Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition

Chun-Ting Zhang* and Kuo-Chen Chou
Computational Chemistry, Upjohn Research Laboratories, Kalamazoo, Michigan 49001 USA

ABSTRACT In the methodology development for statistical prediction of protein structures, the founders of different methods usually selected different sets of proteins to test their predicted results. Therefore, it is hard to make a fair comparison according to the results they reported. Even if the predictions by different methods are performed for the same set of proteins, there is still such a problem: a method better than the other for one set of proteins would not necessarily remain so when applied to another set of proteins. To tackle this problem, a Monte Carlo simulation method is proposed to establish an objective criterion to measure the accuracy of prediction for the protein folding type. Such an objective accuracy is actually corresponding to the asymptotical limit generated during the Monte Carlo simulation process. Based on that, it has been found that the average objective accuracy for predicting the all-α, all-β, $\alpha + \beta$, and α/β proteins by the least Euclid's distance method (Nakashima, H., K. Nishikawa, and T. Ooi. 1986. *J. Biochem.* 99:152–162) is 73.0% and that by the least Minkowski's distance method (Chou, P. Y. 1989. Prediction in Protein Structure and the Principles of Protein Conformation. Plenum Press, New York. 549–586) is 70.9%, indicating that the former is better than the latter. However, according to the original reports, the latter claimed a rate of correct prediction with 79.7% but the former with only 70.2%, leading to a completely opposite conclusion. This indicates the necessity of establishing an objective criterion, and a comparison is meaningful only when it is based on the objective criterion. The simulation method and the idea developed here also can be applied to examine any other statistical prediction methods.

I. INTRODUCTION

Proteins of known structures are usually classified into five folding types: all-α, all-β, $\alpha + \beta$, α/β, and ζ (irregular) proteins (Levitt and Chothia, 1976; Richardson and Richardson, 1989). It has been found that the folding type of a protein is relevant to its amino acid composition (Chou, 1980, 1989; Nakashima et al., 1986). Therefore, the folding type of a protein can be predicted from its amino acid frequencies of occurrence. Such prediction was performed for 64 structurally known proteins, in which there are 19 all-α proteins, 15 all-β proteins, 14 $\alpha + \beta$ proteins, and 16 α/β proteins, and the average rate of correct prediction was reported to be 79.7% (Chou, 1989). On the other hand, Nakashima et al. (1986) considered a set of 135 structurally known proteins, in which there are 31 all-α, 34 all-β, 27 $\alpha + \beta$, 39 α/β, and 4 ζ (irregular) proteins. If the four irregular folding type proteins are not taken into account, the average rate of correct prediction by them would be 70.2%. The prediction by Chou (1989) was based on the least Minkowski's distance principle, and the prediction by Nakashima et al. (1986) was based on the least Euclid's distance principle. The details of the predicted results by means of these two methods are summarized in Table 1.

After carefully analyzing the data listed in Table 1, one may raise the following questions. Although the average accuracy predicted by the least Euclid's distance method (Nakashima et al., 1986) is only 70.2%, which is lower than 79.7%, the average accuracy by the least Minkowski's distance method, it does not necessarily mean that the least Euclid's distance method is poorer because the prediction by it was performed for a set of $135 - 4 = 131$ regular folding type proteins rather than a set of 64 proteins, as done by the least Minkowski's distance method (Chou, 1989). Furthermore, even if the predictions by different methods are performed for an exactly same set of proteins, the accountability of the results thus obtained could still be questionable. This is because a method, which yields the best predicted results for a set of proteins, does not necessarily guarantee to remain so when applied to another set of proteins. In other words, the predicted accuracy is, to some extent, dependent on the set of proteins selected by the predictor. Only when the number of proteins considered is sufficiently large can the bias due to the selection of different sets be eliminated. Unfortunately, so far there are only ~500 proteins whose three-dimensional structures have been determined. In view of this, can we find an approach through which the comparison of various prediction methods can be carried out according to an objective criterion? The present study was initiated in an attempt to tackle such a problem. Below, we shall resort to the Monte Carlo simulation to overcome the difficulty caused by the limited number of structurally known proteins.

