Amino Acids (2006) 30: 49–54
DOI 10.1007/s00726-005-0225-6

Amino Acids
Printed in Austria

Using cellular automata images and pseudo amino acid composition to predict protein subcellular location

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

[1] Bioinformatics Research Center, Donghua University, Shanghai, China
[2] Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, China
[3] Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, California, U.S.A.

Received May 26, 2005
Accepted June 4, 2005
Published online July 28, 2005;

Summary. The avalanche of newly found protein sequences in the post-genomic era has motivated and challenged us to develop an automated method that can rapidly and accurately predict the localization of an uncharacterized protein in cells because the knowledge thus obtained can greatly speed up the process in finding its biological functions. However, it is very difficult to establish such a desired predictor by acquiring the key statistical information buried in a pile of extremely complicated and highly variable sequences. In this paper, based on the concept of the pseudo amino acid composition (Chou, K. C. *PROTEINS: Structure, Function, and Genetics*, 2001, 43: 246–255), the approach of cellular automata image is introduced to cope with this problem. Many important features, which are originally hidden in the long amino acid sequences, can be clearly displayed through their cellular automata images. One of the remarkable merits by doing so is that many image recognition tools can be straightforwardly applied to the target aimed here. High success rates were observed through the self-consistency, jackknife, and independent dataset tests, respectively.

Keywords: Cellular automata images – Pseudo amino-acid composition – Protein subcellular location – Complexity – Covariant-discriminant algorithm

I Introduction

One of the fundamental goals in protein science is to identify the function of a newly found protein. To truly understand the biological function of a protein, it is crucially important to find its localization in a cell. The experimental determination of protein subcellular location is generally accomplished by the following three approaches: cell fractionation, electron microscopy, and fluorescence microscopy (Boland et al., 1998). They are both time-consuming and costly. Besides, these methods also bear some sort of subjectivity (Murphy et al., 2000) and uncertainty. Therefore, it is in high demand to develop an automated method that can rapidly and accurately predict the subcellular localization of a protein just based on the information of its sequence alone. Actually, many efforts have been made in this regard during the last decade. The methods developed in the earlier stage are summarized in some review papers [see, e.g., (Chou, 2000b, 2002; Nakai, 2000)]. New progresses in this area are reflected by a series of papers published in recent years (Chou, 2000a, 2001; Chou and Cai, 2003b, 2004b, c, 2005; Pan et al., 2003; Park and Kanehisa, 2003; Wang et al., 2004a, b; Xiao et al., 2005a; Zhou and Doctor, 2003).

To improve the quality of protein subcellular location, a key problem is how to optimally express the statistical samples for proteins. The following two modes are often used to express a protein: (1) the sequential mode, and (2) the discrete mode. The most straightforward sequential mode is to use the entire sequence of a protein to represent itself. However, because protein sequences are extremely complicated with much variation in both sequence order and length, it is hardly to establish a feasible predictor by using the sequential mode to represent protein samples, as elaborated by Chou (Chou and Cai, 2002). The simplest discrete mode is to use the amino acid composition of a protein to represent it (Chou and Zhang, 1993, 1994; Chou, 1995; Zhou, 1998). By doing so, however, all the information of order and length in an original sequence are totally lost. To cope with such a dilemma, Chou propose a new discrete mode, the so-called pseudo amino acid composition (Chou, 2001), to represent the protein samples. The pseudo amino acid composition consists of $20 + \lambda$

discrete numbers: the 1st 20 numbers represent none but the 20 components of amino acid composition; the numbers from $20+1$ to $20+\lambda$ represent λ factors or functions derived from a protein concerned that bear, at least partially, its sequence order and length information. The introduction of pseudo amino acid composition has greatly stimulated the development of protein subcellular location prediction (Cai and Chou, 2003, 2004a, b; Cai et al., 2002a; Chou and Cai, 2003b, 2004b; Gao et al., 2005; Pan et al., 2003) as well as some related areas (Cai et al., 2002b, 2003, 2005; Chou, 2005; Chou and Cai, 2003a; Wang et al., 2004a, b). The key in successfully using the pseudo amino acid composition is how to choose the functions to derive the λ factors that can optimally reflect the sequence order and length effects of a protein sequence.

In this study, a novel approach – the images of protein cellular automata – is introduced to derive the pseudo amino acid components. The bottom line is that the cellular automaton images can reveal many important features of protein, which are originally hidden in a long and complicated amino acid sequence (Xiao et al., 2005c).

