Does the folding type of a protein depend on its amino acid composition?

Kuo-Chen Chou*

Computer-Aided Drug Discovery, Upjohn Laboratories, Kalamazoo, MI 49007-4940, USA

Received 7 February 1995; revised version received 22 February 1995

Abstract Proteins of known structures are generally classified into one of the following four folding types: α, β, α + β, and α/β proteins. Recent findings [Muskal and Kim (1992) J. Mol. Biol. 225, 713–727] suggested that the folding type of a protein might basically depend on its amino acid composition. If this is true, why is that the predicted results of the protein folding type from amino acid composition always failed to reach the desired accuracy? An examination of the prediction approach indicates that none of the previous algorithms has ever taken into account the coupling effect among different amino acid components. In view of this, a new algorithm has been developed which distinguishes itself from the previous ones by incorporating such a coupling effect. The very high rates, 99.2% and 95.3%, of correct predictions thus obtained for a recently constructed training set of 120 proteins and testing set of 64 proteins, respectively, provide confirmation of the above suggestion.

Key words: α, β, α + β, α/β Protein; Coupling effect; Mahalanobis distance; Cross validation

The overall fold of a protein is generally described in terms of its folding type [1–5]. In order to predict the folding type of a protein, various methods were proposed [6–11]. It is to demonstrate in this letter that, by incorporating the coupling effect among different amino acid components, the prediction accuracy can be significantly improved.

According to its amino acid composition, a protein molecule can be represented by a point or a vector in a 20D (dimensional) space, the so-called composition space [6]. However, the amino acid composition of a protein must be normalized, i.e. constrained by $\sum_{i=1}^{20} x_i = 1$, where x_i is the composition component of the *i*th amino acid in a protein. This indicates that of the 20 amino acid composition components only 19 are independent. Therefore, by leaving out any one of its 20 components, one can stil uniquely represent a protein by a point in a 19D space, as formulated by the following equation:

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

where $x_{k,1}$, $x_{k,2}$, ..., $x_{k,19}$ are respectively the 19 amino acid composition components of the *k*th protein X_k, and *N* the total number of proteins in a given set. The norm of the protein set is defined by

$$\overline{\mathrm{X}} = \begin{bmatrix} \overline{x_1} \\ \overline{x_2} \\ \vdots \\ \overline{x_{19}} \end{bmatrix} \tag{2}$$

where

$$\overline{x_i} = \frac{1}{N}\sum_{k=1}^{N} x_{k,i} \quad (i = 1, 2, \ldots, 19) \tag{3}$$

When the *N* proteins in eq. (3) are all α proteins, $\overline{\mathrm{X}}$ thus defined would become the norm of an α protein set, denoted by $\overline{\mathrm{X}}_\alpha$. Likewise, when the *N* proteins in eq. (3) are all β, or α + β, or α/β proteins, $\overline{\mathrm{X}}$ would become the norm of a β, or α + β, or α/β protein set, denoted by $\overline{\mathrm{X}}_\beta$, $\overline{\mathrm{X}}_{\alpha+\beta}$, or $\overline{\mathrm{X}}_{\alpha/\beta}$, respectively.

Suppose X is a protein whose folding type is to be predicted. It also corresponds to a point $(x_1, x_2, \ldots, x_{19})$ in the 19D space, where x_i is the normalized frequency of its *i*th amino acid. Thus, the Mahalanobis distance, $D(\mathrm{X},\overline{\mathrm{X}})$, between the norm $\overline{\mathrm{X}}$ defined by eq. (2) and X in the 19D space is given by [12]

$$D_2(\mathrm{X},\overline{\mathrm{X}}) = (\mathrm{X} - \overline{\mathrm{X}})^T \mathrm{S}^{-1} (\mathrm{X} - \overline{\mathrm{X}}) \tag{4}$$

where *T* is the transposition operator, and S^{-1} is the inverse matrix of S defined by the following 19 × 19 covariance matrix

$$\mathrm{S} = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,19} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,19} \\ \vdots & \vdots & \ddots & \vdots \\ s_{19,1} & s_{19,2} & \cdots & s_{19,19} \end{bmatrix} \tag{5}$$

where

$$s_{i,j} = \sum_{k=1}^{N} [x_{k,i} - \overline{x_i}]\,[x_{k,j} - \overline{x_j}], \quad (i, j = 1, 2, \ldots, 19) \tag{6}$$

