Available online at www.sciencedirect.com

SCIENCE DIRECT®

Biochemical and Biophysical Research Communications 334 (2005) 288–292

BBRC

www.elsevier.com/locate/ybbrc

Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types ☆

Hongbin Shen [a], Kuo-Chen Chou [a,b,*]

[a] Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai 200030, China
[b] Gordon Life Science Institute, San Diego, CA 92130, USA

Received 9 June 2005
Available online 28 June 2005

Abstract

Knowledge of membrane protein type often provides crucial hints toward determining the function of an uncharacterized membrane protein. With the avalanche of new protein sequences emerging during the post-genomic era, it is highly desirable to develop an automated method that can serve as a high throughput tool in identifying the types of newly found membrane proteins according to their primary sequences, so as to timely make the relevant annotations on them for the reference usage in both basic research and drug discovery. Based on the concept of pseudo-amino acid composition [K.C. Chou, Proteins: Struct. Funct. Genet. 43 (2001) 246–255; Erratum: Proteins: Struct. Funct. Genet. 44 (2001) 60] that has made it possible to incorporate a considerable amount of sequence-order effects by representing a protein sample in terms of a set of discrete numbers, a novel predictor, the so-called "optimized evidence-theoretic K-nearest neighbor" or "OET-KNN" classifier, was proposed. It was demonstrated via the self-consistency test, jackknife test, and independent dataset test that the new predictor, compared with many previous ones, yielded higher success rates in most cases. The new predictor can also be used to improve the prediction quality for, among many other protein attributes, structural class, subcellular localization, enzyme family class, and G-protein coupled receptor type. The OET-KNN classifier will be available as a web-server at www.pami.sjtu.edu.cn/kcchou.

Keywords: Evidence theory; KNN classifier; Pseudo-amino acid composition; Type-I membrane protein; Type-II membrane protein; Multipass transmembrane protein; Lipid-chain-anchored membrane protein; GPI-anchored membrane protein

A cell is highly organized with many functional units or organelles and is enclosed by the plasma membrane. Membrane proteins are crucial for many biological functions and have become attractive targets for drug discovery. Although the basic structure of biological membranes is provided by the lipid bilayer, most of the specific functions are performed by the membrane proteins [1,2]. The way that a membrane-bound protein is associated with the lipid bilayer usually reflects its function. For example, the transmembrane proteins can function on both sides of membrane and transport molecules from one side to the other; whereas the proteins that associated with one side of the lipid monolayer can only function on that side [1,2]. Membrane proteins can generally be classified into the following types [3]: (1) type-I membrane protein; (2) type-II membrane protein; (3) multipass transmembrane proteins; (4) lipid-chain-anchored membrane proteins; (5) GPI-anchored membrane proteins (Fig. 1). With the rapid development in molecular biology, the number of protein sequences has increased significantly, as reflected by the fact that

☆ *Abbreviations:* KNN, K-nearest neighbors; ET-KNN, evidence-theoretic KNN; OET-KNN, optimized evidence-theoretic KNN.

* Corresponding author.
E-mail address: kchou@san.rr.com (K.-C. Chou).

0006-291X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2005.06.087

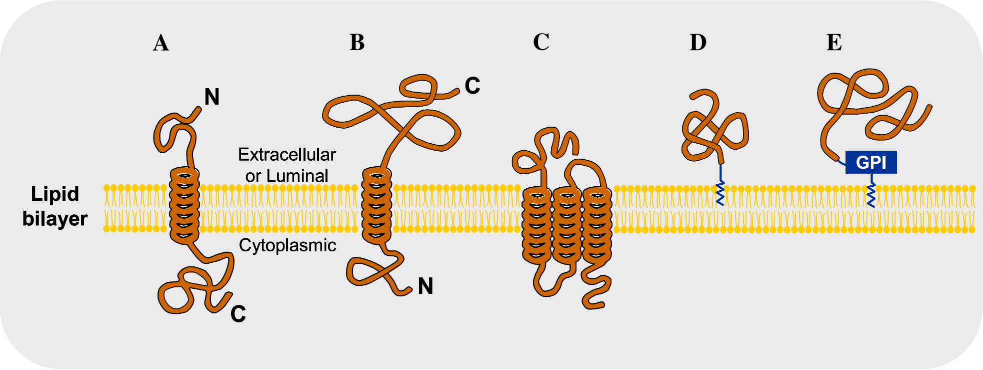

Fig. 1. Schematic drawing showing the following five types of membrane proteins: (A) type-I transmembrane, (B) type-II transmembrane, (C) multipass transmembrane, (D) lipid-chain-anchored membrane, and (E) GPI-anchored membrane. As shown in the figure, although both type-I and type-II membrane proteins 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 [11] with permission.

