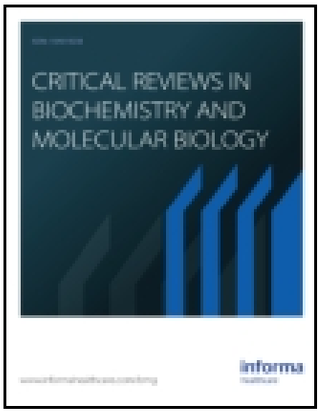

Critical Reviews in Biochemistry and Molecular Biology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ibmg20

Prediction of Protein Structural Classes

Kuo-Chen Chou[a] & Chun-Ting Zhang[b]
[a] Computer-Aided Drug Discovery, Upjohn Laboratories, Kalamazoo, MI, 49007-4940
[b] Department of Physics, Tianjin University, Tianjin, China
Published online: 26 Sep 2008.

To cite this article: Kuo-Chen Chou & Chun-Ting Zhang (1995) Prediction of Protein Structural Classes, Critical Reviews in Biochemistry and Molecular Biology, 30:4, 275-349

To link to this article: http://dx.doi.org/10.3109/10409239509083488

Critical Reviews in Biochemistry and Molecular Biology, 30(4):275–349 (1995)

Prediction of Protein Structural Classes

Kuo-Chen Chou
Computer-Aided Drug Discovery, Upjohn Laboratories, Kalamazoo, MI 49007-4940

Chun-Ting Zhang
Department of Physics, Tianjin University, Tianjin, China

Referee: Gerald D. Fasman, Graduate Dept. of Biochemistry, Brandeis University, Waltham, MA

ABSTRACT: A protein is usually classified into one of the following five structural classes: α, β, $\alpha + \beta$, α/β, and ζ (irregular). The structural class of a protein is correlated with its amino acid composition. However, given the amino acid composition of a protein, how may one predict its structural class? Various efforts have been made in addressing this problem. This review addresses the progress in this field, with the focus on the state of the art, which is featured by a novel prediction algorithm and a recently developed database. The novel algorithm is characterized by a covariance matrix that takes into account the coupling effect among different amino acid components of a protein. The new database was established based on the requirement that the classes should have (1) as many nonhomologous structures as possible, (2) good quality structure, and (3) typical or distinguishable features for each of the structural classes concerned. The very high success rate for both the training-set proteins and the testing-set proteins, which has been further validated by a simulated analysis and a jackknife analysis, indicates that it is possible to predict the structural class of a protein according to its amino acid composition if an ideal and complete database can be established. It also suggests that the overall fold of a protein is basically determined by its amino acid composition.

KEY WORDS: overall fold, amino acid composition, coupling effect, Mahalanobis distance, seed-propagated sampling, jackknife analysis

1040-9238/95/$.50

List of Symbols

C,	The component coefficient (Equation 10);
D^2,	Mahalanobis distance (Equations 20 and 38);
d^{E},	Euclidian distance (Equation 8);
d^{H},	Hamming distance (Equation 5);
d^{M},	Minkowski's distance (Equation 9);
$F_{k,i}(\alpha)$,	The accumulated probability of the ith component of the kth α protein (Equation 41);
N^*,	Number of simulated proteins for a given class;
N_S,	Number of subsampling cycles (Equation 43);
$\mathbf{P}$,	A protein defined in a 19-D space (Equation 38);
$\overline{\mathbf{P}}$,	The norm of a given protein set defined in a 19-D space (Equation 36);
$\mathbf{P}_k$,	The kth protein defined in a 19-D space (Equation 35);
$\mathbf{P}_\alpha$,	The norm of an α protein subset defined in a 19-D space (Equation 39);
$\mathbf{P}_\beta$,	The norm of a β protein subset defined in a 19-D space (Equation 39);
$\mathbf{P}_{\alpha+\beta}$,	The norm of an $\alpha + \beta$ protein subset defined in a 19-D space (Equation 39);
$\mathbf{P}_{\alpha/\beta}$,	The norm of an α/β protein subset defined in a 19-D space (Equation 39);
$\mathbf{P}_\zeta$,	The norm of a ζ (irregular) protein subset defined in a 19-D space (Equation 39);
$\mathbf{Q}$,	Covariance matrix defined in a 19-D space (Equation 37);
$\mathbf{q}$,	The average accuracy (Equation 7);
$\mathbf{S}$,	Covariance matrix defined in a 20-D space (Equation 21);
$s_{i,j}$,	The element of the ith row and jth column in $\mathbf{S}$ (Equation 22);
$\mathbf{U}$,	A unitary matrix formed by eigenvectors (Equation 27);
$u_{i,j}$,	The element of the ith row and jth column in $\mathbf{U}$ (Equation 27);
$\mathbf{Y}$,	A 1×20 transformation matrix associated with eigenvectors, norms, and the protein to be predicted (Equation 30);
y_i,	The ith element of matrix $\mathbf{Y}$ (Equation 31);
$\mathbf{X}$,	A protein defined in a 20-D space (Equation 5);
$\overline{\mathbf{X}}$,	The norm of a given protein set defined in a 20-D space (Equation 3);
$\mathbf{X}_k$,	The kth protein defined in a 20-D space (Equation 2);
$\mathbf{X}_\alpha$,	The norm of an α-protein subset defined in a 20-D space (Equations 6 and 34);
$\mathbf{X}_\beta$,	The norm of a β-protein subset defined in a 20-D space (Equations 6 and 34);
$\mathbf{X}_{\alpha+\beta}$,	The norm of an $\alpha + \beta$ protein subset defined in a 20-D space (Equations 6 and 34);
$\mathbf{X}_{\alpha/\beta}$,	The norm of an α/β protein subset defined in a 20-D space (Equations 6 and 34);
$\mathbf{X}_\zeta$	The norm of a ζ (irregular) protein subset defined in a 20-D space (Equations 6 and 34);
x_i,	The ith component of protein $\mathbf{X}$ (Equation 5);

$x_{k,i}$, The ith component of protein $\mathbf{X}_k$ (Equation 2);
$x_{k,i}(\alpha)$, The ith component of the kth α protein (Equation 40);
$x^*_{k,i}(\alpha)$, The ith component of a simulated protein generated based on the kth α protein (Equation 43);
$\overline{x_i}$ The ith component of protein $\overline{\mathbf{X}}$ (Equation 4);
Θ, The correlation angle (Equations 16 and 18);
λ_i, The ith eigenvalue (Equation 25);
Π, The projection (Equation 16);
Ψ_i, The ith eigenvector (Equation 25);
$\psi_{i,j}$, The jth component of Ψ_i (Equation 25);

I. Introduction

The function of a protein is based on its structure. Knowledge of protein structure plays a key role in molecular biology, cell biology, pharmacology, and medical science. However, despite years of both experimental and theoretical study, the determination of protein structure remains one of the most difficult problems.

Although X-ray crystallographic and nuclear magnetic resonance (NMR) techniques are two powerful experimental tools, both require expensive equipment and take months or even years to determine the structure of a single protein. Moreover, during the working process, there may be some difficulties that cannot always be solved.

Current theoretical approaches to predicting the structure of a protein can be classified into two categories. One is the free-energy minimization method, which is based on the empirical atomic potentials (see, e.g., Scheraga, 1968, 1987; Weiner and Kollman, 1981; Levitt, 1983; Gilson and Honig, 1988; McCammon et al., 1989; Mackay et al., 1989; Rogers, 1989; Chou et al., 1990; Karplus and Shakhnovich, 1992). The other is the statistical method, which was developed based on various statistical data extracted from structure-known proteins (see, e.g., Chou and Fasman, 1974, 1978; Lim, 1974; Garnier et al., 1978; Cid et al., 1982; Fasman, 1989; Orengo et al., 1994; Jones et al., 1994). Meanwhile, various physical theories of protein structures at different levels have been proposed for understanding the mechanism and patterns of protein folding, and for improving the prediction of protein structure as well (Ptitsyn and Rashin, 1975; Ptitsyn and Finkelstein, 1979, 1980, 1989; Ptitsyn et al., 1985; Finkelstein and Ptitsyn, 1987; Chothia and Finkelstein, 1990; Kuwajima et al., 1993; Kolinski and Skolnick, 1994; Vieth et al., 1994; Mitchell et al., 1994; McDonald and Thornton, 1994).

Although one can in principle calculate the detailed atomic coordinates by means of the free-energy minimization method, it is in practice very difficult to find the real global minimum because of the computational time required to search the vast number of conformations accessible to even a short polypeptide. Therefore, in most cases, the energy minimization method was used merely to refine a protein structure determined by X-ray crystallographic and NMR techniques, or to reveal the origin of the structural handedness in proteins (Chou and Scheraga, 1982; Chou et al., 1983, 1989, 1990; Chou and Carlacci, 1991a). Although the

simulated annealing approach is a powerful tool in overcoming the local minimum problem (Kawai et al., 1989; Wilson and Cui, 1990; Chou and Carlacci, 1991b; Chou, 1992), using it to predict the conformation of a protein is still impractical. Therefore, progress in predicting the structure of a protein by *ab initio* energy calculation based only on its primary sequence has been slower. An alternative strategy, the knowledge-based energy calculation approach, which relies on known motifs or folds in sequences, looks more promising (Cohen and Kuntz, 1987; Carlacci et al., 1991; Chou, 1992; Jones et al., 1992; Wodak and Rooman, 1993; Fetrow and Bryant, 1993). However, compared with the *ab initio* calculation, the knowledge-based approach lacks the beauty of unification. Besides, there are not many proteins yet for which one can obtain enough knowledge (input) to result in a successful prediction. On the other hand, although the statistical method can only predict an outline of a structure, it has the merit of simplicity and convenience in application, and hence has been widely applied by biochemists. Besides, the results predicted by the statistical methods, although rough, may be used to reduce the scope of searching conformational space or set good starting points for the energy minimization method. Therefore, in determining the three-dimensional (3-D) structure of a protein, the statistical method has played an important complementary role.

It is instructive to note that although the details of the 3-D structures of proteins are extremely complicated and irregular, their overall folding patterns are surprisingly simple, regular, and even strikingly beautiful from an aesthetic point of view (Finkelstein and Ptitsyn, 1987; Chou and Carlacci, 1991a). Proteins and the domains therein often have similar or identical folding patterns even if they are quite different according to their sequences or biochemical functions (Richardson, 1977, 1981; Ptitsyn and Finkelstein, 1980), suggesting some physical limitation for the folding of protein (Finkelstein and Ptitsyn, 1987; Chou et al., 1990).

During the last decade, the prediction of protein structural classes by the statistical method has become quite active due to the following reasons. The concept of structural class based on the contents of secondary structures in a protein is very useful from both the experimental and the theoretical points of view. The structural class of a protein presents an intuitive description of its overall fold, which can be directly determined by relatively simple spectroscopic methods such as circular dichroism (CD) spectroscopy in the UV absorption range (Johnson, 1990; Perczel et al., 1991; Sreerama and Woody, 1994) and IR Raman spectroscopy (Bussian and Sander, 1989). The restrictions of the structural class of a protein have a high impact on its secondary and tertiary structure prediction. The accuracy of secondary structure prediction from amino acid sequence can be significantly improved by incorporating the effect of knowledge of structural class (Chou, P.Y., 1989; Deléage and Roux, 1987, 1989; Kneller et al., 1990; Muggleton et al., 1992; Cohen et al., 1993). Knowledge of the structural class may also be used to reduce the scope of searching conformational space during energy optimization, and provide useful information for a heuristic approach (e.g., Cohen and Kuntz, 1987; Carlacci et al., 1991) to find the tertiary structure of a protein. In addition to playing an important role in improving the calculation of the hydrophobicity coefficients (Cid et al., 1992), the structural class is related to various properties of a protein such as its location in extra- or intracellular compartments, biological function (being an enzyme or not), or the

existence of disulfide bonds (Nishikawa and Ooi, 1982; Nishikawa et al., 1983a, b). Another important driving force for the progress in this regard has been the unprecedented increase in the number of high-resolution protein structures (over 800 coordinate sets are available in the latest version of the Brookhaven Protein Databank). Statistical analyses aimed at deriving rules or fundamental principles of protein structural classes can now be based on more solid ground than ever before. Various prediction methods have been proposed. This review discusses the progress of these methods, with a focus on the prospects in relation to the current state of the art.

II. Classification of Protein Structural Classes

Proteins of known structures are generally classified into one of the following five structural classes: α, β, $\alpha + \beta$, α/β, and irregular proteins (Levitt and Chothia, 1976; Richardson and Richardson, 1989). The classification is based on the percentage of secondary structure components, although there is no unified quantitative criterion yet. Suppose the percentages of α-helix and β-sheet in a protein are abbreviated by α and β, respectively. The classification by Nakashima et al. (1986) was made according to the following criterion: α-proteins, $\alpha > 15\%, \beta < 10\%$; β-proteins, $\alpha < 15\%, \beta > 10\%$; $\alpha+\beta$-proteins, $\alpha > 15\%, \beta > 10\%$ with dominantly antiparallel β-sheets; α/β-proteins, $\alpha > 15\%, \beta > 10\%$ with dominantly parallel β-sheets; and ζ (irregular)-proteins, $\alpha < 15\%, \beta < 10\%$. According to P.Y. Chou (1989), however, proteins were classified as follows: α-proteins, $\alpha > 45\%, \beta < 5\%$; β-proteins, $\alpha < 5\%, \beta > 45\%$; $\alpha+\beta$-proteins, $\alpha > 30\%, \beta > 20\%$ with dominantly antiparallel β-sheets; α/β-proteins, $\alpha > 30\%, \beta > 20\%$ with dominantly parallel β-sheets. The classification by Nakashima et al. (1986) covered 135 proteins, of which 31 were α-, 34 β-, 27 $\alpha + \beta$-, 39 α/β-, and 4 ζ-proteins. The classification by P.Y. Chou (1989) covered 64 proteins: 19 α-, 15 β-, 14 $\alpha + \beta$-, and 16 α/β-proteins, but no irregular proteins. Although the classification by Nakashima et al. (1986) covers more proteins than those by P.Y. Chou (1989), the relevant percentages set by them 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 other words, α- or β-protein should have at least $\alpha \geq 40\%$ or $\beta \geq 40\%$, respectively. Besides, no quantitative definition whatsoever was given for the term "dominantly" mentioned in both of the two classification methods, and this would certainly cause arbitrariness or ambiguity in discerning $\alpha + \beta$- and α/β-proteins. For example, proteins with PDB codes 0PHH, 1LDX, 2MDH, 2GPD, 2GRS, 4ADH, and 4LDH were classified by Nakashima et al. (1986) as the α/β-structural class. However, an analysis of the secondary structure contents for these proteins indicates that of their β-sheet component, the parallel sheets only occupy 25 to 50%. Obviously, for these proteins, the parallel β-sheet should not be interpreted as a "dominant" part over its antiparallel counterpart. A similar problem has also been found for the α/β-proteins classified by P.Y. Chou (1989).

In view of the above, a new classification has been proposed (Chou, 1995a) that catego-

rizes proteins according to the following quantitative criterion:

$$\begin{cases} \alpha\text{-proteins} & \alpha \geq 40\%,\ \beta \leq 5\%, \\ \beta\text{-proteins} & \alpha \leq 5\%,\ \beta \geq 40\%, \\ \alpha+\beta\text{-proteins} & \alpha \geq 15\%,\ \beta \geq 15\% \\ & \text{with more than 60\% antiparallel } \beta\text{-sheets}, \\ \alpha/\beta\text{-proteins} & \alpha \geq 15\%,\ \beta \geq 15\% \\ & \text{with more than 60\% parallel } \beta\text{-sheets}, \\ \zeta(\text{irregular})\text{-proteins} & \alpha \leq 10\%,\ \beta \leq 10\% \end{cases} \quad (1)$$

Based on the new criteria, 30 representative proteins have been selected for each of the four regular structural classes, together with nine proteins for the irregular structural class. Most of these representative proteins were selected from a list of protein chains with less than 25% sequence identity (Hobohm and Sander, 1994) in order to reduce the redundancy caused by many homologous proteins in the Protein Data Bank. Additionally, the selection was made from those proteins with a good quality of structural determination, as well as a typical or distinguishable feature for classification. To provide an intuitive feeling, five Richardson ribbon drawings are given in Figure 1A-E, each representing one of the five structural classes.

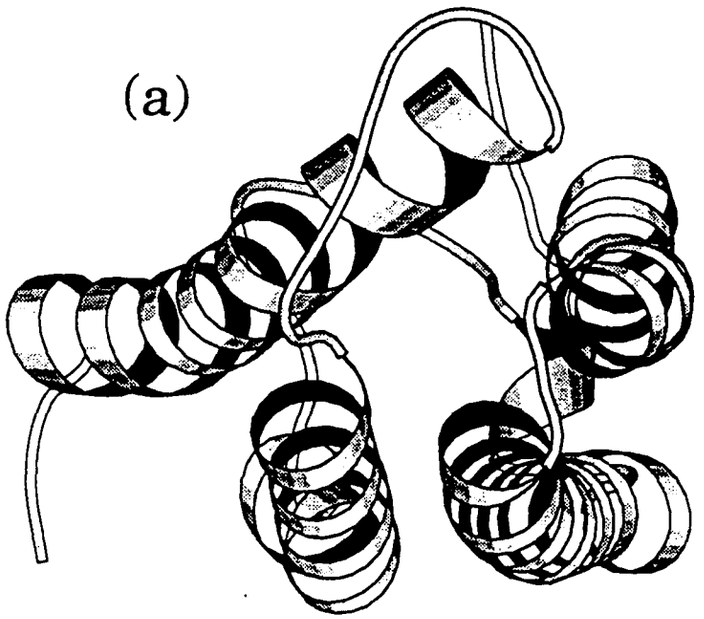

Figure 1A: The Richardson ribbon drawing to show a typical α-protein. The structure represents the A chain of cytochrome c' (Finzel et al., 1985) with PDB code 2CCYA. See Equation 1 for the definition of α-proteins.

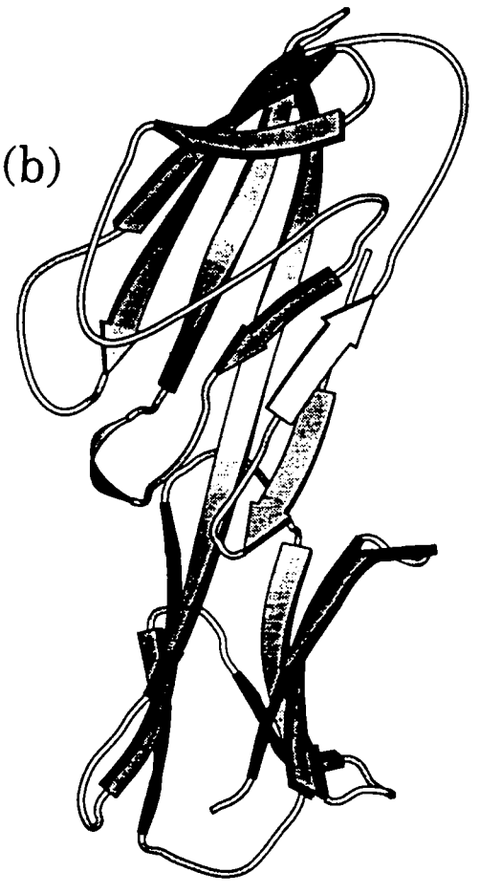

Figure 1B: The Richardson ribbon drawing to show a typical β-protein. The structure represents a T-cell surface glycoprotein (Brady et al., 1995) with PDB code 1CID−. See Equation 1 for the definition of β-proteins.

The $4 \times 30 + 9 = 129$ representative proteins with their PDB codes, the ratios of secondary structure components computed according to the dictionary by Kabsch and Sander (1983), and their amino acid compositions are given in Appendix A, as these data are often requested by readers and they are used as an important database in this article. (Note that a $\pm 5\%$ flexibility should be allowed for the criteria in Equation 1 owing to the errors in calculating the secondary structure components.)

III. A Summary of Existing Prediction Methods

The structural class of a protein is correlated with its amino acid composition (P.Y. Chou, 1980, 1989; Nakashima et al., 1986; Muskal and Kim, 1992; Mao et al., 1994). However,

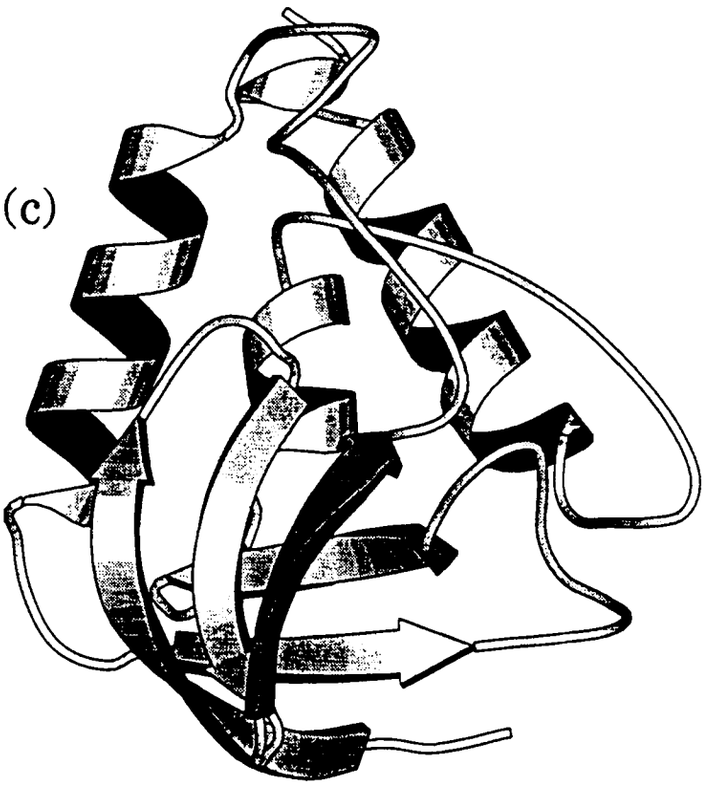

Figure 1C: The Richardson ribbon drawing to show a typical $\alpha + \beta$-protein. The structure represents a hydrolase (Loll and Lattman, 1989) with PDB code 1SNC−. See Equation 1 for the definition of $\alpha + \beta$-proteins.

given the amino acid composition of a protein, how may one predict its structural class? Various methods have been proposed to solve this problem, and they can be summarized as follows.

A. The Least Hamming Distance Method (P.Y. Chou, 1980, 1989)

The premise of this method is that the similarity of any two protein molecules is reflected through their Hamming distance, or city-block metric or Manhattan metric (Mardia et al., 1979), as can be formulated as follows. Suppose there is a set of N proteins, each of which corresponds to a vector or a point in the 20-D space,

$$\mathbf{X}_k = \begin{bmatrix} x_{k,1} \\ x_{k,2} \\ \vdots \\ x_{k,20} \end{bmatrix} \quad (k = 1, 2, \ldots, N) \tag{2}$$

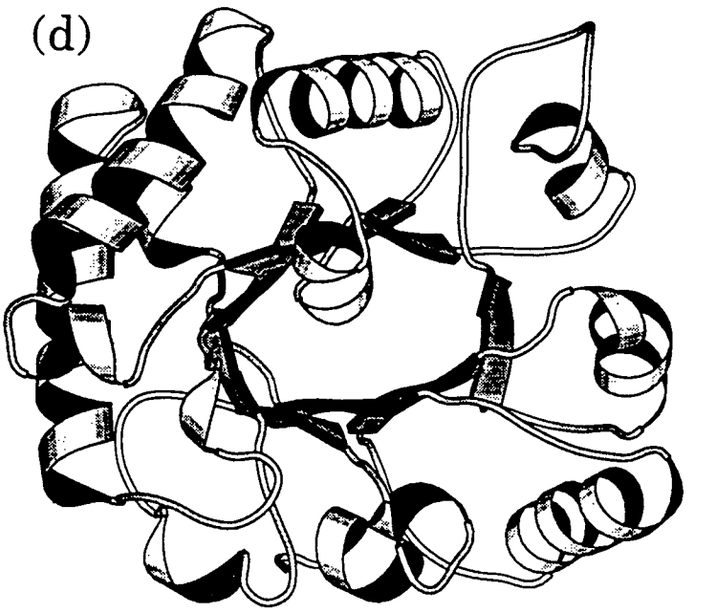

Figure 1D: The Richardson ribbon drawing to show a typical α/β-protein. The structure represents the A chain of triose phosphate isomerase (Banner et al, 1976) with PDB code 1TIMA. See Equation 1 for the definition of α/β-proteins.

where $x_{k,1}$, $x_{k,2}, \ldots x_{k,20}$ are the composition components of the 20 amino acids for the kth protein $\mathbf{X}_k$. The norm of the N proteins is defined by

$$\overline{\mathbf{X}} = \begin{bmatrix} \overline{x_1} \\ \overline{x_2} \\ \vdots \\ \overline{x_{20}} \end{bmatrix} \tag{3}$$

where

$$\overline{x_i} = \frac{1}{N} \sum_{k=1}^{N} x_{k,i} \quad (i = 1, 2, \ldots, 20) \tag{4}$$

Suppose $\mathbf{X}$ is a protein whose structural class is to be predicted. It can be either one of the N proteins in Equation 2 or a protein outside of them. It also corresponds to a point $(x_1, x_2, \ldots, x_{20})$ in the 20-D space, where x_i is the normalized occurrence-frequency of its ith amino acid. The Hamming distance between the norm $\overline{\mathbf{X}}$ and the protein $\mathbf{X}$ in the 20-D space is defined by (Mardia et al., 1979)

$$d^{\mathrm{H}}(\mathbf{X}, \overline{\mathbf{X}}) = \sum_{i=1}^{20} |x_i - \overline{x_i}| \tag{5}$$

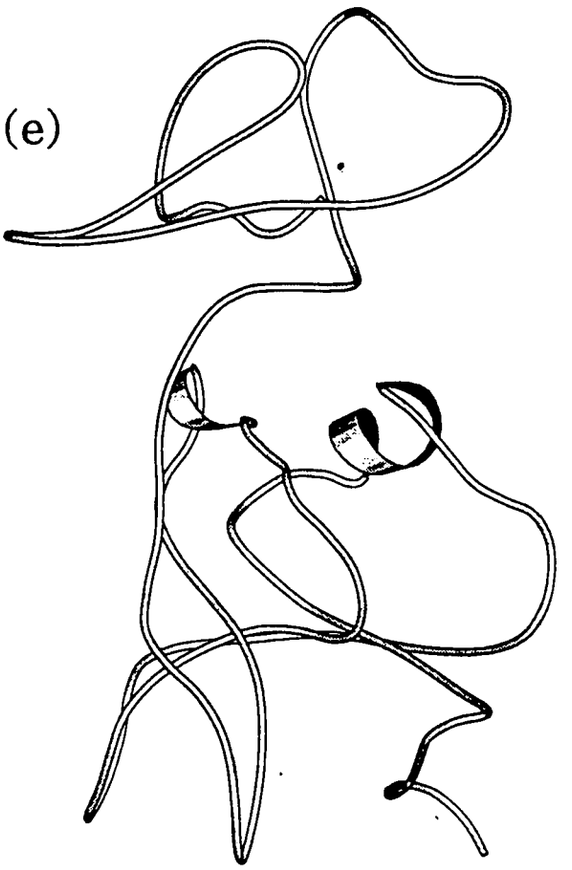

Figure 1E: The Richardson ribbon drawing to show a typical ζ-protein. The structure represents the B chain of aspartate transferase (Honzatko et al., 1982) with PDB code 2ATCB. See Equation 1 for the definition of ζ-proteins.

When the N proteins in Equation 3 are all α-proteins, $\overline{\mathbf{X}}$ thus defined would become the norm (standard point or vector) of an α-protein subset, denoted by $\mathbf{X}_\alpha$, and $d^{\mathrm{H}}(\mathbf{X}, \overline{\mathbf{X}})$ will become $d^{\mathrm{H}}(\mathbf{X}, \mathbf{X}_\alpha)$, the Hamming distance between the protein $\mathbf{X}$ and the norm of the α protein subset. Likewise, when the N proteins in Equation 3 are all β-, $\alpha + \beta$-, α/β-, or ζ-proteins, then the corresponding $\overline{\mathbf{X}}$ will become the norm of the β, $\alpha + \beta$, α/β, or ζ protein subset, denoted by $\mathbf{X}_\beta$, $\mathbf{X}_{\alpha+\beta}$, $\mathbf{X}_{\alpha/\beta}$, or $\mathbf{X}_\zeta$, respectively. The corresponding Hamming distance $d^{\mathrm{H}}(\mathbf{X}, \overline{\mathbf{X}})$ will become $d^{\mathrm{H}}(\mathbf{X}, \mathbf{X}_\beta)$, $d^{\mathrm{H}}(\mathbf{X}, \mathbf{X}_{\alpha+\beta})$, $d^{\mathrm{H}}(\mathbf{X}, \mathbf{X}_{\alpha/\beta})$, or $d^{\mathrm{H}}(\mathbf{X}, \mathbf{X}_\zeta)$, respectively. When $d^{\mathrm{H}}(\mathbf{X}, \overline{\mathbf{X}}_\mu)$ ($\mu = \alpha,\ \beta,\ \alpha+\beta,\ \alpha/\beta,$ or ζ) is smaller, meaning that protein $\mathbf{X}$ is closer to the norm of the μ-protein subset, and hence the likelihood of it belonging to the μ-protein subset is higher, and vice versa. Thus, protein $\mathbf{X}$ is predicted to be the structural class for which the corresponding Hamming distance has the least value, as can be formulated as

follows. Suppose

$$d^{\mathrm{H}}(\mathbf{X}, \mathbf{X}_\mu) = \mathbf{Min}\left\{d^{\mathrm{H}}(\mathbf{X}, \mathbf{X}_\alpha),\ d^{\mathrm{H}}(\mathbf{X}, \mathbf{X}_\beta),\ d^{\mathrm{H}}(\mathbf{X}, \mathbf{X}_{\alpha+\beta}),\ d^{\mathrm{H}}(\mathbf{X}, \mathbf{X}_{\alpha/\beta}),\ d^{\mathrm{H}}(\mathbf{X}, \mathbf{X}_\zeta)\right\} \tag{6}$$

where μ can be α, β, $\alpha + \beta$, α/β, or ζ, and the operator **Min** means taking the least one among those in the parentheses. Then, the subscript μ of Equation 6 will give the structural class to which the predicted protein **X** should belong.

By means of the least Hamming distance algorithm, predictions were performed for the 64 proteins in the training set constructed by P.Y. Chou (1989). Of the 64 proteins, 19 are α-, 15 β-, 14 $\alpha + \beta$-, and 16 α/β-proteins. The percentages of correct prediction were 16/19 (84.2%) for the α-proteins, 12/15 (80%) for the β-proteins, 11/14 (78.6%) for the $\alpha + \beta$-proteins, and 12/16 (75.0%) for the α/β-proteins. 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.,

$$q = \text{average accuracy} = \left(\frac{\text{Total number of correct prediction events}}{\text{Total number of prediction events}}\right)\% \tag{7}$$

we have the average accuracy q of 78.5% for predicting the 64 proteins by the least Hamming distance method. Predictions were also made by P.Y. Chou (1989) for 12 testing proteins not included in the training set; the percentage of correct prediction was 10/12 (83.3%).

B. The Least Euclidean Distance Method (Nakashima et al., 1986)

In this method, rather than Hamming distance, the Euclidean distance (Mardia et al., 1979) is used as a scale to measure the similarity between two proteins molecules. Therefore, instead of Equation 5 the following equation is adopted for predicting the structural class of a protein:

$$d^{\mathrm{E}}(\mathbf{X}, \overline{\mathbf{X}}) = \left[\sum_{i=1}^{20} |x_i - \overline{x_i}|^2\right]^{1/2} \tag{8}$$

According to the report by Nakashima et al. (1986), predictions were performed for the 135 proteins in the training set constructed by them. Of the 135 proteins, 31 are α-, 34 β-, 27 $\alpha + \beta$-, 39 α/β-, and 4 ζ-proteins. The percentages of correct prediction were 27/31 (87.1%) for the α-proteins, 22/34 (64.7%) for the β-proteins, 10/27 (37.0%) for the $\alpha + \beta$-proteins, 33/39 (84.6%) for the α/β-proteins, and 2/4 (50%) for the ζ-proteins, with an average accuracy q of 69.6%. However, no predicted results were reported by them for any testing proteins outside the training set.

Actually, both the Hamming distance and the Euclidean distance are special cases of the Minkowski's distance defined by (Mardia et al., 1979; Gower, 1985)

$$d^{\mathrm{M}}(\mathbf{X}, \overline{\mathbf{X}}) = \left[\sum_{i=1}^{20} |x_i - \overline{x_i}|^q\right]^{1/q} \tag{9}$$

When $q = 1$, Equation 9 reduces to the Hamming distance of Equation 5; when $q = 2$, Equation 9 reduces to the Euclidean distance of Equation 8.

C. The Discriminant Analysis Method (Klein, 1986; Klein and Delisi, 1986)

In this method, the multidimensional statistical technique of discriminant analysis was used to assign a protein to one of the protein structural classes. Moreover, when the Klein and Delisi method (1986) was used to perform prediction, in addition to the amino acid composition, the regular variations in the hydrophobic values of residues along the amino acid sequence was also used as the attribute. This is quite different in selecting parameters from the least Hamming distance method (P.Y. Chou, 1980, 1989) and the least Euclidean distance method (Nakashima et al., 1986), where the amino acid composition of a protein is the only input in performing the prediction of its structural class. Predictions were performed for the 103 proteins in the training set constructed by Klein (1982). Of the 103 proteins, 29 are α-, 27 β-, 16 $\alpha + \beta$-, 26 α/β-, and 5 ζ-proteins. According to his report, the percentages of correct prediction were 20/29 (68.9%) for the α-proteins, 25/27 (92.6%) for the β-proteins, 10/16 (62.5%) for the $\alpha + \beta$-proteins, 20/26 (76.9%) for the α/β-proteins, and 5/5 (100%) for the ζ-proteins, with an average accuracy q of 80/103 (77.7%). However, in the prediction by Klein and Delisi (1986), the $\alpha + \beta$- and α/β-proteins were treated as one structural class, the mixed class. The classification was made for 138 sequences, of which 66 was treated as the training set and 72 as the testing set. According to their report, the average rate of correct prediction was 110/138 (79.7%).

D. The Maximum Component Coefficient Method (Zhang and Chou, 1992a)

The maximum component coefficient method is also called the optimization method. Its basic idea is that a protein denoted by a point (or vector) $\mathbf{X}$ in the 20-D amino acid composition space can be decomposed into five components, each of which corresponds to one of the five norms (or standard vectors) $\mathbf{X}_\alpha$, $\mathbf{X}_\beta$, $\mathbf{X}_{\alpha+\beta}$, $\mathbf{X}_{\alpha/\beta}$, and $\mathbf{X}_\zeta$; i.e.,

$$\mathbf{X} = C_\alpha \mathbf{X}_\alpha + C_\beta \mathbf{X}_\beta + C_{\alpha+\beta} \mathbf{X}_{\alpha+\beta} + C_{\alpha/\beta} \mathbf{X}_{\alpha/\beta} + C_\zeta \mathbf{X}_\zeta \tag{10}$$

where C_α, C_β, $C_{\alpha+\beta}$, $C_{\alpha/\beta}$, and C_ζ are the five component coefficients, which should be subject to the following constraints:

$$\begin{cases} C_\alpha + C_\beta + C_{\alpha+\beta} + C_{\alpha/\beta} + C_\zeta = 1 \\ 0 \le C_\alpha,\ C_\beta,\ C_{\alpha+\beta},\ C_{\alpha/\beta},\ C_\zeta \le 1 \end{cases} \tag{11}$$

Equation 10 can also be written as

$$x_i = C_\alpha x_{\alpha,i} + C_\beta x_{\beta,i} + C_{\alpha+\beta} x_{\alpha+\beta,i} + C_{\alpha/\beta} x_{\alpha/\beta,i} + C_\zeta x_{\zeta,i} \quad (i = 1, 2, \ldots, 20) \tag{12}$$

where x_i, $x_{\alpha,i}$, $x_{\beta,i}$, ... are the ith components of $\mathbf{X}$, $\mathbf{X}_\alpha$, $\mathbf{X}_\beta$, ..., respectively. Equation 12 contains a set of contradictory equations in which the number of equations is greater than

that of the unknown variables. To solve these kinds of contradictory equations, the following least-squares method was adopted.

