Journal of Theoretical Biology 254 (2008) 691–696

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Predicting protein structural classes with pseudo amino acid composition: An approach using geometric moments of cellular automaton image

Xuan Xiao [a,*], Pu Wang [a], Kuo-Chen Chou [b]

[a] Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 33300, China
[b] Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, USA

ARTICLE INFO

Article history:
Received 2 April 2008
Received in revised form
18 June 2008
Accepted 18 June 2008
Available online 24 June 2008

Keywords:
Cellular automaton
Space–time evolution
Image texture
Geometric invariant moment
Pseudo amino acid composition
Covariant-discriminant algorithm
Chou's invariant theorem

ABSTRACT

A novel approach was developed for predicting the structural classes of proteins based on their sequences. It was assumed that proteins belonging to the same structural class must bear some sort of similar texture on the images generated by the cellular automaton evolving rule [Wolfram, S., 1984. Cellular automation as models of complexity. Nature 311, 419–424]. Based on this, two geometric invariant moment factors derived from the image functions were used as the pseudo amino acid components [Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: Struct., Funct., Genet. (Erratum: ibid., 2001, vol. 44, 60) 43, 246–255] to formulate the protein samples for statistical prediction. The success rates thus obtained on a previously constructed benchmark dataset are quite promising, implying that the cellular automaton image can help to reveal some inherent and subtle features deeply hidden in a pile of long and complicated amino acid sequences.

1. Introduction

Although the details of the 3-D (dimensional) structures of proteins are extremely complicated and irregular, their overall topological folding patterns are surprisingly simple and regular. In view of this, proteins are generally classified into a limited number of different structural classes, and typically into four structural classes: all-α, all-β, α/β, and $\alpha+\beta$ (Levitt and Chothia, 1976) although with more data available recently proteins can be also further classified into 11 classes (Andreeva et al., 2004) of which at least 7 classes are highly populated with low sequence homology within the same class (Chou and Cai, 2004).

The structural class is an important attribute used to characterize the overall folding type of a protein. Therefore, prediction of the structural class has attracted many investigators (see, e.g., Cao et al., 2006; Chandonia and Karplus, 1995; Chen et al., 2006a, b, 2008a, b; Chou, 1989, 1995; Chou and Zhang, 1994; Chou and Maggiora, 1998; Deleage and Roux, 1987; Jahandideh et al., 2007; Kedarisetti et al., 2006; Klein, 1986; Klein and Delisi, 1986; Kneller et al., 1990; Kurgan and Homaeian, 2006; Kurgan et al., 2007, 2008; Lin and Li, 2007b; Liu and Chou, 1998; Luo et al., 2002; Mao et al., 1994; Metfessel et al., 1993; Nakashima et al., 1986; Shen et al., 2005; Sun and Huang, 2006; Zhang et al., 1995; Zhang and Ding, 2007; Zhou, 1998; Zhou and Assa-Munt, 2001). Although various different algorithms were used by these investigators, they can be basically categorized into the following two groups. One is based on the amino acid (AA) composition, and the other based on the pseudo amino acid (PseAA) composition (Chou, 2005b). Although the amino acid composition model is simpler and easier to handle, it fails to incorporate any of the sequence-order information in a protein. To avoid the complete loss of the sequence-order information as suffered in the amino acid composition model (Chou, 1995; Nakashima et al., 1986), the PseAA composition was introduced.

The concept of PseAA composition was originally proposed for improving the prediction quality of protein subcellular localization and membrane protein type (Chou, 2001). The essence of PseAA composition is to keep using a discrete model to represent a protein sample, yet without completely losing its sequence-order information. According to its definition, the PseAA composition for a given protein sample is expressed by a set of $20+\lambda$ discrete numbers, where the first 20 represent the 20 components of the classical amino acid composition while the additional λ numbers incorporate some of its sequence-order information via different kinds of coupling modes.

Ever since the concept of PseAA composition was introduced, various PseAA composition approaches have been proposed to deal with different problems in proteins and protein-related systems (see, e.g., Chen et al., 2006a, b; Chen and Li, 2007a, b; Ding et al., 2007; Du and Li, 2006; Fang et al., 2008; Gonzalez-Diaz

* Corresponding author. Tel.: +86 13879809729; fax: +86 798 8499671.
E-mail address: xiaoxuan0326@yahoo.com.cn (X. Xiao).

0022-5193/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2008.06.016

et al., 2007; González-Díaz et al., 2008; Jiang et al., 2008; Li and Li, 2008; Lin, 2008; Lin and Li, 2007a,b; Lin et al., 2008; Mondal et al., 2006; Mundra et al., 2007; Nanni and Lumini, 2008; Pu et al., 2007; Shi et al., 2007; Zhang et al., 2008; Zhou et al., 2007). Owing to its wide usage, recently a very flexible PseAA composition generator, called "PseAAC" (Shen and Chou, 2008), was established at the website http://chou.med.harvard.edu/bioinf/PseAAC/, by which users can generate 63 different kinds of PseAA composition.

To successfully use the PseAA composition for predicting various attributes of proteins, the key is how to optimally extract the features for the PseAA components. The present study was initiated in an attempt to introduce a completely different approach, the so-called "geometric invariant moment" of protein cellular automaton image to address this problem.

