PROTEINS: Structure, Function, and Genetics 21:319–344 (1995)

# A Novel Approach to Predicting Protein Structural Classes in a (20−1)-D Amino Acid Composition Space

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

***ABSTRACT*** **The development of prediction methods based on statistical theory generally consists of two parts: one is focused on the exploration of new algorithms, and the other on the improvement of a training database. The current study is devoted to improving the prediction of protein structural classes from both of the two aspects. To explore a new algorithm, a method has been developed that makes allowance for taking into account the coupling effect among different amino acid components of a protein by a covariance matrix. To improve the training database, the selection of proteins is carried out so that they have (1) as many nonhomologous structures as possible, and (2) a good quality of structure. Thus, 129 representative proteins are selected. They are classified into 30 α, 30 β, 30 α + β, 30 α/β, and 9 ζ (irregular) proteins according to a new criterion that better reflects the feature of the structural classes concerned. The average accuracy of prediction by the current method for the 4 × 30 regular proteins is 99.2%, and that for 64 independent testing proteins not included in the training database is 95.3%. To further validate its efficiency, a jackknife analysis has been performed for the current method as well as the previous ones, and the results are also much in favor of the current method. To complete the mathematical basis, a theorem is presented and proved in Appendix A that is instructive for understanding the novel method at a deeper level.**
© 1995 Wiley-Liss, Inc.

**Key words: α protein, β protein, α + β protein, α/β protein, Mahalanobis distance, seed-propagated sampling, jackknife analysis**

## INTRODUCTION

The function of a protein is based on its structure. The 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 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. Besides, during the working process there may be some difficulties which 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 empirical atomic potentials.[11,23,26,31,34,38,44–47] The other is the statistical method, which was developed based on various statistical data extracted from structure-known proteins.[14,21,22,33] Although by means of the free energy minimization method, one can in principle calculate the detailed atomic coordinates, it is in practice very difficult to find the real global minimum owing to the notorious local minimum problem. Therefore, in most of 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.[3,5,9–11] Although the simulated annealing approach is a powerful tool in overcoming the local minimum problem,[2,4,27,48] using it to predict the conformation of a protein is still impractical. A feasible approach is a combination of the energy minimization and heuristic approach.[1,6,16] However, there are not many proteins from which one can get enough heuristic inputs to result in a successful prediction. On the other hand, the statistical method can predict only an outline of a structure. Nevertheless, the statistical method 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 3-D (dimensional) structure of a protein, the statistical method would play an important complementary role.

This article will be focused on developing a new statistical method that can be used to predict the structural class of a protein.

According to the percentage of secondary struc-

Received September 26, 1994; revision accepted December 8, 1994.
Address reprint requests to Kuo-Chen Chou, Computer-Aided Drug Discovery, Upjohn Laboratories, Kalamazoo, MI 49007-4940.

© 1995 WILEY-LISS, INC.

ture components, proteins of known structures are generally categorized into the following five structural classes: α, β, α+β, α/β, and irregular proteins.[32,43] It is also known that the structural class of a protein is correlated with its amino acid composition.[12,13,36,41] Furthermore, the knowledge of protein structural classes has proven to be useful in improving the prediction of the secondary structures in proteins.[18] However, given the amino acid composition of a protein, how may one predict its structural class? Although various methods were proposed for this problem,[7,12,13,19,28–30,39,41,49] none of them has taken into account the coupling effect among different amino acid components. Actually, in those methods, each of the components of the 20 amino acids was treated as an independent variable, and the contribution owing to the coupling among the different components is completely neglected. Such a treatment is merely a zero-order approximation. Thus, we are confronted with the following problems. Through what avenue can the coupling be effectively taken into account? How should the method be formulated to make the quantitative calculation feasible? How significantly will the predicted results be improved? This paper explores these problems.

## METHOD

According to its amino acid composition, a protein molecule can be represented by a point or a vector[7] in a 20-D space, the so-called composition space.[41] Accordingly, the similarity of any two protein molecules can be reflected through their distance in the 20-D space. However, it is a subtle problem how to define the distance as an effective scale to measure their similarity.

### Distance and Similarity

Suppose there is a set of $N$ proteins, each of which corresponds to a point in the 20-D space, as can be formulated by

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

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

$$\overline{\mathbf{X}} = \begin{bmatrix} \overline{x_1} \\ \overline{x_2} \\ \vdots \\ \overline{x_{20}} \end{bmatrix} \tag{2}$$

where

$$\overline{x_i} = \frac{1}{N}\sum_{k=1}^{N} x_{k,i} \quad (i = 1, 2, \ldots, 20) \tag{3}$$

The distance between the norm $\overline{\mathbf{X}}$ and any point $\mathbf{X}(x_1, x_2, \ldots, x_{20})$ in the 20-D space is ordinarily defined by

$$d(\mathbf{X},\overline{\mathbf{X}}) = \left[\sum_{i=1}^{20} | x_i - \overline{x_i}|^q\right]^{1/q} \tag{4}$$

The distance formulated above is called the Minkowski distance. When $q = 1$, Eq. (4) reduces to the Hamming distance

$$d(\mathbf{X},\overline{\mathbf{X}}) = \left[\sum_{i=1}^{20} | x_i - \overline{x_i} |\right] \tag{5}$$

which is used by Chou[12,13] as a measure to predict the protein structural class. When $q = 2$, Eq. (4) reduces to the Euclidian distance

$$d(\mathbf{X},\overline{\mathbf{X}}) = \left[\sum_{i=1}^{20} | x_i - \overline{x_i} |^2\right]^{1/2} \tag{6}$$

which is used by Nakashima et al.[41] as a measure to predict the protein structural class.

The distances defined by Eqs. (4)–(6) have a common merit, i.e., simple and intuitive, but they also bear the following weaknesses. First, the 20-D composition space is generally not an orthogonal space. Is it valid to extend the definition of the Euclidian distance in a 3-D orthogonal space to a 20-D nonorthogonal space? Second, or more importantly, the coupling among different components on the distance are completely neglected. This might be appropriate when the problem considered is a pure geometric 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 would become important.[8] To incorporate the coupling effect, we resort to the Mahalanobis distance,[35,42] which, as shown below, will take into account the interactions among different amino acid components.

First, let us introduce a covariance matrix given by

$$\mathbf{S} = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,20} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ s_{20,1} & s_{20,2} & \cdots & s_{20,20} \end{bmatrix} \tag{7}$$

where

$$s_{i,j} = \sum_{k=1}^{N} [x_{k,i} - \overline{x_i}]\,[x_{k,j} - \overline{x_j}], \quad (i,j = 1, 2, \ldots, 20) \tag{8}$$

Thus, the Mahalanobis distance, $D^2$ $(\mathbf{X},\overline{\mathbf{X}})$, between the norm defined by Eq. (2) and any point $\mathbf{X}$ in the 20-D space is given by[35,42]

$$D^2(\mathbf{X},\overline{\mathbf{X}}) = (\mathbf{X} - \overline{\mathbf{X}})^T\mathbf{S}^{-1}(\mathbf{X} - \overline{\mathbf{X}}) \quad (9)$$

where $T$ is the transposition operator, and $\mathbf{S}^{-1}$ is the inverse matrix of $\mathbf{S}$ given by Eq. (7). When the nondiagonal elements in $\mathbf{S}$ are zero, the distance defined by eq. (9) will reduce to the weighted Euclidian distance. Actually, $\mathbf{S}$ is a symmetric matrix whose nondiagonal elements are generally not zero. It is through these nondiagonal terms that the coupling effects among different amino acid components are incorporated.

### Mahalanobis Distance Defined in a (20−1)D Space

Since the amino acid composition must be normalized, i.e., constrained by

$$\sum_{i=1}^{20} x_{k,i} = 1 \quad (k = 1, 2, \ldots, N) \quad (10)$$

it follows [cf. Eq. (8)]that

$$\begin{cases} \sum_{j=1}^{20} s_{i,j} = 0, & (i = 1, 2, \ldots, 20) \\ \sum_{i=1}^{20} s_{i,j} = 0, & (j = 1, 2, \ldots, 20) \end{cases} \quad (11)$$

Therefore, $\mathbf{S}$ defined by Eq. (8) is a singular matrix. This indicates that its inverse matrix $\mathbf{S}^{-1}$ actually does not exist, and the Mahalanobis distance formulated by Eq. (9) would completely meaningless. To overcome such a difficulty, let us approach the problem by the following consideration. It is shown from Eq. (10) that of the 20 amino acid components of a protein only 19 are independent. Therefore, by leaving out any 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, then 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 $k$th protein in a given protein set can be expressed by

$$\mathbf{P}_k = \begin{bmatrix} x_{k,1} \\ x_{k,2} \\ \vdots \\ x_{k,19} \end{bmatrix} \quad (k = 1, 2, \ldots, N) \quad (12)$$

where $x_{k,1}, x_{k,2}, \ldots, x_{k,19}$ are the same as in Eq. (1). The only difference between $\mathbf{X}_k$ of Eq. (1) and $\mathbf{P}_k$ of Eq. (12) is that the former contains the component $x_{k,20}$ but the latter not. The norm of the protein set in the 19-D space is defined by

$$\overline{\mathbf{P}} = \begin{bmatrix} \overline{x_1} \\ \overline{x_2} \\ \vdots \\ \overline{x_{19}} \end{bmatrix} \quad (13)$$

where $\overline{x}_i$ $(i = 1, 2, \ldots, 19)$ are the same as in Eq. (2) except $\overline{x}_{20}$, which is not a part of Eq. (13).

Accordingly, the covariance matrix $\mathbf{S}$ of Eq. (7) would reduce to a 19 × 19 covariance matrix given by

$$\mathbf{Q} = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,19} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,19} \\ \vdots & \vdots & \ddots & \vdots \\ s_{19,1} & s_{19,2} & \cdots & s_{19,19} \end{bmatrix} \quad (14)$$

where $s_{i,j}$ is the same as in Eq. (7), as defined by Eq. (8). 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

$$D^2(\mathbf{P},\overline{\mathbf{P}}) = (\mathbf{P} - \overline{\mathbf{P}})^T\mathbf{Q}^{-1}(\mathbf{P} - \overline{\mathbf{P}}) \quad (15)$$

where $T$ is the transposition operator, and $\mathbf{Q}^{-1}$ is the inverse matrix of $\mathbf{Q}$ given by Eq. (14).

A question might be posed. By leaving out a different amino acid component and hence the Mahalanobis distance being defined in a different 19-D space, will it 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 understand the details fully are referred to Appendix A, where a relevant theorem and its mathematical proof are provided, since to the best of the author's knowledge no existing mathematical theorems cover the problem.

