SHORT COMMUNICATION

Prediction and Classification of Domain Structural Classes

Kou-Chen Chou,[1] Wei-Min Liu,[2] Gerald M. Maggiora,[1] and Chun-Ting Zhang[3]*
[1]*Computer-Aided Drug Discovery, Pharmacia & Upjohn, Kalamazoo, Michigan*
[2]*Department of Mathematical Sciences, Indiana University–Purdue University at Indianapolis, Indianapolis, Indiana*
[3]*Department of Physics, Tianjin University, Tianjin, China*

ABSTRACT **Can the coupling effect among different amino acid components be used to improve the prediction of protein structural classes? The answer is yes according to the study by Chou and Zhang (Crit. Rev. Biochem. Mol. Biol. 30:275–349, 1995), but a completely opposite conclusion was drawn by Eisenhaber et al. when using a different dataset constructed by themselves (Proteins 25:169–179, 1996). To resolve such a perplexing problem, predictions were performed by various approaches for the datasets from an objective database, the SCOP database (Murzin, Brenner, Hubbard, and Chothia. J. Mol. Biol. 247:536–540, 1995). According to SCOP, the classification of structural classes for protein domains is based on the evolutionary relationship and on the principles that govern the 3D structure of proteins, and hence is more natural and reliable. The results from both resubstitution tests and jackknife tests indicate that the overall rates of correct prediction by the algorithm incorporated with the coupling effect among different amino acid components are significantly higher than those by the algorithms without using such an effect. It is elucidated through an analysis that the main reasons for Eisenhaber et al. to have reached an opposite conclusion are the result of (1) misusing the component-coupled algorithm, and (2) using a conceptually incorrect rule to classify protein structural classes. The formulation and analysis presented in this article are conducive to clarify these problems, helping correctly to apply the prediction algorithm and interpret the results. Proteins 31:97–103, 1998.** © 1998 Wiley-Liss, Inc.

Key words: **α domains; β domains; α/β domains; $\alpha+\beta$ domains; resubstitution; jackknife; SCOP database**

INTRODUCTION

The concept of protein structural classes was originally introduced by Levitt and Chothia[1] based on a visual inspection of polypeptide chain topologies in a dataset of 31 globular proteins. According to this concept, protein folds can be classified into one of four classes: all-α, all-β, α/β, and α+β. Since then, various quantitative classification rules have been proposed based on the percentages of α-helices and β-sheets in a protein.[2–7] 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 can hardly be without arbitrariness and hence would go short of objectivity, so that, according to different rules, a same protein can be assigned to completely different classes. This can often cause confusion or lead to a wrong conclusion. For example, recently Eisenhaber et al.[8] selected the following rule for the structural classification:

$$\begin{cases} \text{all} - \alpha \text{ proteins} & \rightarrow \alpha > 15\%,\ \beta < 10\% \\ \text{all} - \beta \text{ proteins} & \rightarrow \alpha < 15\%,\ \beta > 10\% \\ \text{mixed class proteins} & \rightarrow \alpha > 15\%,\ \beta > 10\% \\ \text{irregular proteins} & \rightarrow \text{otherwise} \end{cases} \tag{1}$$

where α and β represents the percentages of α-helix and β-sheet in a protein, respectively. According to such a rule, a protein with $\alpha > 15\%$ and $\beta < 10\%$ will be assigned as all-α class, and that with $\alpha < 15\%$ and

Contract grant sponsor: China National Science Foundation; Contract grant numbers: 19577104 and 39570187; Contract grant sponsor: Pandeng Project of China; Contract grant sponsor: NSF; Contract grant number: DUE-9555408.

*Correspondence to: Chun-Ting Zhang, Department of Physics, Tianjin University, Tianjin, China.

Received 22 August 1997; Accepted 12 December 1997

$\beta > 10\%$ as all-β class. This seems hard to reconcile with the pictures generally accepted for the all-α and all-β proteins that are thought to be formed mainly by α-helices or β-sheets, respectively. Another problem is that two proteins with very tiny differences in secondary-structure contents, say 0.01% or even less, will be placed into two completely different classes, obviously reflecting some sort of subjective arbitrariness. This is because the classification rule as formulated in Equation (1) is actually based on a continuous model that is in conflict with the concept of cluster classification itself.