Note that the non-diagonal terms in eq. (5) are generally not equal to zero. It is these terms through which the coupling effect among different amino acid components is incorporated. Actually, a similar treatment has also been used by other investigators [13–16] in developing methods for alignment of protein sequences by using a substitution matrix with scores for all possible exchanges of one amino acid with another. Although the matrix elements introduced by them are different with those of the covariance matrix defined here, they both reflect the importance of coupling effect among different amino acid components in studying the similarity of proteins. It can also be proved that the value of $D(\mathrm{X},\overline{\mathrm{X}})$ is independent of which 19 of the 20 components are chosen as the bases for calculation. In other words, it will lead to a same result of $D(\mathrm{X},\overline{\mathrm{X}})$ by leaving out any one of the 20 normalized components from eq. (1) as long as X, $\overline{\mathrm{X}}$, and S are defined in a same 19D space.

*Corresponding author. Fax: (1) (616) 385-7373.

SSDI 0014-5793(95)00245-6

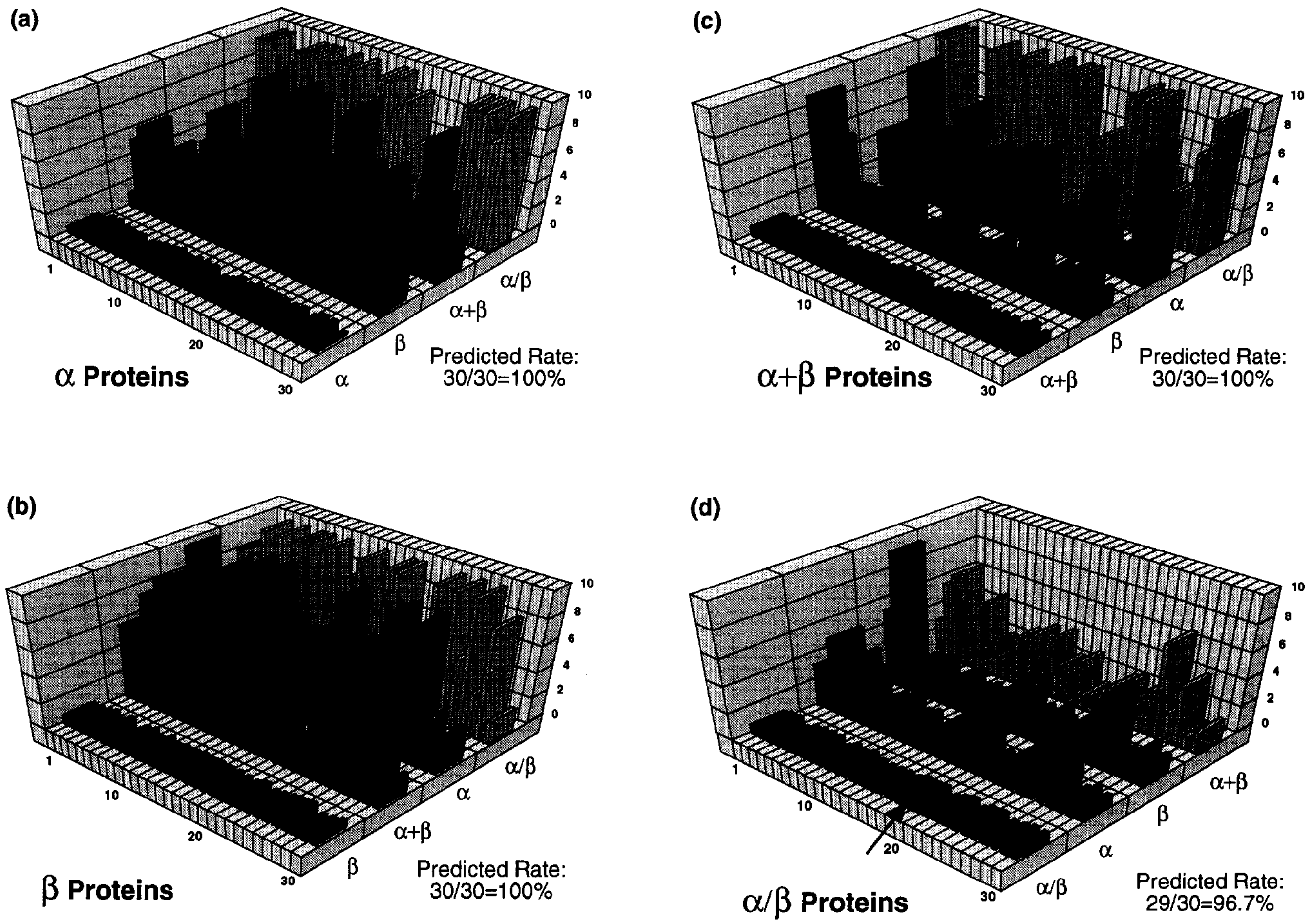

Fig. 1. The 3D histogram to show the Mahalanobis distance from each of (a) the 30 α proteins, (b) the 30 β proteins, (c) the 30 $\alpha+\beta$ proteins, and (d) the 30 α/β proteins to the norms of α, β, $\alpha+\beta$ and α/β types, respectively. The proteins in each folding type are arranged from left to right along the abscissa according to their order in Table 1. 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 (1TMD–, the 19th α/β protein in Table 1) that was incorrectly predicted.

When the N proteins in eqs. (3) and (6) are all α proteins, the S thus defined will become the covariance matrix of an α protein set, denoted by S_α, and $D^2(X,\overline{X})$ will become $D^2(X,\overline{X}_\alpha)$. Likewise, when the N proteins in eqs. (3) and (6) are all β, or $\alpha+\beta$, or α/β proteins, S will become the covariance matrix of a β, or $\alpha+\beta$, or α/β protein set, denoted by