Amino Acids (2005) 28: 373–376
DOI 10.1007/s00726-005-0206-9

Amino Acids
Printed in Austria

# Using pseudo amino acid composition to predict protein subcellular location: Approached with Lyapunov index, Bessel function, and Chebyshev filter

**Y. Gao**[1], **S. Shao**[1], **X. Xiao**[1,2], **Y. Ding**[1], **Y. Huang**[1], **Z. Huang**[1], and **K.-C. Chou**[1,3,4,5]

[1] Bioinformation Research Centre, Donghua University, Shanghai, China
[2] Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, China
[3] Shanghai Jiaotong University, Shanghai, China
[4] Tianjin Institute of Bioinformatics & Drug Discovery, Tianjin, China
[5] Gordon Life Science Institute, Torrey Del Mar, San Diego, California, U.S.A.

Received February 23, 2005
Accepted March 30, 2005
Published online May 17, 2005 

**Summary.** With the avalanche of new protein sequences we are facing in the post-genomic era, it is vitally important to develop an automated method for fast and accurately determining the subcellular location of uncharacterized proteins. In this article, based on the concept of pseudo amino acid composition (Chou, K.C. Proteins: Structure, Function, and Genetics, 2001, 43: 246–255), three pseudo amino acid components are introduced via Lyapunov index, Bessel function, Chebyshev filter that can be more efficiently used to deal with the chaos and complexity in protein sequences, leading to a higher success rate in predicting protein subcellular location.

**Keywords:** Covariant-discriminant algorithm – Pseudo amino acid composition – Chaos – Lyapunov index – Bessel function – Chebyshev filter

## I Introduction

With the rapid increase in the number of new protein sequences entering into data banks, we are confronted with a critical challenge: How to timely (Chou, 2004) use them to stimulate the development of life science and benefit human beings? The knowledge of subcellular location of a protein is very important because it is closely related to its biological function. The traditional way to determine the localization of a protein in a cell is by biochemical experiments, and is both time-consuming and expensive. During the last decade, many theoretical methods were developed in an attempt to predict the subcellular location of a query protein according to its sequence information. The earlier approaches in this regard were based on the amino acid composition (see, e.g., Cedano et al., 1997; Chou, 2000b; Chou and Elrod, 1999; Nakashima and Nishikawa, 1994). However, if the prediction was based the amino acid composition, all the sequence order and sequence length effects of a protein would be lost. To improve the situation, Chou introduced the pseudo amino acid composition to partially incorporate the sequence order effect of a protein (Chou, 2001). The introduction of the concept of pseudo amino acid composition has greatly stimulated the development of this area (see, e.g., Chou and Cai, 2002, 2003a, 2003b; Pan et al., 2003; Wang et al., 2004, 2005; Xiao et al., 2005). In this paper, a different approach to define the pseudo amino acid components is used for predicting protein subcellular location.

## II Method

In addition to the 20 amino acid components as defined in the classical amino acid composition (Chou and Zhang, 1993, 1994; Nakashima et al., 1986), the Lyapunov index, Bessel function, and Chebyshev filter are used to define the additional three components for the pseudo amino acid composition of a protein, as illustrated below.

### *1 Lyapunov index*

Over the past 20 years, it has been a very important problem to measure chaos and noises in many different fields. Lyapunov index is a useful feature to describe a chaos system. Because protein sequences are extremely complicated and disorder, it would be of help to use Lyapunov index to analyze them. Moreover, the Wolf's re-construction method was

adopted here. First of all, the phase space was reconstructed as follows. Given a series of amino acid sequence: $x_1, x_2, \ldots, x_{N-1}$, the reconstructed embed vectors are $X_1, X_2, \ldots, X_m$, where

$$X_m = (x_m, x_{m+D}, \ldots, x_{m+(M-1)*D})^{\mathrm{T}} \tag{1}$$

where $M$ is the number of the re-constructed dimensions, $D$ the reconstructed-delay which is the number of sampling point between any 2 elements of the embedded vector. The Euclidean distance between two random point $X_m$ and $X_n$ in the re-constructed phase space is:

$$L = \sqrt{\begin{array}{l}(x_m - x_n)^2 + (x_{m+D} - x_{n+D})^2 + (x_{m+2D} - x_{n+2D})^2 + \cdots\cdots \\ \quad + (x_{m+(M-1)*D} - x_{n+(M-1)*D})^2\end{array}} \tag{2}$$

Thus, the following equation is used to calculate the Lyapunov index:

$$p_1 = \lambda = \frac{1}{T_Q - T_0} \sum_{k=1}^{Q} \log_2 \frac{L'(t_k)}{L(t_{k-1})} \tag{3}$$

where $Q$ is the number of the points which can be found, $T_Q - T_0 = Q * I * H$ in which $I$ is the number of iterations, $H$ is the time of iteration. The value of $p_1$ is the 1st additional component, i.e., the 21st component in the pseudo amino acid composition.

