Amino Acids (2008) 34: 111–117
DOI 10.1007/s00726-007-0550-z
Printed in The Netherlands

Amino Acids

Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel-Ziv complexity

Y. Diao[1,2], **D. Ma**[1], **Z. Wen**[1,2], **J. Yin**[1], **J. Xiang**[1], and **M. Li**[1,2]

[1] College of Chemistry, Sichuan University, Chengdu, Sichuan, China
[2] State Key Laboratory of Biotherapy, Sichuan University, Chengdu, Sichuan, China

Received February 4, 2007
Accepted March 21, 2007
Published online May 23, 2007;

Summary. Transmembrane (TM) proteins represent about 20–30% of the protein sequences in higher eukaryotes, playing important roles across a range of cellular functions. Moreover, knowledge about topology of these proteins often provides crucial hints toward their function. Due to the difficulties in experimental structure determinations of TM protein, theoretical prediction methods are highly preferred in identifying the topology of newly found ones according to their primary sequences, useful in both basic research and drug discovery. In this paper, based on the concept of pseudo amino acid composition (PseAA) that can incorporate sequence-order information of a protein sequence so as to remarkably enhance the power of discrete models (Chou, K. C., Proteins: Structure, Function, and Genetics, 2001, 43: 246–255), cellular automata and Lempel-Ziv complexity are introduced to predict the TM regions of integral membrane proteins including both α-helical and β-barrel membrane proteins, validated by jackknife test. The result thus obtained is quite promising, which indicates that the current approach might be a quite potential high throughput tool in the post-genomic era. The source code and dataset are available for academic users at liml@scu.edu.cn.

Keywords: Cellular automata – Pseudo amino acid composition – Lempel-Ziv complexity – Augmented covariant-discriminant algorithm – Chou's invariance theorem – Transmembrane regions

1. Introduction

Membrane proteins are found mostly in the cell membranes of both prokaryotic and eukaryotic organisms, performing a variety of biologically important functions such as ion tunnel, nutrient transportation, membrane adhesion, catalytic activity, and receptors of signal molecules including neurotransmitters, peptide hormones, chemokines and cytokines, etc. However, they usually form stable natural conformation with bio-membrane, which makes it difficult to determine their 3D structure with X-ray crystallography or nuclear magnetic resonance spectroscopy. Practically, among tens of thousands of proteins with known 3D structure, membrane proteins constitute only a trivial proportion (Kuhlbrandt and Wang, 1994). As knowledge of topology structure of membrane proteins has important significance in both basic research and drug discovery, people have enormous interest in obtaining structure information of membrane proteins from their primary sequences through theoretical prediction.

Research shows that the structure of membrane proteins can be predicted according to the hydrophobicity of their primary sequence. Kyte and Doolittle (1982) computed the hydrophobicity scales of 20 amino acids, transformed the test sequence into hydrophobicity profile through sliding windows of fixed size, set appropriate threshold and predicted possible TM regions. von Heijine (1986) proposed the so-called ''positive inside rule'', providing further instructs to membrane protein structure prediction methods. In recent years, along with the increase of new-found membrane proteins with determined structure, several statistical prediction methods had been proposed, using artificial neural network (Cao et al., 2006; Chen and Rost, 2002; Rost et al., 1996), hidden Markov model (Tusnady and Simon, 1998; Zhou and Zhou, 2003) and support vector machine (Zheng and John, 2004). Despite preferable prediction accuracy, they had shortcomings such as requiring users to specify the length range of TM segments so as to adjust the size of scanning window, could not recognize β-barrel TM proteins, which limited their application in some cases.

The present study is to develop an integrative method for predicting the topology of TM proteins on the base of PseAA (Chou, 2001). First, scanning the requested protein

sequence with a fixed-size window of 20 amino acids residues; then, the segments thus obtained are transformed into binary sequences by an encoding procedure, upon which the cellular automata are applied to derive PseAA components; finally, the augmented covariant-discriminant algorithm (Chou, 2000a) is used to predict the topology of requested protein. The result suggests this method is an effective tool for the prediction of both α-helical and β-barrel proteins with high accuracy, validated by jackknife cross-validation test. Moreover, based solely on the amino acid sequence, this method does not require any other annotations or sequence alignment information.