In contrast to the continuous model, most of the other rules[2,4–7] are based on a discrete model in which there are gaps for the percentages of secondary structures between different classes so that these rules are conceptually consistent with the discrete prerequisite for classification. However, do these gaps really reflect the objective reality? Or, to what degree can the distribution of protein structures be considered as forming discrete regions in fold-space rather than a continuum? Besides using the percentages of secondary structures as a criterion, is there any better way to classify the classes of proteins? These questions are vitally important to the study of this area. Without realizing these points, one is prone to make the mistake of confusing one thing with another in the structural class assignment, leading to an erroneous classification. This is also important for making a reasonable comparison among various prediction algorithms for structural classes.

Instead of using the percentages of secondary structure as a sole criterion to classify the classes of proteins, Murzin et al.[9] recently proposed a method based on the evolutionary relationships of proteins and on the principles that govern their three-dimensional (3D) structure. The unit of classification is usually 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 database thus constructed is called SCOP (Structural Classification of Proteins). In addition to the information of structural classes, SCOP[9] also provides a detailed and comprehensive description of the structural and evolutionary relationships of proteins whose 3D 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.

In this work, the data from the SCOP database were used to test different prediction algorithms so as to determine whether the rate of correct prediction for the protein structural classes can be significantly improved by taking into account the coupling effect among different amino acid components.

ALGORITHMS AND DATASETS

The class assignments in SCOP[9] are based on known three-dimensional structures; while the class predictions discussed here are based on the amino acid composition. The comparison of prediction quality will be made among the least Hamming distance algorithm,[2,5] the least Euclidean distance algorithm,[3] and the component-coupled algorithm.[7,10] A unified formulation for each of the three algorithms is briefed below.

Suppose there are N domains forming a set S, which is the union of four subsets, so that

$$S = S^{\alpha} \cup S^{\beta} \cup S^{\alpha/\beta} \cup S^{\alpha+\beta} \tag{2}$$

where the subset S^{α} consists of only all-α domains, the subset S^{β} consists of only all-β domains, and so forth. According to the correlation between the structural class of a protein and its amino acid composition, any protein domain in the set S corresponds to a vector or a point in the 20D space:

$$\mathrm{X}_k^{\mu} = \begin{bmatrix} \chi_{k,1}^{\mu} \\ \chi_{k,2}^{\mu} \\ \vdots \\ \chi_{k,20}^{\mu} \end{bmatrix}, \quad (k = 1, 2, \ldots N_{\mu};\ \mu = \alpha, \beta, \alpha/\beta, \alpha+\beta) \tag{3}$$

where $\chi_{k,1}^{\mu}, \chi_{k,2}^{\mu}, \ldots, \chi_{k,20}^{\mu}$ are the normalized occurrence frequencies of the 20 amino acids in the kth protein domain X_k^{μ} of the subset S^{μ}; and N_{μ} is the number of protein domains it contains. The standard vector for the subset S^{μ} is defined by

$$\mathrm{X}^{\mu} = \begin{bmatrix} \chi_{1}^{\mu} \\ \chi_{2}^{\mu} \\ \vdots \\ \chi_{20}^{\mu} \end{bmatrix}, \quad (\mu = \alpha, \beta, \alpha/\beta, \alpha+\beta) \tag{4}$$

where

$$\chi_i^{\mu} = \frac{1}{N_{\mu}} \sum_{k=1}^{N_{\mu}} \chi_{k,i}, \quad (i = 1, 2, \ldots, 20). \tag{5}$$

Suppose X is a domain whose structural class is to be predicted. It can be either one of the N protein domains in the set S, or a domain outside of it. It also corresponds to a point $(\chi_1, \chi_2, \ldots, \chi_{20})$ in the 20D space, where χ_i has the same meaning of $\chi_{k,i}^{\mu}$ but is associated with protein X instead of X_k^{μ}. Thus, the aforementioned three prediction algorithms can be formulated as follows.

The Least Hamming Distance Algorithm[2,5]

The Hamming distance[11] between the standard vector X^μ and the domain X in the 20D space is

$$D_H(X, X^\mu) = \sum_{i=1}^{20} |\chi_i - \chi_i^\mu|, \ (\mu = \alpha, \beta, \alpha/\beta, \alpha + \beta) \quad (6)$$

and the domain **X** is predicted to be the structural class that has the minimal Hamming distance to **X**, as can be formulated as follows. Suppose

$$D_H(X, X^\xi) = \mathrm{Min}\,\{D_H(X, X^\alpha), D_H(X, X^\beta), D_H(X, X^{\alpha/\beta}), D_H(X, X^{\alpha+\beta})\} \quad (7)$$

where ξ can be α, β, α/β, or $\alpha + \beta$ and the operator Min means taking the least one among those in the parentheses, then the superscript ξ of Equation (7) is the predicted structural class for the domain X. If there is a tie case, ξ is not uniquely determined. But it does not occur in our dataset.

