Journal of Theoretical Biology 344 (2014) 31–39

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

A two-stage SVM method to predict membrane protein types by incorporating amino acid classifications and physicochemical properties into a general form of Chou's PseAAC

Guo-Sheng Han[a], Zu-Guo Yu[a,b,*], Vo Anh[b]

[a] School of Mathematics and Computational Science, Xiangtan University, Hunan 411105, China
[b] School of Mathematical Science, Queensland University of Technology, GPO Box 2434, Brisbane Q 4001, Australia

HIGHLIGHTS

- We present a novel SVM method for predicting membrane protein types.
- The SVM method is combined with a two-step optimal feature selection process.
- The performance of the proposed method is evaluated on two benchmark datasets.
- Our method provides better performance as compared to the existing approaches.

ARTICLE INFO

Article history:
Received 16 August 2013
Received in revised form
16 October 2013
Accepted 24 November 2013
Available online 4 December 2013

Keywords:
Membrane protein type
Feature extraction
Amino acid classification
Support vector machine
Hilbert–Huang transform

ABSTRACT

Membrane proteins play important roles in many biochemical processes and are also attractive targets of drug discovery for various diseases. The elucidation of membrane protein types provides clues for understanding the structure and function of proteins. Recently we developed a novel system for predicting protein subnuclear localizations. In this paper, we propose a simplified version of our system for predicting membrane protein types directly from primary protein structures, which incorporates amino acid classifications and physicochemical properties into a general form of pseudo-amino acid composition. In this simplified system, we will design a two-stage multi-class support vector machine combined with a two-step optimal feature selection process, which proves very effective in our experiments. The performance of the present method is evaluated on two benchmark datasets consisting of five types of membrane proteins. The overall accuracies of prediction for five types are 93.25% and 96.61% via the jackknife test and independent dataset test, respectively. These results indicate that our method is effective and valuable for predicting membrane protein types. A web server for the proposed method is available at http://www.juemengt.com/jcc/memty_page.php

1. Introduction

According to cellular anatomy, a cell consists of different functional units or organelles, most of which are enveloped by membranes and they are necessary for many biological functions. Although the lipid bilayer is the basic structure of membranes, most of the specific functions of the cell membrane are performed by the membrane proteins (Alberts et al., 1994; Lodish et al., 1995). Membrane proteins are significantly important for many biological processes, such as contact of the membrane proteins in the cell, recognition of surface, transduction of the signal, and activity of enzyme. Membrane proteins are also attractive targets of drug discovery for various diseases (Tusnady et al., 2004; Feng et al., 2005, 2006; Sanders et al., 2006). Many researchers believe that membrane proteins constitute approximately 50% of possible targets for novel drugs (Gao et al., 2010). Given a particular membrane protein, the elucidation of its membrane protein types can provide clues for understanding the structure and function of proteins (Chou and Elrod, 1999). However, the prediction of membrane protein types from the traditional experimental and manual annotation methods is expensive and time consuming, hence effective computational methods are in urgent need for discriminating uncharacterized proteins.

Based on their functions, membrane proteins can be classified into transmembrane proteins, which span across the membrane, and anchored proteins, which are attached to the membrane on one side (Chou and Elrod, 1999). Transmembrane and anchored

[*] Corresponding author at: School of Mathematics and Computational Science, Xiangtan University, Hunan 411105, China.
E-mail address: yuzg1970@yahoo.com (Z.-G. Yu).

0022-5193/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2013.11.017

proteins have distinct structural features. The transmembrane proteins usually contain one or more hydrophobic segments (i.e. several consecutive hydrophobic amino acids); anchored proteins have a consensus sequence pattern (motif) on either the N- or C-terminus. Transmembrane proteins can be further classified into three sub-types including Type-I transmembrane, Type-II transmembrane and multi-pass transmembrane proteins. Anchored proteins consist of two sub-types, lipid chain-anchored membrane and glycophosphatidylinositol(GPI)-anchored membrane proteins. Five sub-types of membrane proteins have respective significant features. Type-I and Type-II transmembrane proteins are called single-pass transmembrane proteins. Type-I transmembrane proteins are extracellular on N-terminus and cytoplasmic on C-terminus, whereas Type-II transmembrane proteins are extracellular on C-terminus and cytoplasmic on N-terminus. In both types, the lipid bilayer is crossed by polypeptide only once, whereas in multi-pass membrane protein, the lipid bilayer is crossed by polypeptide multiple times. The lipid chain-anchored membrane protein is associated with the bilayer, whereas GPI-anchored membrane protein is directly bound to the membrane by a GPI anchor. Besides the five types of membrane proteins, some investigation has also been carried out to predict them among their seven types (Chou and Shen, 2007); i.e., (1) Type I, (2) Type II, (3) Type III, (4) Type IV, (5) multipass, (6) lipid-chain-anchored, and (7) GPI-anchored.

In the last few decades, various approaches have been developed for predicting membrane protein types. Chou and Elrod (1999) first carried out the prediction of membrane protein types based on the covariant discriminant algorithm (*CDA*) and amino acid composition (*AAC*). Similarly, Cai et al. (2004) used *AAC* in conjunction with support vector machine (*SVM*) to predict membrane protein types. However, when using *AAC* to represent a protein, the sequence-order and sequence-length effects are lost. To avoid this shortcoming, the pseudo-amino acid composition (*PseAAC*) was proposed (Chou, 2001, 2005; Hayat and Khan, 2011) to improve the prediction accuracy of membrane protein types. Then, various forms of the *PseAAC* method were proposed (Wang et al., 2004, 2005, 2006; Shen and Chou, 2005; Shen et al., 2006; Lin, 2008; Mahdavi and Jahandideh, 2011; Wang et al., 2012; Hayat and Khan, 2012a,b; Chen and Li, 2013; Huang and Yuan, 2013) for predicting membrane protein types and related tasks.

A recent comprehensive review for more detailed information about development and applications of *PseAAC* was given by Chou and Shen (2009). Higher order peptide composition is an alternative to improve the *AAC* method in conjunction with stepwise discriminant analysis (Yang et al., 2007) and k-nearest neighbor (Wang et al., 2010). Besides, non-sequence information is also explored for predicting membrane protein types including function domain (Cai et al., 2003), gene ontology (Chou and Cai, 2005), and evolutionary information (Pu et al., 2007; Chou and Shen, 2007; Hayat and Khan, 2012a). Chou and Shen (2007) proposed the Pse-PSSM method and developed a web server for predicting membrane protein types. In addition, time–frequency analysis methods are also applied in this field including the Fourier spectrum (Liu et al., 2005) and the discrete wavelet transform (*DWT*) (Rezaei et al., 2008; Qiu et al., 2010). As well as feature combination, ensemble classifier can be constructed to improve performance (Nanni and Lumini, 2008; Shen and Chou, 2007).