2. Materials and methods

2.1 Data sets

The UniProt/Swiss-Prot at www.ebi.ac.uk/swissprot (Release 46.6) and PDB at www.rcsb.org/pdb were used to construct the dataset used in this study. All sequences with ambiguous words, such as POTENTIAL, PUTATIVE, or HYPOTHETICAL and fragments having less than 50 amino acid residues were excluded. Some sequences with high identity of 90% were not removed in order to provide a wide range prediction, while most sequences were clustered lower than 25% identity using Clustal W program (Thompson et al., 1994). The final dataset consisted of 146 entries that appeared as whole sequences and had reliable experimental annotations for TM regions, the accession numbers of which were given in the supplementary materials.

2.2 Digital coding for amino acid

For the existent 20 native amino acids, a set of digital codes are introduced to represent them (Table 1), which is capable of reflecting the chemical physical properties of amino acids and their degeneracy as well, by means of the similarity rule, complementarity rule and molecular recognition theory (Kuric, 2007; Xiao et al., 2005a, b, 2006a).

2.3 Cellular automata

According to the self-production principle of biological domain, von Neumann (1966) proposed the concept and model of cellular automaton, which was a dynamical system discrete in both time and spatial. Spread in regular lattice, each cell adopted finite discrete state and updated synchronously according to explicit local rule. The evolution of entire dynamical system was implemented through simple and exact interactions between those cells, the characteristic of which was discrete in time, spatial and state, every variable only adopted finite state, and the state transforming rule was local both in time and spatial.

Cellular automata can be used to study many universal phenomena, including communication, information processing, computation, conformation, growth, replication, competition and evolution, etc. Meanwhile, they are effective and powerful modeling tools for investigating system emergent behaviors and complex phenomena in dynamical system theory, such as order, chaos, turbulence, non-symmetry and fractal, etc (Wolfram, 1984, 1986; Martin et al., 1984).

Generally, cellular automata can be formulated as

$$A = (L_d, S, N, f, B) \tag{1}$$

where 1) A represents cellular automata; 2) L_d is the cellular space, d is a positive integer, standing for the dimension of cellular space; 3) S is a finite state set; 4) N represents local neighborhood, which can be denoted as a vector comprising n different cell states:

$$N = (s_1, s_2, \ldots, s_n), \quad s_i \in S, \quad i \in \{1, 2, \ldots, n\} \tag{2}$$

where n is the number of the central cell's neighbor, including itself.

5) f is state transfer function that defines how the state changes from one time to the next. For example, elementary cellular automaton that we adopt in this study is a one-dimensional cellular automaton with its state set S comprising only two states $\{0, 1\}$ and neighbor radius $r = 1$, whose local transfer function f is

$$S_i(t+1) = f(S_{i-1}(t),\ S_i(t),\ S_{i+1}(t)), \quad i \in \{1, 2, \ldots, n\} \tag{3}$$

Considering S_i as the current cell, $S_i(t)$ is its state at time step t; $S_{i-1}(t)$ and $S_{i+1}(t)$ represent the state of its two neighbors at time step t; $S_i(t+1)$ is its state at the next time step $(t+1)$. Moreover, to any i and t, $S_i(t) \in \{0, 1\}$.

In general, if there are K states and if each cell is taken to have N neighbors (including itself), then there should be K^N possible neighborhood configurations and K^{K^N} different local rules. Consequently, for elementary cellular automata, there should be $2^{2^3} = 256$ different rules that can be easily encoded from binary byte into decimal numbers between 0 and 255. For example, rule number 84 that we adopt in this study corresponds to Fig. 1, where this local function is applied simultaneously to all lattice sites.

6) B represents boundary condition, which could be classified into four types named fixed, random, circulating and reflecting boundary condition, respectively. Considering the characteristic of most proteins is self-consistency and self-organization, we adopt the reflecting boundary condition in this study with the iterative formula given by:

$$S_1(t+1) = f(S_2(t), S_1(t), S_2(t)) \tag{4}$$

$$S_n(t+1) = f(S_{n-1}(t), S_n(t), S_{n-1}(t)) \tag{5}$$

where symbols denote the same as in Eq. (3).

