J Membrane Biol (2015) 248:179–186
DOI 10.1007/s00232-014-9755-8

A New Multi-label Classifier in Identifying the Functional Types of Human Membrane Proteins

Hong-Liang Zou · Xuan Xiao

Received: 17 September 2014 / Accepted: 11 November 2014 / Published online: 30 November 2014

Abstract Membrane proteins were found to be involved in various cellular processes performing various important functions, which are mainly associated to their type. Given a membrane protein sequence, how can we identify its type(s)? Particularly, how can we deal with the multi-type problem since one membrane protein may simultaneously belong to two or more different types? To address these problems, which are obviously very important to both basic research and drug development, a new multi-label classifier was developed based on pseudo amino acid composition with multi-label k-nearest neighbor algorithm. The success rate achieved by the new predictor on the benchmark dataset by jackknife test is 73.94 %, indicating that the method is promising and the predictor may become a very useful high-throughput tool, or at least play a complementary role to the existing predictors in identifying functional types of membrane proteins.

Keywords Membrane proteins · Pseudo amino acid composition · ML-KNN algorithm · Jackknife test

Electronic supplementary material The online version of this article (doi:10.1007/s00232-014-9755-8) contains supplementary material, which is available to authorized users.

H.-L. Zou (✉) · X. Xiao
Computer Department, Jing-De-Zhen Ceramic Institute,
Jing-De-Zhen 333046, China
e-mail: hongliangzou@126.com

X. Xiao
e-mail: xxiao@gordonlifescience.org

Introduction

Cell membrane is a key component of the cell and membrane protein is the most important part of cell membrane. Membrane protein plays a key role in physiology and biology, such as a carrier to transport materials into or out of the cells and as receptors of some hormone or chemical substance, and membrane protein is a kind of important drug target (Ding et al. 2012). Therefore, accurately and rapidly identifying the functional types of membrane proteins will be helpful for disease treatment and drug design, because the knowledge about the type of a query membrane protein has a close relationship with its functions.

According to some previous studies (Chou and Shen 2007; Huang and Yuan 2013), membrane proteins are mainly divided into the following eight functional types: single-pass type I, single-pass type II, single-pass type III, single-pass type IV, multipass, lipid-anchor, GPI-anchor, and peripheral membrane proteins.

Although the functional type of a membrane protein may be determined by carrying out various biochemical experiments, these approaches by purely doing experiments are both time consuming and expensive. In the post-genomic age, the gap between the newly found membrane protein sequences and the information of their types is becoming increasingly wide (Wang and Li 2012). Therefore, to bridge such a gap, it is urgent to develop an effective and rapid computational method to identify the functional types of membrane proteins.

In the past several years, many efforts have been made in identifying the functional types of membrane proteins, such as Chou and Elrod (1999) predicted the functional types of membrane proteins based on the covariant discriminant algorithm (CDA) and amino acid composition (AAC); Wang et al. (2005), by using supervised locally linear embedding (SLLE) technique and pseudo amino acid composition (PseAAC) with k-nearest neighbor (KNN) algorithm to identify membrane proteins' types, achieved a success rate of 82.3 % by jackknife test; Shen et al. (2006) predicting membrane protein types by hybridizing pseudo amino acid composition with fuzzy k-nearest neighbor (FKNN) algorithm, achieved a success rate of 85.6 % by jackknife test and 95.7 % by independent dataset test; and Pu et al. (2007) predicting membrane proteins types based on sequence information and evolution information, obtained 92.3 % success rate, and many others.

Although the aforementioned methods have their own advantages and did play a key role in stimulating the development of this field, they were only focused on identifying which type it belongs to for a query membrane protein (Xiao et al. 2013). In fact, there are many membrane proteins that have more than one function or functional type (Xiao et al. 2013). Those proteins should not escape our eyes because they may have some unique biological functions worthy of our special notice (Glory and Murphy 2007; Smith 2008).