The Least Euclidian Distance Algorithm[3]

The squared Euclidean distance[11] between the standard vector X^μ and the domain **X** in the 20D space is given by

$$D_E^2(X, X^\mu) = \sum_{i=1}^{20} [\chi_i - \chi_i^\mu]^2, \ (\mu = \alpha, \beta, \alpha/\beta, \alpha + \beta) \quad (8)$$

and hence, instead of ξ in Equation (7), the prediction is ξ in

$$D_E^2(X, X^\xi) = \mathrm{Min}\,\{ D_E^2(X, X^\alpha), D_E^2(X, X^\beta), D_E^2(X, X^{\alpha/\beta}), D_E^2(X, X^{\alpha+\beta})\}. \quad (9)$$

The Component-Coupled Algorithm[7,10]

The two algorithms above are based on simple geometric distances in which no coupling effect among different amino acid components is taken into account. Much different from them, the component-coupled algorithm is based on the square Mahalanobis distance,[12,13] as given by

$$D_M^2(X, X^\mu) = (X - X^\mu)^T C_\mu^{-1} (X - X^\mu), \ (\mu = \alpha, \beta, \alpha/\beta, \alpha + \beta) \quad (10)$$

where C_μ is a covariance matrix given by

$$C_\mu = \begin{bmatrix} c_{1,1}^\mu & c_{1,2}^\mu & \cdots & c_{1,20}^\mu \\ c_{2,1}^\mu & c_{2,2}^\mu & \cdots & c_{2,20}^\mu \\ \cdots & \cdots & \ddots & \cdots \\ c_{20,1}^\mu & c_{20,2}^\mu & \cdots & c_{20,20}^\mu \end{bmatrix} \quad (11)$$

and the superscript T is the transposition operator, and C_μ^{-1} is the inverse matrix of C_μ. The matrix elements $c_{i,j}^\mu$ in Equation (11) are given by

$$c_{i,j}^\mu = \sum_{k=1}^{N_\mu} [\chi_{k,i}^\mu - \chi_i^\mu]\,[\chi_{k,j}^\mu - \chi_j^\mu], \ (i, j = 1, 2, \ldots, 20). \quad (12)$$

Thus, the component-coupled algorithm is formulated as follows:[7,10]

$$D_M^2(X, X^\xi) = \mathrm{Min}\,\{D_M^2(X, X^\alpha), D_M^2(X, X^\beta), D_M^2(X, X^{\alpha/\beta}), D_M^2(X, X^{\alpha+\beta})\} \quad (13)$$

where ξ can be α, β, α/β, or $\alpha + \beta$ and the operator Min means taking the least one among those in the parentheses, then the superscript ξ of Equation (13) is the predicted structural class for the domain X.

The above algorithm has proved to be very powerful in predicting the structural classes of proteins.[7,10] Moreover, this algorithm was also recently used by Cedano et al.[14] to improve the prediction quality of cellular location of proteins. However, it should be pointed out that in the above applications, all the subset sizes in a training dataset are identical. For example, the training dataset in earlier studies[7,10] consists of 30 α, 30 β, 30 $\alpha + \beta$, and 30 α/β proteins, i.e., $N_\alpha = N_\beta = N_{\alpha/\beta} = N_{\alpha+\beta} = 30$. The training dataset in Cedano et al.[14] consists of 200 intracellular, 200 extracellular, 200 anchored, 200 membrane, and 200 nuclear proteins. In other words, the component-coupled algorithm of Equation (13) can be used only when the subset sizes in the training dataset are the same or approximately the same. When the case is not so, some modification factors must be incorporated, as given below. It is very important to realize this, otherwise, the component-coupled algorithm might be misused,[8] yielding incorrect results.

