Using Grey Dynamic Modeling and Pseudo Amino Acid Composition to Predict Protein Structural Classes

XUAN XIAO,[1] WEI-ZHONG LIN,[1] KUO-CHEN CHOU[2]
[1]*Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333000, China*
[2]*Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, California 92130*

Received 2 October 2007; Revised 7 December 2007; Accepted 3 February 2008
DOI 10.1002/jcc.20955
Published online 31 March 2008 in Wiley InterScience (www.interscience.wiley.com).

Abstract: Using the pseudo amino acid (PseAA) composition to represent the sample of a protein can incorporate a considerable amount of sequence pattern information so as to improve the prediction quality for its structural or functional classification. However, how to optimally formulate the PseAA composition is an important problem yet to be solved. In this article the grey modeling approach is introduced that is particularly efficient in coping with complicated systems such as the one consisting of many proteins with different sequence orders and lengths. On the basis of the grey model, four coefficients derived from each of the protein sequences concerned are adopted for its PseAA components. The PseAA composition thus formulated is called the "grey-PseAA" composition that can catch the essence of a protein sequence and better reflect its overall pattern. In our study we have demonstrated that introduction of the grey-PseAA composition can remarkably enhance the success rates in predicting the protein structural class. It is anticipated that the concept of grey-PseAA composition can be also used to predict many other protein attributes, such as subcellular localization, membrane protein type, enzyme functional class, GPCR type, protease type, among many others.

 J Comput Chem 29: 2018–2024, 2008

Key words: pseudo amino acid composition; grey dynamic model; grey-PseAA; protein structural class; covariant-discriminant algorithm; chou's invariance theorem; misidentification matrix

Introduction

Prediction of protein structural class has attracted the efforts of many investigators[1–21] because the knowledge of structural class of a protein can present an intuitive description of its overall fold, help improve the prediction quality for its secondary structural contents,[22] and reduce the scope of searching its conformational space[23] or provide useful information for conducting the heuristic approach to find its 3D structure.[24,25] Also, prediction of protein structural class has served as a paradigmatic topic; i.e., many concepts and techniques proposed for protein structural class prediction have greatly stimulated the development of predicting various biologically important attributes of proteins.[26] Most of the existing methods for predicting protein structural classification were based on the amino acid (AA) composition.[1,7] As is well known, by using AA composition to represent the sample of a protein, all its sequence-order information would be missing. To avoid complete loss of the sequence order information, the pseudo amino acid (PseAA) composition was introduced[27,28] to represent protein samples. Since then, different kinds of PseAA composition have been developed by many investigators[19,29–68] to deal with varieties of problems in proteins and protein-related systems. Because of its wide usage, recently a very flexible PseAA composition generator, called "PseAAC,"[69] was established at the website http://chou.med.harvard.edu/bioinf/PseAAC/, by which users can generate 63 different kinds of PseAA composition.

In our study, a novel approach called the "grey dynamic modeling" is introduced to derive the PseAA components.

Method

Grey Dynamic Model

In 1982, Deng proposed a grey system theory to study the uncertainty of a system.[70] According to this theory, if the information of a system investigated is fully known, it is called a

Correspondence to: X. Xiao; e-mail: xiaoxuan0326@yahoo.com.cn

Contract/grant sponsor: National Natural Science Foundation of China; contract/grant number: 60661003

Contract/grant sponsor: the Province National Natural Science Foundation; contract/grant number: 0611060

Contract/grant sponsor: the National Education Committee, China; contract/grant number: 20060255006

"white system;" if completely unknown, a "black system;" if partially known, a "grey system."

The grey system theory used for the current study is a special grey dynamic model (GDM) called GM(1,1),[70] which can be concisely described as follows. The great critical feature of the GDM GM(1,1) is the making use of grey generating approaches to reduce the variation of the original data series by transforming the data series linearly. The most commonly applied grey generating approach is the accumulative generation operation (AGO). AGO converts a series lacking any obvious regularity into a strict monotonic increasing series in order to reduce the randomness and increase the smoothness of the series, and minimize interference from the random information. Let us assume that $\mathbf{X}^{(0)} = \left(x^{(0)}(1), x^{(0)}(2), \ldots, x^{(0)}(n)\right)$ is the original series of real numbers with an irregular distribution, and it is a non-negative original data sequence. Then, $\mathbf{X}^{(1)} = \left(x^{(1)}(1), x^{(1)}(2), \ldots, x^{(1)}(n)\right)$ is viewed as the first-order accumulative generation operation (1-AGO) series for $\mathbf{X}^{(0)}$; i.e., the components in $\mathbf{X}^{(1)}$ are given by

$$x^{(1)}(k) = \sum_{i=1}^{k} x^{(0)}(i), \quad k = 1, 2, \ldots, n \tag{1}$$

The GM(1,1) model can be expressed by one variable thru the following first-order grey differential equation:

$$x^{(0)}(k) + az^{(1)}(k) = b, \quad k = 2, \ldots, n \tag{2}$$

where $z^{(1)}(k)$ is called the kth background values for the grey differential equation, and is generated as follows

$$z^{(1)}(1) = x^{(1)}(1) \tag{3}$$

$$z^{(1)}(k) = 0.5x^{(1)}(k) + 0.5x^{(1)}(k-1), \quad k = 2, 3, \ldots, n \tag{4}$$