2. Method

A protein sequence is formed by 20 native amino acids whose single character codes are: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y. It is very difficult to find its characteristic pattern particularly when the sequence is very long. To cope with this situation, we resort to the images derived from the amino acid sequence by means of the space–time evolution of cellular automaton (Wolfram, 1984, 2002).

Suppose a protein **P** consists of N amino acids; i.e.,

$$\mathbf{P} = R_1 R_2 \cdots R_N \tag{1}$$

Table 1
Three different types for coding amino acids

Type	Code									
Character	P	L	Q	H	R	S	F	Y	W	C
Decimal	1	3	4	5	6	9	11	12	14	15
Binary	00001	00011	00100	00101	00110	01001	01011	01100	01110	01111
Character	T	I	M	K	N	A	V	D	E	G
Decimal	16	18	19	20	21	25	26	28	29	30
Binary	10000	10010	10011	10100	10101	11001	11010	11100	11101	11110

where R_1 represents the first residue of the protein, R_2 the second residue, and so forth. To transform a protein sequence from a character code to a numerical one, we adopted the code-converting relation as given in Table 1, which can better reflect the chemical and physical properties of an amino acid, as well as its structure and degeneracy, as detailed in Xiao et al. (2005). If each of the constituent amino acids in the protein **P** is coded in a binary code according to Table 1, the protein sequence will be transformed to a serial of $5N$ elements, where the elements are either 0 or 1. For example, the sequence "PLQHRS…" is accordingly transformed to "000010001100100001010011001001…". Each of these elements can be treated as a pixel with "0" for "white" and "1" for "black", then by following the space–time evolution procedures as described in Xiao et al. (2005), the protein **P** would correspond to a cellular automaton image, as shown in Fig. 1, where panel (a) is the cellular automaton image generated from an all-α protein, panel (b) from an all-β protein, panel (c) from an α/β protein, and panel (d) from an α+β protein.

According to Wolfram's theory (2002), those proteins which belong to the same structural class should have some similar textures in their cellular automaton images. However, how to optimally extract these features and formulate them as a set of parameters is an important problem yet to be solved. Here, we introduce the 2-D geometric moment approach to deal with this problem. First of all, let us define a 2-D image function, as given below:

$$f(x,y) = \begin{cases} 0, & \text{when the pixel at } (x,y) \text{ is white} \\ 1, & \text{when the pixel at } (x,y) \text{ is black} \end{cases} \quad (x = 0,1,\ldots,\Omega;\ y = 1,2,\ldots,5N) \tag{2}$$

where $\Omega = 100$ is the total number of time evolution cycles because the texture of a cellular automaton image thus generated was basically steady after 100 evolution cycles.

Geometric moments are the most popular types of moments and have been frequently used for a number of image processing tasks. The two-dimensional geometric moment τ_{pq} of order $(p+q)$ for image $f(x,y)$ is given by:

$$\tau_{pq} = \sum_{x=0}^{\Omega} \sum_{y=1}^{5N} x^p y^q f(x,y) \quad (p,q = 0,1,2) \tag{3}$$

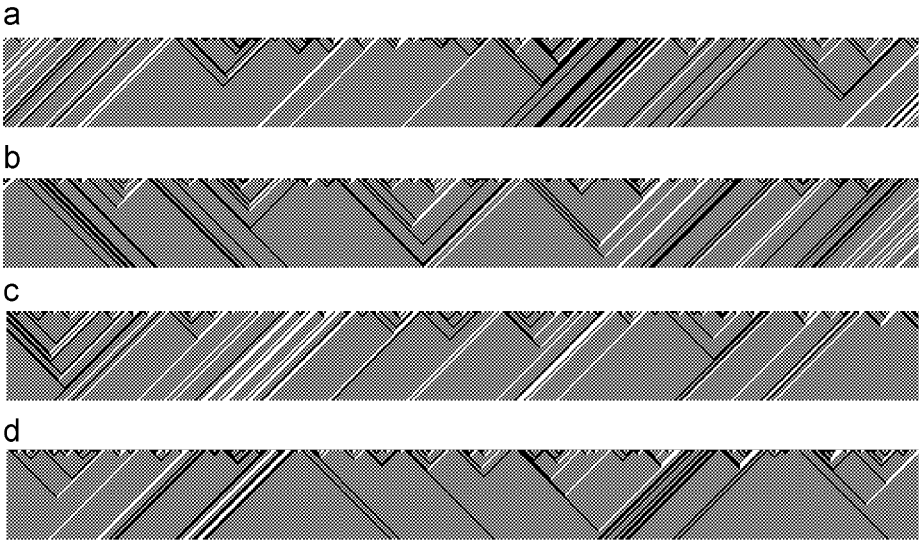

Fig. 1. The image generated by the cellular automaton evolving rule for a protein sequence in: (a) the all-α structural class, (b) the all-β structural class, (c) the α/β structural class, and (d) the α+β structural class.