When the training subset sizes are different, instead of Mahalanobis distance of Equation (10), the component-coupled algorithm should be based on the Mahalanobis discriminant function as given below:[15]

$$F_M(X, X^\mu) = D_M^2(X, X^\mu) + \ln P_\mu \quad (14)$$

where P_μ is the product of all positive eigenvalues of C_μ. Also, instead of Equation (12), the matrix elements $c_{i,j}^\mu$ in Equation (11) should be given by

$$c_{i,j}^\mu = \frac{1}{N_\mu - 1} \sum_{k=1}^{N_\mu} [\chi_{k,i}^\mu - \chi_i^\mu][\chi_{k,j}^\mu - \chi_j^\mu], \quad (i, j = 1, 2, \ldots, 20). \quad (15)$$

It can be proved that, for the covariance matrix C_μ with the elements defined by Equation (15), there is no negative eigenvalue. Thus, the prediction rule is formulated by

$$F_M(\mathrm{X}, \mathrm{X}^{\xi}) = \mathrm{Min}\,\{F_M(\mathrm{X}, \mathrm{X}^{\alpha}), F_M(\mathrm{X}, \mathrm{X}^{\beta}), F_M(\mathrm{X}, \mathrm{X}^{\alpha/\beta}), F_M(\mathrm{X}, \mathrm{X}^{\alpha+\beta})\}. \quad (16)$$

Note that the Mahalanobis discriminant function F_{M} as defined by Equation (14) is no longer a distance because it does not satisfy the condition of $F_{\mathrm{M}}(\mathbf{X},\mathbf{X}^{\mu}) = 0$ when $\mathbf{X} \equiv \mathbf{X}^{\mu}$, and also it may have a negative value, obviously violating the condition that a distance must be nonnegative.

Besides correctly using a prediction algorithm, it is also very important to have a reasonable dataset of protein structural classes. If a dataset 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 dataset would be meaningless. To avoid this, the data in this study are from SCOP database.[9] The SCOP database can be accessed at http://scop.mrc-lmb.cam.ac.uk/scop/. As mentioned above, in the SCOP database the basic unit for classification is the protein domain. Therefore, unless a domain is formed by an entire protein chain, it is usually marked with the starting and end residue positions of the domain. Listed in Table I are 232 domains extracted from the SCOP database. In this table, each domain is expressed by a symbol of A|B, where A is the Protein Data Bank (PDB) code, and B the sequence region. When a domain is formed by a whole chain, B = W.C.; otherwise, B contains two numbers to indicate its starting and end points along the corresponding protein chain. From the PDB files, the corresponding DSSP (definition of the secondary structures of proteins) files were converted by the program DSSP.[16] Based on these DSSP files, the amino acid composition and the secondary-structure contents for each of these protein domains were computed. The former will serve as the only input for the structural class prediction, while the latter will be used for an analysis later.

RESULTS AND DISCUSSION

The prediction quality was examined by both resubstitution and jackknife tests.

Resubstitution Test

The so-called resubstitution test is an examination for the self-consistency of a prediction algorithm. When the resubstitution test is performed for the current study, the structural class for each of the domains in a given dataset is predicted using the rules derived from the same dataset, the so-called development dataset or training dataset. The results thus predicted for the 232 domains in Table I are summarized in Table II, from which we can see that the overall rate of correct prediction by the component-coupled algorithm is more than 30% higher than those by the simple geometry distance algorithms,[2,3,5] indicating a significant improvement in the self-consistency by taking into account the component-coupled effect. This is fully consistent with the results reported earlier[7,10] for the dataset in which the subset sizes are identical. As shown above, by the resubstitution examination, the structural class of each domain from a dataset is predicted using the rules derived from the same dataset, the so-called training dataset. As a consequence, the parameters derived from the training dataset include the information of a domain later plugged back in the test. This will certainly give a somewhat optimistic error estimate because the same domains are used to derive the prediction rules and to test themselves. Nevertheless, the resubstitution examination is absolutely necessary because it reflects the self-consistency of a prediction method, especially for its algorithm part. A prediction algorithm certainly cannot be deemed as a good one if its self-consistency is poor. In other words, the resubstitution examination is necessary but not sufficient for evaluating a prediction method. As a complement, a cross-validation examination for an independent testing dataset is needed because it can reflect the extrapolating effectiveness of a prediction method. This is important especially for checking the validity of a training database: whether it contains sufficient information to reflect all the important features concerned so as to yield a high success rate in application.

Jackknife Test

As is well known, the single-test-set analysis, subsampling, and jackknife analysis are the three tests often used for cross-validation examination. Among these three, the jackknife test is deemed the most effective and objective one. The jackknife test is also called leave-one-out test,[4,11] in which each domain in the dataset is in turn singled out as a tested domain and all the rule parameters are calculated without using this domain. 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 jackknife analysis, both the training dataset and testing dataset are actually open, and a domain will in turn move from one to the other.

