Amino Acids (2005) 28: 57–61
DOI 10.1007/s00726-004-0148-7

Amino Acids
Printed in Austria

Using complexity measure factor to predict protein subcellular location

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

[1] College of Information Sciences and Technology, Donghua University, Shanghai, China
[2] Computer Department, 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, San Diego, California, U.S.A.

Received October 22, 2004
Accepted October 23, 2004
Published online December 22, 2004;

Summary. Recent advances in large-scale genome sequencing have led to the rapid accumulation of amino acid sequences of proteins whose functions are unknown. Because the functions of these proteins are closely correlated with their subcellular localizations, it is vitally important to develop an automated method as a high-throughput tool to timely identify their subcellular location. Based on the concept of the pseudo amino acid composition by which a considerable amount of sequence-order effects can be incorporated into a set of discrete numbers (Chou, K. C., Proteins: Structure, Function, and Genetics, 2001, 43: 246–255), the complexity measure approach is introduced. The advantage by incorporating the complexity measure factor as one of the pseudo amino acid components for a protein is that it can more effectively reflect its overall sequence-order feature than the conventional correlation factors. With such a formulation frame to represent the samples of protein sequences, the covariant-discriminant predictor (Chou, K. C. and Elrod, D. W., Protein Engineering, 1999, 12: 107–118) was adopted to conduct prediction. High success rates were obtained by both the jackknife cross-validation test and independent dataset test, suggesting that introduction of the concept of the complexity measure into prediction of protein subcellular location is quite promising, and might also hold a great potential as a useful vehicle for the other areas of molecular biology.

Keywords: Pseudo amino acid composition – Complexity measure factor – Covariant-discriminant algorithm – Chou's invariance theorem

I Introduction

For a newly found protein sequence, how can one timely identify which subcellular localization it belongs to? This is a big challenge today because a fundamental goal of cell biology is to define the functions of proteins in the context of compartments that organize them in the cellular environment, and the knowledge of the subcellular location of proteins is vitally important to the understanding of their functions and interactions (Chou, 2001; Chou and Cai, 2002, 2004b). Even for a function-known protein, information of its subcellular localization may provide useful insights into the specific enzyme pathway (Nakai and Kanchisa, 1992). Experimental determination of subcellular location is currently mainly accomplished by the following three approaches: cell fractionation, electron microscopy and fluorescence microscopy. These approaches are time-consuming, and might bear some sorts of subjective assumption and uncertainty (Murphy et al., 2000). Accordingly, knowledge of protein subcellular location derived from statistical prediction can play a complemental role to the above biochemical experiments. Furthermore, it can help to screen candidates for drug discovery, expedite the annotation of gene products as well as provide insights in selecting the relevant proteins for further study (Nakai, 2000; Chou, 2002, 2004a; Chou and Elrod, 1999).

To address the challenge, many efforts have been made to develop various approaches on this topic, see, e.g., Horton and Nakai (1997), Cedano et al. (1997), Reinhardt and Hubbard (1998), Chou and Elrod (1999), Yuan (1999), Nakai and Horton (1999), Chou (2000b), Cai and Chou (2000), Feng (2001), Hua and Sun (2001), Chou and Cai (2002), Cai et al. (2002), Emanuelsson et al. (2002), Pan et al. (2003), Zhou and Doctor (2003), Chou and Cai (2003a, c), Bhasin and Raghava (2004), Huang and Li (2004), Chou and Cai (2004a, 2004b). For a systematic description of development in this area at different stages, refer to the reviews by Nakai (2000) and Chou (2000b, 2002) as well as some recent papers (Chou and Cai, 2004a, b).

The present study was initiated in an attempt to develop a different approach for predicting protein subcellular location based on the concept of the pseudo amino acid composition originally proposed by Chou (2001). The introduction of pseudo amino acid composition, by which a considerable amount of the sequence-order effects can be incorporated into a set of discrete components (Chou, 2001), has stimulated the development of several powerful prediction algorithms in the relevant areas (Pan et al., 2003; Cai and Chou, 2003; Chou and Cai, 2002, 2003b, 2004a, b; Wang et al., 2004a, b; Chou, 2004b). Here we attempt to develop a different prediction algorithm by introducing a new concept, the so-called complexity measure factor, into the pseudo amino acid composition in order to make it more effectively reflect the sequence-order effect. The bottom line is that a protein sequence is actually a symbolic sequence for which the complexity measure can be used to reflect its intricate feature. There are several methods to evaluate the text complexity, such as entropy measure (Sadovsky, 2003), evaluation of the alphabet capacity *l*-gram (Gabrielian and Bolshoy, 1999), modification of complexity measure by Lempel and Ziv (Gusev et al., 1999), and stochastic complexity (Orlov et al., 2002). Complexity measure has been successfully used to predict protein-coding regions in DNA, sequences comparison and other areas. Of the known complexity measure approaches so far, the Ziv-Lempel complexity measure (Gusev et al., 1999, 2001) is the most adequate one in reflecting the repeat patterns occurring in the text, and hence was adopted in this study.

