The Protein Journal, Vol. 24, No. 6, August 2005 (© 2005)
DOI: 10.1007/s10930-005-7592-4

Using Fourier Spectrum Analysis and Pseudo Amino Acid Composition for Prediction of Membrane Protein Types

Hui Liu,[1] Jie Yang,[1] Meng Wang,[2] Li Xue,[1] and Kuo-Chen Chou[1,3,4]

Membrane proteins are generally classified into the following five types: (1) type I membrane protein, (2) type II membrane protein, (3) multipass transmembrane proteins, (4) lipid chain-anchored membrane proteins, and (5) GPI-anchored membrane proteins. Given the sequence of an uncharacterized membrane protein, how can we identify which one of the above five types it belongs to? This is important because the biological function of a membrane protein is closely correlated with its type. Particularly, with the explosion of protein sequences entering into databanks, it is in high demand to develop an automated method to address this problem. To realize this, the key is to catch the statistical characteristics for each of the five types. However, it is not easy because they are buried in a pile of long and complicated sequences. In this paper, based on the concept of the pseudo amino acid composition (Chou, K. C. (2001). *PROTEINS: Structure, Function, and Genetics* **43**: 246–255), the technique of Fourier spectrum analysis is introduced. By doing so, the sample of a protein is represented by a set of discrete components that can incorporate a considerable amount of the sequence order effects as well as its amino acid composition information. On the basis of such a statistical frame, the support vector machine (SVM) is introduced to perform predictions. High success rates were yielded by the self-consistency test, jackknife test, and independent dataset test, suggesting that the current approach holds a promising potential to become a high throughput tool for membrane protein type prediction as well as other related areas.

KEY WORDS: Discrete model; Fourier spectrum analysis; jackknife test; pseudo amino acid composition; support vector machines.

1. INTRODUCTION

Owing to the recent success of human genome project, the number of protein sequences entering into biology databases has been rapidly increasing, which challenges the speed and ability to identify uncharacterized proteins (Chou, 2004). Membrane proteins are an important part of proteins. Some automated methods have been proposed to discriminate non-membrane proteins from membrane proteins (see, e.g. Chou and Cai, 2005c); the latter may be further classified into the following five membrane protein types: (1) type I, (2) type II, (3) multipass transmembrane, (4) lipid chain-anchored, and (5) GPI-anchored (Fig. 1). Because the type of a membrane protein is closely correlated with its function (Alberts *et al.*, 1994; Chou and Elrod, 1999), it is also important to characterize the type for a given membrane protein. Actually, many efforts have been made in this regard (Cai *et al.*, 2001, 2002, 2003, 2004; Chou, 2000, 2001; Chou and Cai, 2005b; Chou and Elrod, 1999; Feng and Zhang, 2000; Wang *et al.*, 2004a, b). The present study was devoted to use a different novel approach – the Fourier spectrum analysis and pseudo amino acid

[1] Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, 200030, China.

[2] Microsoft Research Asia, 49 Zhichun Road, Beijing, 100080, China.

[3] Gordon Life Science Institute, San Diego, 92130, CA, USA.