The results of jackknife test obtained by the three different prediction algorithms for the 232 domains of Table I are given in Table III, from which the following phenomena can be observed. First, as expected, the rates of correct prediction by jackknife test are decreased compared with those by resubstitution. Such a decrement is more remarkable for the component-coupled algorithm than for the simple

TABLE I. The 232 Domains Classified into Four Classes*

51 all-α domains

1hbiA|W.C.	1sctA|W.C.	1ytc_|W.C.	1boc_|W.C.	1ctz_|W.C.
1fipA|W.C.	1hddC|W.C.	1dprA|65–136	1tnt_|W.C.	1bbl_|W.C.
1erc_|W.C.	1aca_|W.C.	1vasA|W.C.	1lynA|W.C.	1hme_|W.C.
1hsm_|W.C.	1gnc_|W.C.	1rprA|W.C.	1pou_|W.C.	1cdn_|W.C.
1cih_|W.C.	1argA|W.C.	1mykA|W.C.	1mylA|W.C.	1bpd_|9–91
1olhA|W.C.	1pesA|W.C.	1rpo_|W.C.	1hns_|W.C.	1tag_|57–177
1bod_|W.C.	2pccB|W.C.	4ts1A|228–319	1tyc_|228–319	1lgaA|W.C.
1oxy_|1–379	1nol_|1–379	1pgn_|177–473	1yeb_|W.C.	2utgA|W.C.
3gly_|W.C.	1csi_|W.C.	1csc_|W.C.	1phb_|W.C.	3fisA|W.C.
1troA|W.C.	3wrp_|W.C.	1trrA|W.C.	1grl_|6–136	1raq_|W.C.
1afb1|73–104				

56 all-β domains

1mdtA|381–535	1cgt_|580–684	1cxe_|582–686	1aaj_|W.C.	1mdaA|W.C.
1gcs_|1–85	1pnf_|1–140	1png_|5–140	1gog_|1–150	1tnfA|W.C.
1hivA|W.C.	1thu_|W.C.	2ctvA|W.C.	2tunA|W.C.	1apnA|W.C.
2cna_|W.C.	1bib_|271–317	1ltaD|W.C.	1bfb_|W.C.	2bfh_|W.C.
1bfg_|W.C.	1bas_|W.C.	1fnd_|19–154	1arc_|W.C.	1bcmA|481–560
1hpxA|W.C.	1thv_|W.C.	1hshA|W.C.	1bzm_|W.C.	1cpiA|W.C.
1hvc_|W.C.	1hefE|W.C.	1hvsA|W.C.	1gtsA|339–547	1hbp_|W.C.
1fen_|W.C.	1fga_|W.C.	1erb_|W.C.	1slfB|W.C.	1azm_|W.C.
1srgA|W.C.	1srjA|W.C.	1ptsA|W.C.	1sleB|W.C.	1cyhA|W.C.
3cysA|W.C.	2sim_|W.C.	1gog_|151–537	1cgt_|383–494	1cxe_|383–495
1hug_|W.C.	1mikA|W.C.	1huh_|W.C.	1akl_|247–470	1hpcA|W.C.
1kraC|2–129				

66 α/β domains

1cgt_|1–382	1cxe_|1–382	1cgv_|1–382	1btb_|W.C.	1brsD|W.C.
1cxf_|1–382	1fnd_|155–314	4ts1A|1–217	1selA|W.C.	1cdoA|176–324
1hldA|175–324	1horA|W.C.	2secE|W.C.	1cia_|W.C.	1frn_|155–314
1pnt_|W.C.	2hnp_|W.C.	1tybE|1–217	1tho_|W.C.	1tkbA|535–680
1lam_|1–159	1bllE|1–159	1gdtA|1–140	3hsc_|3–188	1idm_|W.C.
1ngi_|4–188	1atr_|2–188	1cde_|W.C.	1grcA|W.C.	1cddA|W.C.
1mhtA|W.C.	1ama_|W.C.	1alhA|W.C.	1ula_|W.C.	1ngb_|4–188
1rhd_|1–149	1trx_|W.C.	1amn_|W.C.	8atcA|1–150	1acj_|W.C.
1alkA|W.C.	2ctc_|W.C.	1drl_|W.C.	1drj_|W.C.	1hqaA|W.C.
1ajdA|W.C.	1acl_|W.C.	1ngg_|3–188	1ajcA|W.C.	1dbp_|W.C.
1xab_|W.C.	1raiA|1–150	1scnE|W.C.	1ttqB|W.C.	1wsyB|W.C.
1orb_|1–149	1ajaA|W.C.	2anhA|W.C.	5acn_|1–528	5cpa_|W.C.
2bgt_|W.C.	1drk_|W.C.	1acmA|1–150	1ngh_|4–188	1olcA|W.C.
1ctu_|1–150				

