J Membrane Biol (2014) 247:1141–1148
DOI 10.1007/s00232-014-9708-2

A Multi-label Classifier for Prediction Membrane Protein Functional Types in Animal

Hong-Liang Zou

Received: 14 April 2014 / Accepted: 14 July 2014 / Published online: 9 August 2014

Abstract Membrane protein is an important composition of cell membrane. Given a membrane protein sequence, how can we identify its type(s) is very important because the type keeps a close correlation with its functions. According to previous studies, membrane protein can be divided into the following eight types: single-pass type I, single-pass type II, single-pass type III, single-pass type IV, multipass, lipid-anchor, GPI-anchor, peripheral membrane protein. With the avalanche of newly found protein sequences in the post-genomic age, it is urgent to develop an automatic and effective computational method to rapid and reliable prediction of the types of membrane proteins. At present, most of the existing methods were based on the assumption that one membrane protein only belongs to one type. Actually, a membrane protein may simultaneously exist at two or more different functional types. In this study, a new method by hybridizing the pseudo amino acid composition with multi-label algorithm called LIFT (multi-label learning with label-specific features) was proposed to predict the functional types both singleplex and multiplex animal membrane proteins. Experimental result on a stringent benchmark dataset of membrane proteins by jackknife test show that the absolute-true obtained was 0.6342, indicating that our approach is quite promising. It may become a useful high-through tool, or at least play a complementary role to the existing predictors in identifying functional types of membrane proteins.

Keywords Membrane protein · Jackknife test · Absolute-true · Multi-label algorithm

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

Introduction

As the main undertaker of the membrane function, membrane protein plays an essential role in various biochemical process such as transmembrane transport, energy exchange, signal transduction, and so on (Pu et al. 2007). Almost, 30 % encoded proteins in nuclear genome are membrane proteins. In addition, membrane protein constitutes 60 % of drug target, which were crucial to understand the mechanism of cellular activities as well as new drug discover or design. According to the report, the function of a membrane protein is closely correlated with its type(s). In the post-genomic era, the number of newly found protein sequences has been rapidly increasing, such as the number of protein sequence entries was only 3939 in Swiss-Prot in 1986 (Boeckmann et al. 2003), however, the number jumped to 54790250 according to the version released on 19-March-2014 at http://www.uniprot.org/, which is more 10000 times than in 1986. The gap between newly found proteins and the information of functional types is becoming increasingly wide. Although the functional type of a membrane protein may be determined by carrying out various biochemical experiments, it will be time-consuming and costly. Therefore, to bridge such a gap, it is urgent to develop an automatic and effective computational method to identify the types of membrane protein.

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

In the past several years, many efforts have been made in identifying the functional types of membrane proteins based on the sequence information, such as Cai et al.

(2004) predicting membrane protein types using amino acid composition (AAC) with support vector machine (SVM); Hayat and Khan (2012) by hybridizing the split amino acid composition (SAAC) and seven physicochemical properties of protein with SVM to predict membrane protein types; Chou and Shen (2007) using Pse-PSSM to predict eight membrane types with OET-KNN, and obtained overall success rate 85 % by jackknife test, and many others.

Although those methods aforementioned each have their own advantages and did play an important role in stimulating the development of this area (Xiao et al. 2013), they were focused on identifying one of its subtypes, without considering various possible different functional types of membrane protein. In fact, many membrane proteins have two or more functional types or different functions. Membrane proteins with multiple types or dynamic feature of this kind are particularly interesting, because they may have some unique biological functions worthy of our special notice (Glory and Murphy 2007; Smith 2008; Wang and Li 2012).

In this study, to better reflect the characteristics of multiplex proteins, a new predictor has been developed that can be utilized to deal with the systems containing both singleplex and multiplex membrane proteins by introducing a powerful multi-label learning algorithm called LIFT which exploits correlations between the types and by hybridizing the amino acid composition (AAC), CTD (composition, translation, distribution), and EBGW(encoding based on grouped weight) information (Wang and Li 2012).