2.4 Lempel-Ziv complexity

There are several complexity measures to test the randomness of a sequence. Linear complexity, for example, is one of these measures. Being an important measure used in cryptography, Lempel-Ziv (LZ) complexity of a sequence is measured by the minimal number of steps required for its synthesis in a certain process (Lempel and Ziv, 1976). For each step only two operations are allowed in the process: either generating an additional symbol which ensures the uniqueness of each component or copying the longest fragment from the part of a synthesized sequence.

Suppose a string S expressed by $S_1S_2 \ldots S_n$, its substring is expressed by

$$S[i:j] = S_iS_{i+1}S_{i+2} \cdots S_j \quad (1 \le i \le j \le N) \tag{6}$$

Table 1. Digital codes of 20 native amino acids

Amino acid	P	L	Q	H	R	S	F	Y	W	C
Binary notation	00001	00011	00100	00101	00110	01001	01011	01100	01110	01111
Amino acid	T	I	M	K	N	A	V	D	E	G
Binary notation	10000	10010	10011	10100	10101	11001	11010	11100	11101	11110

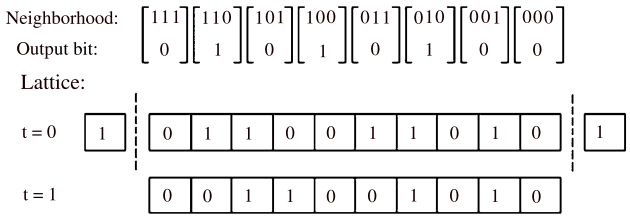

Fig. 1. Illustration of a one-dimensional, two-state, nearest-neighbor ($r = 1$) cellular automaton. Both the lattice and the rule table for updating the lattice are illustrated. The configuration of the cellular automaton is shown at two successive time steps, under reflecting boundary conditions: the left neighbor of the leftmost cell is its right neighbor and vice versa, as if a mirror were placed on the boundary of the lattice

The complexity measure, $C_{LS}(S)$, of string S is the minimal number of steps that are needed to synthesized S according to the following procedure.

$$H(S) = S[1 : i_1]S[i_1 + 1 : i_2]\cdots S[i_{k-1} + 1 : i_k]\cdots S[i_{n-1} + 1 : N] \quad (7)$$

where the uniqueness of every substring is generated by adding an additional symbol to the existent substring. For instance the LZ complexity of string 01000110110000010100011 is 10, because 10 is the minimal number of steps required to synthesize this binary sequence as 0|1|00|01|10|11|000|001|010|011.

2.5 Pseudo amino acid composition

To avoid completely losing the sequence-order information as representing a protein by its amino acid composition, Chou (2001) proposed PseAA that could partially reflect the sequence-order information through a set of correlation factors. Introducing the concept of PseAA has stimulated many follow-up studies for improving the prediction quality in various areas as reflected by a series of recent publications (Chen et al., 2006a, b; Du and Li, 2006; Gao et al., 2005; Liu et al., 2005; Mondal et al., 2006; Pan et al., 2003; Shen and Chou, 2005a, b, 2006a, b, 2007; Shen et al., 2006; Wang et al., 2004, 2006; Xiao et al., 2006a, b; Zhang et al., 2006a).

In this study, LZ complexity factor is used to serve as PseAA components, which could also partially reflect sequence effect and length information of proteins. However, N complexity factors can be derived from the requested protein sequence after N-times cellular automata evolution, the first 26 of which are adopted in this study by the reasons given in the discussion section. Accordingly, by following the procedure defined by Eqs. (8) and (9), a protein can be expressed by a vector or a point in 46-dimensional space:

$$X = (x_1, x_2, x_3, \ldots, x_{46})^{\mathbf{T}} \quad (8)$$

$$x_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{26} p_j} & (1 \le k \le 20) \\ \dfrac{wp_{(k-20)}}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{26} p_j} & (21 \le k \le 46) \end{cases} \quad (9)$$

where f_i ($i = 1, 2, \ldots, 20$) are the occurrence frequencies of the 20 amino acids in the requested sequence, arranged alphabetically according to their single letter codes, p_j ($j = 1, 2, \ldots, 26$) are the complexity factors of the transformed sequence, and w the weight factor. According to previous publications (Chou, 2000a, 2001, 2005a; Chou and Cai, 2003a, b; Xiao et al., 2006b) and several condition tests, we finally chose $w = 1/600$ to make the results of Eq. (9) within the range easier to be handled.

