Amino Acids (2011) 40:991–1002
DOI 10.1007/s00726-010-0727-8

ORIGINAL ARTICLE

Accurate prediction of the burial status of transmembrane residues of α-helix membrane protein by incorporating the structural and physicochemical features

Chengqi Wang · Shuyan Li · Lili Xi · Huanxiang Liu · Xiaojun Yao

Received: 15 February 2010 / Accepted: 13 August 2010 / Published online: 26 August 2010

Abstract Predicting the burial status (the residue exposure to the lipid bilayer or buried within the protein core) of transmembrane (TM) residues of α-helix membrane protein (αHMP) is of great importance for genome-wide annotation and for experimental researchers to elucidate diverse physiological processes. In this work, we developed a new computational model that can be used for predicting the burial status of TM residues of αHMP. By incorporating physicochemical scales and conservation index, an efficient prediction model using least squares support vector machine (LS-SVM) was developed. The model was developed from 43 protein chains and its prediction ability was evaluated by an independent test set of other non-redundant ten protein chains. The prediction accuracy of our method was much better than the results of the reported works. Our results demonstrate that the LS-SVM prediction model incorporating structural and physicochemical features derived from sequence information could greatly improve the prediction accuracy.

Keywords Burial status of transmembrane residues · Recursive feature elimination (RFE) · Least squares support vector machine (LS-SVM)

C. Wang · S. Li · L. Xi · X. Yao
Department of Chemistry, State Key Laboratory of Applied Organic Chemistry, Lanzhou University, Lanzhou 730000, China

H. Liu
School of pharmacy, Lanzhou University, Lanzhou 730000, China

X. Yao (✉)
Key Lab of Preclinical Study for New Drugs of Gansu Province, Lanzhou University, Lanzhou 730000, China
e-mail: xjyao@lzu.edu.cn

Introduction

α-helix membrane protein (αHMP) comprising 20–30% of all proteome play a crucial role in executing biologically importance functions (von Heijine 1999; Liu et al. 2002). These functions include transporting ions and molecules into and out of cells, recognizing the immune system and energy transducers and possessing a characteristic of nature such as membrane adhesion, catalytic activity, and receptors of signal molecules. Despite the biological importance of α-helix membrane proteins, only less than 1% of the structures of this type of proteins are known due to the problems associated with the purification and availability in stable forms suitable for X-ray crystallography and electron microscopy (EM) studies.

Given this circumstance, it is highly desirable to develop sequence-based computational methods to predict the structure of αHMP. Recently, several computational methods for predicting structural aspect of integral membrane protein were proposed (Bernsel et al. 2008; Cao et al. 2006; Chen et al. 2009; Granseth et al. 2006; Jones 2007; Nanni and Lumini 2008; Page et al. 2009; Shen and Chou 2007; Viklund and Elofsson 2004; Yang et al. 2010). Most of these works focus on transmembrane (TM) segment, membrane protein type, or secondary structure prediction. However, because the insertion of transmembrane helices into the membrane is the outcome of molecular interactions among protein, lipids, and water, this requires not only the information of the TM domains of α-helix membrane protein, but also the burial status (exposure of the residue to the lipid bilayer, or to the burial of the residue within the protein core) of each residue of TM domains in order to get the structure of αHMP (Visiers et al. 2002). On the other hand, the knowledge about the burial status of all residues in TM domain is very useful in several cases. For example,

burial status of TM residue would be helpful to design mutation experiments that aimed at identifying catalytically important TM residues (Abramson et al. 2003; Huang et al. 2003; Palczewski et al. 2000; Pebay-Peyroula et al. 1997; Poolman et al. 2007; Sui et al. 2001). Therefore, developing reliable computational method to predict burial status of TM residue is highly desirable to elucidate diverse physiological processes.

Based on knowledge-based propensity scale of the 20 amino acids to be exposed to the membrane, Beuming and Weinstein firstly proposed sequence-based computational method for predicting the burial status of TM residues of α-helix membrane protein (denoted hereafter as the BeWe method) (Beuming and Weinstein 2004). Using the empirical scale based on the canonical helical face model derived from the heptad repeat motif, Adamian and Liang proposed an approach for prediction of helix-lipid interfaces of TM helices (Adamian and Liang 2006). But, this method cannot be used to predict the burial status of the residue located outside of a helix–helix interface. Recently, Yuan (denoted hereafter as the Yuan method) (Yuan et al. 2006) and Illergard (Illergard et al. 2010) utilized support vector machines to predict the relative solvent-accessible surface area (rSASA) of TM residues. These two methods can be used to indirectly predict burial status by using rSASA values, but they cannot be used to predict the burial status directly. More successful burial status prediction method (named as TMX) was built using SVM method that employed conservation patterns. Using conservation patterns, the TMX method achieved a better prediction accuracy of 78.71% on the benchmark (Park et al. 2007). Since it is often difficult to directly compare performance values of different prediction methods reported in different studies, Park and his coworkers evaluated the BeWe method and Yuan method on the same data set as for TMX method. Although three approaches have been proposed to predict burial status, the highest total prediction accuracy is only 78.71%. The performance of these models still needs improvement for the prediction of the membrane protein structures. On the other hand, because the insertion of TM helices into the membrane is the outcome of molecular interactions among protein, lipids and water, it should be possible to predict burial status by methods based on physicochemical features (Bernsel et al. 2008).

In the present study, by using experimental physicochemical scales and conservation index, a simple sequence-based prediction method was proposed. The main process of our prediction method is shown in Fig. 1. First, the sequence information contained in the immediate neighbors of the central residues was extracted by using sliding window. Then, two strategies were used for feature generation to encode the window. The main features used include the conservation index which has already proved to be very important in burial status prediction (Beuming and Weinstein 2004) and structural and physicochemical features got from experimental scales (Bhaskaran and Ponnuswamy 1988; Bigelow 1967; Charton 1981; Charton and Charton 1982; Cid et al. 1992; Kawashima et al. 2008). The features that highly correlated with burial status were then selected using recursive feature elimination (RFE) method. At last, least squares support vector machine (LS-SVM) was used to develop classification model due to its good performance and less time-consuming characteristic in the classification model development. The prediction model is developed based on a development set of 43 protein chains by using a large number of physicochemical features. An independent test set containing ten non-redundant protein chains were used to evaluate the prediction ability of the prediction model. By analyzing the prediction model, we could identify some physicochemical features that could help us understand the main factors influencing the burial status of TM residues. On the other hand, one of the most important advantages of our method than other reported works is that it requires only the information from sequence and is especially useful for the cases with large number of samples.