Define an objective function Q by

$$Q(C_\alpha, C_\beta, C_{\alpha+\beta}, C_{\alpha/\beta}, C_\zeta) = \Sigma_{i=1}^{20}[x_i - (C_\alpha x_{\alpha,i} + C_\beta x_{\beta,i} + C_{\alpha+\beta} x_{\alpha+\beta,i} + C_{\alpha/\beta} x_{\alpha/\beta,i} + C_\zeta x_{\zeta,i})]^2 \quad (13)$$

such that

$$Q(C_\alpha, C_\beta, C_{\alpha+\beta}, C_{\alpha/\beta}, C_\zeta) = \text{minimum} \quad (14)$$

with the constrain imposed by Equation 11. It can be proved that the above optimization problem possesses a unique solution (Rao, 1984). The solution can be easily obtained by means of any existing numerical minimizer, such as the one provided in the IMSL LIBRARY (Fortran subroutines for mathematics and statistics). Once the solutions for these component coefficients are found, the predicted structural class for protein **X** is given by the subscript μ of the following equation:

$$C_\mu = \mathbf{Max}\,(C_\alpha, C_\beta, C_{\alpha+\beta}, C_{\alpha/\beta}, C_\zeta) \quad (15)$$

where μ can be α, β, $\alpha + \beta$, α/β, or ζ, and the operator **Max** means taking the largest one among those in the parentheses. The physical implication of Equation 15 can be illustrated as follows. When, for example, $\mu = \alpha$, i.e., $C_\alpha = \mathbf{Max}\,(C_\alpha, C_\beta, C_{\alpha+\beta}, C_{\alpha/\beta}, C_\zeta)$, it is obvious according to Equation 10 that the weight of the standard vector $\mathbf{X}_\alpha$ in **X** is the largest, and so is the "ingredient" of the α-protein in the protein concerned. Therefore, the predicted protein **X** should be assigned to the α-structural class. Predictions were performed based on the training database consisting of the 64 proteins constructed by P.Y. Chou (1986). According to the report by Zhang and Chou (1992a), the percentages of correct prediction were 19/19 (100%) for the α-proteins, 12/15 (80.0%) for the β-proteins, 10/14 (71.4%) for the $\alpha + \beta$-proteins, 12/16 (75.0%) for the α/β-proteins, with an average accuracy of 53/64 (82.8%). However, no predicted results were reported by them for any testing proteins outside the training set.

E. The Least Correlation Angle Method (Chou and Zhang, 1993)

The least correlation angle method is also called the maximum projection method. The rationale of this method is self-evident, as can be seen by the fact that the correlation angle of two identical vectors must be zero, which is also the case when their projection reaches the maximum. Generally speaking, the larger the projection between two vectors, the smaller their correlation angle, and the higher the similarity between the two vectors. The projections of protein **X** with the norms (or standard vectors) $\mathbf{X}_\alpha$, $\mathbf{X}_\beta$, $\mathbf{X}_{\alpha+\beta}$, $\mathbf{X}_{\alpha/\beta}$, and $\mathbf{X}_\zeta$ are defined

by

$$\begin{cases} \Pi_{\alpha}^{\mathbf{X}} = \mathbf{X}\cdot\mathbf{X}_{\alpha} = |\mathbf{X}||\mathbf{X}_{\alpha}|\cos(\Theta_{\alpha}^{\mathbf{X}}) \\ \Pi_{\beta}^{\mathbf{X}} = \mathbf{X}\cdot\mathbf{X}_{\beta} = |\mathbf{X}||\mathbf{X}_{\beta}|\cos(\Theta_{\beta}^{\mathbf{X}}) \\ \Pi_{\alpha+\beta}^{\mathbf{X}} = \mathbf{X}\cdot\mathbf{X}_{\alpha+\beta} = |\mathbf{X}||\mathbf{X}_{\alpha+\beta}|\cos(\Theta_{\alpha+\beta}^{\mathbf{X}}) \\ \Pi_{\alpha/\beta}^{\mathbf{X}} = \mathbf{X}\cdot\mathbf{X}_{\alpha/\beta} = |\mathbf{X}||\mathbf{X}_{\alpha/\beta}|\cos(\Theta_{\alpha/\beta}^{\mathbf{X}}) \\ \Pi_{\zeta}^{\mathbf{X}} = \mathbf{X}\cdot\mathbf{X}_{\zeta} = |\mathbf{X}||\mathbf{X}_{\zeta}|\cos(\Theta_{\zeta}^{\mathbf{X}}) \end{cases} \tag{16}$$

where $\Theta_{\alpha}^{\mathbf{X}}$ (Figure 2) is the correlation angle of the vector $\mathbf{X}$ with the standard vector $\mathbf{X}_{\alpha}$, $\Theta_{\beta}^{\mathbf{X}}$ the correlation angle of the vector $\mathbf{X}$ with the standard vector $\mathbf{X}_{\beta}$, and so forth.

According to the Cauchy-Schwartz-Buniakowsky inequality (Wylie and Barrell, 1982), for any two arbitrary sets of numbers $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$, we have

$$\left(\Sigma_{k=1}^{n} a_k b_k\right)^2 \leq \left(\Sigma_{k=1}^{n} a_k^2\right)\left(\Sigma_{k=1}^{n} b_k^2\right) \tag{17}$$

The equality holds if, and only if, the sequences $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ are proportional. Therefore, the correlation angles in Equation 16 can be expressed by:

$$\begin{cases} \Theta_{\alpha}^{\mathbf{X}} = \arccos\left\{\dfrac{\Sigma_{i=1}^{20} x_i x_{\alpha,i}}{\{[\Sigma_{i=1}^{20} x_i^2][\Sigma_{i=1}^{20} x_{\alpha,i}^2]\}^{1/2}}\right\} \\ \Theta_{\beta}^{\mathbf{X}} = \arccos\left\{\dfrac{\Sigma_{i=1}^{20} x_i x_{\beta,i}}{\{[\Sigma_{i=1}^{20} x_i^2][\Sigma_{i=1}^{20} x_{\beta,i}^2]\}^{1/2}}\right\} \\ \Theta_{\alpha+\beta}^{\mathbf{X}} = \arccos\left\{\dfrac{\Sigma_{i=1}^{20} x_i x_{\alpha+\beta,i}}{\{[\Sigma_{i=1}^{20} x_i^2][\Sigma_{i=1}^{20} x_{\alpha+\beta,i}^2]\}^{1/2}}\right\} \\ \Theta_{\alpha/\beta}^{\mathbf{X}} = \arccos\left\{\dfrac{\Sigma_{i=1}^{20} x_i x_{\alpha/\beta,i}}{\{[\Sigma_{i=1}^{20} x_i^2][\Sigma_{i=1}^{20} x_{\alpha/\beta,i}^2]\}^{1/2}}\right\} \\ \Theta_{\zeta}^{\mathbf{X}} = \arccos\left\{\dfrac{\Sigma_{i=1}^{20} x_i x_{\zeta,i}}{\{[\Sigma_{i=1}^{20} x_i^2][\Sigma_{i=1}^{20} x_{\zeta,i}^2]\}^{1/2}}\right\} \end{cases} \tag{18}$$

The protein $\mathbf{X}$ is predicted to belong to the structural class for which the projection is the largest, or according to trigonometry, the correlation angle is the smallest. In other words, the predicted class for protein $\mathbf{X}$ is given by the subscript μ of the following equation:

$$\Theta_{\mu}^{\mathbf{X}} = \mathbf{Min}\,(\Theta_{\alpha}^{\mathbf{X}}, \Theta_{\beta}^{\mathbf{X}}, \Theta_{\alpha+\beta}^{\mathbf{X}}, \Theta_{\alpha/\beta}^{\mathbf{X}}, \Theta_{\zeta}^{\mathbf{X}}) \tag{19}$$

where μ can be α, β, $\alpha+\beta$, or α/β, and the operator **Min** means taking the smallest one among those in the parentheses.

The same training set of 64 proteins classified by P.Y. Chou (1986) was used for demonstration, and the results indicated that the rates of correct prediction for the α-, β-, $\alpha+\beta$-,

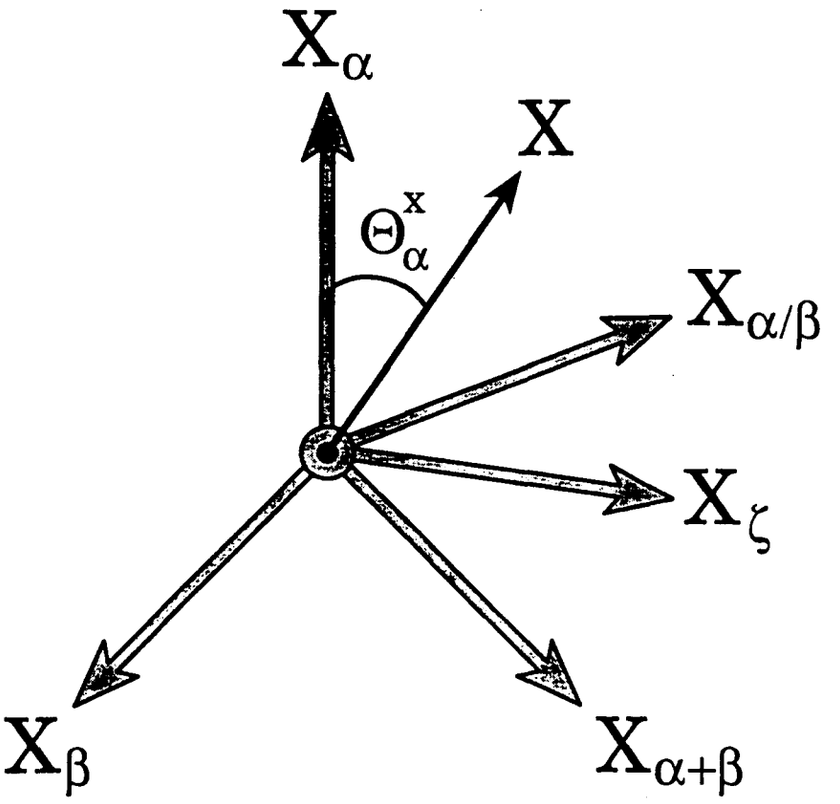

Figure 2: Schematic illustration of the correlation angle between two vectors in a 20-D space. The black single-line vector $\mathbf{X}$ represents the protein to be predicted, while the five shaded double-line vectors $\mathbf{X}_\alpha$, $\mathbf{X}_\beta$, $\mathbf{X}_{\alpha+\beta}$, $\mathbf{X}_{\alpha/\beta}$, and $\mathbf{X}_\zeta$ represents the norms of the five protein structural classes (α, β, $\alpha+\beta$, α/β, and ζ) θ_α^X is the correlation angle between $\mathbf{X}_\alpha$ and $\mathbf{X}$. The correlation angles between $\mathbf{X}$ and the other four standard vectors can likewise be illustrated, although they are not explicitly marked here.

and α/β-proteins were 18.5/19 (97.4%), 12/15 (80.0%), 10/14 (71.4%), and 13/16 (81.3%), respectively, with a rate of 53.5/64 (83.6%) for the average accuracy. Predictions were also made with this method for a testing set of 35 proteins of known X-ray structure not included in the training set; the percentage of correct prediction was 32/35 (91.4%).

F. Other Methods

Recently, a number of other methods have emerged that basically can be categorized as three different approaches.

One methods uses the computer-simulated neural networks, such as the methods proposed by Dubchak et al.(1993), Metfessel et al.(1993), and Rost and Sander (1994). However, as

pointed out recently by Eisenhaber et al. (1995), the prediction accuracy for the structural class thus achieved is not better (and sometimes even lower) than the analytical methods. Moreover, the computational costs for training the network is considerable, and it is scientifically disappointing that no physical explanation for the possible prediction success is afforded by the neural network technique.

The second method uses a different geometric approach, such as the weighted Euclidean distance by Zhou et al., (1992), the average distance maps constructed from protein sequences by Kikuchi (1993), and the elliptically scaled distance by Mao et al. (1994). By means of the weighted Euclidean distance method or the elliptically scaled distance method, the prediction accuracy for the training set data was remarkably improved. However, the prediction accuracy of these methods for the testing set proteins are much lower than a simpler method, the aforementioned least correlation angle method (Chou and Zhang, 1993).

The third approach uses fuzzy clustering to predict the structural class of a protein from its amino acid composition (Zhang et al., 1995). Based on the concept of fuzzy mathematics, each of the structural classes is described by a fuzzy cluster, and each protein is characterized by its membership degree, a number between zero and one, in each of the clusters, with the constraint that the sum of the membership degrees equals unity. A given protein is then classified as belonging to the structural class corresponding to the fuzzy cluster with the maximum membership degree. Calculation of membership degrees is carried out using the fuzzy c-means algorithm developed by Bezdek (1981) on a training set of 64 proteins, the same set used by P.Y. Chou (1989). Results obtained for both the training and testing sets show that the fuzzy clustering approach produces results quite comparable to those obtained by the previous methods. Because the application of fuzzy clustering is based on a new branch of mathematics, it might be worthy of further investigation given its novel concept and elegant formulation, although the results obtained thus far by this method do not seem to be a remarkable improvement in raising the success rate of prediction.

IV. Prediction by the Components-Coupled Method

The components-coupled method is also called the least Mahalanobis distance method (Chou and Zhang, 1994; Chou, 1995a). Much of this review is focused on this method not only because the components-coupled method is a novel approach and bears a profound physical implication, but also because it is much better than all the others according to the results derived from a given training set or testing set, or according to those examined by various statistic analyses (Chou and Zhang, 1994; Chou, 1995a).

A. Coupling Effect and Mahalanobis Distances

As can be seen, none of the methods mentioned in the last section has explicitly taken into account the coupling effect among different amino acid components. Actually, in those methods, the components of the 20 amino acids were each treated as an independent variable, and the contribution owing to the coupling among the different components was completely

neglected. As an illustration, let us analyze the least Euclidean distance method. The merit of Euclidean distance, which was used by Nakashima et al. (1986) as a scale for measuring similarity (see Equation 8), is simple and intuitive. However, it also has the following weaknesses. First, the 20-D composition space is generally not an orthogonal space. Is it validated to extend the definition of Euclidean distance in a 3-D orthogonal space to a 20-D nonorthogonal space? Second, and more importantly, the coupling among different components of the distance is completely neglected. This might be appropriate when the problem considered is a pure geometrical one. However, when the distance is used as a statistical scale to classify different sets of data according to the similarity principle, this kind of effect becomes important (see Appendix B for an illustration). Actually, the scale used in the least Euclidean distance method, like those used by the other methods mentioned in the last section, is merely a zero-order approximation, and it will certainly set a barrier for increasing the accuracy. Thus, the following questions are naturally raised. Through what avenue can the coupling be effectively taken into account? How should the method be formulated to make the quantitative calculation feasible? To deal with these problems, the Mahalanobis distance (Mahalanobis, 1936; Pillai, 1985) has been introduced. As shown below, the introduction of Mahalanobis distance enables one to take into account the interactions among different amino acid components. Thus, the principle underlying the prediction method becomes: the shorter the Mahalanobis distance between two protein molecules, the higher their similarity, and hence the more likely they belong to the same structural class. Given below are a brief introduction to the Mahalanobis distance, its difference from ordinary distances, and a formulation of how to use it to predict the structural class of a protein.

According to the definition, the Mahalanobis distance, $D(\mathbf{X}, \overline{\mathbf{X}})$, between the norm defined by Equation 2 and any point $\mathbf{X}$ in the 20-D space is given by (Mahalanobis, 1936; Pillai, 1985)

$$D^2(\mathbf{X}, \overline{\mathbf{X}}) = (\mathbf{X} - \overline{\mathbf{X}})^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{X} - \overline{\mathbf{X}}) \tag{20}$$

where $\mathbf{S}$ is a covariance matrix given by

$$\mathbf{S} = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,20} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ s_{20,1} & s_{20,2} & \cdots & s_{20,20} \end{bmatrix} \tag{21}$$

the superscript T is the transposition operator, and $\mathbf{S}^{-1}$ is the inverse matrix of $\mathbf{S}$. The matrix elements $s_{i,j}$ in Equation 21 are given by

$$s_{i,j} = \sum_{k=1}^{N} \left[x_{k,i} - \overline{x_i}\right] \left[x_{k,j} - \overline{x_j}\right], \quad (i, j = 1, 2, \ldots, 20) \tag{22}$$

Note that the distance defined by Equation 20 bears all the basic properties of a geometric distance. Suppose $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are any three points in the space where the Mahalanobis distance is defined. It follows that (1) $D(\mathbf{A}, \mathbf{B}) \geq 0$, where an equal sign holds true only when $\mathbf{A} \equiv \mathbf{B}$,

(2) $D(\mathbf{A}, \mathbf{B}) \equiv D(\mathbf{B}, \mathbf{A})$, and (3) $D(\mathbf{A}, \mathbf{B}) + D(\mathbf{B}, \mathbf{C}) \geq D(\mathbf{A}, \mathbf{C})$. However, being different from the ordinary geometric distance, the Mahalanobis distance is unit independent, that is, its value will not be changed by using different units of coordinates. When the nondiagonal elements in the covariance matrix **S** are all zero, the distance defined by Equation 20 will reduce to a weighted Euclidean distance, as used by Zhou et al. (1992) in studying the protein structural class. Actually, **S** is a symmetric matrix whose nondiagonal elements are generally not zero. It is these nondiagonal terms of the covariance matrix **S** that reflect the coupling effect among different amino acid components and enable the Mahalanobis distance to be a more appropriate scale in classifying statistical data, as illustrated by a simple example in Appendix B. Actually, a similar treatment has also been used by other investigators (Dayhoff and Eck, 1968; Henikoff and Henikoff, 1992; Miyazawa and Jernigan, 1993; Henikoff and Henikoff, 1994) in developing methods for the alignment of protein sequences by using a substitution matrix with scores for all possible exchanges of one amino acid with another. Although the matrix elements introduced by them are different from those of the covariance matrix defined here, they both reflect the importance of the coupling effect among different amino acid components in studying the similarity of proteins.

B. Prediction in a 20-D space

1. Formulation

Because the amino acid composition must be normalized, that is, constrained by

$$\sum_{i=1}^{20} x_{k,i} = 1 \quad (k = 1, 2, \ldots, N) \tag{23}$$

it follows (see Equation 22) that

$$\begin{cases} \sum_{j=1}^{20} s_{i,j} = 0, & (i = 1, 2, \ldots, 20) \\ \sum_{i=1}^{20} s_{i,j} = 0, & (j = 1, 2, \ldots, 20) \end{cases} \tag{24}$$

Therefore, **S** defined by Equation 21 is a singular matrix, and its inverse matrix $\mathbf{S}^{-1}$ must be divergent and meaningless. To overcome the divergent difficulty of $\mathbf{S}^{-1}$, here let us provide an eigenvalue-eigenvector approach that can be used to calculate the Mahalanobis distance without the need to go through any inverse matrix. The procedure can be formulated as follows. First, let us find the eigenvalues and eigenvectors of **S** by solving the equation

$$\mathbf{S}\Psi_i = \lambda_i \Psi_i = \lambda_i \begin{bmatrix} \psi_{i,1} \\ \psi_{i,2} \\ \vdots \\ \psi_{i,20} \end{bmatrix} \quad (i = 1, 2, \ldots, 20) \tag{25}$$

where λ_i is the ith eigenvalue of $\mathbf{S}$ and Ψ_i the corresponding eigenvector. Because Equation 24, one of the 20 eigenvalues in Equation 25 must be zero, and its corresponding eigenvector must have equal components. Suppose such a particular eigenvalue and eigenvector are represented by λ_1 and Ψ_1; we thus have

$$\begin{cases} \lambda_1 = 0 \\ \\ \psi_{1,1} = \psi_{1,1} = \ldots = \psi_{1,20} = \sqrt{\frac{1}{20}} \end{cases} \tag{26}$$

The eigenvector Ψ_1 and the other 19 eigenvectors (whose eigenvalues are not zero) are orthonormal, and can be used to construct a unitary matrix

$$\mathbf{U} = \begin{bmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,20} \\ u_{2,1} & u_{2,2} & \cdots & u_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ u_{20,1} & u_{20,2} & \cdots & u_{20,20} \end{bmatrix} = \begin{bmatrix} \psi_{1,1} & \psi_{2,1} & \cdots & \psi_{20,1} \\ \psi_{1,2} & \psi_{2,2} & \cdots & \psi_{20,2} \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{1,20} & \psi_{2,20} & \cdots & \psi_{20,20} \end{bmatrix} \tag{27}$$

which has the property of $\mathbf{U}^{-1} = \mathbf{U}^{\mathrm{T}}$, and hence we have

$$\mathbf{U}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{U} = \left(\mathbf{U}^{\mathrm{T}}\mathbf{S}\mathbf{U}\right)^{-1} = \Lambda^{-1} = \begin{bmatrix} \lambda_1 & & & \mathbf{0} \\ & \lambda_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \lambda_{20} \end{bmatrix}^{-1} \tag{28}$$

Accordingly, Equation 20 can be written as

$$\begin{aligned} D^2(\mathbf{X}, \overline{\mathbf{X}}) &= (\mathbf{X} - \overline{\mathbf{X}})^{\mathrm{T}}\mathbf{U}\mathbf{U}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{U}\mathbf{U}^{\mathrm{T}}(\mathbf{X} - \overline{\mathbf{X}}) \\ &= \left[(\mathbf{X} - \overline{\mathbf{X}})^{\mathrm{T}}\mathbf{U}\right]\left[\mathbf{U}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{U}\right]\left[\mathbf{U}^{\mathrm{T}}(\mathbf{X} - \overline{\mathbf{X}})\right] \\ &= \left[\mathbf{U}^{\mathrm{T}}(\mathbf{X} - \overline{\mathbf{X}})\right]^{\mathrm{T}}\left[\mathbf{U}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{U}\right]\left[\mathbf{U}^{\mathrm{T}}(\mathbf{X} - \overline{\mathbf{X}})\right] \\ &= \mathbf{Y}^{\mathrm{T}}\Lambda^{-1}\mathbf{Y} = \frac{y_1^2}{\lambda_1} + \sum_{i=2}^{20}\frac{y_i^2}{\lambda_i} \end{aligned} \tag{29}$$

where

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{20} \end{bmatrix} = \mathbf{U}^{\mathrm{T}}(\mathbf{X} - \overline{\mathbf{X}}) = \begin{bmatrix} \psi_{1,1} & \psi_{1,2} & \cdots & \psi_{1,20} \\ \psi_{2,1} & \psi_{2,2} & \cdots & \psi_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{20,1} & \psi_{20,2} & \cdots & \psi_{20,20} \end{bmatrix} \begin{bmatrix} (x_1 - \overline{x_1}) \\ (x_2 - \overline{x_2} \\ \vdots \\ (x_{20} - \overline{x_{20}}) \end{bmatrix} \tag{30}$$

or

$$y_i = \Sigma_j^{20}\psi_{i,j}(x_j - \overline{x_j}), \quad (i = 1, 2, \ldots, 20) \tag{31}$$

Note that it can be proved that $\lambda_1 \Rightarrow 0$. Actually, according to Equation 30 and the condition of normalization, it follows (see Equations 26 and 27) that

$$y_1 = \sum_{j=1}^{20} \psi_{1,j}(x_i - \overline{x_j}) = \sqrt{\frac{1}{20}}(1 - 1) \equiv 0 \tag{32}$$

Therefore, Equation 31 can be reduced to

$$D^2(\mathbf{X}, \overline{\mathbf{X}}) \quad = \sum_{i=2}^{20} \frac{y_i^2}{\lambda_i} \tag{33}$$

It can be seen from Equation 32 that one can calculate the Mahalanobis distance without the need to deal with any inverse matrix at all if using the new approach as proposed here.

Suppose that in a training set there are five subsets of proteins. Subset 1 contains only α-proteins whose number is N_α; subsets 2, 3, 4, and 5 contain only β-, $\alpha + \beta$-, α/β-, and ζ-proteins, respectively, whose numbers are N_β, $N_{\alpha+\beta}$, $N_{\alpha/\beta}$, and N_ζ. Based on the data of N_α-proteins in subset 1, the norm of α-proteins, $\mathbf{X}_\alpha$, can be calculated (see Equations 2 and 3). Thus, the covariance matrix, $\mathbf{S}_\alpha$, for the α-proteins in subset 1 is defined, and its 19 nonzero eigenvalues ($i = 1, 2, \ldots, 19$) and the corresponding eigenvectors Ψ_i^α can be calculated by Equation 25. Based on these data, as well as the amino acid composition of protein $\mathbf{X}$, the Mahalanobis distance $D(\mathbf{X}, \mathbf{X}_\alpha)$ between protein $\mathbf{X}$ and the norm of α-proteins can be calculated (see Equations 30 and 32). Similarly, the covariance matrices for the β-, $\alpha + \beta$-, α/β-, and ζ-subsets, denoted by $\mathbf{S}_\alpha$, $\mathbf{S}_\beta$, $\mathbf{S}_{\alpha+\beta}$, $\mathbf{S}_{\alpha/\beta}$, and $\mathbf{S}_\zeta$, respectively, can be defined. Their eigenvalues λ_i^β, $\lambda_i^{\alpha+\beta}$, $\lambda_i^{\alpha/\beta}$, λ_i^ζ and the corresponding eigenvectors Ψ_i^β, $\Psi_i^{\alpha+\beta}$, $\Psi_i^{\alpha/\beta}$, Ψ_i^ζ can be calculated as well. Also, the Mahalanobis distances from protein $\mathbf{X}$ to the norms of the β-, $\alpha+\beta$-, α/β-, and ζ-proteins, denoted by $D^2(\mathbf{X}, \mathbf{X}_\beta)$, $D^2(\mathbf{X}, \mathbf{X}_{\alpha+\beta})$, $D^2(\mathbf{X}, \mathbf{X}_{\alpha/\beta})$, and $D^2(\mathbf{X}, \mathbf{X}_\zeta)$, respectively, can be obtained. Thus, protein $\mathbf{X}$ is predicted to belong to the structural class for which the Mahalanobis distance is the least, which can be formulated as follows. Suppose

$$D^2(\mathbf{X}, \mathbf{X}_\mu) = \mathbf{Min}\left\{D^2(\mathbf{X}, \mathbf{X}_\alpha), D^2(\mathbf{X}, \mathbf{X}_\beta), D^2(\mathbf{X}, \mathbf{X}_{\alpha+\beta}), D^2(\mathbf{X}, \mathbf{X}_{\alpha/\beta}), D^2(\mathbf{X}, \mathbf{X}_\zeta)\right\} \tag{34}$$

where the subscript μ can be α, β, $\alpha + \beta$, α/β, or ζ, and the operator **Min** means taking the least one among those in the parentheses. The subscript μ of Equation 34 will then give the structural class to which the predicted protein $\mathbf{X}$ should belong.

In summary, the prediction can be performed according to the following stepwise procedure:

1. Knowing the amino acid compositions of the database proteins and the unknown protein whose structural class is to be ascertained, normalize their amino acid components by dividing the number of each component amino acid by the total number of amino acids in the protein (see Equations 1 and 23).

2. Calculate the standard points for the α-, β-, $\alpha+\beta$-, α/β-, and ζ-proteins from the database proteins (see Equations 3 and 4).

3. Calculate the 20-row and 20-column covariance matrices $\mathbf{S}_\alpha$, $\mathbf{S}_\beta$, $\mathbf{S}_{\alpha+\beta}$, $\mathbf{S}_{\alpha/\beta}$, and $\mathbf{S}_\zeta$, for the α-, β-, $\alpha + \beta$-, α/β-, and ζ-proteins, respectively, from the database proteins (see Equations 21 and 22 and Appendix A).

4. Calculate the eigenvalues and eigenvectors by solving Equation 25. For each class, there are 20 eigenvalues, of which one must be zero (see Equation 26) and can be left out because it makes no contribution to the Mahalanobis distance (see Equation 32).

5. Calculate the Mahalanobis distance from the point of the unknown protein to each of the above five standard points (see Equations 31 and 33).

6. The unknown protein is predicted to have the same structural class as the one to which the Mahalanobis distance is the least (see Equation 34).

2. Results

Predictions with the above algorithm have been performed for two sets of proteins, the development (or training) set and the testing set. The prediction for the former is a resubstitution examination for checking the self-consistency of the algorithm, while that for the latter is a cross-validation examination for checking its extrapolating effectiveness. A valid new algorithm should give better results for both of these two aspects.

In this article, the $4 \times 30 = 120$ regular proteins in Appendix A are used as a training database; the nine irregular proteins are left out because their number is too small to form a good set of statistical data. Based on the data of the 30 representative α proteins the norm $\mathbf{X}_\alpha$ and the covariance matrix $\mathbf{S}_\alpha$ have been calculated by Equations 3 and 4, and Equations 21 and 22, respectively. According to $\mathbf{S}_\alpha$, the corresponding eigenvalues and eigenvectors, λ_i^α and Ψ_i^α ($i = 1, 2, \ldots, 20$), are calculated (see Equation 25). This can be easily done by calling EIG20D, an eigen analysis subroutine. Similarly, based on the data of the 30 representative β-proteins, 30 representative $\alpha + \beta$-proteins, and 30 representative α/β-proteins, the corresponding norms, covariance matrices, eigenvalues, and eigenvectors have been calculated. All these data thus obtained are given in Appendix C through which the Mahalanobis distance between any protein to be predicted and the norm of the α-, β-, $\alpha + \beta$-, or α/β-protein set can be uniquely calculated (see Equation 33).

The Mahalanobis distance calculated between each of the norms of the four structural classes and each of the 30 α-proteins, 30 β-proteins, 30 $\alpha + \beta$-proteins, and 30 α/β-proteins is listed in Table 1.

To provide an intuitive feeling and facilitate the statistical analysis later, a 3-D histogram illustrates the Mahalanobis distances for the 30 proteins in each of the four structural classes (Figure 3).

As can be seen from Table 1A or Figure 3A, $D^2(\mathbf{P}, \mathbf{X}_\alpha)$ has the least values for all the proteins listed in Table 1A. Therefore, according to Equation 34, all 30 α-proteins are correctly predicted to be of the α-structural class. These results indicate that the rate of correct prediction for the α-proteins is 100%. Tables 1B and C or Figure 3B and C indicate that the rates of correct prediction for the 30 β- and 30 $\alpha + \beta$-proteins are also 100%! Table 1D or

Table 1A

Predicted Results[a] for the 30 α-Proteins in the Training Database by the Component-Coupled Method

PDB code of the 30 α proteins[b]	Mahalanobis distance				Predicted structural class
	$D^2(\mathbf{X}, \mathbf{X}_\alpha)$	$D^2(\mathbf{X}, \mathbf{X}_\beta)$	$D^2(\mathbf{X}, \mathbf{X}_{\alpha+\beta})$	$D^2(\mathbf{X}, \mathbf{X}_{\alpha/\beta})$	
1AVHA	0.51*	1.08	1.55	4.81	α
1BABB	0.46*	5.64	2.29	6.72	α
1BRD–	0.91*	6.78	3.83	31.18	α
1C5A–	0.91*	2.82	6.25	36.11	α
1CPCA	0.59*	1.11	3.12	3.41	α
1CPCL	0.51*	5.63	1.65	3.06	α
1ECO–	0.53*	2.74	4.65	9.10	α
1FCS–	0.21*	4.07	2.49	11.46	α
1FHA–	0.48*	3.70	7.31	11.81	α
1FIAB	0.63*	4.01	26.43	12.06	α
1HBG–	0.74*	5.84	4.11	7.48	α
1HDDC	0.87*	5.69	8.43	28.45	α
1HIGA	0.58*	2.76	2.46	4.54	α
1LE4–	0.81*	4.18	13.44	10.96	α
1LIG–	0.79*	5.58	9.44	4.61	α
1LTSC	0.74*	3.99	4.34	6.87	α
1MBC–	0.25*	5.22	2.61	12.67	α
1MBS–	0.42*	5.58	3.63	12.27	α
1RPRA	0.78*	4.36	9.24	8.21	α
1TROA	0.63*	3.74	10.24	9.01	α
1UTG–	0.89*	1.56	6.02	5.84	α
256BA	0.73*	6.53	2.52	3.65	α
2CCYA	0.73*	8.76	5.35	5.23	α
2LH1–	0.71*	2.01	1.55	4.09	α
2LHB–	0.64*	5.62	2.85	5.76	α
2MHBA	0.73*	4.92	2.13	11.94	α
2MHBB	0.40*	4.45	5.25	11.47	α
2ZTAA	0.82*	6.16	30.04	47.02	α
4MBA–	0.76*	9.84	5.67	8.18	α
4MBN–	0.25*	5.22	2.61	12.67	α

Rate of correct prediction = 30/30 = 100%

[a] The prediction was performed based on Equation 34. The one with the least value of D^2 (marked by *) is assumed to correspond to the structural class for the predicted protein.

[b] The PDB (Protein Data Bank) code is constituted by the first four characters according to the Brookhaven National Laboratory, and the fifth character used here to indicate a specific chain of a protein. If the fifth character is –, the corresponding protein has only one chain. The amino acid compositions of the protein chains listed here are given in Appendix A.

Table 1B

Predicted Results[a] for the 30 β-Proteins in the Training Database by the Component-Coupled Method

PDB code of the 30 β proteins[b]	Mahalanobis distance				Predicted structural class
	$D^2(\mathbf{X}, \mathbf{X}_\alpha)$	$D^2(\mathbf{X}, \mathbf{X}_\beta)$	$D^2(\mathbf{X}, \mathbf{X}_{\alpha+\beta})$	$D^2(\mathbf{X}, \mathbf{X}_{\alpha/\beta})$	
1ACX–	2.08	0.64*	5.63	7.97	β
1AYH–	20.08	0.65*	2.72	3.53	β
1CD8–	2.40	0.66*	3.02	4.46	β
1CDTA	6.71	0.89*	8.54	92.68	β
1CID–	3.27	0.50*	7.09	7.06	β
1DFNA	9.15	0.91*	16.47	153.04	β
1HILA	4.01	0.26*	1.92	4.61	β
1HIVA	6.54	0.79*	6.38	23.29	β
1HLEB	12.42	0.92*	27.16	49.15	β
1MAMH	4.37	0.54*	3.41	4.71	β
1MONA	11.41	0.87*	9.39	13.07	β
1OMF–	6.52	0.57*	1.71	5.69	β
1PHY–	2.79	0.63*	3.02	5.78	β
1REIA	6.00	0.67*	4.91	8.61	β
1TEN–	2.60	0.73*	2.49	11.39	β
1TLK–	3.47	0.65*	3.32	7.94	β
1VAAB	4.65	0.65*	1.77	6.55	β
2ALP–	7.74	0.59*	3.68	16.77	β
2AVIA	25.77	0.64*	6.14	27.43	β
2BPA2	1.09	0.53*	1.16	5.47	β
2HHRC	6.96	0.47*	1.20	5.11	β
2ILA–	0.96	0.71*	1.32	1.55	β
2LALA	4.89	0.43*	5.68	11.10	β
2SNV–	11.76	0.54*	2.03	3.67	β
3CD4A	5.38	0.69*	7.22	10.96	β
4GCR–	6.89	0.76*	3.36	5.41	β
7APIB	11.01	0.85*	5.46	16.94	β
8I1B–	3.12	0.28*	5.16	3.28	β
8FABA	1.86	0.38*	4.61	8.33	β
8FABB	2.63	0.61*	1.49	1.70	β

Rate of correct prediction = 30/30 = 100%

[a] See footnote *a* to Table 1A.

[b] See footnote *b* to Table 1B.

Table 1C

Predicted Results[a] for the 30 $\alpha + \beta$-Proteins in the Training Database by the Component-Coupled Method

PDB code of the 30 $\alpha + \beta$ proteins[b]	Mahalanobis distance				Predicted structural class
	$D^2(\mathbf{X}, \mathbf{X}_\alpha)$	$D^2(\mathbf{X}, \mathbf{X}_\beta)$	$D^2(\mathbf{X}, \mathbf{X}_{\alpha+\beta})$	$D^2(\mathbf{X}, \mathbf{X}_{\alpha/\beta})$	
1AAK–	4.31	8.77	0.79*	13.19	$\alpha + \beta$
1CTF–	2.45	5.38	0.76*	22.03	$\alpha + \beta$
1DNKA	0.99	1.73	0.78*	3.38	$\alpha + \beta$
1EAF–	5.69	1.88	0.50*	2.30	$\alpha + \beta$
1HSBA	10.95	1.91	0.71*	3.09	$\alpha + \beta$
1LTSA	5.01	1.02	0.72*	2.97	$\alpha + \beta$
1LTSD	3.37	2.33	0.79*	10.91	$\alpha + \beta$
1NRCA	2.32	3.20	0.78*	2.03	$\alpha + \beta$
1OVB–	2.14	0.64	0.38*	5.72	$\alpha + \beta$
1POC–	7.85	3.33	0.67*	17.26	$\alpha + \beta$
1PPN–	3.92	0.87	0.57*	8.24	$\alpha + \beta$
1PRF–	2.37	0.79	0.69*	2.83	$\alpha + \beta$
1RND–	5.57	1.62	0.57*	21.79	$\alpha + \beta$
1SNC–	1.87	2.04	0.41*	7.58	$\alpha + \beta$
1TFG–	4.05	1.20	0.63*	33.41	$\alpha + \beta$
1TGSI	6.22	5.61	0.92*	87.89	$\alpha + \beta$
2ACHA	0.89	1.04	0.77*	2.87	$\alpha + \beta$
2ACT–	5.99	0.87	0.65*	5.53	$\alpha + \beta$
2BPA1	1.25	0.63	0.24*	2.31	$\alpha + \beta$
2SNS–	2.67	2.42	0.47*	6.73	$\alpha + \beta$
2SSI–	2.38	1.21	0.49*	4.31	$\alpha + \beta$
3IL8–	3.31	1.35	0.84*	19.27	$\alpha + \beta$
3RUBS	3.67	1.06	0.75*	9.79	$\alpha + \beta$
3SGBI	5.54	11.04	0.76*	68.52	$\alpha + \beta$
3SICI	1.96	1.12	0.41*	4.49	$\alpha + \beta$
4BLMA	0.92	2.05	0.57*	3.58	$\alpha + \beta$
4TMS–	1.77	1.97	0.56*	3.94	$\alpha + \beta$
8CATA	2.09	1.45	0.47*	2.96	$\alpha + \beta$
9RNT–	9.58	2.99	0.87*	7.44	$\alpha + \beta$
9RSAA	5.95	1.42	0.50*	21.73	$\alpha + \beta$

Rate of correct prediction = 30/30 = 100%

[a] See footnote *a* to Table 1A.