II. METHODS

The principle and process of the Monte Carlo simulation can be illustrated as follows.

A. Simulation for the all-α proteins

Suppose that $v_1(\alpha), v_2(\alpha), \ldots, v_{20}(\alpha)$ are, respectively, the frequencies of 20 amino acids in the α folding type proteins. Note that each of

* Sabbatical leave professor from Department of Physics, Tianjin University, Tianjin, China.

TABLE 1 **The rates of correct prediction reported by the founders of the two different methods in literatures**

Method	Rate of correct prediction					
	α type	β type	$\alpha + \beta$ type	α/β type	ζ type	Average accuracy*
Nakashima et al.‡	$\frac{27}{31}$ = 87.1%	$\frac{22}{34}$ = 64.7%	$\frac{10}{27}$ = 37.0%	$\frac{33}{39}$ = 84.6%	$\frac{2}{4}$ = 50.0%§	$\frac{92}{131}$ = 70.2%
Chou‖	$\frac{16}{19}$ = 84.2%	$\frac{12}{15}$ = 80.0%	$\frac{11}{14}$ = 78.6%	$\frac{12}{16}$ = 75.0%	—	$\frac{51}{64}$ = 79.7%

* The average accuracy is the rate of total number of correct prediction for the four regular (i.e., α, β, $\alpha + \beta$, and α/β) type proteins divided by the total number of corresponding prediction events. The denominator in this table represents the number of prediction events and the numerator represents that of the corresponding correct prediction events.

‡ Based on the least Euclid's distance principle (Nakashima et al., 1986).

§ The rate for the ζ (irregular) folding type is based on too few (only 4) protein molecules to justify its statistical meaning, and hence it should not be counted in calculating the average accuracy.

‖ Based on the least Minkowski's distance principle (Chou, 1989).

$v_i(\alpha)$ (i = 1, 2, . . . , 20) is not a constant but a random variable. The mean $M_i(\alpha)$ and standard deviation $D_i(\alpha)$ for each of $v_i(\alpha)$(i = 1, 2, . . . , 20) were calculated based on 31 structurally known all-α proteins (Nakashima et al., 1986), and their values are listed in the third column of Tables 2 and 3, respectively. When the number of the all-α proteins is sufficiently large, each of $v_i(\alpha)$ (i = 1, 2, . . . , 20) may be assumed to be a normal random variable, and hence we have the following distribution density formula for $v_i(\alpha)$ (DeGroot, 1986):

$$N\{v_i(\alpha), M_i(\alpha), D_i(\alpha)\} = \frac{1}{\sqrt{2\pi}D_i(\alpha)} \exp\left\{-\frac{[v_i(\alpha) - M_i(\alpha)]^2}{2D_i^2(\alpha)}\right\} \quad (i = 1, 2, \ldots, 20). \tag{1}$$

Thus, for a sufficiently large set of α proteins, the value of $v_i(\alpha)$ can be treated as a random sample of the normal distribution as formulated by Eq. 1. Therefore, the occurrence frequency of each of the 20 amino acids in any given α protein can be simulated by using the Monte Carlo sampling technique. The concrete steps of Monte Carlo simulation are as follows.

1. Sampling of the standard normal distribution $N\{R, 0, 1\}$

An approximate, but quite accurate, sampling method will be adopted for generating the standard normal distribution, which, by definition, is the one obtained by substituting $v_i(\alpha) = R$, $M_i(\alpha) = 0$, and $D_i(\alpha) = 1$ into Eq. 1; i.e.,

$$N\{R, 0, 1\} = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{R^2}{2}\right\} \tag{1a}$$

Suppose that r_i (i = 1, 2, 3, · · ·) are within the region of [0, 1] and form a random number sequence, then according to the reason given in