Recently we developed a novel system for predicting protein subnuclear localizations (Han et al., 2013). In the present paper, we propose a simplified version of our system for predicting membrane protein types directly from primary protein structures. In this simplified system, we will design a two-stage multi-class support vector machine (*MSVM*) combined with a two-step optimal feature selection process by incorporating amino acid classifications and physicochemical properties into a general form of Chou's PseAAC. In order to reduce computation complexity and feature abundance, we also use a two-step optimal feature selection process to find the optimal feature subset for each binary classification, which is different from conventional methods using the same features for all binary classifiers. In our system, all binary classifiers are constructed using support vector machine with probability output in the first stage (Chang and Lin, 2001). After this, the high dimensional feature vector of each protein is converted into a probability vector. At the second stage, conventional MSVM is used to construct final models. The performance of our method is evaluated on two benchmark datasets consisting of five types of membrane proteins. The overall accuracies of prediction for five types are 93.25% and 96.61% in the jackknife test and independent dataset test, respectively.

As demonstrated by a series of recent publications (Chen et al., 2013; Lin et al., 2013; Xiao et al., 2013a,b) and summarized in a comprehensive review (Chou, 2011) to establish a really useful statistical predictor for a protein system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the attribute to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) establish a user-friendly web-server for the predictor that is accessible to the public. Below, we will describe how to deal with these procedures in our method.

2. Materials and methods

2.1. Dataset

In this work, we use two popular benchmark datasets to evaluate the performance of our method, the training dataset and the independent dataset, which are both taken from Chou and Elrod (1999). The training dataset was constructed through a three-step screening procedure (Chou and Elrod, 1999) which finally contains 2059 protein sequences consisting of 435 type-I, 152 type-II, 1311 multi-pass transmembrane, 51 lipid-chain-anchored, and 110 GPI anchored membrane protein sequences. The independent dataset contains 2625 membrane protein sequences, which consists of 487 type I, 180 type II, 1867 multi-pass, 14 lipid-chain anchored, and 86 GPI anchored membrane proteins.

2.2. Feature extraction

To successfully use the proposed two-stage SVM as a powerful classifier, the key is how to effectively define a feature vector to formulate the statistical samples concerned. According to Eq. (6) of Chou (2011), the feature vector for any protein, peptide, or biological sequence is just a general form of pseudo-amino acid composition or PseAAC (Chou, 2001) that can be formulated as

$$\mathbf{P} = [\psi_1, \psi_2, \ldots, \psi_\mu, \ldots, \psi_\Omega]^{\mathbf{T}} \tag{1}$$

where $\mathbf{T}$ is the transpose operator, the components $\psi_1, \psi_2, \ldots, \psi_\mu, \ldots, \psi_\Omega$ will depend on how to extract the desired information from the statistical samples concerned, while the subscript Ω is an integer representing the dimension of the feature vector $\mathbf{P}$. In this work, we use various features extracted from amino acid classification-based methods and physicochemical property-based methods, and $\Omega = 7270$.

2.2.1. Methods based on amino acid classification

To capture more contextual information, we consider different amino acid classification approaches for local amino acid composition (*LAAC*) and local dipeptide composition (*LDC*) (Höglund et al., 2006), global descriptor (*GD*) (Dubchak et al., 1995; Yang et al., 2008) and Lempel–Ziv complexity (*LZC*) (Lempel and Ziv, 1976). Some of these amino acid classification approaches (Chou and Fasman, 1974; Dill, 1985; Basu et al., 1997; Murphy et al., 2000; Yu et al., 2004a,b; Shen et al., 2007; Alejandro et al., 2008; Li et al., 2008) are listed in Table 1.

The LAAC, LDC, GD, and LZC methods were also used in our previous study (Han et al., 2013), which gave the definitions and descriptions of these methods.

Twenty amino acids are divided into n groups, denoted by $\mathcal{A}$, according to certain classification method listed in Table 1. Note that we select one symbol (boldface letter in Table 1) for representing the group containing two or more amino acids, and this is convenient to generate the feature name. Then, for a given protein sequence S of length N, we may obtain a new sequence S' of n symbols with the same length as S, each symbol corresponding to one group of amino acids.

Using LAAC and LDC, $2 \times (n+n^2)$ features are generated. We get $6 \times n + n \times (n+1)/2$ features from the GD method for S'. Let $S^{'}_{i:j}$ be the subsequence of S' between positions i and j. The LZ complexity of sequence S', usually denoted by $c(S')$, is defined as the minimal number of steps with which S' is synthesized from null sequence according to the rule that at each step only two operations are allowed: either copying the longest fragment from the part of S' that has already been synthesized or generating an additional symbol.

According to Table 1, the number of features generated by LAAC, LDC, GD, and LZC is $4530(=278+2662+1569+21)$.

Table 1
Amino acid classifications.

Method	Num	Amino acid classification
HP (Dill, 1985)	2	(**A**LIMFPWV)(**D**ENCQGSTYRHK)
DHP (Yu et al., 2004a, 2004b)	4	(**A**LVIFWMP)(STYC**N**GQ)(K**R**H)(**D**E)
7-Cat (Shen et al., 2007)	7	(**A**GV)(**I**LFP)(**Y**MTS)(**H**NQW)(**R**K)(**D**E)C
20-Cat	20	AGVILFPYMTSHNQWRKDEC
ms (Alejandro et al., 2008)	6	(**A**VLIMC)(**W**YHF)(**T**QSN)(**R**K)(**E**D) (**G**P)
lesk (Alejandro et al., 2008)	6	(**A**ST)(**C**VILWYMPF)(HQ**N**)(**R**K)(E**D**)G
F-Ic4 (Alejandro et al., 2008)	7	(**A**WM)(**G**ST)(**H**PY)(**C**VIFL)(**D**NQ)(**E**R) K
F-Ic2 (Alejandro et al., 2008)	9	(**A**WM)(**G**S)(**H**PY)(**C**VI)(**F**L)(**D**NQ)(**E**R) KT
F-IIIc4 (Alejandro et al., 2008)	9	(**A**CV)(**H**PL)(**D**Q)S(**E**RGN)F(**I**MT)(**K**W) Y
F-Vc4 (Alejandro et al., 2008)	8	(**A**WHC)G(L**E**PV)(**K**YMT)(**I**N)QDS
Murphy8 (Murphy et al., 2000)	8	(**L**VMIC)(**A**G)(**S**T)P(**F**YW)(D**E**NQ)(**K**R) H
Murphy15 (Murphy et al., 2000)	15	(**L**VIM)CAGSTP(**F**Y)WEDNQ(**K**R)H
Letter6 (Chou and Fasman, 1974)	6	(V**I**M)(**C**YFQLTW)(**R**GD)(**H**KSNP)AE
Letter12 (Basu et al., 1997)	12	(LV**I**M)C(**A**G)(**S**T)P(F**Y**)W(**E**D)NQ(K**R**) H
Hydrophobicity (Li et al., 2008)	3	(RKEDQ**N**)(GAST**P**HY)(CLVIM**F**W)
NvdW (Li et al., 2008)	3	(**G**ASTPD)(**N**VEQIL)(**M**HKFRYW)
Polarity (Li et al., 2008)	3	(**L**IFWCMVY)(**P**ATGS)(**H**QRKNED)
Polarizability (Li et al., 2008)	3	(**G**ASDT)(**C**PNVEQIL)(**K**MHFRYW)
Charge (Li et al., 2008)	3	(**K**R)(AN**C**QGHILMFPSTWYV)(**D**E)
Secondary structure (Li et al., 2008)	3	(EALMQKR**H**)(VIY**C**WFT)(GNP**S**D)
Solvent accessibility (Li et al., 2008)	3	(AL**F**CGIVW)(PKQ**E**ND)(MPS**T**HY)