[b] See footnote *b* to Table 1A.

Table 1D

Predicted Results[a] for the 30 α/β-Proteins in the Training Database by the Component-Coupled Method

PDB code of the 30 α/β proteins[b]	Mahalanobis distance				Predicted structural class
	$D^2(\mathbf{X}, \mathbf{X}_\alpha)$	$D^2(\mathbf{X}, \mathbf{X}_\beta)$	$D^2(\mathbf{X}, \mathbf{X}_{\alpha+\beta})$	$D^2(\mathbf{X}, \mathbf{X}_{\alpha/\beta})$	
1ABA–	3.15	0.86	2.77	0.69*	α/β
1CIS–	3.50	5.58	5.81	0.87*	α/β
1CSEI	5.58	13.23	6.64	0.92*	α/β
1CTC–	2.37	0.74	2.22	0.41*	α/β
1DHR–	4.49	2.03	2.43	0.72*	α/β
1DRI–	2.47	2.09	5.06	0.79*	α/β
1ETU–	1.43	2.36	2.18	0.83*	α/β
1FX1–	1.93	1.18	1.47	0.83*	α/β
1GPB–	1.14	0.63	1.06	0.40*	α/β
1OFV–	2.60	2.17	3.34	0.78*	α/β
1PAZ–	1.45	1.21	2.91	0.59*	α/β
1PFKA	1.66	1.56	4.05	0.81*	α/β
1PGD	0.83	1.44	1.80	0.53*	α/β
1Q21	0.59	1.04	4.21	0.42*	α/β
1S01–	2.14	2.42	2.51	0.46*	α/β
1SBP–	1.84	2.02	0.85	0.49*	α/β
1SBT–	2.33	2.33	3.19	0.44*	α/β
1TIMA	1.21	1.53	1.06	0.70*	α/β
1TMD–	4.19	0.55	0.44*	0.68	$\alpha+\beta$
1TREA	0.87	2.12	1.83	0.74*	α/β
1ULA–	0.92	0.55	2.49	0.34*	α/β
1WSYB	0.69	1.09	2.79	0.61*	α/β
2HAD–	2.07	2.44	1.36	0.79*	α/β
2LIV–	0.82	1.23	0.95	0.41*	α/β
3GBP–	1.78	4.59	3.31	0.70*	α/β
4FXN–	5.27	5.42	6.95	0.79*	α/β
5CPA–	2.37	0.74	2.22	0.41*	α/β
5P21–	0.65	1.10	4.35	0.50*	α/β
8ABP–	0.99	1.90	0.95	0.70*	α/β
8ATCA	0.70	1.25	1.55	0.66*	α/β

Rate of correct prediction = 29/30 = 96.7%

[a] See footnote *a* to Table 1A.

[b] See footnote *b* to Table 1A.

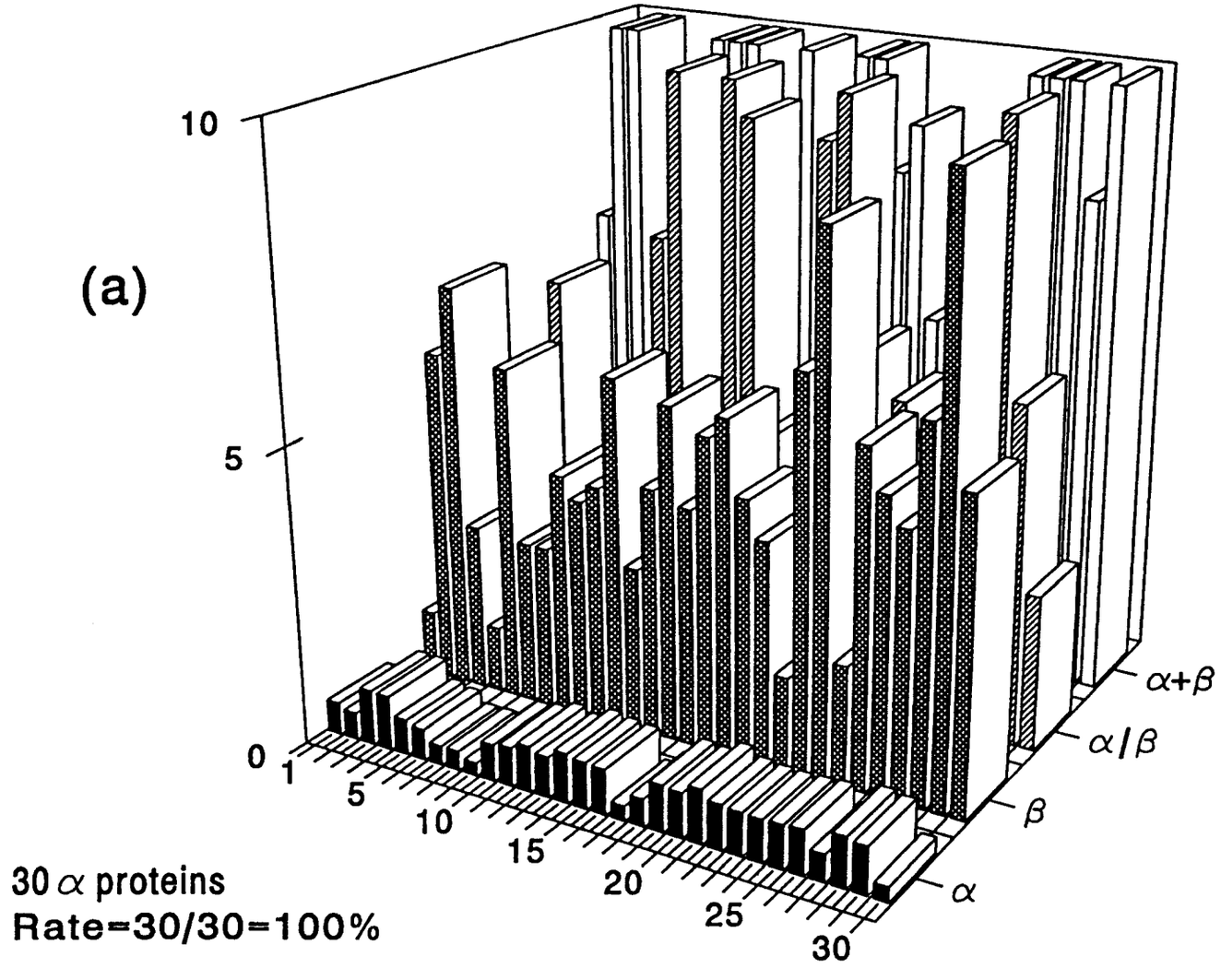

Figure 3A: 3-D histogram showing the Mahalanobis distance from each of the 30 α-proteins in Table 1A to the norms of α-, β-, $\alpha + \beta$-, and α/β-proteins, respectively. The 30 α-proteins are arranged from left to right along the abscissa according to their order in Table 1A. The Mahalanobis distance is shown by the ordinate. Note that any distances with $D^2 > 10$ are cut down to 10.

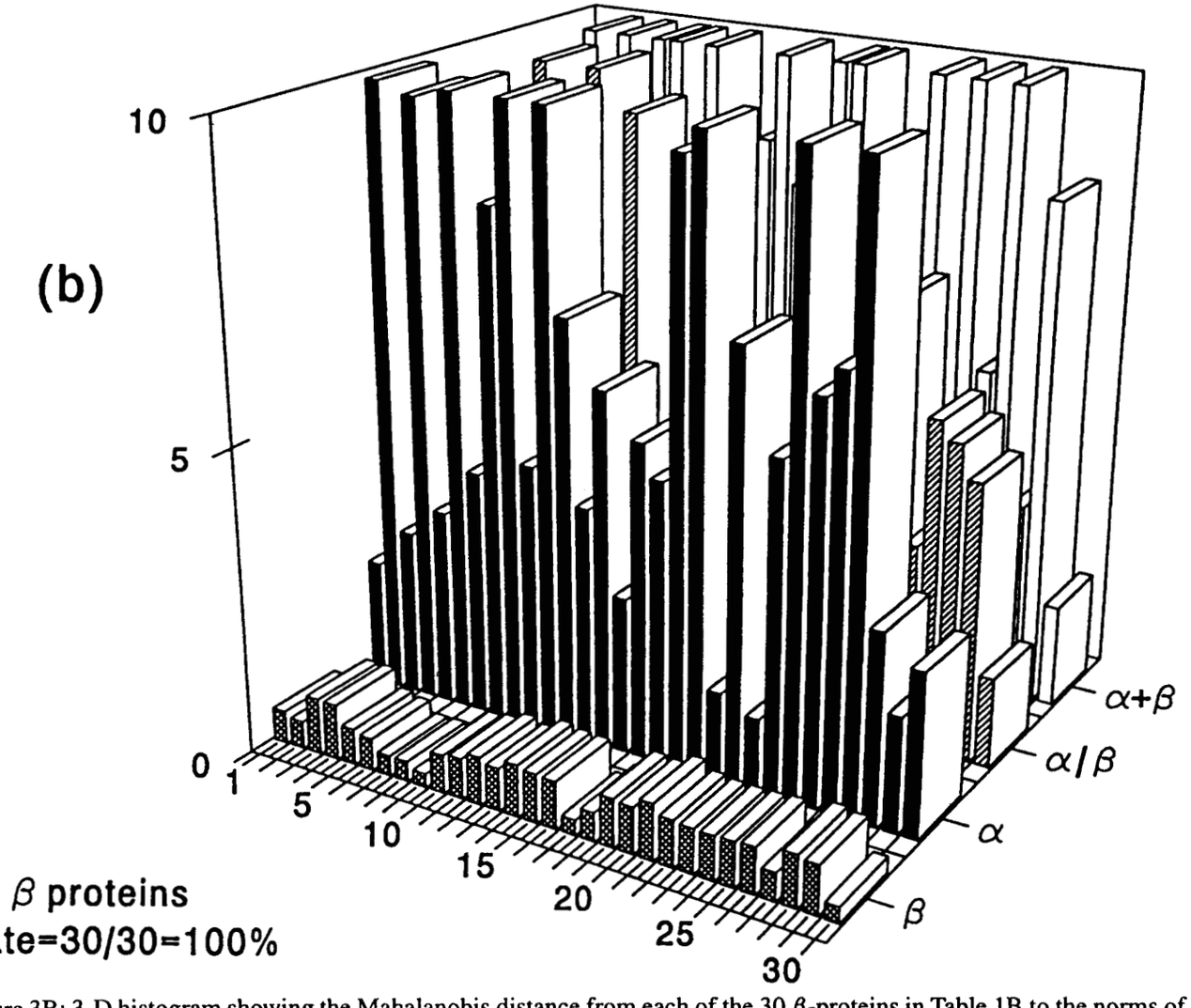

Figure 3B: 3-D histogram showing the Mahalanobis distance from each of the 30 β-proteins in Table 1B to the norms of α-, β-, $\alpha + \beta$-, and α/β-proteins, respectively. The 30 β-proteins are arranged from left to right along the abscissa according to their order in Table 1B. See legend to Figure 3A for further explanation.

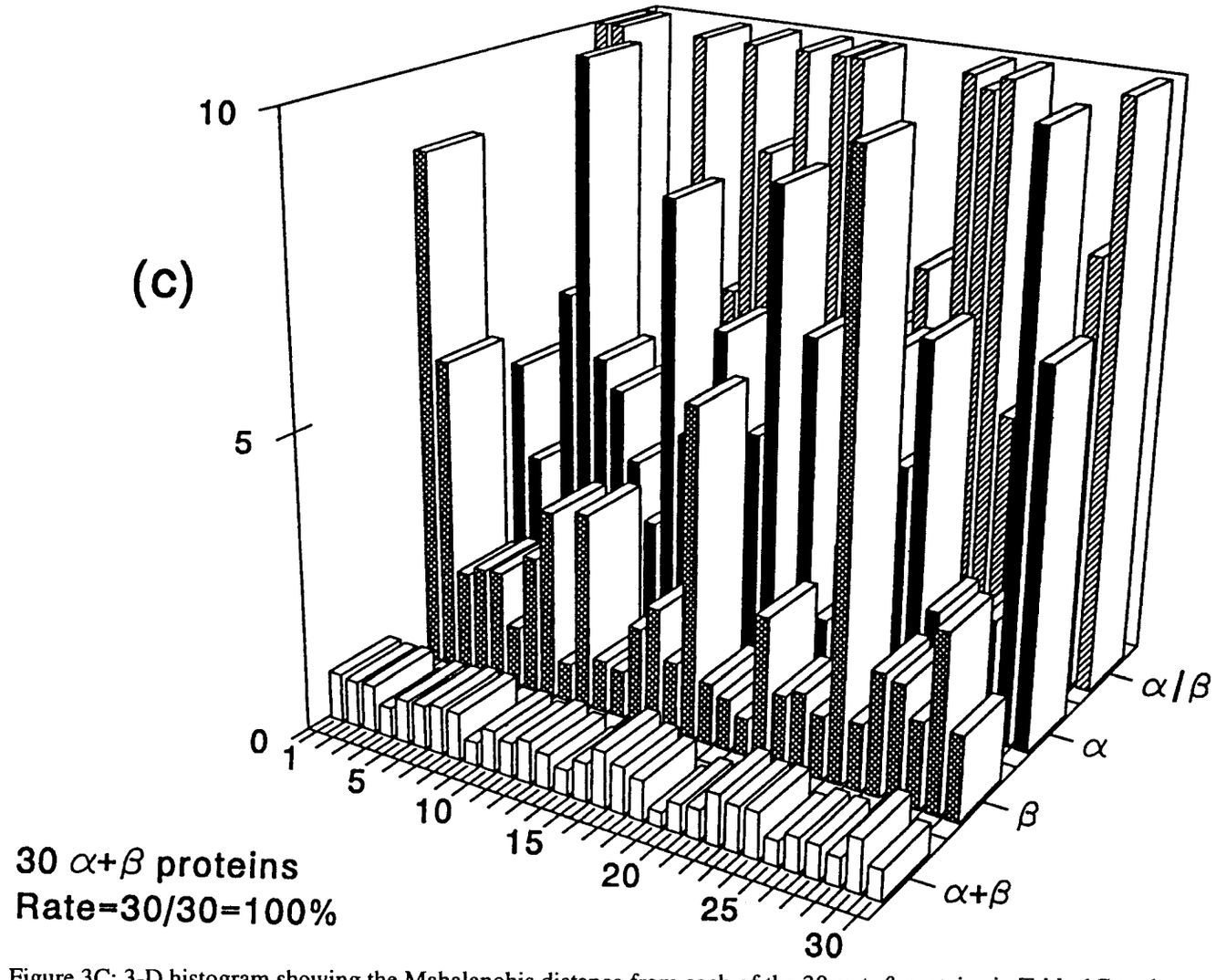

Figure 3C: 3-D histogram showing the Mahalanobis distance from each of the 30 $\alpha + \beta$-proteins in Table 1C to the norms of α-, β-, $\alpha + \beta$-, and α/β-proteins, respectively. The 30 $\alpha + \beta$-proteins are arranged from left to right along the abscissa according to their order in Table 1C. See legend to Figure 3A for further explanation.

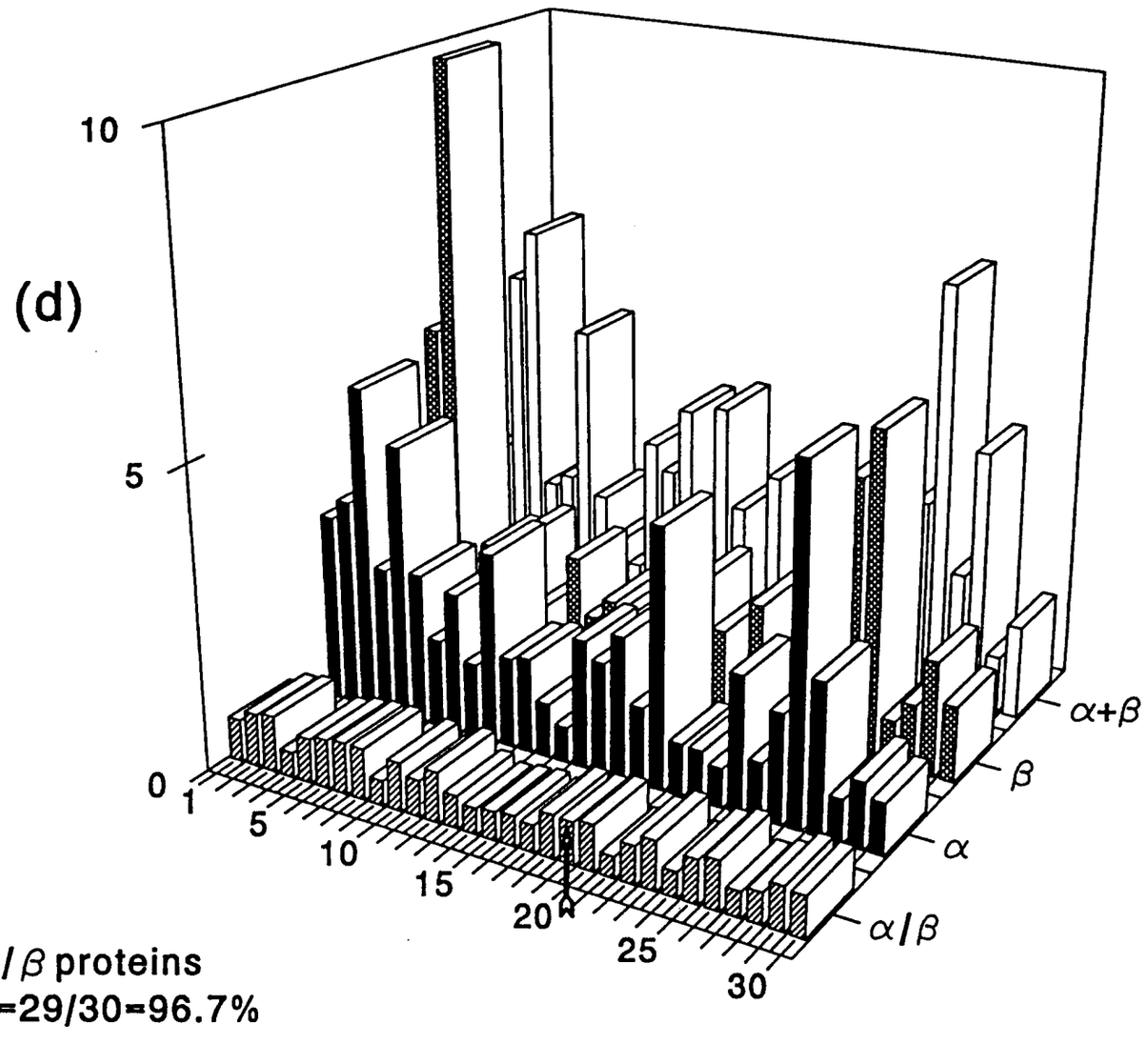

Figure 3D: 3-D histogram showing the Mahalanobis distance from each of the 30 α/β-proteins in Table 1D to the norms of α-, β-, $\alpha + \beta$-, and α/β-proteins, respectively. The 30 α/β-proteins are arranged from left to right along the abscissa according to their order in Table 1D. The arrow indicates the only protein incorrectly predicted. See legend to Figure 3A for further explanation.

Figure 3D indicates that, of the 30 α/β-proteins, 29 have the least values for $D(\mathbf{P}, \overline{\mathbf{P}}_{\alpha/\beta})$, and hence the rate of correct prediction for the α/β-proteins is 29/30=96.7%.

As a cross-validation test, predictions were also performed for a set of 64 independent testing proteins that were not included in the training database of the 4×30 proteins. The 64 testing proteins with their PDB codes as well as amino acid compositions are given in Appendix D. The predicted results for these proteins by the current algorithm are given in Table 2, which indicates that an average accuracy of 61/64 (95.3%) was obtained.

C. Prediction in a (20−1)-D Space

1. Formulation

Here, a different approach is described to overcome the divergence difficulty of the convergence matrix $\mathbf{S}^{-1}$ occurring in the Mahalanobis distance of Equation 20. It is shown from Equation 23 that of the 20 amino acid components of a protein, only 19 are independent. Therefore, by removing one of its 20 components, one can still uniquely represent a protein by a point in a (20−1)-D or 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, the 19-D space will be based on the components of A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, and W. Once the 19-D space is established, the kth protein in a given protein set can be expressed by

$$\mathbf{P}_k = \begin{bmatrix} x_{k,1} \\ x_{k,2} \\ \vdots \\ x_{k,19} \end{bmatrix} \quad (k = 1, 2, \ldots N) \tag{35}$$

where $x_{k,1}, x_{k,2}, \ldots, x_{k,19}$ are the same as in Equation 2. The only difference between $\mathbf{X}_k$ of Equation 2 and $\mathbf{P}_k$ of Equation 35 is that the former contains the component $x_{k,20}$, whereas the latter does not. The norm of the protein set in the 19-D space is defined by

$$\overline{\mathbf{P}} = \begin{bmatrix} \overline{x_1} \\ \overline{x_2} \\ \vdots \\ \overline{x_{19}} \end{bmatrix} \tag{36}$$

where $\overline{x_i}$ $(i = 1, 2, \ldots, 19)$ are the same as in Equation 3 except for $\overline{x_{20}}$, which is not a part of Equation 36.

Accordingly, the covariance matrix $\mathbf{S}$ of Equation 21 would reduce to a 19×19 covariance matrix given by

$$\mathbf{Q} = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,19} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,19} \\ \vdots & \vdots & \ddots & \vdots \\ s_{19,1} & s_{19,2} & \cdots & s_{19,19} \end{bmatrix} \tag{37}$$

Table 2
Predicted Results[a] by the Component-Coupled Algorithm for the 64 Testing Proteins of Known X-Ray Structure Not Included in the Training Database

PDB[b] code of 64 proteins	Mahalanobis distance				Observed type	Predicted type
	$D^2(\mathbf{X}, \mathbf{X}_\alpha)$	$D^2(\mathbf{X}, \mathbf{X}_\beta)$	$D^2(\mathbf{X}, \mathbf{X}_{\alpha+\beta})$	$D^2(\mathbf{X}, \mathbf{X}_{\alpha/\beta})$		
1BBL–	2.20*	3.92	7.47	11.01	α	α
1HBBA	0.85*	4.42	1.88	10.08	α	α
1IFA–	1.81*	2.54	4.69	3.05	α	α
1MRRA	0.53*	0.71	0.63	0.74	α	α
1PDE–	3.27*	3.46	8.56	5.82	α	α
1PRCM	4.38*	4.98	7.09	6.64	α	α
1SAS–	2.88*	3.56	4.96	4.23	α	α
2TMVP	1.16*	2.09	2.10	17.05	α	α
4CPV–	2.83*	6.87	5.94	11.33	α	α
1AAIB	5.18	3.23*	3.48	21.54	β	β
1ATX–	24.33	4.38*	8.37	87.35	β	β
1COBA	5.60	4.26*	4.80	8.81	β	β
1EGF–	17.08	2.66*	7.15	52.36	β	β
1EST–	6.38	1.19*	5.43	10.94	β	β
1GPS–	16.34	5.82*	13.76	159.42	β	β
1HCC–	4.88	4.60*	5.25	14.19	β	β
1IXA–	15.95	7.70*	12.51	88.52	β	β
1MDAA	5.89	2.08*	2.75	4.70	β	β
1PPFE	3.89	2.13*	8.52	19.57	β	β
1R1A2	3.97	1.48*	2.29	4.83	β	β
1SHFA	7.87	0.65*	2.87	6.32	β	β
1TIE–	2.21	0.65*	1.80	3.76	β	β
1TNFA	4.44	1.24*	1.45	5.76	β	β
2ACHB	6.47	4.56*	9.92	83.51	β	β
2CTX–	9.91	3.68*	8.30	139.45	β	β
2MEV1	1.72	0.91*	4.34	5.89	β	β
2PLV1	2.53	0.43*	4.69	3.17	β	β
2SODO	5.60	4.26*	4.80	8.81	β	β
3RP2A	1.28	0.87*	1.02	2.76	β	β
4SGBI	9.59	5.26*	8.22	131.02	β	β
5NN9–	11.46	1.36*	1.45	18.05	β	β
1ABH–	1.97	1.68	1.06*	1.25	$\alpha + \beta$	$\alpha + \beta$
1BBPA	9.76	2.38	2.16*	14.57	$\alpha + \beta$	$\alpha + \beta$

Table 2 *(Continued.)*

PDB[a] code of 64 proteins	Mahalanobis distance $D^2(\mathbf{X}, \mathbf{X}_\alpha)$	$D^2(\mathbf{X}, \mathbf{X}_\beta)$	$D^2(\mathbf{X}, \mathbf{X}_{\alpha+\beta})$	$D^2(\mathbf{X}, \mathbf{X}_{\alpha/\beta})$	Observed type	Predicted type
1BW4–	8.39	6.07	1.79*	21.61	$\alpha+\beta$	$\alpha+\beta$
1COX–	3.72	1.21	0.64*	1.32	$\alpha+\beta$	$\alpha+\beta$
1DNKA	0.99	1.73	0.78*	3.38	$\alpha+\beta$	$\alpha+\beta$
1GLAG	4.45	1.19	1.04*	1.55	$\alpha+\beta$	$\alpha+\beta$
1MS2A	1.56	2.31	0.84*	6.60	$\alpha+\beta$	$\alpha+\beta$
1OVOA	3.93	4.07	1.48*	56.49	$\alpha+\beta$	$\alpha+\beta$
1POC–	7.85	3.34	0.67*	17.26	$\alpha+\beta$	$\alpha+\beta$
1PPBA	3.91	1.30	1.24*	2.25	$\alpha+\beta$	$\alpha+\beta$
1SHAA	1.12	2.38	1.01*	5.04	$\alpha+\beta$	$\alpha+\beta$
1THO–	3.07	3.14	0.85*	2.73	$\alpha+\beta$	$\alpha+\beta$
1TRX–	3.32	3.09	1.00*	2.87	$\alpha+\beta$	$\alpha+\beta$
2AAA–	3.63	1.51	0.57*	2.71	$\alpha+\beta$	$\alpha+\beta$
2PIA–	1.89	0.74*	0.74	7.33	$\alpha+\beta$	β^c
2SN3–	7.91	9.13	2.46*	83.30	$\alpha+\beta$	$\alpha+\beta$
2TAAA	2.70	0.73	0.60*	2.90	$\alpha+\beta$	$\alpha+\beta$
3B5C–	3.72	5.78	1.83*	6.39	$\alpha+\beta$	$\alpha+\beta$
3SC2A	4.22	0.75	0.60*	2.30	$\alpha+\beta$	$\alpha+\beta$
3SC2B	8.64	1.77	1.27*	3.70	$\alpha+\beta$	$\alpha+\beta$
3TLN–	4.34	0.55	0.54*	2.05	$\alpha+\beta$	$\alpha+\beta$
4ENL–	0.42*	1.43	0.45	1.34	$\alpha+\beta$	α^c
4INSB	6.22	21.61	3.86*	25.04	$\alpha+\beta$	$\alpha+\beta$
4RCRH	2.39	1.21	0.91*	2.78	$\alpha+\beta$	$\alpha+\beta$
1GPB–	1.14	0.63	1.06	0.41*	α/β	α/β
1MINA	2.90	1.61	1.18	0.64*	α/β	α/β
1NIPB	1.48	1.71	7.16	1.24*	α/β	α/β
1SBP–	1.84	2.02	0.85	0.48*	α/β	α/β
1WSYA	6.77	1.42	1.22*	1.94	α/β	$\alpha+\beta^c$
4ICD–	1.15	1.34	1.37	0.89*	α/β	α/β
7AATA	1.03	1.47	0.68	0.32*	α/β	α/β
9RUBB	2.20	0.93	0.80	0.79*	α/β	α/β
1GD1O	2.01	1.18	2.65	0.79*	α/β	α/β

Average rate of correct prediction = 61/64 = 95.3%

[a] See footnote *a* to Table 1A.

[b] See footnote *b* to Table 1A.

[c] Incorrect prediction.

where $s_{i,j}$ is the same as in Equation 21, as defined by Equation 22. Thus, the Mahalanobis distance D^2 between the norm $\overline{\mathbf{P}}$ and any protein $\mathbf{P}(x_1, x_2, \ldots, x_{19})$ in the 19-D space is defined by

$$D^2(\mathbf{P}, \overline{\mathbf{P}}) = (\mathbf{P} - \overline{\mathbf{P}})^{\mathrm{T}} \mathbf{Q}^{-1} (\mathbf{P} - \overline{\mathbf{P}}) \tag{38}$$

where T is the transposition operator, and $\mathbf{Q}^{-1}$ is the inverse matrix of $\mathbf{Q}$ given by Equation 37.

A question may be posed. Will leaving out a different amino acid component, and hence the Mahalanobis distance being defined in a different 19-D space, change the value of the distance? The answer is no. Nevertheless, to prove this is by no means a trivial matter. Those who wish to fully understand the details are referred to Appendix E, where a rigorous mathematical proof is provided, because to the best of the authors' knowledge no existing mathematical theorems cover the problem.

Accordingly, the norms $\mathbf{X}_\alpha$, $\mathbf{X}_\beta$, $\mathbf{X}_{\alpha+\beta}$, $\mathbf{X}_{\alpha/\beta}$, and $\mathbf{X}_\zeta$, and the convergence matrices $\mathbf{S}_\alpha$, $\mathbf{S}_\beta$, $\mathbf{S}_{\alpha+\beta}$, $\mathbf{S}_{\alpha/\beta}$, and $\mathbf{S}_\zeta$ defined in the 20-D space should be replaced, respectively, by $\mathbf{P}_\alpha$, $\mathbf{P}_\beta$, $\mathbf{P}_{\alpha+\beta}$, $\mathbf{P}_{\alpha/\beta}$, and $\mathbf{P}_\zeta$ and $\mathbf{Q}_\alpha$, $\mathbf{Q}_\beta$, $\mathbf{Q}_{\alpha+\beta}$, $\mathbf{Q}_{\alpha/\beta}$, and $\mathbf{Q}_\zeta$ defined in the 19-D space. Also, the Mahalanobis distances $D^2(\mathbf{X}, \mathbf{X}_\alpha)$, $D^2(\mathbf{X}, \mathbf{X}_\beta)$, $D^2(\mathbf{X}, \mathbf{X}_{\alpha+\beta})$, $D^2(\mathbf{X}, \mathbf{X}_{\alpha/\beta})$, and $D^2(\mathbf{X}, \mathbf{X}_\zeta)$ defined in the 20-D space should be replaced, respectively, by $D^2(\mathbf{P}, \mathbf{P}_\alpha)$, $D^2(\mathbf{P}, \mathbf{P}_\beta)$, $D^2(\mathbf{P}, \mathbf{P}_{\alpha+\beta})$, $D^2(\mathbf{P}, \mathbf{P}_{\alpha/\beta})$, and $D^2(\mathbf{P}, \mathbf{P}_\zeta)$ as defined in the 19-D space.

Thus, instead of Equation 34, we now have

$$D^2(\mathbf{P}, \mathbf{P}_\mu) = \mathbf{Min}\left\{ D^2(\mathbf{P}, \mathbf{P}_\alpha), D^2(\mathbf{P}, \mathbf{P}_\beta), D^2(\mathbf{P}, \mathbf{P}_{\alpha+\beta}), D^2(\mathbf{P}, \mathbf{P}_{\alpha/\beta}), D^2(\mathbf{P}, \mathbf{P}_\zeta) \right\} \tag{39}$$

where the subscript μ gives the structural class to which the predicted protein $\mathbf{P}$ should belong.

In summary, the prediction can be performed according to the following stepwise procedure:

1. Same as step 1 for the case of the 20-D space. Because the training data in Appendix A are normalized to 100, this step can be easily done by simply dividing each of the values by 100.

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 (see Equation 35).

3. Calculate the 19-D standard point (see Equation 36), and the 19×19 covariance matrix for each of the structural classes concerned from the training database (see Equation 37). For the 4×30 training database in Appendix A, these results are already contained in Appendix C. How to select data from Appendix C depends on which amino acid component is left out in forming the 19-D space, although the final results will be the same regardless of which one is removed (see Appendix E).

4. Calculate the inverse matrix for each of the above 19 matrices. This can easily be done by calling a subroutine DLINDS in the IMSL Library (Fortran Subroutines for Mathematics and Sciences), and the results are given in Appendix F.

5. Calculate the Mahalanobis distance from the point of the unknown protein to each of the standard points (see Equation 38).

6. The unknown protein is predicted to have the same structural class as the one to which the Mahalanobis distance is the least (see Equation 39).

2. Results

As expected, the results predicted by following the above steps for the 4×30 proteins in the training set and 64 proteins in the testing set are exactly the same as those obtained by the eigenvalue-eigenvector approach, as shown in Tables 1 and 2, respectively. Actually, because the Mahalanobis distance defined in the 19-D space does not depend on which component is removed from the amino acid composition space (see Appendix E), we must have $D^2(\mathbf{P}, \mathbf{P}_\alpha) = D^2(\mathbf{X}, \mathbf{X}_\alpha)$, $D^2(\mathbf{P}, \mathbf{P}_\beta) = D^2(\mathbf{X}, \mathbf{X}_\beta)$, $D^2(\mathbf{P}, \mathbf{P}_{\alpha+\beta}) = D^2(\mathbf{X}, \mathbf{X}_{\alpha+\beta})$, and $D^2(\mathbf{P}, \mathbf{P}_{\alpha/\beta}) = D^2(\mathbf{X}, \mathbf{X}_{\alpha/\beta})$.

D. Comparison with Other Methods

To provide an overview, a summary of the predicted accuracies by the least Mahalanobis distance for the 4×30 proteins in the training set is given in Table 3, together with the predicted accuracies for the same database by the least Euclidean distance algorithm (Nakashima et al., 1986) and the least Hamming distance algorithm (P.Y. Chou, 1980, 1989), respectively. As shown in Table 3, the least Mahalanobis distance algorithm yields an average accuracy of 99.2%, much higher than those obtained by the other methods. Prediction of the $\alpha + \beta$-proteins by means of the ordinary geometric distances has been a big problem. As reported by Nakashima et al. (1986), the rate of correct prediction for the $\alpha + \beta$-class was only 37.0% if calculated with their database. Nakashima et al. (1986) attributed the low accuracy to the fact that the $\alpha + \beta$-structural class had a more serious problem in distribution overlapping with all the other structural classes. Even based on the current database, as shown by Table 3, the rate of correct prediction for the $\alpha + \beta$-class by either the least Hamming distance algorithm (P.Y. Chou, 1989) or the least Euclidean distance algorithm (Nakashima et al., 1986) is still smaller than 47%. However, by means of the Mahalanobis distance algorithm, the rate of correct prediction for the $\alpha + \beta$-proteins can reach 100%. Therefore, by taking into account the coupling effect among different amino acid components, the errors caused by the distribution-overlapping problem can be significantly improved.

To facilitate comparison, predictions have also been made by means of the least Hamming distance algorithm (P.Y. Chou, 1980, 1989) and the least Euclidean distance algorithm (Nakashima et al., 1986) for the 64 proteins in the testing set based on the same training database. The average accuracies thus obtained are dramatically decreased to 34/64 (53.1%) (Table 4) and 36/64 (56.3%) (Table 5), or about 42 and 39% lower than that obtained by the least Mahalanobis distance algorithm (Table 2).

Therefore, the results obtained from the resubstitution examination and the cross-validation examination indicate that the least Mahalanobis distance algorithm is much more accurate

Table 3
Comparison of Various Prediction Algorithms for the 4 × 30 Training Database Proteins

Algorithm	Rate of correct prediction				Average accuracy[d]
	α-class	β-class	$\alpha + \beta$-class	α/β-class	q
Component -coupled[a]	$\frac{30}{30} = 100\%$	$\frac{30}{30} = 100\%$	$\frac{30}{30} = 100\%$	$\frac{29}{30} = 96.7\%$	$\frac{119}{120} = 99.2\%$
P. Y. Chou[b] (1989)	$\frac{21}{30} = 70.0\%$	$\frac{22}{30} = 73.3\%$	$\frac{14}{30} = 46.7\%$	$\frac{26}{30} = 86.7\%$	$\frac{83}{120} = 69.2\%$
Nakashima et al.[c] (1986)	$\frac{23}{30} = 76.7\%$	$\frac{20}{30} = 76.7\%$	$\frac{11}{30} = 36.7\%$	$\frac{26}{30} = 86.7\%$	$\frac{76}{120} = 63.3\%$

[a] Based on Mahalanobis distance (Equation 20).
[b] Based on Hamming distance (Equation 5).
[c] Based on Euclidean distance (Equation 8).
[d] The average accuracy is the percentage of the number of correct prediction events for all classes divided by the number of total prediction events (see Equation 7).

than the previous ones even when the tests are based on the same training and testing data.

