr_1	1scs__	1aono_	1prch1	1bmfa2	1smpi_	1msaa_	1ctm_2
1bgla1	2pcda_	1wkt__	1cpn__	2ohxa1	2fgf__	1sfta1	2cpl__	1kapp1	1gpr__
1ggta2	1hoe__	2sblb2	1slaa_	1pdr__	1il1b__	1fiva_	1hxn__	2pec__	1gzi__
1ncia_	1plc__	1hpla1	1saca_	1fgp__	1abrb1	1epne_	2sil__	1idk__	2kaub_
1noa__	1cyx__	1pgs_1	1kit_1	1whi__	1wba__	2eng__	1gof_3	1tsp__	1dupa_
1xsoa_	1aoza1	1slua_	1cela_	1sty__	1hcd__	1bw3__	2bbkh_	1lxa__	1tul__
3dpa_1	1djxa2	1gof_2	1xnb__	1ltsd_	1fnb_1	1cxsa1	2trcb_	1tdta_	2kauc1
1mspa_	1rsy__	1dlc_1	1bgla4	1esfa1	2pia_1	1gtra1	4aaha_	1thja_	
4kbpa1	3dpa_2	1bgla3	1oaca1	1asya1	1fuia1	1mai__	1aofa2	2phla1	
1ddt_1	1who__	1ulo__	1bia_2	1cuk_3	1eft_1	1irsa_	1bplb1	1pmi__	
1exg__	1npoa_	1bvp12	1umua_	3ulla_	1eft_2	1ytfc1	1dkga2	1cgpa2	

3. 184 α+β domains

1gmpa_	1ubi__	1ema__	1gpma3	1afi__	1otga_	1hfc__	1lba__	1seia_	1fjma_
2baa__	1guab_	1fkd__	1ebha2	1psda3	1fim__	1qba_4	3b5c__	1div__	2cmd_2
193l__	1frd__	1ctn_3	1grl_3	1mla_2	1gd1o2	1shaa_	1vcc__	2glt_2	1aiha_
119l__	1esfa2	1grj_2	1fca__	1fwp__	1dapa2	1ptf__	1yua_1	1bnca3	1mrj__
1gbs__	1tif__	1froa_	1fd2__	1regx_	1ofga2	1iba__	1ah6__	1scub2	1lts.1
1sly_2	1lgr_1	1han_1	1xer__	1ab8a_	1oaca4	1af5__	1orda3	1dik_3	1pax_2
1chka_	1coy_2	1mkaa_	1vjw__	1vaoa1	1bp1_1	1gtpa_	1smna_	1ckma2	1def__
2act__	1pbe_2	1csei_	1raab1	1geo_1	2sici_	1gtqa_	1chma2	1vaoa2	1lit__
1ggta4	1kifa2	1alo_3	1pba__	1lbu_2	2ms2a_	1scea_	1lgr_2	1mbb_1	1prtb2
7rsa__	1gal_2	1qapa2	1spbp_	1vhh__	1gesa3	1efnb_	1crka2	1mbb_2	1tsg__
1ag2__	1gnd_2	1tpt_3	1mli__	1tig__	1srsa_	1cby__	1ytba1	1ako__	3fib__
2kaua_	1mola_	1tfe__	1pil__	1uae__	1tbd__	1seta2	3pmga4	1aora2	1msk__
1huma_	1cewi_	1mnga2	1npk__	1kpta_	1nox__	1bia_3	1bv1__	1gdoa_	1pmd_3
1pmd_1	1oaca3	1ctf__	1up1_1	1kvd.1	1kuh__	2vik__	1mxa_1	1pnk.1	
1sso__	1std__	2reb_2	2bopa_	3rubs_	1ezm_2	1ahq__	2pola1	1pmab_	
1kpaa_	1ouna_	1stu__	3rubl2	1dcoa_	1lml__	1pne__	1plq_1	1apy.1	
1hxpa1	1udii_	1pkp_2	1aps__	1xxaa_	1ast__	2phy__	1alo_4	1pya.1	
1dar_3	1cd1a2	1pda_2	1ris__	2chsa_	1atla_	1mut__	1geo_4	1bme__	
1igd__	1aak__	1vig__	1dar_4	1otfa_	1kapp2	1tys__	1hqi__	4kbpa2	