2.2.2. Methods based on physicochemical properties

In order to capture as much information of protein sequences as possible, the autocorrelation descriptor (*AD*) (Li et al., 2008), sequence-order descriptor (*SD*) (Li et al., 2008), and Hilbert–Huang transform (HHT) (Huang et al., 1998; Yu et al., 2010; Han et al., 2011) are used. Thirty physicochemical properties are utilized for AD and HHT, which can be found in AAindex1 of the Amino Acid index (AAindex) database (Kawashima and Kanehisa, 2000). Detailed information of these 30 physicochemical properties is listed in Table 2. For SD, we use two distance matrices for amino acid pairs. One is the *Grantham chemical distance matrix* (Li et al., 2008), and the other is the *Schneider–Wrede physicochemical distance matrix* (Chou, 2000).

In the present study, we only need the AD, SD and HHT methods in this simplified system. In our previous study (Han et al., 2013), we used three more methods: physicochemical property distribution descriptor (PPDD), recurrence quantification analysis (RQA), and discrete wavelet transform (DWT) for predicting protein subnuclear localizations. For the descriptions of AD, SD, and HHT methods, one can also refer to our previous paper (Han et al., 2013).

Three widely used autocorrelation descriptors are selected: normalized Moreau–Broto autocorrelation descriptors, Moran autocorrelation descriptors and Geary autocorrelation descriptors (Li et al., 2008). They are all defined based on the value distributions of 30 physicochemical properties of amino acids along an amino acid sequence (see Table 2). The measurement values of these properties are first standardized to have zero mean and unit standard deviation and then three autocorrelation descriptors are calculated. These descriptors are also used for the classification of G-protein-coupled receptors by Peng et al. (2010). For each AD, we obtain 600 $(=30 \times 20)$ features. In total, 1800 $(=600 \times 3)$ features are obtained to describe a protein sequence.

For SD, we end up with 60 $(=30 \times 2)$ sequence-order-coupling numbers and 100 $(=50 \times 2)$ quasi-sequence-order descriptors. In total, there are 160 features extracted from SD.

The HHT consists of two parts: empirical mode decomposition (EMD) and Hilbert spectral analysis (HSA). For each physicochemical property index selected, $4 \times 5+5+1=26$ features (here 5 is the number of IMFs selected in EMD) are obtained in HHT. In total, $780(=30 \times 26)$ features are obtained from HHT.

2.3. System construction

2.3.1. Support vector machine

Vapnik (1995) introduced the support vector machine (SVM) method to solve the binary classification problem. In order to solve a multi-class classification problem, such as the prediction of membrane protein types, the method must be extended. There are three notable extension strategies: one-against-all, one-against-one and directed acyclic graph SVM (DAGSVM) (Platt et al., 2000). In this paper, we adopted the one-against-one strategy. For a k classification problem, the SVM designed by the one-against-one strategy constructs $k \times (k-1)/2$ classifiers, each of which is trained on data from two different classes. Throughout, we used the radial basis kernel function (RBF) to find the optimal feature subsets and corresponding parameters C and γ by grid search.

Furthermore, we used a weighting scheme as in Blum et al. (2009) for each class in order to reduce the effect of over-prediction when using unbalanced training datasets. The weighting scheme assigns weight 1.0 to the largest class and higher weights to the remaining classes. The weights of these classes are simply calculated by dividing the size of the largest class by that of each smaller class.

Table 2
30 physicochemical properties of amino acids selected from AAindex database.

AAindex	Physicochemical property	Range of property
BULH740101	Transfer free energy to surface	[−2.46, 0.16]
BULH740102	Apparent partial specific volume	[0.558, 0.842]
PONP800102	Average gain in surrounding hydrophobicity	[10.53, 13.86]
PONP800104	Surrounding hydrophobicity in alpha-helix	[10.98, 14.08]
PONP800105	Surrounding hydrophobicity in beta-sheet	[11.79, 16.49]
PONP800106	Surrounding hydrophobicity in turn	[9.93, 15.00]
MANP780101	Average surrounding hydrophobicity	[11.36, 15.71]
EISD840101	Consensus normalized hydrophobicity scale	[−1.76, 0.73]
JOND750101	Hydrophobicity	[0.00, 3.15]
HOPT810101	Hydrophilicity value	[−3.4, 3.00]
PARJ860101	HPLC parameter	[−10.00, 10.00]
JANJ780101	Average accessible surface area	[22.8, 103.0]
PONP800107	Accessibility reduction ratio	[2.12, 7.69]
CHOC760102	Residue accessible surface area in folded protein	[18, 97]
ROSG850101	Mean area buried on transfer	[62.9, 224.6]
ROSG850102	Mean fractional area loss	[0.52, 0.91]
BHAR880101	Average flexibility indices	[0.295, 0.544]
KARP850101	Flexibility parameter for no rigid neighbors	[0.925, 1.169]
KARP850102	Flexibility parameter for one rigid neighbor	[0.862, 1.085]
KARP850103	Flexibility parameter for two rigid neighbors	[0.803, 1.057]
JANJ780102	Percentage of buried residues	[3, 74]
JANJ780103	Percentage of exposed residues	[5, 85]
LEVM780101	Normalized frequency of alpha-helix, with weights	[0.90, 1.47]
LEVM780102	Normalized frequency of beta-sheet, with weights	[0.72, 1.49]
LEVM780103	Normalized frequency of reverse turn, with weights	[0.41, 1.91]
GRAR740102	Polarity	[4.9, 13.0]
GRAR740103	Volume	[3, 170]
MCMT640101	Refractivity	[0.00, 42.35]
PONP800108	Average number of surrounding residues	[4.88, 7.86]
KYTJ820101	Hydropathy index	[−4.5, 4.5]

2.3.2. *Two-step optimal feature selection*

After the feature extraction procedure, all protein sequences are converted into numerical feature vectors with the same dimension. In order to reduce feature abundance and computation complexity, we propose a two-step optimal feature selection process by using an incremental feature selection (IFS) method (Huang et al., 2010; Han et al., 2013).

The IFS is based on the mRMR method originally proposed by Peng et al. (2005) for analyzing microarray data. The detailed information about the mRMR and IFS methods can be found in Peng et al. (2005) and Huang et al. (2010), respectively. In the first step, we consider each feature extraction method separately and construct corresponding models for each binary classification. Supposing that the number of feature extraction methods used is M, there are M optimal feature subsets constructed for each binary classification in this step. In the second step, for each binary classification, we extract the final optimal feature subset on the union of M optimal feature subsets obtained in the first step. We simultaneously find the optimal feature subset and the SVM parameters C and γ using 5-fold cross validation on the training dataset for each turn in the leave-one-out cross validation process.