the total number of protein sequence entries in SWISS-PROT [4] of 2004 is 42 times that of 1986! It is both time-consuming and expensive to determine the types of newly found membrane proteins with traditional experiments. The unbalanced situation has called for finding solutions through bioinformatics. In a pioneering work, Chou and Elrod [3] introduced the covariant discriminant algorithm to predict the types of membrane proteins based on their amino acid composition. The covariant discriminant algorithm is actually a combination of the "Mahalanobis distance" [5–7] and the invariance principle for treating degenerative space [8] that is cited in the literature as "Chou's invariance theorem" (see, e.g. [9,10]). Subsequently, a series of prediction models were proposed in this area [11–22]. Most of the existing prediction methods fall into two categories: one is based on the conventional amino acid composition, and the other based on the "pseudo-amino acid composition." The concept of the "pseudo-amino acid composition" was originally proposed by Chou [11] for incorporating the protein sequence-order effect in terms of a set of discrete numbers so that many existing analytical prediction algorithms can be straightforwardly applied to deal with it.

On the basis of the pseudo-amino acid composition, here we introduce a different classifier, the so-called "optimized evidence-theoretic KNN (K-nearest neighbor) classifier" [23], for predicting the types of membrane proteins. The new classifier is very promising, and may become a powerful tool in bioinformatics and proteomics.

Pseudo-amino acid composition

First of all, let us briefly introduce the concept of "pseudo-amino acid composition" originally proposed by Chou [11]. According to the classical definition, the amino acid composition of a protein is defined by 20 discrete numbers with each representing the occurrence frequency of one of the 20 native amino acids in the protein. Thus, in terms of amino acid composition, a protein can be expressed by a point or a vector in a 20D (dimensional) space as defined by [7,8,24–26]

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{20} \end{bmatrix}, \tag{1}$$

where $x_1, x_2, \ldots, x_{20}$ are the composition components of 20 amino acids for the protein **X**. However, if using the 20D amino acid composition to represent a protein, all its sequence-order and sequence-length effects would be lost. In view of this, instead of the conventional amino acid composition, Chou [11] proposed to represent a protein sample by its pseudo-amino acid composition, which is defined in a $(20 + \lambda)$D space as given below:

$$\mathbf{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_{20} \\ x_{20+1} \\ \vdots \\ x_{20+\lambda} \end{bmatrix}, \tag{2}$$

where the first 20 components are the same as those in the conventional amino acid composition, and $x_{20+1}, \ldots, x_{20+\lambda}$ are the factors related to λ different ranks of sequence-order correlations that can be easily computed by Eqs. (2)–(6) of Chou [11]. For example, $\lambda = 60$ means taking the first 60 ranks of sequence-order correlations into consideration. Thus according to Eq. (2), a protein sample is represented by a $(20 + \lambda)\text{D} = 80\text{D}$ vector.

Algorithms

K-nearest neighbor classifier

Considering the problem of classifying N entities into M classes, which can be formulated as $\Omega = \{w_1, w_2, \ldots, w_M\}$, where w_i denotes the ith class. The available information is assumed to consist in a training dataset $T = \{(\mathbf{X}_1, c_1), \ldots, (\mathbf{X}_N, c_N)\}$ of N patterns $\mathbf{X}_i$ $(i = 1, 2, \ldots, N)$ and their corresponding class labels c_i $(i = 1, 2, \ldots, N)$ taking values in Ω. K-nearest neighbor (KNN) rule [27] is well known in the pattern recognition literature. According to this rule, an unclassified pattern $\mathbf{X}$ is assigned to the class represented by a majority of its K-nearest neighbors in $\mathbf{T}$. This rule is usually called "voting KNN rule." KNN is popular in pattern recognition community due mainly to its good performance and its simple-to-use feature. Since the inception of KNN some modified versions have been proposed in order to improve its performance.

Dempster–Shafer theory

In the Dempster–Shafer (D–S) theory, a problem is represented by a set Θ of mutually exclusive and exhaustive hypotheses called the frame of discernment [28]. Denoting all the subsets of Θ as 2^{Θ}, then we can define a basic belief assignment (BBA) mapping function m from 2^{Θ} to $[0,1]$, which can be formulated as: $m{:}2^{\Theta} \to [0,1]$. BBA function m satisfies $m(\phi) = 0$ and $\sum_{A\subset\Theta} m(A) = 1$, where ϕ denotes the empty set. The quantity $m(A)$ represents the belief that one is willing to commit exactly to A, given the available evidence. If $m(A) > 0$, A is called the focal element of m.