The image centroid is given by:

$$\begin{cases} x_c = \dfrac{\tau_{10}}{\tau_{00}} = \dfrac{\sum_{x=0}^{\Omega}\sum_{y=1}^{5N} xf(x,y)}{\sum_{x=0}^{\Omega}\sum_{y=1}^{5N} f(x,y)} \\ y_c = \dfrac{\tau_{01}}{\tau_{00}} = \dfrac{\sum_{x=0}^{\Omega}\sum_{y=1}^{5N} yf(x,y)}{\sum_{x=0}^{\Omega}\sum_{y=1}^{5N} f(x,y)} \end{cases} \quad (4)$$

When the geometric moments τ_{pq} in Eq. (3) are referred to the image centroid (x_c, y_c), they become the central moments τ^c_{pq}, as given by:

$$\tau^c_{pq} = \sum_{x=0}^{\Omega}\sum_{y=1}^{5N}(x-x_c)^p(y-y_c)^q f(x,y) \quad (p,q=0,1,2) \quad (5)$$

The central moments τ^c_{pq} are invariant to any spatial translation or rotation of the image; to make them invariant to the area scaling as well (Rizon et al., 2006), let us take the normalized form as given by Hu (1962):

$$\tau^0_{pq} = \frac{\tau^c_{pq}}{(\tau^c_{00})^{(p+q+2)/2}} \quad (p,q=0,1,2) \quad (6)$$

Based on Eq. (6), several RTS (rotation, translation, size-scaling) invariant functions were defined by Hu (1962), as given by:

$$\varphi_1 = \tau^0_{20} + \tau^0_{02} \quad (7)$$

$$\varphi_2 = (\tau^0_{20} - \tau^0_{02})^2 + 4(\tau^0_{11})^2 \quad (8)$$

$$\varphi_3 = (\tau^0_{30} - 3\tau^0_{12})^2 + (3\tau^0_{21} - \tau^0_{03})^2 \quad (9)$$

$$\varphi_4 = (\tau^0_{30} + \tau^0_{12})^2 + (\tau^0_{03} + \tau^0_{21})^2 \quad (10)$$

$$\varphi_5 = (\tau^0_{30} - 3\tau^0_{12})(\tau^0_{30} + \tau^0_{12})[(\tau^0_{30} + \tau^0_{12})^2 - 3(\tau^0_{21} + \tau^0_{03})^2] + (3\tau^0_{21} - \tau^0_{03})(\tau^0_{21} + \tau^0_{03})[3(\tau^0_{30} + \tau^0_{12})^2 - (\tau^0_{21} + \tau^0_{03})^2] \quad (11)$$

$$\varphi_6 = (\tau^0_{20} - \tau^0_{02})[(\tau^0_{30} + \tau^0_{12})^2 - (\tau^0_{21} + \tau^0_{03})^2] + 4\tau^0_{11}(\tau^0_{30} + \tau^0_{12})(\tau^0_{21} + \tau^0_{03}) \quad (12)$$

$$\varphi_7 = (3\tau^0_{21} - \tau^0_{03})(\tau^0_{30} + \tau^0_{12})[(\tau^0_{30} + \tau^0_{12})^2 - 3(\tau^0_{03} + \tau^0_{21})^2] - (3\tau^0_{12} - \tau^0_{30})(\tau^0_{03} + \tau^0_{21})[3(\tau^0_{30} + \tau^0_{12})^2 - (\tau^0_{03} + \tau^0_{21})^2] \quad (13)$$

Of the above seven invariant functions formed by the RTS invariant central moments, only the first two as given by Eqs. (7) and (8) were used as the PseAA components (Chou, 2001). This is because preliminary tests indicated that incorporation of the other five invariant functions (Eqs. (9)–(13)) did not yield better results due to that the first two invariant functions might already contain enough information and the other five would be redundant for the current study.

As mentioned above, the advantage of introducing the PseAA components is that they can reflect some important features of a protein sequence through a discrete model (Chou and Shen, 2008). Thus, according to the Chou's PseAA composition (Chou, 2001), a protein sequence can be expressed by a vector or a point in a 20+2 = 22-D space; i.e.,

$$\mathbf{P} = [p_1 \quad p_2 \quad \cdots \quad p_{20} \quad p_{21} \quad p_{22}]^{\mathrm{T}} \quad (14)$$

where T is the transpose operator, and

$$p_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{20} f_i + \sum_{j=1}^{2} w_j\varphi_j}, & (1 \leqslant k \leqslant 20) \\ \dfrac{w_{(k-20)}\varphi_{(k-20)}}{\sum_{i=1}^{20} f_i + \sum_{j=1}^{2} w_j\varphi_j}, & (21 \leqslant k \leqslant 22) \end{cases} \quad (15)$$

where f_i ($i = 1, 2, \ldots, 20$) are the occurrence frequencies of the 20 native amino acids in the protein, arranged alphabetically according to their single letter codes, φ_j ($j = 1, 2$) are derived from geometric moments of the cellular automaton image of protein **P** as given by Eqs. (7) and (8), and the weight factors $w_j = 0.05$ ($j = 1, 2$) (Chou, 2001).

Now the augmented covariant-discriminant algorithm (Chou, 2000; Chou and Elrod, 1999) or CD classifier (Chou and Shen, 2007) was adopted to perform the prediction. For reader's convenience, a brief introduction about the CD classifier is given in Appendix.