V. Statistical Analysis

The development of prediction methods based on statistical theory generally consists of two parts: the exploration of new algorithms and the improvement of the training database. For a given prediction method, what is an appropriate criterion to evaluate it? This is a very common but quite subtle problem. In the protein literature, a prediction method is usually examined by the predicted results for a training data set and testing data set, respectively. According to the statistical terminology, the former is an examination by resubstitution, and the latter is an examination by cross-validation. As shown above, with the resubstitution examination, the structural class of each protein from a training set is predicted using the rules derived from the same set. This gives a somewhat optimistic error estimate because the same proteins 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 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 data set is needed because it can reflect the

Table 4
Predicted Results[a] by the Least Hamming Distance Algorithm for the 64 Testing Proteins of Known X-Ray Structure Not Included in the Training Database

PDB[b] code of 64 proteins	Hamming distance $d^H(\mathbf{X}, \mathbf{X}_\alpha)$	$d^H(\mathbf{X}, \mathbf{X}_\beta)$	$d^H(\mathbf{X}, \mathbf{X}_{\alpha+\beta})$	$d^H(\mathbf{X}, \mathbf{X}_{\alpha/\beta})$	Observed type	Predicted type
1BBL–	0.42*	0.63	0.56	0.48	α	α
1HBBA	0.41*	0.49	0.45	0.47	α	α
1IFA–	0.45*	0.48	0.50	0.53	α	α
1MRRA	0.26*	0.29	0.28	0.26	α	α
1PDE–	0.58	0.64	0.61	0.51*	α	α/β^c
1PRCM	0.48	0.45	0.45	0.40*	α	α/β^c
1SAS–	0.32*	0.33	0.33	0.34	α	α
2TMVP	0.51	0.35*	0.39	0.41	α	β^c
4CPV–	0.51*	0.65	0.60	0.55	α	α
1AAIB	0.49	0.27*	0.32	0.37	β	β
1ATX–	0.77	0.59*	0.62	0.65	β	β
1COBA	0.52	0.40	0.41	0.35*	β	α/β^c
1EGF–	0.76	0.61	0.60*	0.66	β	$\alpha+\beta^c$
1EST–	0.56	0.36*	0.38	0.45	β	β
1GPS–	0.85	0.78*	0.80	0.83	β	β
1HCC–	0.68	0.55	0.54*	0.59	β	$\alpha+\beta^c$
1IXA–	0.84	0.69*	0.73	0.76	β	β
1MDAA	0.43	0.47	0.42	0.41*	β	α/β^c
1PPFE	0.52	0.52	0.49	0.44*	β	α/β^c
1R1A2	0.37	0.24*	0.25	0.29	β	β
1SHFA	0.43	0.33*	0.35	0.39	β	β
1TIE–	0.34	0.29	0.27*	0.28	β	$\alpha+\beta^c$
1TNFA	0.32	0.32	0.27*	0.28	β	$\alpha+\beta^c$
2ACHB	0.92	0.75*	0.83	0.82	β	β
2CTX–	0.72	0.57*	0.60	0.64	β	β
2MEV1	0.48	0.18*	0.28	0.35	β	β
2PLV1	0.41	0.27	0.26*	0.33	β	$\alpha+\beta^c$
2SODO	0.52	0.40	0.41	0.35*	β	α/β^c
3RP2A	0.39	0.27	0.25	0.24*	β	α/β^c
4SGBI	0.75	0.69	0.64*	0.68	β	$\alpha+\beta^c$
5NN9–	0.52	0.25*	0.29	0.39	β	β
1ABH–	0.32	0.29	0.26	0.23*	$\alpha+\beta$	α/β^c
1BBPA	0.52	0.47	0.43*	0.45	$\alpha+\beta$	$\alpha+\beta$

Table 4 *(Continued.)*

PDB[b] code of 64 proteins	Hamming distance $d^{H}(\mathbf{X}, \mathbf{X}_{\alpha})$	$d^{H}(\mathbf{X}, \mathbf{X}_{\beta})$	$d^{H}(\mathbf{X}, \mathbf{X}_{\alpha+\beta})$	$d^{H}(\mathbf{X}, \mathbf{X}_{\alpha/\beta})$	Observed type	Predicted type
1BW4–	0.55	0.45	0.42*	0.46	$\alpha+\beta$	$\alpha+\beta$
1COX–	0.36	0.24*	0.26	0.24	$\alpha+\beta$	β^{c}
1DNKA	0.37	0.33	0.28*	0.30	$\alpha+\beta$	$\alpha+\beta$
1GLAG	0.28	0.29	0.25	0.17*	$\alpha+\beta$	α/β^{c}
1MS2A	0.44	0.30*	0.31	0.31	$\alpha+\beta$	β^{c}
1OVOA	0.61	0.48	0.45*	0.55	$\alpha+\beta$	$\alpha+\beta$
1POC–	0.48	0.37	0.36*	0.44	$\alpha+\beta$	$\alpha+\beta$
1PPBA	0.30	0.27	0.25	0.22*	$\alpha+\beta$	α/β^{c}
1SHAA	0.33	0.26*	0.27	0.37	$\alpha+\beta$	β^{c}
1THO–	0.34	0.44	0.40	0.33*	$\alpha+\beta$	α/β^{c}
1TRX–	0.35	0.45	0.41	0.34*	$\alpha+\beta$	α/β^{c}
2AAA–	0.50	0.27*	0.30	0.33	$\alpha+\beta$	β^{c}
2PIA–	0.36	0.24	0.23*	0.29	$\alpha+\beta$	$\alpha+\beta$
2SN3–	0.72	0.64	0.63*	0.68	$\alpha+\beta$	$\alpha+\beta^{c}$
2TAAA	0.46	0.23*	0.25	0.29	$\alpha+\beta$	β
3B5C–	0.36*	0.39	0.37	0.37	$\alpha+\beta$	α^{c}
3SC2A	0.35	0.29	0.23*	0.25	$\alpha+\beta$	$\alpha+\beta$
3SC2B	0.44	0.36	0.33*	0.37	$\alpha+\beta$	$\alpha+\beta$
3TLN–	0.48	0.35	0.35	0.34*	$\alpha+\beta$	α/β^{c}
4ENL–	0.22	0.33	0.26	0.18*	$\alpha+\beta$	α/β^{c}
4INSB	0.59	0.61	0.59	0.57*	$\alpha+\beta$	α/β^{c}
4RCRH	0.27	0.34	0.28	0.21*	$\alpha+\beta$	α/β^{c}
1GPB–	0.25	0.29	0.25	0.21*	α/β	α/β
1MINA	0.32	0.28	0.26	0.23*	α/β	α/β
1NIPB	0.34	0.40	0.33	0.24*	α/β	α/β
1SBP–	0.27	0.31	0.23	0.17*	α/β	α/β
1WSYA	0.27*	0.36	0.33	0.28	α/β	α^{c}
4ICD–	0.29	0.33	0.26	0.18*	α/β	α/β
7AATA	0.21	0.21	0.15	0.14*	α/β	α/β
9RUBB	0.29	0.33	0.28	0.22*	α/β	α/β
1GD1O	0.27	0.38	0.28	0.21*	α/β	α/β

Average rate of correct prediction = 34/64 = 53.1%

[a] The prediction was performed based on Equation 5. The one with the least value of d^{H} (marked by *) is assumed to correspond to the structural class for the predicted protein.

[b] See footnote *b* to Table 1A.

[c] Incorrect prediction.

Table 5
Predicted Results[a] by the Least Euclidean Distance Algorithm for the 64 Testing Proteins of Known X-Ray Structure Not Included in the Training Database

PDB[b] code of 64 proteins	Euclidean distance $d^{E}(\mathbf{X}, \mathbf{X}_{\alpha})$	$d^{E}(\mathbf{X}, \mathbf{X}_{\beta})$	$d^{E}(\mathbf{X}, \mathbf{X}_{\alpha+\beta})$	$d^{E}(\mathbf{X}, \mathbf{X}_{\alpha/\beta})$	Observed type	Predicted type
1BBL–	0.12*	0.17	0.15	0.14	α	α
1HBBA	0.11*	0.15	0.13	0.13	α	α
1IFA–	0.12*	0.13	0.13	0.14	α	α
1MRRA	0.08	0.08	0.07*	0.08	α	$\alpha+\beta^{c}$
1PDE–	0.15	0.17	0.16	0.14*	α	α/β^{c}
1PRCM	0.14	0.12	0.12	0.11*	α	α/β^{c}
1SAS–	0.10	0.11	0.10	0.09*	α	α/β^{c}
2TMVP	0.13	0.09*	0.10	0.11	α	β^{c}
4CPV–	0.14*	0.19	0.17	0.16	α	α
1AAIB	0.13	0.07*	0.08	0.09	β	β
1ATX–	0.22	0.17*	0.18	0.19	β	β
1COBA	0.15	0.12	0.12	0.11*	β	α/β^{c}
1EGF–	0.22	0.16*	0.17	0.19	β	β
1EST–	0.15	0.10*	0.11	0.11	β	β
1GPS–	0.26	0.23*	0.24	0.25	β	β
1HCC–	0.18	0.15	0.14*	0.15	β	$\alpha+\beta^{c}$
1IXA–	0.24	0.20*	0.21	0.22	β	β
1MDAA	0.12	0.13	0.11	0.10*	β	α/β^{c}
1PPFE	0.15	0.14	0.13	0.12*	β	α/β^{c}
1R1A2	0.11	0.06*	0.07	0.08	β	β
1SHFA	0.11	0.10	0.09*	0.10	β	$\alpha+\beta^{c}$
1TIE–	0.10	0.08	0.07*	0.08	β	$\alpha+\beta^{c}$
1TNFA	0.08	0.09	0.08*	0.09	β	$\alpha+\beta^{c}$
2ACHB	0.23	0.19*	0.20	0.21	β	β
2CTX–	0.23	0.17*	0.18	0.20	β	β
2MEV1	0.13	0.06*	0.08	0.10	β	β
2PLV1	0.12	0.07*	0.08	0.09	β	β
2SODO	0.15	0.12	0.12	0.11*	β	α/β^{c}
3RP2A	0.10	0.07	0.07	0.06*	β	α/β^{c}
4SGBI	0.23	0.20	0.19*	0.21	β	$\alpha+\beta^{c}$
5NN9–	0.14	0.06*	0.07	0.10	β	β
1ABH–	0.09	0.08	0.07	0.06*	$\alpha+\beta$	α/β^{c}
1BBPA	0.15	0.12	0.11*	0.12	$\alpha+\beta$	$\alpha+\beta$

Table 5 *(Continued.)*

PDB[b] code of 64 proteins	Euclidean distance $d^E(X, X_\alpha)$	$d^E(X, X_\beta)$	$d^E(X, X_{\alpha+\beta})$	$d^E(X, X_{\alpha/\beta})$	Observed type	Predicted type
1BW4–	0.14	0.12	0.11*	0.12	$\alpha+\beta$	$\alpha+\beta$
1COX–	0.10	0.08	0.07	0.06*	$\alpha+\beta$	α/β[c]
1DNKA	0.11	0.09	0.08*	0.10	$\alpha+\beta$	$\alpha+\beta$
1GLAG	0.08	0.08	0.06	0.05*	$\alpha+\beta$	α/β[c]
1MS2A	0.12	0.09	0.08*	0.09	$\alpha+\beta$	$\alpha+\beta$
1OVOA	0.17	0.14	0.13*	0.16	$\alpha+\beta$	$\alpha+\beta$
1POC–	0.14	0.10*	0.11	0.13	$\alpha+\beta$	β[c]
1PPBA	0.09	0.08	0.07*	0.08	$\alpha+\beta$	$\alpha+\beta$
1SHAA	0.09	0.07*	0.08	0.10	$\alpha+\beta$	β[c]
1THO–	0.10	0.12	0.10	0.09*	$\alpha+\beta$	α/β[c]
1TRX–	0.10	0.12	0.11	0.09*	$\alpha+\beta$	α/β[c]
2AAA–	0.13	0.08*	0.09	0.09	$\alpha+\beta$	β[c]
2PIA–	0.10	0.08	0.07*	0.08	$\alpha+\beta$	$\alpha+\beta$
2SN3–	0.20	0.18	0.17*	0.19	$\alpha+\beta$	$\alpha+\beta$
2TAAA	0.12	0.07*	0.08	0.08	$\alpha+\beta$	β[c]
3B5C–	0.10*	0.11	0.11	0.11	$\alpha+\beta$	α[c]
3SC2A	0.10	0.07	0.06*	0.07	$\alpha+\beta$	$\alpha+\beta$
3SC2B	0.13	0.10	0.09*	0.10	$\alpha+\beta$	$\alpha+\beta$
3TLN–	0.14	0.10	0.10	0.09*	$\alpha+\beta$	α/β[c]
4ENL–	0.06	0.09	0.07	0.05*	$\alpha+\beta$	α/β[c]
4INSB	0.16	0.16	0.15*	0.16	$\alpha+\beta$	$\alpha+\beta$
4RCRH	0.09	0.09	0.08	0.07*	$\alpha+\beta$	α/β[c]
1GPB–	0.07	0.08	0.07	0.06*	α/β	α/β
1MINA	0.10	0.07	0.07	0.06*	α/β	α/β
1NIPB	0.09	0.11	0.09	0.06*	α/β	α/β
1SBP–	0.07	0.08	0.06	0.05*	α/β	α/β
1WSYA	0.09	0.12	0.10	0.08*	α/β	α/β
4ICD–	0.08	0.09	0.07	0.05*	α/β	α/β
7AATA	0.06	0.06	0.05	0.04*	α/β	α/β
9RUBB	0.09	0.10	0.08	0.06*	α/β	α/β
1GD1O	0.08	0.10	0.08	0.06*	α/β	α/β

Average rate of correct prediction = 36/64 = 56.3%

[a] The prediction was performed based on Equation 8. The one with the least value of d^E (marked by *) is assumed to correspond to the structural class for the predicted protein.

[b] See footnote *b* to Table 1A.

[c] Incorrect prediction.

extrapolating effectiveness of a prediction method. This is important especially for checking the representativeness of a training database: whether it contains sufficient information to yield a high success rate in application. However, how to select the testing data set for the cross-validation is a delicate problem. Also, the concept of cross-validation is worthy of further clarification.

As is well known, the single-test-set analysis, subsampling, and jackknife analysis are the three methods often used for cross-validation examination. According to the single-test-set examination, the prediction rules are derived from the training set, and they are examined by observing the predicted results for the proteins in the test set. However, the selection of a testing set is quite arbitrary, and different selections might yield different results. For example, outside the 64 proteins in his training set, P.Y. Chou (1989) selected 12 proteins as a test set. The reported average rate of correct prediction by the least Hamming distance algorithm for this test set was 83.3%. However, applying the same algorithm to another test set of 35 proteins (Chou and Zhang, 1993) yielded an average rate of correct prediction of only 74.3%. Which of these two accuracy rates should count more? Neither, obviously. Furthermore, it should be realized that even if a comparison is made based on the same database, the accountability of the results thus obtained could still be questionable. This is because an algorithm, which gives the best predicted results for a testing set of proteins, does not necessarily remain so when applied to another testing set of proteins. In other words, the accuracy thus obtained lacks an objective criterion unless the data in the testing set are sufficiently large. On the other hand, even if a testing protein is incorrectly predicted by an algorithm, this would not necessarily mean that anything is wrong with the algorithm because that protein might be just outside the frame of the structural classes defined by the limited number of proteins in the current training database. A problem like this cannot be avoided unless the training database has become an ideal one, that is, a statistically complete one able to represent all the testing proteins concerned. Unfortunately, to date, there are only a few hundred proteins whose 3-D structures have been determined. It is far too premature to constitute a statistically complete training database and a sufficiently large testing database based on the structure-known proteins.

Another approach for cross-validation is the subsampling analysis, according to which a given database is divided into a training set and a testing set. However, the serious problem is how to divide a whole set into a training set and a testing set. As shown below, the number of possible divisions might be extremely large. Suppose a set of N proteins is divided into a training set (with N_1 proteins) and a testing set (with N_2 proteins). The number of such divisions is given by $\frac{(N_1+N_2)!}{(N_1)!N_2!} = \frac{N!}{(N-N_2)!N_2!}$. When $N = 120$ and $N_2 = 10$, the number of possible divisions would be $\sim 1.16 \times 10^{14}$! This is an astronomical figure, too large for any practical application. In practice, therefore, analyses could be carried out only for a very small portion of possible divisions that was selected randomly or arbitrarily.

Compared with the single-set-test examination and subsampling analysis, the jackknife test (Mardia et al., 1979) seems to be most reliable. With the jackknife test, also called the leave-one-out test (Mardia et al., 1979; Klein, 1986), the structural class of each protein is predicted by the rules derived using all other proteins except the one being predicted.

However, when N, the number of proteins in a given set, is not large enough, the leave-one-out test, in which each protein is in turn left out of the set, may result in a severe loss of information. Under such a circumstance, the leave-one-out test cannot be utilized.

In view of this, it would be very helpful if a large number of simulated proteins with a given structural class feature could be generated by Monte Carlo sampling. To accomplish this, an effort was made based on the normal distribution assumption (Zhang and Chou, 1992b). However, one may raise the following question: why was a normal distribution, but not a skewed Gaussian or some other unspecified distribution, used as the underlying distribution? To avoid such an arbitrary assumption, a seed-propagated sampling method is being developed to generate the simulated proteins based only on the experimental database itself without introducing any predetermined assumption about the distribution function.

A. Seed-Propagated Sampling

The process of the seed-propagated sampling method can be illustrated as follows.

1. Sampling of α-Proteins

Suppose that a given α-class set contains N_α proteins whose amino acid compositions are expressed by $\{x_{k,1}(\alpha), x_{k,2}(\alpha), \ldots, x_{k,20}(\alpha)\}$, where $x_{k,i}(\alpha)$ is the occurrence frequency of the ith amino acid of the kth α-protein ($i = 1, 2, \ldots, 20$; $k = 1, 2, \ldots, N_\alpha$). Here, the 20 amino acids are numbered according to the alphabetic order of their single-letter codes: A, C, D, E, F, G, H, I, J, K, L, M, N, P, Q, R, S, T, V, W, and Y. The amino acid composition of a protein must be normalized, and hence we have

$$\sum_{i=1}^{20} x_{k,i}(\alpha) = 1, \qquad (k = 1, 2, \ldots, N_\alpha) \tag{40}$$

The accumulated probability distribution of the 20 amino acid frequencies for the kth α-protein may be described by $F_{k,1}$, $F_{k,2}$, $\ldots$, and $F_{k,20}$, respectively, where

$$\begin{cases} F_{k,1}(\alpha) &= x_{k,1}(\alpha) \\ F_{k,2}(\alpha) &= x_{k,1}(\alpha) + x_{k,2}(\alpha) \\ F_{k,3}(\alpha) &= x_{k,1}(\alpha) + x_{k,2}(\alpha) + x_{k,3}(\alpha) \\ \cdots &= \cdots \\ F_{k,i}(\alpha) &= \sum_{j=1}^{i} x_{k,j}(\alpha) \\ \cdots &= \cdots \\ F_{k,20}(\alpha) &= \sum_{j=1}^{20} x_{k,j}(\alpha) = 1 \end{cases} \tag{41}$$

For convenience, let $F_{k,0}(\alpha) = 0$. Thus, the values in Equation 41 are uniquely defined by the amino acid composition of the kth α-protein of the training database.

The Monte Carlo sampling procedures are illustrated by the following steps:

1. Randomly and uniformly choose a value of k in a set of integers (1, 2, ..., N_α). For example, for the database used here, $N_\alpha = 30$, meaning that there are 30 proteins in the α-protein training subset.

2. Generate a random number $\Re$ uniformly distributed between 0 and 1. If

$$F_{k,i-1}(\alpha) < \Re \leq F_{k,i}(\alpha), \qquad (i = 1, 2, \ldots, 20) \tag{42}$$

the ith amino acid is drawn and counted once. If $i = 1$, A (alanine) will be counted once; if $i = 2$, then C (cysteine) counted once, and so forth. This step is actually a subsampling process.

3. Repeat the above subsampling N_S times, where N_S can be any large integer ($N_S = 10^4$ in this study). It will generate a sampled α-protein whose amino acid composition is given by

$$x^*_{k,i}(\alpha) = \frac{n_{k,i}(\alpha)}{N_S} \qquad (i = 1, 2, \ldots, 20) \tag{43}$$

where $n_{k,i}(\alpha)$ is the total number of the ith amino acid being drawn, and N_S is the number of the subsampling cycles as described in step 2.

The above steps constitute a Monte Carlo sampling cycle for generating a simulated protein of the α-class according to the amino acid compositions derived directly from the α-protein training set (see Appendix G for a further explanation of the rationale of the sampling technique). Repeating steps 1 to 3 N^* times will generate N^* simulated α-proteins.

2. Sampling Proteins of All Other Classes

For the other structural class (i.e., β-, $\alpha + \beta$-, or α/β-proteins), by substituting the α of the aforementioned steps with β, $\alpha + \beta$, or α/β, respectively, we can obtain the corresponding formulations. The sampling process is completely parallel to the one for the α-proteins.

By following the above procedure, for a given training database of α-proteins, one can generate any number of simulated proteins whose amino acid compositions are characterized by the feature of an α-protein class as observed. The same is true for all the other structural classes. It is instructive to associate such an approach with the "seed-propagated mechanism" (Chou, 1993) or "bootstrap mechanism" (Efron, 1990) widely used in statistical mathematics. In this sense, the α-, β-, $\alpha + \beta$-, and α/β-proteins in the training database are the original "seeds" for the α-, β-, $(\alpha+\beta)$-, and (α/β)-classes, while the simulated proteins thus generated are their propagated "progenies". Individually speaking, as a consequence of "mutation", many simulated proteins may not be found at all in the original training database. Statistically speaking, however, the amino acid compositions of the simulated proteins thus generated must

"inherit" the essential characteristics of the "seed" proteins because they are derived from them by the Monte Carlo sampling. This kind of treatment is quite useful, especially when the experimental data are insufficiently sampled for complete statistics.

B. Simulated Accuracy

As demonstrated in a previous paper (Zhang and Chou, 1992b), when the number of simulated proteins generated by Monte Carlo sampling for each class is greater than 3000, the rate of correct prediction will gradually approach the so-called *asymptotical limit*. Under such a circumstance, the errors due to statistical fluctuations can be omitted. Such an asymptotical limit was defined as the *simulated accuracy of prediction*. Using the simulated accuracy can eliminate the noise caused by a limited amount of data (Chou, 1995a), and hence more accurately reflect the objective reality.

In view of this, 5000 simulated proteins were generated for each of the four regular structural classes by the seed-propagated sampling based on the 4×30 proteins in the training database (Appendix A). The number of subsampling cycles selected for each protein was $N_S = 10^4$. For these simulated proteins, predictions were performed by the least Mahalanobis distance algorithm, the least Hamming distance algorithm (P.Y. Chou, 1980, 1989), and the least Euclidean distance algorithm (Nakashima et al., 1986). The predicted results thus obtained are given in Table 6, which shows that the average simulated accuracy by the least Mahalanobis distance algorithm is 97.29%, significantly higher than those by the other two algorithms.

A comparison of Tables 6 and Table 1 indicates that although the rates of correct prediction by the least Mahalanobis distance algorithm for the 30 α-, 30 β-, and 30 $\alpha + \beta$-proteins in the training database are all the same (equal to 100%), those calculated from the corresponding 5000 simulated proteins are different: the accuracies for the α- and β-classes are higher than that of $\alpha + \beta$-class. Why is that? The answer can be found by analyzing Figure 3. As we can see from Figure 1A, all 30 α-proteins have much shorter Mahalanobis distances to the norm of their own class than to the norms of others, meaning that the distribution of α-proteins in the composition space is relatively more concentrated, or with a more clear-cut "border" defined according to the Mahalanobis distance. Under such a circumstance, the error of statistical fluctuation due to a limited number of proteins is relatively smaller, and hence the rate calculated directly from the training database is quite closer to its simulated accuracy. The same is true for the case of β-proteins. However, the situation is quite different for the $\alpha + \beta$-proteins. As shown in Figure 3C, although all 30 $\alpha + \beta$-proteins have the shortest Mahalanobis distances to the norm of their own class, some of them are just slightly shorter than their counterparts, meaning that, even measured according to the Mahalanobis distance, the distribution of $\alpha + \beta$-proteins in the composition space is not so clear-cut, or has a slightly ambiguous "border". The 100% rate of correct prediction obtained according to Figure 3C for the $\alpha + \beta$-proteins may be just a result of statistical errors due to a limited amount of data, and it cannot, of course, be retained for a large amount of statistical data. For the case of α/β-proteins, as can be seen from Figure 1D, this kind of ambiguity is even higher, and

Table 6
Comparison of Various Prediction Algorithms for 5000 × 4 Simulated Proteins

Algorithm	Rate of correct prediction				Average accuracy[d]
	α class	β class	$\alpha + \beta$ class	α/β class	
Component coupled[a]	$\frac{5000}{5000} = 100\%$	$\frac{4995}{5000} = 99.90\%$	$\frac{4895}{5000} = 97.90\%$	$\frac{4530}{5000} = 90.60\%$	$\frac{19420}{20000} = 97.10\%$
P. Y. Chou[b] (1989)	$\frac{3528}{5000} = 70.56\%$	$\frac{3579}{5000} = 71.58\%$	$\frac{2081}{5000} = 41.62\%$	$\frac{4302}{5000} = 86.04\%$	$\frac{13490}{20000} = 67.45\%$
Nakashima et al.[c] (1986)	$\frac{3829}{5000} = 76.58\%$	$\frac{3280}{5000} = 65.60\%$	$\frac{1877}{5000} = 37.54\%$	$\frac{4283}{5000} = 85.66\%$	$\frac{13269}{20000} = 66.35\%$

[a–d] See the corresponding footnotes to Table 3.

hence the statistical error for the rate thus obtained would be even greater.

A similar analysis can also be used to elucidate the results of Table 6 obtained by the least Hamming distance method (P.Y. Chou, 1980, 1989) or the least Euclidean distance method (Nakashima et al., 1986). However, when doing so, the corresponding ambiguity should be defined according to the Hamming or Euclidean distance, respectively.

C. Jackknife Analysis

As mentioned above, the jackknife analysis is also called the leave-one-out test (Klein, 1986). During the process of jackknife analysis, the training set is no longer closed but an opened one, and a protein will in turn be removed from it as a testing sample. However, owing to the definition of the covariance matrix (see Equation 37), in order to make a statistically meaningful leave-one-out test for the least Mahalanobis distance method, the number of proteins in each structural class should be relatively larger. Actually, the more the proteins in each class, the less the loss of information during the leave-one-out test cycles. In view of this, 40 simulated proteins were generated for each of the four structural classes by the seed-propagated sampling based on the 4×30 proteins in the training database (Appendix A). The number of subsampling cycles selected for each protein was $N_S = 10^4$. For a database of such $4 \times 40 = 160$ proteins, a leave-one-out test was performed. The average rate of correct prediction by the least Mahalanobis distance algorithm is 85.0%, and those by the least Hamming distance algorithm (P.Y. Chou, 1980, 1989) and the least Euclidean distance algorithm (Nakashima et al., 1986) are 62.5 and 68.8%, respectively. Interestingly, if the leave-one-out test is carried out for a database of $4 \times 80 = 320$ simulated proteins, the average rate of correct prediction by the least Mahalanobis distance algorithm increases to 94.4% while those by the other two algorithms remain almost unvaried, fluctuating around 62 and 68%, respectively, with an amplitude of less than 1.5%. This indicates that the jackknife-tested rate by the least Mahalanobis distance method is not only much higher than those tested by the others, but the potential of increasing its rate by improving the database is also much greater.

VI. Conclusions

Although the number of protein sequences is extremely large, the number of their folding patterns is quite limited. This is due to the degenerate nature of the sequence-structure relationship. It has been demonstrated by experimental mutagenesis studies (e.g., Sondek and Shortle, 1990) that the overall fold of a protein is much more tolerant to sequence modifications than previously suspected. Analyses of known 3-D structures have also revealed structural similarity for proteins with very different sequences and functions (Farber and Petsko, 1990; Kabsch et al., 1990). The recent studies by Muskal and Kim (1992) further suggested that the structural class of a protein may basically depend on its amino acid composition. Actually, many attempts had been made to predict the structural class of a protein based on its amino acid composition, but failed to reach the desired accuracy. Recently, a

new prediction algorithm was developed that seems quite promising. The unique feature of the new algorithm is that, instead of using the ordinary geometric distances such as Euclidean distance and Minkowski's distance, the Mahalanobis distance is used as a scale to measure the similarity of two proteins in the amino acid composition space. The virtue of the Mahalanobis distance is that it incorporates the coupling of different amino acid components, which distinguishes it from the previous algorithms. The high rates of correct prediction for proteins in both training set and testing set, which have been further verified by the jackknife analysis and simulated accuracy, imply that the new algorithm will become a reliable tool for predicting the structural class of a protein if a statistically complete database for classifying protein structure classes becomes available. How large will the desired database be? According to a recent estimation by Chothia (1992), the large majority of proteins come from about 1000 families. If he is correct, the desired complete database should consist of about 1000 nonhomologous proteins.

The high rates of successful predictions also suggest (Chou, 1995b) that the overall fold of a protein (e.g., the structural class) is basically determined by its overall sequence ingredients (i.e., the amino acid composition).

Acknowledgments

The authors thank Joan Baker and Raymond B. Moeller for their help of drawing Figure 1A–D. K. C. expresses special gratitude to Bi-Kun Luo. Without her encouragement, he would not have been able to overcome the difficulties that occurred during this study.

References

Banner, D.W., Bloomer, A.C., Petsko, G.A., Phillips, D.C., Wilson, I.A. (1976) "Atomic coordinates for triose phosphate isomerase from chicken muscle." *Biochem. Biophys. Res. Commun.*, **72**, 146.

Bezdek, J.C., (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum, New York.

Brady, R.L., Dodson, E.J., Dodson, G.G., Lange, G., Davis, S.J., Williams, A.F., and Barclay, A.N. (1995) "Crystal structure of domains 3 and 4 of rat CD4 and their relationship to the NH_2-terminal domains," in press.

Bussian, B.M. and Sander, C. (1989) "How to determine protein secondary structure in solution by Raman spectroscopy: practical guide and test case DNase I." *Biochemistry*, **28**, 4271–4277.

Carlacci, L., Chou, K.C., and Maggiora, G.M. (1991) "A heuristic approach to predicting the tertiary structure of bovine somatotropin." *Biochemistry*, **30**, 4389–4398.

Chothia, C. and Finkelstein, A.V. (1990) "The classification and origins of protein folding patterns." *Annu. Rev. Biochem.* **59**, 1007–1039.

Chou, K.C. (1992) "Energy-optimized structure of antifreeze protein and its binding mechanism." *J. Mol. Biol.*, **223**, 509–517.

Chou, K.C. (1994) "A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins." *J. Biol. Chem.*, **268**, 16938–16948.

Chou, K.C. (1995a) "A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space." *Proteins Struct. Funct. Genet.*, **21**, 319–344.

Chou, K.C. (1995b) "Does the folding type of a protein depend on its amino acid composition?" *FEBS Lett.*, **363**, 127–131.

Chou, K.C. and Carlacci, L. (1991a) "Energetics approach to the folding of α/β barrels." *Proteins Struct. Funct. Genet.*, **9**, 280–295.

Chou, K.C. and Carlacci, L. (1991b) "Simulated annealing approach to the study of protein structures." *Protein Eng.*, **4**, 661–667.

Chou, K.C. and Scheraga, H.A. (1982) "Origin of the right-handed twist of β-sheets of poly(L-Val) chains." *Proc. Natl. Acad. Sci. U.S.A*, **79**, 7047-7051.

Chou, K.C., Némethy, G. and Scheraga, H.A. (1983) "Effect of amino acid composition on the twist and the relative stability of parallel and antiparallel β-sheets." *Biochemistry*, **22**, 6213–6221.

Chou, K.C., Némethy, G. and Scheraga, H.A. (1989) "Energy of stabilization of the right-handed $\beta\alpha\beta$ crossover in proteins, *J. Mol. Biol.*, **205**, 241-249.

Chou, K.C., Némethy, G. and Scheraga, H.A. (1990) "Energetics of interactions of regular structural elements in proteins." *Acc. Chem. Res.*, **23**, 134–141.

Chou,K.C. and Zhang, C.T. (1993) "A new approach to predicting protein folding types." *J. Protein Chem.*, **12**, 169-178.

Chou,K.C. and Zhang, C.T. (1994) "Predicting protein folding types by distance functions that make allowances for amino acid interaction." *J. Biol. Chem.*, **269**, 22014-22020.

Chou, P.Y. (1980) "Amino acid composition of four classes of proteins." in Abstracts of Papers, Part I, Second Chemical Congress of the North American Continent, Las Vegas.

Chou, P.Y. (1989) "Prediction of protein structural classes from amino acid composition." in *Prediction of Protein Structure and the Principles of Protein Conformation,* Fasman, G.D., Ed., Plenum Press, New York, 549–586.

Chou, P.Y. and Fasman, G.D. (1974) "Prediction of protein conformation." *Biochemistry* **13**, 222–245.

Chou, P.Y. and Fasman, G.D. (1978) "Prediction of secondary structure of proteins from amino acid sequence." *Adv. Enzymol. Relat. Subj. Biochem.* **47**, 45–148.

Chothia, C. (1992) "One thousand families for the molecular biologist." *Nature*, **357**, 543–544.

Cid, H., Bunster, M., Canales, M., and Gazitua, F. (1992) "Hydrophobicity and structural classes in proteins." *Proteins Eng.*, **5**, 373–375.

Cohen, F.E. and Kuntz, I.D. (1987) "Prediction of the three-dimensional structure of human growth hormone." *Proteins Struct. Funct. Genet.*, **2**, 162–166.

Cohen, B., Presnell, S.R., and Cohen, F.E (1993) "Origins of structural diversity within sequentially identical hexapeptides." *Proteins Sci.*, **2**, 2134–2145.

Dayhoff, M.O. and Eck, R.V., Eds. (1968) *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Silver Spring, MD, Vol. 3, 33.

Deléage, G. and Dixon, J.S. (1989) "Use of class prediction to improve protein secondary structure prediction." in *Prediction of Protein Structure and the Principles of Protein Conformation,* Fasman, G.D., Ed., Plenum Press, New York, 587–597.

Deléage, G. and Roux, B. (1987) "An algorithm for protein secondary structure prediction based on class prediction." *Protein Eng.*, **1**, 289–294:

Dubchak, I. , Holbrook, S. R., and Kim, S-H. (1993) "Predicting protein secondary structure content: A tandem neural network approach." *Proteins Struct. Funct. Genet.*, **16**, 79–91.

Efron, B. (1990) *The Jackknife, the Bootstrap and Other Resampling Plans*, Society for Industrial and Applied Mathematics, Philadelphia, chap. 5.

Eisenhaber, F., Persson, B., and Argos, P. (1995) "Protein structure prediction: recognition of primary, secondary, and tertiary structural features from amino acid sequence." *Crit. Rev. Biochem. Mol. Biol.*, **30**, 1–94.

Fasman, G.D. (1989) "The development of the prediction of protein structure." in *Prediction of Protein Structure and the Principles of Protein Conformation,* Fasman, G.D., Ed., Plenum Press, New York, 317–358.

Farber, G.K. and Petsko, G.A., (1990) "The evolution of α/β barrel enzymes." *TIBS*, **15**, 228–234.

Fetrow, J.S. and Bryant, S.H. (1993) "New programs for protein tertiary structure prediction." *Bio/Technology* **11**, 479–484.

Finkelstein, A.V. and Ptitsyn, O.B. (1987) "Why do globular proteins fit the limited set of folding patterns?" *Prog. Biophys. Mol. Biol.*, **50**, 171–190.

Finzel, B.C., Weber, P.C., Hardman, K.D., and Salemme, F.R. (1985) "Structure of ferricytochrome c' from rhodospirillum *Molischianum* at 1.67 Å resolution." *J. Mol. Biol.*, **186**, 627–643.

Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) "Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins." *J. Mol. Biol.* **120**, 97–120.

Gilson, M.K. and Honig, B. (1988) "Energetics of charge-charge interactions in proteins, *Proteins Struct. Funct. and Genet.*, **3**, 32–52.

Gower, J.C. (1985) "Measures of similarity, dissimilarity, and distance." *Encyclopedia of Statistical Sciences,* Vol.5, Kotz, S. and Johnson, N.L., Eds., John Wiley and Sons, New York, 397–405.

Henikoff, S. and Henikoff, J.G. (1992) "Amino acid substitution matrices from protein blocks." *Proc. Natl. Acad. Sci. U.S.A*, **89**, 10915–10919.

Henikoff, S. and Henikoff, J.G. (1994) "Protein family classification based on searching a database of blocks." *Genomics*, **19**, 97–107.

Hobohm, U. and Sander, C. (1994) "Enlarged representative set of protein structures." *Protein Sci.* **3**, 522–524.

Honzatko, R.B., Crawford, J.L., Monaco, H.L., Ladner, J.E., Edwards, B.F.P., Evans, D.R., Warren, S.G., Wiley, D.C., Ladner, R.C., and Lopscomb, W.N. (1982) "Crystal and molecular structures of native and ctP-ligand aspartate carbamoyltransferase from escherichia *coli*." *J. Mol. Biol.*, **160**, 219–263.

Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992) "A new approach to protein fold recognition." *Nature*, **358**, 86–89.