### The Least Mahalanobis Distance Principle

When the $N$ proteins in Eq. (12) are all α proteins, the $\overline{\mathbf{P}}$ defined by Eq. (13) will become $\overline{\mathbf{P}}_\alpha$, the norm for an α protein set, the $\mathbf{Q}$ defined by Eq. (14) will become $\mathbf{Q}_\alpha$, the covariance matrix for the α protein set, and $D^2$ $(\mathbf{P},\overline{\mathbf{P}})$ will become $D^2(\mathbf{P},\overline{\mathbf{P}}_\alpha)$, the Mahalanobis distance between the protein $\mathbf{P}$ and the norm of the α protein set. Likewise, when the $N$ proteins in Eq. (12) are all β, α+β, α/β, or ζ (irregular) proteins, then the corresponding $\overline{\mathbf{P}}$ will become the norm of the β, α+β, α/β, or ζ protein set, denoted by $\mathbf{P}_\beta$, $\mathbf{P}_{\alpha+\beta}$, $\mathbf{P}_{\alpha/\beta}$, or $\mathbf{P}_\zeta$, respectively. The corresponding $\mathbf{Q}$ will then become the covariance matrix for the β, α+β, α/β, or ζ protein set, denoted by $\mathbf{Q}_\beta$, $\mathbf{Q}_{\alpha+\beta}$, $\mathbf{Q}_{\alpha/\beta}$, or $\mathbf{Q}_\zeta$, and hence the corresponding Mahalanobis distance $D^2(\mathbf{P},\overline{\mathbf{P}})$ will become $D^2(\mathbf{P},\overline{\mathbf{P}}_\beta)$, $D^2(\mathbf{P},\overline{\mathbf{P}}_{\alpha+\beta})$, $D^2(\mathbf{P},\overline{\mathbf{P}}_{\alpha/\beta})$, or $D^2(\mathbf{P},\overline{\mathbf{P}}_\zeta)$, respectively. Thus, the Mahalanobis distance between any protein $\mathbf{P}$ and each of the norms for the five structural classes can be generally formulated as follows:

$$D^2(\mathbf{P}, \mathbf{P}_\lambda) = \begin{cases} (\mathbf{P} - \overline{\mathbf{P}}_\alpha)^T\mathbf{Q}_\alpha^{-1}(\mathbf{P} - \overline{\mathbf{P}}_\alpha), & \text{if } \lambda = \alpha \\ (\mathbf{P} - \overline{\mathbf{P}}_\beta)^T\mathbf{Q}_\beta^{-1}(\mathbf{P} - \overline{\mathbf{P}}_\beta), & \text{if } \lambda = \beta \\ (\mathbf{P} - \overline{\mathbf{P}}_{\alpha+\beta})^T\mathbf{Q}_{\alpha+\beta}^{-1}(\mathbf{P} - \overline{\mathbf{P}}_{\alpha+\beta}), & \text{if } \lambda = \alpha + \beta \\ (\mathbf{P} - \overline{\mathbf{P}}_{\alpha/\beta})^T\mathbf{Q}_{\alpha/\beta}^{-1}(\mathbf{P} - \overline{\mathbf{P}}_{\alpha/\beta}), & \text{if } \lambda = \alpha/\beta \\ (\mathbf{P} - \overline{\mathbf{P}}_\zeta)^T\mathbf{Q}_\zeta^{-1}(\mathbf{P} - \overline{\mathbf{P}}_\zeta), & \text{if } \lambda = \zeta \end{cases} \quad (16)$$

When $D^2(\mathbf{P},\overline{\mathbf{P}}_\lambda)$ ($\lambda = \alpha, \beta, \alpha+\beta, \alpha/\beta$, or $\zeta$) is smaller, meaning that the protein **P** is closer to the $\lambda$ protein set, and hence the likelihood of it belonging to the $\lambda$ protein set is higher, and vice versa. Thus, the protein **P** will be predicted to be the structural class for which $D^2$ has the least value, as can be formulated as follows. Suppose

$$D^2(\mathbf{P}, \overline{\mathbf{P}}_\lambda) = \mathbf{Min}\{D^2(\mathbf{P}, \overline{\mathbf{P}}_\alpha), D^2(\mathbf{P}, \overline{\mathbf{P}}_\beta), D^2(\mathbf{P}, \overline{\mathbf{P}}_{\alpha+\beta}), D^2(\mathbf{P}, \overline{\mathbf{P}}_{\alpha/\beta}), D^2(\mathbf{P}, \overline{\mathbf{P}}_\zeta)\} \quad (17)$$

where $\lambda$ can be $\alpha$, $\beta$, $\alpha+\beta$, $\alpha/\beta$, or $\zeta$, and the operator **Min** means taking the least among those in parentheses; the subscript $\lambda$ of Eq. (17) will then give the structural class to which the predicted protein **P** should belong.

### Criteria of Assigning the Structural Class to a Protein

Although proteins of known structure are generally classified into one of the five structural classes, $\alpha$, $\beta$, $\alpha+\beta$, $\alpha/\beta$, and $\zeta$ (irregular) proteins,[32,43] there is no unified quantitative measure for making such a classification. Suppose the percentages of $\alpha$-helix and $\beta$-sheet in a protein are abbreviated by $\alpha$ and $\beta$, respectively. The classification by Nakashima et al.[41] was made according to the following criterion: $\alpha$ proteins, $\alpha > 15\%$, $\beta < 10\%$; $\beta$ proteins, $\alpha < 15\%$, $\beta > 10\%$; $\alpha + \beta$ proteins, $\alpha > 15\%$, $\beta > 10\%$ with dominantly antiparallel $\beta$-sheets; $\alpha/\beta$ proteins, $\alpha > 15\%$, $\beta > 10\%$ with dominantly parallel $\beta$-sheets; and $\zeta$ (irregular) proteins, $\alpha < 15\%$, $\beta < 10\%$. According to Chou,[13] however, proteins were classified as follows: $\alpha$ proteins, $\alpha > 45\%$, $\beta < 5\%$; $\beta$ proteins, $\alpha < 5\%$, $\beta > 45\%$; $\alpha + \beta$ proteins, $\alpha > 30\%$, $\beta > 20\%$ with dominantly antiparallel $\beta$-sheets; $\alpha/\beta$ proteins, $\alpha > 30\%$, $\beta > 20\%$ with dominantly parallel $\beta$-sheets. The classification by Nakashima et al.[41] covered 135 proteins: 31 $\alpha$, 34 $\beta$, 27 $\alpha+\beta$, 39 $\alpha/\beta$, and 4 irregular proteins. The classification by Chou[13] covered 64 proteins, 19 $\alpha$, 15 $\beta$, 14 $\alpha+\beta$, and 16 $\alpha/\beta$, but no irregular proteins. Although the classification by Nakashima et al.[41] covers more proteins than those by Chou,[13] the relevant percentages set by them for $\alpha$ proteins ($\alpha > 15\%$) and $\beta$ proteins ($\beta > 10\%$) do not seem large enough to reflect the real features of the two structural classes. In other words, an $\alpha$ or $\beta$ protein should at least have $\alpha \geq 40\%$ or $\beta \geq 40\%$, respectively. Besides, no quantitative definition 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 $\alpha/\beta$ proteins. For example, proteins with PDB codes 0PHH, 1LDX, 2MDH, 2GPD, 2GRS, 4ADH, and 4LDH were classified by Nakashima et al.[41] as the $\alpha/\beta$ structural class. However, an analysis of the secondary structure contents for these proteins indicates that of their $\beta$-sheet component the parallel sheets occupy only a percentage ranging from 25 to 50%. Obviously, for these proteins the parallel $\beta$-sheet should not be interpreted as a "dominant" part over its antiparallel counterpart. A similar problem has also been found for the $\alpha/\beta$ proteins classified by Chou.[13]

In view of these, a new method is proposed that classifies proteins according to the following quantitative criterion:

$$\begin{cases} \alpha \text{ proteins} & \Rightarrow \alpha \geq 40\%, \beta \leq 5\%, \\ \beta \text{ proteins} & \Rightarrow \alpha \leq 5\%, \beta \geq 40\%, \\ \alpha + \beta \text{ proteins} & \Rightarrow \alpha \geq 15\%, \beta \geq 15\%, \\ & \quad \text{with more than 60\% antiparallel } \beta\text{-sheets}, \\ \alpha / \beta \text{ proteins} & \Rightarrow \alpha \geq 15\%, \beta \geq 15\%, \\ & \quad \text{with more than 60\% parallel } \beta\text{-sheets}, \\ \zeta \text{ proteins} & \Rightarrow \alpha \leq 10\%, \beta \leq 10\% \end{cases} \quad (18)$$

where the contents of protein secondary structures were computed based on the dictionary by Kabsch and Sander.[25]

In order to reduce redundancy in the Protein Data Bank of 3-D protein structures, which is caused by many homologous proteins in the data bank, most of proteins to be classified in this paper are selected from a list of protein chains with less than 25% sequence identity.[24] Besides, the selection is made from those proteins with good quality of structural determination.

According to the new criteria and selection principle, for each of the four regular structural classes 30 representative proteins have been found, together with 9 proteins for the irregular structural class. The $4 \times 30 + 9 = 129$ proteins with their PDB codes as well as amino acid compositions are given in Appendix B since they are often requested by readers and they will serve as a training database for the new prediction algorithm.

## RESULTS AND DISCUSSION

The predictions by the new 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 test for checking the self-consistency of the new algorithm, while that for the latter it is a cross-validation test for checking its extrapolating effectiveness. A valid new algorithm should give a better result in both of these two aspects.

### Predicted Results for the Training Set of Proteins (a Resubstitution Test)

Based on the data of the 30 representative $\alpha$ proteins (Appendix B) the norm $\mathbf{P}_\alpha$ and the covariance matrix $\mathbf{Q}_\alpha$ have been calculated by Eqs. (3) and (13), and Eqs. (8) and (14), respectively. Based on $\mathbf{Q}_\alpha$ the inverse covariance matrix $\mathbf{Q}_\alpha^{-1}$ can be computed by calling a subroutine DLINDS in the IMSL Library (Fortran Subroutines for Mathematics and Sciences). Similarly, based on the data of the 30 representative $\beta$ proteins, 30 representative $\alpha + \beta$ proteins, and 30 representative $\alpha/\beta$ proteins, the

**TABLE I. The Predicted Results* for the 30 α Proteins in the Training Database**

| PDB code of the 30 α proteins† | $D^2(\mathbf{P}, \bar{\mathbf{P}}_\alpha)$ | $D^2(\mathbf{P}, \bar{\mathbf{P}}_\beta)$ | $D^2(\mathbf{P}, \bar{\mathbf{P}}_{\alpha+\beta})$ | $D^2(\mathbf{P}, \bar{\mathbf{P}}_{\alpha/\beta})$ | The predicted structural class |
|---|---|---|---|---|---|
| 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* | 5.63 | 1.65 | 3.06 | 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% | | | | | |

*The prediction was performed based on Eq. (17). The one with the least value of $D^2$ (marked by an asterisk) is assumed to correspond to the structural class for the predicted protein.
†The PDB (Protein Data Bank) code is constituted by the first four characters according to Brookhaven National Laboratory, and the fifth character used here to indicate a specific chain of a protein. If the fifth character is –, it means the corresponding protein has only one chain. The amino acid compositions of the protein chains listed here are given in Appendix B.

corresponding norms, covariance matrices, and inverse covariance matrices have been calculated. The irregular proteins have been left out in this study because their number is only nine, too small to form a good set of statistical data. 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, α + β, or α/β protein set can be uniquely defined in the 19-D space [cf. Eq. (15)].

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 α + β proteins, and 30 α/β proteins is listed in Tables I–IV. For providing an intuitive feeling, and also for facilitating the statistical analysis later, a 3-D histogram is depicted to illustrate the Mahalanobis distances for the 30 proteins in each of the four structural classes (Fig. 1a–d). As we can see from Table I or Figure 1a, $D^2(\mathbf{P},\bar{\mathbf{P}}_\alpha)$ has the least values for all the proteins listed in Table I. Therefore, according to Eq. (17), all of the 30 α proteins are correctly predicted to be the α structural class. These results indicate that the rate of correct prediction for the α proteins is 100%. Tables II and III or Figure 1b and c indicate that the rates of correct prediction for the 30 β and 30 α + β proteins are also 100%! Table IV or Figure 1d indicates that, of the 30 α/β proteins, 29 have the least values for $D^2(\mathbf{P},\bar{\mathbf{P}}_{\alpha/\beta})$, and hence the rate of correct prediction for the α/β proteins is 29/30 = 96.7%.

To provide an overview, a summary of the predicted accuracies for the four structural class proteins is given in Table V, together with the predicted accuracies for the same database by the least Euclidian distance algorithm[41] and the least Hamming distance algorithm,[12,13] respectively. As shown in Table V, the current algorithm yields an average accuracy of 99.2%, much higher than those by the others. The prediction of the α + β proteins by means of the ordinary geometric distances had been difficult. As reported by Nakashima et al.,[41] the rate of correct prediction for the α + β class was only 37.0% if calculated with their database. Nakashima et al.[41] attributed the low accuracy to the fact that the α + β 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 V, the rate of correct prediction for the α + β class by either the least Hamming distance algorithm[13] or the least Euclidian distance algorithm[41] is still smaller than 47%. However, by means of the new algorithm, the rate of correct prediction for the α + β 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.