2.3.3. *Two-stage support vector machine*

Finally, we construct a two-stage support vector machine to predict membrane protein types. In the first stage, $k\times(k-1)/2$ SVM classifiers with probability estimates are constructed based on the two-step optimal feature selection procedure. All optimal feature subsets and SVM parameters for $k\times(k-1)/2$ binary classifiers are simultaneously obtained by the two-step optimal feature selection procedure. We use LIBSVM for probability estimation as in Chang and Lin (2001). After this, each amino acid sequence is represented by a k-dimensional numerical vector, each element of which is the probability of the corresponding class to be predicted. The outputs of this stage are used as inputs for the next stage. In the second stage, we use conventional multi-class SVMs to predict membrane protein types. Here we use LIBSVM (Chang and Lin, 2001) to implement SVMs. It has been demonstrated using permutation analysis in Han et al. (2013) that the two-step optimal feature selection method and two-stage support vector machine are effective. The complete flow chart of the current method is illustrated in Fig. 1, which is slightly simpler than that shown in Figure 1 of Han et al. (2013) (PPDD, RQA, and DWT are removed). Note that if the leave-one-out cross validation is chosen to test this two-stage SVM, a different two-stage SVM is constructed for each turn of the leave-one-out cross validation.

2.4. *Performance evaluation*

In statistical prediction, three validation tests are often used to evaluate the prediction performance: independent dataset test, sub-sampling test and jackknife test. Among these three tests, the jackknife test was thought to be the most rigorous and objective one (Chou, 1995). Hence, we adopted the jackknife test in this paper. That is, each protein sequence in the samples is singled out in turn as a test sample and the remaining protein sequences are used as training samples. In this sense, the jackknife test is also known as the leave-one-out test.

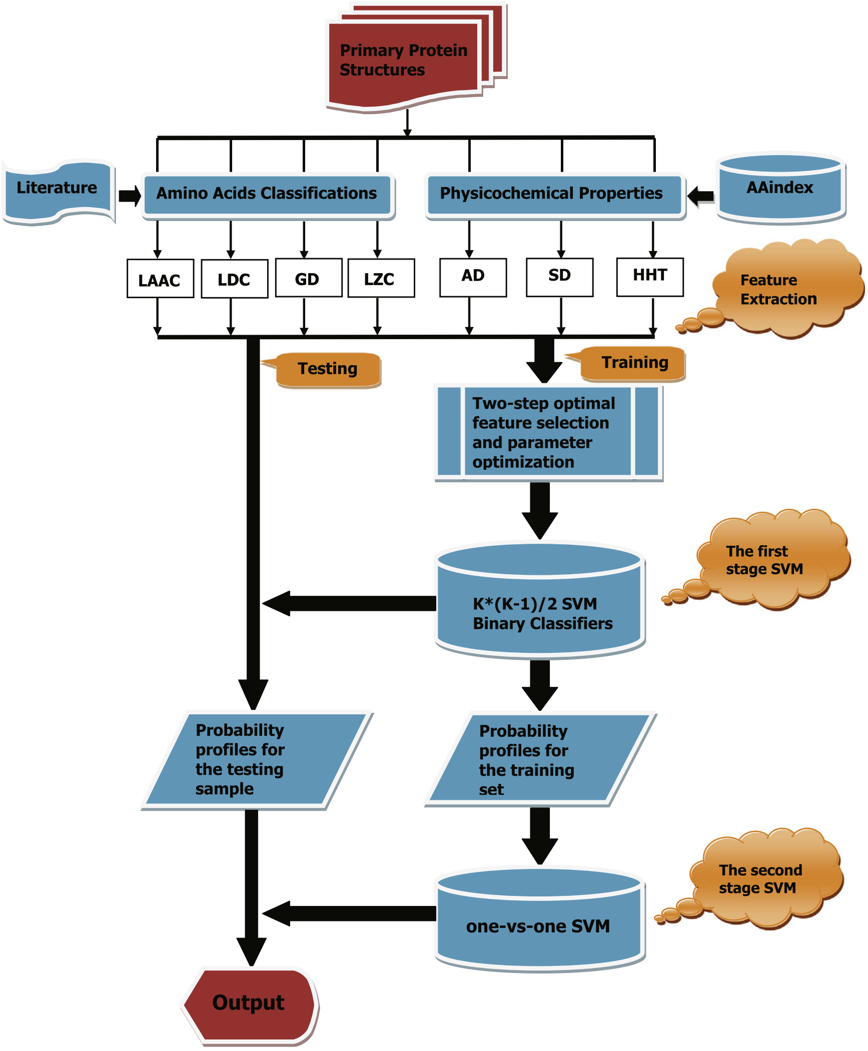
Primary Protein Structures
Literature
Amino Acids Classifications
Physicochemical Properties
AAindex
LAAC
LDC
GD
LZC
AD
SD
HHT
Feature Extraction
Testing
Training
Two-step optimal feature selection and parameter optimization
K*(K-1)/2 SVM Binary Classifiers
The first stage SVM
Probability profiles for the testing sample
Probability profiles for the training set
one-vs-one SVM
The second stage SVM
Output

Fig. 1. The complete flowchart of the proposed method.

The overall prediction accuracy A_c, individual sensitivity S_{in}, individual specificity S_{ip} and Matthew's correlation coefficient MCC_i are used to evaluate the prediction performance of our work as in Xu et al. (2013) and Chen et al. (2013). Their definitions are as follows:

$$S_{in} = TP_i/(TP_i + FN_i), \tag{2}$$

$$S_{ip} = TN_i/(TN_i + FP_i), \tag{3}$$

$$MCC_i = \frac{TP_i \times TN_i - FP_i \times FN_i}{\sqrt{(TP_i + FP_i) \times (TP_i + FN_i) \times (TN_i + FP_i) \times (TN_i + FN_i)}}, \tag{4}$$

$$A_c = \frac{\sum_i TP_i}{N}, \quad i = 1, 2, 3, \ldots, k, \tag{5}$$

where true positives *TP* is the number of positive events that are correctly predicted; true negatives *TN* is the number of negative events that are correctly predicted; false positives *FP* is the number of negative events that are incorrectly predicted to be positive; false negatives *FN* is the number of subjects that are predicted to be negative despite they are positive; *k* is the number of classes.

The above metrics are valid only for the single-label systems. For the multi-label systems whose existence has become more frequent in system biology (Chou et al., 2011, 2012) and system medicine (Chen et al., 2012; Xiao et al., 2013b), a different set of metrics as defined in Chou (2013) is needed.

3. Results and discussion

3.1. Effectiveness of simplified system

In this work, we simplify the system designed in our previous study (Han et al., 2013) removing three methods: PPDD, RQA and DWT. One may ask whether this simplified system is better? In order to show the effectiveness of the simplified system, we compare it with the old system using the jackknife test and independent dataset test. Detailed comparisons for five types are illustrated in Figs. 2 and 3. As seen from these two figures, in most cases, the simplified system is better than the old one. The overall prediction accuracies of the new system via jackknife test and independent dataset test are 93.25% and 96.61%, which are a little better than the old one (92.71% and 96.08%). This may be because these removed features are covered by other features. We should declare that our system is suitable for predicting the type of membrane proteins among the five types, but may not yield meaningful results for non-membrane proteins.