3. Results and discussion

As a demonstration, let us use the benchmark dataset constructed in (Chou, 1999). It consists of 204 proteins, of which 52 are all-α, 61 all-β, 45α/β, and 46α+β. Their PDB codes are given in Table 2 of Chou (1999). In statistical prediction, the independent dataset test, sub-sampling test and jackknife test are the three cross-validation methods often used for examining the accuracy of a predictor (Chou and Zhang, 1995). However, as analyzed in a recent comprehensive review (Chou and Shen, 2007), the independent dataset test and sub-sampling test cannot avoid arbitrariness. Accordingly, the jackknife test has been increasingly and widely adopted by investigators (see, e.g., Chen et al., 2006a, b, 2008c; Chou and Shen, 2008; Ding et al., 2007; Du and Li, 2006; Guo et al., 2006; Jiang et al., 2008; Jin et al., 2008; Kedarisetti et al., 2006; Li and Li, 2008; Lin and Li, 2007a, b; Lin et al., 2008; Mondal et al., 2006; Niu et al., 2006, 2008; Pugalenthi et al., 2007; Shi et al., 2007; Sun and Huang, 2006; Tan et al., 2007; Wen et al., 2007; Xiao and Chou, 2007; Zhang et al., 2006; Zhou, 1998; Zhou and Doctor, 2003) to test the power of various predictors. Therefore, in this study, we also used the jackknife test to examine the performance of the new prediction approach.

The success rates by the 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 other methods on the same benchmark dataset are also listed. It can be seen from Table 2 that the overall success rate by the current approach is 92.6%, indicating that the new approach can remarkably enhance the success rate for the current benchmark dataset, or at least can play a complementary role to

Table 2
Success rates of jackknife cross-validation with different approaches on the 204 proteins from (Chou, 1999)

Method	Input	All-α	All-β	α/β	α+β	Overall
Unsupervised fuzzy clustering (Zhang et al., 1995)	AA composition	$\frac{35}{52}$ = 67.3%	$\frac{55}{61}$ = 90.2%	$\frac{21}{45}$ = 46.7%	$\frac{28}{46}$ = 60.9%	$\frac{139}{204}$ = 68.1%
Supervised fuzzy clustering (Shen et al., 2005)	AA composition	$\frac{38}{52}$ = 73.1%	$\frac{55}{61}$ = 90.2%	$\frac{28}{45}$ = 62.2%	$\frac{29}{46}$ = 63.1%	$\frac{150}{204}$ = 73.5%
Covariant matrix algorithm (Chou and Zhang, 1994)	Correlation analysis approach (Du et al., 2003)	$\frac{49}{52}$ = 94.2%	$\frac{53}{61}$ = 86.9%	$\frac{22}{45}$ = 48.9%	$\frac{41}{46}$ = 89.1%	$\frac{165}{204}$ = 80.9%
Augmented covariant-discriminant algorithm (Chou, 2000)	PseAA composition formulated in this paper[a]	$\frac{45}{52}$ = 86.5%	$\frac{56}{61}$ = 91.8%	$\frac{45}{45}$ = 100%	$\frac{43}{46}$ = 93.5%	$\frac{189}{204}$ = 92.6%

[a] Using the geometric invariant moment factors of protein cellular automaton image for the 21st and 22nd components of PseAA composition.

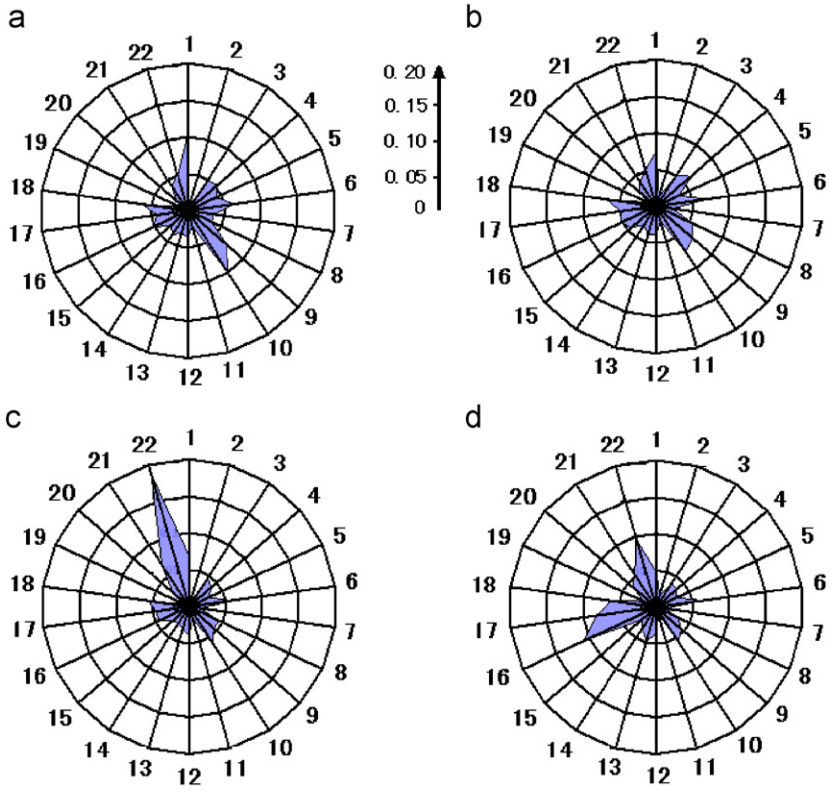

Fig. 2. Radar diagrams to show the difference of the 22-D standard vectors for: (a) all-α, (b) all-β, (c) α/β, and (d) $\alpha+\beta$ structural classes. Here we use the numerical indexes 1, 2, 3, …, 20 to represent the classical 20 amino acid components according to the alphabetical order of the single character codes of amino acids, and use the indices 21 and 22 to represent the PseAA components introduced through the geometric invariant moment factors as defined by Eqs. (7) and (8), respectively.

the existing method in predicting protein structural classification. The enhancement of success rates is particularly obvious for the case of α/β and $\alpha+\beta$ proteins.