4. 180 α/β domains

1cdg_4	1pkya2	1gesa1	1orda1	1hrda1	1rlaa_	1bam__	3pgm__	1thta_	1orb_1
1byb__	1dik_1	1dik_2	1cus__	1scua2	3cla__	1pvua_	1rpa__	1tca__	1cxsa2
1xyza_	3rubl1	1zyma_	1esc__	1bnca2	1phr__	1cfr__	1nula_	1tahb_	1ad3a_
1cbg__	1tpfa_	1aco_1	2naca1	2dln_1	1e2b__	2rsla_	1lfaa_	1hpla2	1aco_2
1hvq__	2xis__	1grl_2	1fnb_2	2glt_1	1vhra_	1pdo__	1gara_	1yasa_	3pmga1
1qba_3	1luca_	1brsd_	2pia_2	1pvda1	2hnp__	1hpm_1	1vid__	2masa_	1fuia2
1add__	1nfp__	1dfji_	2ts1_2	1nbaa_	2trxa_	2yhx_1	1xvaa_	1pbn__	1php__
2kauc2	1qapa1	1nzya_	1gpma1	1deaa_	1mek__	1gtag1	1v39__	2ctb__	1mioa_
1pta__	1djxa3	1rvva_	1tpt_2	1pvda2	1dsba2	1chma1	1hmy__	1obr__	2bgu__
1nal1_	1gym__	1pau.1	1dnpa2	1trka1	1gp1a_	2m2__	1art__	1lam_2	1gpb__
1ucwa_	1reqa1	1udg__	2tmda3	1gky__	2gsta2	1asu__	1tpla_	1amp__	3pga1_
1dosa_	1pud__	1mla_1	1kifa1	5p21__	2trcp_	1bco_2	2dkb__	1draa_	1pfka_
2acr__	1sfta2	3chy__	2ohxa2	1mmd_2	1trka3	1kfd_1	1orda2	1hrda2	1ayl__
1ak5__	1tml__	1scua1	1xel__	1dts__	1pkya3	1noya_	1gpma2	1fua__	2dri__
1ebha1	1rlr_2	2fx2__	1gd1o1	2reb_1	1lam_1	1hjra_	1wht.1	2anha_	2olba_
2mnr_1	2tmda2	1qrda_	2naca2	1ble__	1htta1	1sfe_2	1ede__	1idm__	1lct__
1dora_	1coy_1	1bmta2	2cmd_1	1chd__	1eria_	1tfr__	1din__	1raaa1	1pxta1
1pii_1	1gnd_1	1reqa2	1yvei2	1csee_	1rvaa_	1exna_	1broa_	1wsyb_	1ctt_1

TABLE VI. Prediction Results for the 675 Domains of Table V by the Present Method

Test set	Protein type	Number of domains	Prediction accuracy	
			Self-consistency	Jack-knife
675 protein domains given in Table V (sequence identity <30%)	All α	155	97/155 = 62.6%	83/155 = 53.5%
	All β	156	98/156 = 62.8%	66/156 = 42.3%
	α + β	184	99/184 = 53.8%	52/184 = 28.3%
	α/β	180	156/180 = 86.7%	123/180 = 68.3%
	Average		450/675 = 66.7%	324/675 = 48.0%
675 protein domains given in Table V (sequence identity <30%)	All α	155	97/155 = 62.6%	85/155 = 54.8%
	All β	156	103/156 = 66.0%	76/156 = 48.7%
	Mixed	364	296/364 = 81.3%	255/364 = 70.0%
	Average		496/675 = 73.5%	416/675 = 61.6%