TABLE 2 **The means $M_i(\alpha)$, $M_i(\beta)$, $M_i(\alpha + \beta)$, $M_i(\alpha/\beta)$, and $M_i(\zeta)$ (i = 1, 2, . . . , 20) of the 20 amino acid occurrence frequencies for the five protein folding types***

i‡	Amino acid	$M_i(\alpha)$	$M_i(\beta)$	$M_i(\alpha + \beta)$	$M_i(\alpha/\beta)$	$M_i(\zeta)$
1	Arg	0.0279	0.0322	0.0405	0.0435	0.0108
2	Leu	0.0889	0.0669	0.0637	0.0854	0.0402
3	Ser	0.0544	0.0950	0.0705	0.0589	0.0642
4	Thr	0.0491	0.0783	0.0641	0.0550	0.0435
5	Pro	0.0381	0.0523	0.0429	0.0436	0.0582
6	Ala	0.1163	0.0754	0.0889	0.0883	0.0890
7	Gly	0.0766	0.0987	0.0800	0.0871	0.1049
8	Val	0.0602	0.0748	0.0650	0.0762	0.0489
9	Lys	0.1010	0.0466	0.0718	0.0655	0.0327
10	Asn	0.0379	0.0490	0.0560	0.0413	0.0416
11	Gln	0.0333	0.0412	0.0317	0.0344	0.0403
12	His	0.0279	0.0164	0.0200	0.0219	0.0102
13	Glu	0.0652	0.0375	0.0618	0.0612	0.0685
14	Asp	0.0652	0.0537	0.0576	0.0612	0.0885
15	Tyr	0.0255	0.0367	0.0459	0.0302	0.0395
16	Cys	0.0171	0.0348	0.0294	0.0143	0.1204
17	Phe	0.0422	0.0357	0.0360	0.0388	0.0173
18	Ile	0.0372	0.0476	0.0474	0.0582	0.0699
19	Met	0.0242	0.0124	0.0140	0.0214	0.0053
20	Trp	0.0117	0.0148	0.0128	0.0138	0.0062

* These values were derived based on 135 structurally known proteins (Nakashima et al., 1986), of which 31 proteins are all-α type, 34 all-β, 27 $\alpha + \beta$, 39 α/β, and 4 ζ (irregular).

‡ The order of amino acids, each of which corresponds to a component of the 20-dimensional composition space, is numbered according to the codon usage table compiled by Wada et al. (1990), i.e., the order of an amino acid increases with the decrease of the degenerate degrees of its genetic code. If two amino acids are the same in such a degeneracy, then they are arranged in alphabetical order.

TABLE 3 The standard deviations $D_i(\alpha)$, $D_i(\beta)$, $D_i(\alpha + \beta)$, $D_i(\alpha/\beta)$, and $D_i(\zeta)$ ($i = 1, 2, \ldots, 20$) of the 20 amino acid occurrence frequencies for the five protein folding types*

i‡	Amino acid	$D_i(\alpha)$	$D_i(\beta)$	$D_i(\alpha + \beta)$	$D_i(\alpha/\beta)$	$D_i(\zeta)$
1	Arg	0.0270	0.0253	0.0222	0.0193	0.0083
2	Leu	0.0370	0.0232	0.0243	0.0197	0.0298
3	Ser	0.0230	0.0348	0.0325	0.0214	0.0325
4	Thr	0.0221	0.0272	0.0173	0.0164	0.0468
5	Pro	0.0176	0.0211	0.0202	0.0134	0.0311
6	Ala	0.0493	0.0439	0.0409	0.0265	0.0286
7	Gly	0.0335	0.0329	0.0286	0.0221	0.0833
8	Val	0.0232	0.0241	0.0256	0.0191	0.0321
9	Lys	0.0373	0.0269	0.0346	0.0235	0.0152
10	Asn	0.0174	0.0185	0.0257	0.0161	0.0140
11	Gln	0.0188	0.0238	0.0176	0.0138	0.0163
12	His	0.0246	0.0122	0.0136	0.0101	0.0067
13	Glu	0.0414	0.0249	0.0308	0.0240	0.0425
14	Asp	0.0259	0.0239	0.0170	0.0190	0.0326
15	Tyr	0.0153	0.0198	0.0248	0.0144	0.0150
16	Cys	0.0223	0.0339	0.0372	0.0091	0.0500
17	Phe	0.0258	0.0135	0.0153	0.0143	0.0178
18	Ile	0.0197	0.0229	0.0230	0.0180	0.0426
19	Met	0.0190	0.0110	0.0110	0.0091	0.0053
20	Trp	0.0096	0.0088	0.0121	0.0085	0.0077