In eq. (2), the coefficients a and b are called the grey developing and grey input coefficients, respectively. As a matter of fact, $x^{(0)}(k)$ is the k^{th} observation in the data sequence while the term "derivative" used here comes from a treatment that the term $\left(x^{(1)}(k) - x^{(1)}(k-1)\right)/(k-(k-1)) = x^{(0)}(k)$ is an approximation to the true derivative of the function $x^{(1)}(t)$, i.e., $dx^{(1)}/dt$ when $t = k$. The adjective grey used here indicates the grey uncertainty associated with the derivative approximation. Furthermore, the differential equation

$$dx^{(1)}/dt + ax^{(1)} = b \tag{5}$$

is called the "whitenization differential equation" or the "shadow equation" of the grey differential equation [eq. (2)]. By the least-squares method, the coefficients a and b can be derived by the following procedure. Rewriting eq. (2) as

$$x^{(0)}(k) = b + a\left(-z^{(1)}k\right), \quad k = 2, \ldots, n \tag{6}$$

then we obtain a standard matrix from of the eq. (2) as given by

$$\mathbf{Y} = \mathbf{B}\begin{bmatrix} a \\ b \end{bmatrix} \tag{7}$$

where

$$\mathbf{B} = \begin{bmatrix} -z(2) & 1 \\ -z(3) & 1 \\ \vdots & \vdots \\ -z(n) & 1 \end{bmatrix} \text{ and } \mathbf{Y} = [x^{(0)}(2), x^{(0)}(3), \ldots, x^{(0)}(n)]^{\mathbf{T}} \tag{8}$$

where $\mathbf{T}$ is the transpose operator to a matrix, while the least-square estimate for coefficients (a,b) is

$$\begin{bmatrix} a \\ b \end{bmatrix} = (\mathbf{B}^{\mathbf{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathbf{T}}\mathbf{Y} \tag{9}$$

$$a = \frac{\sum_{k=2}^{n} z^{(1)}(k) \sum_{k=2}^{n} x^{(0)}(k) - (n-1)\sum_{k=2}^{n} z^{(1)}(k)x^{(0)}(k)}{(n-1)\sum_{k=2}^{n}\left[z^{(1)}(k)\right]^2 - \left[\sum_{k=2}^{n} z^{(1)}(k)\right]^2} \tag{10}$$

$$b = \frac{\sum_{k=2}^{n}\left[z^{(1)}(k)\right]^2 \sum_{k=2}^{n} x^{(0)}(k) - \sum_{k=2}^{n} z^{(1)}(k) \sum_{k=2}^{n} z^{(1)}(k)x^{(0)}(k)}{(n-1)\sum_{k=2}^{n}\left[z^{(1)}(k)\right]^2 - \left[\sum_{k=2}^{n} z^{(1)}(k)\right]^2} \tag{11}$$

The GM(1,1) model with the differential equation is established on the basis of discrete data. It should be noticed that the minimum number of sample data in the GM(1,1) model is as few as four; i.e., $n = 4$.

We can also derive four intermediate coefficients C, D, E, F as given by

$$C = \sum_{k=2}^{n} z^{(1)}(k), \quad D = \sum_{k=2}^{n} x^{(0)}(k), \quad E = \sum_{k=2}^{n} z^{(1)}(k)x^{(0)}(k), \quad F = \sum_{k=2}^{n} (x^{(1)}(k))^2 \tag{12}$$

where

$$\mathbf{B}^{\mathbf{T}}\mathbf{B}^{-1} = \frac{1}{(n-1)F - C^2}\begin{bmatrix} n-1 & C \\ C & F \end{bmatrix} \text{ and } \mathbf{B}^{\mathbf{T}}\mathbf{Y} = \begin{bmatrix} -E \\ D \end{bmatrix} \tag{13}$$

The least-square estimator for the coefficients [a, b] and the coefficients of [C, D, E, F] should carry some intrinsic information contained in the discrete data sequence $\mathbf{X}^{(0)}$ sampled from the system investigated.[71] On the basis of such belief, the coefficients can be introduced into the PseAA composition in order to make it more effectively reflect the sequence-order effect as well as the feature of sequence pattern. This is the key of the novel approach. The concrete procedures are as follows.

A protein sequence is composed of 20 different types of native amino acids denoted by A, C, D, E, F, G, H, I, K, L, M,

Table 1. Three Different Types for Coding Amino Acids.[a]

Character	Binary	Decimal
K	00110	6
N	01000	8
D	01001	9
E	01010	10
P	01011	11
Q	01100	12
R	01101	13
S	01110	14
T	01111	15
G	10000	16
A	10001	17
H	10010	18
W	10100	20
Y	10101	21
F	10111	23
L	11000	24
M	11010	26
I	11011	27
V	11100	28
C	11110	30

[a]For the derivation of the data listed in the table, see Ref. 72.

N, P, Q, R, S, T, V, W, and Y.[6] Before using the GDM GM,(1,1) we need to represent the protein sequence by a series of real numbers. Listed in Table 1 are the three different kinds of codes used to represent the 20 amino acids. These codes can better reflect the chemical physical properties of an amino acid, as well as its structure and degeneracy.[72] With the decimal codes in Table 1 we can convert a protein sequence to a series of real numbers. Thus, the six coefficients $[a, b, C, D, E, F]$ for any protein sequence can be derived with the grey accumulative model by following eqs. (1–13).