Now the augmented covariant-discriminant algorithm (Chou, 2000a, 2001) is used to perform the prediction, which is a combination of Mahalanobis distance and Chou's invariance theorem for treating degenerative space (Chou, 1995; Chou et al., 1998; Chou and Zhang, 1994; Zhou, 1998; Zhou and Assa-Munt, 2001; Zhou and Doctor, 2003). It should be noticed that owing to the normalization condition imposed by Eq. (9), a dimension-reduced operation by leaving out one of the 46 components and making the rest completely independent is needed before utilizing this predicting algorithm. Otherwise, the covariant matrix in the covariant-discriminant algorithm would be divergent (Chou and Zhang, 1994). However, which one of the 46 components should be removed? According to Chou's invariance theorem (Chou, 1995), the values of the covariant-discriminant function will remain the same regardless of which one of the 46 components is left out.

3. Results and discussions

As is known, integral membrane proteins are divided into two distinct structural classes, the α-helical membrane proteins and the β-barrel membrane proteins. While the former class is more abundant and well studied, members of the latter class escape attention of most researchers, which are located in the outer membrane of Gram-negative bacteria, presumably in the outer membrane of chloroplasts and mitochondria. These proteins have membrane spanning segments formed by antiparallel β-strands, creating a channel in the form of a barrel that spans the outer membrane (Chou et al., 1990; Pautsch and Schultz, 1998).

As far as α-helical membrane proteins are considered, 20 amino acids residues, on average, are needed to span the bilayer lipid membrane. However, as for β-barrel ones, 10 amino acids residues are enough to do the same job because β-sheet is theoretically more stretched than α-helix, since its lower compression ratio. Fortunately, there exists another secondary structure named β-turn (Chou, 2000c), constituted by four to five amino acids residues, between these concatenated β-strands. Consequently, a compound structure constituted by a β-strand and two β-turns on its both ends now becomes our target of pattern recognition, whose total length is averagely 18–20 amino acids residues. Considering the above-mentioned reasons, the size of sliding window in this study is fixed at 20 residues. Therefore, we can obtain $n - 20 + 1$ segments from the sample sequence, which is supposed to have n amino acids residues in total.

Through the encoding rule given in Table 1, these protein segments are transformed to a serial of binary sequences with a length of 100. Under the evolution rule of 84th and reflecting boundary condition defined by Eqs. (4) and (5), every binary sequence is transformed by a two-state, one-dimensional cellular automaton into a 2D matrix with its data in each row derived from its previous, starting from the second. But how many iterative steps will be adequate in this study? The clue lies in the above-mentioned compound structure, namely a β-sheet and two β-turns on its both ends.

As is mentioned above, the amino acids residues that make up of a β-strand could fill only half of the sliding window, which makes its lateral β-turns of critical importance in recognizing its exact position. Considering the encoding procedure, the 5 amino acids residues in a β-turn are transformed into binary sequence of length 25. At each iteration, considering the localization characteristic of cellular automata, the information contained in the left flank can only diffuse one grid to the right, while that in the right flank diffusing one grid to the left. After 25 times of iteration, the flows of information from both ends converge at the center. At the next step, the flows overlap and that is the least iterative time needed to cover our request.

Computing the LZ complexity of the 26 rows of the 2D matrix generated by the above-mentioned cellular automata, plus the amino acids composition of the same window, we obtain the PseAA needed in augmented covariant-discriminant algorithm, which is then used to predict which class this segment should belong to, α-helix, β-strands or non-TM region. Particularly, as for the first window, an additional prediction is made to determine the N-terminal of the requested protein is located in which side of the membrane, providing more details for the topology structure prediction. Suppose the N-terminal is located inside the cell or organelle, all the singular TM segments should span the membrane from inside to outside and even TM segments from outside to inside.