In the paper, a new method by hybridizing various pseudo amino acid compositions was proposed to identify the functional types of human membrane proteins. A multi-label classifier called multi-label *k*-nearest neighbor (ML-KNN) was introduced, which is derived from classical KNN algorithm. Finally, a promising result was obtained, which indicated that the method is useful, and it may be used in identifying other attributes of proteins.

According to a recent comprehensive review (Chou 2011), to establish a powerful and efficient predictor for a protein system, the following procedures should be considered: (1) establish or select a valid benchmark dataset to train and test the predictor; (2) formulate the protein sequences using an effective mathematic that can truly reflect the intrinsic correlation with the target to be predicted; (3) develop or introduce a powerful algorithm (or engine) to operate the prediction; and (4) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor. We describe the processes in detail.

Materials and Methods

Benchmark Datasets

All of the membrane protein sequences used in the current study were collected from the UniProtKB database released on 04-Apr 16, 2014 at website http://www.uniprot.org/. In order to obtain a high quality and well-defined dataset, the following procedures should be considered: (1) only human membrane protein sequences were collected; (2) sequences annotated with "fragment" were removed; (3) sequences with less than 50 amino acid residues were also removed to avoid the influence of fragment; (4) to reduce the redundancy and homology bias, the program named CD-HIT was used to remove those proteins that have more than 60 % (not 25 %, because some types data too little) pairwise sequence identity to any other protein in the same subset.

Finally, we obtained 3,166 different human membrane protein sequences covered in eight different functional types, which can be formulated as

$$S = S_1 \cup S_2 \cup S_3 \cup S_4 \cup S_5 \cup S_6 \cup S_7 \cup S_8, \qquad (1)$$

where S_1 represents the functional type of "single-pass type I", S_2 for "single-pass type II", and so forth. The symbol $\cup$ represents the "union" in the set theory. For convenience, the numbers from 1 to 8 were used to represent the 8 subsets. A detailed information about the benchmark dataset are listed in Table 1.

Because some membrane proteins may simultaneously belong to two or more functional types, it is instructive to introduce the concept of "virtual protein" (Xiao et al. 2013; Chou et al. 2011, 2012) as briefed below. If a protein possesses two different functional types, it will be counted as two virtual proteins; if it possesses three functional attributes, it will be counted as three virtual proteins, and so forth. Thus, the number of total virtual proteins can be formulated as (Xiao et al. 2013; Lin et al. 2013)

$$N(\mathrm{vir}) = \mathrm{N(seq)} + \sum_{\mathrm{m}=1}^{\mathrm{M}} (\mathrm{m} - 1)\mathrm{N(m)} \qquad (2)$$

where $N(\mathrm{vir})$ is the number of total virtual proteins, $N(\mathrm{seq})$ is the number of total different protein sequences investigated, $N(1)$ is the number of membrane proteins with one functional type, $N(2)$ is the number of membrane proteins

Table 1 Detail of benchmark dataset of human membrane proteins

Order	Functional type	Number of proteins
1	Single-pass type I	605
2	Single-pass type II	195
3	Single-pass type III	25
4	Single-pass type IV	27
5	Multipass	1,444
6	Lipid-anchor	251
7	GPI-anchor	83
8	Peripheral	637
The number of virtual membrane proteins		3,267
The number of different membrane proteins		3,166

Of the 3,166 different membrane proteins, 3,069 belong to one functional type, 93 to two types and 4 to three types, no four or more types with two different functional types, and so forth, while M is the number of total functional types of membrane proteins investigated.

According to Eq. (2), the virtual membrane proteins investigated in the current study can be calculated by the following formulation:

$$\begin{aligned} N(\text{ vir}) &= N(\text{seq}) + (1-1) \times 3069 + (2-1) \\ &\quad \times 93 + (3-1) \times 4 \\ &= 3166 + 0 + 93 + 8 = 3267. \end{aligned} \tag{3}$$

As we can see from Eqs. (2, 3), the number of total virtual membrane proteins is generally greater than the number of different membrane proteins. When and only when all of the membrane proteins belong to one functional type, the two (i.e., virtual proteins and different proteins) will be the same (Lin et al. 2013).