families. Therefore, a more rigorous test for the new method should be based on all representatives selected from each of the 675 families. A largest possible subset of non-homologous structures was established based on the PDB40D_1.37 database by the following procedure: the first domains listed in each of the families was taken as the representative of this family (Table V). The structural class prediction results obtained by our method based on the self-consistency and the jack-knife tests for this data set are listed in Table VI. It can be seen from this table that with this data set of non-homologous structures, the accuracy of the jack-knife test was about 7% lower than a previous result obtained from $4 \times 160 = 640$ domains in Table IV. This result indicates that the existence of homologous domains in the testing data set will significantly affect the prediction accuracy of the composition-coupled method. As expected, the overall rate for 3-type prediction is about 10% higher than that for 4-type prediction. In a recent study, Chou and Maggiogra[20] reported that with increasing the size of data sets, the prediction accuracy for the self-consistency test remains almost unvaried while that for the jack-knife test increases from 63.77% to 84.12%. According to this observation, they concluded that by expanding a database to reduce the information loss, the overall jack-knife rate by the component-coupled algorithm can be improved significantly. This is obviously contradictory to the results of Table IV in this paper. The origin of this puzzling difference between our results and theirs was found to be due to sequence homology in their data sets, which led to differences in accuracies. For example, the 138 domains of Table I, 253 domains of Table II, and 359 domains of Table III in their paper belong to 102, 129, and 130 protein families, respectively. The average number of homologous domains per family in the three data sets are 1.35, 1.96, and 2.76, indicating that even though each domain in the data set is singled out in turn as a "test domain" and all the rule-parameters are determined from the remaining domains, the memorization effects that included in the self-consistency tests cannot be completely removed. One problem with the jack-knife method is that sequence homology within the set may invalidate the assumption that the training set is devoid of information about the tested protein. If the training data and test data are identical or highly homologous, then the prediction accuracy could be misleadingly high. Therefore, a unbiased test in which these algorithms are applied to proteins without significant sequence homology has to be done.

CONCLUSIONS

The accurate prediction of structural classes from amino acid composition alone is an important issue, which has been the object of a number of recent studies. However, the success rates reported for structural class prediction with different methods are contradictory. The problem of recognizing structural class of a protein knowing only its amino acid composition appears completely solved by the least Mahalanobis distance method. The highest reported prediction accuracies are near 100%.[13–15,19,20] This is surprising, because only the amino acid composition has been used in this method, while traditional sequence-based secondary structure prediction achieve success rates of about 75% for structural class only with extensive input information (full sequence of the query protein, its amino acid composition and length multiple alignments with homologous sequences).[34–37] In this article, we resolve the paradox. Our objectives have been to validate the relationship between amino acid composition and structural class by using a Bayes method and to provide the possible upper limit of the prediction rate for structural classes. When applying both the self-consistency and jackknife tests on a larger data set without significant pairwise similarity, we found that knowledge of amino acid composition alone cannot lead to a success rate higher than 60% for a 4-type class prediction by our method. The apparent relatively high accuracy level (more than 90%) attained in the previous studies, which exceed the success rates (75%) of structural class predictions using traditional secondary structure prediction techniques (including those combining evolutionary information and neural networks) was due to the preselection of test sets, which may not be adequately representative of all unrelated proteins.

ACKNOWLEGMENTS

This work was supported in part by grant 103–13–03–02 from the 863 High-Technology Foundation and Pandeng project of the Ministry of Science and Technology.