According to a comprehensive review (Chou 2011), in order to establish a useful and powerful predictor for a biological system based on sequence information, the following procedures should be considered: (1) construct or choose a valid dataset to train and test the predictor; (2) using an effective mathematical expression to formulate the protein sequence, which can truly reflect the intrinsic correlation with the target to be predicted; (3) develop or introduce a powerful algorithm to conduct the prediction processes; (4) properly perform a cross-validation test to objectively evaluate the anticipated accuracy.

Materials

To establish a high quality benchmark dataset for developing a predictor to identify the functional types of membrane proteins, the sequences were collected from UniProtKB/Swiss-Prot release on 2014_03 at http://www.uniprot.org/ according to the following steps (Lin et al. 2013).

Step 1 Only these protein sequences annotation of "metazoa" were collected.

Table 1 The information about the benchmark dataset S constructed in this study

Subset	Type	Number of membrane proteins
1	Single-pass type I	519
2	Single-pass type II	166
3	Single-pass type III	28
4	Single-pass type IV	33
5	Multipass	1,029
6	Lipid-anchor	191
7	GPI-anchor	81
8	Peripheral	602
Total number of virtual proteins		2,649
Total number of different proteins		2,559

Step 2 Those proteins belonging to human beings were removed.

Step 3 Those proteins annotation with "fragment" were removed; meanwhile, those proteins with the length of sequence less than 50 residues were also excluded, in case of the influence of the fragment.

Step 4 Sequences annotated with ambiguous or uncertain terms, such as "potential," "probable," "probably," "maybe," or "by similarity," were removed for further consideration.

Step 5 In order to reduce the influence of the redundancy and homology bias, a software called "CD-HIT" was used to remove these proteins with more than 40 % pairwise sequence identify to any others in the same subset except for the subset of "single-pass III", because there is few sequence in the subset, if not so, the data in the subset may be too few to have statistically significant.

Finally, we obtained 2,559 different animal membrane protein sequences covered in eight functional types, those proteins form the benchmark dataset S for the current study, it can be formulated as

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

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 number varying from 1 to 8 was used to represent the 8 subsets. The detailed information about the benchmark dataset was listed in Table 1.

Among the 2,559 different animal membrane proteins, 2,473 belong to one functional type, 82 to two types, 4 to three types, none in four or more functional types.

Because some proteins may simultaneously have two or more functional types, the concept of "virtual protein" (Lin et al. 2013; Xiao et al. 2013) was introduced, that is if a protein coexist two different types, it will be counted as two virtual proteins; if coexist three different types, it will be counted as three virtual proteins, and so forth. Thus, the number of total virtual proteins can be calculated by the following equation:

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

where N(vir) is the number of total virtual membrane proteins, N(seq) is the number of different membrane protein sequences, $N(1)$ the number of membrane proteins with one functional type, $N(2)$ the number of membrane proteins with two different functional types, and so forth; and M is the number of total functional types investigated. Substituting the afore mentioned data in Eq. (2), we obtained

$$\begin{aligned} N(vir) &= N(seq) + (1-1) \times 2473 + (2-1) \times 82 \\ &\quad + (3-1) \times 4 \\ &= 2559 + 82 + 8 = 2649, \end{aligned} \tag{3}$$

meaning that the total number of virtual membrane proteins is 2,649, which is actually also the sum of the protein numbers for the 8 subsets listed in Table 1. As we can see from Eq. (2) and Eq. (3), the number of total virtual proteins is generally greater than that of total different protein sequences. When and only when, all of the proteins have a single-type, can the two be the same.

Methods

In order to develop a powerful predictor for identifying membrane proteins functional types based on the sequence information, one of the first important things is to formulate the proteins samples with an effective mathematical expression that can truly reflect the intrinsic correlation with the target to be investigated (Chou 2011; Xiao et al. 2013). However, it is by no means a trivial and easy job to realize this because this kind of correlation is usually deeply hidden or "buried" into piles of complicated sequences (Lin et al. 2013; Xiao et al. 2013).