* See the corresponding footnote to Table 2.
‡ See the corresponding footnote to Table 2.

the Appendix, the quantity R defined by the following equation should obey the standard normal distribution $N\{R, 0, 1\}$ of Eq. 1a:

$$R = \sum_{i=1}^{6} r_i - \sum_{i=7}^{12} r_i \quad (2)$$

2. Sampling of the normal distribution $N\{v_i(\alpha), M_i(\alpha), D_i(\alpha)\}$

Once the sample R for the $N\{R, 0, 1\}$ distribution is generated by Eq. 2, the following variable $v_i(\alpha)$ derived from R should obey the normal distribution $N\{v_i(\alpha), M_i(\alpha), D_i(\alpha)\}$:

$$v_i(\alpha) = D_i(\alpha)R + M_i(\alpha) \quad (i = 1, 2, \ldots, 20) \quad (3)$$

This can be proved by substituting $R = [v_i(\alpha) - M_i(\alpha)]/D_i(\alpha)$ into Eq. 1a followed by incorporating a corresponding transformation factor, the Jacobian (DeGroot, 1986), and comparing the equation thus obtained with Eq. 1.

3. Normalization

Since the sum of $v_i(\alpha)$ ($i = 1, 2, \ldots, 20$) thus obtained is generally not equal to 1, the following substitution should be performed:

$$v_i(\alpha) \Leftarrow \frac{v_i(\alpha)}{\sum_{j=1}^{20} v_j(\alpha)} \quad (i = 1, 2, \ldots, 20) \quad (4)$$

4. Prediction of the α folding type proteins

Once the normalized $v_i(\alpha)$ ($i = 1, 2, \ldots, 20$) are generated by Eqs. 2–4, the folding type for the "all-α proteins" is predicted by each of the following methods: (a) the least Minkowski's distance method (Chou, 1989) and (b) the least Euclid's distance method (Nakashima et al., 1986). The prediction by each of the above methods has only two possibilities, i.e., either correct (belonging to all-α type) or incorrect (not belonging to all-α type). For each of the two methods, a counter is set in the program: if the prediction is correct, then the reading of the counter will be increased by 1; otherwise, it remains unchanged.

The above steps 1–4 constitute a cycle of Monte Carlo simulation. The number of the simulation cycles can be set at any integer. The maximum number assigned in this work is 10^5. For each method, the accuracy of prediction is calculated according to the following formula:

$$q = \text{accuracy of prediction} = \left(\frac{\text{total number of correct prediction events}}{\text{total number of prediction events}}\right)\% \quad (5)$$

B. Simulation of all the other folding types

For the other folding types, i.e., all-β, $\alpha + \beta$, α/β, and ζ (irregular) proteins, just substituting α of Eqs. 1, 3, and 4 by β, $\alpha + \beta$, α/β, and ζ, we can get the corresponding formulations, respectively. The simulation process is in a way completely parallel to the one for the all-α proteins. The corresponding means $M_i(\beta)$, $M_i(\alpha + \beta)$, $M_i(\alpha/\beta)$, $M_i(\zeta)$ and standard deviations $D_i(\beta)$, $D_i(\alpha + \beta)$, $D_i(\alpha/\beta)$, $D_i(\zeta)$ are given in Tables 2 and 3, respectively. Note that the mean $M_i(\zeta)$ and standard deviation $D_i(\zeta)$ for the irregular folding type were calculated based on only four structurally known proteins (Nakashima et al., 1986) and their values are statistically insignificant. Therefore, all the data corresponding to the irregular folding type proteins are listed here for reference only.