### Predicted Results for the Testing Set of Proteins (a Cross-Validation Test)

As a cross-validation test, predictions have also been performed for a set of 64 independent testing proteins which are 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 VI, which indicates that an average accuracy of 61/64 = 95.3% is obtained. Also,

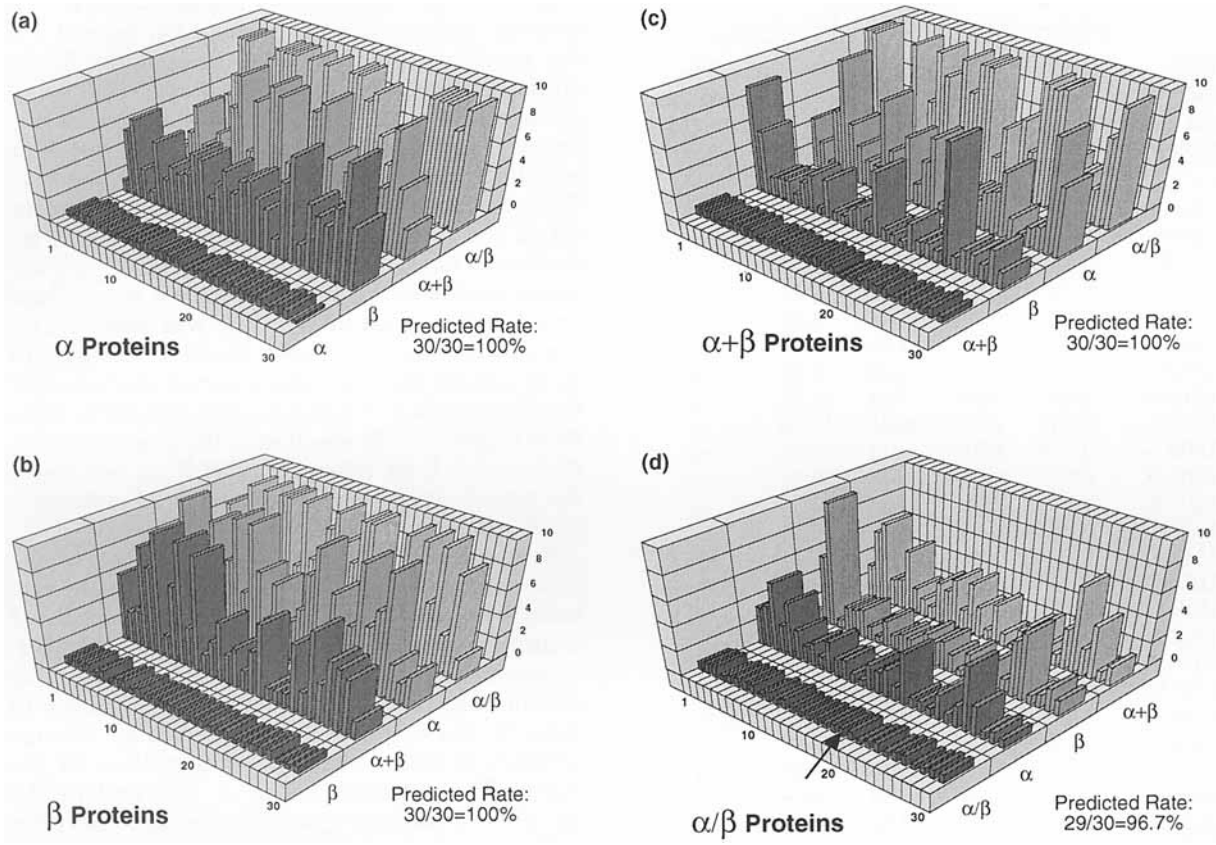


Fig. 1. The 3-D histogram to show the Mahalanobis distance from each of (a) the 30 α proteins, (b) the 30 β proteins, (c) the 30 α + β proteins, and (d) the 30 α/β proteins to the norms of α, β, α + β and α/β (Appendix C1), respectively, derived from the training database of Appendix B. The proteins in each class are arranged from left to right along the abscissa according to their order in Tables 1–4, respectively. The Mahalanobis distance is shown by the ordinate. Note that any distances with $D^2 > 10$ are cut down to 10. The arrow in panel (d) indicates the only protein incorrectly predicted.

based on the same training database, the 64 testing proteins have been predicted by the least Euclidian distance algorithm[41] and the least Hamming distance algorithm,[12,13] respectively. The average accuracies thus obtained dramatically drop to 36/64 = 56.3% (Table VI). Note that although their average accuracies are the same, the result predicted for each individual protein in the testing set by one algorithm is not necessarily the same as that by the other.

Therefore, the results obtained from both the resubstitution test and cross-validation test indicate that the current algorithm is much more accurate than the previous ones even if the tests are performed based on the same training and same testing data.

## STATISTICAL ANALYSIS

Even if a comparison is made based on a completely 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 accuracies for the testing set data obtained by ordinary treatments lack 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 anything 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, i.e., a statistically complete one to be able to represent all the testing proteins concerned. Unfortunately, so far 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. In view of this, a further comparison among these algorithms has been performed in terms of the jackknife test[20] and the simulated accuracy,[50] respectively.

**TABLE II. The Predicted Results* for the 30 β Proteins in the Training Database**

| PDB code of the 30 β proteins† | $D^2(\mathbf{P}, \bar{\mathbf{P}}_\alpha)$ | $D^2(\mathbf{P}, \bar{\mathbf{P}}_\beta)$ | $D^2(\mathbf{P}, \bar{\mathbf{P}}_{\alpha+\beta})$ | $D^2(\mathbf{P}, \bar{\mathbf{P}}_{\alpha/\beta})$ | The predicted structural class |
|---|---|---|---|---|---|
| 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% | | | | | |

*† See corresponding footnotes to Table I.

**TABLE III. The Predicted Results* for the 30 α + β Proteins in the Training Database**

| PDB code of the 30 α + β proteins† | $D^2(\mathbf{P}, \bar{\mathbf{P}}_\alpha)$ | $D^2(\mathbf{P}, \bar{\mathbf{P}}_\beta)$ | $D^2(\mathbf{P}, \bar{\mathbf{P}}_{\alpha+\beta})$ | $D^2(\mathbf{P}, \bar{\mathbf{P}}_{\alpha/\beta})$ | The predicted structural class |
|---|---|---|---|---|---|
| 1AAK– | 4.31 | 8.77 | 0.79* | 13.19 | α + β |
| 1CTF– | 2.45 | 5.38 | 0.76* | 22.03 | α + β |
| 1DNKA | 0.99 | 1.73 | 0.78* | 3.38 | α + β |
| 1EAF– | 5.69 | 1.88 | 0.50* | 2.30 | α + β |
| 1HSBA | 10.95 | 1.91 | 0.71* | 3.09 | α + β |
| 1LTSA | 5.01 | 1.02 | 0.72* | 2.97 | α + β |
| 1LTSD | 3.37 | 2.33 | 0.79* | 10.91 | α + β |
| 1NRCA | 2.32 | 3.20 | 0.78* | 2.03 | α + β |
| 1OVB– | 2.14 | 0.64 | 0.38* | 5.72 | α + β |
| 1POC– | 7.85 | 3.33 | 0.67* | 17.26 | α + β |
| 1PPN– | 3.92 | 0.87 | 0.57* | 8.24 | α + β |
| 1PRF– | 2.37 | 0.79 | 0.69* | 2.83 | α + β |
| 1RND– | 5.57 | 1.62 | 0.57* | 21.79 | α + β |
| 1SNC– | 1.87 | 2.04 | 0.41* | 7.58 | α + β |
| 1TFG– | 4.05 | 1.20 | 0.63* | 33.41 | α + β |
| 1TGSI | 6.22 | 5.61 | 0.92* | 87.89 | α + β |
| 2ACHA | 0.89 | 1.04 | 0.77* | 2.87 | α + β |
| 2ACT– | 5.99 | 0.87 | 0.65* | 5.53 | α + β |
| 2BPA1 | 1.25 | 0.63 | 0.24* | 2.31 | α + β |
| 2SNS– | 2.67 | 2.42 | 0.47* | 6.73 | α + β |
| 2SSI– | 2.38 | 1.21 | 0.49* | 4.31 | α + β |
| 3IL8– | 3.31 | 1.35 | 0.84* | 19.27 | α + β |
| 3RUBS | 3.67 | 1.06 | 0.75* | 9.79 | α + β |
| 3SGBI | 5.54 | 11.04 | 0.76* | 68.52 | α + β |
| 3SICI | 1.96 | 1.12 | 0.41* | 4.49 | α + β |
| 4BLMA | 0.92 | 2.05 | 0.57* | 3.58 | α + β |
| 4TMS– | 1.77 | 1.97 | 0.56* | 3.94 | α + β |
| 8CATA | 2.09 | 1.45 | 0.47* | 2.96 | α + β |
| 9RNT– | 9.58 | 2.99 | 0.87* | 7.44 | α + β |
| 9RSAA | 5.95 | 1.42 | 0.50* | 21.73 | α + β |
| Rate of correct prediction = 30/30 = 100% | | | | | |

*,†See corresponding footnotes to Table I.

### Jackknife Test

The jackknife test, also called leave-one-out test,[29,37] is a method often used for small samples which cannot be divided into training and testing sets without serious loss of information. Here each protein is in turn classified by the rules derived using all other proteins in a given database except the one which is being classified. However, owing to the definition of the covariance matrix [See Eq. (14)], in order to make a statistically meaningful leave-one-out test for the current method, the number of proteins in each structural class should be at least greater than 38. In view of this, 40 simulated proteins have been generated for each of the four structural classes by the seed-propagated sampling,[51] one quite similar to the bootstrap method,[20,40] based on the 4 × 30 proteins in the training database (Appendix B). The details of the resampling scheme are described in Zhang and Chou,[51] and the number of samples selected for each protein is $10^4$. For a database of such 4 × 40 = 160 proteins, a leave-one-out test has been performed. It has been found that the average rate of correct prediction by the current algorithm is 85.0%, and those by the least Hamming distance algorithm[12,13] and the least Euclidian distance algorithm[41] are 62.5 and 68.8%, respectively. Interestingly, if the leave-one-out test is carried out for a database of 4 × 80 = 320 simulated proteins, the average rate of correct prediction by the current algorithm will increase to 94.4%, while those by the other two algorithms remain almost unvaried, fluctuating around 62 and 68%, respectively, with an amplitude less than 1.5%. This indicates that the jackknife tested rate by the current method is not only much higher than those by the others, but the potential of increasing its rate by improving the database is also much greater.

### Simulated Accuracy

As demonstrated in a previous paper,[50] when the number of simulated proteins generated by Monte

Carlo sampling for each class is greater than 3000, the rate of correct prediction would gradually approach a limit, the so-called *asymptotical limit*. Under such a circumstance, the errors due to statistical fluctuations could be omitted. Such an asymptotical limit was defined as the simulated *accuracy of prediction*. However, during the sampling process as described in the previous paper,[50] a normal distribution assumption was imposed that might not be completely consistent with the original database derived from experimental results. To remove such an arbitrary assumption, here the simulated proteins are instead generated by the seed-propagated sampling directly based on the experimental data. The details of the resampling scheme are described in Zhang and Chou.[51] Again, the number of samples selected for each simulated protein is $10^4$. In the current study, 5000 simulated proteins are thus generated for each of the four structural classes based on the 4 × 30 representative proteins of Appendix B. For these simulated proteins, predictions are performed by the current algorithm, the least Hamming distance algorithm,[12,13] and the least Euclidian distance algorithm,[41] respectively. The predicted results thus obtained are given in Table VII, from which it may be seen that the average simulated accuracy by the current algorithm is 97.29%, which is significantly higher than those by the other two algorithms.