59 α + β domains

1fut_|W.C.	2baa_|W.C.	1aec_|W.C.	2rat_|W.C.	2rns_|W.C.
1ras_|W.C.	1ssbA|W.C.	1rbd_|W.C.	1kraA|W.C.	1pgx_|W.C.
1pgb_|W.C.	1igcA|W.C.	2igg_|W.C.	2igh_|W.C.	2secI|W.C.
1coy_|319–450	3monA|W.C.	1frtA|1–178	1fkj_|W.C.	2tecI|W.C.
1lttA|W.C.	1egl_|W.C.	1sbnI|W.C.	3mdsA|93–203	1vig_|W.C.
1egpA|W.C.	1fkl_|W.C.	1mns_|3–132	1grl_|137–190	1fccC|W.C.
1rldS|W.C.	1comA|W.C.	1sphA|W.C.	1gaeO|149–312	1mstA|W.C.
1grb_|364–478	1lklA|W.C.	1lcjA|W.C.	1lckA|117–226	1sceA|W.C.
1setA|111–421	1sibI|W.C.	1tsdA|W.C.	1htlA|W.C.	1bmsA|W.C.
2hpr_|W.C.	1tsy_|W.C.	1tys_|W.C.	3b5c_|W.C.	1tbpA|61–155
1xrc_|1–101	1glv_|123–316	2tscA|W.C.	3dni_|W.C.	1dnkA|W.C.
4mdhA|155–333	1mrk_|W.C.	1ltaA|W.C.	1ltgA|W.C.	

*The four classes are all-α, all-β, α/β, α + β by Murzin et al.,[9] based on the evolutionary relationships and on the principles that govern their 3D structure. Each domain is expressed by a symbol of A|B, where A is the corresponding PDB code, and B the sequence region. The fifth character in the PDB code indicates a specific chain of a protein; if it is _, the corresponding protein has only one chain. When a domain is constituted by a whole chain, B = W.C.; otherwise, B would contain two numbers to indicate its starting and end points along a sequence.

TABLE II. Predicted Results for the 232 Domains of Table I in Resubstitution Tests

	Rate of correct prediction for each class				Overall rate of
Methods	all-α	all-β	α/β	$\alpha + \beta$	correct prediction
Hamminig distance (Eq. 7)	$^{33}/_{51} = 64.71\%$	$^{35}/_{56} = 62.50\%$	$^{42}/_{66} = 63.64\%$	$^{27}/_{59} = 45.76\%$	$^{137}/_{232} = 59.05\%$
Euclidean distance (Eq. 9)	$^{36}/_{51} = 70.59\%$	$^{36}/_{56} = 64.29\%$	$^{32}/_{66} = 48.48\%$	$^{25}/_{59} = 42.37\%$	$^{129}/_{232} = 55.60\%$
Component-coupled (Eq. 16)	$^{49}/_{51} = 96.08\%$	$^{53}/_{56} = 94.64\%$	$^{63}/_{66} = 95.45\%$	$^{55}/_{59} = 93.22\%$	$^{220}/_{232} = 94.83\%$

TABLE III. Predicted Results for the 232 Domains of Table I in Jackknife (Leave-One-Out) Tests

	Rate of correct prediction for each class				Overall rate of
Methods	all-α	all-β	α/β	$\alpha + \beta$	correct prediction
Hamminig distance (Eq. 7)	$^{31}/_{51} = 60.78\%$	$^{35}/_{56} = 62.50\%$	$^{42}/_{66} = 63.64\%$	$^{25}/_{59} = 42.37\%$	$^{133}/_{232} = 57.33\%$
Euclidean distance (Eq. 9)	$^{32}/_{51} = 62.75\%$	$^{34}/_{56} = 60.71\%$	$^{31}/_{66} = 46.97\%$	$^{25}/_{59} = 42.37\%$	$^{122}/_{232} = 52.59\%$
Component-coupled (Eq. 16)	$^{42}/_{51} = 82.35\%$	$^{44}/_{56} = 78.57\%$	$^{51}/_{66} = 77.27\%$	$^{37}/_{59} = 62.71\%$	$^{174}/_{232} = 75.00\%$