[4] To whom correspondence should be addressed. E-mail: kchou@san.rr.com

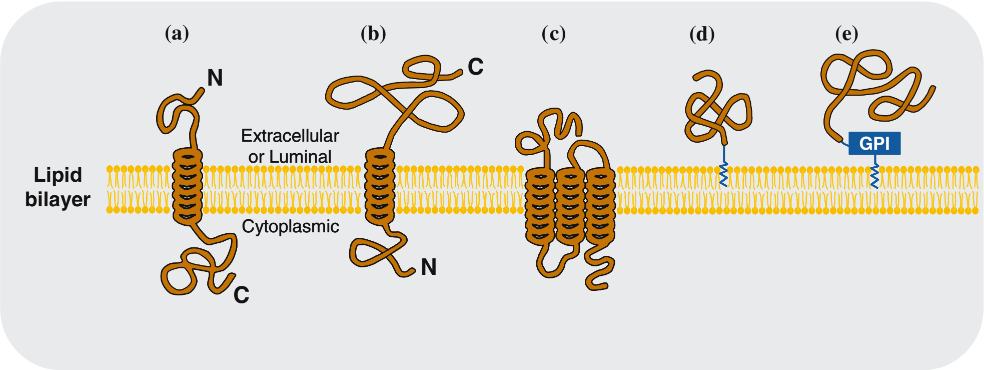

Fig. 1. Schematic drawing showing the following five types of membrane proteins: (a) type I transmembrane, (b) type II transmembrane, (c) multipass transmembrane, (d) lipid-chain anchored membrane, and (e) GPI-anchored membrane. As shown from the figure, although both type I and type II membrane proteins are of single-pass transmembrane, type I has a cytoplasmic C-terminus and an extracellular or luminal N-terminus for plasma membrane or organelle membrane, respectively, while the arrangement of N- and C-termini in type II membrane proteins is just reverse. No such distinction was drawn between the extracellular (or luminal) and cytoplasmic sides for the other three types in the current classification scheme. Reproduced from Chou (2001) with permission.

composition (Chou, 2001) – to predict the types of membrane proteins.

2. METHOD

The conventional amino acid composition is defined as 20 discrete numbers each representing the occurrence frequency of one of the 20 native amino acids (see, e.g., Chou, 1989b, 1995; Chou and Zhang, 1993, 1994; Nakashima *et al.*, 1986; Zhou, 1998). Obviously, if one uses the conventional amino acid composition to represent the sample of a protein, all its sequence order and length effects are lost. In order to include these effects, Chou (2001) introduced the concept of pseudo amino acid composition in a seminal study. In addition to the 20 components defined in the conventional amino acid composition, the pseudo acid composition contains more components. It is through the additional components that some sequence order and length effects can be incorporated. The introduction of pseudo acid composition has greatly stimulated the development of membrane protein type prediction and other related areas (Cai and Chou, 2003; Cai *et al.*, 2005; Chou, 2005; Chou and Cai, 2003, 2004; 2005a, b, c; Pan *et al.*, 2003; Shen and Chou, 2005; Wang *et al.*, 2004a, b; Xiao *et al.*, 2005a).Various approaches have been proposed to formulate these additional components in order to optimally reflect the sequence order effects (Chou, 2005; Chou and Cai, 2005b; Gao *et al.*, 2005; Pan *et al.*, 2003; Wang *et al.*, 2004b; Xiao *et al.*, 2005a, b, c). Here, we would like to propose a different approach to generate these components as formulated below.

Given a protein sequence with L amino acid residues

$$R_1R_2R_3R_4R_5R_6R_7R_8\ldots R_L, \tag{1}$$

suppose $h(R_1)$ is the hydrophilic value of the first residue R_1, $h(R_2)$ that of the second residue R_2, and so forth. The hydrophilic values for the 20 native amino acids were taken from (Hopp and Woods, 1981). Thus, from the sequence of Eq (1), we can generate $2L$ discrete Fourier spectrum numbers as given below:

$$\{F_1, F_2, \ldots, F_L, \varphi_1, \varphi_2, \ldots, \varphi_L\} \tag{3}$$

where the amplitude component F_k and phase component φ_k $(k=1,2,...,L)$ are defined by the following discrete Fourier spectrum transform formula

$$\sum_{l=1}^{L} h(R_i)\exp\left[-i\left(\frac{2\pi l}{L}\right)k\right] = F_k\exp(i\varphi_k), \quad (k = 1, 2, \ldots, L) \tag{4}$$

in which i represents the imaginary number.