**TABLE IV. The Predicted Results* for the 30 α/β Proteins in the Training Database**

| PDB code of the 30 α/β proteins† | $D^2(\mathbf{P}, \overline{\mathbf{P}}_\alpha)$ | $D^2(\mathbf{P}, \overline{\mathbf{P}}_\beta)$ | $D^2(\mathbf{P}, \overline{\mathbf{P}}_{\alpha+\beta})$ | $D^2(\mathbf{P}, \overline{\mathbf{P}}_{\alpha/\beta})$ | The predicted structural class |
|---|---|---|---|---|---|
| 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 |
| 1PAX– | 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 |
| 1SO1– | 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 | α + β |
| 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 |
| 4CPA– | 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% | | | | | |

*,†See corresponding footnotes to Table I.

A comparison of Tables V and VII indicates that although the rates of correct prediction by the current algorithm for the 30 α, 30 β, and 30 α + β proteins in the training database are all the same (equal to 100%), those calculated each from the corresponding 5000 simulated proteins are different: the accuracies for α and β classes are higher than that of α + β class. Why is that? The answer can be found by analyzing Figure 1a–d. As we can see from Figure 1a, all the 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 19-D 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 close to its simulated accuracy. The same is true for the case of β proteins. However, the situation is quite different for the α + β proteins. As shown in Figure 1c, although all the 30 α + β proteins have the shortest Mahalanobis distances to the norm of their own class, some of them are just slightly

**TABLE V. Comparison of Various Prediction Methods for the 4×30 Training Database Proteins**

| Method | Rate of correct prediction: α class | β class | α + β class | α/β class | Average accuracy* |
|---|---|---|---|---|---|
| This paper† | $\frac{30}{30}$ = 100% | $\frac{30}{30}$ = 100% | $\frac{30}{30}$ = 100% | $\frac{29}{30}$ = 96.7% | $\frac{119}{120}$ = 99.2% |
| Chou[13,‡] (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.[41,Σ] | $\frac{23}{30}$ = 76.7% | $\frac{20}{30}$ = 66.7% | $\frac{11}{30}$ = 36.7% | $\frac{26}{30}$ = 86.7% | $\frac{80}{120}$ = 66.7% |

*The average accuracy is the percentage of the number of correct prediction events for all classes divided by the number of total prediction events

†Based on the least Mahalanobis distance principle [Eq. (15)].

‡Based on the least Hamming distance principle [Eq. (5)].

ΣBased on the least Euclidian distance principle [Eq. (6)].

**TABLE VI. The Predicted Results* for the 64 Testing Proteins of Known X-Ray Structure That Are Not Included in the Training Database**

| PDB† code of 64 proteins | $D^2(\mathbf{X},\bar{\mathbf{X}}_\alpha)$ | $D^2(\mathbf{X},\bar{\mathbf{X}}_\beta)$ | $D^2(\mathbf{X},\bar{\mathbf{X}}_{\alpha+\beta})$ | $D^2(\mathbf{X},\bar{\mathbf{X}}_{\alpha/\beta})$ | Observed type | Predicted type |
|---|---|---|---|---|---|---|
| 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 | α+β | α+β |
| 1BBPA | 9.76 | 2.38 | 2.16* | 14.57 | α+β | α+β |
| 1BW4– | 8.39 | 6.07 | 1.79* | 21.61 | α+β | α+β |
| 1COX– | 3.72 | 1.21 | 0.64* | 1.32 | α+β | α+β |
| 1DNKA | 0.99 | 1.73 | 0.78* | 3.38 | α+β | α+β |
| 1GLAG | 4.45 | 1.19 | 1.04* | 1.55 | α+β | α+β |
| 1MS2A | 1.56 | 2.31 | 0.84* | 6.60 | α+β | α+β |
| 1OVOA | 3.93 | 4.07 | 1.48* | 56.49 | α+β | α+β |
| 1POC– | 7.85 | 3.34 | 0.67* | 17.26 | α+β | α+β |
| 1PPBA | 3.91 | 1.30 | 1.24* | 2.25 | α+β | α+β |
| 1SHAA | 1.12 | 2.38 | 1.01* | 5.04 | α+β | α+β |
| 1THO– | 3.07 | 3.14 | 0.85* | 2.73 | α+β | α+β |
| 1TRX– | 3.32 | 3.09 | 1.00* | 2.87 | α+β | α+β |
| 2AAA– | 3.63 | 1.51 | 0.57* | 2.71 | α+β | α+β |
| 2PIA– | 1.89 | 0.74* | 0.74 | 7.33 | α+β | β‡ |
| 2SN3– | 7.91 | 9.13 | 2.46* | 83.30 | α+β | α+β |
| 2TAAA | 2.70 | 0.73 | 0.60* | 2.90 | α+β | α+β |
| 3B5C– | 3.72 | 5.78 | 1.83* | 6.39 | α+β | α+β |
| 3SC2A | 4.22 | 0.75 | 0.60* | 2.30 | α+β | α+β |
| 3SC2B | 8.64 | 1.77 | 1.27* | 3.70 | α+β | α+β |
| 3TLN– | 4.34 | 0.55 | 0.54* | 2.05 | α+β | α+β |
| 4ENL– | 0.42* | 1.43 | 0.45 | 1.34 | α+β | α‡ |
| 4INSB | 6.22 | 21.61 | 3.86* | 25.04 | α+β | α+β |
| 4RCRH | 2.39 | 1.21 | 0.91* | 2.78 | α+β | α+β |
| 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 |
| 2SBP– | 1.84 | 2.02 | 0.85 | 0.48* | α/β | α/β |
| 1WSYA | 6.77 | 1.42 | 1.22* | 1.94 | α/β | α+β‡ |
| 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% | | | | | | |

*,†See corresponding footnotes to Table I.

‡Incorrect prediction.

**TABLE VII. Comparison of Various Prediction Methods for 5000 × 4 Simulated Proteins**

| Method | Rate of correct prediction | | | | |
|---|---|---|---|---|---|
| | α class | β class | α + β class | α/β class | Average accuracy* |
| This paper[†] | $\frac{5000}{5000}$ = 100% | $\frac{4995}{5000}$ = 99.90% | $\frac{4895}{5000}$ = 97.90% | $\frac{4530}{5000}$ = 90.60% | $\frac{19420}{20000}$ = 97.10% |
| Chou[13,‡] | $\frac{3528}{5000}$ = 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.[14,Σ] | $\frac{3829}{5000}$ = 76.58% | $\frac{3280}{5000}$ = 65.60% | $\frac{1877}{5000}$ = 37.54% | $\frac{4283}{5000}$ = 85.66% | $\frac{13269}{20000}$ = 66.35% |

*,†,‡,ΣSee corresponding footnotes to Table V.

shorter than their counterparts, meaning that, even measured according to the Mahalanobis distance, the distribution of α + β proteins in the 19-D space is not so clear-cut, or have a slightly ambiguous "border." The percent 100% rate of correct prediction obtained according to Figure 1c for the α + β proteins may be just a result of statistical errors due to a limited number of data, and it of course cannot be retained for a large number 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 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 Tables V and VII obtained by the least Hamming distance method[12,13] or the least Euclidian distance method.[41] However, when doing so, the corresponding ambiguity should be defined according to the Hamming or Euclidian distance, respectively.

The advantage of using the simulated accuracy can eliminate the errors of statistical fluctuation due to a limited number of data, and hence more accurately reflecting the objective reality.

## CONCLUSION

The unique feature of the new algorithm is the use of the Mahalanobis distance instead of ordinary geometric distances. 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 the training and testing sets 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 in classifying protein structure classes would be available. How large will the desired database be? According to a recent estimation by Chothia,[15] the large majority of proteins come from about 1000 families. If he is correct, then the desired complete database should consist of about 1000 nonhomologous proteins.

As suggested recently by Muskal and Kim,[52] the structural class of a protein may basically depend on its amino acid composition. Their suggestion is supported by the high accuracy of the new algorithm because the only input for the prediction is the amino acid composition. The new algorithm can also serve as a complemental tool for the secondary structure prediction since knowledge of the structural class of a protein is useful in improving the prediction of its secondary structure.[17,18]

## ACKNOWLEDGMENTS

The author would like to thank Joan Baker for her help in drawing Figure 1, and Dr. Gert Vriend for his advice in selecting nonhomologous structures. Valuable discussions with Professor C. T. Zhang and Dr. Donald Tong are gratefully acknowledged. I would particularly like to express special gratitude to Bi-Kun Luo. Without her encouragement I would not have been able to overcome the difficulties that occurred during this study.

## REFERENCES

1. Carlacci, L., Chou, K. C., Maggiora, G. M. A heuristic approach to predicting the tertiary structure of bovine somatotropin. Biochemistry 30:4389–4398, 1991.
2. Chou, K. C. Energy-optimized structure of antifreeze protein and its binding mechanism. J. Mol. Biol. 223:509–517, 1992.
3. Chou, K. C., Carlacci, L. Energetics approach to the folding of α/β barrels. Proteins 9:280–295, 1991.
4. Chou, K. C., Carlacci, L. Simulated annealing approach to the study of protein structures. Protein Eng. 4:661–667, 1991.
5. Chou, K. C., Scheraga, H. A. Origin of the right-handed twist of β-sheets of poly(L-Val) chains. Proc. Natl. Acad. Sci. U.S.A. 79:7047–7051, 1982.
6. Chou, K. C. Carlacci, L., Maggiora, G. M., Parodi, L. A., Schulz, M. W. An energy-based approach to packing the 7-helix bundle of bacteriorhodopsin. Protein Sci. 1:810–827, 1992.
7. Chou, K. C., Zhang, C. T. A new approach to predicting protein folding types. J. Protein Chem. 12:169–178, 1993.
8. Chou, K. C., Zhang, C. T. Predicting protein folding types by distance functions that make allowances for amino acid interactions. J. Biol. Chem., 269:22014–22020, 1994.
9. Chou, K. C., Némethy, G., Scheraga, H. A. Effect of amino acid composition on the twist and the relative stability of parallel and antiparallel β-sheets. Biochemistry 22:6213–6221, 1983.
10. Chou, K. C., Némethy, G., Scheraga, H. A. Energy of stabilization of the right-handed βαβ crossover in proteins. J. Mol. Biol. 205:241–249, 1989.
11. Chou, K. C., Némethy, G., Scheraga, H. A. Energetics of interactions of regular structural elements in proteins. Acc. Chem. Res. 23:134–141, 1990.
12. Chou, P. Y. Amino acid composition of four classes of proteins. Abstracts of Papers, Part I, Second Chemical Congress of the North American Continent, Las Vegas, 1980.
13. Chou, P. Y. Prediction of protein structural classes from amino acid composition. In: "Prediction of Protein Struc-