First of all, let us give a brief introduction about automaton, whose plural is automata. An automaton is a rule-following device. A computer is a typical example in this regard because it operates by following some rules. One type of automaton that has received a lot of attention is cellular automata because they can be used to generate beautiful pictures as well as study very complicated objects such as artificial life and chaos. A cellular automaton is a dynamical system in which space, time, and the states are discrete. Each cell, defined by a point in a regular spatial lattice, can have any one of a finite number of states that are updated according to a local rule; i.e., the state of a cell at a given time depends only on the immediately preceding states of itself and its nearby neighbors. All cells on the lattice are synchronously updated so as to realize the development of the dynamic system in discrete time steps.

The images generated by cellular automata, or in short the CA images, have been applied to the investigation into the sequence character of the severe acute respiratory syndrome coronaviruses (SARS-CoVs), and a remarkable fingerprint for the SARS-CoVs has been revealed (Wang et al., 2005). CA images were also used to predict the effect on the replication ratio by HBV virus gene missense mutation (Xiao et al., 2005b). There are several parameters to evaluate the image feature, such as pseudo zernike moments (Haddadnia et al., 2002), maximum local entropy (Zhu et al., 1997), and complex wavelet coefficients (Portilla and Simoncelli, 2000). Of the known complexity measure approaches so far, the Ziv-Lempel complexity measure is the most adequate one in reflecting the repeat patterns occurring in the character sequence (Gusev et al., 2001), and hence was adopted in this study. The results of both self-consistency and jackknife tests indicate that the protein localization is considerably correlated with its gene CA image.

II Method

1. Cellular automata image

A protein sequence is generally constituted 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 vector particularly when the sequence is very long. To cope with this situation, we resort to the images derived from the amino acid sequence through the space-time evolution of cellular automata. As a first step, the 20 amino acids are coded in a binary mode 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 (Xiao et al., 2005c). Through the above encoding procedure, a protein sequence is transformed to a serial of digital signals. For example, the sequence "MASAAG..." is transformed to "1001111001010011100111001...".

The cellular automaton adopted here is a simple two-state, one-dimensional one, consisting of a line of cells, with the value of 0 or 1 (Wolfram, 2002). The rule is simply implemented as: the nearest cells around the one on which we focus decide its next state. We adopt the circulating boundary condition, with the iterative formula given below:

$$D(i,j) = F(D(i-1,j-1), D(i-1,j), D(i-1,j+1)), \quad (2 \le i \le n, 2 \le j \le 5N-1) \quad (1)$$

$$D(i,1) = F(D(i-1,5N), D(i-1,1), D(i-1,2)), \quad (2 \le i \le n) \quad (2)$$

$$D(i,5N) = F(D(i-1,5N-1), D(i-1,5N), D(i-1,1)), \quad (2 \le i \le n) \quad (3)$$

where F is the iterative rule, n the iterative time, and N the length of the amino acid sequence. Data derived by the process with the evolving rule

Table 1. Digital codes of 20 native amino acids

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

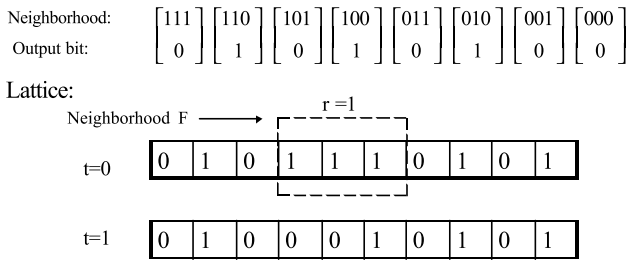

Fig. 1. Illustration to show a one-dimensional, binary-state, and nearest-neighbor (r = 1) cellular automata with $N = 10$. Both the lattice and the rule table F for updating the lattice are illustrated. The lattice configuration is shown at two successive time steps. The cellular automaton has spatially periodic boundary conditions: the lattice is viewed as a circle, with the leftmost cell being the right neighbor of the rightmost cell, and vice versa

are saved in the rows starting from the second, and data in each row are derived from those in its previous row.

The evolution rule for image formation must be able to very obviously distinguish whether the proteins concerned are similar to each other or not. We find the 84th is the best one in serving such a purpose among all the 256 kinds of evolving rules. Rule 84 may be described by Fig. 1. The time that the rule evolves determines the width of the images. It was found that the image structure is basically steady when the time is 300.