Johnson, W.C., Jr. (1990) "Protein secondary structure and circular dichroism: a practical guide." *Proteins Struct. Funct. Genet.*, 7, 205–214.

Jones, D.T., Taylor, W.R., and Thornton, J.M. (1994) "A model recognition approach to the prediction of all-helical membrane protein structure and topology." *Biochemistry*, **33**, 3038–3049.

Kabsch, W., Mannherz, H.G., Suck, D., Pai, E.F., Holms, K.C. (1990) "Atomic structure of the actin:DNase I complex." *Nature*, **347**, 37–44.

Kabsch, W. and Sander, C. (1983) "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." *Biopolymers* **22**, 2577–2637.

Karplus, M. and Shakhnovich, E. (1992) "Theoretical studies of thermodynamics and dynamics." in *Protein Folding*, Creighton, T.E., Ed., Freeman, New York, 127–195.

Kawai, H., Kikuchi, T., and Okamoto, Y. (1989) "A prediction of tertiary structures of peptide by the Monte Carlo simulated annealing method." *Protein Eng.*, **3**, 85–94.

Kikuchi, T. (1993) "Discrimination of folding types of globular proteins based on average distance maps constructed from their sequences." *J. Protein Chem.* **12**, 515–523.

Klein, P. (1986) "Prediction of protein structural class by discriminant analysis." *Biochim. Biophys. Acta*, **874**, 205–215.

Klein, P. and Delisi, C (1986) "Prediction of protein structural class from amino acid sequence." *Biopolymers*, **25**, 1569–1672.

Kneller, D.G., Cohen, F.E., and Langridge, R. (1990) "Improvements in protein secondary structure prediction by enhanced neural networks." *J. Mol. Biol.*, **214**, 171–182.

Kolinski, A. and Skolnick, J. (1994) "Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme." *Proteins Struct. Funct. Genect*, **18**, 338–352.

Kuwajima, K, Semisotnov, G.V., Finkelstein, A.V., Sugai, S. and Ptitsyn, O.B. (1993) "Secondary structure of globular proteins at the early and the final stages in protein folding." *FEBS Lett.*, **334**, 265–268. (Published erratum appears in *FEBS Lett.*, 1993, **336**, 190.)

Levitt, M. (1983) "Protein folding by restrained energy minimization and molecular dynamics." *J. Mol. Biol.*, **170**, 723–764.

Levitt, M. and Chothia, C. (1976) "Structural patterns in globular proteins." *Nature*, **261**, 552–557.

Lim, V.I. (1974) "Structural principles of globular protein secondary structure." *J. Mol. Biol.* **88**, 857–872.

Loll, P.J. and Lattman, E.E. (1989) "The crystal structure of the ternary complex of staphylococcal nuclease." *Proteins Struct. Funct. Genet.*, **5**, 183–201.

Mao, B., Chou, K.C., and Zhang, C.T. (1994) "Protein folding classes: a geometric interpretation of the amino acid composition of globular proteins." *Protein Eng.*, **7**, 319–330.

Mackay, D.H.J., Cross, A.J., and Hagler, A.T. (1989) "The role of energy minimization in simulation strategies of biomolecular systems." in *Prediction of Protein Structure and the Principles of Protein Conformation*, Fasman, G.D., Ed., Plenum Press, New York, 317–358.

Mahalanobis, P.C. (1936) "On the generalized distance in statistics." *Proc. Natl. Inst. Sci. India*, **2**, 49–55.

Mardia, K.V., Kent, J.T. and Bibby, J.M. *Multivariate Analysis*, Academic Press, London, 322 and 381.

McCammon, J.A., Wong, C.F., and Lybrand, T.P. (1989) "Protein stability and function." in *Prediction of Protein Structure and the Principles of Protein Conformation,* Fasman, G.D., Ed., Plenum Press, New York, 149–159.

McDonald, I.K. and Thornton, J.M. (1994) "Satisfying hydrogen bonding potential in proteins." *J. Mol. Biol.* **238**, 777–793.

Metfessel, B. A., Saurugger, P. N., Connelly, D. P., and Rich, S. T. (1993) "Cross-validation of protein structural class prediction using statistical clustering and neural networks." *Protein Sci.*, **2**, 1171–1182.

Mitchell, J.B., Nandi, C.L., McDonald, I.K., Thornton, J.M., Price, S.L. (1994) "Amino/aromatic interactions in proteins: is the evidence stacked against hydrogen bonding?" *J. Mol. Biol.* **239**, 315–331.

Miyazawa, S. and Jernigan, R.I. (1993) "A new substitution matrix for protein sequence searches based on contact frequencies in protein structures." *Protein Eng.*, **6**, 267–278.

Muggleton, S., King, R.D., and Sternberg, M.J.E. (1992) "Protein secondary structure prediction using logic-based machine learning." *Protein Eng.*, **5**, 647–657. (Corrigenda: *Protein Eng.*, (1993), **6**, 549.)

Muskal, S.M.,and Kim, S.-H. (1992) "Predicting protein secondary structure content: A tandem neural network approach." *J. Mol. Biol.*, **225**, 713–727.

Nakashima, H., Nishikawa, K., and Ooi, T. (1986) "The folding type of a protein is relevant to the amino acid composition." *J. Biochem.*, **99**, 152–162.

Nishkawa, K. and Ooi, T. (1982) "Correlation of the amino acid composition of a protein to its structural and biological characters." *J. Biochem.* **91**, 1821–1824.

Nishkawa, K., Kubota, Y., and Ooi, T. (1983a) "Classification of proteins into groups based on amino acid composition and other characters. I. Angular distribution." *J. Biochem.*, **94**, 981–995.

Nishkawa, K., Kubota, Y., and Ooi, T. (1983b) "Classification of proteins into groups based on amino acid composition and other characters, II. Grouping into four types." *J. Biochem.*, **94**, 997–1007.

Orengo, C.A., Jones, D.T., and Thornton, J.M. (1994) "Protein superfamilies and domain superfolds." *Nature*, **372**, 631–634.

Perczel, A., Hollósi, M., Tusnády, G., and Fasman, G.D. (1991) "Convex constraint analysis: a natural deconvolution of circular dichroism curves of proteins." *Protein Eng.*, **4**, 669–679.

Pillai, K.C.S. (1985) "Mahalanobis D^2." *Encyclopedia of Statistical Sciences,* Vol. 5, Kotz, S. and Johnson, N.L., Eds., John Wiley and Sons, New York, 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.)

Ptitsyn, O.B. and Finkelstein, A.V. (1979) "Mechanism of protein folding." *Int. J. Quantum Chem.*, **16**, 407–418.

Ptitsyn, O.B. and Finkelstein, A.V. (1980) "Similarities of protein topologies: evolutionary divergence, functional convergence or principles of folding?" *Q. Rev. Biophys.*, **13**, 339–386.

Ptitsyn, O.B. and Finkelstein, A.V. (1989) "Prediction of protein secondary structure based on physical theory. Histones." *Protein Eng.*, **2**, 443–447.

Ptitsyn, O.B. and Rashin, A.A. (1975) "A model of myoglobin self-organization." *Biophys. Chem.*, **3**, 1–20.

Ptitsyn, O.B., Finkelstein, A.V., and Murzin, A.G. (1985) "Structural model for interactions." *FEBS Lett.*, **186**, 143–148.

Rao, S.S. (1984) *Optimization, Theory and Applications*, 2nd ed., Wiley Eastern Limited, New Delhi, chap. 6.

Richardson, J.S. (1977) "β-sheets topology and the relatedness of proteins." *Nature*, **268**, 495–500.

Richardson, J.S. (1981) "The anatomy and taxonomy of protein structure." *Adv. Protein Chem.*, **34**, 167–339.

Richardson, J.S. and Richardson, D.C. (1989) "Principles and patterns of protein conformation." in *Prediction of Protein Structure and the Principles of Protein Conformation*, Fasman, G.D., Ed., Plenum Press, New York, 1–98.

Rost, B., and Sander, C. (1994) "Combining evolutionary information and neural networks to predict protein secondary structure." *Protein Struct. Funct. Genet.*, **19**, 55–72.

Sreerama, N. and Woody, R.W. (1994) "Protein secondary structure from circular dichroism spectroscopy." *J. Mol. Biol.*, **242**, 497–507.

Sondek, J. and Shortle, D. (1990) "Accomodation of single amino acid insertions by the native state of staphyloccocal nuclease." *Proteins Struct. Funct. Genect.*, **7**, 299–305.

Rogers, N.K. (1989) "The role of electrostatic interactions in the structure of globular proteins." in *Prediction of Protein Structure and the Principles of Protein Conformation*, Fasman, G.D., Ed., Plenum Press, New York, 359–389.

Scheraga, H.A. (1968) "Calculations of conformations of polypeptides." *Adv. Phys. Org. Chem.* **6**, 103–184.

Scheraga, H.A. (1987) "Conformational analysis of polypeptides and proteins for the study of protein folding, molecular recognition, and molecular design." *J. Prot. Chem.* **6**, 61–80.

Vieth, M., Kolinski, A, Brooks, C.L. and Skolnick, J. (1994) "Prediction of the folding pathways and structure of the GCN4 leucine zipper." *J. Mol. Biol.* **237**, 361–367.

Weiner, P.K. and Kollman, P.A. (1981) "AMBER: Assisted model building with energy refinement. a general program for modeling molecules and their interactions." *J. Comp. Chem.* **2**, 287–303.

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

Wodak, S.J. and Rooman, M.J. (1993) "Generating and testing protein folds." *Curr. Opin. Struct. Biol.* **3**, 247–259.

Wylie, C.R. and Barrell, L.C. (1982) *Advanced Engineering Mathematics*, 5th ed., McGraw-Hill, New York, 769.

Zhang, C.T. and Chou, K.C. (1992a) "An optimization approach to predicting protein structural class from amino acid composition." *Protein Sci.* **1**, 401–408.

Zhang, C.T. and Chou, K.C. (1992b). "Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition." *Biophys. J.*, **63**, 1523–1529.

Zhang, C.T., Chou, K.C. and Maggiora, G.M. (1995) "Predicting protein structural classes from amino acid composition: application of fuzzy clustering." *Protein Engineering*, in press.

Zhang, C.T. and Chou, K.C. (1995) "An analysis of predicting protein folding types by seed-propagated sampling and jackknife test." *J. Protein Chem.*, in press.

Zhou, G.F., Xu, X., and Zhang, C.T. (1992) "A weighting method for predicting protein strctural class from amino acid composition." *Eur. J. Biochem.*, **210**, 747–749.

Appendix A

The amino acid composition of the 129 proteins of which 30 are α proteins, 30 β proteins, 30 α+β proteins, 30 α/β proteins, and 9 irregular proteins. The data of each protein contain two lines: the 1st line successively indicates its length, PDB code, the ratios of α, β, parallel β sheets, and antiparallel sheets (see eq.1); and the 2nd line gives the frequencies of 20 amino acids according to the alphabetical order of the single amino acid letter code: ACDEFGHIKLMNPQRSTVWY. The frequencies are normalized to 100. The fifth character in the PDB code indicates a specific chain of a protein; if it is -, meaning the corresponding protein has only one chain.

---

30 α proteins

```
318 1AVHA   0.68  0.00  0.00  0.00
 7.86  0.31  7.86  9.12  4.09  6.92  0.94  5.66  6.92 11.95  2.20  1.89  1.57  3.77  5.97  6.60  7.23  5.03  0.31  3.77
146 1BABB   0.68  0.00  0.00  0.00
10.27  1.37  4.79  5.48  5.48  8.90  6.16  0.00  7.53 12.33  0.68  4.11  4.79  2.05  2.05  3.42  4.79 12.33  1.37  2.05
177 1BRD-   0.88  0.00  0.00  0.00
10.17  0.00  2.82  2.26  5.65  9.04  0.00  6.78  2.82 19.21  3.95  1.69  2.82  0.56  1.69  3.95  8.47  9.04  4.52  4.52
65 1C5A-   0.71  0.00  0.00  0.00
13.85  9.23  6.15 10.77  1.54  3.08  0.00  6.15 15.38  3.08  3.08  3.08  1.54  4.62  6.15  0.00  1.54  3.08  0.00  7.69
162 1CPCA   0.78  0.00  0.00  0.00
14.81  1.23  4.94  3.70  3.70  8.02  0.62  4.94  4.94  8.02  1.23  4.32  3.70  3.09  3.70  9.88  7.41  4.94  0.62  6.17
172 1CPCL   0.72  0.00  0.00  0.00
18.60  1.74  7.56  2.91  2.33  8.14  0.00  5.23  3.49  8.72  3.49  4.07  1.74  2.91  5.81  7.56  4.65  7.56  0.00  3.49
136 1ECO-   0.70  0.00  0.00  0.00
12.50  0.00  6.62  3.68 10.29  8.09  2.94  6.62  7.35  4.41  2.94  3.68  3.68  2.94  2.21  6.62  6.62  6.62  0.74  1.47
154 1FCS-   0.71  0.00  0.00  0.00
11.04  0.00  3.90  9.09  3.90  7.14  7.14  5.84 12.34 11.69  1.95  1.30  2.60  3.25  3.25  3.90  2.60  5.84  1.30  1.95
171 1FHA-   0.73  0.00  0.00  0.00
 7.02  1.75  7.60  8.77  3.51  4.09  5.85  3.51  7.02 12.87  2.34  6.43  1.75  6.43  4.09  5.26  2.34  3.51  0.58  5.26
78 1FIAB   0.67  0.00  0.00  0.00
 6.41  0.00  5.13  3.85  1.28  6.41  0.00  1.28  8.97 15.38  6.41  7.69  2.56  7.69  6.41  1.28  5.13  8.97  0.00  5.13
147 1HBG-   0.73  0.00  0.00  0.00
20.41  0.68  4.08  2.04  2.72 13.61  3.40  5.44  8.16  7.48  3.40  2.04  2.04  4.76  2.04  6.80  0.68  6.80  1.36  2.04
57 1HDDC   0.63  0.00  0.00  0.00
 7.02  0.00  0.00 10.53  5.26  1.75  0.00  5.26 10.53 10.53  0.00  7.02  1.75  8.77 15.79  8.77  3.51  0.00  1.75  1.75
138 1HIGA   0.98  0.00  0.00  0.00
 5.07  0.00  7.25  6.52  7.25  3.62  1.45  5.07 14.49  7.25  2.90  7.25  1.45  5.80  4.35  7.25  3.62  5.80  0.72  2.90
139 1LE4-   0.83  0.00  0.00  0.00
 9.35  0.00  5.04 12.23  0.72  3.60  0.72  0.00  5.04 17.27  2.88  0.00  0.72  9.35 12.23  5.04  4.32  6.47  2.16  2.88
149 1LIG-   0.80  0.00  0.00  0.00
14.09  0.67  4.03  4.70  4.03  4.70  1.34  2.01  2.01 12.08  6.04  7.38  2.01 10.74  4.70  6.71  6.04  2.01  0.67  4.03
41 1LTSC   0.73  0.00  0.00  0.00
 0.00  2.44  7.32  9.76  2.44  2.44  0.00  9.76  4.88  4.88  0.00  7.32  0.00  9.76  7.32  9.76  7.32  4.88  0.00  9.76
153 1MBC-   0.74  0.00  0.00  0.00
11.11  0.00  4.58  9.15  3.92  7.19  7.84  5.88 12.42 11.76  1.31  0.65  2.61  3.27  2.61  3.92  3.27  5.23  1.31  1.96
153 1MBS-   0.73  0.00  0.00  0.00
 9.15  0.00  5.23  9.15  4.58  7.84  8.50  5.23 12.42 12.42  1.31  1.96  2.61  1.96  3.27  4.58  3.27  3.92  1.31  1.31
63 1RPRA   0.78  0.00  0.00  0.00
 9.52  3.17 11.11 11.11  3.17  3.17  3.17  3.17  4.76 15.87  3.17  4.76  0.00  4.76  6.35  4.76  6.35  0.00  0.00  1.59
104 1TROA   0.78  0.00  0.00  0.00
 9.62  0.00  3.85 12.50  0.96  4.81  1.92  2.88  2.88 18.27  2.88  4.81  3.85  3.85  8.65  5.77  3.85  4.81  1.92  1.92
70 1UTG-   0.71  0.00  0.00  0.00
 2.86  2.86  5.71  8.57  2.86  4.29  1.43  5.71 10.00 11.43  7.14  2.86  7.14  2.86  2.86  7.14  8.57  4.29  0.00  1.43
106 256BA   0.76  0.00  0.00  0.00
16.04  0.00 11.32  7.55  1.89  2.83  1.89  2.83 12.26  9.43  2.83  5.66  3.77  5.66  3.77  1.89  4.72  3.77  0.00  1.89
127 2CCYA   0.71  0.00  0.00  0.00
22.05  1.57  3.94  7.87  3.15  7.87  0.79  1.57 11.81  9.45  2.36  1.57  5.51  5.51  1.57  3.94  4.72  2.36  2.36  0.00
153 2LH1-   0.70  0.00  0.00  0.00
13.73  0.00  3.92  9.15  4.58  4.58  3.27  5.88  9.15  9.15  0.65  3.92  3.27  2.61  0.65  5.88  5.23 11.11  1.96  1.31
149 2LHB-   0.67  0.00  0.00  0.00
14.09  0.67  7.38  5.37  5.37  4.03  1.34  5.37  8.72  6.71  3.36  1.34  4.03  1.34  3.36  8.72  6.71  8.05  1.34  2.68
141 2MHBA   0.67  0.00  0.00  0.00
11.35  0.71  6.38  2.13  4.96  7.09  7.09  0.00  7.80 14.89  0.71  2.84  4.26  0.71  2.13  9.22  6.38  8.51  0.71  2.13
146 2MHBB   0.67  0.00  0.00  0.00
 9.59  0.68  4.79  6.85  5.48  9.59  6.16  0.00  7.53 13.01  0.68  4.79  3.42  2.74  2.74  4.11  2.05 12.33  1.37  2.05
31 2ZTAA   0.94  0.00  0.00  0.00
 3.23  0.00  3.23 16.13  0.00  3.23  3.23  0.00 16.13 19.35  3.23  6.45  0.00  3.23  6.45  3.23  0.00  9.68  0.00  3.23
146 4MBA-   0.73  0.00  0.00  0.00
19.86  0.00  5.48  3.42 10.27  7.53  0.68  2.74  7.53  7.53  2.05  6.16  4.11  1.37  2.74  8.90  1.37  6.85  1.37  0.00
153 4MBN-   0.75  0.00  0.00  0.00
11.11  0.00  4.58  9.15  3.92  7.19  7.84  5.88 12.42 11.76  1.31  0.65  2.61  3.27  2.61  3.92  3.27  5.23  1.31  1.96
```

30 β proteins

```
107 1ACX-   0.00  0.44  0.00  1.00
18.69  3.74  4.67  0.93  4.67 12.15  0.93  0.93  0.93  3.74  0.00  3.74  5.61  3.74  0.93 14.02  9.35  8.41  0.00  2.80
214 1AYH-   0.02  0.48  0.00  1.00
 5.61  0.93  5.61  3.74  6.07 12.15  1.87  3.74  6.54  4.67  1.40  8.88  3.27  1.87  0.93  7.01  8.88  5.14  3.74  7.94
114 1CD8-   0.00  0.46  0.00  1.00
 6.14  2.63  3.51  4.39  8.77  6.14  0.88  0.88  3.51 13.16  0.88  4.39  6.14  4.39  6.14 11.40  6.14  5.26  1.75  3.51
60 1CDTA   0.00  0.45  0.00  1.00
```

5.00 13.33 1.67 1.67 0.00 3.33 0.00 5.00 16.67 8.33 5.00 8.33 6.67 0.00 3.33 5.00 3.33 8.33 0.00 5.00
177 1CID- 0.00 0.56 0.19 0.81
4.52 1.13 1.69 9.60 2.82 5.08 0.00 3.95 7.34 11.30 2.26 3.95 5.08 9.60 2.82 11.30 6.78 7.91 1.69 1.13
30 1DFNA 0.00 0.53 0.00 1.00
10.00 20.00 3.33 3.33 3.33 10.00 0.00 10.00 0.00 3.33 0.00 0.00 3.33 3.33 13.33 0.00 3.33 0.00 3.33 10.00
217 1HILA 0.04 0.51 0.08 0.92
4.15 1.84 5.53 4.15 3.69 6.45 0.92 3.23 6.45 6.45 1.38 5.07 5.07 5.07 2.76 14.29 11.06 5.99 1.84 4.61
99 1HIVA 0.04 0.56 0.25 0.75
3.03 2.02 4.04 4.04 2.02 13.13 1.01 12.12 6.06 12.12 2.02 3.03 6.06 6.06 4.04 1.01 8.08 7.07 2.02 1.01
31 1HLEB 0.00 0.61 0.21 0.79
6.45 0.00 3.23 6.45 19.35 3.23 6.45 9.68 0.00 6.45 0.00 12.90 9.68 0.00 6.45 9.68 0.00 0.00 0.00 0.00
217 1MAMH 0.00 0.48 0.08 0.92
6.45 2.30 3.23 3.23 2.76 9.22 0.92 1.38 3.69 8.29 1.38 1.84 6.45 3.23 3.23 14.29 11.98 8.29 2.30 5.53
44 1MONA 0.00 0.71 0.00 1.00
4.55 0.00 6.82 6.82 4.55 6.82 0.00 4.55 9.09 9.09 0.00 2.27 9.09 2.27 11.36 4.55 2.27 4.55 0.00 11.36
340 1OMF- 0.04 0.58 0.02 0.98
8.53 0.00 7.94 4.12 5.59 14.12 0.29 3.53 5.29 5.88 0.88 8.82 1.18 3.82 3.24 4.71 6.18 6.76 0.59 8.53
126 1PHY- 0.00 0.76 0.00 1.00
7.14 0.79 9.52 6.35 7.14 10.32 1.59 3.97 8.73 5.56 3.97 4.76 3.17 3.17 1.59 3.97 5.56 7.94 0.79 3.97
107 1REIA 0.00 0.48 0.18 0.82
5.61 1.87 4.67 1.87 2.80 7.48 0.00 7.48 3.74 7.48 0.93 1.87 5.61 12.15 1.87 13.08 10.28 2.80 0.93 7.48
89 1TEN- 0.00 0.54 0.00 1.00
4.49 0.00 11.24 8.99 2.25 5.62 0.00 8.99 5.62 7.87 1.12 3.37 5.62 2.25 3.37 6.74 13.48 4.49 1.12 3.37
103 1TLK- 0.00 0.59 0.26 0.74
6.80 4.85 10.68 11.65 3.88 4.85 1.94 2.91 7.77 4.85 2.91 2.91 4.85 0.97 1.94 4.85 6.80 9.71 0.97 3.88
99 1VAAB 0.00 0.51 0.00 1.00
3.03 2.02 6.06 6.06 4.04 2.02 5.05 7.07 9.09 3.03 5.05 4.04 9.09 5.05 3.03 6.06 8.08 5.05 2.02 5.05
198 2ALP- 0.04 0.53 0.12 0.88
12.12 3.03 1.01 2.02 3.03 16.16 0.51 4.04 1.01 5.05 1.01 6.57 2.02 4.55 6.06 10.10 9.09 9.60 1.01 2.02
121 2AVIA 0.00 0.51 0.00 1.00
3.31 1.65 4.13 4.96 5.79 9.09 0.83 5.79 6.61 5.79 1.65 8.26 1.65 1.65 4.96 7.44 16.53 5.79 3.31 0.83
175 2BPA2 0.00 0.49 0.29 0.71
6.86 2.86 5.71 1.14 6.86 5.71 1.71 4.00 4.00 8.00 1.71 6.29 5.71 3.43 2.86 9.14 9.14 10.86 0.57 3.43
203 2HHRC 0.00 0.55 0.00 1.00
2.46 2.96 4.43 9.36 3.45 4.43 1.97 4.93 6.40 6.40 1.48 5.42 5.91 3.94 4.43 7.39 7.88 7.88 3.94 4.93
145 2ILA- 0.00 0.47 0.00 1.00
8.97 0.69 4.83 5.52 5.52 3.45 2.07 8.97 6.90 9.66 2.07 6.21 4.14 4.83 2.07 4.83 8.28 4.83 1.38 4.83
181 2LALA 0.00 0.46 0.00 1.00
7.18 0.00 6.08 2.76 7.18 6.63 1.10 5.52 6.08 4.42 0.00 8.29 3.87 3.31 2.21 7.73 13.26 8.84 1.66 3.87
151 2SNV- 0.00 0.49 0.12 0.88
5.96 0.00 5.30 6.62 3.97 12.58 3.97 3.97 6.62 5.30 3.31 3.31 4.64 1.32 5.30 6.62 7.95 8.61 1.99 2.65
178 3CD4A 0.00 0.46 0.12 0.88
2.81 2.25 5.06 5.06 2.81 7.30 1.12 5.62 10.11 11.80 0.00 5.06 3.37 8.43 2.81 10.11 7.30 6.74 1.69 0.56
174 4GCR- 0.03 0.46 0.00 1.00
1.15 4.02 7.47 5.17 5.17 8.05 2.87 3.45 1.15 7.47 4.02 3.45 4.60 5.75 11.49 7.47 2.87 3.45 2.30 8.62
37 7APIB 0.00 0.49 0.17 0.83
0.00 2.70 0.00 5.41 10.81 2.70 0.00 5.41 13.51 5.41 5.41 8.11 13.51 5.41 0.00 5.41 5.41 10.81 0.00 0.00
146 8I1B- 0.00 0.49 0.00 1.00
2.74 0.68 4.79 7.53 5.48 4.79 2.05 4.79 10.27 9.59 2.74 6.16 5.48 7.53 2.05 8.90 2.74 7.53 0.68 3.42
206 8FABA 0.05 0.49 0.09 0.91
9.71 1.94 2.91 4.37 1.94 5.83 0.97 3.40 5.34 5.34 0.49 2.91 8.25 6.80 2.43 13.11 10.19 7.28 1.94 4.85
214 8FABB 0.02 0.48 0.08 0.92
6.54 1.87 3.74 1.87 3.74 8.88 1.40 2.34 5.14 7.48 1.40 3.74 5.61 3.27 4.21 12.15 7.94 11.21 2.34 5.14

30 α+β proteins

151 1AAK- 0.30 0.18 0.00 1.00
7.28 1.32 7.95 3.31 3.97 3.97 0.66 5.96 3.31 6.62 3.31 6.62 9.27 5.30 7.28 9.27 3.97 4.64 2.65 3.31
68 1CTF- 0.51 0.26 0.00 1.00
22.06 0.00 5.88 13.24 1.47 8.82 0.00 2.94 14.71 10.29 0.00 1.47 1.47 0.00 1.47 2.94 1.47 11.76 0.00 0.00
250 1DNKA 0.26 0.27 0.19 0.81
8.80 0.80 6.80 3.60 4.40 2.80 2.40 4.80 3.60 9.20 1.60 4.40 3.60 3.60 4.80 11.60 6.00 10.00 1.20 6.00
243 1EAF- 0.32 0.23 0.18 0.82
13.58 0.82 4.53 4.53 3.70 6.17 2.06 6.17 6.58 13.17 2.06 2.47 6.58 3.70 4.53 5.35 4.12 7.41 0.82 1.65
270 1HSBA 0.24 0.39 0.00 1.00
8.52 1.48 7.41 7.41 2.59 7.41 3.70 1.85 3.33 5.56 2.22 1.48 4.07 6.67 8.52 5.19 8.15 5.93 3.33 5.19
226 1LTSA 0.31 0.18 0.00 1.00
5.31 0.88 6.19 5.75 2.65 8.41 3.54 6.19 1.33 6.19 1.33 5.75 5.31 4.87 9.29 7.08 3.98 4.87 1.33 9.73
103 1LTSD 0.22 0.37 0.00 1.00
4.85 1.94 3.88 7.77 1.94 2.91 0.97 11.65 9.71 4.85 3.88 5.83 2.91 4.85 3.88 7.77 11.65 3.88 0.97 3.88
85 1NRCA 0.26 0.22 0.00 1.00
5.88 1.18 4.71 3.53 7.06 3.53 1.18 9.41 10.59 8.24 3.53 5.88 3.53 4.71 5.88 8.24 4.71 3.53 0.00 4.71
159 1OVB- 0.19 0.18 0.18 0.82
9.43 5.03 5.66 6.92 3.14 9.43 1.89 4.40 8.81 5.66 0.63 2.52 3.14 3.77 4.40 6.92 5.66 6.92 1.89 3.77
134 1POC- 0.27 0.17 0.22 0.78
2.99 7.46 8.21 3.73 3.73 8.21 5.22 2.99 8.96 6.72 2.24 3.73 3.73 0.75 4.48 7.46 8.21 3.73 1.49 5.97
212 1PFN- 0.23 0.18 0.18 0.82
6.60 3.30 2.83 4.72 1.89 13.21 0.94 5.66 4.72 5.19 0.00 6.13 4.72 4.72 5.66 6.13 3.77 8.49 2.36 8.96
124 1PRF- 0.24 0.26 0.00 1.00
13.71 0.00 5.65 1.61 2.42 12.10 0.81 5.65 4.84 7.26 0.00 4.84 2.42 4.84 2.42 6.45 9.68 9.68 1.61 4.03
124 1RND- 0.18 0.33 0.12 0.88
13.71 6.45 4.03 4.03 2.42 2.42 3.23 2.42 4.84 1.61 3.23 8.06 3.23 5.65 3.23 11.29 8.06 7.26 0.00 4.84
135 1SNC- 0.24 0.30 0.00 1.00
8.89 0.00 4.44 8.15 2.22 6.67 2.96 3.70 15.56 8.15 2.96 3.70 4.44 3.70 3.70 2.22 5.93 6.67 0.74 5.19
112 1TFG- 0.19 0.38 0.07 0.93
7.14 8.04 5.36 3.57 2.68 3.57 1.79 6.25 6.25 9.82 0.89 6.25 6.25 3.57 3.57 8.93 3.57 3.57 2.68 6.25
56 1TGSI 0.16 0.20 0.00 1.00
1.79 10.71 1.79 7.14 0.00 7.14 0.00 5.36 7.14 3.57 0.00 5.36 8.93 5.36 3.57 10.71 10.71 7.14 0.00 3.57
372 2ACHA 0.28 0.34 0.06 0.94
7.80 0.27 5.91 6.45 6.45 3.49 1.88 5.38 6.99 13.17 3.23 4.30 3.23 4.03 3.76 7.53 7.26 5.65 0.81 2.42

```
218 2ACT-   0.26  0.18  0.17  0.82
 8.26  3.21  7.34  5.05  2.29 12.84  0.46  7.80  2.75  3.67  0.92  4.59  3.21  4.59  2.29  5.50  8.26  7.80  2.75  6.42
426 2BPA1   0.20  0.27  0.00  1.00
 6.10  0.70  6.57  3.52  5.16  6.57  3.05  5.40  3.52  8.22  2.58  4.69  6.57  5.16  5.87  6.57  7.98  5.16  1.41  5.16
141 2SNS-   0.18  0.20  0.11  0.89
 9.22  0.00  3.55  7.80  2.13  6.38  2.84  3.55 16.31  7.80  2.84  4.26  4.26  3.55  3.55  2.84  7.09  6.38  0.71  4.96
107 2SSI-   0.16  0.24  0.00  1.00
14.95  3.74  4.67  4.67  2.80 10.28  1.87  0.00  1.87  7.48  2.80  2.80  6.54  0.93  3.74  7.48  7.48 12.15  0.93  2.80
 68 3IL8-   0.22  0.25  0.00  1.00
 2.94  5.88  2.94 10.29  4.41  2.94  2.94  7.35 11.76  8.82  0.00  4.41  5.88  2.94  7.35  5.88  2.94  7.35  1.47  1.47
123 3RUBS   0.22  0.22  0.00  1.00
 4.88  2.44  3.25  9.76  4.07  5.69  0.81  4.88  7.32  8.13  2.44  4.07  6.50  5.69  3.25  4.07  4.07  6.50  4.07  8.13
 50 3SGBI   0.20  0.22  0.00  1.00
 4.00 12.00  4.00  6.00  4.00  8.00  2.00  0.00  8.00  8.00  0.00 10.00  6.00  0.00  2.00  8.00  8.00  4.00  0.00  6.00
107 3SICI   0.15  0.36  0.00  1.00
14.95  3.74  4.67  4.67  2.80 10.28  1.87  0.00  2.80  7.48  1.87  2.80  6.54  0.93  3.74  7.48  7.48 12.15  0.93  2.80
256 4BLMA   0.36  0.19  0.00  1.00
10.16  0.00  9.38  7.42  2.73  5.47  0.39  5.47  8.20 10.55  1.17  4.69  4.30  2.73  5.86  4.30  7.81  5.86  1.17  2.34
316 4TMS-   0.36  0.21  0.00  1.00
 6.33  0.63  9.18  4.75  6.01  5.70  6.01  4.75  6.33 11.08  1.90  2.53  6.01  4.11  3.80  4.11  5.06  5.06  2.22  4.43
498 8CATA   0.28  0.15  0.00  1.00
 6.83  0.80  7.63  4.62  6.22  6.83  4.22  3.61  5.02  7.03  2.01  5.82  7.43  4.22  6.22  5.02  4.42  6.83  1.20  4.02
104 9RNT-   0.15  0.27  0.00  1.00
 6.73  3.85  5.77  5.77  3.85 11.54  2.88  1.92  1.92  2.88  0.00  8.65  3.85  1.92  0.96 14.42  5.77  7.69  0.96  8.65
124 9RSAA   0.18  0.33  0.12  0.88
 9.68  6.45  4.03  4.03  2.42  2.42  3.23  2.42  8.06  1.61  3.23  8.06  3.23  5.65  3.23 12.10  8.06  7.26  0.00  4.84
```