### *2 Besseli function*

First, we use the numeric amino acid sequence of a protein as the input of the Besseli function, the output of the Besseli function is the vector with the same dimension number as the input vector: $y = \beta(1, x)$ where $\beta$ is the improved Besseli function, and $x$ amino acid sequence of a protein. Thus, it follows

$$p_2 = \text{corr. coef.}(x, y) = \frac{\mathrm{cov}(x, y)}{\mathrm{std}(x) \cdot \mathrm{std}(y)} \tag{4}$$

where std is standard deviation, and cov the covariance; their expressions are as follows:

$$\mathrm{std}(x) = \frac{1}{N} \sum_{i=1}^{N} |x(i) - M(x)| \tag{5}$$

$$\mathrm{cov}(x) = \frac{1}{N-1} \sum_{i=1}^{N} \big(x(i) - M(x)\big)\big(y(i) - M(y)\big) \tag{6}$$

The value of $p_2$ is the 2nd additional component, i.e., the 22nd component in the pseudo amino acid composition.

### *3 Chebyshev filter*

Likewise, if choosing the digital amino acid sequence $x$ of the protein as the input of the Chebyshev filter, we can get an output vector $z = \gamma(1, x)$, where $\gamma$ is the Chebyshev filter, and $x$ amino acid sequence of a protein. Thus, it follows

$$p_3 = \text{corr. coef.}(x, z) = \frac{\mathrm{cov}(x, z)}{\mathrm{std}(x) \cdot \mathrm{std}(z)} \tag{7}$$

The value of $p_3$ is the 3rd additional component, i.e., the 23rd component in the pseudo amino acid composition.

### *4 Numerical coding of protein sequence*

To make the current method work, each amino acid in a protein sequence must have a numerical value. Here, let us assign a numerical value for each of the 20 native amino acids; i.e., A = 5, C = 10, D = 15, E = 20, F = 25, G = 30, H = 35, I = 40, K = 45, L = 60, M = 65, N = 70, P = 75, Q = 80, R = 85, S = 90, T = 95, V = 100, W = 105, and Y = 110.

As an illustration, let us consider the following two sequences:

ADFGHIK (8)

and

AFDGIHK (9)

According to the numerical codes, we can transfer above two sequences into discrete stochastic time sequences, which can be converted into a curve as shown in Figs. 1 and 2, respectively.

### *5 Pseudo amino acid prediction*

After we get those three parameters stated above, we can predict protein's subcellular location by the pseudo amino acid prediction approach. According to Chou (2001), a protein can now be represented by a vector in a 23-D (dimensional) space; i.e.,

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{20} \\ x_{21} \\ x_{22} \\ x_{23} \end{bmatrix} \tag{10}$$