We transform the 2D array (matrix) into an image with visualization techniques. The basic bitmap format is chosen owing to its easily handled property. In this way, if the matrix element is zero, the color of the counterpart pixel bit is black; otherwise, white. For a systematic description of CA images, refer to the paper (Xiao et al., 2005a). Thus, all the existing tools in the area of image processing can be straightforwardly used for the current study. Figures 2 and 3 show proteins gene images based on CA 84th rule, and it can be seen from these figures that different type protein have completely different texture feature.

2. Predicting algorithm

Image recognition is concerned with the automatic detection and classification of image. Its techniques can be divided into two main categories: (1) those employing geometrical features, and (2) those using gray-level information. The texture characteristics of these gene images are very complicated and it is very difficult to characterize these images by using either deterministic or statistical models. Nevertheless, these protein images can be saved in 2D arrays. Thus, we can simply regard the Ziv-Lempel complexity (Ziv and Lempel, 1976) of these sequences as pseudo amino acid composition. The Ziv-Lempel complexity measure reflects most adequately repeats occurring in a sequence text (Ziv and Lempel, 1976).

The Ziv-Lempel complexity of a sequence can be measured by the minimal number of steps required for its synthesis in a certain process (Gusev et al., 2001). For each step only two operations were allowed in the process: either generating an additional symbol which ensures the uniqueness of each component $S[i_{k-1}+1:i_k]$ or copying the longest fragment from the part of a synthesized sequence. Suppose a string S expressed by

$$S = a_1 a_2 a_3 \cdots a_M \tag{4}$$

Its substring is expressed by

$$S[i:j] = a_i a_{i+1} a_{i+2} \cdots a_j \quad (1 \le i \le j \le M) \tag{5}$$

The complexity measure, $C_{LS}(S)$, of a non-empty sequence S synthesized according to the following procedure is defined by the minimal number of steps

$$H(S) = S[1:i_1]S[i_1+1:i_2]\cdots S[i_{k-1}+1:i_k]\cdots S[i_{m-1}+1:M] \tag{6}$$

At each step k the sequence is extended by concatenating a fragment $S[i_{k-1}+1:i_k]$. The length of this fragment is equal to 1 if some symbol at position $i_{k-1}+1$ occurs for the very first time. Otherwise, a component of length

$$i_k - i_{k-1} = \max_{j \le i_{k-1}} \{l_j : S[i_{k-1}+1:i_{k-1}+l_j] = S[j:j+l_j-1]\} \tag{7}$$

is copied from the proper prefix $S[1:i_{k-1}+l_j-1]$ where l_j is the length of the fragment being copied.

We can derive N complexity factors if the image has N rows. These complexity factors can all be used to serve as the pseudo amino acid components. However, it was observed that the highest jackknife success rates were resulted if the first 28 complexity factors were used. Accordingly, by following exactly the same procedure as described by Chou (Chou, 2001), a protein can be expressed by a vector or a point in a (20 + 28)D = 48D (dimensional) space; i.e.

$$\mathbf{X} = (x_1, x_2, x_3, \ldots, x_{48})^{\mathbf{T}} \tag{8}$$

$$x_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{20} f_i + \sum_{j=1}^{28} w_j p_j}, & (1 \le k \le 20) \\ \dfrac{w_{(k-20)} p_{(k-20)}}{\sum_{i=1}^{20} f_i + \sum_{j=1}^{28} w_j p_j}, & (21 \le k \le 48) \end{cases} \tag{9}$$

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, p_j $(j = 1, 2, \ldots, 28)$ are the complexity factors the protein sequence, w_j are the weight factor for the j^{th} complexity factor p_j, and $\mathbf{T}$ represents the transpose operator.

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 (Chou, 1995; Chou et al., 1998; Chou and Zhang, 1994; Zhou, 1998; Zhou and Assa-Munt, 2001). It is instructive to point out that owing to the normalization condition imposed by Eq. (9), the 48 components in Eq. (8) are not independent. Therefore, a dimension-reduced operation (Chou and Zhang, 1994) by leaving out one of the components and making the rest completely independent is needed when using the augmented covariant discriminant algorithm; i.e., a protein should be defined in a (48–1)D space instead of 48D space. Otherwise, a divergence difficulty will occur. However, which one of the 48 components should be removed? Anyone. The reason is that according to a theorem given and proved by Chou (Chou, 1995), which is generally quoted as "Chou's *Invariance Theorem*" (Pan et al., 2003; Zhou and