Materials and methods

Data sets

In this work, we used the same benchmark data set as TMX method (Park et al. 2007) to develop the prediction model. This data can be used to directly compare the performance of different prediction methods. This data set contained more samples than the former datasets used in the works of BeWe and Yuan. 43 protein chains with less than 25% pairwise identify and a resolution better than 3.0 Å were used in our study. A detailed list of α-helix membrane proteins with known structure was supplied by White (http://blanco.biomol.uci.edu/Membraneproteinsxtal.html) and Michel (http://www.mpibpfrankfurt.mpg.de/michel/public/memprotstruct.html). The data sets can be downloaded from (http://service.bioinformatik.unisaarland.de/tmx) (Park et al. 2007). On the other hand, to further evaluate the prediction ability of our prediction model, we used an independent test set containing ten helical membrane proteins (sequence identity less than 30% pairwise and resolution better than 3.0 Å) as the external test set. This data set is extracted from the MPtopo database (Jayasinghe et al. 2001). These ten 3D structures of helical proteins were published in 2008 or later and proved to be non-redundant (sequence identity less than 30%) with the development set using ClustalW (Thompson et al. 1994). The protein samples used in the development set and the independent test set are listed in Tables 1 and 2.

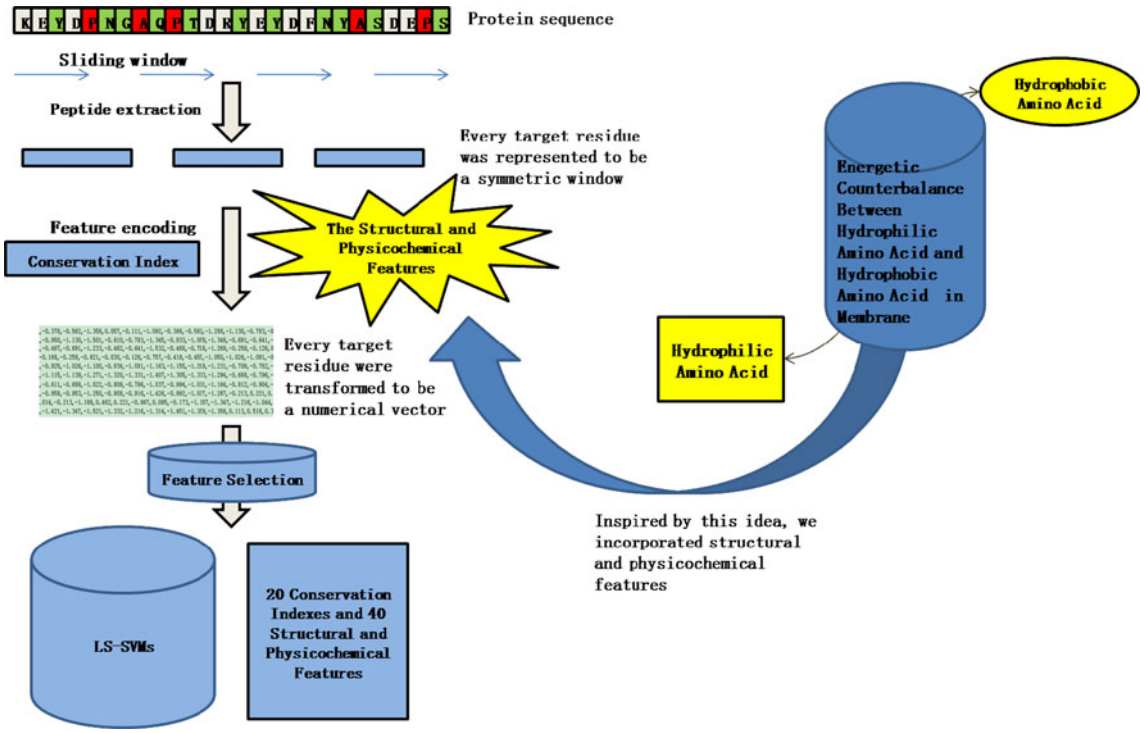

Fig. 1 Burial status prediction of TM residue was performed in the following order: in order to consider the immediate neighbors of the target residue, positions of $i \pm n$ with respect to each target residue i in the sliding window were included and we set $n = 10$ (window size of 21) for feature generation. Then, conservation index, structural and physicochemical features were encoded for each window. By using recursive feature elimination (RFE) method, 20 conservation indexes and 40 structural and physicochemical features were selected to build LS-SVM model to predict burial status of TM residues of αHMP

Feature generation

To capture the information of the target residue (central residue for window), the residues near to the central residues were extracted by sliding window. Given the position of the target residue, positions of $i \pm n$ with respect to each target residue i in the sliding window were included. Here, the length of sliding window is set to be 21; for each target residue in position i, the position of first and last residues in the window are $i - 10$ and $i + 10$, respectively. As a result, each target residue was represented by a symmetric window. A window is referred to as a positive one if the central residue of the window is buried within the protein core; whereas, it is considered to be negative one if the central residue of the window is exposed to the lipid bilayer. To build a predictor that can distinguish burial status of TM residues, three blocks of features were used to describe window: conservation, amino acid composition, and the distribution of amino acid properties along the sequence. Among them, amino acid properties were got from experimental scales (Bhaskaran and Ponnuswamy 1988; Bigelow 1967; Charton 1981; Charton and Charton 1982; Cid et al. 1992; Kawashima et al. 2008). At last, each window was converted into a feature vector. These input features were described in detail as follows.