ture and the Principles of Protein Conformation." Fasman, G. D. ed. New York: Plenum Press, 1989: 549–586.
14. Chou, P. Y., Fasman, G. D. Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. Biochemistry 13:222–245, 1974.
15. Chothia, C. One thousand families for the molecular biologist. Nature (London) 357:543–544, 1992.
16. Cohen, F. E., Kuntz, I. E. Prediction of the three-dimensional structure of human growth hormone. Proteins 2:162–166, 1987.
17. Cohen, B., Presnell, S. R., Cohen, F. E. Origins of structural diversity within sequentially identical hexapeptides. Proteins Sci. 2:2134–2145,1993.
18. Deléage, G., Roux, B. In: "Prediction of Protein Structure and the Principles of Protein Conformation." Fasman, G. D., ed. New York: Plenum Press, 1989: 587–597.
19. Dubchak, I., Holbrook, S. R., Kim, S. H. Prediction of protein folding class from amino acid composition. Proteins 16:79–91, 1993.
20. Efron, B. "The Jackknife, the Bootstrap and Other Resampling Plans." Philadelphia: Society for Industrial and Applied Mathematics, 1990): Chap. 5.
21. Fasman, G. D. The development of the prediction of protein structure. In: "Prediction of Protein Structure and the Principles of Protein Conformation." Fasman, G. D., ed. New York: Plenum Press, 1989: 317–358.
22. Garnier, J., Osguthorpe, D. J., Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. 120:97–120, 1978.
23. Gilson, M. K., Honig B. Energetics of charge-charge interactions in proteins. Proteins 3:32–52, 1988.
24. Hobohm, U., Sander, C. Enlarged representative set of protein structures. Protein Sci. 3:522–524, 1994.
25. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
26. Karplus, M., Shakhnovich, E. Theoretical studies of thermodynamics and dynamics. In: "Protein Folding." Creighton, T. E., ed. New York: Freeman, 1992: 127–195.
27. Kawai, H., Kikuchi, T., Okamoto, Y. A prediction of tertiary structures of peptide by the Monte Carlo simulated annealing method. Protein Eng. 3:85–94, 1989.
28. Kikuchi, T. Discrimination of folding types of globular proteins based on average distance maps constructed from their sequences. J. Protein Chem. 12:515–523, 1993.
29. Klein, P. Prediction of protein structural class by discriminant analysis. Biochim. Biophys. Acta 874:205–215, 1986.
30. Klein, P., Delisi, C. Prediction of protein structural class from amino acid sequence. Biopolymers 25:1569–1672, 1986.
31. Levitt, M. Protein folding by restrained energy minimization and molecular dynamics. J. Mol. Biol. 170:723–764, 1983.
32. Levitt, M., Chothia, C. Structural patterns in globular proteins. Nature (London) 261:552–557, 1976.
33. Lim, V. I. Structural principles of globular protein secondary structure. J. Mol. Biol. 88:857–872, 1974.
34. Mackay, D. H. J., Cross, A. J., Hagler, A. T. 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. New York: Plenum Press, 1989: 317–358.
35. Mahalanobis, P. C. On the generalized distance in statistics. Proc. Natl. Inst. Sci. India 2:49–55, 1936.
36. Mao, B., Chou, K. C., Zhang, C. T. Protein folding classes: A geometric interpretation of amino acid composition of globular proteins. Protein Eng. 7:319–330, 1994.
37. Mardia, K. V., Kent, J. T., Biby, J. M. "Multivariate Analysis." London: Academic Press, 1979.
38. McCammon, J. A., Wong, C. F., Lybrand, T. P. Protein stability and function. In "Prediction of Protein Structure and the Principles of Protein Conformation." Fasman, G. D., ed. ny Structure and the Principles of Protein Conformation." Fasman, G. D., ed. ny: Plenum Press, 1989: 149–159.
39. Metfessel, B. A., Saurugger, P. N., Connelly, D. P., Rich, S. S. Cross-validation of protein structural class prediction using statistical clustering and neural networks. Protein Sci. 2:1171–1182, 1993.
40. Mooney, C. Z., Duval, R D. "Bootstrapping: A Nonparametric Approach to Statistical Inference." London: Sage Publications, 1993.
41. Nakashima, H., Nishikawa, K., Ooi, T. The folding type of a protein is relevant to the amino acid composition. J. Biochem. 99;152–162, 1986.
42. Pillai, K. C. S. Mahalanobis $D^2$. In: "Encyclopedia of Statistical Sciences," Vol. 5. Kotz, S., Johnson, N. L., eds. New York: John Wiley, 1985: 176–181. (This reference also presents a brief biography of Mahalanobis, who was a man of great originality and who made considerable contributions to statistics.)
43. Richardson, J. S., Richardson, D. C. Principles and patterns of protein conformation. In: "Prediction of Protein Structure and the Principles of Protein Conformation." ed. Fasman, G. D., ed. New York: Plenum Press, 1989: 1–98.
44. Rogers, N. K. 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. New York: Plenum Press, 1989: 359–389.
45. Scheraga, H. A. Calculations of conformations of polypeptides. Adv. Phys. Org. Chem. 6:103–184, 1968.
46. Scheraga, H. A. Conformational analysis of polypeptides and proteins for the study of protein folding, molecular recognition, and molecular design. J. Prot. Chem. 6:61–80, 1987.
47. Weiner, P. K., Kollman, P. A. AMBER: Assisted model building with energy refinement. a general program for modeling molecules and their interactions. J. Comp. Chem. 2:287–303, 1981.
48. Wilson, S. R., Cui, W. Applications of simulated annealing to peptides. Biopolymers 29:225–235, 1990.
49. Zhang, C. T., Chou, K. C. An optimization approach to predicting protein structural class from amino acid composition. Protein Sci. 1:401–408, 1992.
50. Zhang, C. T., Chou, K. C. Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition. Biophys. J. 63:1523–1529, 1992.
51. Zhang, C. T., Chou, K. C. An analysis of protein folding type prediction by seed-propagated sampling and jackknife test. Protein Eng., submitted.
52. Muskal, Kim. Predicting protein secondary structure content: a tandem neural network approach. J. Mol. Biol. 225: 713–727, 1992.

## APPENDIX A

**Theorem.** *Suppose there is a set of N points* $\mathbf{X}_k$ $(x_{k,1}, x_{k,2}, \ldots, x_{k,N})$ *defined in an m-dimensional space, and of their m components only* $m - 1$ *are independent because they are constrained by the normalization condition* $\sum_{i=1}^{m} x_{k,i} = 1$; *thus, the Mahalanobis distance based on such a set of N points would be a divergent quantity if defined in the m-dimensional space. However, it may be instead defined in a reduced* $(m - 1)$*-dimensional space formed by leaving out any one of the m components. And the distance thus defined will not depend on which component is taken off for establishing such an* $(m - 1)$*-dimensional space.*

Now, let us prove this theorem. Suppose $\mathbf{X}$ and $\mathbf{X}_k$ $(k = 1, 2, \ldots, N)$ are defined in an $m$-D (dimensional) space by

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}, \quad \mathbf{X}_k = \begin{bmatrix} x_{k,1} \\ x_{k,2} \\ \vdots \\ x_{k,m} \end{bmatrix} \quad (k = 1, 2, \cdots, N) \quad \text{(A1)}$$

where $x_j$ and $x_{k,j}$ $(j = 1, 2, \ldots, m)$ are the $j$th normalized components of $\mathbf{X}$ and $\mathbf{X}_k$ in the $m$-D space, respectively. $\mathbf{X}$ can be either one of $\mathbf{X}_k$ $(k = 1, 2, \ldots, N)$ or not. The norm $\overline{\mathbf{X}}$ and the covariance matrix $\mathbf{S}$ for the set of $\mathbf{X}_k$ $(k = 1, 2, \ldots, N)$ are defined by

$$\overline{\mathbf{X}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_m \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,m} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,m} \\ \vdots & \vdots & \ddots & \cdots \\ s_{m,1} & s_{m,2} & \cdots & s_{m,m} \end{bmatrix} \quad \text{(A2)}$$

where

$$\begin{cases} \bar{x}_i = \dfrac{1}{N}\sum_{k=1}^{N} x_{k,i} \\ s_{i,j} = \sum_{k=1}^{N} [x_{k,i} - \bar{x}_i]\,[x_{k,j} - \bar{x}_j] \end{cases} \quad (i, j = 1, 2, \ldots, m). \quad \text{(A3)}$$

The constraint due to the normalization yields

$$\sum_{i=1}^{m} x_i = 1, \quad \sum_{i=1}^{m} x_{k,i} = 1, \quad \sum_{i=1}^{m} \bar{x}_i = 1 \quad \text{(A4)}$$

From Eqs. (A3)–(A4), it follows

$$\begin{cases} s_{i,j} = s_{j,i}, & (i, j = 1, 2, \ldots, m) \\ \sum_{i=1}^{m} s_{i,j} = 0, & (j = 1, 2, \ldots, m) \\ \sum_{j=1}^{m} s_{i,j} = 0, & (i = 1, 2, \ldots, m). \end{cases} \quad \text{(A5)}$$

The Mahalanobis distance between $\mathbf{X}$ and $\overline{\mathbf{X}}$ in the $m$-D space is[35,42]:

$$D^2(\mathbf{X}, \overline{\mathbf{X}}) = (\mathbf{X} - \overline{\mathbf{X}})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{X} - \overline{\mathbf{X}}) \quad \text{(A6)}$$

Since Eq. (A5), the matrix $\mathbf{S}$ is singular, i.e., its determinant $\det \mathbf{S} = 0$, and hence Eq. (A6) is divergent. We therefore instead define the Mahalanobis distance in an $(m - 1)$-D space. To realize this, let us suppose

$$\mathbf{X}_{\bar{i}} = \begin{bmatrix} x_1 \\ \vdots \\ x_{i-1} \\ x_{i+1} \\ \vdots \\ x_m \end{bmatrix}, \quad \overline{\mathbf{X}}_{\bar{i}} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_{i-1} \\ \bar{x}_{i+1} \\ \vdots \\ \bar{x}_m \end{bmatrix} \quad \text{(A7)}$$

and

$$\mathbf{S}_{\bar{i},\bar{i}} = \begin{bmatrix} s_{1,1} & \cdots & s_{1,i-1} & s_{1,i+1} & \cdots & s_{1,m} \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ s_{i-1,1} & \cdots & s_{i-1,i-1} & s_{i-1,i+1} & \cdots & s_{i-1,m} \\ s_{i+1,1} & \cdots & s_{i+1,i-1} & s_{i+1,i+1} & \cdots & s_{i+1,m} \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ s_{m,1} & \cdots & s_{m,i-1} & s_{m,i+1} & \cdots & s_{m,m} \end{bmatrix} \quad \text{(A8)}$$

Thus, the Mahalanobis distance in an $(m - 1)$-D space is given by

$$D^2(\mathbf{X}_{\bar{i}}, \overline{\mathbf{X}}_{\bar{i}}) = (\mathbf{X}_{\bar{i}} - \overline{\mathbf{X}}_{\bar{i}})^T \mathbf{S}_{\bar{i},\bar{i}}^{-1} (\mathbf{X}_{\bar{i}} - \overline{\mathbf{X}}_{\bar{i}}). \quad \text{(A9)}$$

Our task is to prove the following:

$$D^2(\mathbf{X}_{\bar{1}}, \overline{\mathbf{X}}_{\bar{1}}) = D^2(\mathbf{X}_{\bar{2}}, \overline{\mathbf{X}}_{\bar{2}}) = \cdots = D^2(\mathbf{X}_{\bar{m}}, \overline{\mathbf{X}}_{\bar{m}}) \quad \text{(A10)}$$