III. RESULTS AND DISCUSSION

The main aim of this study is to find the objective accuracy for each of the two prediction methods. To realize this, the number of "proteins" generated by the simulation process should be sufficiently large. Note that according to the principle of Monte Carlo simulation, the error of the result thus obtained is proportional to $1/\sqrt{n}$,

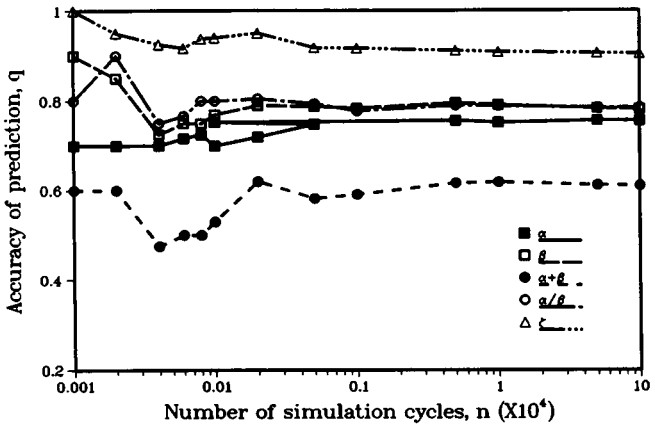

FIGURE 1 Plot of the rate of correct prediction q by the least Euclid's distance method (Nakashima et al., 1986) versus the number of simulation cycles n for each of the five protein folding types. When $n < 10^3$, the statistical fluctuations are remarkable. When $n \geq 3 \times 10^3$, however, each of the five curves approaches to its asymptotical limit, which is defined as the corresponding objective accuracy of prediction.

where n is the number of simulation cycles (DeGroot, 1986). Therefore, to obtain a reliable result, the value of n should be considerably large. The accuracy of prediction for each of the five folding types as a function of n by the least Euclid's distance method and least Minkowski's distance method is shown in Figs. 1 and 2, respectively, where there are five curves, each of which corresponds to a folding type of protein. It is seen through the two figures that when n is $>3 \times 10^3$, all of these curves gradually approach to a respective limit, the so-called asymptotical limit. In this case, the errors due to fluctuation can be omitted. Such an asymptotical limit is defined as the objective accuracy of prediction, whose value obtained for each of the five folding types and calculated by each of the two methods is listed in Table 4. Below, let us examine the results thus obtained according to the five different folding types.

A. Objective accuracy of prediction for the all-α proteins

It is seen from Table 4 that the asymptotical limit calculated by the least Euclid's distance method (Nakashima et al., 1986) is 76.3%, whereas the asymptotical limit by the least Minkowski's distance method (Chou, 1989) is only 74.1%, indicating that the objective accuracy of prediction for the all-α proteins by means of the least Euclid's method is better than that by the least Minkowski's method.

B. Objective accuracy of prediction for the all-β proteins

The asymptotical limit obtained by the least Euclid's method is 78.2% and that by the Minkowski's distance method is lower, with a value of only 74.5%.

C. Objective accuracy of prediction for the $\alpha + \beta$ proteins

The asymptotical limits obtained by the least Euclid's distance method and the least Minkowski's method are 59.5 and 58.1%, respectively. These figures indicate that the accuracy of prediction for the $\alpha + \beta$ proteins is poor regardless of which one of the two methods is used, although the former is slightly better than the latter. The poor predicted results for $\alpha + \beta$ proteins are consistent with the observation of Nakashima et al. (1986). They have found from a set of 135 structurally known proteins that the distribution of amino acid composition of $\alpha + \beta$ proteins extensively overlaps with those of all the other folding types of proteins. This might be the intrinsical reason why the accuracy is always poor when prediction is made for the $\alpha + \beta$ proteins only according to their amino acid composition.

D. Objective accuracy of prediction for the α/β proteins

The asymptotical limits obtained by the lease Euclid's distance method and the least Minkowski's method are 78.2 and 76.7%, respectively. Again, we can see that the result obtained by the former is better than that by the latter.

E. Objective accuracy of prediction for the ζ proteins

As pointed out previously, the statistical parameters for this type of proteins are derived from only four structurally known proteins (Nakashima et al., 1986), a number too small to justify their statistical meaning. Therefore, the Monte Carlo simulated results for the irregular proteins are statistically insignificant and they are reported here just for reference only.