Block1: Amino acid composition and the distribution of amino acid properties along the sequence

For each sequence contained in the window, seven feature groups composed of 1,487 structural and physicochemical features were calculated by a web-version of PROFEAT (Protein Features) (Li et al. 2006). Several publications have used some of the features covered by PROFEAT (Cai and Chou 2005; Chou and Cai 2005; Dubchak et al. 1995; Han et al. 2004; Han et al. 2007; Li et al. 2009; Li et al. 2008; Shen and Chou 2007; Xiao et al. 2006; Zhang et al. 2006). These 1,487 structural and physicochemical features include (a) amino acid composition, dipeptide composition (Reczko et al. 1997) (b) normalized Moreau-Broto autocorrelation (Berman et al. 2000; Lin and Pan 2001), (c) Moran autocorrelation (David 1988), (d) Geary autocorrelation (Sokal and Thomson 2006), (e) sequence-order-coupling number, quasi sequence-order descriptors(Chou 2000), (f) pseudo amino acid composition (Cai and Chou 2005), and (g) the composition and transition and distribution of various structural and physicochemical properties (Han et al. 2007). The first group is the amino acid composition and dipeptide composition. The second, third, and fourth group are the autocorrelation features. Eight amino acid indices are used here: (I) hydrophobicity scale (Cid et al. 1992), (II) average

Table 1 α-Helix membrane protein used in generating the development set

PDB ID	Chain used	Description
1DXR	L, M, H	Photosynthetic reaction center
1EHK	A, B	Cytochrome C oxidase (ba3 type)
1GZM	A	Rhodopsin
1J4N	A	Aquaporin
1KF6	C, D	Fumarate Reductase
1KQF	B, C	Formate dehydrogenase,nitrate-inducible,iron-sulfur subunit
1LDF	A	Glycerol uptake facilitator protein
1M0L	A	Bacteriorhodopsin
1NEK	C, D	Succinate dehydrogenase
1OKC	A	Mitochondrial ADP/ATP carrier
1OTS	A	Voltage-gated ClC-type chloride channel eriC
1PP9	D, E, G, J	Cytochrome bc1 complex
1Q16	C	Respiratory nitrate reductase 1 gamma chain
1QLA	C	Quinol-fumarate reductase diheme cytochrome B subunit C
1R3J	C	Voltage-gated potassium channel
1SU4	A	Sarcoplasmic/endoplasmic reticulum calcium ATPase 1
1V55	B, D, G, I, J, L, M	Cytochrome c oxidase polypeptide
1XQF	A	Ammonia channel
1YEW	B, C	Particulate methane monooxygenase
1ZOY	C, D	Cytochrome binding protein
2A65	A	Na(+): neurotransmitter symporter
2BL2	A	V-Type sodium ATP syntase subunit K
2CFQ	A	Lactose permease
2GIF	A	Acriflavine resistance protein B
2IC8	A	Protein glpG
2NQ2	A	Hypothetical ABC transporter permease protein HI1471

Table 2 α-Helix membrane protein used in generating the independent test set

PDB ID	Chain used	Description
2ZJS	Y	Preprotein translocase SecE subunit
2ZW3	A	Gap junction beta-2 protein
2Z73	A	Rhodopsin
3DD1	A	Glycogen phosphorylase, liver form
3D31	C	Sulfate/molybdate ABC transporter, ATP-binding protein
3DH4	A	Sodium/glucose cotransporter
3DHW	A	D-methionine transport system permease protein metI
3EFF	M	Voltage-gated potassium channel
3FH6	F	Maltose transport system permease protein malF
3KBC	A	Hypothetical proton glutamate symport protein

flexibility index(Bhaskaran and Ponnuswamy 1988), (III) polarizability parameter (Charton and Charton 1982), (IV) free energy of amino acid (Charton and Charton 1982), (V) residue accessible surface areas (Charton 1981), (VI) amino acid residue volumers (Bigelow 1967), (VII) residue accessible surface areas(Charton 1981), and (VIII) relative mutability(Dayhoff and Calderone 1978). The fifth group is composition, transition, and distribution features represent the amino acid distribution pattern of a specific structural or physicochemical property along the window. As to group f and g, they are derived from physicochemical distance matrix between each pair of the 20 amino acids. The physicochemical properties computed include hydrophobicity, polarity, and side-chain volume.

Block2: conservation index

It has been shown previously that conserved residues in MPs tend to be buried within the protein core (Stevens and Arkin 2001). In order to consider the influence of the conservation property on the burial status, we calculated conservation index for each residue using program AL2CO (Pei and Grishin 2001). The details about the calculation of the conservation index have been described elsewhere (Park et al. 2007; Park and Helms 2006, 2007). Here, we only give a brief description. First of all, amino acid frequencies at each position were calculated. Briefly, for each protein, a maximum of 1,000 similar sequences were retrieved from BLAST (http://www.ncbi.nih.gov/BLAST). ClustalW were then applied to build the initial multiple sequence alignments (MSAs) (Thompson et al. 1994). Sequence fragments and sequences that less than 25% identical to the query sequence were removed. Remaining sequences were then realigned using ClustalW to yield final MSAs. At last, we calculated amino acid frequencies from an MSA using three weighting schemes. These three weighting schemes are unweighted amino acid frequencies; weighted amino acid frequencies (Altschul et al. 1997; Henikoff and Henikoff 1994), and independent-count-based amino acid frequencies (Sunyaev et al. 1999). In the next step, as described by Pei and Grishin (Pei and Grishin 2001), the conservation index was calculation from amino acid frequencies by the following strategies. The first strategy is entropy-based one (denoted thereafter as ENTR)

$$C^{\mathrm{ENTR}}(i) = \sum_{a=1}^{20} f_a(i) \ln f_a(i) \tag{1}$$

In Eq. 1, $C^{\mathrm{ENTR}}(i)$ is the conservation index and $f_a(i)$ is the frequency of amino acid a at position i.

The second strategy is variance-based measure (denoted thereafter as VARI).

$$C^{\mathrm{VARI}}(i) = \sqrt{\sum_{a=1}^{20} (f_a(i) - f_a)^2} \tag{2}$$

where f_a is the overall frequency of amino acid a in the alignment. The third method is based on the sum-of-pairs measure using identity matrix (denoted thereafter as SOPid). Sum-of-pairs measure using BLOSUM62 matrix is the fourth method (denoted thereafter as SOPbl_62). As described in Eq. 3, Sopid and SOPbl_62 differ in S_{ab}. In Sopid method, S_{ab} is identity matrix, while, in SOPbl_62 method S_{ab} is a BLOSUM62 matrix.

$$C^{\mathrm{SOPid/SOPbl\_62}}(i) = \sum_{a}^{20} \sum_{b}^{20} f_a(i) f_b(i) S_{ab} \tag{3}$$

.

