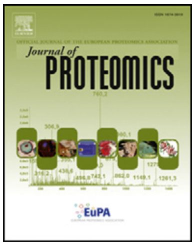
Available online at www.sciencedirect.com

SciVerse ScienceDirect

www.elsevier.com/locate/jprot



Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions

Chen Ding[a], Lu-Feng Yuan[a], Shou-Hui Guo[a], Hao Lin[a,*], Wei Chen[b,*]

[a]Key Laboratory for NeuroInformation of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China
[b]Department of Physics, Center for Genomics and Computational Biology, College of Sciences, Hebei United University, Tangshan 063000, China

ARTICLE INFO

Article history:
Received 11 May 2012
Accepted 8 September 2012
Available online 20 September 2012

Keywords:
Mycobacterium
Membrane protein
Tripeptide
Support vector machine
Feature selection

ABSTRACT

Mycobacterium can cause many serious diseases, such as tuberculosis and leprosy. Its membrane proteins play a critical role for multidrug-resistance and its tenacious survival ability. Knowing the types of membrane proteins will provide novel insights into understanding their functions and facilitate drug target discovery. In this study, a novel method was developed for predicting mycobacterial membrane protein and their types by using over-represented tripeptides. A total of 295 non-membrane proteins and 274 membrane proteins were collected to evaluate the performance of proposed method. The results of jackknife cross-validation test show that our method achieves an overall accuracy of 93.0% in discriminating between mycobacterial membrane proteins and mycobacterial non-membrane proteins and an overall accuracy of 93.1% in classifying mycobacterial membrane protein types. By comparing with other methods, the proposed method showed excellent predictive performance. Based on the proposed method, we built a predictor, called MycoMemSVM, which is freely available at http://lin.uestc.edu.cn/server/MycoMemSVM. It is anticipated that MycoMemSVM will become a useful tool for the annotation of mycobacterial membrane proteins and the development of anti-mycobacterium drug design.

1. Introduction

Mycobacterium is a genus of Actinobacteria and is notoriously known for its pathogenicity. *Mycobacterium* can cause serious diseases such as tuberculosis (TB) and leprosy, which result in millions of cases of infection and deaths every year. Although scientists have made great efforts to develop bacterin and drugs for the treatment of these diseases, the appearance of multidrug-resistant of TB makes some drugs lost their luster and brings a great challenge to drug design. Mycobacteria possess an especially complex cell envelope that consists of a cell wall and a cytoplasmic membrane, which plays a critical role for its multidrug-resistance and tenacious survival ability under harsh conditions [1–3]. Membrane proteins are the most important part of cell membrane. These membrane proteins exert their many crucial physiological and biological functions, including as carrier to transport materials into or out of cells and as receptors of some hormone or chemical substance. In particular, membrane proteins are important drug targets [2–4]. Therefore, accurately identifying the types of mycobacterial

* *Corresponding authors.* Tel.: +86 28 83208232; fax: +86 28 83208238.
E-mail addresses: hlin@uestc.edu.cn (H. Lin), chenwei_imu@yahoo.com.cn (W. Chen).

1874-3919/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jprot.2012.09.006

membrane proteins will be helpful for the analysis of the functions of mycobacterial proteins and the development of antimicrobial drugs [5].

It is highly reliable to use experimental methods to recognize the types of membrane proteins. However, because most of membrane proteins are difficult to crystallize and dissolve in majority solvents, we can not get their structure by crystal diffraction or nuclear magnetic resonance (NMR) spectroscopy [6]. Furthermore, it is also both costly and time-consuming for experimental approaches to identify membrane proteins. Thus, it is necessary to develop effective computational methods to predict mycobacterial membrane proteins and their types.