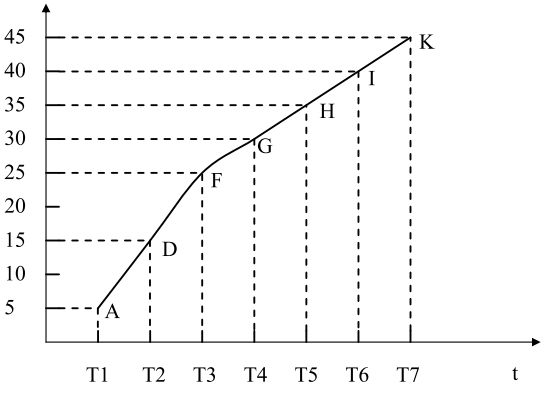


**Fig. 1.** The curve shape for the sequence of Eq. (8) according to the numerical codes assigned in this paper

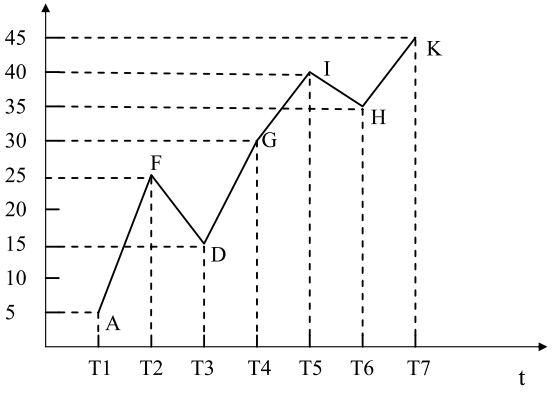


**Fig. 2.** The curve shape for the sequence of Eq. (9) according to the numerical codes assigned in this paper

where $x_1 \sim x_{20}$ are 20 components defined in the classical amino acid composition space (Chou, 1995; Chou and Zhang, 1994), $x_{21} \sim x_{23}$ are 3 additional components for the pseudo amino acid composition. After imposing the normalization condition, the 23 components are expressed as follows:

$$x_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{20} f_i + \sum_{j=1}^{3} w_j p_j}, & (1 \le k \le 20) \\ \dfrac{w_{k-20} \cdot p_{k-20}}{\sum_{i=1}^{20} f_i + \sum_{j=1}^{3} w_j p_j}, & (21 \le k \le 23) \end{cases} \tag{11}$$

where $f_k$ is the normalized occurrence frequency of the $k$th amino acid in the protein $X$, $p_1, p_2, p_3$ are the 3 parameters derived from Eqs. (3), (4), and (7), respectively, and $w_1, w_2, w_3$ are the corresponding weight factors. For the current study, the weight factors were generated thru an optimal process as given below

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} 0.1 \\ 0.005 \\ 0.1 \end{bmatrix} \tag{12}$$

Now the augmented covariant-discriminant algorithm (Chou, 2000a, 2001) was used to perform the prediction. For the details of the algorithm, the reader is referred to the previous papers (Chou, 1995; Chou et al., 1998; Chou and Zhang, 1994; Zhou, 1998; Zhou and Assa-Munt, 2001).

## III Results and discussion

To test the result of the current method, the dataset constructed by Chou and Elrod (1999) was used, which involves 12 subcellular locations (Fig. 3). However, for the reason as explained in Chou (2001), of the 2,319 proteins originally listed in Appendix A of Chou and Elrod (1999), only 2,191 protein sequences were retrieved that contains 145 chloroplast proteins, 571 cytoplasm, 34 cytoskeleton, 49 endoplasmic reticulum, 224 extracellular, 25 Golgi apparatus, 37 lysosome, 84 mitochondria, 272 nucleus proteins, 27 peroxisome, 699 plasma membrane, and 24 vacuole.

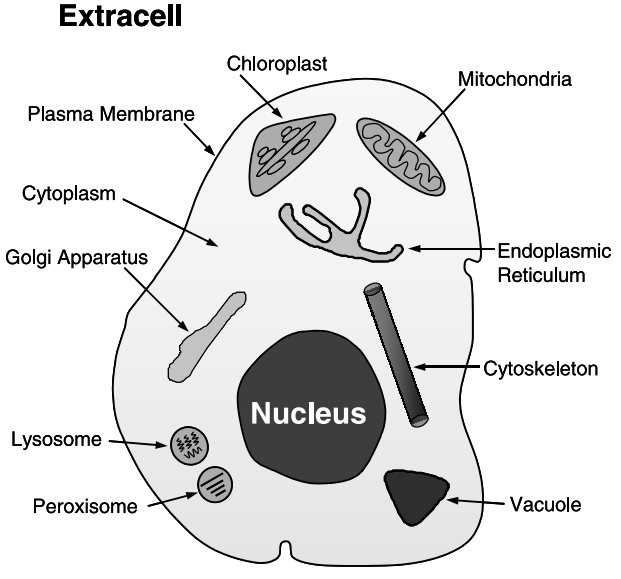


**Fig. 3.** Schematic illustration to show the twelve subcellular locations of proteins: chloroplast, cytoplasm, cytoskeleton, endoplasmic reticulum, extracell, Golgi apparatus, lysosome, mitochondria, nucleus, peroxisome, plasma membrane, and vacuole. Note that the vacuole and chloroplast proteins exist only in a plant. Reproduced from Fig. 2 of Chou (2001) with permission

**Table 1.** The overall success rates obtained by different methods[a]

| | Pan et al. (2003) | This paper |
|---|---|---|
| Re-substitution test | 81.5% | 82.3% |
| Jackknife test | 67.7% | 69.6% |

[a] The predictions were performed on the dataset taken from Chou and Elrod (1999)