30 α/β proteins

```
 87 1ABA-   0.34  0.18  0.63  0.38
 3.45  2.30  8.05  5.75  8.05  9.20  2.30  6.90 10.34  8.05  3.45  3.45  5.75  4.60  3.45  2.30  4.60  4.60  0.00  3.45
 66 1CIS-   0.17  0.15  1.00  0.00
10.61  0.00  7.58 10.61  0.00  3.03  0.00  7.58 10.61  7.58  1.52  3.03  4.55  6.06  6.06  1.52  1.52 13.64  1.52  3.03
 63 1CSEI   0.17  0.30  0.68  0.32
 1.59  0.00  3.17  4.76  6.35  6.35  4.76  0.00  3.17  6.35  0.00  7.94  9.52  3.17  6.35  3.17  6.35 17.46  0.00  9.52
307 1CTC-   0.36  0.16  0.60  0.40
 6.84  0.65  3.91  4.56  5.21  7.49  2.61  6.84  4.89  7.49  0.98  5.54  3.26  3.58  3.58 10.42  8.47  5.21  2.28  6.19
236 1DHR-   0.37  0.24  0.88  0.13
11.44  1.69  4.66  4.66  2.97 10.17  1.69  3.81  6.36  8.90  2.97  3.39  3.39  3.39  3.39  8.05  7.20  7.63  2.97  1.27
271 1DRI-   0.45  0.23  0.95  0.05
13.65  0.00  7.75  4.06  2.58  8.86  1.11  4.80  8.86  9.23  1.48  5.90  3.32  6.27  2.21  3.32  4.80 10.70  0.00  1.11
177 1ETU-   0.44  0.20  0.83  0.17
 8.47  1.13  7.91  7.34  2.26  7.91  4.52  6.78  3.95  9.04  2.82  2.26  5.08  2.26  3.39  2.26  9.04  9.04  0.56  3.95
147 1FX1-   0.29  0.22  0.88  0.12
11.56  2.72 10.88  8.16  4.08 12.24  0.68  6.12  2.72  8.16  0.00  2.04  2.04  2.72  4.76  5.44  4.76  6.12  1.36  3.40
823 1GPB-   0.45  0.15  0.60  0.40
 7.65  1.09  5.83  7.78  4.62  5.71  2.67  6.08  5.59  9.23  2.55  5.59  4.13  3.40  7.29  3.16  4.37  7.41  1.46  4.37
169 1OFV-   0.27  0.22  0.73  0.27
 6.51  0.59  9.47  6.51  4.73 10.65  0.00  7.69  4.73  7.10  0.59  6.51  1.78  7.10  1.18  7.10  4.73  5.92  2.37  4.73
120 1PAZ-   0.14  0.37  0.60  0.40
10.00  0.83  4.17  8.33  2.50  6.67  2.50  9.17 10.00  5.00  4.17  5.83  6.67  2.50  0.83  3.33  4.17 10.00  0.00  3.33
320 1PFKA   0.43  0.18  0.83  0.17
 8.44  1.87  7.19  6.56  3.13 11.87  2.19  8.75  4.37  7.81  3.75  2.50  2.50  1.56  6.87  4.37  4.69  7.81  0.31  3.44
469 1PGD    0.53  0.09  0.74  0.26
 9.17  1.92  6.61  4.90  4.90 10.45  2.35  7.46  7.46  9.81  2.77  3.62  3.20  3.62  4.26  4.69  3.62  4.90  1.71  2.56
171 1Q21    0.41  0.27  0.68  0.32
 6.43  1.75  8.19  7.60  2.92  6.43  1.75  6.43  5.85  7.60  2.34  2.34  1.75  6.43  7.02  4.68  6.43  8.77  0.00  5.26
275 1S01-   0.30  0.17  0.66  0.34
13.82  0.36  4.00  1.82  1.45 11.64  2.18  4.73  4.36  5.45  1.45  5.82  5.09  3.27  0.73 13.82  4.73 10.91  1.09  3.27
309 1SBP-   0.45  0.17  0.60  0.40
10.36  0.00  8.41  6.15  3.88  6.80  1.94  5.83  9.06  6.80  0.00  5.83  3.88  3.24  3.24  5.50  5.18  7.77  2.27  3.88
275 1SBT-   0.30  0.18  0.60  0.40
13.45  0.00  4.00  1.45  1.09 12.00  2.18  4.73  4.00  5.45  1.82  6.18  5.09  4.00  0.73 13.45  4.73 10.91  1.09  3.64
247 1TIMA   0.43  0.17  1.00  0.00
11.34  1.62  5.26  6.88  3.24 10.93  3.24  6.88  8.91  6.88  0.81  2.43  2.83  3.64  3.24  4.86  4.05  9.31  2.02  1.62
729 1TMD-   0.29  0.17  0.77  0.23
 9.05  1.37  7.27  7.96  2.74  9.05  2.88  5.49  5.08  6.04  2.19  2.74  5.49  3.02  5.49  5.90  5.62  5.90  2.19  4.53
254 1TREA   0.44  0.15  1.00  0.00
17.72  1.18  3.54  8.27  2.36  8.66  3.15  7.48  6.30  6.69  2.76  3.94  2.76  3.94  3.15  4.33  3.54  7.48  0.79  1.97
289 1ULA-   0.26  0.19  0.65  0.35
 7.61  1.38  4.50  6.23  5.19  9.34  2.77  4.15  4.15  9.00  4.15  3.81  5.19  4.50  5.88  4.84  5.19  7.96  1.04  3.11
385 1WSYB   0.38  0.17  0.69  0.31
11.17  1.30  4.68  7.01  3.38 10.91  3.64  5.97  4.94  9.35  3.64  2.60  4.42  4.42  4.68  4.94  4.94  4.94  0.26  2.86
310 2HAD-   0.34  0.14  0.60  0.40
 9.68  1.29  8.39  5.81  7.42  5.81  1.61  4.52  3.87  9.03  3.55  3.23  7.42  4.19  4.84  4.19  5.16  4.52  1.94  3.55
344 2LIV-   0.39  0.19  0.78  0.22
12.79  0.58  7.56  4.94  2.91  9.88  1.16  5.81  8.43  6.69  1.45  4.07  4.36  5.81  2.03  4.07  5.23  7.56  0.87  3.78
305 3GBP-   0.42  0.19  0.95  0.05
12.79  0.00  9.18  4.59  1.97  7.21  0.98  4.92  9.84  7.54  2.30  6.89  3.28  4.59  1.97  4.59  4.26  8.85  1.64  2.62
138 4FXN-   0.34  0.21  1.00  0.00
 4.35  2.17  6.52 13.77  3.62 10.14  0.00 10.87  7.25  5.80  3.62  5.80  2.17  1.45  1.45  5.80  3.62  7.25  2.17  2.17
307 5CPA-   0.36  0.16  0.60  0.40
 6.84  0.65  3.91  4.56  5.21  7.49  2.61  6.84  4.89  7.49  0.98  5.54  3.26  3.58  3.58 10.42  8.47  5.21  2.28  6.19
166 5P21-   0.34  0.27  0.77  0.23
 6.63  1.81  8.43  7.83  3.01  6.63  1.81  6.63  4.82  6.63  2.41  2.41  1.81  6.63  6.63  4.82  6.63  9.04  0.00  5.42
305 8ABP-   0.43  0.21  0.94  0.06
10.16  0.33  6.89  6.56  3.93  9.51  0.98  5.25  9.84  7.21  2.95  3.28  4.92  3.61  2.62  4.59  5.25  8.52  1.64  1.97
310 8ATCA   0.32  0.14  1.00  0.00
10.97  0.32  6.77  4.52  3.87  4.84  3.55  4.84  4.84 12.26  2.58  4.84  3.87  4.52  4.84  6.45  5.81  7.10  0.65  2.58
```

9 Irregular proteins

```
 55 1AAF-    0.00  0.04  0.00  1.00
  5.45 10.91  1.82  5.45  3.64 10.91  3.64  5.45 14.55  0.00  3.64  9.09  1.82  7.27 12.73  0.00  1.82  0.00  1.82  0.00
 38 1ERP-    0.00  0.00  0.00  0.00
  2.63 15.79  5.26 10.53  2.63  7.89  2.63  0.00  2.63 10.53  2.63  7.89 10.53  7.89  0.00  5.26  0.00  2.63  2.63  0.00
152 2ATCB    0.04  0.03  0.00  1.00
  6.58  2.63  6.58  8.55  3.29  3.95  2.63  7.89  6.58  9.87  1.32  6.58  4.61  1.97  5.26  7.24  4.61  7.89  0.00  1.97
 43 2BDS-    0.00  0.05  1.00  0.00
  4.65 13.95  2.33  0.00  2.33 16.28  2.33  4.65  4.65  4.65  0.00  4.65 11.63  0.00  4.65  4.65  4.65  0.00  4.65  9.30
 36 2BPA3    0.00  0.00  0.00  0.00
  5.56  0.00  0.00  0.00  2.78 22.22  0.00  0.00 16.67  5.56  0.00  0.00  8.33  8.33 16.67  2.78  2.78  2.78  2.78  2.78
 36 2CBH-    0.00  0.08  1.00  0.00
  2.78 11.11  0.00  0.00  0.00 16.67  2.78  2.78  0.00  5.56  0.00  2.78  5.56 11.11  0.00 11.11 11.11  5.56  0.00 11.11
 58 2MEV4    0.07  0.00  0.00  0.00
  8.62  0.00  3.45  3.45  5.17  8.62  0.00  5.17  1.72 10.34  1.72 15.52  5.17  5.17  0.00 13.79  3.45  3.45  0.00  5.17
 50 4TGP-    0.00  0.08  1.00  0.00
  8.00 12.00  8.00  4.00  8.00  6.00 10.00  0.00  2.00  6.00  0.00  2.00  4.00  4.00  4.00  6.00  4.00 10.00  0.00  2.00
170 9WGAA    0.09  0.09  0.00 1.00
  5.88 18.82  2.94  2.35  1.76 23.53  1.18  1.18  4.12  2.35  1.18  5.88  3.53  5.88  2.35  8.24  2.35  0.59  1.76  4.12
```

Appendix B

Here, a simple example is presented that illustrates the advantage of using Mahalanobis distance as a scale to measure the similarity among a set of points. Suppose in a 2-D space there are ten points whose coordinates in the Cartesian system (x_1, x_2) are

$$\begin{pmatrix} x_1 = & 0.907 & 1.596 & -0.808 & -0.529 & 1.116 & -0.585 & -1.236 & 0.663 & -0.278 & -0.848 \\ x_2 = & 0.803 & 1.467 & -0.978 & -0.748 & 0.868 & -0.463 & -0.474 & 1.121 & -0.267 & -1.329 \end{pmatrix} \tag{B1}$$

The mean of the ten points is (see Equations 3 and 4)

$$\overline{\mathbf{X}} = \begin{bmatrix} \overline{x_1} \\ \overline{x_2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{B2}$$

The covariance **S** is (see Equations 21 and 22)

$$\mathbf{S} = \begin{bmatrix} s_{1,1} & s_{1,2} \\ s_{2,1} & s_{2,2} \end{bmatrix} \approx \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix} . \tag{B3}$$

from which it follows that the inverse matrix of **S** is

$$\mathbf{S}^{-1} \approx \frac{1}{0.19} \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix} . \tag{B4}$$

Given $\mathbf{S}^{-1}$, the normal distribution density is generally given by

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{2\pi\sqrt{|\mathbf{S}|}} \exp\left\{ -\frac{1}{2\times 0.19} \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right\} \\ &= \frac{1}{2\pi\sqrt{0.19}} \exp\left\{ -\frac{1}{0.38}(x_1^2 - 1.8x_1x_2 + x_2^2) \right\} \end{aligned} \tag{B5}$$

Now suppose we have two points **A** and **B** whose coordinates in the 2-D space are given by

$$\mathbf{A} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{B6}$$

and whose Euclidean distances to the central point $\overline{\mathbf{X}} = (0, 0)$ are the same (i.e., both equal to 2). However, according to Equation B5, the distribution densities at these two points are, respectively,

$$f(\mathbf{A}) = f(1, 1) = 0.216, \quad f(\mathbf{B}) = f(1, -1) = 0.0000166 \tag{B7}$$

indicating that from the viewpoint of classification, point **A** should be much closer to the central point (0, 0) than point **B** because the distribution density $f(\overline{\mathbf{X}}) = f(0, 0)$ at the central point is the maximum. Such

a fact cannot be reflected by their Euclidean distances to the central points, but can be well elucidated by the Mahalanobis as follows. The Mahalanobis distance between point **A** and $\overline{\mathbf{X}}$ is (see Equation 20)

$$\begin{aligned} D^2(\mathbf{A}, \overline{\mathbf{X}}) &= \left[\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right]^T \mathbf{S}^{-1} \left[\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right] \\ &= \begin{bmatrix} 1 & 1 \end{bmatrix} \frac{1}{0.19} \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{0.2}{0.19} \end{aligned} \tag{B8}$$

and

$$\begin{aligned} D^2(\mathbf{B}, \overline{\mathbf{X}}) &= \left[\begin{pmatrix} 1 \\ -1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right]^T \mathbf{S}^{-1} \left[\begin{pmatrix} 1 \\ -1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right] \\ &= \begin{bmatrix} 1 & -1 \end{bmatrix} \frac{1}{0.19} \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \frac{3.8}{0.19} \end{aligned} \tag{B9}$$

Therefore, the ratio of these two Mahalanobis distances is

$$\frac{D(\mathbf{B}, \overline{\mathbf{X}})}{D(\mathbf{A}, \overline{\mathbf{X}})} = \sqrt{\frac{3.8}{0.2}} \approx 4.4 \tag{B10}$$

indicating that the Mahalanobis distance from **B** to $\overline{\mathbf{X}}$ is more than four times that from **A** to $\overline{\mathbf{X}}$ although their corresponding Euclidean distances are the same. According to the Mahalanobis distance, point **A** is much closer to the central point $\overline{\mathbf{X}}$ than is point **B**, and hence the distribution density at point **A** is much higher than that at **B**.

Therefore, in classifying a set of statistical data that generally belong to a normal distribution, it would better reflect the essential aspect of the problem to use the Mahalanobis distance as a scale to measure the similarity.

Appendix C

The data of (1) norms, (2) the elements of covariance matrices, and (3) eigenvalues and eigenvectors derived for α, β, α+β, and α/β structural classes from the training database of Appendix A.

(1) Norms of the four structural classs. The 20 components of each of the four norms in the 20-D space are normalized to 100, and they are printed according to the alphabetical order of the single amino acid letter code.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
α class	11.06	0.97	5.55	7.45	3.98	6.03	2.86	4.02	8.59	11.27	2.55	3.92	2.73	4.32	4.58	5.63	4.53	5.97	1.04	2.94
β class	6.00	2.74	4.96	4.97	4.98	7.59	1.41	5.05	6.12	7.11	1.82	5.13	5.49	4.24	4.04	8.08	7.67	6.70	1.53	4.34
α+β class	8.45	3.10	5.47	5.79	3.39	6.84	2.19	4.60	6.84	7.27	1.76	4.87	4.91	3.75	4.41	7.10	6.38	6.84	1.32	4.72
α/β class	9.48	1.03	6.49	6.33	3.65	8.60	2.13	6.11	6.32	7.66	2.20	4.31	4.09	4.04	3.86	5.55	5.24	8.08	1.22	3.63

(2) The covariance matrices of the four structural classes.

$S_\alpha = [s_{i,j}(\alpha)]$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	783.0	5.9	4.9	-257.7	71.1	200.6	-29.4	-40.1	-55.0	-212.2	-4.8	-106.9	78.4	-99.3	-182.5	-0.3	-61.5	-7.6	35.0	-121.7
2	5.9	95.5	30.6	28.2	-33.4	-34.0	-34.6	24.1	26.6	-84.4	11.3	0.3	-7.6	10.5	8.7	-32.6	-2.7	-47.6	-18.9	53.9
3	4.9	30.6	152.3	-15.1	-4.1	-40.5	-7.6	9.5	-19.0	-65.1	14.0	13.9	-13.0	-3.6	-36.9	-4.4	37.7	-35.5	-37.2	19.2
4	-257.7	28.2	-15.1	362.6	-124.5	-172.3	26.6	-13.8	146.5	127.6	-28.5	0.0	-64.4	74.5	164.2	-82.2	-70.9	-93.4	-14.2	6.6
5	71.1	-33.4	-4.1	-124.5	165.5	53.3	23.9	30.8	-15.9	-110.2	-29.1	6.4	34.0	-68.7	-82.4	68.3	17.9	32.6	17.0	-52.5
6	200.6	-34.0	-40.5	-172.3	53.3	209.1	70.5	-9.5	-53.1	-6.1	-8.1	-73.3	40.4	-100.4	-144.4	0.7	-18.8	105.0	25.6	-44.6
7	-29.4	-34.6	-7.6	26.6	23.9	70.5	230.9	-31.0	93.8	71.5	-57.0	-64.9	13.5	-83.9	-111.2	-42.9	-60.9	51.9	3.2	-62.4
8	-40.1	24.1	9.5	-13.8	30.8	-9.5	-31.0	185.0	6.5	-167.9	-12.0	-25.6	-16.4	-6.2	-16.6	48.6	46.4	-65.7	0.6	53.2
9	-55.0	26.6	-19.0	146.5	-15.9	-53.1	93.8	6.5	426.3	-99.7	-24.3	-26.2	3.5	-40.0	-40.3	-121.0	-129.9	-12.5	-22.8	-43.3
10	-212.2	-84.4	-65.1	127.6	-110.2	-6.1	71.5	-167.9	-99.7	523.5	38.3	-17.6	-23.1	-17.8	79.0	-87.6	0.3	64.3	37.5	-50.5
11	-4.8	11.3	14.0	-28.5	-29.1	-8.1	-57.0	-12.0	-24.3	38.3	88.2	12.0	11.6	22.3	-6.5	-27.8	25.0	-18.4	-9.6	3.3
12	-106.9	0.3	13.9	0.0	6.4	-73.3	-64.9	-25.6	-26.2	-17.6	12.0	157.0	-27.0	81.5	62.7	24.6	-13.7	-18.1	-28.2	43.2
13	78.4	-7.6	-13.0	-64.4	34.0	40.4	13.5	-16.4	3.5	-23.1	11.6	-27.0	77.1	-59.9	-77.6	9.3	30.4	28.8	10.0	-47.9
14	-99.3	10.5	-3.6	74.5	-68.7	-100.4	-83.9	-6.2	-40.0	-17.8	22.3	81.5	-59.9	211.1	157.5	-1.4	-3.6	-119.8	-17.7	65.0
15	-182.5	8.7	-36.9	164.2	-82.4	-144.4	-111.2	-16.6	-40.3	79.0	-6.5	62.7	-77.6	157.5	305.6	23.9	-19.5	-120.8	-10.2	47.1
16	-0.3	-32.6	-4.4	-82.2	68.3	0.7	-42.9	48.6	-121.0	-87.6	-27.8	24.6	9.3	-1.4	23.9	185.7	59.3	-20.9	-1.6	2.2
17	-61.5	-2.7	37.7	-70.9	17.9	-18.8	-60.9	46.4	-129.9	0.3	25.0	-13.7	30.4	-3.6	-19.5	59.3	150.5	-16.2	3.4	26.7
18	-7.6	-47.6	-35.5	-93.4	32.6	105.0	51.9	-65.7	-12.5	64.3	-18.4	-18.1	28.8	-119.8	-120.8	-20.9	-16.2	287.0	18.9	-11.8
19	35.0	-18.9	-37.2	-14.2	17.0	25.6	3.2	0.6	-22.8	37.5	-9.6	-28.2	10.0	-17.7	-10.2	-1.6	3.4	18.9	28.1	-18.7
20	-121.7	53.9	19.2	6.6	-52.5	-44.6	-62.4	53.2	-43.3	-50.5	3.3	43.2	-47.9	65.0	47.1	2.2	26.7	-11.8	-18.7	132.9

$S_\beta = [s_{i,j}(\beta)]$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	377.4	73.5	-9.4	-116.9	-19.5	171.8	-18.3	-66.4	-199.9	-98.0	-79.0	-28.9	-80.3	-46.0	-15.4	88.8	47.2	16.0	-24.2	27.4
2	73.5	486.3	-78.8	-76.6	-114.9	-7.7	-46.8	54.0	-10.9	-61.6	15.4	-87.0	-19.8	-51.8	159.7	-154.6	-125.7	-72.5	19.3	100.3
3	-9.4	-78.8	195.7	80.6	-20.3	11.6	14.3	-0.2	-11.5	-27.8	-2.1	-38.0	-59.9	-46.1	-2.3	-74.1	30.0	-25.5	-3.2	66.7
4	-116.9	-76.6	80.6	205.0	19.8	-109.5	26.8	25.3	69.8	22.4	26.6	-16.5	24.7	-10.4	2.6	-78.3	-51.3	-19.1	6.5	-31.7
5	-19.5	-114.9	-20.3	19.8	355.4	-85.4	82.0	29.7	-85.0	-26.1	-11.3	177.1	85.6	-76.1	7.2	-0.3	-128.4	-73.3	-31.6	-84.8
6	171.8	-7.7	11.6	-109.5	-85.4	389.7	-36.1	-44.4	-170.9	-45.5	-52.0	-39.5	-166.6	-17.6	50.1	-19.2	75.7	30.5	21.9	43.1
7	-18.3	-46.8	14.3	26.8	82.0	-36.1	67.0	18.2	-30.9	-25.6	13.4	38.1	22.8	-33.0	7.5	-2.0	-44.9	-29.8	2.6	-25.2
8	-66.4	54.0	-0.2	25.3	29.7	-44.4	18.2	216.4	-3.3	14.4	-3.8	15.0	20.2	14.2	45.3	-169.2	-26.6	-118.1	3.6	-24.3
9	-199.9	-10.9	-11.5	69.8	-85.0	-170.9	-30.9	-3.3	407.3	52.9	102.7	53.3	78.3	-3.7	-141.7	-111.9	-39.3	119.0	-27.4	-46.9
10	-98.0	-61.6	-27.8	22.4	-26.1	-45.5	-25.6	14.4	52.9	203.7	-10.7	-21.6	7.1	76.1	7.5	31.5	-44.1	-0.2	-9.0	-45.3
11	-79.0	15.4	-2.1	26.6	-11.3	-52.0	13.4	-3.8	102.7	-10.7	73.7	8.2	34.7	-6.7	-31.9	-55.8	-35.6	39.2	-6.1	-19.1
12	-28.9	-87.0	-38.0	-16.5	177.1	-39.5	38.1	15.0	53.3	-21.6	8.2	219.3	3.7	-74.4	-73.8	-8.9	-29.0	9.4	-19.9	-86.5
13	-80.3	-19.8	-59.9	24.7	85.6	-166.6	22.8	20.2	78.3	7.1	34.7	3.7	190.3	2.8	-16.7	21.1	-87.8	3.3	-30.7	-32.7
14	-46.0	-51.8	-46.1	-10.4	-76.1	-17.6	-33.0	14.2	-3.7	76.1	-6.7	-74.4	2.8	216.4	-48.0	101.3	20.2	-8.8	1.9	-10.4
15	-15.4	159.7	-2.3	2.6	7.2	50.1	7.5	45.3	-141.7	7.5	-31.9	-73.8	-16.7	-48.0	286.2	-106.5	-138.1	-138.4	21.1	125.5
16	88.8	-154.6	-74.1	-78.3	-0.3	-19.2	-2.0	-169.2	-111.9	31.5	-55.8	-8.9	21.1	101.3	-106.5	417.4	134.0	57.2	-5.8	-64.9
17	47.2	-125.7	30.0	-51.3	-128.4	75.7	-44.9	-26.6	-39.3	-44.1	-35.6	-29.0	-87.8	20.2	-138.1	134.0	384.9	81.3	42.7	-65.4
18	16.0	-72.5	-25.5	-19.1	-73.3	30.5	-29.8	-118.1	119.0	-0.2	39.2	9.4	3.3	-8.8	-138.4	57.2	81.3	231.1	-14.7	-86.7
19	-24.2	19.3	-3.2	6.5	-31.6	21.9	2.6	3.6	-27.4	-9.0	-6.1	-19.9	-30.7	1.9	21.1	-5.8	42.7	-14.7	35.9	16.8
20	27.4	100.3	66.7	-31.7	-84.8	43.1	-25.2	-24.3	-46.9	-45.3	-19.1	-86.5	-32.7	-10.4	125.5	-64.9	-65.4	-86.7	16.8	244.1

$S_{\alpha+\beta} = [s_{i,j}(\alpha+\beta)]$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	546.8	-148.8	17.9	30.9	-52.0	86.2	-44.5	-120.6	21.2	60.8	0.1	-100.7	-85.3	-67.2	-80.5	-67.4	-46.1	217.7	-31.8	-136.8
2	-148.8	315.8	-75.3	-17.3	-41.7	-3.2	-0.7	-68.8	-20.9	-114.9	-40.1	94.0	42.7	-37.2	-48.7	134.9	55.0	-36.5	-18.5	30.2
3	17.9	-75.3	106.8	-34.8	33.4	7.8	27.9	-3.6	-64.7	41.8	5.8	-27.8	-8.4	-6.6	26.4	-24.3	-9.2	-24.6	17.7	-6.5
4	30.9	-17.3	-34.8	176.2	-30.5	-11.4	-24.2	0.5	167.8	33.0	-18.2	-49.0	-22.3	-27.3	-11.2	-95.7	-44.0	23.5	-1.9	-43.9
5	-52.0	-41.7	33.4	-30.5	73.9	-47.4	27.5	14.5	-19.9	61.4	17.8	3.6	9.8	0.5	20.2	-0.7	-33.4	-37.1	3.3	-3.1
6	86.2	-3.2	7.8	-11.4	-47.4	299.2	-22.4	-73.1	-109.7	-54.2	-66.0	-32.4	-14.6	-49.8	-43.5	-48.6	9.2	99.8	18.1	56.0

7	-44.5	-0.7	27.9	-24.2	27.5	-22.4	63.6	-34.3	-11.3	-0.2	12.8	-4.2	2.3	-0.3	20.3	-0.8	-5.5	-27.4	0.9	20.5
8	-120.6	-68.8	-3.6	0.5	14.5	-73.1	-34.3	201.2	45.5	26.5	12.6	1.5	-13.0	63.4	40.2	-21.7	-4.5	-80.8	14.2	0.3
9	21.2	-20.9	-64.7	167.8	-19.9	-109.7	-11.3	45.5	462.7	85.5	12.4	-41.8	-65.4	-41.4	-46.8	-158.9	-36.4	-42.4	-40.8	-94.6
10	60.8	-114.9	41.8	33.0	61.4	-54.2	-0.2	26.5	85.5	248.5	1.5	-84.3	17.7	-46.9	15.6	-126.1	-81.8	-8.2	5.3	-81.1
11	0.1	-40.1	5.8	-18.2	17.8	-66.0	12.8	12.6	12.4	1.5	46.9	-3.4	0.7	22.4	14.3	-1.4	18.3	-25.9	-2.5	-8.1
12	-100.7	94.0	-27.8	-49.0	3.6	-32.4	-4.2	1.5	-41.8	-84.3	-3.4	124.3	9.4	9.6	-19.7	108.5	23.1	-52.4	-17.9	59.5
13	-85.3	42.7	-8.4	-22.3	9.8	-14.6	2.3	-13.0	-65.4	17.7	0.7	9.4	104.7	8.5	34.9	4.7	-18.7	-17.8	11.5	-1.3
14	-67.2	-37.2	-6.6	-27.3	0.5	-49.8	-0.3	63.4	-41.4	-46.9	22.4	9.6	8.5	91.3	40.8	10.9	19.2	-40.4	19.5	30.8
15	-80.5	-48.7	26.4	-11.2	20.2	-43.5	20.3	40.2	-46.8	15.6	14.3	-19.7	34.9	40.8	112.6	-26.5	-31.4	-43.3	17.5	8.8
16	-67.4	134.9	-24.3	-95.7	-0.7	-48.6	-0.8	-21.7	-158.9	-126.1	-1.4	108.5	4.7	10.9	-26.5	242.7	51.6	-3.8	-24.3	46.7
17	-46.1	55.0	-9.2	-44.0	-33.4	9.2	-5.5	-4.5	-36.4	-81.8	18.3	23.1	-18.7	19.2	-31.4	51.6	166.3	-12.7	-16.1	-2.9
18	217.7	-36.5	-24.6	23.5	-37.1	99.8	-27.4	-80.8	-42.4	-8.2	-25.9	-52.4	-17.8	-40.4	-43.3	-3.8	-12.7	170.5	-11.2	-47.0
19	-31.8	-18.5	17.7	-1.9	3.3	18.1	0.9	14.2	-40.8	5.3	-2.5	-17.9	11.5	19.5	17.5	-24.3	-16.1	-11.2	31.9	24.9
20	-136.8	30.2	-6.5	-43.9	-3.1	56.0	20.5	0.3	-94.6	-81.1	-8.1	59.5	-1.3	30.8	8.8	46.7	-2.9	-47.0	24.9	147.5

$S_{\alpha/\beta} = [s_{ij}(\alpha/\beta)]$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	335.6	-20.9	-11.2	-62.7	-105.5	51.1	-12.9	-5.5	39.1	2.3	-2.8	-16.3	-32.4	10.4	-75.2	58.8	-48.2	-1.9	5.6	-107.3
2	-20.9	18.1	10.8	24.3	12.3	19.8	-0.8	18.1	-10.8	4.6	10.4	-24.6	-14.0	-8.3	14.5	-14.0	-0.1	-32.6	-0.6	-6.2
3	-11.2	10.8	122.4	37.3	3.6	-8.1	-39.1	22.0	21.7	19.8	-9.5	-31.4	-37.4	26.8	14.2	-71.7	-13.4	-42.4	1.0	-15.5
4	-62.7	24.3	37.3	167.1	-10.2	-35.0	-23.1	82.6	33.2	-13.5	22.2	-38.2	-29.7	-15.2	35.3	-114.3	-35.8	-13.1	2.6	-13.8
5	-105.5	12.3	3.6	-10.2	89.0	-6.8	11.7	-18.6	-22.9	23.0	3.0	6.1	25.9	-5.1	23.4	-23.4	20.1	-59.0	2.7	30.7
6	51.1	19.8	-8.1	-35.0	-6.8	158.5	-7.7	15.4	-33.8	-22.5	0.6	-16.2	-31.7	-26.4	-53.7	80.5	-7.0	-47.6	6.5	-36.1
7	-12.9	-0.8	-39.1	-23.1	11.7	-7.7	42.2	-24.7	-31.3	10.6	3.6	-1.9	27.9	-16.7	19.2	-2.4	24.3	11.8	-10.6	20.2
8	-5.5	18.1	22.0	82.6	-18.6	15.4	-24.7	109.3	40.5	-15.9	21.1	-22.4	-54.8	-12.9	-20.6	-19.7	-25.1	-63.4	4.3	-29.8
9	39.1	-10.8	21.7	33.2	-22.9	-33.8	-31.3	40.5	167.6	-15.6	9.3	6.4	-5.0	17.6	-44.5	-82.3	-52.1	16.1	1.7	-54.7
10	2.3	4.6	19.8	-13.5	23.0	-22.5	10.6	-15.9	-15.6	70.6	12.2	-20.4	-8.8	9.6	36.2	-37.1	10.3	-44.3	-1.9	-19.3
11	-2.8	10.4	-9.5	22.2	3.0	0.6	3.6	21.1	9.3	12.2	44.3	-19.0	7.0	-9.9	6.2	-28.5	-7.7	-28.9	-8.4	-25.3
12	-16.3	-24.6	-31.4	-38.2	6.1	-16.2	-1.9	-22.4	6.4	-20.4	-19.0	80.2	22.1	-1.1	-43.1	46.2	-2.2	46.4	4.4	24.9
13	-32.4	-14.0	-37.4	-29.7	25.9	-31.7	27.9	-54.8	-5.0	-8.8	7.0	22.1	91.0	-16.4	1.3	-21.8	3.0	55.5	-10.1	28.7
14	10.4	-8.3	26.8	-15.2	-5.1	-26.4	-16.7	-12.9	17.6	9.6	-9.9	-1.1	-16.4	60.1	6.0	-14.3	-6.6	6.7	-7.3	2.9
15	-75.2	14.5	14.2	35.3	23.4	-53.7	19.2	-20.6	-44.5	36.2	6.2	-43.1	1.3	6.0	114.8	-72.5	9.3	5.0	-11.6	35.8
16	58.8	-14.0	-71.7	-114.3	-23.4	80.5	-2.4	-19.7	-82.3	-37.1	-28.5	46.2	-21.8	-14.3	-72.5	259.0	39.1	-25.2	29.1	14.4
17	-48.2	-0.1	-13.4	-35.8	20.1	-7.0	24.3	-25.1	-52.1	10.3	-7.7	-2.2	3.0	-6.6	9.3	39.1	72.5	-19.8	1.8	37.4
18	-1.9	-32.6	-42.4	-13.1	-59.0	-47.6	11.8	-63.4	16.1	-44.3	-28.9	46.4	55.5	6.7	5.0	-25.2	-19.8	227.7	-25.8	34.9
19	5.6	-0.6	1.0	2.6	2.7	6.5	-10.6	4.3	1.7	-1.9	-8.4	4.4	-10.1	-7.3	-11.6	29.1	1.8	-25.8	23.3	-6.6
20	-107.3	-6.2	-15.5	-13.8	30.7	-36.1	20.2	-29.8	-54.7	-19.3	-25.3	24.9	28.7	2.9	35.8	14.4	37.4	34.9	-6.6	84.9

(3) Eigenvalues and eigenvectors. The eigenvalue (in boldface) is listed in the 2nd column, followed by the 20 components of its eigenvector that are printed according to the alphabetical order of the single amino acid letter code.

		A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
α class																					
1	**0.0**	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22
2	**1.6**	-0.06	-0.13	0.01	0.04	-0.21	0.06	0.14	-0.25	-0.03	-0.29	0.23	0.16	-0.25	-0.25	0.06	-0.01	0.14	-0.10	0.73	0.00
3	**18.9**	0.10	-0.46	-0.13	0.23	0.21	0.05	0.11	-0.20	-0.04	-0.12	0.27	-0.04	0.05	-0.24	0.11	-0.11	0.13	-0.10	-0.36	0.52
4	**13.4**	0.10	0.22	-0.26	0.06	-0.03	0.08	0.15	0.00	0.06	-0.04	0.08	0.19	-0.54	-0.04	0.05	-0.02	0.51	0.10	-0.36	-0.30
5	**30.8**	0.10	0.02	-0.13	-0.26	-0.11	-0.57	0.43	0.28	-0.14	-0.04	0.37	-0.02	0.09	-0.09	0.22	-0.06	-0.17	0.19	-0.03	-0.07
6	**33.4**	0.00	-0.21	0.08	-0.02	-0.33	0.22	0.05	0.23	-0.06	-0.09	-0.30	0.41	0.41	-0.20	0.28	-0.33	0.17	-0.03	-0.11	-0.18
7	**53.7**	-0.19	0.09	0.11	0.29	0.03	0.40	-0.06	0.01	-0.25	-0.24	0.51	-0.07	0.11	0.01	0.06	0.03	-0.32	0.07	-0.17	-0.40
8	**69.6**	-0.11	0.34	-0.01	-0.10	0.48	-0.01	0.09	-0.25	-0.14	-0.14	-0.15	-0.25	0.12	0.06	0.33	-0.51	0.16	0.01	0.13	-0.04
9	**74.5**	0.06	-0.57	0.28	0.08	0.15	-0.07	-0.11	0.28	0.01	-0.03	0.01	-0.17	-0.28	0.36	-0.02	-0.30	0.11	0.26	0.12	-0.19
10	**78.8**	-0.07	-0.06	0.39	-0.42	-0.13	0.20	-0.03	-0.08	0.31	0.08	0.05	-0.31	-0.21	-0.19	0.51	0.18	-0.04	-0.02	-0.16	0.01
11	**82.9**	0.05	0.11	0.09	0.00	0.38	0.03	-0.20	0.35	-0.07	0.25	0.03	0.32	-0.34	-0.45	0.03	-0.22	-0.32	-0.17	0.07	0.06
12	**148.8**	-0.16	-0.10	-0.34	-0.42	0.03	0.31	0.12	0.16	0.20	0.12	0.30	0.07	0.03	0.37	-0.11	-0.21	-0.01	-0.43	0.04	0.03
13	**169.3**	0.01	0.04	0.29	0.00	-0.08	0.18	0.59	-0.08	-0.31	-0.09	-0.28	0.14	-0.30	0.26	-0.07	-0.01	-0.29	-0.09	-0.12	0.21
14	**197.3**	-0.02	-0.13	0.43	-0.12	0.28	-0.31	0.09	-0.38	0.12	0.08	0.14	0.37	0.16	0.02	-0.17	0.08	0.10	-0.27	-0.13	-0.35
15	**256.3**	-0.01	-0.18	-0.34	-0.17	0.19	0.05	-0.14	-0.29	0.16	-0.15	-0.12	0.43	-0.06	0.19	0.27	0.13	-0.37	0.41	0.01	-0.01
16	**263.5**	-0.01	0.26	0.27	-0.18	-0.32	-0.02	-0.26	-0.19	0.07	-0.03	0.28	0.19	-0.03	0.07	-0.24	-0.44	0.01	0.33	-0.12	0.33
17	**510.6**	-0.62	-0.03	0.07	-0.21	0.26	0.05	0.16	0.27	0.07	-0.28	-0.07	0.04	0.05	-0.22	-0.30	0.18	0.19	0.29	-0.03	0.12
18	**672.3**	-0.08	-0.11	-0.02	-0.30	0.02	0.08	-0.20	-0.08	-0.73	0.36	0.09	0.09	0.00	0.07	0.11	0.22	0.27	0.10	0.05	0.07
19	**774.5**	0.09	0.15	0.12	-0.07	0.04	-0.20	-0.34	0.26	-0.11	-0.60	0.02	0.15	-0.06	0.27	0.21	0.20	0.11	-0.35	-0.07	0.19
20	**1315.4**	0.66	-0.02	0.01	-0.40	0.18	0.29	0.04	0.02	-0.07	-0.28	-0.01	-0.12	0.12	-0.20	-0.33	0.06	0.00	0.12	0.04	-0.11
β class																					
1	**0.0**	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22
2	**2.8**	-0.12	0.16	-0.16	0.16	0.02	0.09	0.50	-0.15	-0.03	0.10	-0.23	0.02	0.01	0.08	-0.08	-0.11	0.24	0.01	-0.68	0.17
3	**9.0**	-0.05	-0.04	-0.23	0.11	0.27	0.00	-0.15	0.43	0.08	-0.05	0.24	-0.34	-0.40	-0.24	-0.05	0.30	-0.04	0.18	-0.25	0.24
4	**21.4**	-0.02	0.01	0.02	-0.16	0.22	-0.01	0.36	0.06	0.33	-0.12	-0.62	-0.34	-0.17	0.04	0.02	0.02	-0.08	0.14	0.34	-0.06
5	**32.3**	0.27	-0.14	-0.17	-0.15	0.22	-0.11	0.19	-0.32	0.25	0.08	0.44	-0.28	-0.21	0.12	0.20	-0.22	0.22	-0.32	0.04	-0.12
6	**38.3**	0.01	-0.04	0.30	-0.13	-0.01	-0.04	-0.22	0.07	0.23	-0.32	-0.08	0.10	-0.08	0.24	0.53	0.14	0.01	0.07	-0.45	-0.31
7	**51.4**	-0.13	0.23	0.15	0.14	0.30	0.41	-0.36	-0.16	0.24	0.02	-0.11	-0.23	0.25	-0.10	-0.22	0.14	0.09	-0.43	-0.08	-0.13
8	**63.8**	0.09	-0.02	-0.27	0.29	0.31	-0.16	-0.50	-0.20	-0.06	-0.04	-0.28	0.07	-0.05	0.28	0.09	-0.31	0.20	0.23	0.07	0.23
9	**77.8**	-0.22	0.23	0.45	-0.42	0.34	-0.13	-0.06	-0.09	-0.30	0.21	0.16	-0.19	-0.05	0.14	-0.18	-0.16	0.05	0.32	-0.08	-0.02
10	**120.7**	-0.22	0.24	0.00	0.28	0.05	0.10	0.16	-0.16	-0.10	-0.42	0.22	0.14	-0.40	0.44	-0.22	0.12	-0.20	-0.08	0.16	-0.09
11	**149.8**	0.15	0.21	0.21	0.08	-0.03	-0.26	-0.10	-0.11	0.13	0.47	-0.21	0.34	-0.55	-0.17	-0.05	0.16	-0.02	-0.21	-0.02	-0.02
12	**179.2**	-0.07	0.18	-0.06	0.45	-0.03	0.00	0.08	-0.16	-0.24	0.14	0.04	-0.24	-0.01	-0.30	0.29	0.03	0.07	0.30	0.09	-0.56
13	**201.1**	-0.58	0.05	-0.17	-0.27	0.06	0.07	0.00	-0.25	0.03	-0.10	0.04	0.26	-0.07	-0.29	0.34	0.18	0.31	-0.01	0.19	0.22
14	**296.8**	-0.12	-0.20	-0.23	-0.13	0.02	0.55	-0.07	-0.12	0.11	0.43	0.02	0.11	-0.14	0.17	0.12	-0.11	-0.47	0.20	-0.06	-0.09
15	**320.6**	0.12	-0.17	0.33	0.19	-0.01	-0.07	0.05	-0.57	0.10	-0.13	0.06	-0.15	0.08	-0.21	0.08	0.16	-0.36	0.18	-0.08	0.41
16	**466.6**	0.33	0.37	-0.22	-0.24	0.13	0.05	-0.01	-0.12	0.27	-0.28	0.11	0.32	0.05	-0.34	-0.25	-0.13	-0.08	0.32	-0.08	-0.18
17	**487.9**	-0.05	-0.34	0.41	0.20	0.04	0.33	0.05	0.14	0.06	-0.17	0.03	0.15	-0.28	-0.29	-0.08	-0.51	0.23	0.04	0.04	0.00
18	**682.9**	-0.14	0.25	0.03	0.01	-0.65	0.03	-0.19	-0.05	0.42	0.07	0.11	-0.29	-0.10	0.14	-0.11	-0.13	0.22	0.23	0.05	0.11
19	**979.5**	0.09	-0.49	-0.02	-0.06	-0.02	0.06	-0.02	-0.22	0.01	0.04	-0.04	0.08	-0.03	0.12	-0.39	0.47	0.40	0.27	-0.01	-0.24
20	**1822.8**	0.45	0.20	0.01	-0.23	-0.19	0.46	-0.07	-0.09	-0.45	-0.13	-0.15	-0.17	-0.25	-0.01	0.17	0.12	0.17	-0.06	0.05	0.18
α+β class																					
1	**0.0**	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22
2	**2.7**	0.23	-0.09	-0.15	-0.01	0.39	-0.16	-0.03	0.06	0.08	-0.21	-0.56	-0.19	0.26	-0.25	0.08	0.03	0.25	-0.18	0.20	0.24
3	**6.3**	0.03	0.17	0.27	0.04	0.20	-0.06	-0.24	0.04	0.00	-0.03	-0.01	-0.24	0.11	0.20	-0.04	-0.11	-0.01	0.10	-0.74	0.33