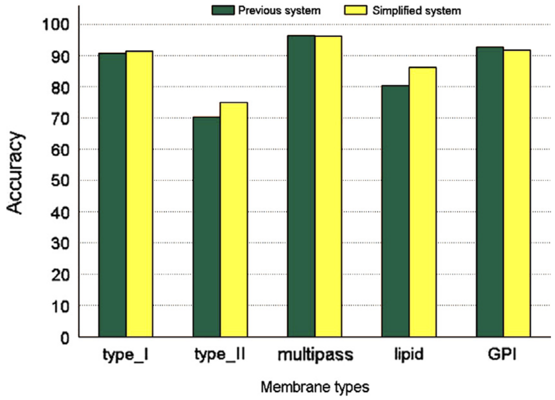

Fig. 2. Comparison for five types between the new system and the old one using jackknife test.

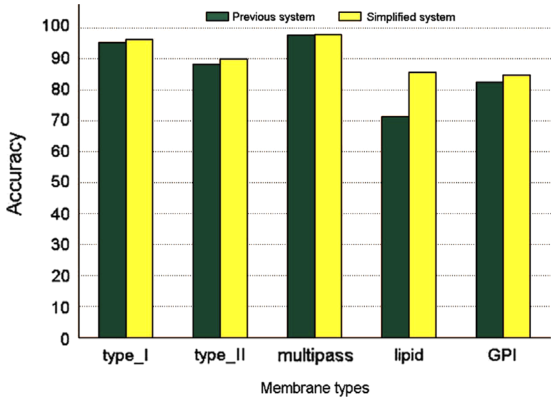

Fig. 3. Comparison for five types between the new system and the old one using independent dataset test.

3.2. Comparison of feature extraction methods

Now, we intend to perform each feature extraction method separately to see which method is more effective. During the training process, we used a grid search approach to find optimal feature subsets and optimize the SVM parameters C and γ using 5-fold cross validation for all binary classifications of each feature extraction method. It is found that the number of elements of the optimal feature subset for each binary classification is generally less than 300. So we chose the top-rank 300 features as the upper bound for optimal feature subset search. The top-rank 10 features are used as an initial feature subset. The size of the feature subset is increased by 10, obtaining 10, 20, 30,…, 300 features. At each size, we searched for a pair (C,γ) with the best 5-fold cross validation (e.g. $\log C = -5, -3, -1, \ldots, 15; \log \gamma = -15, -13, -11, \ldots, 3$). From this process, each binary classification corresponds to an optimal feature subset and a parameter pair (C,γ). Thus we can construct all binary classification models and make preparation for training the second stage model. The training method for the second stage model is identical to the first stage except that it does not need feature selection. The final prediction system can be constructed as follows: the training dataset is used; the optimal feature subsets for each binary classification are taken as the union of all optimal feature subsets obtained from the leave-one-out cross validation; and the optimal value for each parameter of the SVMs for the training set is taken as the average value of the optimal parameters obtained from the leave-one-out cross validation. The final system is then tested on the independent dataset. Note that all parameters of the final system including optimal features and SVM parameters are not re-parameterized to apply on the independent dataset.

Table 3
Comparison of the overall prediction accuracy between different feature extraction methods on the training dataset and the independent dataset.

Method	Jackknife test (%)	Independent dataset test (%)
Combination1	89.46	90.82
HHT	86.16	92.19
Combination2	93.25	96.61

We calculated the overall prediction accuracies for all feature extraction methods on the training dataset and the independent dataset. As far as the individual feature extraction methods are concerned, we found that the HHT method is the best. We combine the feature extraction methods LAAC, LDC, GD, LZC, AD, SD as one method, named *Combination1*. In the following, the values on the independent dataset are shown in the parentheses. From Table 3, the overall prediction accuracies of HHT are 86.16% (92.19%). The overall prediction accuracy on the training dataset is worse than that of *Combination1* (89.46%). But, for the independent dataset, the overall prediction accuracy of the HHT method is better than that of *Combination1* (90.82%). Finally, we evaluate the overall prediction accuracies of the combination of all feature extraction methods, named *Combination*2. As shown in Table 3, *Combination*2 achieves the overall prediction accuracy of 93.25% (96.61%), with accuracy increases against individual methods between 3.78% (4.42%) and 7.09% (5.79%).

3.3. Assessment of the reliability of two-step optimal feature selection by permutation analysis

In order to evaluate the reliability of the two-step optimal feature selection method, two kinds of randomization studies are performed for each binary classification as in Han et al. (2013). They are given the number K, randomly select K features from original features (case 1) or suboptimal features (case 2) of the samples from two different membrane protein types, while keeping the class memberships unchanged. Then the newly generated feature set is analyzed by using the same five-fold cross validation as applied before to the original feature set. Here, the given numbers of features K are set as one fourth, half or all of the number of optimal features. This procedure for case 2 is carried out 50 rounds and the error rates (±standard deviation) over 50 permutations are shown in Fig. 4, and compared with the minimum error rates obtained from optimal features. For case 1, similar results are obtained. In each case, the estimated error rate obtained by optimal features is significantly lower than that obtained by the randomization study. Especially when we compare the experimental results for case 1 and 2, the misclassification error rates obtained by using features selected randomly from suboptimal features are also much lower than that estimated by using those from the original features. If we perform these two randomization analysis on the whole original feature set 50 times, the overall error rates on average are 43.24% (5.23%) and 20.56% (1.80%), which are both significantly higher than the error rate 7.92% obtained by optimal features. Therefore, it can be concluded that the two-step optimal feature selection method is effective and reliable.

3.4. Analysis of optimal feature

Since we select different optimal feature subsets for each binary classification, we have 10 final selected feature subsets corresponding to 10 binary classifications. Top 10 features for each binary classification are listed in Table 4. Fragments of all feature names are linked by underline. For LAAC and LDC, the template of feature names is *(name of amino acid classification)_(LAAC or LDC)_(C or N terminal)_(amino acid or dipeptide)*. For GD, there are four templates, namely, *(name of amino acid classification)_(GD)_(COMP)_(amino acids)*, *(name of amino acid classification)_(GD)_(TRAN)_(dipeptide)_(reverse dipeptide)*, *(name of amino acid classification)_(GD)_(DIST)_(amino acids)_(percentage)* and *(name of amino acid classification)_(GD)_(EDP)_(amino acids)*. For LZC, the template is *(name of amino acid classification)_(LZ)*. For AD, the template is *(index of physicochemical property)_(AD)_(NMB or M or G)_(distance)*. For HHT, the template is *(index of physicochemical property)_(HHT)_(MAX, Min, AVG, STD, or ENG)_(IMF)_(number)* or is *(index of physicochemical property)_(HHT)_(SE)*. For SD, the template is *(SD)_(Tau)_(SW or GR)_(tau)* or is *(SD)_(QSO)_(SW or GR)_(amino acids or tau)*. From Table 4, we can find that the distribution of features for different binary classifications is significantly different. This suggests that our idea works well to determine which different optimal feature subsets should be constructed for different binary classifications. Surprisingly, for all binary classifications about the GPI-anchored membrane type (i.e., I vs G, II vs G, M vs G, and L vs G), the features from LAAC and LDC on C-terminal constitute most of the top 10 features. It is verified that LAAC and LDC methods can represent the signal for GPI-anchoring to some extent, which is confined to be C-terminal of the target protein. Except these four binary classifications, some HHT features appear in the list of top 10 features of other binary classifications, although most features are from GD. Especially, HHT features are effective for identifying multipass type, which may be due to the special topology of multipass transmembrane proteins.