Note that the asymptotical limits listed in Table 4 are

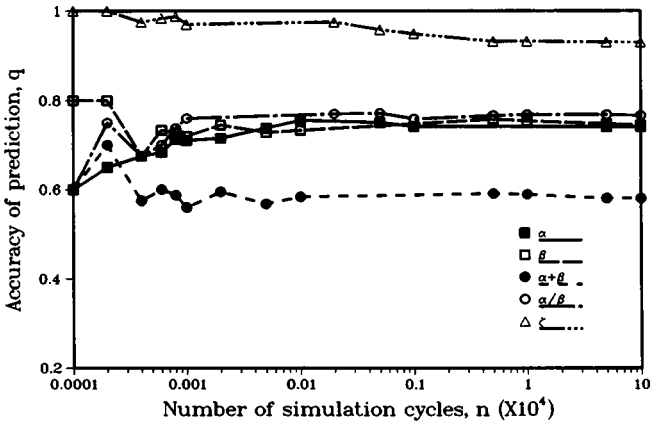

FIGURE 2 Plot of the rate of correct prediction q by the least Minkowski's distance method (Chou, 1989) versus the number of simulation cycles n for each of the five protein folding types. See legend to Fig. 1 for further explanation.

TABLE 4 **The objective accuracies obtained by Monte Carlo simulation for the two different methods***

Method	Objective accuracy					
	α type	β type	$\alpha + \beta$ type	α/β type	ζ type‡	Average accuracy§
			%			%
Nakashima et al.‖	76.3	78.2	59.5	78.2	90.9	73.0
Chou¶	74.1	74.5	58.1	76.7	90.9	70.9

* The number of simulation cycles in 10^5.

‡ The data for the ζ (irregular) folding type is statistically insignificant, and they are listed here for reference only. See text for further explanation about this.

§ The average accuracy is the rate of total number of correct prediction for the four regular (i.e., α, β, $\alpha + \beta$, and α/β) type proteins divided by the total number of corresponding prediction events.

‖ Based on the least Euclid's distance principle (Nakashima et al., 1986).

¶ Based on the least Minkowski's distance principle (Chou, 1989).

generally lower than the corresponding rates of correct prediction as listed in Table 1. This is because the number of proteins from which the data of Table 1 were derived is not large enough to really form a statistically significant set, and hence any data thus obtained must bear considerable statistical fluctuation. Such a statistical fluctuation also can be seen easily through Fig. 1. When the number of simulation cycles is small, the rate of correct prediction fluctuates remarkably with the number of simulation cycles. After the number of simulation cycles $n \geq 3 \times 10^3$, however, the amplitude of fluctuation tends to almost zero, indicating the predicted accuracy reaches an asymptotical limit. Accordingly, it is our belief that with the continuous increase of structurally known proteins, the predicted accuracy for the realistic proteins will gradually get close to the corresponding asymptotical limit, namely, the objective accuracy of prediction.

According to the Monte Carlo simulation results as listed in Table 4, the least Euclid's distance method is better than the least Minkowski's distance method in the objective accuracy. This is obviously contradictory to the results of Table 1, where the average accuracy reported by Chou (1989) using the least Minkowski's distance method is 79.7%, higher than the average accuracy of 70.2% obtained by means of the least Euclid's distance method as reported by Nakashima et al. (1986). Such a contradiction implies that the accuracy of prediction based on different set of proteins would bear considerable arbitrariness and hence could not be regarded as an objective standard. In the report by Chou (1989), the average accuracy was derived based on a set of 64 proteins, whereas in the report by Nakashima et al. (1986) that rate was derived on a set of 131 proteins, and hence they are not comparable. The Monte Carlo simulation results indicate that the phenomenon as shown in Table 1, that Chou's average accuracy is higher than Nakashima et al.'s, is none but an artifact. Such a contradiction also further indicates the necessity to use the asymptotical limit, namely the objective accuracy, as a criterion to judge which method is really more accurate than the other. According to the objective accuracy, the least Euclid's distance method is better than the least Minkowski's distance method.