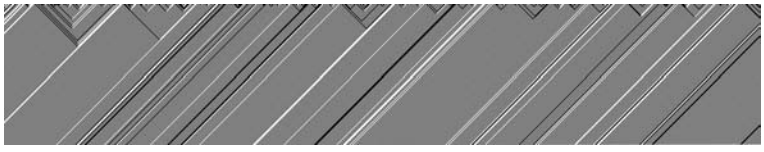

Fig. 2. Images of ACR1_YEAST, which is located in mitochondria, were generated by Cellular Automata 84th rule. The time of evolving was 300, the sequence was obtained from NCBI GenBank (P33303). The compression ratio was 2:2

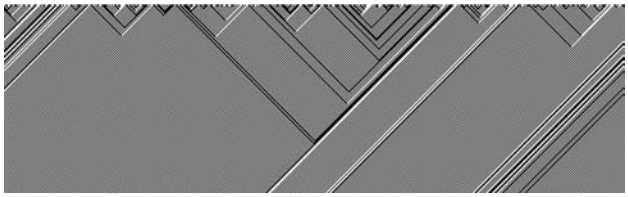

Fig. 3. Images of ABP1_ARATH, which is located in Endoplasmic reticulum, were generated by Cellular Automata 84[th] rule. The time of evolving was 300, the sequence was obtained from NCBI GenBank (P33487). The compression ratio was 2:2

Assa-Munt, 2001; Zhou and Doctor, 2003), the values of the covariant discriminant function will remain the same regardless of which one of the 48 components is left out.

III Results and discussion

The prediction quality was examined by the standard testing procedure in statistics, which is a combination of the self-consistency, jackknife, and independent dataset tests.

In the self-consistency test, the subcellular location of each protein in a given dataset was predicted by using the parameters derived from the same dataset, the so-called training dataset. The proteins studied here were taken from the dataset S^{12} of Chou (Chou, 2001) which is slightly different from the dataset originally constructed by Chou and Elrod (Chou and Elrod, 1999) for the reason given in (Chou, 2001). The training dataset S^{12} contains 2191 protein sequences, of which 145 are chloroplast, 571 cytoplasmic, 34 cytoskeletal, 49 endoplasmic reticulum, 224 extracellular, 25 Golgi apparatus, 37 lysosomal, 84 mitochondrial, 272 nuclear, 27 peroxisomal, 699 plasma membrane and 24 vacuoles (Fig. 4).

In the jackknife test, each of the proteins S^{12} is in turn singled out as a tested protein and all the rule-parameters

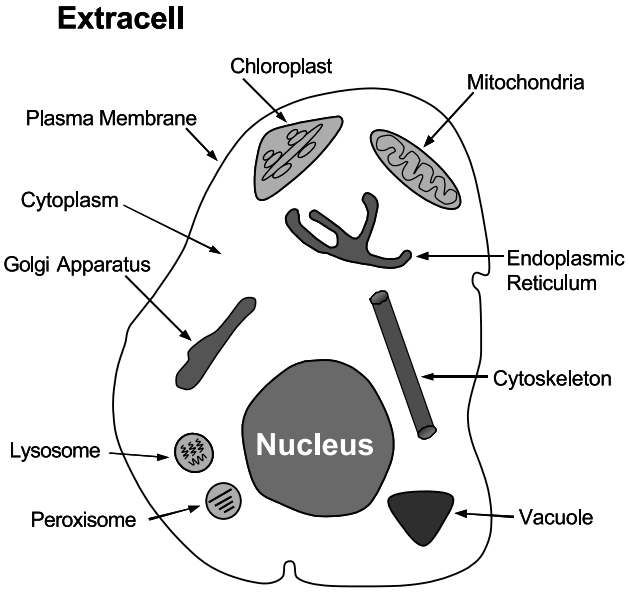

Fig. 4. Schematic illustration to show the twelve subcellular locations of proteins: chloroplast, cytoplasm, cytoskeleton, endoplasmic reticulum, extracell, Golgi apparatus, lysosome, mitochondria, nucleus, peroxisome, plasma membrane, and vacuole. Note that the vacuole and chloroplast proteins exist only in a plant. Reproduced from Fig. 2 of Chou (Chou, 2001) with permission

are calculated based on the remaining proteins without including the one being identified. Therefore, both the training data set and testing data set during the jackknifing process are actually open, and a sample will in turn move from one to the other.