In the past decade, many computational methods have been proposed to predict membrane proteins according to their sequence information [7–14]. The statistical-based prediction models contain two pivotal procedures. One is the representation of protein sequences. The information-rich description of protein is a key step in building precise and robust model. Researchers have presented various feature extraction strategies, such as amino acid composition (AAC) [15], pseudo amino acid composition (PseAAC) [13,16–19], functional domain composition [8], supervised locally linear embedding (SLLE) [20], evolution information [12,21,22], split amino acid composition (SAAC) [6,17,21,23], wavelet analysis [24] and so on. The other is the classification algorithm. A prominent algorithm is also requisite for constructing a wonderful model. Up to now, many algorithms such as support vector machine (SVM) [25,26], artificial neural networks (ANN) [24,27], K-nearest neighbor (KNN) [10,28,29], hidden Markov model (HMM) [30], ensemble classification [6,11] and Mahalanobis discriminant algorithm [16,31] have been applied in protein structure and function prediction. The success of these predictors implied that the machine learning method can be used for the prediction of mycobacterial membrane proteins and their types. However, few works have focused on the prediction of membrane protein in mycobacterium. Yeh and Mao [32] have used SVM with N-terminal sequence patterns to distinguish between membrane proteins and soluble proteins in *Mycobacterium tuberculosis*. Pajón et al. [33] have developed a predictor, called PROB, to identify beta-barrel outer membrane proteins of *M. tuberculosis* and employed it to predict 79 new proteins which were not annotated by experiment. Song et al. [34] have analyzed outer membrane proteins of *M. tuberculosis* and found two function-undefined outer membrane proteins. Results of these works are pretty, but none of them paid attention to the prediction of mycobacterial membrane proteins types. Recently, Fan and Li [35] presented a PseAAC-based method to predict three types of mycobacterial membrane proteins. The overall accuracy of 85.0% with the average accuracy of 63.4% was achieved in jackknife cross-validation. However, this accuracy is far from satisfactory. Thus, it is urgent to construct accurate model to predict mycobacterial membrane proteins and their types.

In view of this, the present study was attempted to develop an effective method for predicting mycobacterial membrane proteins and their types. The binomial distribution was used to pick out the over-represented tripeptides. The SVM was used to perform prediction. In the jackknife cross-validation, our method achieved an overall accuracy of 93.0% for the prediction of mycobacterial membrane proteins and an overall accuracy of 93.1% for the identification of the types of mycobacterial membrane proteins. To further demonstrate its advantages, we also compared the performance of the proposed method with other methods.

2. Materials and methods

2.1. Dataset

The data of mycobacterial proteins used in this paper were extracted from Universal Protein Resource (UniProt) database [36]. In order to obtain high quality and well defined dataset, we selected protein sequences which strictly following the procedures below: first, we chose sequences which were reviewed and manually annotated by experts; second, we excluded the proteins whose type is undefined or ambiguous; third, we eliminated the sequences whose protein existence is uncertain or predicted; fourth, we dislodged the sequences which are fragments of other proteins; finally, the sequence redundancy was reduced to 90%. After the above screening procedures, we obtained a dataset containing 2439 mycobacterial non-membrane proteins and 1645 mycobacterial membrane proteins. In order to objectively validate the method and compare its predictive performance with existing algorithms, the 1645 mycobacterial membrane proteins were randomly divided into a training dataset (including 1039 membrane proteins) and an independent dataset (including 606 membrane proteins).

It is well known that protein dataset with high similarity always contains redundancy, which can overestimate the performance and reduce the generalization ability of a proposed model. To get non-redundant data, the CD-HIT [37] program was utilized to remove the redundant sequences using 40% sequence identity as the cutoff. As a result, we obtained 295 non-membrane proteins. The training set of membrane proteins contains 274 sequences, of which 32 are single-pass membrane proteins, 192 are multi-pass membrane proteins, 20 are lipid-anchor membrane proteins, and 30 are peripheral membrane proteins. The independent set of membrane proteins contains 125 membrane protein sequences, of which 18 are single-pass membrane proteins, 75 are multi-pass membrane proteins, 8 are lipid-anchor membrane proteins, and 24 are peripheral membrane proteins. These data can be freely downloaded from our web site http://lin.uestc.edu.cn/server/MycoMemSVM.