Predicting Algorithm

The six coefficients obtained in the above section can be used for the PseAA components.[27,73] However, of the six coefficients, four are more important because preliminary tests indicated that the highest success rates could be yielded by just using $[b, C, D, E]$. Thus, according Chou's PseAA composition,[27] a protein sample can be expressed by a vector or a point in a 24D (dimensional) space; i.e.

$$\mathbf{X} = (x_1, x_2, \ldots, x_{20}, x_{21}, \ldots, x_{24})^{\mathbf{T}} \tag{14}$$

where

$$x_k = \begin{cases} \frac{f_k}{\sum_{i=1}^{20} f_i + \sum_{j=1}^{4} w_j p_j}, & (1 \le k \le 20) \\ \frac{w_{(k-20)} p_{(k-20)}}{\sum_{i=1}^{20} f_i + \sum_{j=1}^{4} w_j p_j}, & (21 \le k \le 24) \end{cases} \tag{15}$$

where f_i $(i = 1, 2, \ldots, 20)$ are the occurrence frequencies of the 20 amino acids in the protein arranged alphabetically according to their single letter codes,[7] p_j $(j = 1, 2, 3, 4)$ represent the coefficients $b, C, D,$ and E, respectively, and w_j the weight factor for the j^{th} coefficient p_j.

Now the augmented covariant-discriminant (CD) algorithm[16] was adopted to perform the prediction. For reader's convenience, a brief introduction about the augmented CD algorithm is given below.

Suppose a system containing N proteins $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N)$, which have been classified into M subsets (structural classes); i.e.

$$\mathbf{S} = S_1 \cup S_2 \cup S_3 \cup \cdots \cup S_M \tag{16}$$

where each subset S_m $(m = 1, 2, \ldots, M)$ is composed of proteins with the same structural class and its size (the number of proteins therein) is N_m. Obviously, we have $N = N_1 + N_2 + \cdots + N_M$. Now, for a query protein $\mathbf{X}$ [eq. (14)], how can we identify which subset it belongs to. According to the augmented CD algorithm,[74] we can suppose without losing generality that the u-th protein in the subset S_m of eq. (16) is formulated by [see eq. (14)]

$$\mathbf{X}_m^u = \left[x_{m,1}^u \;\; x_{m,2}^u \;\; \cdots \;\; x_{m,20}^u \;\; \cdots \;\; x_{m,24}^u \right]^{\mathbf{T}} \tag{17}$$

where $x_{m,j}^u$ $(j = 1, 2, \ldots, 24)$ is the j-th component of the u-th protein in S_m, and the standard vector for the subset S_m is defined by

$$\bar{\mathbf{X}}_m = \left[\bar{x}_{m,1} \;\; \bar{x}_{m,2} \;\; \cdots \;\; \bar{x}_{m,20} \;\; \cdots \;\; \bar{x}_{m,24} \right]^{\mathbf{T}} \tag{18}$$

where

$$\bar{x}_{m,i} = \frac{1}{N_m} \sum_{u=1}^{N_m} x_{m,i}^u, \quad (i = 1, 2, \ldots, 24) \tag{19}$$

Actually, $\bar{\mathbf{X}}_m$ as defined above can be deemed as a standard protein for the subset S_m. Thus, the similarity between a query protein $\mathbf{X}$ and $\bar{\mathbf{X}}_m$ is defined by the following covariant discriminant function:

$$\mathrm{F}(\mathbf{X}, \bar{\mathbf{X}}_m) = D_{\text{Mah}}^2(\mathbf{X}, \bar{\mathbf{X}}_m) + \ln |\mathbf{C}_m|, \quad (m = 1, 2, \ldots, M) \tag{20}$$

where

$$D_{\text{Mah}}^2(\mathbf{X}, \bar{\mathbf{X}}_m) = (\mathbf{X} - \bar{\mathbf{X}}_m)^{\mathbf{T}} \mathbf{C}_m^{-1} (\mathbf{X}, \bar{\mathbf{X}}_m) \tag{21}$$

is the squared Mahalanobis distance[6,75,76] (The Ref. 76 presents a brief biography of Mahalanobis who was a man of great originality and who made considerable contributions to statistics.) between $\mathbf{X}$ and $\bar{\mathbf{X}}_m$;

$$\mathbf{C}_m = \begin{bmatrix} c_{1,1}^m & c_{1,2}^m & \cdots & c_{1,24}^m \\ c_{2,1}^m & c_{2,2}^m & \cdots & c_{2,24}^m \\ \vdots & \vdots & \ddots & \vdots \\ c_{24,1}^m & c_{24,2}^m & \cdots & c_{24,24}^m \end{bmatrix} \tag{22}$$