Feature Extraction

In order to develop an effective predictor for identifying the functional types of human membrane proteins based on the sequence information, one of the most important things is to formulate the protein sequence with an efficient mathematical expression that can truly reflect the intrinsic correlation with the target to be predicted (Xiao et al. 2013; Chou 2011). However, it is not an easy work to realize this because this kind of correlation is usually deeply hidden or "buried" into piles of complicated sequences (Xiao et al. 2013).

As is well known, the most straightforward method is to formulate the protein sequence using its entire amino acid composition. For a protein sequence P with L amino acids, it can be expressed as

$$P = R_1 R_2 R_3 \cdots R_L, \tag{4}$$

where R_1 is the first residue of the sequence, R_2 is the second residue, and so forth. Each of the residues belongs to the 20 native amino acids. In order to identify its attribute(s), the sequence similarity search-based tools, such as BLAST (Zhang et al. 1997; Wootton and Federhen 1993), were utilized to search the protein database for those proteins that have high sequence similarity to the query protein P. Then, the attribute(s) of the proteins thus found were used to deduce the attribute(s) for the query P. However, this kind of straightforward sequential model, although quite intuitive and has the ability to contain the entire sequence information, failed to work when the query protein P did not have significant sequence similarity to any attribute-known proteins.

Thus, to overcome the above difficulty, various discrete models were proposed in a hope to enhance the power of the predictor.

Amino Acid Composition (AAC)

Among the various discrete models, the simplest one is the AAC-discrete model that represents the protein sample using its AAC (Nakashima et al. 1986). According to the AAC-discrete model, the protein P can be formulated as (Chou 1995)

$$P = [f_1 f_2 f_3 \cdots f_{20}]^{\mathrm{T}}, \tag{5}$$

where $f_i (i = 1, 2, \ldots, 20)$ represents the normalized occurrence frequencies of the 20 native amino acids in the protein and T stands for the transposing operator. However, as we can see from Eq. (5), if only the AAC-discrete model was used to represent the protein sequence P, all of the sequence-order information would be lost.

In order to avoid completely losing the sequence-order information, a new model was proposed to replace the simple amino acid composition that is pseudo amino acid composition (PseAAC) (Chou 2001).

Since the concept of PseAAC was proposed by Chou in 2001, it has been widely recognized and used by many investigators to identify various attributes of proteins, such as identifying subcellular location of proteins (Xiao et al. 2005; Shen and Chou 2007; Li and Li 2008; Park and Kanehisa 2003), predicting subcellular location of apoptosis proteins (Chen and Li 2007; Jian et al. 2008; Lin et al. 2009), predicting enzyme classes or subclasses (Zhou et al. 2007; Chou and Elrod 2003), identifying the functional types of antimicrobial peptides (Xiao et al. 2013; Khosravian and Kazemi 2013), and among many others.

Chou's Pseudo Amino Acid Composition (CPseAAC)

According to the concept of Chou's pseudo amino acid composition, a protein sequence can be represented by a $20 + \lambda$ dimension vector. The first 20 elements represent the amino acid composition, and the latter λ elements represent the sequence-order information. The sequence-order information can be indirectly represented by the following equation:

$$\begin{aligned} \delta_\theta &= \frac{1}{L-\theta} \sum_{i=1}^{L-\theta} \Omega(R_i, R_{i+\theta}), \\ &(\theta = 1, 2, \ldots, \lambda \text{ and } \lambda < L\ ), \end{aligned} \tag{6}$$

where L denotes the length of the protein sequence and δ_θ is called the θth correlation factor which harbors the sequence-order information between all the θ most contiguous residues. The correlation function $\Omega(R_i, R_{i+\theta})$ is defined by

$$\begin{aligned} \Omega(R_i, R_{i+\theta}) = \frac{1}{3}\Big\{ &[F(R_{i+\theta}) - F(R_i)]^2 + [G(R_{i+\theta}) \\ &- G(R_i)]^2 \quad + [H(R_{i+\theta}) - H(R_i)]^2 \Big\}, \end{aligned} \tag{7}$$