II Method

The key to improve the prediction quality is how to effectively define the components of pseudo amino acid composition for different cases studied (Chou, 2004b; Chou and Cai, 2003c). In this study, we are to use the complexity measure and auto correlation functions to define the pseudo amino acid components for a protein sequence.

The complexity of a sequence can be measured by the minimal number of steps required for its synthesis in a certain process (Ziv and Lempel, 1976). For each step only two operations were allowed in the process: either generating a new symbol or copying a fragment from the part of a synthesized sequence. Suppose a string S expressed by

$$S = \alpha_1\alpha_2\alpha_3\ldots\alpha_N \tag{1}$$

Its substring is expressed by

$$S[i:j] = \alpha_i\alpha_{i+1}\alpha_{i+2}\ldots\alpha_j \quad (1 \le i < j \le N) \tag{2}$$

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

$$H(S) = S[1:i_1]S[i_1+1:i_2]\cdots S[i_{k-1}+1:i_k]\cdots S[i_{m-1}+1:N] \quad (1 < k < m \le N) \tag{3}$$

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 given by

$$i_k - i_{k-1} = \max_{j \le i_{k-1}} \{l_j : S[i_{k-1}+1 : i_{k-1}+l_j] = S[j : j+l_j-1]\} \tag{4}$$

is copied from the prefix $S[1:i_{k-1}+l_j-1]$, where l_j is the length of the fragment being copied. The Lempel-Ziv complexity algorithm in its simplest form is given in Box 1. For example, for the string $S = 0001101001000101$, the Lempel-Ziv schema of synthesis gives the following components and corresponding complexity:

$$\begin{cases} H(S) = 0 \bullet 001 \bullet 10 \bullet 100 \bullet 1000 \bullet 101 \\ C_{LS}(S) = 6 \end{cases} \tag{5}$$

Now, let us represent the sequence of a protein by a stochastic digital signal, i.e.,

$$S = \{x_i\}, \quad i = 1, 2, \ldots, N \tag{6}$$

where x_i is the numerical code for ith amino acid in protein sequence, and N the length of sequence. Its complexity factor C_{LS} (S) can be easily derived according to Table 1 and Box 1. And the value thus obtained is used to represent the 1st pseudo amino acid component Φ_1 for the protein concerned. The other pseudo amino acid components are defined as follows:

$$\Phi_{\lambda+1} = \frac{1}{N-\lambda}\sum_{i=1}^{N-\lambda} x(i)x(i+\lambda), \quad \lambda = 1, 2, \ldots, 5 \tag{7}$$

The five factors in Eq. 7 can be easily computed according to Table 1 that, to some degree, reflect the sequence-order effect, as illustrated in Fig. 1 of (Chou, 2000a).

Given a protein sequence, we can generate a set of 6 numbers Φ_1, Φ_2, Φ_3, Φ_4, Φ_5, and Φ_6 according to Table 1, Box 1 and Eq. 6. Thus, following

Table 1. Digital codes for the 20 native amino acids

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

```
Routine Lempel-Ziv complexity
    STRING = get input character
    CHARACTER = ''
    C_LS = 1
    WHILE there are still input characters DO
            Character = get input character
            CHARACTER = CHARACTER + Character
            Q = dele (STRING + CHARACTER)
        IF CHARACTER is not substring of the Q
            STRING = STRING + CHARACTER
            C_LS = C_LS + 1
            CHARACTER = ''
        END of IF
    END of WHILE
```

[a] The function of the dele (STRING + CHARACTER) is to delete the last character in the character string of STRING + CHARACTER.