The test was performed by re-substitution and jackknife. The former is to test the self-consistency (Chou and Zhang, 1995) of the prediction method, while the latter to test its extrapolating effectiveness (Chou and Zhang, 1993). Besides, among the jackknife test, independent dataset test, and sub-sampling test, the jackknife test is deemed the most objective and rigorous examination for cross-validation (Chou and Zhang, 1995; Zhou, 1998; Zhou and Assa-Munt, 2001; Zhou and Doctor, 2003). Therefore, a combination of the results obtained by the re-substitution and jackknifing can well indicate the power of a statistical prediction method. The predicted results thus obtained are given in Table 1, where for facilitating comparison, the corresponding results obtained by Pan et al. (2003) are also shown. As we can see from the table, the current success rates by both the re-substitution and jackknife tests are higher than those of Pan et al. (2003).

## IV Conclusions

It is a feasible and effective approach to introduce the pseudo amino acid composition (Chou, 2001) to incorporate, at least partially, the information of the protein sequence order and sequence length for improving the quality of predicting protein subcellular location. It is demonstrated thru the present study that it can better reflect the sequence order effect by using Lyapunov index, Bessel function, and Chebyshev filter to construct the 3 additional components for the pseudo amino acid composition.

## Acknowledgments

The work in this research was supported in part by the Doctoral Foundation from the National Education Committee (20030255009), China.

## References

Cedano J, Aloy P, P'erez-Pons JA, Querol E (1997) Relation between amino acid composition and cellular location of proteins. J Mol Biol 266: 594–600

Chou JJ, Zhang CT (1993) A joint prediction of the folding types of 1490 human proteins from their genetic codons. J Theor Biol 161: 251–262

Chou KC (1995) A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins: Structure, Function, and Genetics 21: 319–344

Chou KC (2000a) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochemical & Biophysical Research Communications 278: 477–483

Chou KC (2000b) Review: prediction of protein structural classes and subcellular locations. Current Protein and Peptide Science 1: 171–208

Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino-acid-composition. Proteins: Structure, Function, and Genetics (Erratum: ibid. (2001) 44: 60) 43: 246–255

Chou KC (2004) Review: structural bioinformatics and its impact to biomedical science. Curr Med Chem 11: 2105–2134

Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 277: 45765–45769

Chou KC, Cai YD (2003a) A new hybrid approach to predict subcellular localization of proteins by incorporating Gene ontology. Biochem Biophys Res Commun 311: 743–747

Chou KC, Cai YD (2003b) Prediction and classification of protein subcellular location: sequence-order effect and pseudo amino acid composition. J Cell Biochem 90: 1250–1260 (Addendum, ibid. (2004) 91/5: p 1085)

Chou KC, Elrod DW (1999) Protein subcellular location prediction. Protein Engineering 12: 107–118

Chou KC, Zhang CT (1993) A new approach to predicting protein folding types. J Protein Chem 12: 169–178

Chou KC, Zhang CT (1994) Predicting protein folding types by distance functions that make allowances for amino acid interactions. J Biol Chem 269: 22014–22020

Chou KC, Zhang CT (1995) Review: prediction of protein structural classes. Crit Rev Biochem Mol Biol 30: 275–349

Chou KC, Liu W, Maggiora GM, Zhang CT (1998) Prediction and classification of domain structural classes. Proteins: Structure, Function, and Genetics 31: 97–103

Nakashima H, Nishikawa K (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J Mol Biol 238: 54–61

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

Pan YX, Zhang ZZ, Guo ZM, Feng GY, Huang ZD, He L (2003) Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. J Protein Chem 22: 395–402

Wang M, Yang J, Liu GP, Xu ZJ, Chou KC (2004) Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition. Protein Eng Design Sel 17: 509–516

Wang M, Yang J, Xu ZJ, Chou KC (2005) SLLE for predicting membrane protein types. J Theor Biol 232: 7–15

Xiao X, Shao S, Ding Y, Huang Z, Huang Y, Chou KC (2005) Using complexity measure factor to predict protein subcellular location. Amino Acids 28: 57–61

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17: 729–738

Zhou GP, Assa-Munt N (2001) Some insights into protein structural class prediction. Proteins: Structure, Function, and Genetics 44: 57–59

Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins: Structure, Function, and Genetics 50: 44–48

**Authors' address:** Prof. Kuo-Chen Chou, Gordon Life Science Institute, 13784 Torrey Del Mar, San Diego, CA 92130, U.S.A.,

E-mail: kchou@san.rr.com or Prof. Shihuang Shao, Donghua University, Shanghai, China, E-mail: shshao@dhu.edu.cn