It is instructive to point out that there is always some uncertainty in predicting the folding type of a protein according to its amino acid composition. The essence for this is that, although the folding type of a protein is correlated to its amino acid composition, the former cannot be uniquely defined by the latter. In other words, the amino acid composition of a protein is not yet sufficiently perfect to form a set of "complete parameters" to uniquely define its folding type. The effect of some "hidden variables," such as the amino acid order along the sequence of a protein, was not taken into account in any of the aforementioned two prediction methods. Obviously, these "hidden variables" will certainly somehow affect the folding of a protein. Therefore, merely based on the amino acid composition of proteins, it is impossible to raise the objective accuracy of folding type prediction to 100%. This also can be addressed rather quantitatively from a statistical point of view. It is clear from Table 3 that the standard deviations of the 20 amino acid frequencies for any of the five protein folding types are not equal to zero. This is the essence of why the objective accuracy of prediction will never reach 100% unless all these standard deviations are reduced to zero. As a demonstration to show the consistency of our theory and program, operations were performed according to the following hypothetical conditions. Suppose that the standard deviations listed in Table 3 are reduced by being multiplied with a factor of 0.9^{λ}, where $\lambda = 0, 2, 4, 6, 8, 10$, and 12, respectively. For each value of λ, the least Euclid's distance method was used to perform the Monte Carlo simulation as described above. As we see from Fig. 3, the rates of correct prediction are raised with the increase of λ, i.e., the reduction of standard deviations. When $\lambda = 12$, all the standard deviations would be reduced by a factor of 0.9^{12}, i.e., almost equal to zero, and hence the rates of correct prediction for all the folding types are very close to 100%, as expected. It should be underscored, however, that this is just a hypothetical

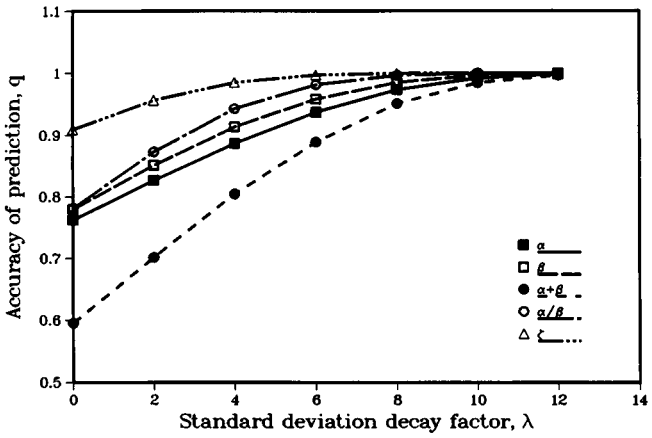

FIGURE 3 Plot of the rate of correct prediction q by the least Euclid's distance method versus the hypothetical decay exponent λ imposed on the standard deviations. During the simulation process, each of the standard deviations listed in Table 3 is multiplied by 0.9^{λ}. As shown in the figure, when $\lambda \geq 12$, the rate of correct prediction for the all folding types would get close to 100%, implying that the hypothetical standard deviations would almost be zero and the corresponding statistical fluctuations vanish.

case, because in reality the standard deviations must exist and cannot be eliminated.

Besides the least Euclid's distance method (Nakashima et al., 1986) and the least Minkowski's distance method (Chou, 1989), there are some other methods, such as those proposed by Klein (1986) and Klein and Delisi (1986), for predicting the folding type of a protein. However, in their method the multidimensional statistical technique of discriminant analysis was used to assign a protein to one of the protein folding types. Accordingly, their methods actually belong to a discriminant analysis method in which the multidimensional statistical technique is used. Moreover, when the Klein and Delisi (1986) method was used to perform prediction, in addition to the amino acid composition, the regular variations in the hydrophobic values of residues along the amino acid sequence was also used as the attribute. This is quite different in selecting parameters from the two demonstrated methods, where the amino acid composition of a protein is the only input in performing the prediction of its folding type. For brevity, we would rather not include their methods as examples for demonstration here since the main goal of this article is in developing the idea of defining an objective criterion in terms of Monte Carlo simulation. Once established, its basic principle and procedure can be applied to examine any other statistical prediction methods.