Table 2 Details of the physiochemical descriptor

Physiochemical property	Class one	Class two	Class three
Secondary structure	E,A,L,M,Q,K,R,H	V,I,Y,C,W,F,T	G,N,P,S,D
Solvent accessibility	A,L,F,C,G,I,V,W	P,K,Q,E,N,D	M,R,S,T,H,Y
Normalized van der Walls volume	G,A,S,T,P,D,C	N,V,E,Q,I,L	M,H,K,F,R,Y,W
Hydrophobicity	R,K,E,D,Q,N	G,A,S,T,P,H,Y	C,L,V,I,M,F,W
Charge	K,R	A,N,C,Q,G,H,I,L,M,F, P,S,T,W,Y,V	D,E
Polarizability	G,A,S,D,T	C,P,N,V,E,Q,I,L	K,M,H,F,R,Y,W
Polarity	L,I,F,W,C,M,V,Y	P,A,T,G,S	H,Q,R,K,N,E,D
Surface accessibility	G,Q,D,N,A,H,R	K,T,S,E,C	I,L,M,F,P,W,Y,V

where $F(R_i)$, $G(R_i)$ and $H(R_i)$ are the evaluated values of hydrophobicity, hydrophilicity, and mass, respectively. There are also three types of values that can be used. Before we use these values, a standard conversion should be conducted using Eq. (4) of Huang and Yuan (2013). The numerical values of the three physical–chemical properties for each of the 20 native amino acids can be obtained from (Shen and Chou 2008).

Then a sample protein P can be represented as

$$P = [x_1, x_2, \ldots, x_{20}, x_{20+1}, \ldots, x_{20+\lambda}]^T \quad \lambda < L \tag{8}$$

where

$$x_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w \sum_{\theta=1}^{\lambda} \delta_\theta}, & (1 \le u \le 20) \\ \dfrac{w\delta_{u-20}}{\sum_{i=1}^{20} f_i + w \sum_{\theta=1}^{\lambda} \delta_\theta}, & (20 + 1 \le u \le 20 + \lambda;\ \lambda < L) \end{cases} \tag{9}$$

where w is the weight factor, $f_i (i = 1, 2, \ldots, 20)$ represents the normalized occurrence frequencies of the 20 native amino acids in the sample protein P, and δ_θ is the θ-tier sequence-correlation factor, computed according to Eq. (6). In this study, we choose $w = 0.05$, $\lambda = 20$ after careful consideration of easy handling; they can be assigned other values, of course, but the impact on the result would be small.

Information Entropy (IE)

Shannon proposed that any information is redundant, and redundant size is related with the occurrence probability or uncertainty of each symbol such as numbers, letters, or words among the information. The information entropy for a system can be defined as

$$H = -\sum_{i=1}^{20} f(i) \log_2 f(i) \tag{10}$$

where $f(i) (i = 1, 2, \ldots, 20)$ represents the occurrence probability of amino acid i. The information entropy H is a measured value of the amount of information. For example, for the digital sequence $P = 100100011010010$, the information entropy H is obtained as given below:

$$\begin{cases} P(0) = 9/15 = 0.6 \\ P(1) = 6/15 = 0.4 \\ H = -(0.6 \times \log_2 0.6 + 0.4 \times \log_2 0.4) = 0.971. \end{cases} \tag{11}$$

Distribution (D)

According to Zou et al. (2013), based on the different physiochemical properties, the 20 native amino acids can be divided into 3 groups. In this study, the following eight different physiochemical properties were utilized: secondary structure, solvent accessibility, normalized van der Waals volume, hydrophobicity, charge, polarizability, polarity, and surface tension (Zou et al. 2013) (listed in Table 2). The descriptor called distribution was utilized to describe the global composition of each of those properties. In this study, five distributions were assigned—position percentage of first, 25, 50, 75, and 100 % residue occurrence in the entire sequence. Therefore, the distribution D_x for the descriptor E_i is calculated as below (Saravanan and Lakshmi 2013):

$$E_i 1D_x = \frac{P_1}{L}, \tag{12}$$

$$E_i 25D_x = \frac{P_{25}}{L}, \tag{13}$$

$$E_i 50D_x = \frac{P_{50}}{L}, \tag{14}$$

$$E_i75D_x = \frac{P_{75}}{L}, \tag{15}$$

$$E_i100D_x = \frac{P_{100}}{L}(i = 1,2,\ldots,8;\ x = 1,2,3), \tag{16}$$

where P_1, P_{25}, P_{50}, P_{75}, P_{100} indicate the position of first occurrence of x, and positions of 25, 50, 75, 100 % occurrence of x, respectively.