Associated with BBA m are a belief or credibility function Bel and a plausibility function Pl, defined, respectively as: $Bel(A) = \sum_{B\subset A} m(B)$ and $Pl(A) = \sum_{A\cap B\neq\phi} m(B)$. The quantity $Bel(A)$ can be interpreted as a global measure of one's belief that hypothesis A is true, while $Pl(A)$ may be viewed as the amount of belief that could potentially be placed in A. For further information, please refer to [28].

Two BBAs m^1 and m^2 on Θ, induced by two independent items of evidence, can be combined by the so-called Dempster's rule of combination to yield a new BBA: $m = m^1 \oplus m^2$, where $\oplus$ is called the orthogonal sum of m^1 and m^2, and defined as:

$$m(A) = \frac{\sum_{A_1\cap A_2 = A} m^1(A_1) m^2(A_2)}{\sum_{A_1\cap A_2\neq\phi} m^1(A_1) m^2(A_2)}. \tag{3}$$

The orthogonal sum of $\oplus$ is commutative and associative.

Evidence-theoretic KNN rule

The evidence-theoretic K-nearest neighbor (ET-KNN) rule is a pattern classification method based on the Dempster–Shafer theory of belief functions [29]. In this approach, each neighbor of a pattern to be classified is considered as an item of evidence supporting certain hypotheses concerning the class membership of that pattern. Based on this evidence, basic belief masses (BBA) are assigned to each subset of the set of classes. Such masses are obtained for each of the K-nearest neighbors of the pattern under consideration and aggregated using the Dempster's rule of combination [28]. A decision is made by assigning a pattern to the class with the maximum credibility.

Let $\Omega = \{w_1, w_2, \ldots, w_M\}$ denote the M classes. Considering $\mathbf{X}$ to be classified and Ψ_k is the set of its K-nearest neighbors in a training set T. For any $\mathbf{X}^i \in \Psi_k$, the knowledge that $\mathbf{X}^i$ belongs to class $w_q (w_q \in \Omega)$ can be considered as a piece of evidence that increases our belief that $\mathbf{X}$ also belongs to w_q. This item of evidence can be represented by a BBA m^i assigning a fraction α_q^i of the unit mass to the singleton $\{w_q\}$ and the rest to Ω:

$$m^i(\{w_q\}) = \alpha_q^i, \tag{4}$$

$$m^i(\Omega) = 1 - \alpha_q^i, \tag{5}$$

with $m^i(A) = 0$ for all $A \subseteq \Omega$ and $A \notin \{\Omega, \{w_q\}\}$ [29]. The mass α_q^i is chosen as a decreasing function of the Euclidean distance d^i between $\mathbf{X}$ and $\mathbf{X}^i$

$$\alpha_q^i = \alpha_0 \exp(-\gamma_q^2 (d^i)^2), \tag{6}$$

where γ_q is a parameter associated to class w_q and α_0 is a fixed parameter. The BBAs $m^1, m^2, \ldots, m^k$ corresponding to the K-nearest neighbors of $\mathbf{X}$ can then be combined using Dempster's rule (see Eq. (3)): $m = m^1 \oplus m^2 \oplus \ \ldots \ \oplus m^k$. Hence, we can obtain the credibility and plausibility of each class w_q $(q = 1, 2, \ldots, M)$: $Bel(\{w_q\}) = m(\{w_q\})$ and $Pl(\{w_q\}) = m(\{w_q\}) + m(\Omega)$.

A decision is made by assigning a query sample $\mathbf{X}$ to the class $w_{q\max}$ with maximum credibility (or equivalently, maximum plausibility); i.e.,

$$\mathbf{X} \Rightarrow \arg\max_{q} m(\{w_q\}). \tag{7}$$

Optimized evidence-theoretic KNN rule

In the above description of ET-KNN rule, how to select the parameters α_0 and $R = (\gamma_1, \ldots, \gamma_M)^t$ is not addressed. Whereas the value of α_0 proves in practice not to be too critical, the tuning of $R = (\gamma_1, \ldots, \gamma_M)^t$ was found experimentally to have significant influence on classification accuracy. In 1998 an optimization procedure to determine the optimal or near-optimal parameter values was proposed from the data by minimizing an error function [23]. It was observed that the Optimized evidence-theoretic KNN (OET-KNN) rule obtained through such an optimization treatment would lead to a substantial improvement in classification accuracy.