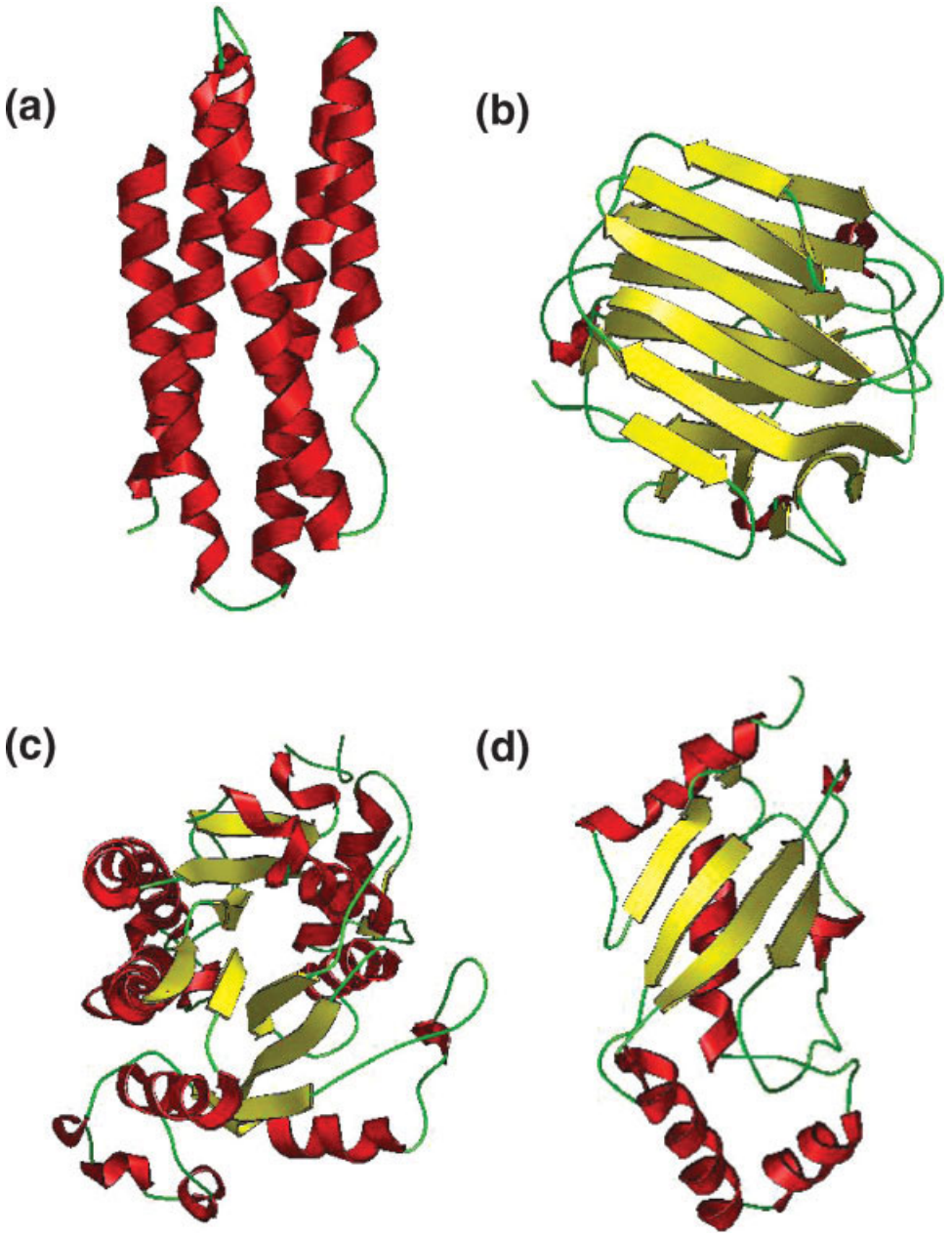

Figure 1. Ribbon drawings to show the four structural classes of proteins: (a) all-α, (b) all-β, (c) α/β, and (d) $\alpha + \beta$. Reproduced from Ref. 16 with permission.

is the covariance matrix for the subset S_m; the 24 × 24 elements in $\mathbf{C}_m$ are given by

$$c_{i,j}^{m} = \frac{1}{N_m - 1} \sum_{u=1}^{N_m} \left(x_{m,i}^{u} - \bar{x}_{m,i} \right) \left(x_{m,j}^{u} - \bar{x}_{m,j} \right), \quad (i,j = 1,2,\ldots,24) \tag{23}$$

and $|\mathbf{C}_m|$ is the determinant of the matrix $\mathbf{C}_\mathrm{m}$. The smaller the value of F $(\mathbf{X}, \bar{\mathbf{X}}_m)$, the higher the similarity between $\mathbf{X}$ and $\bar{\mathbf{X}}_m$. Therefore, the query protein is predicted belonging to the subset S_μ or the μ-th class if

$$\mu = \mathbf{arg\,min}_m \{ \mathrm{F}(\mathbf{X}, \bar{\mathbf{X}}_m) \}, \quad (m = 1,2,\ldots,M) \tag{24}$$

where μ is the argument of m that minimizes $\mathrm{F}(\mathbf{X}, \bar{\mathbf{X}}_m)$. If there are two and more arguments leading to a same minimum value for $\mathrm{F}(\mathbf{X}, \bar{\mathbf{X}}_m)$, the query protein will be randomly assigned to one of the subcellular locations associated with these arguments although this kind of tie case rarely happens. Note that

Table 2. Success Rates of Jackknife Cross-Validation with Different Approaches on the 204 Proteins From Ref. 16.