Why could the overall success rate be improved so remarkably by introducing the geometric moment factors of protein cellular automaton image? To address this problem, let us consider the standard vectors for the four structural classes, $\bar{\mathbf{P}}^{\alpha}$, $\bar{\mathbf{P}}^{\beta}$, $\bar{\mathbf{P}}^{\alpha/\beta}$, and $\bar{\mathbf{P}}^{\alpha+\beta}$, as defined in Eq. (A3) in the Appendix. Each of the four standard vectors in the current approach contains 22 components. To provide an intuitive picture, each of the four 22-D standard vector is projected onto a 2-D radar diagram (Chou, 1993) as shown in Fig. 2, from which we can see that, by introducing the geometric invariant moment factor into the representation for protein samples, the standard vectors for the four structural classes have become remarkably distinct from each other. In contrast to that, the 20-D standard proteins for the same dataset are given in Fig. 1 of Du et al. (2003), from which we can see that the difference between $\bar{\mathbf{P}}^{\alpha/\beta}$ and $\bar{\mathbf{P}}^{\alpha+\beta}$ is trivial, meaning that the geometric moments as introduced in this paper are important for distinctly characterizing the structural class of proteins.

4. Conclusions

Using the geometric moments of protein cellular automaton images as the PseAA components (Chou, 2001) can more effectively reflect the overall protein sequence patterns, yielding a higher overall success rate in predicting protein structural classification. These kinds of subtle overall patterns are hidden in a pile of long and complicated sequences and are very difficult to extract without resorting to the cellular automaton approach (Wolfram, 1984). The procedures proposed in this study can be briefly summarized as follows. (1) For each of the protein sequences concerned, generate a 2-D image function according to the cellular automaton evolving rule. (2) Based on the image function, derive the geometric invariant moments. (3) Use the moments thus obtained as the PseAA components to formulate the protein sample. (4) The augmented covariant-discriminant algorithm or CD classifier was utilized as operation engine to perform the prediction. (5) The jackknife cross-validation test was adopted to examine the prediction quality.

Here, the protein structural class is just a paradigm for demonstration. It is instructive to point out that the current novel approach can also be used to predict a series of other protein attributes, such as subcellular localization, enzyme functional class, membrane protein type, proteinase type, and GPCR type, among many others.

Acknowledgments

The authors wish to express their gratitude to the two anonymous reviewers whose constructive comments were very helpful for strengthening the presentation of this paper. The work in this research was supported by the grants from the National Natural Science Foundation of China (no. 60661003), the Province National Natural Science Foundation of Jiangxi (no. 0611060), and the plan for training youth scientists (stars of Jing-Gang) of province Jiangxi.

Appendix

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

$$\mathbb{S} = S_1 \cup S_2 \cup S_3 \cup \cdots \cup S_M \tag{A.1}$$

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 $\mathbf{P}$ (Eq. (1)), how can we identify which subset it belongs to? According to the CD classifier (Chou and Shen, 2007), we can suppose without losing generality that the uth protein in the subset S_m of Eq. (A.1) is formulated by (see Eq. (14)):

$$\mathbf{P}_m^u = \left[\begin{array}{cccccc} p_{m,1}^u & p_{m,2}^u & \cdots & p_{m,20}^u & p_{m,21}^u & p_{m,22}^u \end{array}\right]^{\mathrm{T}} \tag{A.2}$$

where $p_{m,j}^u (j = 1, 2, \ldots, 22)$ is the jth component of the uth protein in S_m, and the standard vector for the subset S_m is defined by:

$$\bar{\mathbf{P}}_m = \left[\begin{array}{cccccc} \bar{p}_{m,1} & \bar{p}_{m,2} & \cdots & \bar{p}_{m,20} & \bar{p}_{m,21} & \bar{p}_{m,22} \end{array}\right]^{\mathrm{T}} \tag{A3}$$

where

$$\bar{p}_{m,i} = \frac{1}{N_m}\sum_{u=1}^{N_m} p_{m,i}^u, \quad (i = 1, 2, \ldots, 22) \tag{A.4}$$

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

$$\mathbb{F}(\mathbf{P}, \bar{\mathbf{P}}_m) = D_{Mah}^2(\mathbf{P}, \bar{\mathbf{P}}_m) + \ln|\mathbf{C}_m| \quad (m = 1, 2, \ldots, M) \tag{A.5}$$

where

$$D_{Mah}^2(\mathbf{P}, \bar{\mathbf{P}}_m) = (\mathbf{P} - \bar{\mathbf{P}}_m)^{\mathrm{T}}\mathbf{C}_m^{-1}(\mathbf{P} - \bar{\mathbf{P}}) \tag{A.6}$$

is the squared Mahalanobis distance (Chou and Zhang, 1994; Mahalanobis, 1936; Pillai, 1985) between $\mathbf{P}$ and $\bar{\mathbf{P}}_m$;