We give an example to explain in detail the distribution in the following. Assuming that there is a protein sequence, its amino acid composition is AEAAAEAEEAAAAAE AEEEAAEEAEEEAAE, which has 16 alanines and 14 glutamic acids. The first, 25, 50, 75, and 100 % of A are located in the first, 5th, 12th, 20th, and 29th residue. The D descriptor for A is $1/30 = 0.0333$, $5/30 = 0.1667$, $12/30 = 0.4000$, $20/30 = 0.6667$, $29/30 = 0.9667$. Similarly, the D descriptor for E is 0.0667, 0.2667, 0.6000, 0.7667, 1.0000. Overall, the D descriptor for this sequence is $D = (0.0333, 0.1667, 0.4000, 0.6667, 0.9667, 0.0667, 0.2667, 0.6000, 0.7667, 1.0000)$.

Position-Specific Scoring Matrix (PSSM)

The position-specific scoring matrix (PSSM) is often used to describe the sequence evolution information of protein. A protein sequence P with L amino acid residues can be formulated by an $L \times 20$ matrix, it can be expressed as follows:

$$P^{(0)}_{\text{PSSM}} = \begin{bmatrix} n^{(0)}_{1,1} & n^{(0)}_{1,2} & \cdots & n^{(0)}_{1,20} \\ n^{(0)}_{2,1} & n^{(0)}_{2,2} & \cdots & n^{(0)}_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ n^{(0)}_{L,1} & n^{(0)}_{L,2} & \cdots & n^{(0)}_{L,20} \end{bmatrix} \tag{17}$$

where $n^{(0)}_{i,j}$ stands for the initial score of amino acid residues during the evolution process the i-th $(i = 1,2,\ldots,L)$ sequential position has been changed into type $j(j = 1,2,\ldots,20)$ amino acid. The numbers 1, 2, …, 20, respectively, represent the 20 native amino acid types based on the alphabetical order considering only their single character codes (Chou et al. 2012). We can obtain the $L \times 20$ scores in Eq. (17) using PSI-BLAST (Schäffer et al. 2001) to search the UniProtKB/Swiss-Prot database (Boutet et al. 2007; UniPort Consortium 2008). There is an important problem to be noticed, when only Eq. (17) was used directly, because the data have a significant variation, it gives inaccurate results; in order to solve the problem, we should make each element in Eq. (17) change from 0 to 1, and thus a standard conversion was performed. Through the conversion, Eq. (17) will become this

$$P^{(1)}_{\text{PSSM}} = \begin{bmatrix} n^{(1)}_{1,1} & n^{(1)}_{1,2} & \cdots & n^{(1)}_{1,20} \\ n^{(1)}_{2,1} & n^{(1)}_{2,2} & \cdots & n^{(1)}_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ n^{(1)}_{L,1} & n^{(1)}_{L,2} & \cdots & n^{(1)}_{L,20} \end{bmatrix} \tag{18}$$

where

$$n^{(1)}_{i,j} = \frac{1}{1 + \mathrm{e}^{-n^{(0)}_{i,j}}} \tag{19}$$

After getting the PSSM matrix, we compute the average replaced possibility for all 20 types of amino acids, and finally 20 features are obtained. It can be formulated as (Zou et al. 2013)

$$P'_{\text{PSSM}} = \left[\overline{E_1}, \overline{E_2}, \ldots, \overline{E_{20}}\right]^{\mathrm{T}} \tag{20}$$

where T is the symbol of transpose operator

$$\overline{E_j} = \frac{1}{L}\sum_{i=1}^{L} n^{(1)}_{i,j} \tag{21}$$

where $\overline{E_j}$ represents the average score of the amino acid residues in the protein sequence being changed to amino acid type j during the evolution process.