Method	All- α	All- β	α/β	$\alpha+\beta$	Overall
Supervised fuzzy clustering [a]	$\frac{38}{52} = 73.1\%$	$\frac{55}{61} = 90.2\%$	$\frac{28}{45} = 62.2\%$	$\frac{29}{46} = 63.1\%$	$\frac{150}{204} = 73.5\%$
AA correlation approach[b]	$\frac{49}{52} = 94.2\%$	$\frac{53}{61} = 86.9\%$	$\frac{22}{45} = 48.9\%$	$\frac{41}{46} = 89.1\%$	$\frac{165}{204} = 80.9\%$
AAPCA[c]	$\frac{42}{52} = 82.69\%$	$\frac{59}{61} = 96.72\%$	$\frac{35}{45} = 77.78\%$	$\frac{38}{46} = 82.61\%$	$\frac{174}{204} = 85.29\%$
Complexity measure factor[d]	$\frac{43}{52} = 82.7\%$	$\frac{55}{61} = 90.2\%$	$\frac{45}{45} = 100\%$	$\frac{40}{46} = 87.0\%$	$\frac{183}{204} = 89.7\%$
Grey modeling[e]	$\frac{46}{52} = 88.46\%$	$\frac{59}{61} = 96.72\%$	$\frac{45}{45} = 100\%$	$\frac{41}{46} = 89.13\%$	$\frac{191}{204} = 93.63\%$

[a]Based on the AA composition.[17]
[b]Considering amino acid correlation[80] and using the covariant discriminant algorithm.[78,81]
[c]Using the AA principal component analysis.[21]
[d]Using the complexity measure factor for the PseAA component[19] and augmented covariant discriminant algorithm for the prediction engine.[74]
[e]Using the four coefficients derived from the grey system theory for PseAA components as described in this paper and the augmented covariant discriminant algorithm[74] for the prediction engine. The weight factors for the four PseAA components are $w_1 = 1/1000$, $w_2 = 1/(3 \times 10^7)$, $w_3 = 1/10^6$, and $w_4 = 1/10^9$ respectively [see eq. (15)].

owing to the normalization condition imposed by eq. (15), of the 24 components in eq. (14), only 23 are independent, and hence the covariance matrix $\mathbf{C}_m$ as defined by eq. (23) must be a singular one.[6] This would lead the Mahalanobis distance defined by eq. (21) and the covariant discriminant function by eq. (20) to be divergent and meaningless. To cope with such a situation, the dimension-reducing procedure[7] was adopted in practical calculations; i.e., instead of 24D space, a protein sample is defined in a (24-1)-D space by leaving out one of its 24 components. The remaining 24-1 components would be completely independent, thereby the corresponding covariance matrix $\mathbf{C}_m$ being no longer singular. In other words, the Mahalanobis distance [eq. (21)] and the covariant discriminant function [eq. (20)] based on such a (24-1)-D space can be uniquely defined without any trouble. However, a question might be raised: which one of the 24 components can be left out? The answer is: anyone. Will it lead to a different predicted result by leaving out a different component? The answer is: no. According to the Chou's invariance theorem (see Appendix A of Ref. 7), both the value of the Mahalanobis distance and the value of the determinant of $\mathbf{C}_m$ will remain exactly the same regardless of which one of the 24 components is left out. Accordingly, the final value of the covariant discriminant function [eq. (20)] can be uniquely defined through such a dimension-reducing procedure. For more details of the CD algorithm and augmented CD algorithm, the reader is referred to the papers.[19,26,42,77–79]

Results and Discussion

As a demonstration, let us use the same dataset[16] studied by many previous investigators (see, e.g., Refs. 17, 19, 21, 80). It consists of 204 proteins, of which 52 all-α, 61 all-β, 45 α/β, and 46 $\alpha+\beta$ (see Fig. 1). Their PDB codes are given in Table 2 of Chou.[16]

In our study, we adopt the jackknife test to examine the grey-PseAA approach. As is well known, in statistical prediction the sub-sampling test and jackknife test are the two methods often used in literatures for examining the accuracy of a predictor.[82] However, as illustrated by eq. 50 in a recent review article,[83] the sub-sampling (e.g., 5-fold cross-validation) test cannot avoid arbitrariness even for a very simple benchmark dataset. Accordingly, the jackknife test has been widely and rapidly increasingly adopted by investigators (for example, see Refs. 11, 14, 20, 33, 39, 40, 42, 44–48, 50–52, 56, 58, 60–62, 72, 73, 79, 84–102) to test the power of varieties of predictors. In the jackknife test, each protein sample in the dataset is singled out in turn as a "test sample" and all the rule-parameters are determined from the remaining $N-1$ samples. The success rates by jackknife test for the aforementioned 204 proteins classified into four structural classes are given in Table 2, where for facilitating comparison the corresponding rates obtained by the recently developed algorithms, such as the correlation analysis approach,[80] the supervised fuzzy clustering approach,[17] the amino acid principal component analysis (AAPCA),[21] and the complexity factor approach[19] are also listed. It can be seen from Table 2 that the overall success rate by the current approach is 93.63%, which is remarkably higher than those by the other approaches.