S_β, $S_{\alpha+\beta}$, or $S_{\alpha/\beta}$, respectively. And the corresponding $D^2(X,\overline{X})$ will become $D^2(X,\overline{X}_\beta)$, $D^2(X,\overline{X}_{\alpha+\beta})$, or $D^2(X,\overline{X}_{\alpha/\beta})$. The protein X will be predicted to be the folding type for which D^2 has the least value, as can be formulated as follows. Suppose

$$D^2(X,\overline{X}_\ell) = \\ = \mathrm{Min}\,\{D^2(X,\overline{X}_\alpha),\ D^2(X,\overline{X}_\beta), D^2(X,\overline{X}_{\alpha+\beta}),\ D^2(X,\overline{X}_{\alpha/\beta})\} \quad \textbf{(8)}$$

where the index ℓ can be α, β, $\alpha+\beta$, or α/β, and the operator Min means taking the least one among those in the parentheses, then the index ℓ of eq. (8) will represent which folding type the protein X should belong to. Therefore, the new algorithm is based on the least Mahalanobis distance principle.

In addition to a more efficient algorithm [17] a better training database is also important for improving the accuracy of prediction. In view of this, the selection of proteins for the training database is carried out according to that they should have (i) as many nonhomologous structures as possible, (ii) a good quality of structure, and (iii) a typical or distinguishable feature for each of the folding types concerned. 120 structure-known proteins were thus selected and classified into 30 α, 30 β, 30 $\alpha+\beta$, and 30 α/β proteins (Table 1). Based on such a training database, the norms and the inverse covariance matrices for the α, β, $\alpha+\beta$, and α/β are derived. Since these data are important for any practical calculations, they are given in Table 2. The predicted results thus obtained indicate that the rates of correct prediction for the 30 α, 30 β, 30 $\alpha+\beta$, and 30 α/β proteins are 100%, 100%, 100%, and 96,7%, respectively (Fig. 1a–d), with an average accuracy of 119/120 = 99.2%. However, for the same set of proteins, if predicted by the least Euclidean distance algorithm [6] or the least Hamming distance algorithm [7], the average accuracy was only 76/120 = 63.3% or 83/120 = 69.2%, respectively.