Table 1
Overall rates of correct prediction for the five membrane protein types by different algorithms and test methods

Algorithm	Sample representation	Test method (%)		
		Self-consistency[a]	Jackknife[a]	Independent dataset[b]
Least Hamming distance [25]	Amino acid composition	1293/2059=62.8	1279/2059=62.1	1751/2625=66.7
Least Euclidean distance [26]	Amino acid composition	1307/2059=63.5	1293/2059=62.8	1816/2625=69.2
ProtLock [37]	Amino acid composition	1372/2059=66.6	1348/2059=65.5	1674/2625=63.8
Covariant-discriminant [3]	Amino acid composition	1670/2059=81.1	1573/2059=76.4	2085/2625=79.4
OET-KNN ($K = 4$)	Pseudo-amino acid composition[c] [11]	2049/2059=99.5	1743/2059=84.7	2472/2625=94.2

[a] Conducted for the 2059 membrane proteins classified into five different types as described in the text.

[b] Conducted based on the rule parameters derived from the 2059 membrane proteins for the 2625 independent membrane proteins.

[c] See Eq. (2) with $\lambda = 60$.

Results and discussion

The same training dataset originally constructed by Chou and Elrod [3] was used for the current study. It contains 2059 membrane protein sequences, of which 435 are type-I transmembrane proteins, 152 are type-II transmembrane proteins, 1311 are multipass transmembrane proteins, 51 are lipid-chain-anchored transmembrane proteins, and 110 are GPI-anchored transmembrane proteins.

The demonstration of the optimized evidence-theoretic KNN was conducted by three most typical approaches in statistical prediction: re-substitution, jackknife test, and independent dataset test [30]. The independent dataset was also taken from Chou and Elrod [3] that contains 2625 proteins, of which 478 are type-I transmembrane proteins, 180 are type-II transmembrane proteins, 1867 are multipass transmembrane proteins, 14 are lipid-chain-anchored transmembrane proteins, and 86 are GPI-anchored transmembrane proteins. For OET-KNN classifier, there are many options for the distance metric, such as Euclidean distance metric, Hamming distance metric, etc. In this paper, we adopted the metric of Euclidean distance for calculation. The results thus obtained are given in Table 1. For facilitating comparison, the results by other approaches using different sample representations are also listed in the same table.

As we can see from Table 1, the success rates by the current OET-KNN classifier are higher than those by the previous approaches, including the case obtained by the jackknife test which is deemed the most rigorous and objective cross validation procedure in comparison with the other test procedures [9,30–32].

Conclusions

The optimized evidence-theoretic K-nearest neighbor classifier, or OET-KNN classifier, is a very powerful predictor in identifying the type of an uncharacterized membrane protein. It is anticipated that the power of predicting membrane protein types will be further enhanced if the OET-KNN classifier can be effectively complemented with other existing powerful algorithms, such as the covariant discriminant algorithms [3,33], the pseudo-amino acid composition approach [10–12], the functional domain composition approach [14], the weighted-support vector machine approach [15], the supervise locally linear embedding (SLLE) approach [16], cellular automata approach [18], and the GO-PseAA approach [13]. Also, it has not escaped our notice that the OET-KNN classifier as introduced here may have a positive impact in improving the prediction quality for many other protein attributes [34], such as protein structural class [7,8,25,26,32,35,36], protein subcellular localization [9,10,17,37–43], enzyme family and subfamily class [44–46], G-protein coupled receptor type [47,48], and protein quaternary structure types [49], among many others.

Acknowledgments

Hongbin Shen wishes to thank Jie Yang and Xiaojun Liu for their encouragement and discussions.

References

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J.D. Watson, Molecular Biology of the Cell, Garland Publishing, New York, London, 1994, Chapter 1.

[2] H. Lodish, D. Baltimore, A. Berk, S.L. Zipursky, P. Matsudaira, J. Darnell, Molecular Cell Biology, Scientific American Books, New York, 1995, Chapter 3.

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

[4] A. Bairoch, R. Apweiler, Nucleic Acids Res. 25 (2000) 31–36.

[5] P.C. Mahalanobis, Proc. Natl. Inst. Sci. India 2 (1936) 49–55.

[6] K.C.S. Pillai, in: S. Kotz, N.L. Johnson (Eds.), Encyclopedia of Statistical Sciences, Wiley, New York, 1985, pp. 176–181.

[7] K.C. Chou, C.T. Zhang, J. Biol. Chem. 269 (1994) 22014–22020.

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

[9] G.P. Zhou, K. Doctor, Proteins: Struct. Funct. Genet. 50 (2003) 44–48.

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

[11] K.C. Chou, Proteins: Struct. Funct. Genet. 43 (2001) 246–255, Erratum: Proteins: Struct. Funct. Genet. 44 (2001) 60.