$$\mathbf{C}_m = \left[\begin{array}{cccc} c_{1,1}^m & c_{1,2}^m & \cdots & c_{1,22}^m \\ c_{2,1}^m & c_{2,2}^m & \cdots & c_{2,22}^m \\ \vdots & \vdots & \ddots & \vdots \\ c_{22,1}^m & c_{22,2}^m & \cdots & c_{22,22}^m \end{array}\right] \tag{A.7}$$

is the covariance matrix for the subset S_m; the 22×22 elements in $\mathbf{C}_m$ are given by:

$$c_{i,j}^m = \frac{1}{N_m - 1}\sum_{u=1}^{N_m}(p_{m,i}^u - \bar{p}_{m,i})(p_{m,j}^u - \bar{p}_{m,j}) \quad (i, j = 1, 2, \ldots, 22) \tag{A.8}$$

and $|\mathbf{C}_m|$ is the determinant of the matrix $\mathbf{C}_m$ that is always positive as proved in Appendix B of Chou and Elrod (1999). The smaller the value of $\mathbb{F}(\mathbf{X}, \bar{\mathbf{X}}_{mm})$, the higher the similarity between $\mathbf{P}$ and $\bar{\mathbf{P}}_m$. Therefore, the query protein is predicted to belong to the subset S_μ or the μth type if:

$$\mu = \mathbf{arg\ min}_m\{\mathbb{F}(\mathbf{P}, \bar{\mathbf{P}}_m)\}, \quad (m = 1, 2, \ldots, M) \tag{A.9}$$

where μ is the argument of m that minimizes $\mathbb{F}(\mathbf{P}, \bar{\mathbf{P}}_m)$. If there are two or more arguments leading to a same minimum value for $\mathbb{F}(\mathbf{P}, \bar{\mathbf{P}}_m)$, the query protein will be randomly assigned to one of the structural classes associated with these arguments although this kind of tie case rarely happens. Note that owing to the normalization condition imposed by Eq. (15), of the 22 components in Eq. (A.2), only 21 are independent, and hence the covariance matrix $\mathbf{C}_m$ as defined by Eq. (A.8) must be a singular one (Chou and Zhang, 1994). This would lead the Mahalanobis distance defined by Eq. (A.6) and the covariant-discriminant function given by Eq. (A.5) to be divergent and meaningless. To cope with such a situation, the dimension-reducing procedure (Chou, 1995) was adopted in practical calculations; i.e., instead of 22-D space, a protein sample is defined in a (22-1)-D space by leaving out one of its 22 components. The remaining (22-1) components would be completely independent, thereby the corresponding covariance matrix $\mathbf{C}_m$ being no longer singular. In other words, the Mahalanobis distance (Eq. (A.6)) and the covariant-discriminant function (Eq. (A.5)) based on such a (22-1)-D space can be uniquely defined without any trouble. However, a question might be raised: which one of the 22 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 Chou, 1995), 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 22 components is left out. Accordingly, the final value of the covariant-discriminant function (Eq. (A.5)) can be uniquely defined through such a dimension-reducing procedure. For more details about the CD classifier, the reader is referred to the papers (Chou, 2005a; Chou and Elrod, 1999; Chou and Shen, 2007).

References

Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C., Murzin, A.G., 2004. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res. 32, D226–D229.

Cao, Y., Liu, S., Zhang, L., Qin, J., Wang, J., Tang, K., 2006. Prediction of protein structural class with rough sets. BMC Bioinformatics 7, 20.

Chandonia, J.M., Karplus, M., 1995. Neural networks for secondary structure and structural class prediction. Protein Sci. 4, 275–285.

Chen, Y.L., Li, Q.Z., 2007a. Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. J. Theor. Biol. 248, 377–381.

Chen, Y.L., Li, Q.Z., 2007b. Prediction of the subcellular location of apoptosis proteins. J. Theor. Biol. 245, 775–783.

Chen, C., Zhou, X., Tian, Y., Zou, X., Cai, P., 2006a. Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal. Biochem. 357, 116–121.

Chen, C., Tian, Y.X., Zou, X.Y., Cai, P.X., Mo, J.Y., 2006b. Using pseudo-amino acid composition and support vector machine to predict protein structural class. J. Theor. Biol. 243, 444–448.

Chen, K., Kurgan, L.A., Ruan, J., 2008a. Prediction of protein structural class using novel evolutionary collocation-based sequence representation. J. Comput. Chem. 29, 1596–1604.

Chen, K., Kurgan, M., Kurgan, L., 2008b. Sequence based prediction of relative solvent accessibility using two-stage support vector regression with confidence values. J. Biomed. Sci. Eng. 1, 1–9.

Chen, W.J., Huang, P.T., Cheng, Y.C., Liao, T.H., 2008c. Putative secondary active site of bovine pancreatic deoxyribonuclease I. Protein Pept. Lett. 15, 640–646.

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

Chou, K.C., 1993. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. J. Biol. Chem. 268, 16938–16948.

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

Chou, K.C., 1999. A key driving force in determination of protein structural classes. Biochem. Biophys. Res. Commun. 264, 216–224.

Chou, K.C., 2000. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem. Biophys. Res. Commun. 278, 477–483.

Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: Struct., Funct., Genet. (Erratum: ibid., 2001, vol. 44, 60) 43, 246–255.

Chou, K.C., 2005a. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21, 10–19.