The most straightforward method to formulate the sample of a query protein sequence P with L-amino acids is to use its entire amino acid sequence, it can be expressed by

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

where R_1 represent the first residue in the protein sequence, R_2 the second residue,…, R_L the L–th residue, each of them belongs to one of the 20 native amino acids. To identify its functional type(s), the sequence similarity-search-based tools, such as BLAST (Altschul 1997; Wootton and Federhen 1993) was utilized to search the protein database for those proteins that have high sequence similarity to the query protein P. Subsequently, the function annotations of the targeted proteins thus found were used to infer the function for the query protein P. However, this kind of straightforward sequential model, although quite intuitive and able to contain the entire sequence information of a protein sample, failed to work when the query protein P did not have any significant sequence similarity to any attribute-known proteins (Chou and Shen 2007; Xiao et al. 2013).

To overcome the above difficulty, which is inherent to the sequence model, various non-sequential or discrete models were introduced to represent the sample of a protein with a series of discrete numbers, which in hope to enhance the prediction power (Xiao et al. 2013).

Among the various discrete models, the simplest one is to represent the sample of a protein with its amino acid composition or AAC (Nakashima et al., 1986). According to the AAC-discrete model, the protein P can be formulated by (Chou 1999; Chou 1995)

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

where f_i ($i = 1,2,\ldots,20$) are the normalized occurrence frequencies of the 20 native amino acids in protein P according to the alphabetic order, and T represents the transposing operator. The AAC-discrete model was widely utilized in prediction protein attributes. However, as we can see from Eq. (5), if only AAC model was used to represent the protein P, all of its sequence-order effects would be lost, and hence might considerably limit the prediction quality.

In order to avoid completely losing the sequence-order information, the pseudo amino acid composition (PseAAC, also called "Chou's PseAAC" (Lin and Lapointe 2013)) was proposed to represent the protein sample by Chou (2001). After that, the pseudo amino acid composition was widely used in bioinformatics, proteomics and system biology, such as predicting G-Protein-Coupled Receptor classes (Chou 2005; Xiao et al. 2009), prediction subcellular location of proteins (Chou et al. 2011; Gao et al. 2005; Shen and Chou 2007b; Xiao et al. 2005), subnuclear location of proteins prediction (Jiang et al. 2008; Li and Li 2008; Mundra et al., 2007; Shen and Chou, 2005), predicting GABA(A) receptor proteins (Mohabatkar et al. 2011), predicting enzyme family and sub-family classes (Qiu et al. 2010; Shen and Chou 2007a; Wang et al. 2010), identifying the functional types of antimicrobial peptides (Khosravian et al. 2013; Xiao et al. 2013), predicting subcellular location of apoptosis proteins (Chen and Li 2007; Jian et al. 2008; Lin et al. 2009; Saravanan and

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 Waals 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

Lakshmi 2013; Zhang et al. 2006; Zhou and Doctor 2003), among many others.

According to previous studies, the general form of PseAAC for a protein P is formulated by (Chou 2001)

$$P = [\xi_1, \xi_2, \ldots, \xi_\Omega]^T, \tag{6}$$

where the subscript Ω is an integer, and its value as well as the components $\xi_1, \xi_2, \ldots$ will depend on how to extract the desired information from the amino acid sequence of P. In the following, we will describe in detail the process of how to define the elements in Eq. (6).

CTD of Physiochemical Descriptors

The 20 native amino acids can be divided into 3 groups according to the following eight different physiochemical properties (as show in Table 2) (Hua and Sun 2001; Nanni and Lumini 2006; Saravanan and Lakshmi 2013; Zou et al. 2013): secondary structure, solvent accessibility, normalized van der waals volume, hydrophobicity, charge, polarizability, polarity, surface tension. Three descriptors, composition (C_S), transition (T_{xy}), and distribution (D_S), are utilized to describe the global composition of each of these properties, they can be calculated using the following equations:

$$C_S = n_s/L\,(S = 1, 2, 3), \tag{7}$$

where n_s represent the number of s in the encoded sequence, and L is the length of the protein sequence.

$$T_{xy} = \frac{n_{xy} + n_{yx}}{L - 1} xy = [12], [13], [23], \tag{8}$$

where n_{xy} is the number of dipeptide encoded as "xy" and "yx", respectively.