[12] K.C. Chou, Y.D. Cai, J. Chem. Inf. Model. 45 (2005) 407–413.

[13] K.C. Chou, Y.D. Cai, Biochem. Biophys. Res. Commun. 327 (2005) 845–847.

[14] Y.D. Cai, G.P. Zhou, K.C. Chou, Biophys. J. 84 (2003) 3257–3263.

[15] M. Wang, J. Yang, G.P. Liu, Z.J. Xu, K.C. Chou, Protein Eng. Design Selection 17 (2004) 509–516.

[16] M. Wang, J. Yang, Z.J. Xu, K.C. Chou, J. Theor. Biol. 232 (2005) 7–15.

[17] X. Xiao, S. Shao, Y. Ding, Z. Huang, Y. Huang, K.C. Chou, Amino Acids 28 (2005) 57–61.

[18] X. Xiao, S. Shao, Y. Ding, Z. Huang, X. Chen, K.C. Chou, Amino Acids 28 (2005) 29–35.

[19] Z.P. Feng, C.T. Zhang, J. Protein Chem. 19 (2000) 269–275.

[20] Z.P. Feng, Biopolymers 58 (2001) 491–499.

[21] Y.D. Cai, X.J. Liu, K.C. Chou, J. Biomol. Struct. Dyn. 18 (2001) 607–610.

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

[23] L.M. Zouhal, T. Denoeux, IEEE Trans. Syst. Man Cybernet. 28 (1998) 263–271.

[24] J.J. Chou, C.T. Zhang, J. Theor. Biol. 161 (1993) 251–262.

[25] P.Y. Chou, in: G.D. Fasman (Ed.), Prediction of Protein Structure and the Principles of Protein Conformation, Plenum Press, New York, 1989, pp. 549–586.

[26] H. Nakashima, K. Nishikawa, T. Ooi, J. Biochem. 99 (1986) 152–162.

[27] T.M. Cover, P.E. Hart, IEEE Trans. Inf. Theory IT-13 (1967) 21–27.

[28] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, Princeton, NJ, 1976.

[29] T. Denoeux, IEEE Trans. Syst. Man Cybernet. 25 (1995) 804–813.

[30] K.C. Chou, C.T. Zhang, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.

[31] G.P. Zhou, N. Assa-Munt, Proteins: Struct. Funct. Genet. 44 (2001) 57–59.

[32] G.P. Zhou, J. Protein Chem. 17 (1998) 729–738.

[33] K.C. Chou, Curr. Protein Peptide Sci. 1 (2000) 171–208.

[34] K.C. Chou, in: P.W. Weinrer, Q. Lu (Eds.), Gene Cloning & Expression Technologies, Eaton Publishing, Westborough, MA, 2002, pp. 57–70, Chapter 4.

[35] K.C. Chou, Y.D. Cai, Biochem. Biophys. Res. Commun. 321 (2004) 1007–1009, Corrigendum: Biochem. Biophys. Res. Commun. 329 (2005) 1362.

[36] K.C. Chou, G.M. Maggiora, Protein Eng. 11 (1998) 523–538.

[37] J. Cedano, P. Aloy, J.A. P'erez-Pons, E. Querol, J. Mol. Biol. 266 (1997) 594–600.

[38] K. Nakai, P. Horton, Trends Biochem. Sci. 24 (1999) 34–36.

[39] K. Nakai, Adv. Protein Chem. 54 (2000) 277–344.

[40] K.C. Chou, D.W. Elrod, Protein Eng. 12 (1999) 107–118.

[41] K.C. Chou, Y.D. Cai, J. Cell. Biochem. 90 (2003) 1250–1260. Addendum: J. Cell. Biochem. 1291 (1255) (2004) 1085.

[42] K.C. Chou, Y.D. Cai, Bioinformatics 21 (2005) 944–950.

[43] K.C. Chou, Y.D. Cai, J. Biol. Chem. 277 (2002) 45765–45769.

[44] K.C. Chou, Y.D. Cai, Protein Sci. 13 (2004) 2857–2863.

[45] K.C. Chou, Bioinformatics 21 (2005) 10–19.

[46] K.C. Chou, D.W. Elrod, J. Proteome Res. 2 (2003) 183–190.

[47] D.W. Elrod, K.C. Chou, Protein Eng. 15 (2002) 713–715.

[48] K.C. Chou, D.W. Elrod, J. Proteome Res. 1 (2002) 429–433.

[49] K.C. Chou, Y.D. Cai, Proteins: Struct. Funct. Genet. 53 (2003) 282–289.