Prediction Engine

In this study, the ML-KNN classifier was adopted to perform the prediction, which is derived from the classical KNN algorithm. The detailed description about how the classifier works is clearly described in Zhang and Zhou (2007), and hence there is no need for repeating it here. The predictor established in this study can be used to predict the functional types of both singleplex and multiplex human membrane proteins.

Performance Metrics

It is worthy to point out that for a multi-label learning system like the current system, which is different from the classical single-label learning system, the existing metrics, which were used to evaluate the quality of a predictor on a single-label system would fail to work when a multi-label learning system like this is faced. The metrics will be much more complicated for a multi-label learning system. We now describe the metrics used in multi-label system in the following section.

For a multi-label learning system containing N protein sequences, which belong to M functional types, L is the label set that contains all of the possible functional types concerned. Thus, the $i-$th sequence P_i and its corresponding functional type(s) can be expressed by

$$\{P_i, L_i\}(i = 1, 2, \ldots, N) \tag{22}$$

where L_i is the subset that includes all class label (s) for the $i-$th protein. Obviously, we have

$$L_1 \cup L_2 \cup \ldots \cup L_N \subseteq L \;=\; \{l_1, l_2, \ldots l_M\} \tag{23}$$

where $l_i (i = 1, 2, \ldots, M)$ corresponds to the label for the $i-$th functional type. In this study, the value of N is 3,166, the value of M is 8. Assume that L_i^* is the all predicted label(s) for the $i-$th sample. Thus, the following five metrics can be used to measure the prediction quality of the multi-label system:

$$\begin{cases} \text{Absolute} - \text{False} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\left(\dfrac{\left\|L_i \cup L_i^*\right\| - \left\|L_i \cap L_i^*\right\|}{M}\right) \\ \text{Accuracy} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\left(\dfrac{\left\|L_i \cap L_i^*\right\|}{\left\|L_i \cup L_i^*\right\|}\right) \\ \text{Precision} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\left(\dfrac{\left\|L_i \cap L_i^*\right\|}{\left\|L_i^*\right\|}\right) \\ \text{Recall} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\left(\dfrac{\left\|L_i \cap L_i^*\right\|}{\left\|L_i\right\|}\right) \\ \text{Absolute true} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\Delta(L_i, L_i^*) \end{cases} \tag{24}$$

where N is the number of different membrane proteins, M is the total number of functional types, here $N = 3,166$ and $M = 8$. The symbol $\cup$ and $\cap$ represent "union" set theory and intersection, respectively. $\| \;\|$ is the operator acting on the set therein to count the number of its elements, and

$$\begin{cases} \Delta(L_i, L_i^*) = 1, & \text{if all the labels in } \mathrm{L_i} \text{ are identified} \\ & \quad\text{to those in } \mathrm{L_i^*} \\ \Delta(L_i, L_i^*) = 0, & \text{otherwise} \end{cases} . \tag{25}$$

Among the five evaluation measures, the rate of absolute-false is opposite to those of the other four. As can be easily seen from Eq. (22), when the multi-labels for all of the samples are correctly predicted, i.e., $L_i \equiv L_i^*$ or $\left\|L_i \cup L_i^*\right\| = \left\|L_i \cap L_i^*\right\|$ $(i = 1, 2, \ldots, N)$, the rate of absolute-false equals to 0. When each of P $(i = 1, 2, \ldots, N)$ is predicted completely wrong, i.e., belonging to all the possible categories except its own category or categories; i.e., $L_i \cup L_i^* = L$ and $L_i \cap L_i^* = \emptyset$, or $\left\|L_i \cup L_i^*\right\| = M$ and $\left\|L_i \cap L_i^*\right\| = 0$, the rate of absolute-false is equal to 1. Therefore, the lower the absolute-false is, the better the prediction quality will be. However, for the other four metrics, the meanings of their rates are just opposite; i.e., the higher their rates are, the better the prediction quality will be.