There are totally five distributions that 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:

$$E_i 1D_x = \frac{P_1}{L} \tag{9}$$

$$E_i 25D_x = \frac{P_{25}}{L} \tag{10}$$

$$E_i 50D_x = \frac{P_{50}}{L} \tag{11}$$

$$E_i 75D_x = \frac{P_{75}}{L} \tag{12}$$

$$E_i 100D_x = \frac{P_{100}}{L}\,(i = 1, 2, \ldots, 8; x = 1, 2, 3), \tag{13}$$

where P_1, P_{25}, P_{50}, P_{75}, and P_{100} were the position of first occurrence of x, position of 25, 50, 75, and 100 % occurrence of x, respectively. The values of composition, translation, and distribution were calculated for all the eight descriptors, and the corresponding feature vector CTD was expressed as

$$CTD = [Ci_{s[1,2,3]}, Ti_{xy[12,13,23]}, Ei_{[1,2,3,4,5]}] \quad (i = 1, 2, \ldots, 8) \tag{14}$$

Now, let us give an example to explain the CTD in detail in the following (Cai et al. 2003). Assuming that there is a protein sequence, its amino acid composition is AEAAAEAEEAAAAAEAEEEAAEEAEEEAAE, which has 16 alanines, i.e., $n_1 = 16$; and 14 glutamic acids, i.e., $n_2 = 14$. The composition of the two kind of amino acids is $C_1 = n_1/(n_1 + n_2) = 16/(16 + 14) = 0.5333$, and meanwhile the C_2 can be formulated as $C_2 = n_2/(n_1 + n_2) = 14/(16 + 14) = 0.4667$, there are total 15 transitions from A to E or from E to A in the sequence, that is $n_{xy} + n_{yx} = 15$; thus, the percent frequency of these transitions is T = 15/(20–1) = 0.5172. The first, 25, 50, 75, and 100 % of A are located in the first, 5-th, 12-th, 20-th, and 29-th 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. Similar, the D descriptor for E is 0.0667, 0.2667, 0.6000, 0.7667, and 1.0000. Overall, the amino acid composition descriptors for this sequence are $C = (0.5333, 0.4667)$, $T = (0.5172)$, and $D = (0.0333, 0.1667, 0.4000, 0.6667, 0.9667, 0.0667, 0.2667, 0.6000, 0.7667, 1.0000)$.

Encoding Based On Grouped Weight (EBGW)

There existed a situation that is for some different things, we can treat them as the one if they have some same features. This is the concept of coarse-gained. According to the charged and hydrophobicity character, the 20 native amino acid residues can be divided into the following four classes (Zhang et al. 2006):

Neutral and non-polarity residue	CG1 = {A,F,G,I,L,M,P,V,W}
Neutral and polarity residue	CG2 = {C,N,Q,S,T,Y}
Acidic residue	CG3 = {D,E}
Basic residue	CG4 = {H,K,R}

Thus, we will obtain three groups, one of which can divide the 20 native amino acids into two disjoint combinations: CG1+CG2 versus CG3+CG4, CG1+CG3 versus CG2+CG4, and CG1+CG4 versus CG2+CG3.