In the prediction of TM regions, if the amino acids composition of the sliding window is completely uniform, or the probability of each amino acid equals to $1/20$, the LZ complexity factor gets its maximum; if the sliding window only contains one kind of amino acid, the LZ complexity factor gets its minimum. So, the magnitude of LZ complexity factor partially reflects the deviation extent between the sliding window and average expectation of all proteins in the training set. The permutation of LZ complexity factors generated from TM region, including its lateral 40 amino acids residues, 20 at each side, could provide crucial clue to its exact location in the requested proteins.

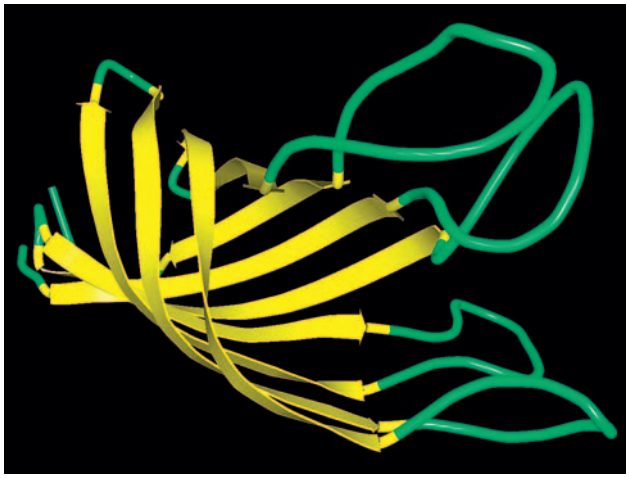

Fig. 2. Ribbon drawing to show the structure of protein *OmpA*

Table 2 shows the prediction result of this method in recognizing 8 TM segments of *Outer membrane protein A* (*OmpA*), whose 3D structure is also shown in Fig 2. While the predictor HMMTOP (Tusnady and Simon, 1998) and TMpred (Hofmann and Stoffel, 1993) fail the test, the current method obtains preferable results compared with another predictor named PRED-TMBB (Bagos et al., 2004), which is specialized in recognizing β-barrel proteins.

Upon all the 914 TM segments of 146 TM proteins in our data set, the prediction result of the current method is compared with those of three other algorithms (Table 3), using the jackknife cross-validation test. As is known, this test has been considered as one of the most objective and rigorous test methods in examining the power of a prediction method (Chou and Zhang, 1995).

In the process of predicting, only primary sequence of the test protein is inputted, without any other parameters, annotations or sequence alignment information. Although prediction accuracy may decrease by doing so, experiment result becomes more consistent and replicable. Moreover, because of the suppleness of biological molecules and practical limits of experiments, margins often exist when defining the ends of TM segments. Accordingly, in this study, we define no more than 5 residues deviation, in each end, is acceptable. The underlying physical significance of this is that every 20-residue segment comprising more than 5 consecutive

Table 2. Comparing prediction results of the current method with those of three other algorithms in recognizing 8 TM segments of Outer membrane protein A (*OmpA*)

	Observed	This paper	HMMTOP	TMpred	PRED-TMBB
TM1	6–16	5–17	–	–	7–15
TM2	34–45	32–62	–	–	43–51
TM3	49–60	–	–	–	55–63
TM4	75–86	72–84	–	–	77–85
TM5	91–103	93–105	–	–	91–101
TM6	121–130	118–128	–	–	119–129
TM7	135–143	133–143	–	–	133–143
TM8	161–170	159–169	–	–	160–170

Table 3. Performance of the current method, as compared to three other algorithms in recognizing TM segments in our data set