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 predicted results by the current algorithm are given in Table 3, which indicates an average accuracy of 61/64 = 95.3%. How-

Table 1
The PDB (Protein Data Bank) codes of the 4 × 30 = 120 representative proteins in the training database[a]

	PDB code[b]			
Number	α type	β type	$\alpha + \beta$ type	α/β type
1	1AVHA	1ACX–	1AAK–	1ABA–
2	1BABB	1AYH–	1CTF–	1CIS–
3	1BRD–	1CD8–	1DNKA	1CSEI
4	1C5A–	1CDTA	1EAF–	1CTC–
5	1CPCA	1CID–	1HSBA	1DHR–
6	1CPCL	1DFNA	1LTSA	1DRI–
7	1ECO–	1HILA	1LTSD	1ETU–
8	1FCS–	1HIVA	1NRCA	1FX1–
9	1FHA–	1HLEB	1OVB–	1GPB–
10	1FIAB	1MAMH	1POC–	1OFV–
11	1HBG–	1MONA	1PPN–	1PAZ–
12	1HDDC	1OMF–	1PRF–	1PFKA
13	1HIGA	1PHY–	1RND–	1PGD
14	1LE4–	1REIA	1SNC–	1Q21
15	1LIG–	1TEN–	1TFG–	1S01–
16	1LTSC	1TLK–	1TGSI	1SBP–
17	1MBC–	1VAAB	2ACHA	1SBT–
18	1MBS–	2ALP–	2ACT–	1TIMA
19	1RPRA	2AVIA	2BPA1	1TMD–
20	1TROA	2BPA2	2SNS–	1TREA
21	1UTG–	2HHRC	2SSI–	1ULA–
22	256BA	2ILA–	3IL8–	1WSYB
23	2CCYA	2LALA	3RUBS	2HAD–
24	2LH1–	2SNV–	3SGBI	2LIV–
25	2LHB–	3CD4A	3SICI	3GBP–
26	2MHBA	4GCR–	4BLMA	4FXN–
27	2MHBB	7APIB	4TMS–	5CPA–
28	2ZTAA	8I1B–	8CATA	5P21–
29	4MBA–	8FABA	9RNT–	8ABP–
30	4MBN–	8FABB	9RSAA	8ATCA

[a] The classification was made according to the following criteria: α proteins, $\alpha > 40\%$, $\beta < 5\%$; β proteins: $\alpha < 5\%$, $\beta > 40\%$; $\alpha + \beta$ proteins, $\alpha > 15\%$, $\beta > 15\%$ with more than 60% antiparallel β-sheets; and α/β proteins, $\alpha > 15\%$, $\beta > 15\%$ with more than 60% parallel β-sheets. (Here, for brevity, the percentages of α-helix and β-sheet in a protein are abbreviated by α and β, respectively.) The amino acid composition for each of the proteins listed here, and the ratios of its α helix and β sheet (parallel or antiparallel) components, are available upon request.

[b] The PDB 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.

ever, if the same 64 testing proteins were predicted by the least Euclidian distance algorithm [6] or the least Hamming distance algorithm [7], the average accuracy would be only 36/64 = 56.3% or 34/64 = 53.1%, respectively.

The above results indicate that for the same training and testing data the rates of correct prediction by the current algorithm are about 30–40% higher than those by the previous algorithms. 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 training data. The very high rates of correct prediction obtained here imply that the new algorithm will become a reliable tool for predicting the protein folding types if a statistically complete database in classifying protein structures would be available. How large will the desired database be? According to a recent estimation by Chothia [18], the large majority of proteins come from about one thousand families. If he is correct, then the desired complete database should consist of about one thousand nonhomologous proteins.

Since the only input for the new method is the amino acid composition of a protein, the high rate itself would further confirm the suggestion by Muskal and Kim [5] that the knowledge of sequence information is not necessary for highly accurate predictions of protein secondary structure content, implying that the folding type of a protein may basically depend on its amino acid composition.

Acknowledgements: I would 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] Levitt, M. and Chothia, C. (1976) Nature 261, 552–557.

[2] Richardson, J.S. and Richardson, D.C. (1989) in: Prediction of Protein Structure and the Principles of Protein Conformation (Fasman, G.D. Ed.) pp. 1–98, Plenum Press, New York.