At last, $4 \times 3 = 12$ empirical scales encoded an amino acid and its position relative to central residue in a sliding window, leading to $12 \times 21(\mathrm{WindowSize}) = 252$ features. Totally, 1,739 features can be calculated for each target residue.

Feature selection

Due to the high intercorrelation and high noise level in such large number features, feature selection should be performed to select the features relevant with the burial status. Recursive feature elimination (RFE) method proposed by Guyon et al. (2002) was used in the feature selection The main steps include (1) train the classifier (2) compute the ranking criterion for all features, and (3) remove the feature with smallest ranking criterion. The number of features selected to build the final model was based on the prediction accuracy of leave-one-out cross-validation.

Least squares support vector machines

Recently, the support vector machine (SVM), as a powerful tool for data classification and function estimation has been developed (Cortes and Vapnik 1995; Cristianini and Shawe-Taylor 2000; lkopf et al. 1999; Vapnik 1998). One reason that SVM often performs better than earlier methods is that SVM was designed to minimize structural risk whereas previous techniques are usually based on minimization of empirical risk. However, it is often time-consuming to find the appropriate kernel function and optimal-free parameters of an SVM for very large training datasets (Vapnik 1998). Therefore, Suykens and Vandewalle proposed a modified version of SVM called least square SVM (LS-SVM) (Suykens and Vandewalle 1999), which resulted in a set of linear equations instead of a quadratic programming problem. It has been shown that LS-SVM can give readily excellent generalization performance at low computation cost (Suykens and Vandewalle 1999). In this work, LS-SVM is used to effectively utilize the extracted features for the prediction of burial status. The radial basis function (RBF) kernel was utilized and then LOO-CV was used to tune the optimized values of the two parameters γ (the relative weight of the classification error) and σ (the kernel parameter of the RBF kernel).

Results and discussion

Evaluation criteria and cross validation results of the development set

In this work, structural and physicochemical features are combined with conservation index for training the

LS-SVM model on the development set. The radial basis function (RBF) kernel was utilized as kernel function. LOO-CV was used to tune the optimized values of the two parameters γ (the relative weight of the classification error) and σ (the kernel parameter of the RBF kernel).

Prediction performance of the model were evaluated using the following indicators: sensitivity, specificity, total accuracy, Matthew's correlation coefficient (MCC) (Matthews 1975), and receiver operating characteristic (ROC) curve (Yang and Johnson 2005).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \tag{4}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100 \tag{5}$$

$$\text{Total accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \times 100 \tag{6}$$

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \times 100 \tag{7}$$

where TP and TN denote correctly predicted target residue buried within the protein core and exposed to the membrane. FP and FN are wrongly predicted target residue buried within the protein core and exposed to the membrane. The total accuracy is the ratio of the number of correctly predicted target residue (both positive and negative) to the total target residues. ROC curve is a two-dimensional plot of the sensitivity versus 1-specificity for a binary classifier as its decision boundary is moved. Sensitivity measures the capability of predicting positive samples (buried to the protein core) correctly, and specificity determines if any exposed residues are incorrectly predicted as buried residue. Additionally, MCC, ranging from -1 to $+1$, is also a good measure. Its value is -1 for the worst possible predictor, $+1$ for the best possible predictor, and 0 for a random predictor.

We used the results of cross-validation on the development set of 43 protein chains to estimate the stability and robustness of the model. Two cross-validation analyses were performed in our work: (1) k-fold cross-validation, where $k = 5$, 10; (2) leave-one-out cross validation (LOO-CV). The LS-SVM model using selected features achieved total accuracies of 85.50% and a 0.930 AUC (area under the ROC curve) value. The MCC, sensitivity, and specificity are 70.21, 79.71, and 89.79%, respectively. The prediction accuracy based on the different cross validation was listed in the Table 3. The ROC curves for different cross validation were shown in Fig. 2.

Prediction performance of the independent test set

In order to further evaluate the performance of our model, we applied our model to predict the samples in the independent test set containing ten proteins. These ten protein chains are non-redundant (sequence identity less than 30%) with development set. Our model can give a prediction accuracy of 80.80%, sensitivity of 71.19%, specificity of 87.96, and MCC of 60.5 for the independent test set (Table 4). In order to visualize the classification ability of our model, we also gave the ROC curve of the test set on Fig. 3. It is shown that similar classification performance can be obtained when our model is applied on the independent test set.

Performance comparison of our method with the reported works

For predicting the burial status of TM residue, the BeWe method utilized a knowledge-based propensity scale for predicting which residues of a TM helix are exposed to the lipid bilayer or buried within the protein core. For each target residue, the BeWe method computed its position score and compared the score with a threshold to predict the burial status (Beuming and Weinstein 2004). However, it is difficult to set the threshold manually when implementing BeWe method. Yuan (Yuan et al. 2006) and Illergard (Illergard et al. 2010) proposed two hybrid methods using SVMs in conjunction with evolutionary information of amino acid sequences in terms of their position-specific scoring matrices (PSSMs) for prediction of relative solvent-accessible surface area (rSASA). The biologist could utilize predicted rSASA values to predict the burial status of TM residue. Recently, Park and his coworkers combined the conservation index to the SVM to predict the burial status of TM residue (TMX method) (Park et al. 2007). In order to compare the performance of our method with the reported works, PSSMs and conservation index were also utilized to build the LS-SVM model. The prediction results of the LS-SVM models using different features are shown in Table 5. Figure 4 shows the ROC curve for LS-SVM model with different features. The combination of conservation index and structural and physicochemical features produced the best performance with an 85.50% total accuracy and a 0.930 AUC (area under the ROC curve) value. Therefore, the LS-SVM model combined conservation indexes with structural and physicochemical features outperforms other two LS-SVM model only using conservation or PSSMs. The results indicate that the combination of all features is capable of capturing more information for burial status of TM residue.