Data set	Method	N_{obs}	N_{pred}	N_{cor}	Q_P (%)	M (%)	C (%)
α-helical	This paper	690	721	574	81.4	83.2	79.6
	HMMTOP		678	517	75.6	74.9	76.3
	TMpred		703	458	65.7	66.4	65.1
	PRED-TMBB[a]		2	0	–	–	–
β-barrel	This paper	224	236	175	76.1	78.1	74.2
	HMMTOP		1	0	–	–	–
	TMpred		65	6	5.0	2.7	9.2
	PRED-TMBB[a]		230	186	81.9	83.0	80.9
Total	This paper	914	957	749	80.1	81.9	78.3
	HMMTOP		679	517	65.6	56.6	76.1
	TMpred		768	464	55.4	50.8	60.4
	PRED-TMBB[a]		232	186	40.4	20.4	80.2

N_{obs}, N_{pred}, and N_{cor} are the number of observed, predicted, and correctly predicted TM segments, respectively; M (sensitivity) is the percentage of correctly predicted segments over the observed segments; C (specificity) is the percentage of correctly predicted segments over the predicted segments; and $Q_p = \sqrt{M \cdot C}$ (Tusnady and Simon, 1998)
[a] Using the standard Viterbi algorithm.

residues that belong to TM regions, should be identified as a TM comprising segment out of a background noise of $20^5 = 320{,}000$ segments.

To show more details, we list the prediction result of α-helical and β-barrel proteins separately. From Table 3 we can summarize that 1) in the prediction of α-helical proteins, the current method obtains a jackknife success rate slightly higher than that of HMMTOP and obviously higher than that of TMpred, while PRED-TMBB fails this test; 2) in the prediction of β-barrel proteins, the current method obtains a jackknife success rate mildly less than that of PRED-TMBB, while the other two predictors almost fail the test. However, upon all the proteins in the data set, the current method obtains an overall success rate considerably higher than the other three predictors.

4. Conclusions

Having the dynamical characteristics of decentralized decision-making and highly parallel information processing, cellular automata has been the subject of interest from the computer scientists for many years, especially in the domain of artificial intelligence and artificial life. However, it is also an intriguing and promising approach especially useful for investigating complicated biological sequences. It is demonstrated in this study that using cellular automata to derive PseAA components can effectively reflect the overall sequence-order feature of a protein, useful for the prediction of TM regions in proteins. Meanwhile, it does not escape our attention that the possible usage of this method on improving the prediction quality for a series of other protein attributes, such as subcellular localization (Cai and Chou, 2004a, b; Chou, 2000b; Chou and Cai, 2002, 2003b, 2004b, c, 2005; Chou and Elrod, 1999; Chou and Shen, 2006a, b, c; Zhang et al., 2006b), enzyme family classes (Cai et al., 2005), G protein coupled receptor classification (Chou, 2005b; Wen et al., 2006), and protein quaternary structure types (Chou and Cai, 2004a), among many others.

Acknowledgements

This work was supported by Sichuan University Student Innovation Found 2006L012, and the State Key Laboratory of Chemo-Biosensing and Chemometrics, Hunan University.

References

Bagos PG, Liakopoulos TD, Spyropoulas IC, Hamodrakas SJ (2004) PRED-TMBB: a web server for predicting the topology of β-barrel outer membrane proteins. Nucleic Acids Res 32: W400–W404

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

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

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

Cao BQ, Porollo A, Adamczak R, Jarrell M, Meller J (2006) Enhanced recognition of protein transmembrane domains with prediction-based structural profiles. Bioinformatics 22: 303–309

Chen CP, Rost B (2002) State-of-the-art in membrane protein prediction. Applied Bioinformatics 1: 21–35

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

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

Chou KC, Carlacci L, Maggiora GM (1990) Conformational and geometrical properties of idealized beta-barrels in proteins. J Mol Biol 213: 315–326

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 (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 KC, Liu W, Maggiora GM, Zhang CT (1998) Prediction and classification of domain structural classes. Proteins Struct Funct Genet 31: 97–103

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

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 (2000c) Review: prediction of tight turns and their types in proteins. Anal Biochem 286: 1–16

Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins Struct Funct Genet 43: 246–255

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 Struct Funct Genet 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

Chou KC, Cai YD (2004a) Predicting protein structural class by functional domain composition. Biochem Biophys Res Commun 321: 1007–1009

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 predictor. Biochem Biophys Res Commun 320: 1236–1239