2.2. Tripeptide compositions

It is one of the most important parts for pattern recognition to extract a set of informative parameters. The tripeptides play important roles in biology. Three contiguous amino acids constitute a useful and minimal biological recognition signal. This may form a useful paradigm for discovering peptides and small organic molecule mimics that are useful modulators of biological function [38]. Anishetty et al. [39] have also showed that the tripeptide may be used to predict plausible structures for oligopeptides as well as denovo protein design. Therefore, in this work, tripeptide compositions were utilized to represent the sample of membrane proteins. By scanning one sequence using a sliding window of three residues with one step, we calculated the frequency of each tripeptide appeared in the protein sequence according to the following equation:

$$f_i = n_i/\sum_{i=1}^{8000} n_i = n_i/(L–2) \tag{1}$$

where n_i and L denote the number of the i-th tripeptide and the length of the protein sequence, respectively.

Then the protein can be described by an 8000 dimensional vector as follows:

$$F_{8000} = [f_1, f_2, \cdots, f_i, \cdots, f_{8000}]^T \tag{2}$$

where symbol T denotes the transposition of vector and f_i is the frequency of the i-th tripeptide.

2.3. *Feature selection*

Theoretically, if all 8000 tripeptides are selected as the feature of membrane proteins, there will be three problems: one is over-fitting which results in low generalization ability of prediction model; another is information redundancy or noise which results in bad prediction accuracy; the other is dimension disaster which results in a handicap for the computation. Using feature selection techniques to optimize feature set can not only economize the time for computation, but also build robust prediction model [40]. Many techniques such as principal component analysis (PCA) [41,42], diffusion Maps [43], minimal-redundancy-maximal-relevance (mRMR) [29,44], analysis of variance (ANOVA) [45,46], local linear discriminant analysis (LLDA) [47] and geometry preserving projections (GPP) [22] have been proposed and used in sequence analysis and prediction.

In this study, the binomial distribution was proposed to optimize tripeptides [48] for eliminating the redundant features and improving the efficiency and performance of prediction. Due to 8000 kinds of tripeptides may occur in benchmark dataset, and each kind of tripeptide occurring in one class may be a stochastic event, we must judge whether each kind of tripeptide occurring in one class is a stochastic event or not.

Suppose the total occurrence frequency of all tripeptides in the dataset is M, the probability of tripeptides occurred in class i is denoted as p_i and calculated as:

$$p_i = m_i/M \tag{3}$$

where the m_i is the number of tripeptides that appear in the class i.

Let N_j represent the total occurrence number of a given tripeptide j in dataset, the probability of tripeptide j randomly occurring n_{ij} or more times in the class i can be defined by:

$$P(n_{ij}) = \sum_{m=n_{ij}}^{N_j} \frac{N_j!}{m!(N_j–m)!} p_i^m (1–p_i)^{N_j–m}. \tag{4}$$

If $P(n_{ij})$ is a small value, it means the tripeptide j appearing in class i should not be random. The confidence level of this case is defined by CL_{ij}:

$$CL_{ij} = 1–P(n_{ij}). \tag{5}$$

If there are k tripeptides whose CL_{ij} is larger than a given cutoff CL_0, the frequencies of these tripeptides are selected as optimized features that can be described as:

$$F_k = [f_1, f_2, \cdots, f_i, \cdots, f_k]^T \tag{6}$$

Based on confidence (Eq. (5)), high-dimensional feature parameters can be projected into low-dimensional space. The parameter k or CL_0 can be chosen by the use of five-fold cross-validation.