Table 3 Cross validation results on the development set for LS-SVM model

Method	Accuracy (%)	AUC (area under the ROC curve)
Fivefold cross validation	77.400	0.845
Tenfold cross validation	77.690	0.849
Leave-one-out cross validation	78.020	0.850

Since it is often difficult to directly compare performance values of different prediction methods reported in different studies, Park and his coworkers chose a data set of 3,138 residues to evaluate the BeWe method and Yuan method. We also used utilized this data set to evaluate the performance of our model and compare its performance with the reported works. The total accuracy is 68.67, 71.06 and 78.71% for the BeWe, Yuan, and TMX methods. Compared with these methods, our model achieves an accuracy of 85.50%. It shows considerable improvements to the proposed methods. Because prediction accuracy of the TMX method is much higher than the methods of BeWe and Yuan, we make further comparison between our prediction results with TMX method. Compared with TMX method, the MCC, specificity and sensitivity of our work attained 70.21% (56.21% with TMX method) 79.71% (70.61% with TMX method) and 89.79% (84.77% with TMX method). The results clearly demonstrate that our method is capable of predicting burial status with significantly better performance than those previous methods.

The interpretation of the selected features

In this work, we combined the structural and physicochemical features together with conservation index to develop LS-SVM model for the prediction of the burial status of TM residue. But, such a large number of features may be high intercorrelation. Generally speaking, choosing a relevant subset of features leads to a simpler model and can lead to better performance (Han et al. 2007; Park et al. 2007). Here, 20 conservation indexes and 40 structural and physicochemical features were selected. The ranking of the features according to their relative importance by RFE algorithm is shown in Tables 6 and 7. As shown in Table 6,

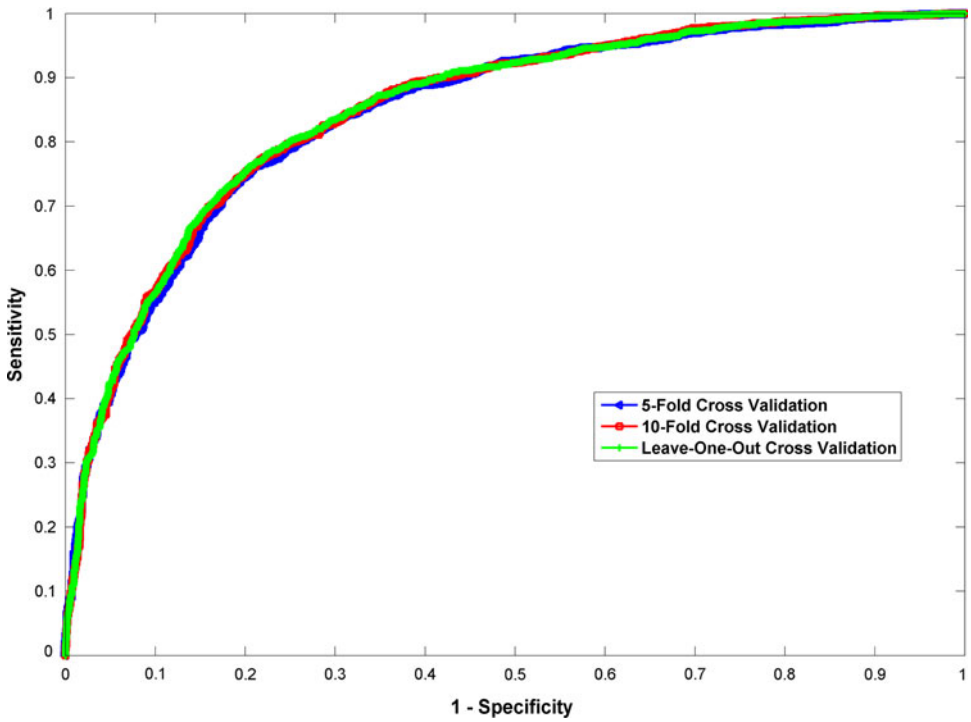

Fig. 2 Receiver operating characteristic (ROC) curves of fivefold, tenfold and Leave one out cross validation for the development set

Table 4 Burial status prediction accuracy of the development set and independent test set

Method	MCC	Sensitivity (%)	Specificity (%)	Total accuracy (%)	AUC (area under the ROC curve)
Development set	70.21	79.71	89.79	85.5	0.93
Independent test set	60.5	71.19	87.96	80.8	0.89

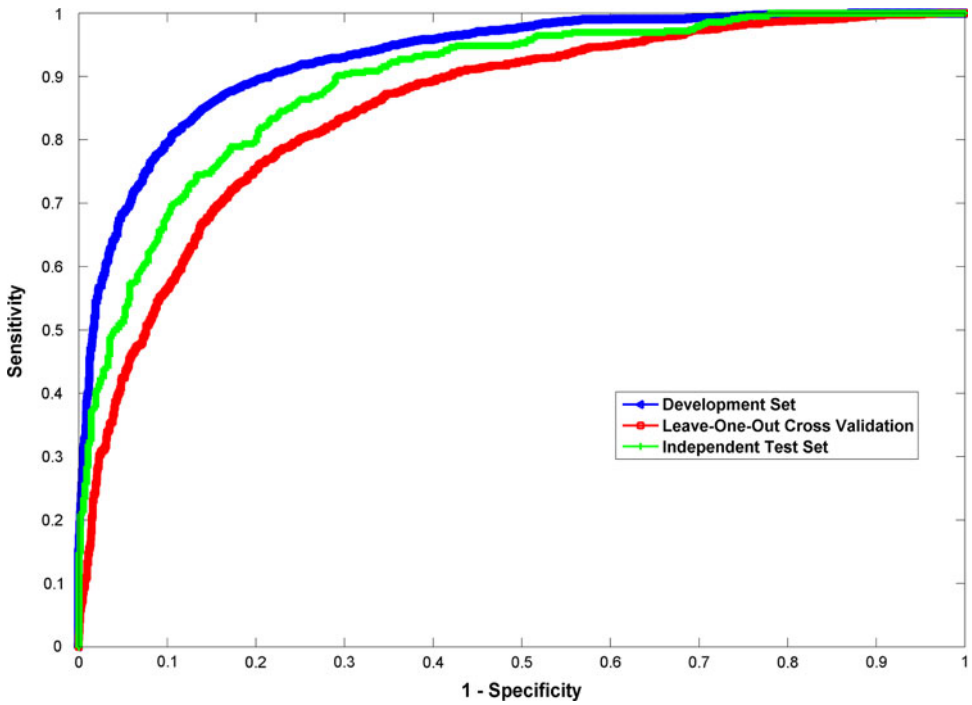

Fig. 3 Receiver operating characteristic (ROC) on development set and independent test set