To illustrate the difference between the current grey dynamic modeling (GDM) and the complexity measure factor[19] (CMF), the misidentification matrices are constructed according to the jackknife test results by the two methods on the 204 proteins, as shown below:

$$\text{GDM} \Rightarrow \begin{bmatrix} & 6 \in \alpha & 2 \in \beta & 0 \in \alpha/\beta & 5 \in \alpha+\beta \\ \alpha & 0 & 0 & 0 & 4 \\ \beta & 0 & 0 & 0 & 1 \\ \alpha/\beta & 0 & 0 & 0 & 0 \\ \alpha+\beta & 6 & 2 & 0 & 0 \end{bmatrix} \quad (25)$$

and

$$\text{CMF} \Rightarrow \begin{bmatrix} & 9 \in \alpha & 6 \in \beta & 0 \in \alpha/\beta & 6 \in \alpha+\beta \\ \alpha & 0 & 1 & 0 & 5 \\ \beta & 0 & 0 & 0 & 0 \\ \alpha/\beta & 0 & 1 & 0 & 1 \\ \alpha+\beta & 9 & 4 & 0 & 0 \end{bmatrix} \quad (26)$$

Table 3. The MCC Indexes Obtained by the Jackknife Tests with the Current Classifier and the CMF Approach on the 204 Proteins From Ref. 16.

	Matthew's correlation coefficient	
Protein structural classes	Complexity measure factor[a]	Grey modeling[b]
all-α	0.803	0.870
all-β	0.930	0.965
α/β	0.972	1.000
$\alpha+\beta$	0.750	0.822

[a]Using the complexity measure factor for the PseAA component[19] and augmented covariant discriminant algorithm for the prediction engine.[74]
[b]Using the four coefficients derived from the grey system theory for PseAA components as described in this article and the augmented covariant discriminant algorithm.[74]

where the figures in the 1st row indicate the numbers of proteins failed to be predicted to the class they actually belong to, and the figures in the other rows indicate the numbers of proteins incorrectly predicted to the corresponding class. It can be seen by comparing eqs. (16) and (17) that the misallocation errors are remarkably reduced by GDM against CMF, particularly for β and α proteins.

List in Table 3 are the Matthew's correlation coefficient (MCC) indexes for the four structural classes obtained by the jackknife tests with the GDM and CMF predictor, respectively. The definition of MCC is given by

$$\mathrm{MCC} = \frac{(\mathrm{TP})(\mathrm{TN}) - (\mathrm{FP})(\mathrm{FN})}{\sqrt{[\mathrm{TP}+\mathrm{FP}][\mathrm{TP}+\mathrm{FN}][\mathrm{TN}+\mathrm{FP}][\mathrm{TN}+\mathrm{FN}]}} \quad (27)$$

where TP represents the true positive; TN, the true negative; FP, the false positive and FN, the false negative.[89,103] It can be seen from Table 3 that the results obtained by the current predictor not only possess higher success rates but also are more stable than those by the CMF approach, indicating that the new approach is indeed very powerful and promising.

Conclusion

The grey system theory was developed to deal with those systems for which only partial information is available, and hence is particularly useful to deal with biological problems. It is demonstrated in this study that the overall success rate in predicting protein structural classes can be remarkably improved by using the grey dynamic modeling and PseAA composition approach to represent the protein samples. It has not escaped our notice that the Grey-PseAA approach can be also used to deal with many other complicated problems in biology, such as predicting protein subcellular localization,[83] membrane protein type,[91] enzyme functional class,[104,105] GPCR type,[81] signal peptides,[89,106] protease type,[107] among many others.

Acknowledgments

The author would like to take this opportunity to express their gratitude to the two anonymous reviewers whose constructive comments were very helpful for strengthening the presentation of this study.