Chou, K.C., 2005b. Review: progress in protein structural class prediction and its impact to bioinformatics and proteomics. Curr. Protein Pept. Sci. 6, 423–436.

Chou, K.C., Cai, Y.D., 2004. Predicting protein structural class by functional domain composition. Biochem. Biophys. Res. Commun. (Corrigendum: ibid., 2005, vol. 329, 1362) 321, 1007–1009.

Chou, K.C., Elrod, D.W., 1999. Protein subcellular location prediction. Protein Eng. 12, 107–118.

Chou, K.C., Maggiora, G.M., 1998. Domain structural class prediction. Protein Eng. 11, 523–538.

Chou, K.C., Shen, H.B., 2007. Review: recent progresses in protein subcellular location prediction. Anal. Biochem. 370, 1–16.

Chou, K.C., Shen, H.B., 2008. Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms. Nat. Protoc. 3, 153–162.

Chou, K.C., Zhang, C.T., 1994. Predicting protein folding types by distance functions that make allowances for amino acid interactions. J. Biol. Chem. 269, 22014–22020.

Chou, K.C., Zhang, C.T., 1995. Review: prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349.
Deleage, G., Roux, B., 1987. An algorithm for protein secondary structure prediction based on class prediction. Protein Eng. 1, 289–294.
Ding, Y.S., Zhang, T.L., Chou, K.C., 2007. Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein Pept. Lett. 14, 811–815.
Du, P., Li, Y., 2006. Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC Bioinformatics 7, 518.
Du, Q.S., Wei, D.Q., Chou, K.C., 2003. Correlation of amino acids in proteins. Peptides 24, 1863–1869.
Fang, Y., Guo, Y., Feng, Y., Li, M., 2008. Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. Amino Acids 34, 103–109.
Gonzalez-Diaz, H., Vilar, S., Santana, L., Uriarte, E., 2007. Medicinal chemistry and bioinformatics-current trends in drugs discovery with networks topological indices. Curr. Top. Med. Chem. 10, 1015–1029.
González-Díaz, H., González-Díaz, Y., Santana, L., Ubeira, F.M., Uriarte, E., 2008. Proteomics, networks, and connectivity indices. Proteomics 8, 750–778.
Guo, Y.Z., Li, M., Lu, M., Wen, Z., Wang, K., Li, G., Wu, J., 2006. Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform. Amino Acids 30, 397–402.
Hu, M.K., 1962. Visual pattern recognition by moments invariants. IRE Trans. Inf. Theory 8, 179–187.
Jahandideh, S., Abdolmaleki, P., Jahandideh, M., Asadabadi, E.B., 2007. Novel two-stage hybrid neural discriminant model for predicting proteins structural classes. Biophys. Chem. 128, 87–93.
Jiang, X., Wei, R., Zhang, T.L., Gu, Q., 2008. Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. Protein Pept. Lett. 15, 392–396.
Jin, Y., Niu, B., Feng, K.Y., Lu, W.C., Cai, Y.D., Li, G.Z., 2008. Predicting subcellular localization with AdaBoost learner. Protein Pept. Lett. 15, 286–289.
Kedarisetti, K.D., Kurgan, L.A., Dick, S., 2006. Classifier ensembles for protein structural class prediction with varying homology. Biochem. Biophys. Res. Commun. 348, 981–988.
Klein, P., 1986. Prediction of protein structural class by discriminant analysis. Biochim. Biophys. Acta 874, 205–215.
Klein, P., Delisi, C., 1986. Prediction of protein structural class from amino acid sequence. Biopolymers 25, 1659–1672.
Kneller, D.G., Cohen, F.E., Langridge, R., 1990. Improvements in protein secondary structure prediction by an enhanced neural network. J. Mol. Biol. 214, 171–182.
Kurgan, L., Homaeian, L., 2006. Prediction of structural classes for protein sequences and domains—impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. Pattern Recognit. Lett. 39, 2323–2343.
Kurgan, L.A., Stach, W., Ruan, J., 2007. Novel scales based on hydrophobicity indices for secondary protein structure. J. Theor. Biol. 248, 354–366.
Kurgan, L.A., Zhang, T., Zhang, H., Shen, S., Ruan, J., 2008. Secondary structure-based assignment of the protein structural classes. Amino Acids. doi:18427716.
Levitt, M., Chothia, C., 1976. Structural patterns in globular proteins. Nature 261, 552–557.
Li, F.M., Li, Q.Z., 2008. Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. Protein Pept. Lett. 15, 612–616.
Lin, H., 2008. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J. Theor. Biol. 252, 350–356.
Lin, H., Li, Q.Z., 2007a. Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem. Biophys. Res. Commun. 354, 548–551.
Lin, H., Li, Q.Z., 2007b. Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. J. Comput. Chem. 28, 1463–1466.
Lin, H., Ding, H., Feng-Biao Guo, F.B., Zhang, A.Y., Huang, J., 2008. Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein Pept. Lett. 15, 739–744.
Liu, W., Chou, K.C., 1998. Prediction of protein structural classes by modified Mahalanobis discriminant algorithm. J. Protein Chem. 17, 209–217.
Luo, R.Y., Feng, Z.P., Liu, J.K., 2002. Prediction of protein structural class by amino acid and polypeptide composition. Eur. J. Biochem. 269, 4219–4225.