[3] Deléage, G. and Roux, B. (1989) in: Prediction of Protein Structure and the Principles of Protein Conformation (Fasman, G.D. Ed.) pp. 587–597, Plenum Press, New York.

[4] Cohen, B.I., Presnell, S.R. and Cohen, F.E. (1993) Protein Sci. 2, 2134–2145.

[5] Muskal, S.M. and Kim, S.H. (1992) J. Mol. Biol. 225, 713–727.

[6] Nakashima, H., Nishikawa, K. and Ooi, T. (1986) J. Biochem. 99, 152–162.

[7] Chou, P.Y. (1989) in: Prediction of Protein Structure and the Principles of Protein Conformation (Fasman, G.D. Ed) pp. 549–586, Plenum Press, New York.

[8] Klein, P. (1986) Biochim. Biophys. Acta 874, 205–215.

[9] Zhang, C.T. and Chou, K.C. (1992) Protein Sci. 1, 401–408.

[10] Metfessel, B.A., Saurugger, P.N., Connelly, D.P. and Rich, S.S. (1993) Protein Sci. 2, 1171–1182.

[11] Dubchak, I., Holbrook, S.R. and Kim, S.H. (1993) Proteins: Struct. Funct. Genet. 16, 79–91.

[12] Pillai, K.C.S. (1985) in: Encyclopedia of Statistical Sciences (Kotz, S. and Johnson, N.L. Ed.) Vol. 5, pp. 176–181, John Wiley & Sons, New York. (This reference also presents a brief biography of Mahalanobis who was a man of great originality and who made considerable contributions to statistics.)

[13] Dayhoff, M.O. and Eck, R.V. (1968) Atlas of Protein Sequence and Structure, Natl. Biomed. Res. Found., Silver Spring, MD, Vol. 3, p. 33.

[14] Henikoff, S. and Henikoff, J.G. (1992) Proc. Natl. Acad. Sci. USA 89, 10915–10919.

[15] Miyazawa and Jernigan, J.G. (1993) Prot. Eng. 6, 267–278.

[16] Henikoff, S. and Henikoff, J.G. (1994) Genomics 19, 97–107.

[17] Chou, K.C. and Zhang, C.T. (1994) J. Biol. Chem. 269, 22014–22020.

[18] Chothia, C. (1992) Nature 357, 543–544.

Table 2
The data of (1) the norms of protein folding types and (2) the elements of the inverse covariance matrices, derived from the 4 × 30 representative proteins listed in Table 1

(1) The norms of the four folding types. 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
α	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
β	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
α+β	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
α/β	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 inverse covariance matrices of the four folding types.