2.4. *Support vector machine*

Support vector machine (SVM) is a machine learning algorithm based on statistical learning theory. It has been widely used in the field of protein structure and functional predictions [49–53]. The basic idea of SVM is to transform the data into a high dimensional feature space, and then determine the optimal separating hyperplane. For handling a multi-class problem, “one-versus-one (OVO)” and “one-versus-rest (OVR)” are generally applied to extend the traditional SVM. In this study, OVO strategy was employed. Usually, four kinds of kernel functions, i.e. linear function, polynomial function, sigmoid function and radial basis function (RBF), can be available to perform prediction. Empirical studies have demonstrated that the RBF outperforms the other three kinds of kernel functions. Hence, in this work we used the RBF to perform prediction. The grid search method is applied to tune the regularization parameter C and the kernel width parameter γ by using five-fold cross-validation. The software toolbox used to implement SVM is LibSVM written by Lin's lab and can be freely downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm [54].

2.5. *Performance assessment*

In order to assess the capability of the prediction method, the sensitivity (Sn), specificity (Sp), overall accuracy (OA), and average accuracy (AA) were used in this study. These measures are defined as follows:

$$\mathrm{Sn}_i = \mathrm{TP}_i/(\mathrm{TP}_i + \mathrm{FN}_i) \tag{7}$$

$$\mathrm{Sp}_i = \mathrm{TN}_i/(\mathrm{TN}_i + \mathrm{FP}_i) \tag{8}$$

$$\mathrm{OA} = \sum_{i=1}^{\mu} \mathrm{TP}_i/N \tag{9}$$

$$\mathrm{AA} = \sum_{i=1}^{\mu} \mathrm{Sn}_i/\mu \tag{10}$$

where TP_i and FN_i represent the number of true positives and false negatives, while FP_i and TN_i represent false positives and true negatives for class i. μ is the number of protein classes. N is the total number of sequences in benchmark dataset.

3. Results and discussion

3.1. *Prediction performance*

In statistical prediction, various methods such as n-fold cross-validation test, sub-sampling test, independent dataset test and jackknife cross-validation test have been adopted to evaluate the performance of a prediction model. Because the jackknife test can achieve unique outcome [55], in this study, the jackknife cross-validation was used to investigate the performance of the prediction model. In the jackknife test, each protein sequence is in turn singled out as an independent test sample and all the rule parameters are calculated based on the remaining samples.

As described in feature selection section, each membrane protein sequence was translated into a set of over-represented tripeptides. If we select a high CL_0, the results are robust and credible. The selected tripeptides are also informative. However, they are not the optimized features for prediction because the number of these tripeptides is too small to reflect enough information of membrane proteins. They can only describe part of membrane protein properties. For example, by using >99.99% as the confidence level, we obtained 31 tripeptides, but the overall accuracy was only 78.8% for predicting four types of membrane proteins. On the contrary, feature sets with low confidence level contain so many components that the cluster-tolerant capacity of the prediction model reduces so as to lower down the cross-validation accuracy. For instance, 4173 tripeptides with >50% confidence level produced the overall accuracy of 70.8% for predicting four types of membrane proteins. Therefore, it is a key step to choose an appropriate confidence level or the number of features for a robust and high accuracy model.

At first, we used our method to discriminate mycobacterial membrane proteins from non-membrane proteins. By adding the tripeptides one by one according to the CLs calculated by Eq. (5), we built a set of individual predictors for these sub-feature sets using SVM. We then examined the predictive performance for each of these predictors using five-fold cross-validation and plotted the 3-dimension (3-D) curve for CL, feature dimension and overall accuracy in Fig. 1. We found that the overall accuracy reached its maximum 93.0% when the CL_0 was selected as 90.81%. The sub-feature set contains 1884 tripeptides. The jackknife cross-validated results were listed in Table 1. As it can be seen from Table 1, our method can correctly identify 89.8% (246/274) mycobacterial membrane proteins and 95.9% (283/295) non-membrane proteins. These results indicate that the proposed method is indeed very powerful in identifying mycobacterial membrane proteins.