Table 3 The results obtained by jackknife test with ML-KNN algorithm in the benchmark dataset

Methods	Evaluation measures				
	Absolute-false	Accuracy	Precision	Recall	Absolute-true
CPseAAC	0.0824	0.6376	0.6390	0.6379	0.6361
AAC + IE	0.0895	0.6095	0.6105	0.6098	0.6083
D + PSSM	0.0648	0.7467	0.7532	0.7478	0.7394

Results and Discussion

In statistical prediction, it is meaningless to simply say the success rate of a predictor without specifying what methods and benchmark dataset were utilized to test its accuracy (Wu et al. 2012). As is well known, there are three methods that are often used to examine the quality of a predictor: they are jackknife test, sub-sampling test, and independent dataset test, respectively. Among the three approaches, the jackknife test was considered as the least but most objective one, yielding a unique result for a given benchmark dataset, and hence it has been widely recognized and increasingly used to examine the power of various predictors. Therefore, the jackknife test was also adopted in this study to evaluate the quality of the predictor.

However, even though the jackknife test method has been used, the same predictor may also generate obviously different results for different benchmark datasets. The reason is that the more stringent a benchmark dataset in excluding homologous and high similarity sequences, the more difficult it becomes for a predictor to achieve a high overall success rate (Chou and Shen 2010). Also, the more the number of subsets a benchmark dataset covers, the more difficult it is to achieve a high overall success rate.

In this study, the results obtained are listed in Table 3. As we can see from Table 3, comparing the other two methods, the combination of D + PSSM provides better results; the overall absolute-true is 73.94 %, while the absolute-false is 6.48 %, i.e., the overall absolute-false rates are very low, while the absolute-true rates are quite higher; all of these results are indeed promising, indicating that the method is useful in identifying the functional types of membrane proteins.

Now, let us consider that a benchmark dataset consists of two subsets with each containing the same number of proteins. The overall success rate in identifying their attribute categories by random assignment would be 1/2 = 50 %; however, when the protein samples distributed among the eight different types are completely random, the overall success rate by random assignments would be 1/8 = 12.5 %; if the assignments are weighted as its sizes of subsets (Table 1), then the overall success rate would be:

$$\left(605^2+195^2+25^2+27^2+1444^2+251^2 +83^2+6377^2\right)/3267^2 \approx 27.79\,\% \quad (26)$$

Apparently, even the overall success rate by the worst solution in the benchmark dataset is overwhelmingly higher than the completely randomized rate and weighted randomized rate, so the models presented in this paper are indeed very encouraging (Huang and Yuan 2013).

Conclusion

Although many investigators made efforts in identifying the functional types of membrane proteins, it is still a challenge in this area with the explosion of newly found protein sequences entering into protein databanks. In this study, a new method by fusing various pseudo amino acid compositions was proposed, and the results obtained indicate that the new method has a very high potential for becoming a useful high-throughput tool for identifying the functional types of membrane proteins (Xiao et al. 2013). We hope it may play a key complementary role to the existing predictors in this area. In the future, we will investigate other methods for the sake of enhancing the powerful of the prediction.

Since user-friendly and publicly accessible web-servers provide direction for developing practically more useful models, simulated methods, or predictors (Chou and Shen 2009), we shall make efforts in our future work to provide a web server for the method presented in this study.

References

Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A (2007) Uniprotkb/swiss-prot. Springer, Plant Bioinformatics, pp 89–112

Chen Y-L, Li Q-Z (2007) Prediction of the subcellular location of apoptosis proteins. J Theor Biol 245:775–783

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

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

Chou K-C (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 273:236–247

Chou KC, Elrod DW (1999) Prediction of membrane protein types and subcellular locations. Proteins 34:137–153

Chou K-C, Elrod DW (2003) Prediction of enzyme family classes. J Proteome Res 2:183–190

Chou K-C, Shen H-B (2007) MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Commun 360:339–345

Chou K-C, Shen H-B (2009) Review: recent advances in developing web-servers for predicting protein attributes. Nat Sci 1:63