4	18.0	-0.12	0.12	0.18	-0.20	0.21	-0.16	-0.59	-0.13	0.10	-0.10	0.23	0.02	-0.19	-0.20	0.17	-0.16	0.00	0.34	0.36	0.11
5	21.6	-0.01	0.06	0.09	-0.03	-0.19	-0.29	0.38	0.34	-0.15	-0.11	-0.03	0.24	0.07	-0.40	-0.10	-0.39	0.05	0.39	-0.08	0.15
6	24.6	0.11	0.02	-0.19	0.12	0.06	0.12	-0.01	0.19	-0.09	-0.14	0.64	-0.29	0.13	-0.48	-0.01	0.11	-0.06	-0.29	-0.05	0.10
7	31.2	0.07	0.19	-0.32	0.16	0.59	0.08	0.10	0.04	-0.25	-0.09	0.06	0.24	-0.33	0.12	0.07	-0.35	0.02	-0.06	-0.09	-0.25
8	41.1	0.07	-0.03	-0.24	0.01	-0.27	0.00	-0.09	-0.09	-0.04	0.32	-0.08	0.17	-0.27	-0.23	0.64	0.04	0.22	-0.09	-0.26	0.22
9	56.8	0.01	0.31	-0.25	-0.22	-0.03	-0.12	0.28	0.10	-0.02	0.19	-0.01	-0.57	-0.39	0.19	-0.06	0.03	-0.06	0.15	0.19	0.27
10	62.9	-0.18	-0.08	-0.02	-0.33	0.19	0.24	0.23	0.13	0.28	-0.25	-0.07	-0.15	0.00	-0.16	0.36	0.24	-0.16	0.31	-0.23	-0.35
11	78.8	-0.33	-0.42	-0.38	0.19	0.21	-0.05	0.02	-0.22	-0.03	0.20	0.12	-0.01	0.08	-0.03	-0.26	0.19	0.24	0.43	-0.09	0.14
12	87.4	0.15	-0.01	-0.43	-0.58	-0.04	0.08	-0.04	-0.07	0.19	0.05	0.18	0.20	0.41	0.22	-0.03	-0.30	-0.05	-0.04	-0.04	0.14
13	146.4	0.00	-0.03	0.02	0.24	-0.12	-0.24	0.29	-0.49	-0.02	-0.40	0.19	-0.22	0.19	0.23	0.39	-0.21	0.08	0.05	0.07	-0.01
14	159.1	0.10	-0.20	0.05	0.03	0.16	-0.05	0.21	-0.21	0.17	-0.17	0.00	0.30	-0.26	-0.05	-0.05	0.16	-0.58	-0.09	-0.01	0.48
15	169.9	-0.05	-0.07	0.36	-0.33	0.14	0.12	0.30	-0.30	0.20	0.06	0.14	0.00	-0.26	-0.17	-0.21	-0.13	0.50	-0.26	-0.08	0.03
16	266.7	-0.09	0.38	0.08	-0.02	0.20	-0.09	0.14	-0.42	-0.12	0.49	-0.09	-0.03	0.33	-0.33	0.02	0.01	-0.31	0.00	-0.01	-0.14
17	337.9	-0.36	0.03	-0.02	0.22	-0.10	0.68	-0.01	-0.03	0.05	-0.03	-0.19	-0.13	0.05	-0.08	0.02	-0.42	-0.09	0.01	0.13	0.27
18	511.0	-0.06	-0.52	0.29	-0.23	0.20	0.00	0.07	0.19	-0.50	0.26	0.07	-0.14	0.09	0.15	0.27	-0.13	-0.13	-0.08	0.13	0.08
19	743.4	-0.43	-0.09	-0.02	0.19	0.10	-0.39	0.03	0.29	0.54	0.23	0.07	-0.03	0.02	0.07	0.12	-0.23	-0.09	-0.31	0.01	-0.06
20	968.5	0.60	-0.33	0.03	0.20	-0.03	0.08	-0.05	-0.07	0.31	0.27	0.00	-0.24	-0.11	-0.11	-0.06	-0.32	-0.13	0.23	-0.02	-0.23
α/β class																					
1	0.0	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22
2	1.3	-0.06	0.30	-0.04	-0.07	-0.16	0.00	-0.07	0.19	-0.08	-0.18	-0.37	0.30	0.32	0.16	0.29	-0.04	0.23	-0.10	-0.11	-0.53
3	2.5	0.00	0.81	-0.11	-0.11	-0.26	-0.10	-0.17	-0.12	0.02	0.17	-0.01	0.02	0.00	-0.05	-0.22	-0.09	-0.11	-0.05	0.08	0.30
4	4.9	-0.08	0.14	0.10	-0.25	0.04	-0.14	0.43	0.08	-0.19	-0.33	0.18	-0.01	-0.18	0.13	-0.08	-0.14	-0.15	0.10	0.60	-0.23
5	8.2	-0.06	0.19	0.22	0.26	-0.04	-0.09	0.58	-0.36	0.14	-0.09	-0.04	0.10	-0.14	0.00	-0.07	0.26	-0.17	-0.11	-0.42	-0.14
6	11.8	0.17	0.30	0.01	-0.07	0.63	-0.23	-0.23	0.08	-0.05	-0.20	0.07	-0.11	-0.25	-0.17	-0.03	0.08	0.15	0.29	-0.27	-0.17
7	15.4	-0.07	0.01	-0.03	0.02	0.16	-0.09	0.21	0.33	-0.08	0.34	-0.48	-0.43	0.23	0.14	-0.30	0.14	-0.21	0.16	0.00	-0.07
8	28.5	0.00	0.02	-0.29	0.45	0.12	0.07	-0.13	-0.42	-0.08	-0.01	-0.02	-0.09	-0.05	0.47	-0.30	-0.07	0.31	0.01	0.20	-0.18
9	39.0	-0.14	0.00	0.22	-0.06	-0.20	-0.15	-0.18	0.09	-0.21	-0.04	0.60	-0.18	0.38	0.28	-0.20	0.19	-0.05	0.03	-0.26	-0.14
10	41.8	0.15	-0.03	-0.21	-0.08	0.10	0.10	0.18	0.31	-0.24	-0.11	0.01	0.29	-0.16	0.44	-0.14	-0.32	-0.10	-0.11	-0.39	0.31
11	45.7	0.14	0.06	-0.09	-0.01	0.12	-0.01	-0.06	-0.10	0.21	-0.56	-0.12	-0.40	0.26	0.23	0.29	0.12	-0.22	-0.23	0.05	0.30
12	54.9	0.07	-0.02	0.20	-0.08	-0.21	-0.03	0.25	0.12	0.12	-0.24	-0.08	-0.27	0.07	-0.16	-0.30	-0.19	0.69	-0.04	-0.09	0.19
13	63.5	-0.27	0.07	-0.32	-0.31	-0.16	0.17	0.09	0.08	0.40	0.15	0.15	-0.29	-0.36	0.28	0.24	0.09	0.13	0.16	-0.16	-0.13
14	116.2	-0.15	0.09	0.39	-0.18	0.12	0.71	-0.10	-0.24	-0.04	-0.09	-0.10	-0.09	0.08	0.02	-0.10	-0.26	-0.14	0.24	-0.10	-0.05
15	143.6	-0.03	-0.07	0.52	-0.03	-0.20	-0.23	-0.28	-0.01	-0.11	-0.02	-0.32	0.05	-0.41	0.38	0.09	0.22	0.07	0.12	0.05	0.19
16	165.6	-0.13	-0.11	0.14	-0.39	0.34	-0.19	-0.10	-0.12	0.51	0.12	-0.03	0.28	0.14	0.18	-0.32	-0.03	0.01	-0.32	0.07	-0.05
17	221.2	-0.39	0.00	-0.13	0.30	-0.08	0.16	-0.16	0.36	0.25	-0.41	-0.04	0.24	-0.07	-0.13	-0.35	0.25	-0.11	0.17	0.07	0.07
18	363.4	0.22	-0.15	-0.11	-0.04	-0.21	-0.30	0.03	-0.21	0.27	-0.10	-0.06	0.16	0.23	0.07	-0.02	-0.25	-0.14	0.69	-0.08	0.02
19	465.0	-0.39	-0.05	-0.24	-0.31	0.16	-0.02	0.15	-0.29	-0.37	-0.04	-0.09	0.18	0.20	-0.04	0.07	0.35	0.23	0.19	-0.01	0.33
20	546.6	0.60	-0.06	-0.15	-0.33	-0.19	0.28	-0.02	-0.05	-0.04	-0.07	-0.05	0.08	-0.07	-0.02	-0.29	0.50	-0.02	-0.01	0.05	-0.16

Appendix D

The amino acid composition of the 64 testing proteins not included in the training database of Appendix A. Of the 64 proteins 9 are α proteins, 22 β proteins, 24 α+β proteins, and 9 α/β proteins. The data of each protein contain two lines: the 1st line successively indicates its length, PDB code, the ratios of α, β, parallel β sheets, and antiparallel sheets (see eq.1); and the 2nd line gives the frequencies of 20 amino acids according to the alphabetical order of the single amino acid letter code: ACDEFGHIKLMNPQRSTVWY. The frequencies are normalized to 100. The fifth character in the PDB code indicates a specific chain of a protein; if it is -, meaning the corresponding protein has only one chain.

---

9 α proteins

```
37 1BBL-   0.38  0.00  0.00  0.00
10.81  0.00  5.41  8.11  0.00 10.81  5.41  5.41  5.41 16.22  0.00  2.70  2.70  0.00 10.81  5.41  5.41  5.41  0.00  0.00
140 1HBBA   0.66  0.00  0.00  0.00
15.00  0.71  5.71  2.86  5.00  5.00  7.14  0.00  7.14 12.86  1.43  2.86  5.00  0.71  2.14  7.86  6.43  9.29  0.71  2.14
158 1IFA-   0.50  0.00  0.00  0.00
 4.43  0.63  1.27  8.86  5.70  1.27  1.27  5.06  6.96 12.03  4.43  6.33  0.63  8.23  8.23  4.43  6.96  5.70  2.53  5.06
340 1MRRA   0.66  0.03  0.00  1.00
 6.47  1.47  5.88  8.53  4.71  3.53  2.35  7.94  4.71 10.00  2.35  4.41  3.53  5.29  5.00  6.18  5.59  5.59  2.06  4.41
33 1PDE-   0.42  0.00  0.00  0.00
 6.06  0.00  9.09  6.06  3.03 12.12  0.00  6.06 12.12  9.09  0.00  3.03  0.00  3.03 12.12  0.00  3.03 12.12  0.00  3.03
323 1PRCM   0.53  0.03  0.00  1.00
11.15  1.55  3.72  2.17  7.74 11.76  3.10  8.05  1.55  8.36  1.86  1.86  5.88  2.48  4.02  4.95  4.95  4.64  5.57  4.64
185 1SAS-   0.57  0.02  0.00  1.00
 6.49  2.16 12.43  5.95  5.95  5.95  0.54  3.78  7.57  8.65  3.78  5.95  2.70  4.86  3.24  4.32  4.32  4.86  2.16  4.32
154 2TMVP   0.42  0.05  0.00  1.00
 8.44  0.65  5.19  4.55  5.19  3.25  0.00  5.84  1.30  7.79  0.00  6.49  4.55  5.84  7.14 10.39  9.74  9.09  1.95  2.60
108 4CPV-   0.50  0.00  0.00  0.00
18.52  0.93 12.96  5.56  9.26  7.41  0.93  4.63 12.04  8.33  0.00  2.78  0.00  1.85  0.93  4.63  4.63  4.63  0.00  0.00
```

22 β proteins

```
262 1AAIB   0.00  0.34  0.00  1.00
 6.11  3.44  6.49  1.91  1.53  7.63  0.76  6.87  2.67  9.16  1.15  8.02  4.58  5.73  4.96  7.63  8.02  6.49  3.44  3.44
46 1ATX-   0.00  0.33  0.00  1.00
 6.52 13.04  2.17  2.17  2.17 17.39  0.00  6.52  4.35  2.17  2.17  8.70  4.35  2.17  4.35  8.70  4.35  2.17  4.35  2.17
151 1COBA   0.00  0.39  0.00  1.00
 5.96  1.99  7.28  5.30  2.65 16.56  5.30  5.96  6.62  5.30  0.66  3.97  3.97  1.99  2.65  5.30  7.95  9.93  0.00  0.66
53 1EGF-   0.00  0.21  0.00  1.00
 0.00 11.32  7.55  3.77  0.00 11.32  1.89  3.77  0.00  7.55  1.89  5.66  3.77  1.89  7.55 11.32  3.77  3.77  3.77  9.43
240 1EST-   0.05  0.34  0.00  1.00
 7.08  3.33  2.92  1.67  1.25 10.42  2.50  4.17  1.25  7.50  0.83  7.08  2.92  6.25  5.00  9.17  7.92 11.25  2.92  4.58
47 1GPS-   0.00  0.26  1.00  0.00
 4.26 17.02  2.13  2.13  4.26 14.89  0.00  4.26  8.51  0.00  2.13  6.38  4.26  8.51 12.77  4.26  0.00  2.13  2.13  0.00
59 1HCC-   0.00  0.32  0.00  1.00
 5.08  6.78  3.39 10.17  3.39 11.86  5.08  6.78  6.78  3.39  1.69  0.00 10.17  1.69  0.00 10.17  1.69  5.08  1.69  5.08
39 1IXA-   0.00  0.21  0.00  1.00
 0.00 15.38 10.26 10.26  5.13 12.82  0.00  2.56  5.13  5.13  0.00 10.26  5.13  2.56  0.00  7.69  0.00  2.56  2.56  2.56
103 1MDAA   0.00  0.31  0.44  0.56
12.62  0.97  3.88  8.74  3.88  6.80  4.85  3.88  6.80  3.88  4.85  1.94  6.80  0.97  1.94  2.91  7.77 11.65  0.97  3.88
218 1PPFE   0.04  0.35  0.00  1.00
10.09  3.67  2.29  1.38  4.13 11.01  2.29  5.50  0.00  9.63  1.38  7.34  4.13  5.50  8.72  5.50  2.75 12.39  1.38  0.92
253 1R1A2   0.03  0.31  0.10  0.90
 6.72  1.19  5.53  2.77  3.16  6.72  3.95  6.32  3.95  7.91  2.77  6.32  5.93  3.95  4.35 10.28  6.32  5.93  2.77  3.16
59 1SHFA   0.00  0.41  0.21  0.79
 6.78  0.00  8.47 10.17  5.08  6.78  1.69  3.39  3.39  8.47  0.00  3.39  3.39  1.69  3.39  8.47  8.47  6.78  3.39  6.78
166 1TIE-   0.00  0.39  0.00  1.00
 4.22  2.41  6.63  7.83  3.01  8.43  1.20  3.61  7.23  9.64  0.00  2.41  6.02  5.42  4.22  6.63  4.82  9.64  1.81  4.82
152 1TNFA   0.00  0.43  0.00  1.00
 8.55  1.32  2.63  6.58  2.63  7.24  1.97  5.26  3.95 12.50  0.00  4.61  6.58  6.58  5.26  6.58  3.95  7.89  1.32  4.61
35 2ACHB   0.00  0.40  0.21  0.79
 2.86  0.00  2.86  0.00 11.43  0.00  0.00 11.43  5.71  2.86  5.71  8.57  8.57  5.71  8.57  2.86 14.29  8.57  0.00  0.00
71 2CTX-   0.00  0.35  0.00  1.00
 4.23 14.08  8.45  0.00  4.23  5.63  1.41  7.04  7.04  1.41  0.00  4.23  8.45  1.41  7.04  4.23 12.68  5.63  1.41  1.41
268 2MEV1   0.04  0.23  0.00  1.00
 5.60  1.87  4.85  4.48  7.46  7.46  1.49  2.61  4.85  6.72  0.75  4.85  9.70  3.36  3.36  8.21  8.96  7.84  1.49  4.10
288 2PLV1   0.05  0.28  0.02  0.97
 8.33  0.69  5.56  3.13  4.17  5.21  2.43  3.82  5.21  5.90  1.04  4.51  7.29  2.43  4.86  9.03 10.76  8.68  1.39  5.56
151 2SOD0   0.00  0.38  0.00  1.00
 5.96  1.99  7.28  5.30  2.65 16.56  5.30  5.96  6.62  5.30  0.66  3.97  3.97  1.99  2.65  5.30  7.95  9.93  0.00  0.66
224 3RP2A   0.05  0.37  0.00  1.00
 7.14  2.68  4.02  5.36  2.68  8.04  4.02  8.04  5.80  7.14  2.23  2.23  6.70  2.23  5.36  5.80  5.36  9.82  1.34  4.02
51 4SGBI   0.00  0.24  0.00  1.00
 9.80 15.69  3.92  1.96  1.96  7.84  1.96  5.88  9.80  1.96  0.00  7.84 11.76  1.96  1.96  5.88  1.96  0.00  0.00  7.84
388 5NN9-   0.01  0.43  0.00  1.00
 4.12  4.64  5.93  5.41  2.32  7.73  1.80  6.44  4.12  4.64  1.29  6.70  5.67  2.06  6.19  9.02  7.73  5.93  3.61  4.64
```

24 α+β proteins

```
321 1ABH-   0.37  0.21  0.48  0.52
10.90  0.00  5.92  4.98  3.74 10.90  0.31  5.30  9.66  7.48  0.00  5.30  4.36  4.36  1.25  5.92  6.54  6.85  2.49  3.74
173 1BBPA   0.10  0.48  0.00  1.00
 4.62  2.31  5.78  6.94  2.89  8.67  3.47  4.05 11.56  3.47  0.00  6.36  3.47  1.16  0.58  6.94  4.05 11.56  2.89  9.25
125 1EW4-   0.14  0.32  0.42  0.57
12.80  4.80  8.00  0.80  2.40  9.60  1.60  3.20  3.20  4.80  0.00  6.40  4.80  6.40  5.60  3.20  7.20  5.60  4.00  5.60
502 1COX-   0.24  0.19  0.41  0.59
11.55  0.20  4.98  4.18  4.18 11.55  0.80  4.78  4.58  6.18  2.59  5.58  4.38  2.99  3.39  7.37  7.37  7.77  1.79  3.78
250 1DNKA   0.26  0.27  0.19  0.81
 8.80  0.80  6.80  3.60  4.40  2.80  2.40  4.80  3.60  9.20  1.60  4.40  3.60  3.60  4.80 11.60  6.00 10.00  1.20  6.00
489 1GLAG   0.32  0.22  0.34  0.66
 9.41  1.02  5.32  7.77  2.86  7.98  1.84  6.95  4.09  7.36  2.86  4.09  2.86  4.09  6.34  4.29  7.16  7.36  2.66  3.68
129 1MS2A   0.13  0.42  0.00  1.00
10.85  1.55  3.10  3.88  3.10  6.98  0.00  6.20  4.65  5.43  1.55  7.75  4.65  4.65  3.10 10.08  6.98 10.85  1.55  3.10
56 1OVOA    0.18  0.21  0.25  0.75
 7.14 10.71  5.36  3.57  3.57  5.36  1.79  0.00  8.93  5.36  0.00 10.71  7.14  0.00  1.79  8.93  5.36  8.93  0.00  5.36
134 1POC-   0.27  0.17  0.22  0.78
 2.99  7.46  8.21  3.73  3.73  8.21  5.22  2.99  8.96  6.72  2.24  3.73  3.73  0.75  4.48  7.46  8.21  3.73  1.49  5.97
295 1PPBA   0.11  0.27  0.00  1.00
 4.41  2.71  6.44  7.12  3.73  8.81  1.69  5.76  7.46  8.47  2.37  3.73  5.08  2.71  7.12  5.76  4.07  5.76  3.05  3.73
108 1SHAA   0.15  0.30  0.25  0.75
 4.63  2.78  3.70  6.48  4.63  6.48  2.78  2.78  6.48 11.11  0.93  4.63  3.70  2.78  7.41  8.33  7.41  5.56  0.93  6.48
109 1THO-   0.36  0.26  0.29  0.71
11.01  1.83 10.09  4.59  3.67  8.26  0.92  8.26  9.17 11.93  0.92  3.67  4.59  2.75  1.83  2.75  5.50  4.59  1.83  1.83
108 1TRX-   0.32  0.24  0.23  0.77
11.11  1.85 10.19  4.63  3.70  8.33  0.93  8.33  9.26 12.04  0.93  3.70  4.63  2.78  0.93  2.78  5.56  4.63  1.85  1.85
476 2AAA-   0.27  0.18  0.35  0.65
 7.35  1.89  8.61  3.36  2.94  7.98  1.47  6.09  2.52  7.56  1.68  5.25  3.99  2.52  2.10 11.13  7.77  6.30  2.31  7.14
321 2PIA-   0.15  0.30  0.35  0.65
 7.48  2.49  7.48  6.54  4.98  7.17  2.18  4.05  4.05  7.48  1.87  3.43  5.30  2.18  8.10  9.35  6.85  6.23  1.25  1.56
65 2SN3-    0.12  0.18  0.00  1.00
 4.62 12.31  3.08  9.23  1.54 13.85  0.00  0.00 12.31  7.69  0.00  4.62  6.15  1.54  0.00  6.15  4.62  1.54  1.54  9.23
478 2TAAA   0.21  0.14  0.45  0.55
 7.74  2.09  9.21  2.51  2.93  8.58  1.26  5.65  4.18  7.11  1.88  5.44  4.39  3.77  2.09  7.74  8.37  6.07  1.88  7.11
86 3B5C-    0.24  0.22  0.37  0.63
 4.65  0.00  6.98 12.79  3.49  6.98  5.81  5.81  8.14  9.30  0.00  3.49  2.33  2.33  3.49  5.81  8.14  4.65  1.16  4.65
402 3SC2A   0.33  0.20  0.41  0.59
 9.20  1.49  6.47  3.98  3.98  8.46  3.23  3.48  2.24  8.21  1.74  3.98  5.97  2.74  5.22  6.72  5.97  6.97  2.74  7.21
152 3SC2B   0.37  0.21  0.41  0.59
 8.55  1.32  5.26  2.63  1.97  7.89  3.29  3.29  1.32  8.55  2.63  2.63  6.58  3.95  6.58  5.92  9.21  7.24  3.95  7.24
316 3TLN-   0.37  0.16  0.19  0.81
 8.86  0.00  7.91  2.53  3.16 11.39  2.53  5.70  3.48  5.06  0.63  6.01  2.53  4.11  3.16  8.23  7.91  6.96  0.95  8.86
436 4ENL-   0.39  0.16  0.41  0.59
12.61  0.23  7.11  5.73  3.67  8.49  2.52  5.05  8.26  9.17  1.15  4.36  3.44  2.06  3.21  7.34  4.59  7.80  1.15  2.06
30 4INSB    0.37  0.10  0.00  1.00
 6.67  6.67  0.00  6.67 10.00 10.00  6.67  0.00  3.33 13.33  0.00  3.33  3.33  3.33  3.33  3.33  3.33 10.00  0.00  6.67
237 4RCRH   0.17  0.11  0.11  0.89
 9.28  0.84  5.06  6.33  3.38  9.70  2.53  5.49  5.49  9.70  2.95  2.95  9.70  2.11  4.64  4.64  4.64  6.75  1.27  2.53
```

9 α/β proteins

```
823 1GPB-   0.45  0.15  0.53  0.47
 7.65  1.09  5.83  7.78  4.62  5.71  2.67  6.08  5.59  9.23  2.55  5.59  4.13  3.40  7.29  3.16  4.37  7.41  1.46  4.37
468 1MINA   0.45  0.13  0.89  0.11
 5.56  1.92  6.20  8.12  4.06  9.40  2.35  7.69  7.48  5.98  3.21  3.21  4.06  2.14  5.34  6.20  3.85  6.84  1.92  4.49
287 1NIPB   0.40  0.12  1.00  0.00
 9.76  2.44  5.92  9.76  2.09  9.76  0.70  7.67  5.92  7.32  4.88  4.53  2.79  2.44  4.53  3.48  4.18  8.71  0.00  3.14
309 1SBP-   0.45  0.17  0.57  0.43
10.36  0.00  8.41  6.15  3.88  6.80  1.94  5.83  9.06  6.80  0.00  5.83  3.88  3.24  3.24  5.50  5.18  7.77  2.27  3.88
248 1WSYA   0.50  0.13  1.00  0.00
14.92  1.21  4.84  6.05  4.84  6.85  2.02  7.26  3.23 10.48  2.02  4.03  6.45  4.03  4.84  5.24  2.42  6.45  0.00  2.82
414 4ICD-   0.37  0.18  0.55  0.45
 9.18  1.45  6.04  7.97  2.42  9.66  1.21  8.94  7.49  7.49  2.90  3.62  4.83  2.90  4.11  3.14  4.35  7.25  1.45  3.62
401 7AATA   0.46  0.14  0.57  0.43
 8.98  1.25  4.99  5.74  3.99  7.48  2.24  5.99  6.98  7.98  3.24  4.24  4.24  3.74  5.99  7.23  4.24  5.99  1.75  3.74
458 9RUBB   0.34  0.16  0.56  0.44
13.10  1.09  6.33  5.90  5.02 10.92  2.62  4.59  3.93  6.99  2.84  3.28  4.59  2.84  5.68  3.93  5.24  6.11  1.31  3.71
334 1GD1O   0.22  0.28  0.51  0.49
11.68  0.60  5.99  6.89  1.50  6.89  2.69  5.99  6.89  8.08  2.10  5.99  3.29  0.90  4.49  5.39  5.69 11.98  0.60  2.40
```

Appendix E

Now, let us prove that the value of the Mahalanobis distance (see Equation 38) defined in the (20-1)-D (dimensional) space is unchanged regardless of which one of the 20 normalized components is removed from the 20-D composition space.

Suppose $\mathbf{X}$ and $\mathbf{X}_k$ $(k = 1, 2, \cdots, N)$ are defined in an 20-D space by

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{20} \end{bmatrix}, \quad \mathbf{X_k} = \begin{bmatrix} x_{k,1} \\ x_{k,2} \\ \vdots \\ x_{k,20} \end{bmatrix} \quad (k = 1, 2, \cdots, N) \tag{E1}$$

where x_j and $x_{k,j}$ $(j = 1, 2, \cdots, 20)$ are the jth normalized components of $\mathbf{X}$ and $\mathbf{X}_k$ (see Equation 2) in the 20-D space, respectively. $\mathbf{X}$ can be either one of $\mathbf{X}_k$ $(k = 1, 2, \cdots, N)$ or not. The norm $\overline{\mathbf{X}}$ and the covariance matrix $\mathbf{S}$ for the set of $\mathbf{X}_k$ $(k = 1, 2, \cdots, N)$ are defined by (see Equations 4, 21, and 22)

$$\overline{\mathbf{X}} = \begin{bmatrix} \overline{x_1} \\ \overline{x_2} \\ \vdots \\ \overline{x_{20}} \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,20} \\ s_{2,1} & s_{2,2} & \cdots & s_{1,20} \\ \vdots & \vdots & \ddots & \vdots \\ s_{20,1} & s_{20,2} & \cdots & s_{20,20} \end{bmatrix} \tag{E2}$$

where

$$\begin{cases} \overline{x_i} = \frac{1}{N} \sum_{k=1}^{N} x_{k,i} \\ s_{i,j} = \sum_{k=1}^{N} \left[x_{k,i} - \overline{x_i}\right]\left[x_{k,j} - \overline{x_j}\right] \end{cases} \quad (i, j = 1, 2, \ldots, 20) \tag{E3}$$

The constraint due to the normalization yields

$$\sum_{i=1}^{20} x_i = 1, \quad \sum_{i=1}^{20} x_{k,i} = 1, \quad \sum_{i=1}^{20} \overline{x_i} = 1 \tag{E4}$$

From Equations E3 and E4, it follows that

$$\begin{cases} s_{i,j} = s_{j,i}, & (i, j = 1, 2, \ldots, 20) \\ \sum_{i=1}^{20} s_{i,j} = 0, & (j = 1, 2, \ldots, 20) \\ \sum_{j=1}^{20} s_{i,j} = 0, & (i = 1, 2, \ldots, 20) \end{cases} \tag{E5}$$

The Mahalanobis distance between $\mathbf{X}$ and $\overline{\mathbf{X}}$ in the 20-D space is (see Equation 20)

$$D^2(\mathbf{X}, \overline{\mathbf{X}}) = (\mathbf{X} - \overline{\mathbf{X}})^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{X} - \overline{\mathbf{X}}) \tag{E6}$$

Because Equation E5, the matrix $\mathbf{S}$ is singular, that is, its determinant $\mathbf{S} = 0$, and hence $\mathbf{S}^{-1}$ in Equation E6 is divergent. We therefore instead define the Mahalanobis distance in a 19-D space. To realize this, let us

assume that

$$\mathbf{X}_{\overline{p}} = \begin{bmatrix} x_1 \\ \vdots \\ x_{p-1} \\ x_{p+1} \\ \vdots \\ x_{20} \end{bmatrix}, \quad \overline{\mathbf{X}}_{\overline{p}} = \begin{bmatrix} \overline{x_1} \\ \vdots \\ \overline{x_{p-1}} \\ \overline{x_{p+1}} \\ \vdots \\ \overline{x_{20}} \end{bmatrix} \tag{E7}$$

and

$$\mathbf{S}_{\overline{p},\overline{p}} = \begin{bmatrix} s_{1,1} & \cdots & s_{1,p-1} & s_{1,p+1} & \cdots & s_{1,20} \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ s_{p-1,1} & \cdots & s_{p-1,p-1} & s_{p-1,p+1} & \cdots & s_{p-1,20} \\ s_{p+1,1} & \cdots & s_{p+1,p-1} & s_{p+1,p+1} & \cdots & s_{p+1,20} \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ s_{20,1} & \cdots & s_{20,p-1} & s_{20,p+1} & \cdots & s_{20,20} \end{bmatrix}, \quad (p = 1, 2, \ldots, 20) \tag{E8}$$

Thus, the Mahalanobis distance in a 19-D space is given by

$$D^2(\mathbf{X}_{\overline{p}}, \overline{\mathbf{X}}_{\overline{p}}) = (\mathbf{X}_{\overline{p}} - \overline{\mathbf{X}}_{\overline{p}})^{\mathrm{T}} \mathbf{S}^{-1}_{\overline{p},\overline{p}} (\mathbf{X}_{\overline{p}} - \overline{\mathbf{X}}_{\overline{p}}) \tag{E9}$$

Our task is to prove the following

$$D^2(\mathbf{X}_{\overline{1}}, \overline{\mathbf{X}}_{\overline{1}}) = D^2(\mathbf{X}_{\overline{2}}, \overline{\mathbf{X}}_{\overline{2}}) = \ldots = D^2(\mathbf{X}_{\overline{20}}, \overline{\mathbf{X}}_{\overline{20}}) \tag{E10}$$

Without losing generality, let us prove $D^2(\mathbf{X}_{\overline{1}}, \overline{\mathbf{X}}_{\overline{1}}) = D^2(\mathbf{X}_{\overline{20}}, \overline{\mathbf{X}}_{\overline{20}})$. According to the definition of an inverse matrix, it follows

$$\begin{aligned} D^2(\mathbf{X}_{\overline{20}}, \overline{\mathbf{X}}_{\overline{20}}) = \; & (\mathbf{X}_{\overline{20}} - \overline{\mathbf{X}}_{\overline{20}})^{\mathrm{T}} \mathbf{S}^{-1}_{\overline{20},\overline{20}} (\mathbf{X}_{\overline{20}} - \overline{\mathbf{X}}_{\overline{20}}) \\ = & \begin{bmatrix} x_1 - \overline{x_1} \\ x_2 - \overline{x_2} \\ \vdots \\ x_{19} - \overline{x_{19}} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} s^{-1}_{1,1} & s^{-1}_{1,2} & \cdots & s^{-1}_{1,19} \\ s^{-1}_{2,1} & s^{-1}_{2,2} & \cdots & s^{-1}_{2,19} \\ \vdots & \vdots & \ddots & \vdots \\ s^{-1}_{19,1} & s^{-1}_{19,2} & \cdots & s^{-1}_{19,19} \end{bmatrix} \begin{bmatrix} x_1 - \overline{x_1} \\ x_2 - \overline{x_2} \\ \vdots \\ x_{19} - \overline{x_{19}} \end{bmatrix} \end{aligned} \tag{E11}$$

where $s^{-1}_{i,j}$ $(i, j = 1, 2, \ldots, 19)$ are the elements of matrix $\mathbf{S}^{-1}_{\overline{20},\overline{20}}$. According to the theory of linear

algebra, the above equation can be further written as

$$
\begin{aligned}
D^2(\mathbf{X}_{\overline{20}}, \overline{\mathbf{X}}_{\overline{20}}) = \; & (\mathbf{X}_{\overline{20}} - \overline{\mathbf{X}}_{\overline{20}})^{\mathrm{T}} \mathbf{S}^{-1}_{\overline{20},\overline{20}} (\mathbf{X}_{\overline{20}} - \overline{\mathbf{X}}_{\overline{20}}) \\
& = \frac{1}{|\mathbf{S}_{\overline{20},\overline{20}}|} \begin{bmatrix} x_1 - \overline{x_1} \\ x_2 - \overline{x_2} \\ \vdots \\ x_{19} - \overline{x_{19}} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} A_{1,1} & A_{2,1} & \cdots & A_{19,1} \\ A_{1,2} & A_{2,2} & \cdots & A_{19,2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1,19} & A_{2,19} & \cdots & A_{19,19} \end{bmatrix} \begin{bmatrix} x_1 - \overline{x_1} \\ x_2 - \overline{x_2} \\ \vdots \\ x_{19} - \overline{x_{19}} \end{bmatrix} \\
& = \frac{1}{|\mathbf{S}_{\overline{20},\overline{20}}|} \sum_{p=1}^{19} \sum_{q=1}^{19} (x_p - \overline{x_p}) A_{p,q} (x_q - \overline{x_q}) \\
& = \frac{1}{|\mathbf{S}_{\overline{20},\overline{20}}|} \sum_{p=1}^{19} (x_p - \overline{x_p}) \sum_{q=1}^{19} A_{p,q} (x_q - \overline{x_q})
\end{aligned}
\tag{E12}
$$

where $A_{p,q}$ is a cofactor defined by

$$
A_{p,q} = (-1)^{p+q} \det \left[\mathbf{S}_{\overline{20},\overline{20}} \right]_{\overline{p},\overline{q}} \tag{E13}
$$

in which $\left[\mathbf{S}_{\overline{20},\overline{20}} \right]_{\overline{p},\overline{q}}$ is the matrix obtained by deleting the pth row and qth column from the matrix $\mathbf{S}_{\overline{20},\overline{20}}$. Thus, according to the basic principle of a determinant, it follows that

$$
\begin{aligned}
D^2(\mathbf{X}_{\overline{20}}, \overline{\mathbf{X}}_{\overline{20}}) = \; & \frac{1}{|\mathbf{S}_{\overline{20},\overline{20}}|} (x_1 - \overline{x_1}) \begin{vmatrix} x_1 - \overline{x_1} & x_2 - \overline{x_2} & \cdots & x_{19} - \overline{x_{19}} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,19} \\ \vdots & \vdots & \ddots & \vdots \\ s_{19,1} & s_{19,2} & \cdots & s_{19,19} \end{vmatrix} + \\
& \frac{1}{|\mathbf{S}_{\overline{20},\overline{20}}|} (x_2 - \overline{x_2}) \begin{vmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,19} \\ x_1 - \overline{x_1} & x_2 - \overline{x_2} & \cdots & x_{19} - \overline{x_{19}} \\ \vdots & \vdots & \ddots & \vdots \\ s_{19,1} & s_{19,2} & \cdots & s_{19,19} \end{vmatrix} + \cdots + \\
& \frac{1}{|\mathbf{S}_{\overline{20},\overline{20}}|} (x_{19} - \overline{x_{19}}) \begin{vmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,19} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,19} \\ \vdots & \vdots & \ddots & \vdots \\ x_1 - \overline{x_1} & x_2 - \overline{x_2} & \cdots & x_{19} - \overline{x_{19}} \end{vmatrix} \\
= \; & \frac{1}{|\mathbf{S}_{\overline{20},\overline{20}}|} \sum_{p=1}^{19} (x_p - \overline{x_p}) \Delta_{\overline{20},p}
\end{aligned}
\tag{E14}
$$

where

$$\Delta_{\overline{20},p} = \begin{vmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,19} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p-1,1} & s_{p-1,2} & \cdots & s_{p-1,19} \\ x_1 - \overline{x_1} & x_2 - \overline{x_2} & \cdots & x_{19} - \overline{x_{19}} \\ s_{p+1,1} & s_{p+1,2} & \cdots & s_{p+1,19} \\ \vdots & \vdots & \vdots & \vdots \\ s_{19,1} & s_{19,2} & \cdots & s_{19,19} \end{vmatrix} \tag{E15}$$