Table 5 The prediction performance of the LS-SVM model based on different features

Features	MCC	Sensitivity (%)	Specificity (%)	Total accuracy (%)	AUC (area under the ROC curve)
PSSMs	34.62	50.91	81.78	68.63	0.74
Conservation index	51.37	68.31	82.40	76.40	0.84
Conservation index + structural and physicochemical features	70.21	79.71	89.79	85.50	0.93

the first top-ranking element is conservation index of target residue. This result is consistent with that of previous study that conservation property of TM residue correlates strongly with their burial status (Park and Helms 2006; Park et al. 2007). Table 6 also shows that the conservation properties of C4 (the 4th residue C terminal to the target residue), N4 (the 4th residue N terminal to the target residue) are highly associated with the burial status of TM residue, which is also consistent with previous works (Russ and Engelman 2000; Senes et al. 2000; Walters and Degrado 2006). On the other hand, two conservation indexes of C6 and N9 were also used in the model development. This is largely owing to the helix conformation of TM segment, and one should consider the spatial position of each residue in TM domain as mentioned earlier.

Except the conservation index, 40 structural and physicochemical features were also involved to develop burial status prediction model. As shown in Table 7, more than one-third of the selected features are hydrophobicity autocorrelation features, indicating the relevance of the distribution of hydrophobic amino acid in αHMP with burial status. The works of several groups have compiled statistical information about the burial status of TM residue (Beuming and Weinstein 2004; Park and Helms 2007). The results show that hydrophobic amino acids (F, I, L, V) are more likely exposed to the membrane whereas hydrophilic amino acids (D, N, Q, R) are more likely buried within the protein core. In order to maintain energetic counterbalance between hydrophobic amino acid and hydrophilic amino acid in TM segment, hydrophobic amino acids are always favored exposed to the membrane whereas hydrophilic amino acids are always favored buried within the protein core (MacKinnon 2005; White and von Heijine 2008; Hessa et al. 2007). Therefore, the distribution of hydrophobic amino acid can directly influence the burial status of amino acid in TM segment. This can

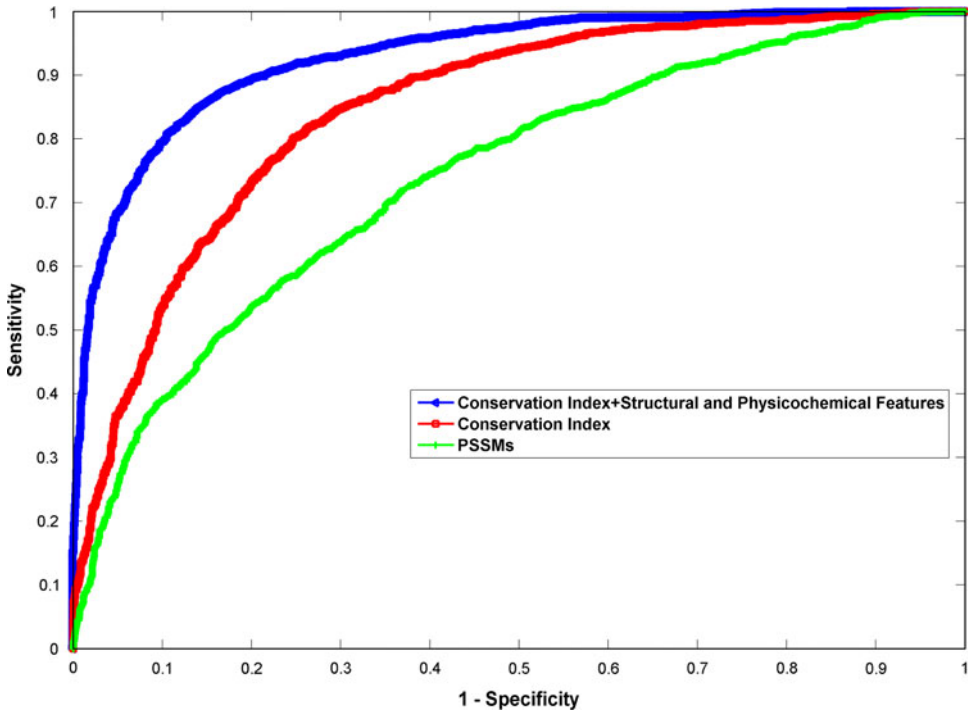

Fig. 4 Receiver operating characteristic (ROC) curves by using different features

Table 6 Twenty significant conservation index selected to build model

Rank	Type	Position
1	Independent count ENTR	T[a]
2	Independent count SOPid	T
3	Weighted SOPbl_62	T
4	Independent count VARI	N9
5	Independent count SOPid	C4[b]
6	Independent count VARI	N4[c]
7	Independent count VARI	N6
8	Unweighted ENTR	N7
9	Independent count VARI	C4
10	Independent count SOPid	N9
11	Weighted SOPbl_62	C4
12	Independent count SOPid	C6
13	Independent count VARI	N2
14	Independent count SOPid	C1
15	Independent count VARI	C6
16	Unweighted SOPid	N4
17	Weighted SOPid	C7
18	Unweighted VARI	N5
19	Independent count VARI	N10
20	Independent count ENTR	N4

[a] The target residue

[b] The 4th residue C terminal to the target residue

[c] The 4th residue N terminal to the target residue

explain why hydrophobicity autocorrelation features play a crucial role in burial status prediction. Table 7 contains seven flexibility features and account for almost one-fifth of selected features. The symmetric/asymmetric distribution (flexibility scale) of amino acid residues in the protein molecules is an important factor to reflect the environmental constraints (Bhaskaran and Ponnuswamy 1988). Considering the special environmental of TM segment in αHMP, the flexibility scale can indirectly represent the distribution of hydrophobic amino acids, which can help one to understand why flexibility features strongly correlate with burial status.