geometry distance algorithms, especially for a small dataset. This is because the component-coupled algorithm needs more training data to make its prediction mechanism work properly. Therefore, the information loss resulting from jackknife will have greater impact on the predicted results by the component-coupled algorithm than those by the simple geometry distance algorithms. Nevertheless, the overall jackknife rate for the dataset of 232 domains by the component-coupled algorithm is still about 18–22% higher than those by the simple geometry distance algorithms. Furthermore, it was observed that, by enlarging the dataset for the jackknife test, the overall rate by the component-coupled algorithm was increased much faster than the corresponding rates by the simple geometry distance algorithms. For example, if the dataset is expanded from 232 to 359 domains, the overall jackknife rate by the component-coupled algorithm is increased to 84%, while the corresponding rates by the least Hamming and Eucleadian distance algorithms become 52% and 41%; the former is about 32–43% higher than the latter. Accordingly, by improving or expanding a database to reduce the information loss, the difference in favor of the component-coupled algorithm will become even larger. This also indicates that the jackknife-tested rate will be close to the resubstitution-tested rate when the datasets approach an ideal one where every entry has a good representative other than itself. Therefore, while jackknife tests are usually deemed as objective tests to compare different prediction algorithms for a given dataset, the resubstitution test is useful for testing the potential of a new algorithm when the working dataset is small and not ideal.

The protein structural class prediction has also been studied by other investigators via neural networks.[17–19] However, in these studies no jackknife-tested rates but only the rate for a training dataset and the rate for a selected testing dataset were reported. Therefore, their results can not be compared with the results here. Actually, for a selected independent testing dataset, the rate of correct prediction by means of the component-coupled algorithm can be higher than 80%[20] and even 90%.[7,10]

The above results by both resubstitution and jackknife tests indicate that the incorporation of the coupling effect among different amino acid components is significant for improving the prediction quality of the domain structural classes. A question might be raised here as to why a complete opposite conclusion was reached by Eisenharber et al.[8] The reasons are as follows. (1) Misuse of algorithm: the component-coupled algorithm used by them was Equation (13), which is valid only for training sets consisting of subsets of the same or approximately the same size, while the subset sizes in their training sets are not so but very much different. For example, the Table I of their paper[8] contains four training datasets: the first one consists of 41 α, 54 β, 70 mixed, 1 irregular; the second consists of 55 α, 78 β, 127 mixed, 2 irregular; the third consists of 84 α, 103 β, 206 mixed, 5 irregular; and the fourth consists of 99 α, 140 β, 232 mixed, 4 irregular. When dealing with cases like these, two important factors must be explicitly incorporated as formulated in Equations (14) and (15), and predictions should be based on Equation (16) instead of Equation (13), as described above. Moreover, all the above four datasets contain a very small subset that is statistically insignificant and should be removed from training data according to common sense. Without realizing these two points, one might misapply the component-coupled algorithm and get incorrect results. For example, the self-consistency by them for the 262 proteins (the second training set in Table I of their paper) was only 60.7%. However, if computed by following a correct procedure as described above, the corresponding self-consistency is 88.89%, which is about 28% higher than the result reported by them. Similar big discrepancies exist for all the other results. Therefore, all the results reported by them[8] are not the true results obtained when correctly using the component-

TABLE IV. Average (Mean ± Standard Deviation) Percentages of α-Helices, β-Strands, Parallel β-Sheets, and Antiparallel β-Sheets Derived From the DSSP Files of the 232 Domains in Table I for Each of the Four Structural Classes

Class	α-helices	β-strands	Parallel β-sheets	Antiparallel β-sheets
all-α	57 ± 16%	2 ± 5%	—	—
all-β	5 ± 6%	43 ± 9%	8 ± 13%	92 ± 13%
α/β	34 ± 7%	19 ± 5%	64 ± 25%	36 ± 25%
α + β	24 ± 10%	28 ± 10%	16 ± 21%	83 ± 24%

coupled algorithm. (2) Incorrect datasets: the datasets constructed by Eisenhaber et al.[8] were based on the rule of Equation (1), which is questionable, as can be further illustrated as follows. Although the structural class for each of the 232 domains in Table I from SCOP[9] was not based on the percentages of its secondary-structure contents, it is instructive to examine their average (mean ± standard deviation) percentage values for each class. These values can be derived from the corresponding 232 DSSP files, as given in Table IV, from which we can see that the average percentages of α-helices for the α protein class is 57% (with a standard deviation of 16%), and the average percentages of β-sheets for the β protein class is 43% (with a standard deviation of 9%). These values are about four times as large as the corresponding values defined in Equation (1) for all-α and all-β proteins, respectively. Accordingly, it is not surprising that many proteins assigned by Eisenhaber et al.[8] as belonging to the α or β class actually do not belong to either. A dataset thus formed cannot correctly reflect the relationship between the structural class of a protein and its amino acid composition.