Mahalanobis, P.C., 1936. On the generalized distance in statistics. Proc. Natl. Inst. Sci. India 2, 49–55.
Mao, B., Chou, K.C., Zhang, C.T., 1994. Protein folding classes: a geometric interpretation of the amino acid composition of globular proteins. Protein Eng. 7, 319–330.
Metfessel, B.A., Saurugger, P.N., Connelly, D.P., Rich, S.T., 1993. Cross-validation of protein structural class prediction using statistical clustering and neural networks. Protein Sci. 2, 1171–1182.
Mondal, S., Bhavna, R., Mohan Babu, R., Ramakumar, S., 2006. Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. J. Theor. Biol. 243, 252–260.
Mundra, P., Kumar, M., Kumar, K.K., Jayaraman, V.K., Kulkarni, B.D., 2007. Using pseudo amino acid composition to predict protein subnuclear localization: approached with PSSM. Pattern Recognit. Lett. 28, 1610–1615.
Nakashima, H., Nishikawa, K., Ooi, T., 1986. The folding type of a protein is relevant to the amino acid composition. J. Biochem. 99, 152–162.
Nanni, L., Lumini, A., 2008. Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. Amino Acids 34, 653–660.
Niu, B., Cai, Y.D., Lu, W.C., Zheng, G.Y., Chou, K.C., 2006. Predicting protein structural class with AdaBoost learner. Protein Pept. Lett. 13, 489–492.
Niu, B., Jin, Y.H., Feng, K.Y., Liu, L., Lu, W.C., Cai, Y.D., Li, G.Z., 2008. Predicting membrane protein types with bagging learner. Protein Pept. Lett. 15, 590–594.
Pillai, K.C.S., 1985. Mahalanobis D2. In: Kotz, S., Johnson, N.L. (Eds.), Encyclopedia of Statistical Sciences, vol. 5. John Wiley & Sons. This reference also presents a brief biography of Mahalanobis who was a man of great originality and who made considerable contributions to statistics, New York, pp. 176–181.
Pu, X., Guo, J., Leung, H., Lin, Y., 2007. Prediction of membrane protein types from sequences and position-specific scoring matrices. J. Theor. Biol. 247, 259–265.
Pugalenthi, G., Tang, K., Suganthan, P.N., Archunan, G., Sowdhamini, R., 2007. A machine learning approach for the identification of odorant binding proteins from sequence-derived properties. BMC Bioinformatics 8, 351.
Rizon, M., Yazid, H., Saad, P., 2006. Object detection using geometric invariant moment. Am. J. Appl. Sci. 2, 1876–1878.
Shen, H.B., Chou, K.C., 2008. PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. Anal. Biochem. 373, 386–388.
Shen, H.B., Yang, J., Liu, X.J., Chou, K.C., 2005. Using supervised fuzzy clustering to predict protein structural classes. Biochem. Biophys. Res. Commun. 334, 577–581.
Shi, J.Y., Zhang, S.W., Pan, Q., Cheng, Y.-M., Xie, J., 2007. Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. Amino Acids 33, 69–74.
Sun, X.D., Huang, R.B., 2006. Prediction of protein structural classes using support vector machines. Amino Acids 30, 469–475.
Tan, F., Feng, X., Fang, Z., Li, M., Guo, Y., Jiang, L., 2007. Prediction of mitochondrial proteins based on genetic algorithm—partial least squares and support vector machine. Amino Acids 33, 669–675.
Wen, Z., Li, M., Li, Y., Guo, Y., Wang, K., 2007. Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. Amino Acids 32, 277–283.
Wolfram, S., 1984. Cellular automation as models of complexity. Nature 311, 419–424.
Wolfram, S., 2002. A New Kind of Science. Wolfram Media Inc., Champaign, IL.
Xiao, X., Chou, K.C., 2007. Digital coding of amino acids based on hydrophobic index. Protein Pept. Lett. 14, 871–875.
Xiao, X., Shao, S., Ding, Y., Huang, Z., Chen, X., Chou, K.C., 2005. Using cellular automata to generate Image representation for biological sequences. Amino Acids 28, 29–35.
Zhang, T.L., Ding, Y.S., 2007. Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. Amino Acids 33, 623–629.
Zhang, C.T., Chou, K.C., Maggiora, G.M., 1995. Predicting protein structural classes from amino acid composition: application of fuzzy clustering. Protein Eng. 8, 425–435.
Zhang, S.W., Pan, Q., Zhang, H.C., Shao, Z.C., Shi, J.Y., 2006. Prediction protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and naive Bayes feature fusion. Amino Acids 30, 461–468.
Zhang, S.W., Zhang, Y.L., Yang, H.F., Zhao, C.H., Pan, Q., 2008. Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies. Amino Acids 34, 565–572.
Zhou, G.P., 1998. An intriguing controversy over protein structural class prediction. J. Protein Chem. 17, 729–738.
Zhou, G.P., Assa-Munt, N., 2001. Some insights into protein structural class prediction. Proteins: Struct., Funct., Genet. 44, 57–59.
Zhou, G.P., Doctor, K., 2003. Subcellular location prediction of apoptosis proteins. Proteins: Struct., Funct., Genet. 50, 44–48.
Zhou, X.B., Chen, C., Li, Z.C., Zou, X.Y., 2007. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J. Theor. Biol. 248, 546–551.