For a protein sequence with L amino acid residues, it can be expressed by

$$X = x_1 x_2 \ldots x_L, \tag{15}$$

where x_1 represents the first residue of the sequence, x_2 represents the second residue, and so forth. Then, the sequence can be transformed into three binary sequences by three homomorphic maps $\psi_i(X(L)) = \psi_i(x_1)\psi_i(x_2)\ldots\psi_i(x_L)(i = 1, 2, 3)$, they can be expressed as below:

$$\psi_1(x_i) = \begin{cases} 1 & \text{if } x_i \in CG1 \cup CG2 \\ 0 & \text{if } x_i \in CG3 \cup CG4 \end{cases} (i = 1,\ 2,\ 3,\ \ldots,\ L) \tag{16}$$

$$\psi_2(x_i) = \begin{cases} 1 & \text{if } x_i \in CG1 \cup CG3 \\ 0 & \text{if } x_i \in CG2 \cup CG4 \end{cases} (i = 1,\ 2,\ 3,\ \ldots,\ L) \tag{17}$$

$$\psi_3(x_i) = \begin{cases} 1 & \text{if } x_i \in CG1 \cup CG4 \\ 0 & \text{if } x_i \in CG2 \cup CG3 \end{cases} (i = 1,\ 2,\ 3,\ \ldots,\ L) \tag{18}$$

Defined $U(L)^j = \psi(X(L)) = U_1^j, U_2^j, \ldots, U_L^j (j = 1, 2, 3)$, we called $U(L)^1$, $U(L)^2$, and $U(L)^3$ as the 1-, 2- and 3-characteristic sequences of the proteins, respectively.

For convenience, in the following section, we use $U(L) = U_1, U_2,\ldots, U_L$ as any character sequence of the three defined above.

Let $U(L) = U_1, U_2,\ldots, U_L$ be a character sequence, the weight of $U(L)$ can be defined as the number of occurrences of digit 1 in $U(L)$. From the above, we can know that the weight of sequences rely on the length of sequence. Due to the length of sequences is different for different protein sequences, to make sure the effectiveness of features what we extract, we should make the weight normal. It can be described as $w(L) = v/L$, where the v represents the weight of $U(L)$. For a character sequence $U(L) = U_1, U_2,\ldots, U_L$, given a positive integer n, the protein sequence can be divided into n subsequences. The length of every subsequence is gradually increasing. The subsequence of $U(L)$ can be represented as $U([K \times L/n])$ $(k = 1,2,3,\ldots,n)$, whose length is $([K \times L/n])$ $(k = 1,2,3,\ldots,n)$, in above expression, the symbol [.] represent the operation return the value which is down to the nearest integer. The normalized weight of $U([K \times L/n])$ $(k = 1,2,3,\ldots,n)$ can be written as $w([K \times L/n])$ $(k = 1,2,3,\ldots,n)$. Thus, through the above processes, we can obtain the EBGW string of a character sequence as the following equation:

$$W = (w[L/n], w[2 \times L/n], \ldots, w[n \times L/n]). \tag{19}$$

Therefore, for a given sequence $X(L) = x_1, x_2,\ldots, x_L$, we can convert it into three character sequences, i.e., $U(L) = U_1, U_2,\ldots, U_L$. Thus, for a character sequence, we can use L-dimension vector to represent it. A protein sequence can be transformed into a 3L-dimension vector. From above, we can see that the value of n is important to the result. In this, by preliminary computation and analyses, we find when $n = 50$, the best result would be obtained, thus we can use 150 elements to represent a protein sequence, it can be expressed as

$$EBGW = [\zeta_1 \zeta_2 \ldots \zeta_{150}]^T. \tag{20}$$

Combination all of the features aforementioned, it can be formulated as

$$P = [\psi_1, \psi_2, \ldots, \psi_{338}], \tag{21}$$

where the first 20 elements from AAC, $\psi_{21}, \psi_{22}, \ldots \psi_{188}$ from CTD, the rest elements from EBGW.

LIFT Classifier

In this study, the multi-label classifier called LIFT was used to perform the prediction. The detailed description of how the classifier works is clearly described in (Zhang 2011), and hence, there is no need to repeat here. The predictor established in this study has the ability to predict the functional types of both singleplex and multiplex animal membrane proteins. To provide an intuitive picture, a flowchart is provided in Fig. 1 to illustrate the prediction process.