In the independent dataset test, the prediction is made for a testing dataset containing only those proteins that do not occur in the training dataset S^{12}. In the current study, the independent dataset test, denoted by $\bar{S}^{12}$, was also taken from (Chou, 2001). It contains 2,494 protein sequences, of

Table 2. Overall success rates obtained with different methods by self-consistency, jackknife and independent dataset tests, respectively

Algorithm	Test method		
	Self-consistency[a]	Jackknife[a]	Independent dataset[b]
ProtLock (Cedano et al., 1997)	$\frac{1023}{2191} = 46.7\%$	$\frac{971}{2191} = 44.3\%$	$\frac{1018}{2494} = 40.8\%$
Digital signal (Pan et al., 2003)	$\frac{1785}{2191} = 81.5\%$	$\frac{1483}{2191} = 67.7\%$	$\frac{1842}{2494} = 73.9\%$
This paper[c]	$\frac{1893}{2191} = 86.4\%$	$\frac{1590}{2191} = 72.6\%$	$\frac{1865}{2494} = 74.8\%$

[a] Using the dataset S^{12} taken from Chou (Chou, 2001)
[b] Using the independent dataset $\bar{S}^{12}$ taken from Chou (Chou, 2001)
[c] Using the augmented covariant-discriminant algorithm (Chou, 2000a) and the pseudo amino acid composition approach with 28 pseudo components generated by CA images as described in this paper and their weights are $w_j = 1/27000$ $(j = 1, 2, \ldots, 28)$

which 112 are chloroplast proteins, 761 cytoplasm, 19 cytoskeleton, 106 endoplasmic reticulum, 95 extracellular, 4 Golgi apparatus, 31 lysosome, 163 mitochondria, 418 nucleus proteins, 23 peroxisome, and 762 plasma membrane.

The overall success rates thus obtained by the self-consistency test, jackknife test, and independent dataset test are given in Table 2. For facilitating comparison, the corresponding success rates obtained by ProtLoc algorithm (Cedano et al., 1997) and digital signal algorithm (Pan et al., 2003) are listed in the same table as well. Among the above three test methods, the jackknife test is deemed the most rigorous and objective (Chou and Zhang, 1995), and thereby the jackknife test has been used by more and more investigators to examine the power of a statistical predictor (Cai, 2001; Cai et al., 2003; Chou and Cai, 2004a; Gao et al., 2005; Pan et al., 2003; Wang et al., 2004a, b; Xiao et al., 2005a, b; Zhou, 1998; Zhou and Assa-Munt, 2001; Zhou and Doctor, 2003). As we can see from Table 2, the overall jackknife success rate by the current method is over 28% higher than that by ProtLoc (Cedano et al., 1997) and about 5% higher that by the digital signal approach (Pan et al., 2003).

IV Conclusion

One of the fundamental goals in cell biology and proteomics is to identify the functions of proteins in the context of compartments that organize them in the cellular environment. To realize this, it is indispensable to first identify the subcellular locations of proteins. However, it is time-consuming and expensive to determine the localization of a newly-found protein in cells purely based on experiments. With the explosion of newly found protein sequences in the post-genomic, it is in high demand to develop a fast and powerful method for predicting the subcellular location of a query protein according to its sequence. To realize this, one has to first find the key statistical information from a great pile of protein sequences that is closely correlated with their subcellular locations.

Unfortunately, this is very difficult owing to the extreme complexity and variety of these sequences. The introduction of the pseudo amino acid composition (Chou, 2001) represents one step forward in this regard that has stimulated many follow-up studies to derive various pseudo amino acid components by different approaches [see, e.g., (Cai and Chou, 2003, 2004b; Gao et al., 2005; Pan et al., 2003; Wang et al., 2004a, b; Xiao et al., 2005a)]. The essence of the problem is what kind of pseudo amino acid components can optimally reflect the statistical features of protein sequences so as to enhance the prediction quality. It is demonstrated in the present study that the novel approach by using the cellular automata images to derive the pseudo amino acid components is a very intriguing and promising avenue, as reflected by the informative pictures as well as high success rates via the self-consistency, jackknife, and independent dataset tests.

References

Boland MV, Markey MK, Murphy RF (1998) Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. Cytometry 33: 366–375

Cai YD (2001) Is it a paradox or misinterpretation. PROTEINS: Structure, Function, and Genetics 43: 336–338

Cai YD, Chou KC (2003) Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. Biochem Biophys Res Comm 305: 407–411

Cai YD, Chou KC (2004a) Predicting 22 protein localizations in budding yeast. Biochem Biophys Res Comm 323: 425–428

Cai YD, Chou KC (2004b) Predicting subcellular localization of proteins in a hybridization space. Bioinformatics 20: 1151–1156

Cai YD, Liu XJ, Xu XB, Chou KC (2002a) Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. J Cell Biochem 84: 343–348