The $2L$ Fourier spectrum numbers contain a substantial amount of information about the digit signal (Oppenheim *et al.*, 1985), and thereby can also be used to reflect characters of the sequence order of a protein. Furthermore, in the L phase components $\{\varphi_1,\varphi_2,...,\varphi_L\}$, the high-frequency components are more noisy and hence only the low-frequency components are more important. This is just like the case of protein internal motions where the low-frequency components imply more biological functions (Chou, 1988; 1989a). Accordingly, we only need to consider the first λ phase components and their corresponding amplitudes, i.e.,

$$\{F_1, F_2, \ldots, F_\lambda, \varphi_1, \varphi_2, \ldots, \varphi_\lambda\} \quad (\lambda < L) \tag{5}$$

After incorporating the above components into the classical 19D (dimensional) amino acid composition (Chou, 1995; Chou and Zhang, 1994), we obtain a pseudo amino acid composition with (19 + 2λ) components. In other words, the representation for a protein sample **P** is now formulated as

$$\mathbf{P} = \begin{bmatrix} p_1 \\ \vdots \\ p_{19} \\ p_{19+1} \\ \vdots \\ p_{19+\lambda} \\ p_{19+\lambda+1} \\ \vdots \\ p_{19+2\lambda} \end{bmatrix}, \tag{6}$$

where

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{j=1}^{19} f_j + w\sum_{j=1}^{\lambda} F_j + w\sum_{j=1}^{\lambda} \varphi_j}, & (1 \le u \le 19) \\ \dfrac{F_{u-19}}{\sum_{j=1}^{19} f_j + w\sum_{j=1}^{\lambda} F_j + w\sum_{j=1}^{\lambda} \varphi_j}, & (20 \le u \le 19 + \lambda) \\ \dfrac{w\varphi_{u-19-\lambda}}{\sum_{j=1}^{19} f_j + w\sum_{j=1}^{\lambda} F_j + w\sum_{j=1}^{\lambda} \varphi_j}, & (19 + \lambda + 1 \le u \le 19 + 2\lambda) \end{cases} \tag{7}$$

where f_j ($j = 1, 2, \ldots, 19$) are the normalized occurrence frequencies of the 19 amino acid components in the protein, w is the weight factor, and λ the threshold for the low-frequency passing filter. In this study, we took $w = 0.15$ and $\lambda = 30$.

The statistical prediction was operated by the SVM (support vector machine). For the mathematical principles, see (Cortes and Vapnik, 1995; Vapnik, 1998). The detailed procedures how to use SVM and the formulation of pseudo amino acid composition to predict protein subcellular location and membrane protein type can be found in (Chou and Cai, 2002) and (Cai *et al.*, 2003), respectively.

3. RESULTS AND DISCUSSION

The training dataset constructed by Chou and Elrod (1999) was used for the current method. It contains 2059 membrane protein sequences, of which 435 are type I membrane proteins, 152 type II membrane proteins, 1311 multi-pass transmembrane proteins, 51 lipid chain-anchored membrane proteins and 110 GPI-anchored membrane proteins.

During the prediction process, the regulation parameter *c* for SVM (cf. Eq. 12 of Chou and Cai, 2002) was set at 100. After being trained, the hyperplane was built in the feature space and thus the output could be obtained. The prediction quality was examined by three typical methods in statistical prediction (Chou and Zhang, 1995): the re-substitution test, the jackknife test, and the independent dataset test.

3.1. Re-substitution Test

The re-substitution test is to examine the self-consistency of a prediction method. In this test, the type of each membrane protein in the training dataset is in turn predicted by the SVM hyper-plane derived from the same dataset. The overall success rate thus obtained is 99.0% (Table 1), indicating that after being trained, the SVM model has grasped the complicated relationship between the pseudo-amino acid composition and the membrane protein types. However, during the resubstitution process, the rule parameters derived from the train-

Table 1. Overall Rates of Correct Prediction for the Five Membrane Protein Types by Different Algorithms and Test Methods.