Although the rates of correct prediction by the component-coupled algorithm for either resubstitution or jackknife test are quite high, the use of the present training datasets for practical application should be cautioned with the caveat that some protein domains might be mispredicted if they are outside the "frame" defined by the current limited database. Nevertheless, for a given dataset in which as long as the structural classes are correctly classified, the predicted rate by the component-coupled algorithm is generally much higher than those by the simple geometry algorithms. Such a fact has also been confirmed recently by Bahar et al.[20] How far the prediction quality can be improved by the component-coupled algorithm will also depend on how complete a training dataset can be formed. The above results and analysis have demonstrated that the component-coupled algorithm as formulated in this study can become a powerful tool for predicting the structural classes of domains if an ideal training database is available.

ACKNOWLEDGMENTS

We thank Dr. F. Eisenhaber for sending us his results, which stimulated the current study, especially for the problems such as how to classify protein structural classes reasonably and how to consider the effect when the numbers of proteins in the training dataset are much different for different classes.

REFERENCES

1. Levitt, M., Chothia, C. Structural patterns in globular proteins. Nature 261:552–557, 1976.
2. Chou, P.Y. 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.
3. Nakashima, H., Nishikawa, K., Ooi, T. The folding type of a protein is relevant to the amino acid composition. J. Biochem. 99:152–162, 1986.
4. Klein, P., Delisi, C. Prediction of protein structural class from amino acid sequence. Biopolymers 25:1659–1672, 1986.
5. Chou, P.Y. Prediction of protein structural classes from amino acid composition. In "Prediction of Protein Structure and the Principles of Protein Conformation." Fasman, G.D. (ed.). New York: Plenum Press, 1989:549–586.
6. Kneller, D.G., Cohen, F.E., Langridge, R. Improvements in protein secondary-structure prediction by enhanced neural networks. J. Mol. Biol. 214:171–182, 1990.
7. Chou, K.C. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins 21:319–344, 1995.
8. Eisenhaber, F., Frömmel, C., Argos, P. Prediction of secondary-structural content of proteins from their amino acid composition alone. II. The paradox with secondary-structural class. Proteins 25:169–179, 1996.
9. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C. SCOP: A structural classification of protein database for the investigation of sequence and structures. J. Mol. Biol. 247:536–540, 1995.
10. Chou, K.C., Zhang, C.T. Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30:275–349, 1995.
11. Mardia, K.V., Kent, J.T., Bibby, J.M. "Multivariate Analysis." London: Academic Press, 1979:322, 381.
12. Mahalanobis, P.C. On the generalized distance in statistics. Proc. Natl. Inst. Sci. India 2:49–55, 1936.
13. Pillai, K.C.S. Mahalanobis D/u^2. In: "Encyclopedia of Statistical Sciences." Vol. 5. Kotz, S., Johnson, N.L. (eds.). New York: John Wiley & Sons, Inc., 1985:176–181. (This reference also presents a brief biography of Mahalanobis, who was a man of great originality and who made considerable contributions to statistics.)
14. Cedano, J., Aloy, P., Pérez-Pons, J.A., Querol, E. Relation between amino acid composition and cellular location of proteins. J. Mol. Biol. 266:594–600, 1997.
15. Liu, W., Chou, K.C. Prediction of protein structural classes by modified Mahalanobis discriminant algorithm. J. Protein Chem. 17:in press.
16. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
17. Metfessel, B.A., Saurugger, P.N., Connelly, D.P., Rich, S.T. Cross-validation of protein structural class prediction using statistical clustering and neural networks. Protein Sci. 2:1171–1182, 1993.
18. Dubchak, I., Holbrook, S.R., Kim, S.-H. Predicting protein secondary-structure content: A tandem neural network approach. Proteins 16:79–91, 1993.
19. Chandonia, J.M., Karplus, M. Neural networks for secondary structure and structural class prediction. Protein Sci. 4:275–285, 1995.
20. Bahar, I., Atilgan, A.R., Jernigan, R.L., Erman, B. Understanding the recognition of protein structural classes by amino acid composition. Proteins 29:172–185, 1997.