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

$$\begin{aligned} & D^2(\mathbf{X}_{\bar{m}}, \overline{\mathbf{X}}_{\bar{m}}) \\ &= (\mathbf{X}_{\bar{m}} - \overline{\mathbf{X}}_{\bar{m}})^T \mathbf{S}_{\bar{m},\bar{m}}^{-1} (\mathbf{X}_{\bar{m}} - \overline{\mathbf{X}}_{\bar{m}}) \\ &= \begin{bmatrix} x_1 - \bar{x}_1 \\ x_2 - \bar{x}_2 \\ \vdots \\ x_{m-1} - \bar{x}_{m-1} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} s_{\bar{1},\bar{1}}^{-1} & s_{\bar{1},\bar{2}}^{-1} & \cdots & s_{\bar{1},m-1}^{-1} \\ s_{\bar{2},\bar{1}}^{-1} & s_{\bar{2},\bar{2}}^{-1} & \cdots & s_{\bar{2},m-1}^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m-1,1}^{-1} & s_{m-1,2}^{-1} & \cdots & s_{m-1,m-1}^{-1} \end{bmatrix} \begin{bmatrix} x_1 - \bar{x}_1 \\ x_2 - \bar{x}_2 \\ \vdots \\ x_{m-1} - \bar{x}_{m-1} \end{bmatrix} \\ &= \frac{1}{|S_{\bar{m},\bar{m}}|} \sum_{i=1}^{m-1} \sum_{j=1}^{m-1} (x_i - \bar{x}_i) A_{i,j} (x_j - \bar{x}_j) \\ &= \frac{1}{|S_{\bar{m},\bar{m}}|} \sum_{i=1}^{m-1} (x_i - \bar{x}_i) \sum_{j=1}^{m-1} A_{i,j} (x_j - \bar{x}_j) \end{aligned} \quad \text{(A11)}$$

where $A_{i,j}$ is a cofactor defined by

$$A_{i,j} = (-1)^{i+j} \det[\mathbf{S}_{\bar{m},\bar{m}}]_{\bar{i},\bar{j}} \quad \text{(A12)}$$

in which $[\mathbf{S}_{\bar{m},\bar{m}}]_{\bar{i},\bar{j}}$ is the matrix obtained by deleting the $i$th row and $j$th column from the matrix $\mathbf{S}_{\bar{m},\bar{m}}$. Thus,

according to the basic principle of a determinant, it follows

$$D^2(\mathbf{X}_{\overline{m}}, \overline{\mathbf{X}}_{\overline{m}}) = \frac{1}{|\mathbf{S}_{\overline{m},\overline{m}}|}(x_1 - \overline{x_1}) \begin{bmatrix} x_1 - \overline{x_1} & x_2 - \overline{x_2} & \cdots & x_{m-1} - \overline{x}_{m-1} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m-1,1} & s_{m-1,2} & \cdots & s_{m-1,m-1} \end{bmatrix}$$

$$+ \frac{1}{|\mathbf{S}_{\overline{m},\overline{m}}|}(x_2 - \overline{x_2}) \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,m-1} \\ x_1 - \overline{x_1} & x_2 - \overline{x_2} & \cdots & x_{m-1} - \overline{x}_{m-1} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m-1,1} & s_{m-1,2} & \cdots & s_{m-1,m-1} \end{bmatrix}$$

$$+ \cdots + \frac{1}{|S_{\overline{m},\overline{m}}|}(x_{m-1} - \overline{x}_{m-1}) \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,m-1} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ x_1 - \overline{x}_1 & x_2 - \overline{x}_2 & \cdots & x_{m-1} - \overline{x}_{m-1} \end{bmatrix}$$

$$= \frac{1}{|S_{\overline{m},\overline{m}}|} \sum_{i=1}^{m-1} (x_i - \overline{x}_i)\Delta_{\overline{m},i} \qquad \text{(A13)}$$

where

$$\Delta_{\overline{m},i} = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,m-1} \\ \vdots & \vdots & \vdots & \vdots \\ s_{i-1,1} & s_{i-1,2} & \cdots & s_{i-1,m-1} \\ x_1 - \overline{x}_1 & x_2 - \overline{x_2} & \cdots & x_{m-1} - \overline{x}_{m-1} \\ s_{i+1,1} & s_{i+1,2} & \cdots & s_{i+1,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m-1,1} & s_{m-2,2} & \cdots & s_{m-1,m-1} \end{bmatrix}. \qquad \text{(A14)}$$

Similarly, we have

$$D^2(\mathbf{X}_{\overline{1}}, \overline{\mathbf{X}}_{\overline{1}}) = \frac{1}{|\mathbf{S}_{\overline{1},\overline{1}}|}(x_2 - \overline{x_2}) \begin{bmatrix} x_2 - \overline{x_2} & x_3 - \overline{x_3} & \cdots & x_m - \overline{x_m} \\ s_{3,2} & s_{3,3} & \cdots & s_{3,m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m,2} & s_{m,3} & \cdots & s_{m,m} \end{bmatrix}$$

$$+ \frac{1}{|\mathbf{S}_{\overline{1},\overline{1}}|}(x_3 - \overline{x_3}) \begin{bmatrix} s_{2,2} & s_{2,3} & \cdots & s_{2,m} \\ x_2 - \overline{x_2} & x_3 - \overline{x_3} & \cdots & x_m - \overline{x_m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m,2} & s_{m,3} & \cdots & s_{m,m} \end{bmatrix}$$

$$+ \cdots + \frac{1}{|\mathbf{S}_{\overline{1},\overline{1}}|}(x_m - \overline{x_m}) \begin{bmatrix} s_{2,2} & s_{2,3} & \cdots & s_{2,m} \\ s_{3,2} & s_{3,3} & \cdots & s_{3,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_2 - \overline{x_2} & x_3 - \overline{x_3} & \cdots & x_m - \overline{x_m} \end{bmatrix}$$

$$= \frac{1}{|\mathbf{S}_{\overline{1},\overline{1}}|} \sum_{i=2}^{m} (x_i - \overline{x_i})\Delta_{\overline{1},i} \qquad \text{(A15)}$$

where

$$\Delta_{\overline{1},i} = \begin{bmatrix} s_{2,2} & s_{2,3} & \cdots & s_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ s_{i-1,2} & s_{i-1,3} & \cdots & s_{i-1,m} \\ x_2 - \overline{x}_2 & x_3 - \overline{x_3} & \cdots & x_m - \overline{x}_m \\ s_{i+1,2} & s_{i+1,3} & \cdots & s_{i+1,m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m,2} & s_{m,3} & \cdots & s_{m,m} \end{bmatrix}. \qquad \text{(A16)}$$

To prove $D^2(\mathbf{X}_{\overline{m}}, \overline{\mathbf{X}}_{\overline{m}}) = D^2(\mathbf{X}_{\overline{1}}, \overline{\mathbf{X}}_{\overline{1}})$ let us first prove their denominators are equal to each other, i.e., $|\mathbf{S}_{\overline{m},\overline{m}}| = |\mathbf{S}_{\overline{1},\overline{1}}|$. According to Eq. (A5) as well Eq. (A8), it follows

$$|\mathbf{S}_{\overline{1},\overline{1}}| = \begin{vmatrix} s_{2,2} & s_{2,3} & \cdots & s_{2,m} \\ s_{3,2} & s_{3,3} & \cdots & s_{3,m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m,2} & s_{m,3} & \cdots & s_{m,m} \end{vmatrix} = \begin{vmatrix} s_{2,2} & s_{2,3} & \cdots & -s_{2,1} \\ s_{3,2} & s_{3,3} & \cdots & -s_{3,1} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m,2} & s_{m,3} & \cdots & -s_{m,1} \end{vmatrix}$$

$$= (-1)^m \begin{vmatrix} s_{2,1} & s_{2,2} & \cdots & s_{2,m-1} \\ s_{3,1} & s_{3,2} & \cdots & s_{3,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m,1} & s_{m,2} & \cdots & s_{m,m-1} \end{vmatrix} = (-1)^m \begin{vmatrix} s_{2,1} & s_{2,2} & \cdots & s_{2,m-1} \\ s_{3,1} & s_{3,2} & \cdots & s_{3,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ -s_{1,1} & -s_{1,2} & \cdots & -s_{1,m-1} \end{vmatrix}$$

$$= (-1)^{2m} \begin{vmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,m-1} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m-1,1} & s_{m-1,2} & \cdots & s_{m-1,m-1} \end{vmatrix} = |\mathbf{S}_{\overline{m},\overline{m}}| \qquad \text{(A17)}$$

which indicates that the denominator of Eq. (A13) is equal to that of Eq. (A15).

Now let us prove their numerators are also equal to each other. Note that the first term of the numerator in Eq. (A15) is

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

$$= (x_2 - \overline{x_2}) \begin{vmatrix} x_2 - \overline{x_2} & x_3 - \overline{x_3} & \cdots & -(x_1 - \overline{x}_1) \\ s_{3,2} & s_{3,3} & \cdots & -s_{3,1} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m,2} & s_{m,3} & \cdots & -s_{m,1} \end{vmatrix}$$

$$= (-1)^{m-1}(x_2 - \overline{x_2}) \begin{vmatrix} x_1 - \overline{x}_1 & x_2 - \overline{x_2} & \cdots & -(x_{m-1} - \overline{x}_{m-1}) \\ s_{3,1} & s_{3,2} & \cdots & s_{3,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m,1} & s_{m,2} & \cdots & s_{m,m-1} \end{vmatrix}$$

$$= (-1)^{m-1}(x_2 - \bar{x}_2)\begin{vmatrix} x_1 - \bar{x}_1 & x_2 - \bar{x}_2 & \cdots & -(x_{m-1} - \bar{x}_{m-1}) \\ s_{3,1} & s_{3,2} & \cdots & s_{3,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m-1,1} & s_{m-1,2} & \cdots & s_{m-1,m-1} \\ -(s_{1,1} + s_{2,1}) & -(s_{1,2} + s_{2,2}) & \cdots & -(s_{1,m-1} + s_{2,m-1}) \end{vmatrix}$$

$$= (-1)^{2(m-1)}(x_2 - \bar{x}_2)\begin{vmatrix} s_{1,1} + s_{2,1} & s_{1,2} + s_{2,2} & \cdots & s_{1,m-1} + s_{2,m-1} \\ x_1 - \bar{x}_1 & x_2 - \bar{x}_2 & \cdots & -(x_{m-1} - \bar{x}_{m-1}) \\ s_{3,1} & s_{3,2} & \cdots & s_{3,m-1} \\ \vdots & \vdots & \vdots & \vdots \\ s_{m-1,1} & s_{m-1,2} & \cdots & s_{m-1,m-1} \end{vmatrix}$$

$$= (x_2 - \bar{x}_2)\begin{vmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,m-1} \\ x_1 - \bar{x}_1 & x_2 - \bar{x}_2 & \cdots & -(x_{m-1} - \bar{x}_{m-1}) \\ s_{3,1} & s_{3,2} & \cdots & s_{3,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m-1,1} & s_{m-1,2} & \cdots & s_{m-1,m-1} \end{vmatrix}$$

$$- (x_2 - \bar{x}_2)\begin{vmatrix} x_1 - \bar{x}_1 & x_2 - \bar{x}_2 & \cdots & -(x_{m-1} - \bar{x}_{m-1}) \\ s_{2,1} & s_{2,2} & \cdots & s_{2,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m-1,1} & s_{m-1,2} & \cdots & s_{m-1,m-1} \end{vmatrix}$$

$$= (x_2 - \bar{x}_2)\Delta_{\bar{m},2} - (x_2 - \bar{x}_2)\Delta_{\bar{m},1}. \tag{A18}$$

Similarly, we have

$$\begin{cases} (x_3 - \bar{x}_3)\Delta_{\bar{1},3} = (x_3 - \bar{x}_3)\Delta_{\bar{m},3} - (x_3 - \bar{x}_3)\Delta_{\bar{m},1} \\ (x_4 - \bar{x}_4)\Delta_{\bar{1},4} = (x_4 - \bar{x}_4)\Delta_{\bar{m},4} - (x_4 - \bar{x}_4)\Delta_{\bar{m},1} \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ (x_{m-1} - \bar{x}_{m-1})\Delta_{\bar{1},m-1} = (x_{m-1} - \bar{x}_{m-1})\Delta_{\bar{m},m-1} - (x_{m-1} - \bar{x}_{m-1})\Delta_{\bar{m},1} \end{cases} \tag{A19}$$

The last numerator term of Eq. (A15) is

$$(x_m - \bar{x}_m)\Delta_{\bar{1},m} = (x_m - \bar{x}_m)\begin{vmatrix} s_{2,2} & s_{2,3} & \cdots & s_{2,m} \\ s_{3,2} & s_{3,3} & \cdots & s_{3,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_2 - \bar{x}_2 & x_3 - \bar{x}_3 & \cdots & x_m - \bar{x}_m \end{vmatrix}$$

$$= (x_m - \bar{x}_m)\begin{vmatrix} s_{2,2} & s_{2,3} & \cdots & -s_{2,1} \\ s_{3,2} & s_{3,3} & \cdots & -s_{3,1} \\ \vdots & \vdots & \ddots & \vdots \\ x_2 - \bar{x}_2 & x_3 - \bar{x}_3 & \cdots & -(x_1 - \bar{x}_1) \end{vmatrix}$$

$$= -(X_m - \bar{X}_m)\Delta_{\bar{m}1}. \tag{A20}$$

Substituting Eqs. (A17)–(A20) into Eq. (A15), we obtain

$$D^2(\mathbf{X}_{\bar{1}}, \bar{\mathbf{X}}_{\bar{1}}) = \frac{1}{|\mathbf{S}_{\bar{m},\bar{m}}|}\left\{\sum_{i=2}^{m-1}(x_i - \bar{x}_i)\Delta_{\bar{m},i} - \Delta_{\bar{m},1}\sum_{i+2}^{m}(x_i - \bar{x}_i)\right\}$$

$$= \frac{1}{|\mathbf{S}_{\bar{m},\bar{m}}|}\left\{\sum_{i=2}^{m-1}(x_i - \bar{x}_i)\Delta_{\bar{m},i} + (x_1 - \bar{x}_1)\Delta_{\bar{m},1}\right\} \tag{A21}$$

$$= \frac{1}{|\mathbf{S}_{\bar{m},\bar{m}}|}\sum_{i=1}^{m-1}(x_i - \bar{x}_i)\Delta_{\bar{m},i} = D^2(\mathbf{X}_m, \bar{\mathbf{X}}_{\bar{m}})$$

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

In the above proof, the reason $\bar{\mathbf{X}}$ was chosen as one of the end points of the distance is because it would be more intuitive to associate with the norm of a folding type which is the main topic of this paper. Actually, $\bar{\mathbf{X}}$ can be replaced by any point defined in the $m$-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 Eq. (A4). Thus, the proof of the theorem is completed.