Conclusion

In this work, a novel method for predicting burial status of TM residues of α-helix membrane protein was developed. Structural and physicochemical features together with conservation index from sequence information were used to describe the features of the target residues. The RFE algorithm was used to select the significant features relevant to the burial status of the residues. LS-SVM was utilized to build the prediction models for the burial status of TM residue. The LS-SVM model achieves 85.50% overall accuracy with Matthew's correlation coefficient of 70.21%. The result of our LS-SVM model is much better

Table 7 Forty significant structural and physicochemical features selected to build model

Rank	Feature index in PROFEAT	Type
1	F3.1	Hydrophobicity scale-Moran autocorrelation
2	F2.1	Hydrophobicity scale-Normalized Moreau-Broto autocorrelation
3	F2.1	Hydrophobicity scale-Normalized Moreau-Broto autocorrelation
4	F2.1	Hydrophobicity scale-Normalized Moreau-Broto autocorrelation
5	F2.1	Average flexibility index- Normalized Moreau-Broto autocorrelation
6	F4.1	Average flexibility index- Geary autocorrelation
7	F2.1	Hydrophobicity scale-Normalized Moreau-Broto autocorrelation
8	F6.2	Quasi-sequence-order descriptor
9	F2.1	Hydrophobicity scale-Normalized Moreau-Broto autocorrelation
10	F3.1	Average flexibility index- Moran autocorrelation
11	F2.1	Hydrophobicity scale-Normalized Moreau-Broto autocorrelation
12	F1.2	Dipeptide composition
13	F2.1	Hydrophobicity scale-Normalized Moreau-Broto autocorrelation
14	F3.1	Hydrophobicity scale-Moran autocorrelation
15	F3.1	Hydrophobicity scale-Moran autocorrelation
16	F1.2	Dipeptide composition
17	F3.1	Hydrophobicity scale-Moran autocorrelation
18	F2.1	Polarizability parameter-Normalized Moreau-Broto autocorrelation
19	F2.1	Amino acid residue volumes-Normalized Moreau-Broto autocorrelation
20	F2.1	Free energy of amino acid solution in water-Normalized Moreau-Broto autocorrelation
21	F1.2	Dipeptide composition
22	F2.1	Hydrophobicity scale-Normalized Moreau-Broto autocorrelation
23	F3.1	Average flexibility index -Moran autocorrelation
24	F3.1	Hydrophobicity scale-Moran autocorrelation
25	F2.1	Hydrophobicity scale-Normalized Moreau-Broto autocorrelation
26	F3.1	Hydrophobicity scale-Moran autocorrelation
27	F2.1	Hydrophobicity scale-Normalized Moreau-Broto autocorrelation
28	F2.1	Amino acid residue volumes-Normalized Moreau-Broto autocorrelation
29	F2.1	Amino acid residue volumes-Normalized Moreau-Broto autocorrelation
30	F3.1	Hydrophobicity scale-Moran autocorrelation
31	F3.1	Average flexibility index- Moran autocorrelation
32	F4.1	Average flexibility index- Geary autocorrelation
33	F6.2	Quasi-sequence-order descriptor
34	F1.2	Dipeptide composition
35	F2.1	Amino acid residue volumes-Normalized Moreau-Broto autocorrelation
36	F2.1	Relative mutability-Normalized Moreau-Broto autocorrelation
37	F6.2	Quasi-sequence-order descriptor
38	F2.1	Average flexibility index-Moran autocorrelation
39	F1.2	Dipeptide composition
40	F1.2	Dipeptide composition

than those reported in the references for predicting burial status of TM residues of α-helix membrane protein. Importantly, the position and the composition of hydrophobic amino acid properties were proved to be very important features influencing the burial status of a TM residue.

Acknowledgments This work was supported by the Program for New Century Excellent Talents in University (Grant No. NCET-07-0399) and the National Natural Science Foundation of China (Grant No. 20905033).

References

Abramson J, Smirnova I, Kasho V, Verner G, Kaback HR, Iwata S (2003) Structure and mechanism of the lactose permease of *Escherichia coli*. Science 301:610–615

Adamian L, Liang J (2006) Prediction of transmembrane helix orientation in polytopic membrane proteins. BMC Struct Biol 6:13

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242

Bernsel A, Viklund H, Falk J, Lindahl E, von Heijne G, Elofsson A (2008) Prediction of membrane-protein topology from first principles. Proc Natl Acad Sci USA 105:7177–7181

Beuming T, Weinstein H (2004) A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins. Bioinformatics 20:1822–1835

Bhaskaran R, Ponnuswamy PK (1988) Positional flexibilities of amino acid residues in globular proteins. J Peptide Protein Res 32:241–255

Bigelow CC (1967) On the average hydrophobicity of proteins and the relation between it and protein structure. J Theor Biol 16:187–211

Cai Y-D, Chou K-C (2005) Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. J Proteome Res 4:967–971

Cao B, Porollo A, Adamczak R, Jarrell M, Meller J (2006) Enhanced recognition of protein transmembrane domains with prediction-based structural profiles. Bioinformatics 22:303–309

Charton M (1981) Protein folding and the genetic code: an alterative quantitative model. J Theor Biol 91:115–123

Charton M, Charton BI (1982) The structural dependence of amino acid hydrophobicity parameters. J Theor Biol 99:629–644

Chen K, Jiang Y, Du L, Kurgan L (2009) Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs. J Comput Chem 30:163–172

Chou KC (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem Biophys Res Commun 278:477–483

Chou KC, Cai YD (2005) Predicting protein–protein interactions from sequences in a hybridization space. J Proteome Res 5:316–322

Cid H, Bunster M, Canales M, Gazitua F (1992) Hydrophobicity and structural classes in proteins. Protein Eng 5:373–375

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297

Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines. Cambridge University Press, Cambridge

David CW (1988) Voronoi polyhedra as structure probes in large molecular systems. Biopolymers 27:339–344

Dayhoff H, Calderone H (1978) Composition of proteins. Altas Protein Seq Struct 5:363–373

Dubchak I, Muchnik I, Holbrook SR, Kim SH (1995) Prediction of protein folding class using global description of amino acid sequence. Proc Natl Acad Sci USA 92:8700–8704

Granseth E, Viklund H, Elofsson A (2006) ZPRED: predicting the distance to the membrane center for residues in {alpha}-helical membrane proteins. Bioinformatics 22:e191–e196

Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46:389–422

Han LY, Cai CZ, Ji ZL, Cao ZW, Cui J, Chen YZ (2004) Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. Nucleic Acids Res 32:6437–6444

Han LY, Zheng CJ, Xie B, Jia J, Ma XH, Zhu F, Lin HH, Chen X, Chen YZ (2007) Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness. Drug Disco Today 12:304–313

Henikoff S, Henikoff JG (1994) Position-based sequence weights. J Mol Biol 243:574–578

Hessa T, Meindl-Beinker NM, Bernsel A, Kim H, Sato Y, Lerch-Bader M, Nilsson I, White SH, Von Heijne G (2007) Molecular code for transmembrane-helix recognition by the Sec61 translocon. Nature 450:1026–1030

Huang Y, Lemieux MJ, Song J, Auer M, Wang D-N (2003) Structure and mechanism of the Glycerol-3-Phosphate transporter from *Escherichia coli*. Science 301:616–620

Illergard K, Callegari S, Elofsson A (2010) MPPAP: an accessibility predictor for alpha-helical transmembrane proteins that performs well inside and outside the membrane. BMC Bioinformatics 11:333

Jayasinghe S, Hristova K, White SH (2001) MPtopo: a database of membrane protein topology. Protein Sci 10:455–458

Jones DT (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics 23:538–544

Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M (2008) AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 36:D202–D205

Li ZR, Lin HH, Han LY, Jiang L, Chen X, Chen YZ (2006) PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res 34:W32–W37

Li ZC, Zhou XB, Lin YR, Zou XY (2008) Prediction of protein structure class by coupling improved genetic algorithm and support vector machine. Amino Acids 35:581–590

Li S, Xi L, Wang C, Li J, Lei B, Liu H, Yao X (2009) A novel method for protein-ligand binding affinity prediction and the related descriptors exploration. J Comput Chem 30:900–909

Lin Z, Pan X-M (2001) Accurate prediction of protein secondary structural content. J Protein Chem 20:217–220

Liu Y, Engelman D, Gerstein M (2002) Genomic analysis of membrane protein families: abundance and conserved motifs. Genome Biol 3: research0054.0051–research0054.0012

lkopf BS, Burges C, Smola A (1999) Advances in Kernel methods—support vector learning. MIT Press, Cambridge

MacKinnon R (2005) STRUCTURAL BIOLOGY: membrane protein insertion and stability. Science 307:1425–1426

Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica Biophys Acta 405:442–451

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

Page RC, Lee S, Moore JD, Opella SJ, Cross TA (2009) Backbone structure of a small helical integral membrane protein: a unique structural characterization. Protein Sci 18:134–146

Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Trong IL, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M (2000) Crystal structure of rhodopsin: a G Protein-coupled receptor. Science 289:739–745

Park Y, Helms V (2006) How strongly do sequence conservation patterns and empirical scales correlate with exposure patterns of transmembrane helices of membrane proteins? Biopolymers 83:389–399

Park Y, Helms V (2007) On the derivation of propensity scales for predicting exposed transmembrane residues of helical membrane proteins. Bioinformatics 23:701–708

Park Y, Hayat S, Helms V (2007) Prediction of the burial status of transmembrane residues of helical membrane proteins. BMC Bioinformatics 8:302

Pebay-Peyroula E, Rummel G, Rosenbusch JP, Landau EM (1997) X-ray Structure of bacteriorhodopsin at 2.5 angstroms from microcrystals grown in lipidic cubic phases. Science 277:1676–1681

Pei J, Grishin NV (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 17:700–712

Poolman B, Geertsma ER, Slotboom D-J (2007) BIOCHEMISTRY: a missing link in membrane protein evolution. Science 315:1229–1231

Reczko M, Karras D, Bohr H (1997) An update of the DEF database of protein fold class predictions. Nucleic Acids Res 25:235

Russ WP, Engelman DM (2000) The GxxxG motif: a framework for transmembrane helix-helix association. J Mol Biol 296:911–919

Senes A, Gerstein M, Engelman DM (2000) Statistical analysis of amino acid patterns in transmembrane helix: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. J Mol Biol 296:921–936

Shen HB, Chou KC (2007) Using ensemble classifier to identify membrane protein types. Amino Acids 32:483–488

Sokal RR, Thomson BA (2006) Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. Am J Phys Anthropol 129:121–131

Stevens TJ, Arkin IT (2001) Substitution rates in alpha-helical transmembrane proteins. Protein Sci 10:2507–2517

Sui H, Han B-G, Lee JK, Walian P, Jap BK (2001) Structural basis of water-specific transport through the AQP1 water channel. Nature 414:872–878

Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng 12:387–394

Suykens JAK, Vandewalle J (1999) Least squares support vector machine classifiers. Neural Process Lett 9:293–300

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Vapnik VN (1998) Statistical learning theory. Wiley, New York

Viklund H, Elofsson A (2004) Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. Protein Sci 13:1908–1917

Visiers I, Ballesteros JA, Weistein H (2002) Three-dimensional representations of G protein-coupled receptor structures and mechanisms. Methods Enzymol 343:329–371

von Heijine G (1999) Recent advances in the understanding of membrane protein assembly and structure. Q Rev Biophys 32:285–307

Walters RFS, DeGrado WF (2006) Helix-packing motif in membrane proteins. Proc Natl Acad Sci USA 103:13658–13663

White SH, von Heijine G (2008) How translocons select transmembrane helices. Annu Rev Biophys 37:23–42

Xiao X, Shao S, Ding Y, Huang Z, Chou KC (2006) Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30:49–54

Yang ZR, Johnson FC (2005) Prediction of T-cell epitopes using biosupport vector machines. J Chem Inf Model 45:1424–1428

Yang L, Li Y, Xiao R, Zeng Y, Xiao J, Tan F, Li M (2010) Using auto covariance method for functional discrimination of membrane proteins based on evolution information. Amino Acids 38:1497–1503

Yuan Z, Zhang F, Davis MJ, Bodén M, Teasdale RD (2006) Predicting the solvent accessibility of transmembrane residues from protein sequence. J Proteome Res 5:1063–1070

Zhang SW, Pan Q, Zhang HC, Shao ZC, Shi JY (2006) Prediction of protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and Naive Bayes Feature Fusion. Amino Acids 30:461–468