References

1. Nakashima, H.; Nishikawa, K.; Ooi, T. J Biochem 1986, 99, 152.
2. Chou, P. Y. In Prediction of Protein Structure and the Principles of Protein Conformation; Fasman, G. D., Ed.; Plenum: New York, 1989; pp. 549–586.
3. Klein, P.; Delisi, C. Biopolymers 1986, 25, 1659.
4. Klein, P. Biochim Biophys Acta 1986, 874, 205.
5. Deleage, G.; Roux, B. Protein Eng 1987, 1, 289.
6. Chou, K. C.; Zhang, C. T. J Biol Chem 1994, 269, 22014.
7. Chou, K. C. Prot Struct Funct Genet 1995, 21, 319.
8. Kneller, D. G.; Cohen, F. E.; Langridge, R. J Mol Biol 1990, 214, 171.
9. Metfessel, B. A.; Saurugger, P. N.; Connelly, D. P.; Rich, S. T. Prot Sci 1993, 2, 1171.
10. Chandonia, J. M.; Karplus, M. Prot Sci 1995, 4, 275.
11. Zhou, G. P. J Prot Chem 1998, 17, 729.
12. Zhou, G. P.; Assa-Munt, N. Prot Struct Funct Genet 2001, 44, 57.
13. Luo, R. Y.; Feng, Z. P.; Liu, J. K. Eur J Biochem 2002, 269, 4219.
14. Kedarisetti, K. D.; Kurgan, L. A.; Dick, S. Biochem Biophys Res Commun 2006, 348, 981.
15. Kurgan, L.; Homaeian, L. Pattern Recogn Lett 2006, 39, 2323.
16. Chou, K. C. Biochem Biophys Res Commun 1999, 264, 216.
17. Shen, H. B.; Yang, J.; Liu, X. J.; Chou, K. C. BBRC 2005, 334, 577.
18. Feng, K. Y.; Cai, Y. D.; Chou, K. C. Biochem Biophys Res Commun 2005, 334, 213.
19. Xiao, X.; Shao, S. H.; Huang, Z. D.; Chou, K. C. J Comput Chem 2006, 27, 478.
20. Niu, B.; Cai, Y. D.; Lu, W. C.; Zheng, G. Y.; Chou, K. C. Prot Pept Lett 2006, 13, 489.
21. Du, Q. S.; Jiang, Z. Q.; He, W. Z.; Li, D. P.; Chou, K. C. J Biomol Struct Dynam 2006, 23, 635.
22. Zhang, C. T.; Zhang, Z.; He, Z. J Prot Chem 1998, 17, 261.
23. Chou, K. C. J Mol Biol 1992, 223, 509.
24. Carlacci, L.; Chou, K. C.; Maggiora, G. M. Biochemistry 1991, 30, 4389.
25. Chou, K. C. Curr Med Chem 2004, 11, 2105.
26. Chou, K. C. Curr Prot Pept Sci 2005, 6, 423.
27. Chou, K. C. Prot Struct Funct Genetics (Erratum: ibid., 2001, 44, 60; 2001, 43, 246.
28. Chou, K. C. Bioinformatics 2005, 21, 10.
29. Pan, Y. X.; Zhang, Z. Z.; Guo, Z. M.; Feng, G. Y.; Huang, Z. D.; He, L. J Prot Chem 2003, 22, 395.
30. Wang, M.; Yang, J.; Liu, G. P.; Xu, Z. J.; Chou, K. C. Prot Eng Des Sel 2004, 17, 509.
31. Liu, H.; Wang, M.; Chou, K. C. Biochem Biophys Res Commun 2005, 336, 737.
32. Shen, H. B.; Chou, K. C. Biochem Biophys Res Commun 2005, 337, 752.
33. Gao, Y.; Shao, S. H.; Xiao, X.; Ding, Y. S.; Huang, Y. S.; Huang, Z. D.; Chou, K. C. Amino Acids 2005, 28, 373.
34. Shen, H. B.; Chou, K. C. Biochem Biophys Res Commun 2005, 334, 288.
35. Liu, H.; Yang, J.; Wang, M.; Xue, L.; Chou, K. C. Prot J 2005, 24, 385.
36. Wang, S. Q.; Yang, J.; Chou, K. C. J Theor Biol 2006, 242, 941.
37. Shen, H. B.; Chou, K. C. Bioinformatics 2006, 22, 1717.