## APPENDIX B

The amino acid compositions 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 are given. The data of each protein contain two lines: the first line successively indicates its length, PDB code, the ratios of α, β, parallel β sheets, and antiparallel sheets [see Eq. (18)] and the second lines 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 minus, 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 1PPN- 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 4TGF- 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 C

The data of (1) the norms of protein structural classes, (2) the elements of covariance matrices, and (3) the elements of the inverse covariance matrices are given. Data in (1) and (2) for $\alpha$, $\beta$, $\alpha+\beta$, and $\alpha/\beta$ classes were derived from the $4 \times 30$ regular proteins given in Appendix B. Data in (3) were derived from (2) by calling a subroutine DLINDS in the IMSL Library (Fortran Subroutines for Mathematics and Statistics).

(1) The norms of the four structural classes. The 19 components of each of the four norms in the 19-D space are normalized to 100, and they are listed according to the alphabetical order of the single amino acid code.

| | A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
| $\alpha$ | 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 |
| $\beta$ | 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 |
| $\alpha+\beta$ | 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 |
| $\alpha/\beta$ | 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 |

(2) The covariance matrices of the four structural classes

$$\mathbf{Q}_\alpha = [s_{i,j}(\alpha)]$$

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 782.6 | 5.9 | 4.8 | −257.7 | 70.9 | 200.4 | −29.3 | −40.1 | −55.1 | −212.3 | −4.8 | −106.9 | 78.3 | −99.3 | −182.4 | −0.5 | −61.6 | −7.7 | 35.0 |
| 2 | 5.9 | 95.6 | 30.6 | 28.3 | −33.4 | −34.0 | −34.6 | 24.2 | 26.6 | −84.4 | 11.3 | 0.3 | −7.6 | 10.5 | 8.7 | −32.5 | −2.7 | −47.6 | −18.9 |
| 3 | 4.8 | 30.6 | 152.3 | −15.1 | −4.1 | −40.5 | −7.6 | 9.6 | −19.0 | −65.2 | 13.9 | 13.9 | −13.0 | −3.6 | −36.9 | −4.3 | 37.7 | −35.5 | −37.2 |
| 4 | −257.7 | 28.3 | −15.1 | 362.7 | −124.5 | −172.2 | 26.6 | −13.8 | 146.6 | 127.8 | −28.4 | 0.1 | −64.4 | 74.6 | 164.3 | −82.2 | −70.9 | −93.2 | −14.2 |
| 5 | 70.9 | −33.4 | −4.1 | −124.5 | 165.4 | 53.2 | 23.9 | 30.8 | −15.9 | −110.3 | −29.1 | 6.3 | 34.0 | −68.7 | −82.4 | 68.3 | 17.9 | 32.5 | 17.0 |
| 6 | 200.4 | −34.0 | −40.5 | −172.2 | 53.2 | 209.0 | 70.5 | −9.5 | −53.2 | −6.2 | −8.1 | −73.3 | 40.3 | −100.4 | −144.4 | 0.6 | −18.9 | 104.8 | 25.5 |
| 7 | −29.3 | −34.6 | −7.6 | 26.6 | 23.9 | 70.5 | 230.9 | −30.9 | 93.8 | 71.5 | −56.9 | −64.9 | 13.5 | −83.9 | −111.2 | −42.9 | −60.9 | 51.9 | 3.2 |
| 8 | −40.1 | 24.2 | 9.6 | −13.8 | 30.8 | −9.5 | −30.9 | 185.0 | 6.6 | −167.9 | −12.0 | −25.5 | −16.4 | −6.2 | −16.6 | 48.7 | 46.4 | −65.7 | 0.6 |
| 9 | −55.1 | 26.6 | −19.0 | 146.8 | −15.9 | −53.2 | 93.8 | 6.6 | 426.4 | −99.5 | −24.2 | −26.1 | 3.5 | −40.0 | −40.2 | −121.0 | −129.9 | −12.4 | −22.8 |
| 10 | −212.3 | −84.4 | −65.2 | 127.8 | −110.3 | −6.2 | 71.5 | −167.9 | −99.5 | 523.4 | 38.3 | −17.7 | −23.1 | −17.8 | 79.1 | −87.6 | 0.2 | 64.3 | 37.5 |
| 11 | −4.8 | 11.3 | 13.9 | −28.4 | −29.1 | −8.1 | −56.9 | −12.0 | −24.2 | 38.3 | 88.2 | 12.0 | 11.6 | 22.3 | −6.5 | −27.8 | 25.0 | −18.4 | −9.6 |
| 13 | 78.3 | −7.6 | −13.0 | −64.4 | 34.0 | 40.3 | 13.5 | −16.4 | 3.5 | −23.1 | 11.6 | −27.0 | 77.1 | −59.9 | −77.6 | 9.3 | 30.4 | 28.7 | 10.0 |
| 14 | −99.3 | 10.5 | −3.6 | 74.6 | −68.7 | −100.4 | −83.9 | −6.2 | −40.0 | −17.8 | 22.3 | 81.5 | −59.9 | 211.1 | 157.5 | −1.3 | −3.6 | −119.8 | −17.7 |
| 15 | −182.4 | 8.7 | −36.9 | 164.3 | −82.4 | −144.4 | −111.2 | −16.6 | −40.2 | 79.1 | −6.5 | 62.7 | −77.6 | 157.5 | 305.7 | 24.0 | −19.5 | −120.7 | −10.2 |
| 16 | −0.5 | −32.5 | −4.3 | −82.2 | 68.3 | 0.6 | −42.9 | 48.7 | −121.0 | −87.6 | −27.8 | 24.6 | 9.3 | −1.3 | 24.0 | 185.7 | 59.3 | −20.9 | −1.6 |
| 17 | −61.6 | −2.7 | 37.7 | −70.9 | 17.9 | −18.9 | −60.9 | 46.4 | −129.9 | 0.2 | 25.0 | −13.7 | 30.4 | −3.6 | −19.5 | 59.3 | 150.4 | −16.3 | 3.4 |
| 18 | −7.7 | −47.6 | −35.5 | −93.2 | 32.5 | 104.8 | 51.9 | −65.7 | −12.4 | 64.3 | −18.4 | −18.1 | 28.7 | −119.8 | −120.7 | −20.9 | −16.3 | 286.8 | 18.8 |
| 19 | 35.0 | −18.9 | −37.2 | −14.2 | 17.0 | 25.5 | 3.2 | 0.6 | −22.8 | 37.5 | −9.6 | −28.2 | 10.0 | −17.7 | −10.2 | −1.6 | 3.4 | 18.8 | 28.0 |

$$\mathbf{Q}_\beta\ [s_{i,j}(\beta)]$$

$\mathbf{Q}_{\beta}\ [s_{i,j}(\beta)]$

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 377.3 | 73.4 | −9.4 | −116.9 | −19.5 | 171.7 | −18.2 | −66.4 | −199.8 | −97.9 | −79.0 | −28.8 | −80.3 | −46.0 | −15.4 | 88.8 | 47.2 | 16.0 | −24.2 |
| 2 | 73.4 | 486.1 | −78.8 | −76.6 | −114.9 | −7.8 | −46.8 | 53.9 | −10.9 | −61.6 | 15.4 | −87.0 | −19.8 | −51.8 | 159.6 | −154.5 | −125.7 | −72.4 | 19.3 |
| 3 | −9.4 | −78.8 | 195.7 | 80.6 | −20.3 | 11.6 | 14.3 | −0.2 | −11.5 | −27.7 | −2.1 | −38.0 | −59.9 | −46.1 | −2.3 | −74.1 | 30.0 | −25.5 | −3.2 |
| 4 | −116.9 | −76.6 | 80.6 | 204.9 | 19.8 | −109.5 | 26.8 | 25.3 | 69.7 | 22.4 | 26.6 | −16.5 | 24.7 | −10.4 | 2.7 | −78.4 | −51.3 | −19.1 | 6.5 |
| 5 | −19.5 | −114.9 | −20.3 | 19.8 | 355.5 | −85.4 | 82.0 | 29.8 | −85.0 | −26.0 | −11.3 | 177.1 | 85.7 | −76.1 | 7.2 | −0.3 | −128.4 | −73.2 | −31.5 |
| 6 | 171.7 | −7.8 | 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 | −19.1 | 75.8 | 30.5 | 21.9 |
| 7 | −18.2 | −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 |
| 8 | −66.4 | 53.9 | −0.2 | 25.3 | 29.8 | −44.4 | 18.2 | 216.4 | −3.3 | 14.4 | −3.8 | 15.0 | 20.2 | 14.2 | 45.2 | −169.2 | −26.5 | −118.0 | 3.6 |
| 9 | −199.8 | −10.9 | −11.5 | 69.7 | −85.0 | −170.9 | −30.9 | −3.3 | 407.2 | 52.9 | 102.7 | 53.3 | 78.3 | −3..7 | −141.7 | −111.9 | −39.2 | 119.0 | −27.4 |
| 10 | −97.9 | −61.6 | −27.7 | 22.4 | −26.0 | −45.5 | −25.6 | 14.4 | 52.9 | 203.6 | −10.7 | −21.5 | 7.1 | 76.1 | 7.5 | 31.5 | −44.1 | −0.3 | −9.0 |
| 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 |
| 12 | −28.8 | −87.0 | −38.0 | −16.5 | 177.1 | −39.5 | 38.1 | 15.0 | 53.3 | −21.5 | 8.2 | 219.3 | 3.7 | −74.4 | −73.7 | −8.9 | −28.9 | 9.4 | −19.9 |
| 13 | −80.3 | −19.8 | −59.9 | 24.7 | 85.7 | −166.6 | 22.8 | 20.2 | 78.3 | 7.1 | 34.7 | 3.7 | 190.3 | 2.7 | −16.7 | 21.1 | −87.8 | 3.3 | −30.7 |
| 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 | 2.7 | 216.3 | −48.0 | 101.3 | 20.3 | −8.8 | 1.9 |
| 15 | −15.4 | 159.6 | −2.3 | 2.7 | 7.2 | 50.1 | 7.5 | 45.2 | −141.7 | 7.5 | −31.9 | −73.7 | −16.7 | −48.0 | 286.2 | −106.4 | −138.1 | −138.4 | 21.1 |
| 16 | 88.8 | −154.5 | −74.1 | −78.4 | −0.3 | −19.1 | −2.0 | −169.2 | −111.9 | 31.5 | −55.8 | −8.9 | 21.1 | 101.3 | −106.4 | 417.4 | 134.0 | 57.2 | −5.8 |
| 17 | 47.2 | −125.70 | 30.0 | −51.3 | −128.4 | 75.8 | −44.9 | −26.5 | −39.2 | −44.1 | −35.6 | −28.9 | −87.8 | 20.3 | −138.1 | 134.0 | 384.9 | 81.4 | 42.8 |
| 18 | 16.0 | −72.4 | −25.5 | −19.1 | −73.2 | 30.5 | −29.8 | −118.0 | 119.0 | −0.3 | 39.2 | 9.4 | 3.3 | −8.8 | −138.4 | 57.2 | 81.4 | 231.0 | −14.7 |
| 19 | −24.2 | 19.3 | −3.2 | 6.5 | −31.5 | 21.9 | 2.6 | 3.6 | −27.4 | −9.0 | −6.1 | −19.9 | −30.7 | 1.9 | 21.1 | −5.8 | 42.8 | −14.7 | 35.9 |