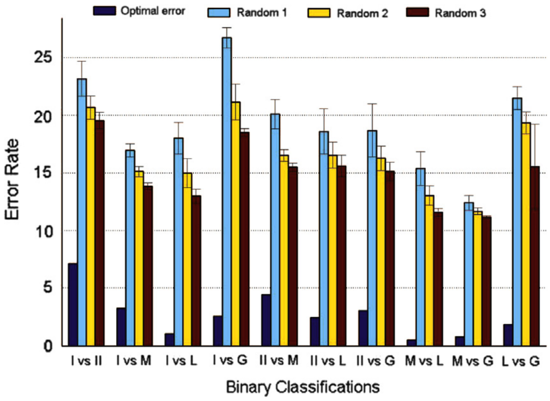

Fig. 4. Comparisons of error rate (percentage of misclassified samples) over 50 runs of randomization analysis. Random 1: selecting randomly features subsets from suboptimal features, whose size is one-fourth of the number of optimal features; Random 2: one half of the number of optimal features; Random 3: equal to the number of optimal features. The upper letters I, II, M, L and G correspond to five membrane protein types, namely, type I, type II, multipass, lipid-chain-anchored, and GPI-anchored, respectively.

3.5. Comparison with the existing methods

The detailed performances of our method (*Combination*2) on the training dataset and the independent dataset are illustrated in Table 5 and a comparison of the performance of this method against other existing methods is illustrated in Table 6, where better results are highlighted in bold. As shown in Table 5, *Combination*2 achieves highest prediction results on type I and multipass transmembrane proteins, which are 91.49% (96.64%) and 96.34% (97.91%), respectively. This may be due to large sample sizes of these two classes in the data. However, the training dataset used is highly unbalanced. The lipid-chain-anchored type only contains 51 samples, while the multi-pass type contains 1311 samples. So the model trained on this benchmark dataset may cause bias to large-size types, although we use a weight scheme. It can be seen from Table 5 that the results for small-size types, such as type-II and lipid, are not as good as the other types. Other more effective strategies for solving imbalance problem may contribute to improving our method.

Table 4
Top 10 features for all binary classifications.

Binary	Top 10 features
I vs II	MS_GD_DIST_W_1, polarizability_GD_COMP_K, 7-cat_GD_DIST_H_50, F-Vc4_GD_DIST_K_75, Letter12_LAAC_C_N, MANP780101_HHT_ENG_4, Letter6_GD_DIST_C_50, Murphy15_GD_EDP_F, Murphy8_GD_DIST_S_75, SD_QSO_SW_23
I vs M	F-Ic4_GD_TRAN_AC_CA, BULH740101_AD_NMB_4, Murphy8_GD_TRAN_PE_EP, F-IIIc4_GD_TRAN_AF_FA, Letter20_GD_EDP_M, KYTJ820101_AD_NMB_7, BULH740101_HHT_STD_IMF_2, GRAR740102_AD_G_4, Letter12_GD_TRAN_YA_AY, DHP_GD_TRAN_DN_ND
I vs L	Letter12_GD_TRAN_IW_WI, Letter12_LDC_N_AW, SD_QSO_SW_K, KYTJ820101_HHT_MAX_IMF_5, polarizability_GD_COMP_C, SD_QSO_GR_N, MS_GD_DIST_W_1, Murphy15_GD_TRAN_SH_HS, JANJ780103_HHT_ENG_IMF_2, Letter20_LDC_N_MK
I vs G	sol-acc_LAAC_C_F, F-IIIc4_LAAC_C_E, hydro_LDC_C_FF, MS_LDC_C_AA, HP_LAAC_C_A, Letter20_LAAC_C_L, charge_LAAC_C_C, polarity_LAAC_C_L, Murphy8_LAAC_C_E, Letter12_LDC_C_II
II vs M	SD_QSO_GR_4, SD_QSO_SW_K, 7-cat_GD_TRAN_AY_YA, HOPT810101_HHT_ENG_IMF_1, Murphy8_GD_DIST_S_50, JANJ780103_HHT_ENG_IMF_1, HOPT810101_HHT_STD_IMF_3, Letter6_GD_DIST_R_50, KARP850102_AD_G_4, 7-cat_GD_COMP_H
II vs L	Letter12_LDC_N_AW, EISD840101_HHT_STD_IMF_5, F-IIIc4_GD_COMP_H, Letter20_LDC_N_MK, hydro_GD_TRAN_NP_PN, MS_LAAC_N_W, 7-cat_GD_DIST_D_50, Murphy15_LDC_N_GC, Murphy8_GD_EDP_H, GRAR740102_AD_M_1
II vs G	HP_LDC_C_AA, Letter20_LAAC_C_L, sol-acc_LDC_C_EF, MS_C_LAAC_A, 7-cat_LDC_C_II, MS_LDC_C_EA, DHP_LAAC_C_R, F-Ic4_LDC_C_CC, hydro_GD_EDP_P, sol-acc_LAAC_C_F
M vs L	SD_QSO_GR_K, Letter12_LDC_N_AW, DHP_GD_DIST_A_25, ROSG850101_AD_NMB_11, KARP850103_AD_NMB_16, JOND750101_HHT_MAX_IMF_5, Letter12_GD_TRAN_IW_WI, SD_QSO_GR_D, Letter20_GD_TRAN_DK_KD, KYTJ820101_AD_M_4
M vs G	MS_LAAC_C_A, F-IIIc4_LAAC_C_H, BULH740101_HHT_STD_IMF_2, charge_LDC_C_CC, HP_GD_DIST_A_50, F-Ic4_LDC_C_CC, F-IIIc4_GD_COMP_F, HP_LAAC_C_D, Letter20_GD_TRAN_AI_IA, Letter20_LDC_C_LL
L vs G	DHP_LAAC_C_D, 7-cat_LAAC_C_I, Letter12_LAAC_C_Q, Murphy8_LDC_C_LL, LESK_LDC_C_CD, 7-cat_GD_DIST_C_100, F-IIIc4_LAAC_C_E, polarity_LDC_C_HH, charge_LAAC_C_C, hydro_LDC_C_PN

Table 5
Detailed classification performances for the jackknife and independent dataset tests using our method *Combination*2.