Box 1. The Lempel-Ziv complexity algorithm[a]

exactly the same procedure as described by Chou (2001), a protein **X** can be expressed by a vector or a point in a 26D (dimensional) space, i.e.,

$$\mathbf{X} = (x_1, x_2, x_3, \ldots, x_{26})^{\mathbf{T}} \tag{8}$$

where **T** is the transpose operator, and

$$x_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{20} f_i + \sum_{j=1}^{6} w_j \Phi_j}, & (1 \le k \le 20) \\ \dfrac{w_{(k-20)} \Phi_{(k-20)}}{\sum_{i=1}^{20} f_i + \sum_{j=1}^{6} w_j \Phi_j}, & (21 \le k \le 26) \end{cases} \tag{9}$$

where $f_i (i = 1, 2, \ldots, 20)$ are the occurrence frequencies of the 20 native amino acids in the protein (Chou and Zhang, 1993), and w_j the weight factor for the jth factor

$$[w_1, w_2, w_3, w_4, w_5, w_6] = \left[\frac{1}{7000}, \frac{1}{7000}, \frac{1}{7000}, \frac{1}{7000}, \frac{1}{7000}, \frac{1}{8000}\right] \tag{10}$$

Now we can directly use the augmented covariant-discriminant algorithm develop by Chou (2001) to perform prediction. It is instructive to point out that owing to the normalization condition imposed by Eq. 9, the 20 + 6 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 (26-1)D space instead of 26D space. Otherwise, a divergence difficulty will occur. However, which one of the 26 components should be removed? Anyone. The reason is that according to a theorem given and proved by Chou (1995), which is generally quoted as "Chou's *Invariance Theorem*" (Pan et al., 2003; Zhou and Assa-Munt, 2001; Zhou and Doctor, 2003), the values of the covariant discriminant function will remain the same regardless of which one of the 26 components is left out.

III Results and discussion

The training dataset and independent dataset taken from Chou (2001) are used to test the current method. The training dataset consists of 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. The independent dataset consists of 2,494 protein sequences, of 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 prediction quality was examined by the resubstitution, jackknife, and independent dataset tests, respectively.

(1) Resubstitution test

This method is for testing the self-consistency of a method. During the test process, the subcellular location of each protein in a dataset was predicted by using the parameters derived from the same dataset, the so-called training dataset. The success rates thus obtained on the training dataset are given in Table 2, form which we can see that the overall success rate is 86%. This is 40% higher than the rate by the ProtLock predictor (Cedano et al., 1997) and 4.5% higher than the rate by Pan et al. (2003), respectively. However, to really examine the power of a predictor, a cross-validation test is needed, as described below.

(2) Jackknife test

As is well known, the single independent dataset test, subsampling test and jackknife test are the three procedures often used for cross-validation in literature (Chou and Zhang, 1995). Of these three, the jackknife test is regarded as the most objective and effective one (Zhou, 1998; Zhou and Assa-Munt, 2001). The mathematical principle and a comprehensive discussion about this can be found in a monograph (Mardia et al., 1979) and a review paper (Chou and Zhang, 1995), respectively. Accordingly, the real power of a predictor should be measured by the success rate of jackknife test. The success rates by the jackknife test on the training dataset are given in Table 3, from which we can see that overall success rate is 73.1%, which is 6% higher than the rate by Pan et al. (2003).

(3) Independent dataset test

Finally, as a paradigm to show how to use the present method in practical application, prediction was also performed the aforementioned independent dataset from Chou (2001). The success rates were also given in Table 3, from which we can see that the overall success rate is 79.8%, which also is 6% higher that the rate by Pan et al. (2003).

Table 2. Success rates by the resubstitution test for the training dataset from Chou (2001)

Method	Success rate (%) for each subcellular location												
	Chloro-plast	Cytop-lasmic	Cytoske-letal	ER	Extra-cellular	Golgi apparatus	Lydo-some	Mitochon-drial	Nuc-lear	Peroxi-some	Plasmer membrane	Vacuo-lar	Overall
ProtLock (Cedano et al.,1997)	42.9%	30.7%	40.5%	50.9%	28.3%	50.0%	63.2%	52.3%	54.2%	34.4%	59.8%	32.0%	$1006/2191 = 45.9\%$
Digital signal (Pan et al., 2003)	87.6%	73.7%	100%	89.8%	71.4%	100%	100%	78.6%	76.8%	100%	87.4%	100%	$1785/2191 = 81.5\%$
This paper	82.8%	83.4%	100%	87.8%	74.1%	100%	100%	81.0%	77.9%	100%	93.3%	100%	$1884/2191 = 86.0\%$