Cai YD, Liu XJ, Xu XB, Chou KC (2002b) SVM for predicting membrane protein types by incorporating quasi-sequence-order effect. Internet. Electronic Journal of Molecular Design 1: 219–226

Cai YD, Zhou GP, Chou KC (2003) Support vector machines for predicting membrane protein types by using functional domain composition. Biophys J 84: 3257–3263

Cai YD, Zhou GP, Chou KC (2005) Predicting enzyme family classes by hybridizing gene product composition and pseudo-amino acid composition. J Theor Biol 234: 145–149

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

Chou KC (2000a) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem Biophys Res Commun 278: 477–483

Chou KC (2000b) Review: Prediction of protein structural classes and subcellular locations. Curr Protein Pept Sci 1: 171–208

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

Chou KC (2002) A new branch of proteomics: prediction of protein cellular attributes. In: Weinrer PW, Lu Q (eds) Gene cloning & expression technologies, Chapter 4. Eaton Publishing, Westborough, MA, pp 57–70

Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21: 10–19

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) Predicting protein quaternary structure by pseudo amino acid composition. PROTEINS: Structure, Function, and Genetics 53: 282–289

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: 1085)

Chou KC, Cai YD (2004a) Predicting protein structural class by functional domain composition. Biochem Biophys Res Comm 321: 1007–1009 (Corrigendum: ibid. (2005) 329: 1362)

Chou KC, Cai YD (2004b) Predicting subcellular localization of proteins by hybridizing functional domain composition and pseudo-amino acid composition. J Cell Biochem 91: 1197–1203

Chou KC, Cai YD (2004c) Prediction of protein subcellular locations by GO-FunD-PseAA predicor. Biochem Biophys Res Commun 320: 1236–1239

Chou KC, Cai YD (2005) Predicting protein localization in budding yeast. Bioinformatics 21: 944–950

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

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, 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

Gao Y, Shao SH, Xiao X, Ding YS, Huang YS, Huang ZD, Chou KC (2005) Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. Amino Acids 28: 373–376

Gusev VD, Nemytikova LA, Chuzhanova NA (2001) A rapid method for detecting interconnections between functionally and/or evolutionary close biological sequences. Mol Biol (Mosk) 35: 1015–1022

Haddadnia J, Faez K, Ahmadi M (2002) A neural based human face recognition system using an efficient feature extraction method with pseudo zernike moment. J Circuits, Systems, and Computers 11: 283–304

Murphy RF, Boland MV, Velliste M (2000) Towards a systematics for protein subcellular location: quantitative description of protein localization patterns and automated analysis of fluorescence microscope images. Proc Int Conf Intell Syst Mol Biol 8: 251–259

Nakai K (2000) Protein sorting signals and prediction of subcellular localization. Adv Protein Chem 54: 277–344

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

Park KJ, Kanehisa M (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acid and amino acid pairs. Bioinformatics 19: 1656–1663

Portilla J, Simoncelli EP (2000) A parametric texture model based on joint statistics of complex wavelet coefficients. Int J Comput Vision 40: 49–71

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

Wang M, Yang J, Xu ZJ, Chou KC (2004b) SLLE for predicting membrane protein types. J Theor Biol 232: 7–15

Wang M, Yao JS, Huang ZD, Xu ZJ, Liu GP, Zhao HY, Wang XY, Yang J, Zhu YS, Chou KC (2005) A new nucleotide-composition based fingerprint of SARS-CoV with visualization analysis. Med Chem 1: 39–47

Wolfram S (2002) A new kind of science. Wolfram Media Inc., Champaign, IL

Xiao X, Shao S, Ding1 Y, Huang Z, Huang Y, Chou KC (2005a) Using complexity measure factor to predict protein subcellular location. Amino Acids 28: 57–61

Xiao X, Shao S, Ding Y, Huang Z, Chen X, Chou KC (2005b) An application of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation. J Theor Biol 235: 555–565

Xiao X, Shao S, Ding Y, Huang Z, Chen X, Chou KC (2005c) Using cellular automata to generate Image representation for biological sequences. Amino Acids 28: 29–35

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

Zhu SC, Wu Y, Mumford D (1997) Minimax entropy principle and its application to texture modeling. Neural comput 9: 1627–1660

Ziv J, Lempel A (1976) On the complexity of finite sequences. IEEE Trans Inf Theor IT-22: 75–81

Authors' address: Prof. Kuo-Chen Chou, Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, U.S.A.,
E-mail: kchou@san.rr.com