Types	Size	Jackknife test			Independent dataset test		
		S_n	S_p	MCC	S_n	S_p	MCC
Type-I	435(487)	91.49	97.63	0.89	96.44	99.00	0.95
Type-II	152(180)	75.00	97.89	0.73	90.00	98.71	0.86
Multipass	1311(1867)	96.34	93.06	0.89	97.91	95.93	0.94
Lipid	51(14)	86.27	99.79	0.89	85.71	100	0.93
GPI	110(86)	91.82	99.45	0.91	84.88	99.72	0.88
OA		93.25			96.61		

Table 6
Comparison of the overall prediction accuracy with existing approaches.

Method	Jackknife test (%)	Independent dataset test (%)
CDA (Chou and Elrod, 1999)	76.4	79.4
CDA and PseAA (Chou, 2001)	80.9	87.5
AA composition (Cai et al., 2004)	80.4	85.4
Fourier spectrum (Liu et al., 2005)	78.0	87
PseAA (Wang et al., 2004)	82.3	90.3
Stacking (Wang et al., 2006)	85.4	94.3
Wavelet (Rezaei et al., 2008)	81.3	91.4
Discrete wavelet (Qiu et al., 2010)	78.1	–
Dipeptide (Wang et al., 2010)	82.0	90.1
CPSR (Hayat and Khan, 2011)	86.0	95.2
MemHyb (Hayat and Khan, 2012a)	91.1	–
Combination2	**93.25**	**96.61**

As shown in Table 6, *Combination*2 achieves better performance on all types of two datasets compared with several existing methods. *Combination*2 significantly improves performances of the jackknife test. Interestingly, it can be seen from Tables 3 and 6 that HHT outperforms Fourier spectrum (Liu et al., 2005) and wavelet method (Rezaei et al., 2008; Qiu et al., 2010), implying that HHT is more effective. Note that the HHT and wavelet method are both time–frequency analysis methods and use similar definitions of statistical features. Overall, the results show that our method has good generalization abilities in predicting membrane protein types regardless of the size of types.

Furthermore, comparing with Hayat and Khan (2012a), our method only uses information on amino acids of the protein sequence, and do not use non-sequence information such as evolutionary information (e.g. PSI-BLAST profile), which makes our method more general since the PSI-BLAST profile is computationally more intensive than sequence-based methods, and is also time consuming for training on a novel query sequence.

However, although our method makes encouraging identification among the five types on two benchmark datasets, there are still some proteins that fail to be grouped into the correct types. In addition to inaccuracy in predicting small-size types, some hidden complicated signals cannot be captured by our model. Therefore, some other good features should be incorporated to make potential improvement for our system.

4. Conclusions

In this work, we not only proposed a novel feature extraction method based on HHT but also established a two-stage SVM system by incorporating amino acid classifications and physicochemical properties into the general form of Chou's PseAAC to predict membrane protein types. Numerical results show that this method is more effective than the other individual feature extraction methods under consideration. The results further demonstrate that combination of the feature extraction methods significantly increases the performance compared with individual methods. The reason may be that diverse amino acid classifications and physicochemical properties can capture more contextual and environmental information of the protein sequence, and different feature extraction methods effectively integrate multidirectional information of the protein sequence. Furthermore, our method also has better generalization abilities than non-sequence information-based methods such as MemHyb (Hayat and Khan, 2012a). Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors (Chou and Shen, 2009; Lin and Lapointe, 2013), here we have provided a web-server for the method presented in this paper at http://www.juemengt.com/jcc/memty_page.php.

Acknowledgments

We thank all anonymous reviewers for providing constructive suggestions and comments that improved our manuscript. This project was supported by the Natural Science Foundation of China (Grant nos. 11071282 and 11371016), the Chinese Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) (Grant no. IRT1179), the Research Foundation of Education Commission of Hunan Province of China (Grant no. 11A122), Science and Technology Planning Project of Hunan Province of China (Grant no. 2011FJ2011), the Lotus Scholars Program of Hunan Province of China, the Aid program for Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province of China, and the Australian Research Council (Grant no. DP0559807).

References

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

Alejandro, S., Ernesto, P., Segovia, L., 2008. Protein homology detection and fold inference through multiple alignment entropy profiles. Proteins 70, 248–256.

Basu, S., Pan, A., Dutta, C., Das, J., 1997. Chaos game representation of proteins. J. Mol. Graph. Model. 15, 279–289.

Blum, T., Briesemeister, S., Kohlbacher, O., 2009. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. BMC Bioinforma. 10, 274.

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.

Cai, Y.D., Zhou, G.P., Chou, K.C., 2003. Support vector machines for predicting membrane protein types by using functional domain composition. Biophys. J. 84, 3257–3263.

Chang, C.C., Lin, C.J., 2001. LIBSVM: A Library for Support Vector Machines. ⟨http://www.csie.ntu.edu.tw/∼cjlin/papers/libsvm.pdf⟩.

Chen, L., Zeng, W.M., Cai, Y.D., Feng, K.Y., Chou, K.C., 2012. Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical–chemical interactions and similarities. PLoS ONE 7, e35254.

Chen, W., Feng, P.M., Lin, H., Chou, K.C., 2013. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res. 41, e69, Open access at http://dx.doi.org/10.1093/nar/gks1450.

Chen, Y.K., Li, K.B., 2013. Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition. J. Theor. Biol. 318, 1–12.

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

Chou, K.C., 2000. Prediction of protein subcellar locations by incorporating quasi-sequence-order effect. Biochem. Biophys. Res. Commun. 278, 477–483.

Chou, K.C., 2001. Prediction of protein subcellular attributes using pseudo-amino acid composition. Proteins: Struct. Funct. Genet. 43, 246–255.

Chou, K.C., 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21, 10–19.

Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition (50th anniversary year review). J. Theor. Biol. 273, 236–247.

Chou, K.C., 2013. Some remarks on predicting multi-label attributes in molecular biosystems. Mol. Biosyst. 9, 1092–1100.

Chou, K.C., Cai, Y.D., 2005. Using GO-PseAA predictor to identify membrane proteins and their types. Biochem. Biophys. Res. Commun. 327, 845–847.

Chou, K.C., Elrod, D.W., 1999. Prediction of membrane protein types and subcellular location. Proteins: Struct. Funct. Genet. 34, 137–153.

Chou, P.Y., Fasman, G.D., 1974. Prediction of protein conformation. Biochemistry 13, 222–245.

Chou, K.C., Shen, H.B., 2007. MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through PsePSSM. 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. 2, 63–92.

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 accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol. Biosyst. 8, 629–641.

Dill, K.A., 1985. Theory for the folding and stability of globular proteins. Biochemistry 24, 1501–1509.

Dubchak, I., Muchanikt, I., Holbrook, S.R., Kim, S.H., 1995. Prediction of protein folding class using global description of amino acid sequence. Proc. Natl. Acad. Sci. 92, 8700–8704.

Feng, Z.P., Keizer, D.W., Stevenson, R.A., Yao, S., Babon, J.J., Murphy, V.J., Anders, R.F., Norton, R.S., 2005. Structure and inter-domain interactions of domain II from the blood-stage malarial protein, apical membrane antigen 1. J. Mol. Biol. 350, 641–656.