Chou C-H, Shen H-B (2010) Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms. Eng, 2

Chou K-C, Wu Z-C, Xiao X (2011) iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. PLoS One 6:e18258

Chou K-C, Wu Z-C, Xiao X (2012) iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol BioSyst 8:629–641

Ding C, Yuan L-F, Guo S-H, Lin H, Chen W (2012) Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions. J Proteomics 77:321–328

Glory E, Murphy RF (2007) Automated subcellular location determination and high-throughput microscopy. Dev Cell 12:7–16

Huang C, Yuan J-Q (2013a) A multilabel model based on Chou's pseudo-amino acid composition for identifying membrane proteins with both single and multiple functional types. J Membr Biol 246:327–334

Huang C, Yuan J-Q (2013b) Predicting protein subchloroplast locations with both single and multiple sites via three different modes of Chou's pseudo amino acid compositions. J Theor Biol 335:205–212

Jian X, Wei R, Zhan T, Gu Q (2008) Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. Protein Pept Lett 15:392–396

Khosravian M, Kazemi Faramarzi F, Mohammad Beigi M, Behbahani M, Mohabatkar H (2013) Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods. Protein Pept Lett 20:180–186

Li F-M, Li Q-Z (2008) Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. Protein Pept Lett 15:612–616

Lin H, Wang H, Ding H, Chen Y-L, Li Q-Z (2009) Prediction of subcellular localization of apoptosis protein using Chou's pseudo amino acid composition. Acta Biotheor 57:321–330

Lin W-Z, Fang J-A, Xiao X, Chou K-C (2013) iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. Mol BioSyst 9:634–644

Nakashima H, Nishikawa K, Tatsuo O (1986) The folding type of a protein is relevant to the amino acid composition. J Biochem 99:153–162

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

Pu X, Guo J, Leung H, Lin Y (2007) Prediction of membrane protein types from sequences and position-specific scoring matrices. J Theor Biol 247:259–265

Saravanan V, Lakshmi P (2013) APSLAP: an adaptive boosting technique for predicting subcellular localization of apoptosis protein. Acta Biotheor 61:481–497

Schäffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29:2994–3005

Shen H-B, Chou K-C (2007) Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem Biophys Res Commun 355:1006–1011

Shen H-B, Chou K-C (2008) PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Anal Biochem 373:386–388

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

Smith C (2008) Subcellular targeting of proteins and drugs. http://www.biocompare.com/Articles/TechnologySpotlight/976/Subcellular-Target-ing-Of-Proteins-An

UniProt Consortium (2008) The universal protein resource (UniProt). Nucleic Acids Res 36:D190–D195

Wang X, Li G-Z (2012) A multi-label predictor for identifying the subcellular locations of singleplex and multiplex eukaryotic proteins. PLoS One 7:e36317

Wang M, Yang J, Xu Z-J, Chou K-C (2005) SLLE for predicting membrane protein types. J Theor Biol 232:7–15

Wootton JC, Federhen S (1993) Statistics of local complexity in amino acid sequences and sequence databases. Comput Chem 17:149–163

Wu Z-C, Xiao X, Chou K-C (2012) iLoc-Gpos: a multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins. Protein Pept Lett 19:4–14

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

Xiao X, Wang P, Lin W-Z, Jia J-H, Chou K-C (2013) iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal Biochem 436:168–177

Zhang M-L, Zhou Z-H (2007) ML-KNN: a lazy learning approach to multi-label learning. Pattern Recogn 40:2038–2048

Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:17

Zhou X-B, Chen C, Li Z-C, Zou X-Y (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J Theor Biol 248:546–551

Zou Q, Wang Z, Guan X, Liu B, Wu Y, Lin Z (2013) An Approach for identifying cytokines based on a novel ensemble classifier. BioMed Res Int 2013

Zou Q, Chen W, Huang Y, Liu X, Jiang Y (2013b) Identifying multi-functional enzyme by hierarchical multi-label classifier. J Comput Theor Nanosci 10:1038–1043