REFERENCES

1. Levitt M, Chothia C. Structural patterns in globular proteins. Nature 1976;261: 552–558.

2. Richardson JS, Richardson DC. Principles and patterns of protein conformation. In: Fasman GD, editor. Prediction of protein structure and the principles of protein conformation. New York:Plenum Press; 1989. p 1–98.
3. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: A structural classification of protein database for the investigation of sequence and structures. J Mol Biol 1995;247:536–540.
4. Michie AD, Orengo CA, Thornton JJ. Analysis of domain structural class using an automated class assignment protocol. J Mol Biol 1996;262:168–185.
5. Deleage G, Roux B. Use of class prediction to improve protein secondary structure prediction. In: Fasman GD, editor. Prediction of protein structure and the principles of protein conformation. New York: Plenum Press; 1989. p 417–465.
6. Kneller DG, Cohen FE, Langridge R. Improvements in protein secondary-structure prediction by enhanced neural networks. J Mol Biol 1990;214:171–182.
7. Klein P, Delisi C. Prediction of protein structural class from amino acid sequence. Biopolymers 1986;25:1659–1672.
8. Klein P. Prediction of protein structural class by discriminant analysis. Biochim Biophys Acta 1986;874:205–215.
9. Chou PY. Amino acid composition of four classes of proteins. In: Abstracts of papers, Part I, Second Chemical Congress of the North American Continent. Las Vegas, 1980.
10. Chou PY. Prediction of protein structural classes from amino acid composition. In: Fasman GD, editor. Prediction of protein structure and the principles of protein conformation. New York: Plenum Press; 1989. p 549–586.
11. Nakashima H, Nishikawa K, Ooi T. The folding type of a protein is relevant to the amino acid composition. J Biochem 1986;99:153–162.
12. Zhang CT, Chou KC. An optimization approach to predicting protein structural class from amino acid composition. Protein Sci 1992;1:401–408.
13. Chou KC, Zhang CT. A new approach to predicting protein folding types. J Protein Chem 1993;12:169–178.
14. Chou KC, Zhang CT. Predicting protein folding types by distance functions that make allowances for amino acid interactions. J Biol Chem 1994;269:22014–22020.
15. Chou KC. A novel approach to predicting protein structural classes in a (20–1)-D amino acid composition space. Proteins 1995;21:319–344.
16. Chou KC, Zhang CT. Prediction of protein structural classes. Crit Rev Biochem Mol Biol 1995;30:275–349.
17. Eisenhaber F, Frömmel C, Argos P. Prediction of secondary structural content of proteins from their amino acid composition along. II. The paradox with secondary structural class. Proteins 1996;25:169–179.
18. Bahar I, Atilgan AR, Jernigan RL, Erman B. Understanding the recognition of protein structural classes by amino acid composition. Proteins 1997;29:172–185.
19. Chou KC, Liu WM, Maggiora GM, Zhang CT. Prediction and classification of domain structural classes. Proteins 1998;31:97–103.
20. Chou KC, Maggiora GM. Doman structural class prediction. Protein Eng 1998;11:523–538.
21. Tou JT, Gonzalez RC. Pattern recognition principles. London: Addison-Wesley Publishing Company; 1974.
22. Duda RO, Hart PE. Pattern classification and scene analysis. New York: John Wiley & Sons; 1973.
23. Andrew HC. Introduction to mathematical techniques in pattern recognition. New York: John Wiley & Sons; 1972.
24. DeGroot MH. Probability and statistics. 2nd ed. Reading, MA: Addision-Wesley Publishing Co.; 1986. p 267–278.
25. Zhang CT, Chou KC. Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition. Biophys J 1992;63:1523–1529.
26. Anderson TW. Introduction to multivariate statistics. New York: John Wiley & Sons; 1958.
27. Sander C, Schneider R. Database of homology-derived protein structures and structural meaning of sequence alignment. Proteins 1991;9:56–58.
28. Blundell TL, Johnson MS. Catching a common fold. Protein Sci 1993;2:877–883.
29. Flores TP, Orengo CA, Moss D, Thornton JM. Comparison of conformational characteristics in structurally similar protein pairs. Protein Sci 1993;2:1811–1826.
30. Hilbert M, Bohm G, Jaenicke R. Structural relationships of homologous proteins as a fundamental principle in homology modeling. Proteins 1993;17:138–151.
31. Brenner SE, Chhothia C, Hubbard TJP. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci USA 1998;95:6073–6078.
32. Kabsch W, Sander C. Dictionary of protein secondary structures: pattern recognitioin of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
33. Metfessel BA, Saurugger PN, Connelly DP, Rich SS. Cross-validation of protein structural class prediction using statistical clustering and neural networks. Protein Sci 1993;2:1171–1182.
34. Chandonia JM, Karplus M. Neural networks for secondary structure and structural class predictions. Protein Sci 1995;4:275–285.
35. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
36. Levin JM, Pascarella S, Argos P, Garnier J. Quantification of secondary structure prediction improvement using multiple alignments. Protein Eng 1993;6:849–854.
37. Boberg J, Salakoski T, Vihinen M. Accurate prediction of protein secondary structural class with fuzzy structural vectors. Protein Eng 1995;8:505–512.