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

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

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

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

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

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

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

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	0.22	0.26	0.24	0.22	0.24	0.24	0.25	0.32	0.20	0.14	0.12	0.34	0.34	0.30	0.36	0.23	0.33	0.22	0.20
2	0.26	0.69	0.27	0.23	0.18	0.26	0.27	0.39	0.25	0.20	0.09	0.49	0.48	0.40	0.43	0.26	0.41	0.25	0.28
3	0.24	0.27	0.30	0.25	0.26	0.27	0.32	0.35	0.23	0.15	0.16	0.38	0.37	0.34	0.40	0.28	0.36	0.25	0.24
4	0.22	0.23	0.25	0.26	0.24	0.26	0.27	0.30	0.22	0.16	0.12	0.35	0.35	0.31	0.37	0.26	0.34	0.22	0.18
5	0.24	0.18	0.26	0.24	0.31	0.25	0.28	0.33	0.21	0.13	0.16	0.31	0.31	0.30	0.36	0.26	0.33	0.26	0.21
6	0.24	0.26	0.27	0.26	0.25	0.29	0.28	0.36	0.23	0.17	0.13	0.39	0.39	0.34	0.42	0.26	0.38	0.24	0.23
7	0.25	0.27	0.32	0.27	0.28	0.28	0.41	0.36	0.24	0.15	0.17	0.39	0.36	0.37	0.40	0.29	0.35	0.27	0.28
8	0.32	0.39	0.35	0.30	0.33	0.36	0.36	0.50	0.29	0.22	0.17	0.50	0.52	0.45	0.53	0.33	0.49	0.33	0.32
9	0.20	0.25	0.23	0.22	0.21	0.23	0.24	0.29	0.21	0.14	0.11	0.32	0.32	0.28	0.35	0.23	0.31	0.20	0.18
10	0.14	0.20	0.15	0.16	0.13	0.17	0.15	0.22	0.14	0.14	0.06	0.23	0.25	0.20	0.24	0.16	0.22	0.15	0.12
11	0.12	0.09	0.16	0.12	0.16	0.13	0.17	0.17	0.11	0.06	0.14	0.18	0.15	0.17	0.19	0.13	0.17	0.14	0.15
12	0.34	0.49	0.38	0.35	0.31	0.39	0.39	0.50	0.32	0.23	0.18	0.59	0.56	0.48	0.58	0.36	0.53	0.33	0.32
13	0.34	0.48	0.37	0.35	0.31	0.39	0.36	0.52	0.32	0.25	0.15	0.56	0.60	0.49	0.58	0.37	0.53	0.33	0.31
14	0.30	0.40	0.34	0.31	0.30	0.34	0.37	0.45	0.28	0.20	0.17	0.48	0.49	0.46	0.50	0.32	0.46	0.31	0.31
15	0.36	0.43	0.40	0.37	0.36	0.42	0.40	0.53	0.35	0.24	0.19	0.58	0.58	0.50	0.63	0.39	0.56	0.36	0.33
16	0.23	0.26	0.28	0.26	0.26	0.26	0.29	0.33	0.23	0.16	0.13	0.36	0.37	0.32	0.39	0.28	0.35	0.24	0.20
17	0.33	0.41	0.36	0.34	0.33	0.38	0.35	0.49	0.31	0.22	0.17	0.53	0.53	0.46	0.56	0.35	0.54	0.33	0.30
18	0.22	0.25	0.25	0.22	0.26	0.24	0.27	0.33	0.20	0.15	0.14	0.33	0.33	0.31	0.36	0.24	0.33	0.25	0.23
19	0.20	0.28	0.24	0.18	0.21	0.23	0.28	0.32	0.18	0.12	0.15	0.32	0.31	0.31	0.33	0.20	0.30	0.23	0.32

Table 3
The predicted results[a] for the 64 testing proteins of known X-ray structure that are not included in the training database