Results and Discussion

In statistical prediction, it would be meaningless to simply say a success rate of a predictor without considering what

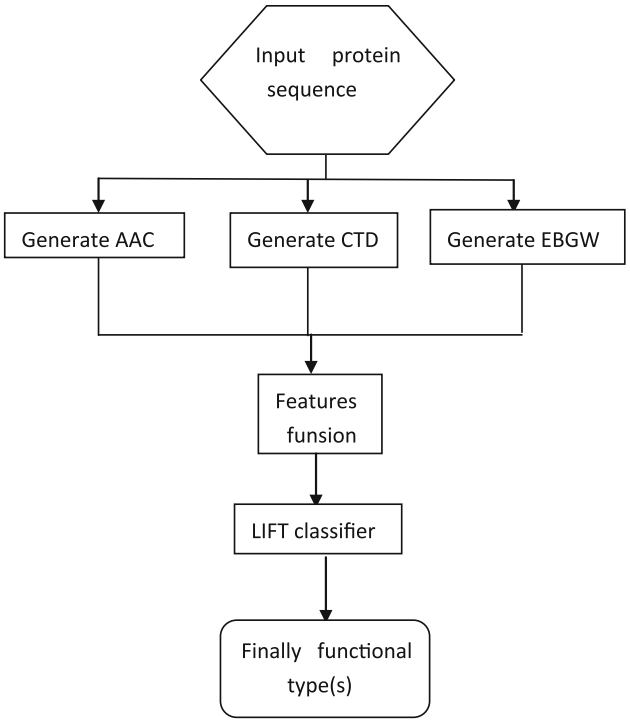

Fig. 1 The flowchart to show the prediction process

method and benchmark dataset were used to test its accuracy (Wu et al. 2012). As is well known, the following three methods are often utilized to examine the quality of a predictor: jackknife test, subsample test, and independent dataset test. Among the three methods, the jackknife test, also called Leave-One-Out (LOO) cross-validation, was considered as the least arbitrary that can always yield a unique result for a given benchmark dataset and hence has been widely used by various investigators. Accordingly, the jackknife test was also used in this study to evaluate the power of the predictor.

However, even though using the jackknife test approach for cross-validation, it still may generate obviously different success rates for a same predictor when tested by different benchmark datasets. This is because the more stringent of a benchmark dataset in excluding homologous and high similarity sequences, the more difficult for a predictor to achieve a high overall success rate. Meanwhile, the more number of subsets (functional types) a benchmark dataset covered, the more difficult to achieve a high overall success rate. This can be easily conceivable via the following example. Assuming a benchmark dataset consists of four subsets with each containing a same number of protein sequences, the success rate would be $1/4 = 25\ \%$ by random guess, however, if the benchmark dataset consists of eight subsets, and each of the subset has the same proteins, the corresponding overall success rate would be $1/8 = 12.5\ \%$.

Table 3 The result obtained in this study

Hamming loss	One-error	Coverage	Ranking loss	Average precision
0.0632	0.2493	0.5979	0.0792	0.8443

For such a complicated dataset containing both single-plex and multiplex membrane proteins distributed among eight functional types, this is the first try in predicting animal membrane functional types, the result obtained are listed in Table 3.

To provide a more intuitive and easier-to-understand measurement, a new scale called absolute-true was introduced to reflect the accuracy of the predictor, it can be formulated as

$$Absolute{-}True = \frac{1}{N}\sum_{i=1}^{N}\Delta(i), \tag{22}$$

where N is the number of different virtual membrane proteins, in here $N = 2559$.

$$\Delta(i) = \begin{cases} 1 & \text{if all the functional types of the ith membrane proteinis correctly predicted without any overprediction} \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

According to the above definition, we can see that if a protein has three functional types, only two are correctly predicted, or in fact the predicted result contains a type not belongs to the three, the prediction score will be counted as 0. In other words, when and only when all types have been correctly predicted for a query protein, the prediction score can be counted as 1. It is instructive to point out that, for a multi-label system like this, the absolute-true success rate for an individual membrane protein functional type is meaningless and misleading. Therefore, instead of the absolute-true success rate for each of the individual functional types, the overall absolute-true success rate achieved in this study is 0.6342, indicating that the predictor is a quite promising multi-label predictor in identifying the functional types of animal membrane proteins.