Feng, Z.P., Zhang, X., Han, P., Arora, N., Anders, R.F., Norton, R.S., 2006. Abundance of intrinsically unstructured proteins in *P. falciparum* and other apicomplexan parasite proteomes. Mol. Biochem. Parasitol. 150, 256–267.

Gao, Q.B., Ye, X.F., Jin, Z.C., He, J., 2010. Improving discrimination of outer membrane proteins by fusing different forms of pseudo amino acid composition. Anal. Biochem. 398, 52–59.

Han, G.S., Yu, Z.G., Anh, V., 2011. Predicting the subcellular location of apoptosis proteins based on recurrence quantification analysis and the Hilbert–Huang transform. Chin. Phys. B 20, 100504.

Han, G.S., Yu, Z.G., Anh, V., Krishnajith, A.P.D., Tian, Y.C., 2013. An ensemble method for predicting subnuclear localizations from primary protein structures. PLoS ONE 8 (2), e57225.

Hayat, M., Khan, A., 2011. Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. J. Theor. Biol. 271, 10–17.

Hayat, M., Khan, A., 2012a. MemHyb: predicting membrane protein types by hybridizing SAAC and PSSM. J. Theor. Biol. 292, 93–102.

Hayat, M., Khan, A., 2012b. Discriminating outer membrane proteins with fuzzy K-nearest neighbor algorithms based on the general form of Chou's PseAAC. Protein Pept. Lett. 19, 411–421.

Höglund, A., Dönnes, P., Blum, T., Adolph, H.W., Kohlbacher, O., 2006. MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. Bioinformatics 22, 1158–1165.

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.

Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, S.H., Zheng, Q., Yen, N.C., Tung, C.C., Liu, H.H., 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series analysis. Proc. R. Soc. A 454, 903–995.

Huang, T., Shi, X.H., Wang, P., He, Z.S., Feng, K.Y., Hu, L.L., Kong, X.Y., Li, Y.X., Cai, Y.D., Chou, K.C., 2010. Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks. PLoS ONE 5, e10972.

Kawashima, S., Kanehisa, M., 2000. AAindex: amino acid index database. Nucleic Acids Res. 28 374–374.

Lempel, A., Ziv, J., 1976. On the complexity of finite sequence. IEEE Trans. Inf. Theory. 22, 75–81.

Li, Z.R., Lin, H.H., Han, L.Y., Jiang, L., Chen, X., Chen, Y.Z., 2008. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res. 34, W32–W37.

Lin, H., 2008. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J. Theor. Biol. 252, 350–356.

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

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.

Liu, H., Wang, M., Chou, K.C., 2005. Low-frequency Fourier spectrum for predicting membrane protein types. Biochem. Biophys. Res. Commun. 336, 737–739.

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

Mahdavi, A., Jahandideh, S., 2011. Application of density similarities to predict membrane protein types based on pseudo-amino acid composition. J. Theor. Biol. 276, 132–137.

Murphy, L.R., Wallqvist, A., Levy, R.M., 2000. Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng. 13, 149–152.

Nanni, L., Lumini, A., 2008. An ensemble of support vector machines for predicting the membrane protein type directly from the amino acid sequence. Amino Acids 35, 573–580.

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 27, 1226–1238.

Peng, Z.L., Yang, J.Y., Chen, X., 2010. An improved classification of G-proteincoupled receptors using sequence-derived features. BMC Bioinformatics 11, 420.

Platt, J.C., Cristianini, N., Shawe-Taylor, J., 2000. Large margin DAGs for multiclass classification. Adv. Neural Inf. Process. Syst. 12, 547–553.

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

Qiu, J.D., Sun, X.U., Huang, J.H., Liang, R.P., 2010. Prediction of the types of membrane proteins based on discrete wavelet transform and support vector machines. Protein J. 29, 114–119.

Rezaei, M.A., Maleki, P.A., Karami, Z., Asadabadi, E.B., Sherafat, M.A., Moghaddam, K.A., Fadaie, M., Forouzanfar, M., 2008. Prediction of membrane protein types by means of wavelet analysis and cascaded neural network. J. Theor. Biol. 255, 817–820.

Sanders, P.R., Kats, L.M., Drew, D.R., O'Donnell, R.A., O'Neill, M., Maier, A.G., Coppel, R.L., Crabb, B.S., 2006. A set of glycosylphosphatidyl inositol-anchored membrane proteins of *Plasmodium falciparum* is refractory to genetic deletion. Infect. Immun. 74, 4330–4338.

Shen, H.B., Chou, K.C., 2005. Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. Biochem. Biophys. Res. Commun. 334, 288–292.

Shen, H.B., Chou, K.C., 2007. Using ensemble classifier to identify membrane protein types. Amino Acids 32, 483–488.

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.

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., Jiang, H., 2007. Predicting protein–protein interactions based only on sequences information. Proc. Natl. Acad. Sci. 104, 4337–4341.

Tusnady, G.E., Dosztanyi, Z., Simon, I., 2004. Transmembrane proteins in the Protein DataBank: identification and classification. Bioinformatics 20, 2964–2972.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer.

Wang, J.Y., Li, Y.P., Wang, Q.Q., You, X.G., Man, J.J., Wang, C., Gao, X., 2012. ProClusEnsem: predicting membrane protein types by fusing different modes of pseudo amino acid composition. Comput. Biol. Med. 42, 564–574.

Wang, L., Yuan, Z., Chen, X., Zhou, Z., 2010. The prediction of membrane protein types with NPE. IEICE Electron. Express 7, 397–402.

Wang, M., Yang, J., Liu, G.P., Xu, Z.J., Chou, K.C., 2004. Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition. Protein Eng. Des. Sel. 17, 509–516.

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

Wang, S.Q., Yang, J., Chou, K.C., 2006. Using stacking generalization to predict membrane protein types based on pseudo amino acid composition. J. Theor. Biol. 242, 941–946.

Xiao, X., Min, J.L., Wang, P., Chou, K.C., 2013a. iGPCR-Drug: a web server for predicting interaction between GPCRs and drugs in cellular networking. PLoS ONE 8, e72234.

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

Xu, Y., Ding, J., Wu, L.Y., Chou, K.C., 2013. iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PLoS ONE 8, e55844.

Yang, X.G., Luo, R.Y., Feng, Z.P., 2007. Using amino acid and peptide composition to predict membrane protein types. Biochem. Biophys. Res. Commun. 353, 164–169.

Yang, J.Y., Zhou, Y., Yu, Z.G., Anh, V., Zhou, L.Q., 2008. Human Pol II promoter recognition based on primary sequences and free energy of dinucleotides. BMC Bioinformatics 9, 11.

Yu, Z.G., Anh, V., Lau, K.S., 2004a. Fractal analysis of measure representation of large proteins based on the detailed HP model. Physica A 337, 171–184.

Yu, Z.G., Anh, V., Lau, K.S., 2004b. Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses. J. Theor. Bol. 226, 341–348.

Yu, Z.G., Anh, V., Wang, Y., Mao, D., Wanliss, J., 2010. Modelling and simulation of the horizontal component of the geomagnetic field by fractional stochastic differential equations in conjunction with empirical mode decomposition. J. Geophys. Res. 115, A10219.