Table 1 – The results for predicting mycobacterial membrane proteins.

Method	Sn (%)		OA (%)	AA (%)
	Mem	Non-mem		
SVM (1884 tripeptides)	89.8	95.9	93.0	92.9
Naïve Bayes (1243 tripeptides)	88.3	93.9	91.2	91.1
RBF Network (895 tripeptides)	86.5	91.9	89.3	89.2
Random forest (1443 tripeptides)	81.8	75.3	78.4	78.5
J48 tree (841 tripeptides)	71.2	66.4	68.7	68.8
SVM (PseAAC)	82.8	91.9	87.5	87.4
SVM (amino acids)	79.6	88.8	84.4	84.2
SVM (dipeptide)	74.8	91.9	83.7	83.3

Subsequently, the proposed method was used to predict the types of mycobacterial membrane proteins. We repeated the feature selection process for finding the optimized sub-feature set. By plotting the 3-D curve for CL, feature dimension and overall accuracy in Fig. 2, we found that the overall accuracy reached its maximum 93.1% when the CL_0 was selected as 99.29%. The sub-feature set contains 261 tripeptides. The results were recorded in Table 2. It shows that 75% (24/32) single-pass, 99.5% (191/192) multi-pass, 80% (16/20) lipid-anchor, and 80% (24/30) peripheral membrane proteins can be correctly predicted by our method. Such high accuracies demonstrate that the proposed method is an effective and powerful approach for predicting the types of mycobacterial membrane proteins.

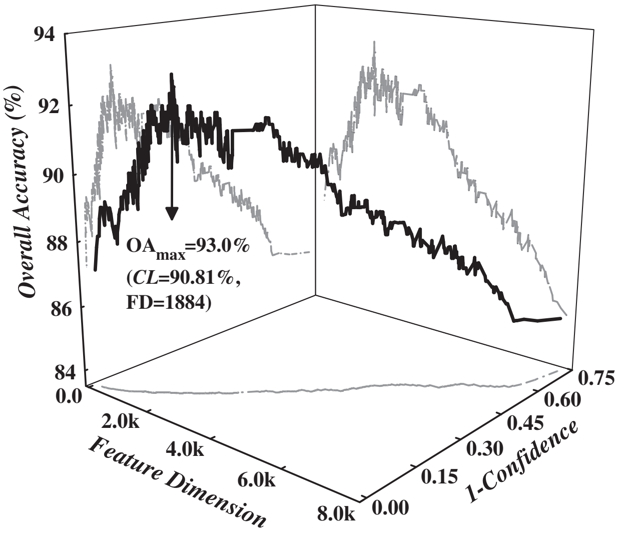


Fig. 1 – The 3-D graph for discriminating mycobacterial membrane proteins from non-membrane proteins.

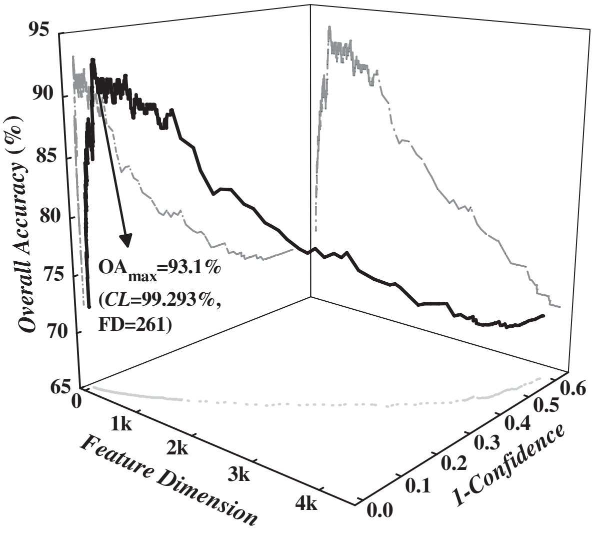


Fig. 2 – The 3-D graph for predicting the types of mycobacterial membrane proteins.