38. Zhang, T.; Ding, Y.; Chou, K. C. Comput Biol Chem 2006, 30, 367.
39. Chen, C.; Zhou, X.; Tian, Y.; Zou, X.; Cai, P. Anal Biochem 2006, 357, 116.
40. Chen, C.; Tian, Y. X.; Zou, X. Y.; Cai, P. X.; Mo, J. Y. J Theor Biol 2006, 243, 444.
41. Cai, Y. D.; Chou, K. C. J Theor Biol 2006, 238, 395.
42. Xiao, X.; Shao, S. H.; Ding, Y. S.; Huang, Z. D.; Chou, K. C. Amino Acids 2006, 30, 49.
43. Shen, H. B.; Yang, J.; Chou, K. C. J Theor Biol 2006, 240, 9.
44. Zhang, S. W.; Pan, Q.; Zhang, H. C.; Shao, Z. C.; Shi, J. Y. Amino Acids 2006, 30, 461.
45. Du, P.; Li, Y. BMC Bioinform 2006, 7, 518.
46. Mondal, S.; Bhavna, R.; Babu, M. R.; Ramakumar, S. J Theor Biol 2006, 243, 252.
47. Lin, H.; Li, Q. Z. Biochem Biophys Res Commun 2007, 354, 548.
48. Lin, H.; Li, Q. Z. J Comput Chem 2007, 28, 1463.
49. Pu, X.; Guo, J.; Leung, H.; Lin, Y. J Theor Biol 2007, 247, 259.
50. Shi, J. Y.; Zhang, S. W.; Pan, Q.; Cheng, Y.-M.; Xie, J. Amino Acids 2007, 33, 69.
51. Shen, H. B.; Yang, J.; Chou, K. C. Amino Acids 2007, 33, 57.
52. Shen, H. B.; Chou, K. C. Amino Acids 2007, 32, 483.
53. Chen, Y. L.; Li, Q. Z. J Theor Biol 2007, 248, 377.
54. Chen, Y. L.; Li, Q. Z. J Theor Biol 2007, 245, 775.
55. Kurgan, L. A.; Stach, W.; Ruan, J. J Theor Biol 2007, 248, 354.
56. Zhou, X. B.; Chen, C.; Li, Z. C.; Zou, X. Y. J Theor Biol 2007, 248, 546.
57. Mundra, P.; Kumar, M.; Kumar, K. K.; Jayaraman, V. K.; Kulkarni, B. D. Pattern Recogn Lett 2007, 28, 1610.
58. Zhang, T. L.; Ding, Y. S. Amino Acids 2007, 33, 623–629.
59. Diao, Y.; Li, M.; Feng, Z.; Yin, J.; Pan, Y. J Theor Biol 2007, 247, 608.
60. Li, F. M.; Li, Q. Z. Amino Acids 2008, 34, 119–125.
61. Fang, Y.; Guo, Y.; Feng, Y.; Li, M. Amino Acids 2008, 34, 103–109.
62. Chou, K. C.; Shen, H. B. Nature Protocols 2008, 3, 153–162.
63. Gonzalez-Diaz, H.; Vilar, S.; Santana, L.; Uriarte, E. Curr Top Med Chem 2007, 10, 1015.
64. Gonzalez-Diaz, H.; Perez-Castillo, Y.; Podda, G.; Uriarte, E. J Comput Chem 2007, 28, 1990.
65. Gonzalez-Diaz, H.; Aguero-Chapin, G.; Varona, J.; Molina, R.; Delogu, G.; Santana, L.; Uriarte, E.; Podda, G. J Comput Chem 2007, 28, 1049.
66. Caballero, J.; Fernandez, L.; Garriga, M.; Abreu, J. I.; Collina, S.; Fernandez, M. J Mol Graph Model 2007, 26, 166.
67. Aguero-Chapin, G.; Gonzalez-Diaz, H.; Molina, R.; Varona-Santos, J.; Uriarte, E.; Gonzalez-Diaz, Y. FEBS Lett 2006, 580, 723.
68. Gonzalez-Diaz, H.; Perez-Bello, A.; Uriarte, E.; Gonzalez-Diaz, Y. Bioorg Med Chem Lett 2006, 16, 547.
69. Shen, H. B.; Chou, K. C. Anal Biochem 2008, 373, 386–388.
70. Deng, J. L. Sys Control Lett 1985, 1, 288.
71. Guo, R. Comput Intell Reliab Eng 2007, 40, 387.
72. Xiao, X.; Chou, K. C. Prot Pept Lett 2007, 14, 871.
73. Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Chen, X.; Chou, K. C. Amino Acids 2005, 28, 29.
74. Chou, K. C. Biochem Biophys Res Commun 2000, 278, 477.
75. Mahalanobis, P. C. Proc Natl Inst Sci India 1936, 2, 49.
76. Pillai, K. C. S. In Encyclopedia of Statistical Sciences, Vol. 5; Kotz, S.; Johnson, N. L., Eds.; Wiley: New York, 1985; pp. 176–181.
77. Chou, K. C.; Maggiora, G. M. Prot Eng 1998, 11, 523.
78. Chou, K. C.; Elrod, D. W. Prot Eng 1999, 12, 107.
79. Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Huang, Y.; Chou, K. C. Amino Acids 2005, 28, 57.
80. Du, Q. S.; Wei, D. Q.; Chou, K. C. Peptides 2003, 24, 1863.
81. Chou, K. C. J Proteome Res 2005, 4, 1413.
82. Chou, K. C.; Zhang, C. T. Crit Rev Biochem Mol Biol 1995, 30, 275.
83. Chou, K. C.; Shen, H. B. Anal Biochem 2007, 370, 1.
84. Zhou, G. P.; Doctor, K. Prot Struct Funct Genet 2003, 50, 44.
85. Guo, Y. Z.; Li, M.; Lu, M.; Wen, Z.; Wang, K.; Li, G.; Wu, J. Amino Acids 2006, 30, 397.
86. Sun, X. D.; Huang, R. B. Amino Acids 2006, 30, 469.
87. Wen, Z.; Li, M.; Li, Y.; Guo, Y.; Wang, K. Amino Acids 2006, 32, 277.
88. Liu, D. Q.; Liu, H.; Shen, H. B.; Yang, J.; Chou, K. C. Amino Acids 2007, 32, 493.
89. Chen, J.; Liu, H.; Yang, J.; Chou, K. C. Amino Acids 2007, 33, 423.
90. Chou, K. C.; Shen, H. B. Biochem Biophys Res Commun 2007, 357, 633.
91. Chou, K. C.; Shen, H. B. Biochem Biophys Res Comm 2007, 360, 339.
92. Ding, Y. S.; Zhang, T. L.; Chou, K. C. Prot Pept Lett 2007, 14, 811.
93. Wang, M.; Yang, J.; Chou, K. C. Amino Acids (Erratum, ibid. 2005, 29, 301) 2005, 28, 395.
94. Diao, Y.; Ma, D.; Wen, Z.; Yin, J.; Xiang, J.; Li, M. Amino Acids 2008, 34, 111–117.
95. Tan, F.; Feng, X.; Fang, Z.; Li, M.; Guo, Y.; Jiang, L. Amino Acids 2007, 33, 669–675.
96. Chou, K. C.; Shen, H. B. Biochem Biophys Res Commun 2006, 347, 150.
97. Chou, K. C.; Shen, H. B. J Proteome Res 2006, 5, 1888.
98. Chou, K. C.; Shen, H. B. J Cell Biochem 2007, 100, 665.
99. Shen, H. B.; Chou, K. C. Biopolymers 2007, 85, 233.
100. Shen, H. B.; Chou, K. C. Biochem Biophys Res Commun 2007, 355, 1006.
101. Chou, K. C.; Shen, H. B. J Proteome Res 2007, 6, 1728.
102. Shen, H. B.; Chou, K. C. Prot Eng Design Sel 2007, 20, 39.
103. Shen, H. B.; Chou, K. C. Prot Eng Design Sel 2007, 20, 561.
104. Chou, K. C.; Elrod, D. W. J Proteome Res 2003, 2, 183.
105. Shen, H. B.; Chou, K. C. Biochem Biophys Res Commun 2007, 364, 53.
106. Shen, H. B.; Chou, K. C. Biochem Biophys Res Commun 2007, 363, 297.
107. Chou, K. C.; Cai, Y. D. Biochem Biophys Res Commun 2006, 339, 1015.