PDB[b]code of 64 proteins	Mahalanobis distance[c]				Observed type	Predicted type
	$D^2(X,\bar{X}_\alpha)$	$D^2(X,\bar{X}_\beta)$	$D^2(X,\bar{X}_{\alpha+\beta})$	$D^2(X,\bar{X}_{\alpha/\beta})$		
1BBL–	2.20*	3.92	7.47	11.01	α	α
1HBBA	0.85*	4.42	1.88	10.08	α	α
1IFA–	1.81*	2.54	4.69	3.05	α	α
1MRRA	0.53*	0.71	0.63	0.74	α	α
1PDE–	3.27*	3.46	8.56	5.82	α	α
1PRCM	4.38*	4.98	7.09	6.64	α	α
1SAS–	2.88*	3.56	4.96	4.23	α	α
2TMVP	1.16*	2.09	2.10	17.05	α	α
4CPV–	2.83*	6.87	5.94	11.33	α	α
1AAIB	5.18	3.23*	3.48	21.54	β	β
1ATX–	24.33	4.38*	8.37	87.35	β	β
1COBA	5.60	4.26*	4.80	8.81	β	β
1EGF–	17.08	2.66*	7.15	52.36	β	β
1EST–	6.38	1.19*	5.43	10.94	β	β
1GPS–	16.34	5.82*	13.76	159.42	β	β
1HCC–	4.88	4.60*	5.25	14.19	β	β
1IXA–	15.95	7.70*	12.51	88.52	β	β
1MDAA	5.89	2.08*	2.75	4.70	β	β
1PPFE	3.89	2.13*	8.52	19.57	β	β
1R1A2	3.97	1.48*	2.29	4.83	β	β
1SHFA	7.87	0.65*	2.87	6.32	β	β
1TIE–	2.21	0.65*	1.80	3.76	β	β
1TNFA	4.44	1.24*	1.45	5.76	β	β
2ACHB	6.47	4.56*	9.92	83.51	β	β
2CTX–	9.91	3.68*	8.30	139.45	β	β
2MEV1	1.72	0.91*	4.34	5.89	β	β
2PLV1	2.53	0.43*	4.69	3.17	β	β
2SODO	5.60	4.26*	4.80	8.81	β	β
3RP2A	1.28	0.87*	1.02	2.76	β	β
4SGBI	9.59	5.26*	8.22	131.02	β	β
5NN9–	11.46	1.36*	1.45	18.05	β	β
1ABH–	1.97	1.68	1.06*	1.25	$\alpha+\beta$	$\alpha+\beta$
1BBPA	9.76	2.38	2.16*	14.57	$\alpha+\beta$	$\alpha+\beta$
1BW4–	8.39	6.07	1.79*	21.61	$\alpha+\beta$	$\alpha+\beta$
1COX–	3.72	1.21	0.64*	1.32	$\alpha+\beta$	$\alpha+\beta$
1DNKA	0.99	1.73	0.78*	3.38	$\alpha+\beta$	$\alpha+\beta$
1GLAG	4.45	1.19	1.04*	1.55	$\alpha+\beta$	$\alpha+\beta$
1MS2A	1.56	2.31	0.84*	6.60	$\alpha+\beta$	$\alpha+\beta$
1OVOA	3.93	4.07	1.48*	56.49	$\alpha+\beta$	$\alpha+\beta$
1POC–	7.85	3.34	0.67*	17.26	$\alpha+\beta$	$\alpha+\beta$
1PPBA	3.91	1.30	1.24*	2.25	$\alpha+\beta$	$\alpha+\beta$
1SHAA	1.12	2.38	1.01*	5.04	$\alpha+\beta$	$\alpha+\beta$
1THO–	3.07	3.14	0.85*	2.73	$\alpha+\beta$	$\alpha+\beta$
1TRX–	3.32	3.09	1.00*	2.87	$\alpha+\beta$	$\alpha+\beta$
2AAA–	3.63	1.51	0.57*	2.71	$\alpha+\beta$	$\alpha+\beta$
2PIA–	1.89	0.74*	0.74	7.33	$\alpha+\beta$	β^c
2SN3–	7.91	9.13	2.46*	83.30	$\alpha+\beta$	$\alpha+\beta$
2TAAA	2.70	0.73	0.60*	2.90	$\alpha+\beta$	$\alpha+\beta$
3B5C–	3.72	5.78	1.83*	6.39	$\alpha+\beta$	$\alpha+\beta$
3SC2A	4.22	0.75	0.60*	2.30	$\alpha+\beta$	$\alpha+\beta$
3SC2B	8.64	1.77	1.27*	3.70	$\alpha+\beta$	$\alpha+\beta$
3TLN–	4.34	0.55	0.54*	2.05	$\alpha+\beta$	$\alpha+\beta$
4ENL–	0.42*	1.43	0.45	1.34	$\alpha+\beta$	α^c
4INSB	6.22	21.61	3.86*	25.04	$\alpha+\beta$	$\alpha+\beta$
4RCRH	2.39	1.21	0.91*	2.78	$\alpha+\beta$	$\alpha+\beta$
1GPB–	1.14	0.63	1.06	0.41*	α/β	α/β
1MINA	2.90	1.61	1.18	0.64*	α/β	α/β
1NIPB	1.48	1.71	7.16	1.24*	α/β	α/β
1SBP–	1.84	2.02	0.85	0.48*	α/β	α/β
1WSYA	6.77	1.42	1.22*	1.94	α/β	$\alpha+\beta^c$
4ICD–	1.15	1.34	1.37	0.89*	α/β	α/β
7AATA	1.03	1.47	0.68	0.32*	α/β	α/β
9RUBB	2.20	0.93	0.80	0.79*	α/β	α/β
1GD1O	2.01	1.18	2.65	0.79*	α/β	α/β
	Average Rate of correct prediction = 61/64 = 95.3%					

[a] See footnote a to Table 1.

[b] See footnote b to Table 1.

[c] See eq. (4) for the definition of the Mahalanobis distance, which is different for different folding type. The one with the least value (marked by *) is assumed to correspond to the folding type for the predicted protein (cf. eq. **(8)**)