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

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

Chou KC (2005b) Prediction of G-protein-coupled receptor classes. J Proteome Res 4: 1413–1418

Chou KC, Shen HB (2006a) Predicting protein subcellular location by fusing multiple classifiers. J Cell Biochem 99: 517–527

Chou KC, Shen HB (2006b) Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem Biophys Res Commun 347: 150–157

Chou KC, Shen HB (2006c) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J Proteome Res 5: 1888–1897

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

Gao QB, Wang ZZ, Yan C, Du YH (2005) Prediction of protein subcellular location using a combined feature of sequence. FEBS Lett 579: 3444–3448

Hofmann K, Stoffel W (1993) A database of membrane spanning proteins segments. Biol Chem 374: 166

Kuhlbrandt W, Wang D (1994) Three-dimensional structure of plant light-harvesting complex determined by electron crystallography. Nature 350: 130–134

Kuric L (2007) The digital language of amino acids. Amino Acids. doi: 10.1007/s00726-006-0476-x

Kyte J, Doolittle RF (1982) A simple method for displaying the hydroparthic character of a protein. J Mol Biol 157: 105–132

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

Liu H, Wang M, Chou KC (2005) Low-frequency Fourier spectrum for predicting membrane protein types. Biochem Biophys Res Commun 336: 737–739

Martin O, Odlyzko AM, Wolfram S (1984) Algebraic properties of cellular automata. Commun Math Phys 93: 219–258

Mondal S, Bhavna R, Mohan BR, Ramakumar S (2006) Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. J Theor Biol 243: 252–260

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

Pautsch A, Schultz GE (1998) Structure of the outer membrane protein A transmembrane domain. Nat Struct Biol 5: 1013–1017

Rost B, Fariselli P, Casadio R (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. Protein Sci 5: 1704–1718

Shen HB, Chou KC (2005a) Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochem Biophys Res Commun 337: 752–756

Shen HB, Chou KC (2005b) Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. Biochem Biophys Res Commun 334: 288–292

Shen HB, Yang J, Chou KC (2006) Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. J Theor Biol 240: 9–13

Shen HB, Chou KC (2006a) Ensemble classifier for protein fold pattern recognition. Bioinformatics 22: 1717–1722

Shen HB, Chou KC (2006b) Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers 85: 233–240

Shen HB, Chou KC (2007) Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. Protein Eng Des Sel 20: 39–46

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680

Tusnady GE, Simon I (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol 283: 489–506

von Heijine G (1986) The distribution of positively charged residues in bacterial inner membrane proteins correlates with the transmembrane topology. EMBO J 5: 3021–3027

von Neumann J (1966) Theory of self-reproducing automata. University of Illinois Press

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 Des Sel 17: 509–516

Wang SQ, Yang J, Chou KC (2006) Using stacked generalization to predict membrane protein types based on pseudo amino acid composition. J Theor Biol 242: 941–946

Wen Z, Li M, Li Y, Guo Y, Wang K (2006) 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 automata as models of complexity. Nature 311: 419

Wolfram S (1986) Cellular automaton fluid: basic theory. J Stat Phys 45: 471

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

Xiao X, Shao S, Ding Y, Huang Z, Chou KC (2006a) Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30: 49–54

Xiao X, Shao S, Huang Z, Chou KC (2006b) Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J Comput Chem 27: 478–482

Zhang SW, Pan Q, Zhang HC, Shi JY (2006a) 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 T, Ding Y, Chou KC (2006b) Prediction of protein subcellular location using hydrophobic patterns of amino acid sequence. Comput Biol Chem 30: 367–371

Zheng Y, John S (2004) SVMtm: support vector machines to predict transmembrane segments. J Comput Chem 25: 632–636

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 Struct Funct Genet 44: 57–59

Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins Struct Funct Genet 50: 44–48

Zhou HY, Zhou YQ (2003) Predicting the topology of transmembrane helical proteins using mean burial propensity and a hidden-Markov-model-based method. Protein Sci 12: 1547–1555

Authors' address: Menglong Li, College of Chemistry, Sichuan University, Chengdu, Sichuan 610064, P.R. China,
Fax: +86-28-85412356, E-mail: liml@scu.edu.cn