Algorithm	Self-Consistency	Jackknife	Independent dataset
Least Hamming distance (Chou, 1989b)	$\frac{1293}{2059} = 62.8\%$	$\frac{1279}{2059} = 62.1\%$	$\frac{1751}{2625} = 66.7\%$
Least Euclidean distance (Nakashima *et al.*, 1986)	$\frac{1307}{2059} = 63.5\%$	$\frac{1293}{2059} = 62.8\%$	$\frac{1816}{2625} = 69.2\%$
ProtLock (Cedano *et al.*, 1997)	$\frac{1372}{2059} = 66.6\%$	$\frac{1348}{2059} = 65.5\%$	$\frac{1674}{2625} = 63.8\%$
Pseudo amino acid composition and Fourier transform approach	$\frac{2038}{2059} = 99.0\%$	$\frac{1613}{2059} = 78.3\%$	$\frac{2274}{2625} = 86.6\%$

ing dataset harbor the information of the query protein later plugged back for testing. This will certainly overestimate the success rate (Cai, 2001; Zhou, 1998; Zhou and Assa-Munt, 2001; Zhou and Doctor, 2003). Therefore, to more objectively determine the success rate, a cross-validation test should be conducted as described below.

3.2. Jackknife Test

The independent dataset test, sub-sampling test, and jackknife test are the three methods often used for cross-validation in statistical prediction. Of these three, the jackknife test is deemed as the most objective and rigorous one (Chou and Zhang, 1995). The overall jackknifing success rate is also given in Table 1

3.3. Independent Dataset Test

As a demonstration for practical application, prediction was also conducted for a set of independent proteins, none of which is included in the training dataset. The independent dataset was also taken from (Chou and Elrod, 1999), which contains 2625 protein sequences, of which 478 are type I membrane proteins, 180 type II membrane proteins, 1867 multi-pass transmembrane proteins, 14 lipid-chain anchored membrane proteins, and 86 GPI-anchored membrane proteins. The overall success rate is over 86% (Table 1).

Meanwhile, for facilitating comparison, the success rates by the other approaches are also listed in Table 1, from which we can see that the pseudo amino acid composition and Fourier transform approach remarkably outperform the least Hamming distance approach (Chou, 1989b), the least Euclidean distance approach (Nakashima *et al.*, 1986), and the ProtLoc predictor (Cedano *et al.*, 1997) in all the three tests.

4. CONCLUSIONS

Pseudo amino acid composition (Chou, 2001) is a very useful concept that allows us more effectively using the discrete model to statistically deal with the problems with many long and complicated sequences. The introduction of the Fourier spectrum analysis has further enriched the power of pseudo amino acid composition. The current approach can play a complementary role to the existing methods in this area.

REFERENCES

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J.D. (1994) *Molecular Biology of the Cell, Chap. 1.* 3rd edn. New York London: Garland Publishing.

Cai, Y. D. (2001) *Proteins: Struct, Funct, Genet* **43:** 336–338.

Cai, Y. D., and Chou, K. C. (2003). *Biochem. Biophys. Res. Commun.* **305:** 407–411.

Cai, Y. D., Liu, X. J., and Chou, K. C. (2001). *J. Biomol. Struct. Dyn.* **18:** 607–610.

Cai, Y. D., Liu, X. J., Xu, X. B., and Chou, K .C. (2002). *Internet Electron. J. Mol. Design* **1:** 219–226.

Cai, Y. D., Pong-Wong, R., Feng, K., Jen, J. C. H., and Chou, K. C. (2004). *J. Theor. Biol.* **226:** 373–376.

Cai, Y. D., Zhou, G. P., and Chou, K. C. (2003). *Biophys. J.* **84:** 3257–3263.

Cai, Y. D., Zhou, G. P., and Chou, K. C. (2005). *J. Theor. Biol.* **234:** 145–149.

Cedano, J., Aloy, P., P'erez-Pons, J. A., and Querol, E. (1997). *J. Mol. Biol* **266:** 594–600.

Chou, K. C. (1988) *Biophys. Chem.* **30:** 3–48.

Chou, K. C. (1989a) *Trends Biochem. Sci.* **14:** 212.

Chou, P. Y. (1989b). In: Fasman, G. D. (ed.), *Prediction of Protein Structure and the Principles of Protein Conformation* Plenum Press, New York, pp. 549–586.

Chou, K. C. (1995) *Proteins: Struct. Funct. Genet.* **21:** 319–344.

Chou, K. C. (2000) *Curr. Protein Peptide Sci.* **1:** 171–208.

Chou, K. C. (2001) *Proteins: Struct. Funct. Genet. (Erratum: ibid., 2001 Vol. 44, 60)* **43:** 246–255.