Table 3. Overall success rates by the jackknife and independent dataset tests

Algorithm	Test method	
	Jackknife[a]	Independent dataset[b]
ProtLock (Cedano et al., 1997)	$971/2191 = 44.3\%$	$1018/2494 = 40.8\%$
Digital signal (Pan et al., 2003)	$1483/2191 = 67.7\%$	$1842/2494 = 73.9\%$
This paper	$1612/2191 = 73.6\%$	$1990/2494 = 79.8\%$

[a] Using the training dataset taken from (Chou and Elrod, 1999; Chou 2001).
[b] Using the independent dataset taken from (Chou and Elrod, 1999; Chou 2001).

IV Conclusions

Different with the 20D conventional amino acid composition, the pseudo amino acid composition contains more than 20 components. It is thru the additional components that a considerable amount of the sequence-order effects can be reflected in terms of a set of discrete numbers (Chou, 2001). The pseudo amino acid composition also possesses the merit of flexibility, by which the additional components can be defined according to the case investigated to optimize the desired results. It is demonstrated in this study that the introduction of the complexity measure factor as one of the pseudo amino acid components can more effectively reflect the overall sequence-order feature of a protein, leading to higher success rates in predicting the subcellular location of proteins. It is anticipated that the concept of complexity measure might be of use to the other areas of molecular biology.

Acknowledgements

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

References

Bhasin M, Raghava GPS (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res 32: 414–419

Cai YD, Chou KC (2000) Using neural networks for prediction of subcellular location of prokaryotic and eukaryotic proteins. Molecular Cell Biology Research 4: 172–173

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

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

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

Chou KC (2004b) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics, Advance Access published on August 12, 2004; doi: 10.1093/bioinformatics/bth466

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

Chou KC, Cai YD (2003c) 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 subcellular localization of proteins by hybridizing functional domain composition and pseudo-amino acid composition. J Cell Biochem 91: 1197–1203

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

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

Emanuelsson O (2002) Predicting protein subcellular localization from amino acid sequence information. Brief Bioinform 3: 361–376

Feng ZP (2001) Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. Biopolymers 58: 491–499

Gabrielian A, Bolshoy A (1999) Sequence complexity and DNA curvature. Comput Chem 23: 263–274

Gusev VD, Nemytikova LA, Chuzhanova NA (1999) On the complexity measures of genetic sequences. Bioinformatics 15: 994–999

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

Horton P, Nakai K (1997) Better prediction of protein cellular localization sites with the k nearest neighbor classifier. Proc Int Conf Intellig Syst Mol Biol 5: 147–152

Hua S, Sun Z (2001) Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17: 721–728

Huang Y, Li YD (2004) Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics 20: 21–28

Mardia KV, Kent JT, Bibby JM (1979) Multivariate Analysis: Chapter 11: Discriminant analysis; Chapter 12: Multivariate analysis of variance; Chapter 13: Cluster analysis. Academic Press, London, pp 322–381

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

Nakai K, Horton P (1999) PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization. Trends Biochem Sci 24: 34–36

NaKai K, Kanehisa M (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics 14: 897–911

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

Orlov YL, Filippov VP, Potapov VN, Kolchanov NA (2002) Construction of stochastic context trees of genetic texts. In Silico Biology 2: 233–247

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

Ponnuswamy PK, Prabhakaran K, Manavalan P (1980) Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. Biochim Biophys Acta 623: 301–316

Reinhardt A, Hubbard T (1998) Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 26: 2230–2236

Sadovsky MG (2003) The method to compare nucleotide sequences based on minimum entropy principle. Bull Math Biol 65: 309–322

Xiao X, Shao SH, Ding YS, Chen XJ (2004) Digital coding for amino acid based on cellular automata. 2004 IEEE Int. Conf. Systems, Man, and Cybernetics, Oct. 10–13, 2004, The Hague, The Netherlands (in press)

Yuan Z (1999) Prediction of protein subcellular locations using Markov chain models. FEBS Lett 14: 23–26

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 Engineering, Design, and Selection (in press)

Wang M, Yang J, Xu ZJ, Chou KC (2004b) SLLE for predicting membrane protein types. J Theor Biol (in press)

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

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

Authors' address: 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