Conclusion

Membrane protein is a kind of important proteins in most of the creatures. Though, there are many models have been proposed in the past several years, it is still a challenging task to predict the functional types of membrane proteins with multiple membrane types.

In this study, a new model was proposed to predict animal membrane proteins with single or multiple types. From the result obtained listed in Table 3, we can see that the new predictor holds very high potential to become a useful high throughput tool for identifying animal membrane protein functional type, we hope it will play an important complementary role to the existing predictors in this area. Though a promising result has obtained, there is still much room for further improvement in future studies. This is our direction of the research in the future.

References

Altschul SF (1997) Evaluating the statistical significance of multiple distinct local alignments. Theoretical and computational methods in genome research. Springer, New York, pp 1–14

Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31:365–370

Cai C, Han L, Ji ZL, Chen X, Chen YZ (2003) SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res 31:3692–3697

Cai Y-D, Ricardo P-W, Jen C-H, Chou K-C (2004) Application of SVM to predict membrane protein types. J Theor Biol 226:373–376

Chen Y-L, Li Q-Z (2007) Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo-amino acid composition. J Theor Biol 248:377–381

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

Chou K-C (1999) A key driving force in determination of protein structural classes. Biochem Biophys Res Commun 264:216–224

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

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

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

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, 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

Gao Y, Shao S, Xiao X, Ding Y, Huang Y, Huang Z, 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 28:373–376

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

Hayat M, Khan A (2012) Mem-P Hybrid: hybrid features-based prediction system for classifying membrane protein types. Anal Biochem 424:35–44

Hua S, Sun Z (2001) A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J Mol Biol 308:397–407

Huang C, Yuan J-Q (2013) 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

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

Jiang X, Wei R, Zhao Y, Zhang T (2008) Using Chou's pseudo amino acid composition based on approximate entropy and an ensemble of AdaBoost classifiers to predict protein subnuclear location. Amino Acids 34:669–675

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) Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. Amino Acids 34:119–125

Lin S-X, Lapointe J (2013) Theoretical and experimental biology in one. J Biomed Sci Eng 6:435–442

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

Mohabatkar H, Mohammad Beigi M, Esmaeili A (2011) Prediction of GABA < sub > A </sub > receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. J Theor Biol 281:18–23

Mundra P, Kumar M, Kumar KK, Jayaraman VK, Kulkarni BD (2007) Using pseudo amino acid composition to predict protein subnuclear localization: Approached with PSSM. Pattern Recogn Lett 28:1610–1615

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

Nanni L, Lumini A (2006) MppS: an ensemble of support vector machine based on multiple physicochemical properties of amino acids. Neurocomputing 69:1688–1690

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

Qiu J-D, Huang J-H, Shi S-P, Liang R-P (2010) Using the concept of chou's pseudo amino acid composition to predict enzyme family classes: an approach with support vector machine based on discrete wavelet transform. Protein Pept Lett 17:715–722

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

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

Shen H-B, Chou K-C (2007a) EzyPred: a top–down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Commun 364:53–59

Shen H-B, Chou K-C (2007b) 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

Smith C (2008) Subcellular targeting of proteins and drugs. http://www.biocompare.com/Editorial-Articles/41619-Subcellular-Targeting-Of-Proteins-And-Drugs/

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 Y-C, Wang X-B, Yang Z-X, Deng N-Y (2010) Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature. Protein Pept Lett 17:1441–1449

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, Chou KC (2009) GPCR-CA: a cellular automaton image approach for predicting G-protein–coupled receptor functional classes. J Comput Chem 30:1414–1423

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 (2011) LIFT: Multi-label learning with label-specific features. In: Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Two, AAAI Press, pp 1609–1614

Zhang Z-H, Wang Z-H, Zhang Z-R, Wang Y-X (2006) A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. FEBS Lett 580:6169–6174

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

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 2013:686090