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

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 546.8 | −148.8 | 17.9 | 30.9 | −51.9 | 86.2 | −44.5 | −120.5 | 21.2 | 60.8 | 0.1 | −100.7 | −85.4 | −67.1 | −80.5 | −67.3 | −46.1 | 217.6 | −31.8 |
| 2 | −148.8 | 315.8 | −75.2 | −17.3 | −41.7 | −3.3 | −0.7 | −68.8 | −20.9 | −114.9 | −40.1 | 94.0 | 42.7 | −37.2 | −48.7 | 134.8 | 55.0 | −36.5 | −18.5 |
| 3 | 17.9 | −75.2 | 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 | −9.1 | −24.6 | 17.7 |
| 4 | 30.9 | −17.3 | −34.8 | 176.2 | −30.5 | −11.4 | −24.2 | 0.4 | 167.8 | 33.0 | −18.2 | −49.0 | −22.3 | −27.3 | −11.2 | −95.7 | −44.0 | 23.4 | −1.9 |
| 5 | −51.9 | −41.7 | 33.4 | −30.5 | 73.9 | −47.4 | 27.5 | 14.6 | −19.9 | 61.4 | 17.8 | 3.6 | 9.7 | 0.5 | 20.2 | −0.7 | −33.4 | −37.1 | 3.3 |
| 6 | 86.2 | −3.3 | 7.8 | −11.4 | −47.4 | 299.2 | −22.4 | −73.1 | −109.7 | −54.1 | −66.0 | −32.4 | −14.6 | −49.8 | −43.5 | −48.6 | 9.2 | 99.8 | 18.2 |
| 7 | −44.5 | −0.7 | 27.9 | −24.2 | 27.5 | −22.4 | 63.6 | −34.3 | −11.2 | −0.2 | 12.8 | −4.2 | 2.3 | −0.2 | 20.3 | −0.8 | −5.5 | −27.4 | 0.9 |
| 8 | −120.5 | −68.8 | −3.6 | 0.4 | 14.6 | −73.1 | −34.3 | 201.2 | 45.5 | 26.5 | 12.6 | 1.4 | −13.0 | 63.4 | 40.1 | −21.7 | −4.5 | −80.8 | 14.2 |
| 9 | 21.2 | −20.9 | −64.7 | 167.8 | −19.9 | −109.7 | −11.2 | 45.5 | 462.7 | 85.5 | 12.4 | −41.8 | −65.4 | −41.4 | −46.8 | −158.9 | −36.3 | −42.4 | −40.8 |
| 10 | 60.8 | −114.9 | 41.8 | 33.0 | 61.4 | −54.1 | −0.2 | 26.5 | 85.5 | 248.5 | 1.5 | −84.3 | 17.7 | −46.9 | 15.6 | −126.1 | −81.7 | −8.2 | 5.3 |
| 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 | 0.6 | 22.4 | 14.3 | −1.3 | 18.3 | −25.9 | −2.5 |
| 12 | −100.7 | 94.0 | −27.8 | −49.0 | 3.6 | −32.4 | −4.2 | 1.4 | −41.8 | −84.3 | −3.4 | 124.3 | 9.4 | 9.7 | −19.7 | 108.5 | 23.1 | −52.4 | −17.9 |
| 13 | −85.4 | 42.7 | −8.4 | −22.3 | 9.7 | −14.6 | 2.3 | −13.0 | −65.4 | 17.7 | 0.6 | 9.4 | 104.6 | 8.5 | 34.8 | 4.7 | −18.7 | −17.8 | 11.5 |
| 14 | −67.1 | −37.2 | −6.6 | −27.3 | 0.5 | −49.8 | −0.2 | 63.4 | −41.4 | −46.9 | 22.4 | 9.7 | 8.5 | 91.3 | 40.8 | 10.9 | 19.2 | −40.3 | 19.5 |
| 15 | −80.5 | −48.7 | 26.4 | −11.2 | 20.2 | −43.5 | 20.3 | 40.1 | −46.8 | 15.6 | 14.3 | −19.7 | 34.8 | 40.8 | 112.6 | −26.5 | −31.4 | −43.3 | 17.5 |
| 16 | −67.3 | 134.8 | −24.3 | −95.7 | −0.7 | −48.6 | −0.8 | −21.7 | −158.9 | −126.1 | −1.3 | 108.5 | 4.7 | 10.9 | −26.5 | 242.6 | 51.6 | −3.8 | −24.3 |
| 17 | −46.1 | 55.0 | −9.1 | −44.0 | −33.4 | 9.2 | −5.5 | −4.5 | −36.3 | −81.7 | 18.3 | 23.1 | −18.7 | 19.2 | −31.4 | 51.6 | 166.3 | −12.7 | −16.1 |
| 18 | 217.6 | −36.5 | −24.6 | 23.4 | −37.1 | 99.8 | −27.4 | −80.8 | −42.4 | −8.2 | −25.9 | −52.4 | −17.8 | −40.3 | −43.3 | −3.8 | −12.7 | 170.5 | −11.2 |
| 19 | −31.8 | −18.5 | 17.7 | −1.9 | 3.3 | 18.2 | 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 |

$\mathbf{Q}_{\alpha/\beta} = [s_{i,j}(\alpha/\beta)]$

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 335.7 | −20.9 | −11.2 | −62.6 | −105.5 | 51.0 | −12.9 | −5.5 | 39.1 | 2.4 | −2.8 | −16.3 | −32.4 | 10.5 | −75.1 | 58.7 | −48.2 | −1.9 | 5.6 |
| 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 | −32.7 | −0.6 |
| 3 | −11.2 | 10.8 | 122.4 | 37.3 | 3.6 | −8.1 | −39.1 | 22.0 | 21.7 | 19.8 | −9.5 | −31.3 | −37.4 | 26.8 | 14.2 | −71.7 | −13.4 | −42.4 | 1.0 |
| 4 | −62.6 | 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 | −31.0 | 2.6 |
| 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 | −59.1 | 2.7 |
| 6 | 51.0 | 19.8 | −8.1 | −35.0 | −6.8 | 158.4 | −7.7 | 15.3 | −33.8 | −22.6 | 0.6 | −16.2 | −31.7 | −26.4 | −53.7 | 80.5 | −7.0 | −47.7 | 6.5 |
| 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 | 27.8 | −16.7 | 19.1 | −2.4 | 24.3 | 11.8 | −10.6 |
| 8 | −5.5 | 18.1 | 22.0 | 82.6 | −18.6 | 15.3 | −24.7 | 109.3 | 40.5 | −15.9 | 21.1 | −22.4 | −54.8 | −12.9 | −20.6 | −19.7 | −25.1 | −63.4 | 4.2 |
| 9 | 39.1 | −10.8 | 21.7 | 33.2 | −22.9 | −33.8 | −31.3 | 40.5 | 167.7 | −15.6 | 9.3 | 6.4 | −4.9 | 17.7 | −44.4 | −82.4 | −52.1 | 16.2 | 1.7 |
| 10 | 2.4 | 4.6 | 19.8 | −13.5 | 23.0 | −22.6 | 10.6 | −15.9 | −15.6 | 70.6 | 12.2 | −20.4 | −8.7 | 9.7 | 36.2 | −37.1 | 10.3 | −44.3 | −1.9 |
| 11 | −2.8 | 10.4 | −9.5 | 22.2 | 3.0 | 0.6 | 3.6 | 21.1 | 9.3 | 12.2 | 44.4 | −19.0 | 7.0 | −9.9 | 6.2 | −28.5 | −7.7 | −28.9 | −8.4 |
| 12 | −16.3 | −24.6 | −31.3 | −38.2 | 6.1 | −16.2 | −1.9 | −22.4 | 6.4 | −20.4 | −19.0 | 80.2 | 22.0 | −1.1 | −43.1 | 46.2 | −2.2 | 46.3 | 4.4 |
| 13 | −32.4 | −14.0 | −37.4 | −29.7 | 25.9 | −31.7 | 27.8 | −54.8 | −4.9 | −8.7 | 7.0 | 22.0 | 91.0 | −16.4 | 1.3 | −21.8 | 3.0 | 55.4 | −10.1 |
| 14 | 10.5 | −8.3 | 26.8 | −15.2 | −5.1 | −26.4 | −16.7 | −12.9 | 17.7 | 9.7 | −9.9 | −1.1 | −16.4 | 60.2 | 6.0 | −14.3 | −6.6 | 6.7 | −7.3 |
| 15 | −75.1 | 14.5 | 14.2 | 35.3 | 23.4 | −53.7 | 19.1 | −20.6 | −44.4 | 36.2 | 6.2 | −43.1 | 1.3 | 6.0 | 114.8 | −72.5 | 9.2 | 5.0 | −11.6 |
| 16 | 58.7 | −140.0 | −71.7 | −114.3 | −23.4 | 80.5 | −2.4 | −19.7 | −82.4 | −37.1 | −28.5 | 46.2 | −21.8 | −14.3 | −72.5 | 258.9 | 39.1 | −25.3 | 29.1 |
| 17 | −48.2 | −0.1 | −134.4 | −35.8 | 20.1 | −7.0 | 24.3 | −25.1 | −52.1 | 10.3 | −7.7 | −2.2 | 3.0 | −6.6 | 9.2 | 39.1 | 72.5 | −19.8 | 1.8 |
| 18 | −1.9 | −32.7 | −42.4 | −13.0 | −59.1 | −47.7 | 11.8 | −63.4 | 16.2 | −44.3 | −28.9 | 46.3 | 55.4 | 6.7 | 5.0 | −25.3 | −19.8 | 227.7 | −25.8 |
| 19 | 5.6 | −0.6 | 1.0 | 2.6 | 2.7 | 6.5 | −10.6 | 4.2 | 1.7 | −1.9 | −8.4 | 4.4 | −10.1 | −7.3 | −11.6 | 29.1 | 1.8 | −25.8 | 23.3 |

(3) The inverse covariance matrices of the four structural classes.

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

| | 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 | 0.09 | 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 | 0.04 | 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 |

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

| | 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 |

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

| | 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 | 0.14 | 0.05 | 0.07 | 0.02 | 0.09 | 0.05 |
| 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 | 0.09 | 0.09 | 0.06 | 0.06 | 0.20 |

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

| | 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 | 0.37 | 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 | .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 D

The amino acid compositions of the 64 testing proteins of which 9 are α proteins, 22 β proteins, 24 α + β proteins, and 9 α/β proteins are given. The data of each protein contain two lines: the first line successively indicates its length, PDB code, the ratios of α, β, parallel β sheets, and antiparallel sheets [see Eq. (18)] and the second line gives the frequencies of 20 amino acids according to the alphabetical order of the single amino acide letter code: ACDEFGHIKLMNPQRSTVWY. The frequencies are normalized to 100. The fifth character in the PDB code incicates a specific chain of a protein; if it is minus, 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 2SODO | 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 1BW4- | 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 |

## Note Added in Proof

It was reported in a recent publication[8] that a similar effort was made to improve the predicted accuracy for a completely different proteins database classified by Nakashima et al.[41] However, in that report neither a cross-validation for an independent set of testing proteins nor a jackknife analysis wasperformed. Also, the simulated accuracy reportedthere was computed according to a normaldistribution assumption, which was arbitrarily imposed and might be inconsistent with the experimental data. Finally, no mathematical proof was given in that report why the predicted result is unique no matter which one of the 20 amino acid compositions is left out during calculation.