Similarly, we have

$$\begin{aligned} D^2(\mathbf{X}_{\overline{1}}, \overline{\mathbf{X}}_{\overline{1}}) &= \frac{1}{|\mathbf{S}_{\overline{1},\overline{1}}|}(x_2 - \overline{x_2}) \begin{vmatrix} x_2 - \overline{x_2} & x_3 - \overline{x_3} & \cdots & x_{20} - \overline{x_{20}} \\ s_{3,2} & s_{3,3} & \cdots & s_{3,20} \\ \vdots & \vdots & \ddots & \vdots \\ s_{20,2} & s_{20,3} & \cdots & s_{20,20} \end{vmatrix} + \\ &\quad \frac{1}{|\mathbf{S}_{\overline{1},\overline{1}}|}(x_3 - \overline{x_3}) \begin{vmatrix} s_{2,2} & s_{2,3} & \cdots & s_{2,20} \\ x_2 - \overline{x_2} & x_3 - \overline{x_3} & \cdots & x_{20} - \overline{x_{20}} \\ \vdots & \vdots & \ddots & \vdots \\ s_{20,2} & s_{20,3} & \cdots & s_{20,20} \end{vmatrix} + \cdots + \\ &\quad \frac{1}{|\mathbf{S}_{\overline{1},\overline{1}}|}(x_{20} - \overline{x_{20}}) \begin{vmatrix} s_{2,2} & s_{2,3} & \cdots & s_{2,20} \\ s_{3,2} & s_{3,3} & \cdots & s_{3,20} \\ \vdots & \vdots & \ddots & \vdots \\ x_2 - \overline{x_2} & x_3 - \overline{x_3} & \cdots & x_{20} - \overline{x_{20}} \end{vmatrix} \\ &= \frac{1}{|\mathbf{S}_{\overline{1},\overline{1}}|} \sum_{p=2}^{20} (x_p - \overline{x_p}) \Delta_{\overline{1},p} \end{aligned} \tag{E16}$$

where

$$\Delta_{\overline{1},p} = \begin{vmatrix} s_{2,2} & s_{2,3} & \cdots & s_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p-1,2} & s_{p-1,3} & \cdots & s_{p-1,20} \\ x_2 - \overline{x_2} & x_3 - \overline{x_3} & \cdots & x_{20} - \overline{x_{20}} \\ s_{p+1,2} & s_{p+1,3} & \cdots & s_{p+1,20} \\ \vdots & \vdots & \vdots & \vdots \\ s_{20,2} & s_{20,3} & \cdots & s_{20,20} \end{vmatrix} \tag{E17}$$

To prove that $D^2(\mathbf{X}_{\overline{20}}, \overline{\mathbf{X}}_{\overline{20}}) = D^2(\mathbf{X}_{\overline{1}}, \overline{\mathbf{X}}_{\overline{1}})$, let us first prove their denominators are equal to each

other, that is, $|S_{\overline{20},\overline{20}}| = |S_{\overline{1},\overline{1}}|$. According to Equation E5 as well as Equation E8, it follows that

$$
\begin{aligned}
|S_{\overline{1},\overline{1}}| &= \begin{vmatrix} s_{2,2} & s_{2,3} & \cdots & s_{2,20} \\ s_{3,2} & s_{3,3} & \cdots & s_{3,20} \\ \vdots & \vdots & \ddots & \vdots \\ s_{20,2} & s_{20,3} & \cdots & s_{20,20} \end{vmatrix} = \begin{vmatrix} s_{2,2} & s_{2,3} & \cdots & -s_{2,1} \\ s_{3,2} & s_{3,3} & \cdots & -s_{3,1} \\ \vdots & \vdots & \ddots & \vdots \\ s_{20,2} & s_{20,3} & \cdots & -s_{20,1} \end{vmatrix} \\
&= (-1)^{20} \begin{vmatrix} s_{2,1} & s_{2,2} & \cdots & s_{2,19} \\ s_{3,1} & s_{3,2} & \cdots & s_{3,19} \\ \vdots & \vdots & \ddots & \vdots \\ s_{20,1} & s_{20,2} & \cdots & s_{20,19} \end{vmatrix} = (-1)^{20} \begin{vmatrix} s_{2,1} & s_{2,2} & \cdots & s_{2,19} \\ s_{3,1} & s_{3,2} & \cdots & s_{3,19} \\ \vdots & \vdots & \ddots & \vdots \\ -s_{1,1} & -s_{1,2} & \cdots & -s_{1,19} \end{vmatrix} \qquad (E18) \\
&= (-1)^{40} \begin{vmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,19} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,19} \\ \vdots & \vdots & \ddots & \vdots \\ s_{19,1} & s_{19,2} & \cdots & s_{19,19} \end{vmatrix} = |S_{\overline{20},\overline{20}}|
\end{aligned}
$$

which indicates that the denominator of Equation E14 is equal to that of Equation E16.

Now let us prove that their numerators are also equal to each other. Note that the first term of the numerator in Equation E16 is

$$
\begin{aligned}
(x_2 - \overline{x_2})\Delta_{\overline{1},2} &= (x_2 - \overline{x_2}) \begin{vmatrix} x_2 - \overline{x_2} & x_3 - \overline{x_3} & \cdots & x_{20} - \overline{x_{20}} \\ s_{3,2} & s_{3,3} & \cdots & s_{3;20} \\ \vdots & \vdots & \ddots & \vdots \\ s_{20,2} & s_{20,3} & \cdots & s_{20,20} \end{vmatrix} \\
&= (x_2 - \overline{x_2}) \begin{vmatrix} x_2 - \overline{x_2} & x_3 - \overline{x_3} & \cdots & -(x_1 - \overline{x_1}) \\ s_{3,2} & s_{3,3} & \cdots & -s_{3,1} \\ \vdots & \vdots & \ddots & \vdots \\ s_{20,2} & s_{20,3} & \cdots & -s_{20,1} \end{vmatrix} \\
&= (-1)^{19}(x_2 - \overline{x_2}) \begin{vmatrix} x_1 - \overline{x_1} & x_2 - \overline{x_2} & \cdots & -(x_{19} - \overline{x_{19}}) \\ s_{3,1} & s_{3,2} & \cdots & s_{3,19} \\ \vdots & \vdots & \ddots & \vdots \\ s_{20,1} & s_{20,2} & \cdots & s_{20,19} \end{vmatrix}
\end{aligned}
$$

$$
\begin{aligned}
&= (-1)^{19}(x_2-\overline{x_2})\begin{vmatrix} x_1-\overline{x_1} & x_2-\overline{x_2} & \cdots & -(x_{19}-\overline{x_{19}}) \\ s_{3,1} & s_{3,2} & \cdots & s_{3,19} \\ \vdots & \vdots & \ddots & \vdots \\ s_{19,1} & s_{19,2} & \cdots & s_{19,19} \\ -(s_{1,1}+s_{2,1}) & -(s_{1,2}+s_{2,2}) & \cdots & -(s_{1,19}+s_{2,19}) \end{vmatrix} \\
&= (-1)^{38}(x_2-\overline{x_2})\begin{vmatrix} s_{1,1}+s_{2,1} & s_{1,2}+s_{2,2} & \cdots & s_{1,19}+s_{2,19} \\ x_1-\overline{x_1} & x_2-\overline{x_2} & \cdots & -(x_{19}-\overline{x_{19}}) \\ s_{3,1} & s_{3,2} & \cdots & s_{3,19} \\ \vdots & \vdots & \vdots & \vdots \\ s_{19,1} & s_{19,2} & \cdots & s_{19,19} \end{vmatrix} \\
&= (x_2-\overline{x_2})\begin{vmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,19} \\ x_1-\overline{x_1} & x_2-\overline{x_2} & \cdots & -(x_{19}-\overline{x_{19}}) \\ s_{3,1} & s_{3,2} & \cdots & s_{3,19} \\ \vdots & \vdots & \ddots & \vdots \\ s_{19,1} & s_{19,2} & \cdots & s_{19,19} \end{vmatrix} \\
&\quad -(x_2-\overline{x_2})\begin{vmatrix} x_1-\overline{x_1} & x_2-\overline{x_2} & \cdots & -(x_{19}-\overline{x_{19}}) \\ s_{2,1} & s_{2,2} & \cdots & s_{2,19} \\ \vdots & \vdots & \ddots & \vdots \\ s_{19,1} & s_{19,2} & \cdots & s_{19,19} \end{vmatrix}. \\
&= (x_2-\overline{x_2})\Delta_{\overline{20},2} - (x_2-\overline{x_2})\Delta_{\overline{20},1}
\end{aligned} \tag{E19}
$$

Similarly, we have

$$
\begin{cases}
(x_3-\overline{x_3})\Delta_{\overline{1},3} = (x_3-\overline{x_3})\Delta_{\overline{20},3} - (x_3-\overline{x_3})\Delta_{\overline{20},1} \\
(x_4-\overline{x_4})\Delta_{\overline{1},4} = (x_4-\overline{x_4})\Delta_{\overline{20},4} - (x_4-\overline{x_4})\Delta_{\overline{20},1} \\
\vdots \qquad\qquad \vdots \qquad\qquad\qquad \vdots \\
(x_{19}-\overline{x}_{19})\Delta_{\overline{1},19} = (x_{19}-\overline{x}_{19})\Delta_{\overline{20},19} - (x_{19}-\overline{x_{19}})\Delta_{\overline{20},1}
\end{cases} \tag{E20}
$$

The last numerator term of Equation E16 is

$$(x_{20}-\overline{x_{20}})\Delta_{\overline{1},20} = (x_{20}-\overline{x_{20}})\begin{vmatrix} s_{2,2} & s_{2,3} & \cdots & s_{2,20} \\ s_{3,2} & s_{3,3} & \cdots & s_{3,20} \\ \vdots & \vdots & \ddots & \vdots \\ x_2-\overline{x_2} & x_3-\overline{x_3} & \cdots & x_{20}-\overline{x_{20}} \end{vmatrix}$$

$$= (x_{20}-\overline{x_{20}})\begin{vmatrix} s_{2,2} & s_{2,3} & \cdots & -s_{2,1} \\ s_{3,2} & s_{3,3} & \cdots & -s_{3,1} \\ \vdots & \vdots & \ddots & \vdots \\ x_2-\overline{x_2} & x_3-\overline{x_3} & \cdots & -(x_1-\overline{x_1}) \end{vmatrix} = -(x_{20}-\overline{x}_{20})\Delta_{\overline{20},1} \tag{E21}$$

Substituting Equations A17 to 20 into Equation E16, we obtain

$$\begin{aligned} D^2(\mathbf{X}_{\overline{1}},\overline{\mathbf{X}}_{\overline{1}}) &= \frac{1}{|\mathbf{S}_{\overline{20},\overline{20}}|}\left\{\sum_{p=2}^{19}(x_p-\overline{x_p})\Delta_{\overline{20},p}-\Delta_{\overline{20},1}\sum_{p=2}^{20}(x_p-\overline{x_p})\right\} \\ &= \frac{1}{|\mathbf{S}_{\overline{20},\overline{20}}|}\left\{\sum_{p=2}^{19}(x_p-\overline{x_p})\Delta_{\overline{20},p}+(x_1-\overline{x_1})\Delta_{\overline{20},1}\right\} \qquad (E22) \\ &= \frac{1}{|\mathbf{S}_{\overline{20},\overline{20}}|}\sum_{p=1}^{19}(x_p-\overline{x_p})\Delta_{\overline{20},p} = D^2(\mathbf{X}_{\overline{20}},\overline{\mathbf{X}}_{\overline{20}}) \end{aligned}$$

Following the same procedure, we can also prove that $D^2(\mathbf{X}_{\overline{1}},\overline{\mathbf{X}}_{\overline{1}}) = D^2(\mathbf{X}_{\overline{2}},\overline{\mathbf{X}}_{\overline{2}})$, $D^2(\mathbf{X}_{\overline{1}},\overline{\mathbf{X}}_{\overline{1}}) = D^3(\mathbf{X}_{\overline{3}},\overline{\mathbf{X}}_{\overline{3}})$, and so forth.

In the above proof, the reason $\overline{\mathbf{X}}$ was chosen as one of the end points of the distance is because it would be more intuitively to associate with the norm of a structural class that is the main topic of this paper. Actually, $\overline{\mathbf{X}}$ can be replaced by any point defined in the 20-D space, and the same conclusion can be reached by following exactly the same procedure as long as its components are constrained by the normalization condition as formulated in Equation E4.

It is obvious from the proof of Equation E10 that the Mahalanobis distance thus defined in a 19-D space must have a unique value, and it is none but the value calculated based on the 20-D space by means of the eigenvalue-eigenvector approach (see Equation 33).

Finally, the above conclusion can be extended to an m-D space (Chou, 1995a) as long as the corresponding m components are normalized, that is, constrained by

$$\sum_{i=1}^{m} x_i = 1, \qquad \sum_{i=1}^{m} x_{k,i} = 1 \tag{E23}$$

Appendix F

The elements of the inverse covariance matrices Q^{-1}_{α}, Q^{-1}_{β}, $Q^{-1}_{\alpha+\beta}$, and $Q^{-1}_{\alpha/\beta}$ in the 19-D space formed by removing the last amino acid (Y) component. These data were derived from the data $s_{i,j}$ (i=1, 2, ..., 19; j=1, 2, ..., 19) given in Appendix C by calling a subroutine DLINDS in the IMSL Library (Fortran Subroutines for Mathematics and Statistics).

$Q^{-1}_{\alpha} = [s^{-1}_{i,j}(\alpha)]$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	0.04	0.06	0.03	0.03	0.03	0.03	0.03	0.05	0.04	0.05	0.02	0.04	0.03	0.05	0.03	0.04	0.04	0.05	0.01
2	0.06	0.13	0.06	0.04	0.06	0.06	0.05	0.10	0.07	0.09	0.03	0.06	0.06	0.10	0.05	0.07	0.06	0.08	0.02
3	0.03	0.06	0.06	0.03	0.02	0.04	0.03	0.05	0.04	0.04	0.03	0.05	0.03	0.05	0.04	0.04	0.04	0.05	0.06
4	0.03	0.04	0.03	0.04	0.02	0.04	0.03	0.03	0.03	0.02	0.03	0.04	0.01	0.03	0.03	0.03	0.04	0.03	0.05
5	0.03	0.06	0.02	0.02	0.06	0.02	0.01	0.06	0.03	0.06	-0.01	0.01	0.04	0.07	0.02	0.03	0.01	0.04	-0.06
6	0.03	0.06	0.04	0.04	0.02	0.06	0.04	0.04	0.04	0.03	0.04	0.06	0.02	0.04	0.04	0.04	0.05	0.04	0.07
7	0.03	0.05	0.03	0.03	0.01	0.04	0.06	0.03	0.03	0.01	0.05	0.06	0.00	0.02	0.05	0.04	0.06	0.04	0.10
8	0.05	0.10	0.05	0.03	0.06	0.04	0.03	0.11	0.05	0.10	0.00	0.04	0.07	0.10	0.04	0.05	0.03	0.08	-0.05
9	0.04	0.07	0.04	0.03	0.03	0.04	0.03	0.05	0.05	0.05	0.03	0.04	0.03	0.05	0.04	0.04	0.05	0.05	0.04
10	0.05	0.09	0.04	0.02	0.06	0.03	0.01	0.10	0.05	0.10	-0.01	0.02	0.07	0.10	0.03	0.05	0.02	0.06	-0.07
11	0.02	0.03	0.03	0.03	-0.01	0.04	0.05	0.00	0.03	-0.01	0.08	0.06	-0.02	0.00	0.04	0.03	0.05	0.03	0.13
12	0.04	0.06	0.05	0.04	0.01	0.06	0.06	0.04	0.04	0.02	0.06	0.08	0.01	0.03	0.05	0.04	0.07	0.04	0.12
13	0.03	0.06	0.03	0.01	0.04	0.02	0.00	0.07	0.03	0.07	-0.02	0.01	0.08	0.07	0.02	0.03	-0.01	0.04	-0.06
14	0.05	0.10	0.05	0.03	0.07	0.04	0.02	0.10	0.05	0.10	0.00	0.03	0.07	0.11	0.03	0.06	0.03	0.07	-0.05
15	0.03	0.05	0.04	0.03	0.02	0.04	0.05	0.04	0.04	0.03	0.04	0.05	0.02	0.03	0.05	0.03	0.05	0.04	0.07
16	0.04	0.07	0.04	0.03	0.03	0.04	0.04	0.05	0.04	0.05	0.03	0.04	0.03	0.06	0.03	0.06	0.04	0.05	0.05
17	0.04	0.06	0.04	0.04	0.01	0.05	0.06	0.03	0.05	0.02	0.05	0.07	-0.01	0.03	0.05	0.04	0.09	0.04	0.10
18	0.05	0.08	0.05	0.03	0.04	0.04	0.04	0.08	0.05	0.06	0.03	0.04	0.04	0.07	0.04	0.05	0.04	0.07	0.01
19	0.01	0.02	0.06	0.05	-0.06	0.07	0.10	-0.05	0.04	-0.07	0.13	0.12	-0.06	-0.05	0.07	0.05	0.10	0.01	0.40

$Q^{-1}_{\beta} = [s^{-1}_{i,j}(\beta)]$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	0.05	0.02	0.06	0.01	0.02	0.02	-0.01	0.03	0.04	0.02	0.05	0.04	0.04	0.03	0.05	0.03	0.01	0.02	0.11
2	0.02	0.02	0.03	0.01	0.01	0.02	0.02	0.00	0.02	0.02	0.01	0.02	0.03	0.02	0.02	0.01	0.02	0.01	0.02
3	0.06	0.03	0.08	0.01	0.03	0.03	-0.01	0.04	0.05	0.03	0.05	0.06	0.06	0.05	0.06	0.05	0.01	0.03	0.13
4	0.01	0.01	0.01	0.02	0.00	0.01	0.01	0.00	0.01	0.01	0.01	0.02	0.02	0.01	0.01	0.01	0.01	0.00	0.01
5	0.02	0.01	0.03	0.00	0.03	0.01	-0.01	0.02	0.03	0.01	0.03	0.01	0.01	0.02	0.02	0.02	0.01	0.01	0.05
6	0.02	0.02	0.03	0.01	0.01	0.02	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.02	0.02	0.02	0.01	0.01	0.04
7	-0.01	0.02	-0.01	0.01	-0.01	0.01	0.08	-0.04	0.00	0.01	-0.04	0.01	0.01	0.02	0.00	-0.03	0.03	-0.01	-0.06
8	0.03	0.00	0.04	0.00	0.02	0.01	-0.04	0.05	0.03	0.01	0.05	0.01	0.01	0.01	0.03	0.04	-0.01	0.03	0.09
9	0.04	0.02	0.05	0.01	0.03	0.02	0.00	0.03	0.04	0.02	0.03	0.02	0.03	0.03	0.04	0.03	0.01	0.02	0.08
10	0.02	0.02	0.03	0.01	0.01	0.02	0.01	0.01	0.02	0.02	0.02	0.03	0.03	0.02	0.02	0.01	0.01	0.01	0.04
11	0.05	0.01	0.05	0.01	0.03	0.02	-0.04	0.05	0.03	0.02	0.09	0.03	0.03	0.02	0.05	0.05	0.00	0.02	0.12
12	0.04	0.02	0.06	0.02	0.01	0.02	0.01	0.01	0.02	0.03	0.03	0.06	0.05	0.04	0.04	0.02	0.02	0.02	0.07
13	0.04	0.03	0.06	0.02	0.01	0.03	0.01	0.01	0.03	0.03	0.03	0.05	0.07	0.04	0.04	0.02	0.02	0.02	0.08
14	0.03	0.02	0.05	0.01	0.02	0.02	0.02	0.01	0.03	0.02	0.02	0.04	0.04	0.05	0.04	0.01	0.02	0.02	0.06
15	0.05	0.02	0.06	0.01	0.02	0.02	0.00	0.03	0.04	0.02	0.05	0.04	0.04	0.04	0.06	0.04	0.02	0.03	0.09
16	0.03	0.01	0.05	0.01	0.02	0.02	-0.03	0.04	0.03	0.01	0.05	0.02	0.02	0.01	0.04	0.04	0.00	0.02	0.09
17	0.01	0.02	0.01	0.01	0.01	0.01	0.03	-0.01	0.01	0.01	0.00	0.02	0.02	0.02	0.02	0.00	0.02	0.00	0.00
18	0.02	0.01	0.03	0.00	0.01	0.01	-0.01	0.03	0.02	0.01	0.02	0.02	0.02	0.02	0.03	0.02	0.00	0.03	0.06
19	0.11	0.02	0.13	0.01	0.05	0.04	-0.06	0.09	0.08	0.04	0.12	0.07	0.08	0.06	0.09	0.09	0.00	0.06	0.30

$Q^{-1}_{\alpha+\beta} = [s^{-1}_{i,j}(\alpha+\beta)]$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	0.03	0.02	0.01	0.03	0.02	0.04	0.04	0.03	0.02	0.03	0.03	0.04	0.02	0.03	0.03	0.03	0.03	0.01	0.06
2	0.02	0.06	0.06	0.05	0.01	0.07	0.06	0.04	0.04	0.07	0.11	0.08	0.01	0.08	0.04	0.04	0.02	0.07	0.04
3	0.01	0.06	0.08	0.05	-0.01	0.07	0.05	0.04	0.04	0.07	0.12	0.08	0.01	0.09	0.04	0.05	0.01	0.07	0.03
4	0.03	0.05	0.05	0.06	0.01	0.07	0.07	0.04	0.04	0.07	0.10	0.08	0.02	0.07	0.04	0.06	0.03	0.05	0.06
5	0.02	0.01	-0.01	0.01	0.05	0.01	0.01	0.01	0.01	-0.01	-0.02	0.01	0.02	0.00	0.02	0.01	0.02	0.00	0.04
6	0.04	0.07	0.07	0.07	0.01	0.11	0.09	0.06	0.06	0.10	0.15	0.11	0.03	0.11	0.06	0.08	0.03	0.08	0.08
7	0.04	0.06	0.05	0.07	0.01	0.09	0.13	0.07	0.05	0.09	0.11	0.11	0.04	0.08	0.05	0.07	0.04	0.07	0.10
8	0.03	0.04	0.04	0.04	0.01	0.06	0.07	0.05	0.03	0.05	0.08	0.07	0.02	0.05	0.03	0.04	0.02	0.05	0.06
9	0.02	0.04	0.04	0.04	0.01	0.06	0.05	0.03	0.04	0.05	0.07	0.07	0.02	0.06	0.04	0.05	0.03	0.05	0.07
10	0.03	0.07	0.07	0.07	-0.01	0.10	0.09	0.05	0.05	0.11	0.15	0.11	0.02	0.11	0.06	0.07	0.03	0.09	0.07
11	0.03	0.11	0.12	0.10	-0.02	0.15	0.11	0.08	0.07	0.15	0.28	0.16	0.02	0.15	0.07	0.09	0.02	0.14	0.08
12	0.04	0.08	0.08	0.08	0.01	0.11	0.11	0.07	0.07	0.11	0.16	0.15	0.03	0.11	0.07	0.08	0.04	0.10	0.11
13	0.02	0.01	0.01	0.02	0.02	0.03	0.04	0.02	0.02	0.02	0.02	0.03	0.04	0.02	0.02	0.03	0.02	0.01	0.04
14	0.03	0.08	0.09	0.07	0.00	0.11	0.08	0.05	0.06	0.11	0.15	0.11	0.02	0.14	0.05	0.07	0.02	0.09	0.05
15	0.03	0.04	0.04	0.04	0.02	0.06	0.05	0.03	0.04	0.06	0.07	0.07	0.02	0.05	0.06	0.05	0.03	0.05	0.07
16	0.03	0.04	0.05	0.06	0.01	0.08	0.07	0.04	0.05	0.07	0.09	0.08	0.03	0.07	0.05	0.08	0.03	0.05	0.09
17	0.03	0.02	0.01	0.03	0.02	0.03	0.04	0.02	0.03	0.03	0.02	0.04	0.02	0.02	0.03	0.03	0.03	0.02	0.06
18	0.01	0.07	0.07	0.05	0.00	0.08	0.07	0.05	0.05	0.09	0.14	0.10	0.01	0.09	0.05	0.05	0.02	0.10	0.06
19	0.06	0.04	0.03	0.06	0.04	0.08	0.10	0.06	0.07	0.07	0.08	0.11	0.04	0.05	0.07	0.09	0.06	0.06	0.20

$Q^{-1}_{\alpha/\beta} = [s^{-1}_{i,j}(\alpha/\beta)]$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	0.22	0.26	0.24	0.22	0.24	0.24	0.25	0.32	0.20	0.14	0.12	0.34	0.34	0.30	0.36	0.23	0.33	0.22	0.20
2	0.26	0.69	0.27	0.23	0.18	0.26	0.27	0.39	0.25	0.20	0.09	0.49	0.48	0.40	0.43	0.26	0.41	0.25	0.28
3	0.24	0.27	0.30	0.25	0.26	0.27	0.32	0.35	0.23	0.15	0.16	0.38	0.37	0.34	0.40	0.28	0.36	0.25	0.24
4	0.22	0.23	0.25	0.26	0.24	0.26	0.27	0.30	0.22	0.16	0.12	0.35	0.35	0.31	0.37	0.26	0.34	0.22	0.18
5	0.24	0.18	0.26	0.24	0.31	0.25	0.28	0.33	0.21	0.13	0.16	0.31	0.31	0.30	0.36	0.26	0.33	0.26	0.21
6	0.24	0.26	0.27	0.26	0.25	0.29	0.28	0.36	0.23	0.17	0.13	0.39	0.39	0.34	0.42	0.26	0.38	0.24	0.23
7	0.25	0.27	0.32	0.27	0.28	0.28	0.41	0.36	0.24	0.15	0.17	0.39	0.36	0.37	0.40	0.29	0.35	0.27	0.28

8	0.32	0.39	0.35	0.30	0.33	0.36	0.36	0.50	0.29	0.22	0.17	0.50	0.52	0.45	0.53	0.33	0.49	0.33	0.32
9	0.20	0.25	0.23	0.22	0.21	0.23	0.24	0.29	0.21	0.14	0.11	0.32	0.32	0.28	0.35	0.23	0.31	0.20	0.18
10	0.14	0.20	0.15	0.16	0.13	0.17	0.15	0.22	0.14	0.14	0.06	0.23	0.25	0.20	0.24	0.16	0.22	0.15	0.12
11	0.12	0.09	0.16	0.12	0.16	0.13	0.17	0.17	0.11	0.06	0.14	0.18	0.15	0.17	0.19	0.13	0.17	0.14	0.15
12	0.34	0.49	0.38	0.35	0.31	0.39	0.39	0.50	0.32	0.23	0.18	0.59	0.56	0.48	0.58	0.36	0.53	0.33	0.32
13	0.34	0.48	0.37	0.35	0.31	0.39	0.36	0.52	0.32	0.25	0.15	0.56	0.60	0.49	0.58	0.37	0.53	0.33	0.31
14	0.30	0.40	0.34	0.31	0.30	0.34	0.37	0.45	0.28	0.20	0.17	0.48	0.49	0.46	0.50	0.32	0.46	0.31	0.31
15	0.36	0.43	0.40	0.37	0.36	0.42	0.40	0.53	0.35	0.24	0.19	0.58	0.58	0.50	0.63	0.39	0.56	0.36	0.33
16	0.23	0.26	0.28	0.26	0.26	0.26	0.29	0.33	0.23	0.16	0.13	0.36	0.37	0.32	0.39	0.28	0.35	0.24	0.20
17	0.33	0.41	0.36	0.34	0.33	0.38	0.35	0.49	0.31	0.22	0.17	0.53	0.53	0.46	0.56	0.35	0.54	0.33	0.30
18	0.22	0.25	0.25	0.22	0.26	0.24	0.27	0.33	0.20	0.15	0.14	0.33	0.33	0.31	0.36	0.24	0.33	0.25	0.23
19	0.20	0.28	0.24	0.18	0.21	0.23	0.28	0.32	0.18	0.12	0.15	0.32	0.31	0.31	0.33	0.20	0.30	0.23	0.32

Appendix G

The seed-propagated sampling principle as formulated in Section IV.A can be briefly explained as follows.

Because $\Re$ in Equation 42 is a random number within the range of 0 and 1, its density function can be expressed as (DeGroot, 1986):

$$f(\Re) = \begin{cases} 1, & \Re \in [0, 1] \\ 0, & \text{otherwise} \end{cases} \tag{G1}$$

The density function is a nonnegative function, and its meaning can be illustrated as follows. Assuming $0 \leq A < B \leq 1$, the probability of $\Re$ occurring in $[A, B]$ will be

$$P(A \leq \Re \leq B) = \int_A^B f(\Re) d\Re \tag{G2}$$

Therefore, the probability of $\Re$ occurring between $F_{k,i-1}(\alpha)$ and $F_{k,i}(\alpha)$ will be given by (see Equation 41)

$$\begin{aligned} P\left(F_{k,i-1}(\alpha) \leq \Re \leq F_{k,i}(\alpha)\right) &= \int_{F_{k,i-1}(\alpha)}^{F_{k,i}(\alpha)} f(\Re) d\Re = \int_{F_{k,i-1}(\alpha)}^{F_{k,i}(\alpha)} 1 \cdot d\Re = F_{k,i}(\alpha) - F_{k,i-1}(\alpha) \\ &= \sum_{j=1}^{i} x_{k,j}(\alpha) - \sum_{j=1}^{i-1} x_{k,j}(\alpha) = x_{k,i}(\alpha) \end{aligned} \tag{G3}$$

which indicates that the probability of $\Re$ occurring between $F_{k,i-1}(\alpha)$ and $F_{k,i}(\alpha)$ is exactly equal to $x_{k,i}(\alpha)$. Accordingly, amino acids drawn according to Equation 42 can be treated as discrete stochastic variables that satisfy the amino acid frequency distribution of the kth α-protein.

The sampling principle according to Equation 42 can also be understood through an intuitive geometrical illustration. Suppose that in Figure 4 the coordinates of the 20 amino acids A, C, D, E, ..., Y in the kth α protein are given by $\sum_{j=1}^{1} x_{k,j}(\alpha)$, $\sum_{j=1}^{2} x_{k,j}(\alpha)$, $\sum_{j=1}^{3} x_{k,j}(\alpha)$, $\sum_{j=1}^{4} x_{k,j}(\alpha)$, ..., $\sum_{j=1}^{20} x_{k,j}(\alpha)$, respectively. Thus, the distance between 0 and A is $x_{k,1}(\alpha)$, that between A and C is $x_{k,2}(\alpha)$, that between C and D is $x_{k,3}(\alpha)$, that between C and D is $x_{k,4}(\alpha)$, and so forth. Generally speaking, the distance between the $(i-1)$th and the ith amino acids is $x_{k,i}(\alpha)$ (Figure 4). Because the random number $\Re$ is evenly distributed, the probability of finding it in the range between $\sum_{j=1}^{i-1} x_{k,j}(\alpha)$ and $\sum_{j=1}^{i} x_{k,j}(\alpha)$ must be proportional to the range's length $x_{k,i}(\alpha)$, that is, the occurrence frequency of the ith amino acid of the kth α-protein in the training database. This is exactly what is imposed by Equation 42.

Furthermore, according to the central limit theorem (DeGroot, 1986), the deviation of the amino acid compositions of the simulated proteins thus sampled from those of the proteins in the training database should be proportional to $1/\sqrt{N_S}$, where N_S is the number of the total subsampling cycles (see steps 2 and 3 of Section IV.A1). The larger the N_S, the smaller the deviation. In the study reported here, $N_S = 10^4$, and hence the corresponding deviation is on the order of 1/100; hence, the simulated proteins thus generated are very close but not identical to the proteins in the training database in the sense of amino acid composition.

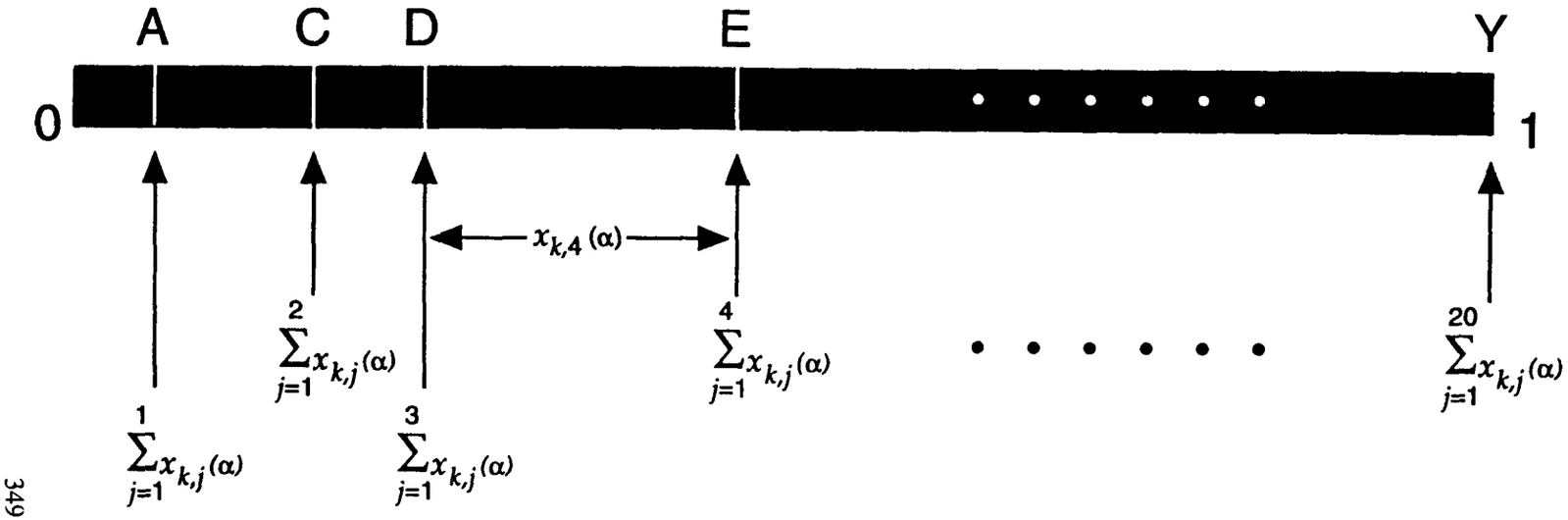

Figure 4: An intuitive geometric drawing illustrating the principles of seed-propagated sampling as described in Section IV.A. Suppose that the 1-D coordinates of amino acids A, C, D, E, . . . , Y of the kth α-protein in the training database are successively given by $\sum_{j=1}^{1} x_{k,j}(\alpha)$, $\sum_{j=1}^{2} x_{k,j}(\alpha)$, $\sum_{j=1}^{3} x_{k,j}(\alpha)$, $\sum_{j=1}^{4} x_{k,j}(\alpha)$, . . . , $\sum_{j=1}^{20} x_{k,j}(\alpha)$, respectively. Thus, the distance between 0 and A is $x_{k,1}(\alpha)$, that between A and C is $x_{k,2}(\alpha)$, that between C and D is $x_{k,3}(\alpha)$, that between D and E is $x_{k,4}(\alpha)$, and so forth. Generally speaking, the distance between the $(i-1)$th and the ith amino acids in the 1-D space thus defined is $x_{k,i}(\alpha)$. Therefore, a simulated protein generated according to Equation 42 must statistically "inherit" the intrinsical feature of a "seed" protein in the sense of amino acid composition.