IV. CONCLUSION

A Monte Carlo simulation method is proposed in an attempt to establish an objective criterion for measuring the accuracy of predicting the folding type of a protein according to its amino acid composition. It has been shown that the asymptotical limit generated by the simulation process can be well defined as the objective accuracy of prediction. Based on that, it has been found that the objective accuracies of prediction for α, β, $\alpha + \beta$, and α/β proteins by the least Euclid's distance method (Nakashima et al., 1986) are 76.3, 78.2, 59.5, and 78.2%, respectively, and those by the least Minkowski's distance method (Chou, 1989) are 74.1, 74.5, 58.1, and 76.7%, respectively. Accordingly, the least Euclid's distance method is better than the least Minkowski's distance method.

The simulation method and the idea developed here also can be applied to examine any other statistical prediction methods.

APPENDIX

The sampling principle as formulated in Eq. 2 can be briefly explained as follows.

The probability of any random variable $R \leq x$ can be expressed as (DeGroot, 1986)

$$P(R \leq x) = \int_{-\infty}^{x} f(R)\, dR \tag{A.1}$$

where $x \in (-\infty, \infty)$, $f(R)$ is the density function of R, and its concrete form will depend on the distribution nature of the random variable R. For a random number $r \in [0, 1]$ as defined in Eq. 2 of the text, the distribution density should be

$$f(r) = \begin{cases} 1, & r \in [0, 1] \\ 0, & \text{otherwise} \end{cases} \tag{A.2}$$

from which the mean of the random number r is

$$\mu = \int_0^1 r f(r)\, dr = \frac{1}{2} \tag{A.3}$$

and its standard deviation is

$$D = \left[\int_0^1 (r-\mu)^2 f(r)\, dr\right]^{1/2} = \left[\int_0^1 (r - 1/2)^2 f(r)\, dr\right]^{1/2} = \sqrt{\frac{1}{12}} \tag{A.4}$$

Setting

$$R(m) = \frac{\sqrt{m}}{D}\left[\frac{1}{m}\sum_{i=1}^{m} r_i - \mu\right] \tag{A.5}$$

then we have that the probability for $R(m) \leq x$ can be written as

$$P\{R(m) \leq x\} = \int_{-\infty}^{x} g(R)\, dR \tag{A.6}$$

where $g(R)$ is the distribution density of $R(m)$. When $m \gg 1$, according to the central limit theorem (DeGroot, 1986), we have

$$\lim_{m \gg 1} P\{R(m) \leq x\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-R^2/2}\, dR \tag{A.7}$$

Comparing Eq. A.6 with Eq. A.5 indicates that, when $m \gg 1$, $g(R)$ is actually a standard normal distribution density, and hence $R = R(m)$ becomes a standard normal random variable when $m \gg 1$. For convenience, letting $m = 12$, we have

$$R(12) = \sum_{i=1}^{12} r_i - 6 = \sum_{i=1}^{6} r_i - \sum_{i=7}^{12} (1 - r_i) \qquad (A.8)$$

Since the density of r and the density of $(1 - r)$ are the same according to Eq. A.1, as far as the statistical distribution is concerned, $R(12)$ also can be equivalently expressed as

$$R(12) = R = \sum_{i=1}^{6} r_i - \sum_{i=7}^{12} r_i \qquad (A.9)$$

which has exactly the same form as Eq. 2 in the text. Therefore, R as given by Eq. 2 belongs to the standard normal distribution.

We are greatly indebted to Dr. Ken Nishikawa, Dr. Hiroshi Nakashima, and Professor Tasuo Ooi for kindly providing the standard deviation data. We would also like to thank Dr. Wei-Zhu Zhong for providing her software to draw Figs. 1–3.

Received for publication 12 May 1992 and in final form 31 July 1992.

REFERENCES

Chou, P. Y. 1980. Amino acid composition of four classes of proteins. *Second Chemical Congress of the North American Continent, Las Vegas.*

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

DeGroot, M. H. 1986. Probability and Statistics. 2nd ed., Addison-Wesley Publishing Company, Reading, MA. 267–278.

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

Klein, P., and C. Delisi. 1986. Prediction of protein structural class from amino acid sequence. *Biopolymers.* 25:1659–1672.

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

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

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

Wada, K., S. Aota, R. Tsuchiya, F. Ishibashi, T. Gojobori, and T. Ikemura. 1990. Codon usage tabulated from the GenBank genetic sequence data. *Nucleic Acids Res.* 18:2367–2411.