Chou, K. C. (2004) *Curr. Med. Chem.* **11:** 2105–2134.

Chou, K. C. (2005) *Bioinformatics* **21:** 10–19.

Chou, K. C., and Cai, Y. D. (2002). *J. Biol. Chem.* **277:** 45765–45769.

Chou, K. C., and Cai, Y. D. (2003). Predicting protein quaternary structure by pseudo amino acid composition. *Proteins: Struct. Funct. Genet.* **53:** 282–289.

Chou, K. C., and Cai, Y. D. (2004). *J. Cell. Biochem.* **91:** 1197–1203.

Chou, K. C., and Cai, Y. D. (2005a). *Bioinformatics* **21:** 944–950.

Chou, K. C., and Cai, Y. D. (2005b). *J. Chem. Inf. Model.* **45:** 407–413.

Chou, K. C., and Cai, Y. D. (2005c). *Biochem. Biophys. Res. Commun.* **327:** 845–847.

Chou, K. C., and Elrod, D. W. (1999). *Proteins: Struct. Funct. Genet.* **34:** 137–153.

Chou, J. J., and Zhang, C. T. (1993). *J. Theor. Biol* **161:** 251–262.

Chou, K. C., and Zhang, C. T. (1994). *J. Biol. Chem.* **269:** 22014–22020.

Chou, K. C., and Zhang, C. T. (1995). *Crit. Rev. Biochem. Mol. Biol.* **30:** 275–349.

Cortes, C., and Vapnik, V. (1995). *Mach. Learn.* **20:** 273–293.

Feng, Z. P., and Zhang, C. T. (2000). *J. Protein Chem.* **19:** 269–275.

Gao, Y., Shao, S. H., Xiao, X., Ding, Y. S., Huang, Y. S., Huang, Z. D., and Chou, K. C. (2005). Using pseudo amino acid composition to predict protein subcellular location: approached with lyapunov index, bessel function, and chebyshev filter. *Amino Acids, in press.*

Hopp, T. P., and Woods, K. R. (1981). *Proc. Natl. Acad. Sci. USA* **78:** 3824–3828.

Nakashima, H., Nishikawa, K., and Ooi, T. (1986). *J. Biochem* **99:** 152–162.

Oppenheim, A. V., Willsky, A. S., and Nawab, S. H. (1985) *Signals and Systems.* New York: Prentice Hall.

Pan, Y. X., Zhang, Z. Z., Guo, Z. M., Feng, G. Y., Huang, Z. D., and He, L. (2003). *J. Protein Chem.* **22:** 395–402.

Shen, H. B., and Chou, K. C. (2005). *Biochem. Biophys. Res. Commun.* **334**: 288–292.

Vapnik, V. (1998) *Statistical Learning Theory* New York: Wiley-Interscience.

Wang, M., Yang, J., Liu, G. P., Xu, Z. J., and Chou, K. C. (2004a). *Protein Eng. Des. Select.* **17:** 509–516.

Wang, M., Yang, J., Xu, Z. J., and Chou, K. C. (2004b). *J. Theor. Biol.* **232:** 7–15.

Xiao, X., Shao, S., Ding, Y., Huang, Z., Huang, Y., and Chou, K. C. (2005a). *Amino Acids* **28:** 57–61.

Xiao, X., Shao, S., Ding, Y., Huang, Z., Chen, X., and Chou, K. C. (2005b). *Amino Acids* **28:** 29–35.

Xiao, X., Shao, S. H., Ding, Y. S., Huang, Z. D., and Chou, K. C. (2005c). *Amino Acids*, DOI: 10.1007/s00726-005-0225-6.

Zhou, G. P. (1998) *J. Protein Chem.* **17:** 729–738.

Zhou, G. P., and Assa-Munt, N. (2001). Some insights into protein structural class prediction. *Proteins: Struct. Funct. Genet.* **44:** 57–59.

Zhou, G. P., and Doctor, K. (2003). *Proteins: Struct. Funct. Genet.* **50:** 44–48.