Table 2 – The results for predicting four types of mycobacterial membrane proteins on training set.

Method	Sn (%)				OA (%)	AA (%)
	Single	Multi	Lipid	Peripheral		
SVM (261 tripeptides)	75.0	99.5	80.0	80.0	93.1	83.6
Naïve Bayes (278 tripeptides)	53.1	95.3	90.0	46.7	84.7	71.3
RBF Network (265 tripeptides)	34.4	97.9	35.0	46.7	80.3	53.5
Random forest (283 tripeptides)	9.3	99.5	30.0	23.3	75.5	40.5
J48 tree (280 tripeptides)	37.5	83.3	50.0	16.7	68.2	46.9
SVM (PseAAC)	56.3	93.2	70.0	70.0	84.7	72.4
SVM (amino acids)	37.5	96.4	70.0	40.0	81.4	61.0
SVM (dipeptides)	34.3	95.8	45.0	36.7	78.5	53.0

3.2. *Comparing with other methods*

For testifying the superiority of our model, it is necessary to compare the capability of prediction models with that of other methods. Here, we performed many comparisons, including using different kinds of parameters such as optimized tripeptides, amino acid composition (AAC), dipeptide composition and pseudo amino acid composition (PseAAC) as well as using different kinds of algorithms such as J48 tree, Random forest, RBF Network, and Naïve Bayes. It should be mentioned that when using optimized tripeptides as inputs of these algorithms, we repeated the process of feature selection for obtaining the best feature sets. All jackknife test results were recorded in Tables 1 and 2. We noticed that the proposed method is superior to other methods, suggesting our proposed method is an effective method. Because the types of membrane proteins closely correlate with the biological function of membrane proteins, we did more experiments on the data of mycobacterial membrane protein as follows to compare the performance of the optimized tripeptides with other parameters.

A whole set of tripeptide contains 8000 components which accommodate enough information for classification. For tetrapeptide, there are 160,000 kinds. If using feature selection to optimize tetrapeptides, it is possible to achieve higher accuracy. To examine this hypothesis, binominal distribution was used again to filter tetrapeptides. For predicting four types of mycobacterial membrane proteins, results in Table 3 show that the maximum overall accuracy of 93.4% was achieved with an average accuracy of 83.6% by using 3256 tetrapeptides with CL>93.42%. We found that this result is almost as high as the accuracy (OA=93.1%, AA=83.6%) of optimized tripeptides. However, it is time-consuming for this feature set to search optimized parameters and build model. Moreover, the dimension of optimized tripeptide set is far less than that of optimized tetrapeptide set. The less the parameters are, the more robust the model is. In addition, the CL_0 of optimized tripeptides is 99.29% which is larger than that of optimized tetrapeptides, suggesting that the model built by optimized tripeptides is more credible. Thus, we propose using optimized tripeptides to perform prediction.

Position specific scoring matrix (PSSM) generated from PSI-BLAST is usually used for representation of biological sequences. Many works [12,21,56] have used PSSM to predict the types of membrane proteins and the subcellular localization of mycobacterial proteins. Thus we investigated the performance of PSSM for comparison. We executed the PSI-BLAST to generate PSSM by searching the protein dataset in GenBank. The *E*-value cutoff is set to 0.002. By inputting PSSM into SVM, we achieved the jackknife cross-validated accuracy of 81.4% for the prediction of four types of mycobacterial membrane proteins (Table 3). These results are not better than that of the proposed method.

Hybridizing different parameters to represent protein sequence is also an important strategy for improving predictive accuracy. By combining amino acid composition, sequence length and physiochemical properties of amino acids, Hayat and Khan have successfully predicted the membrane protein types [19,21]. Motivated by their work, we also evaluated the predictive performance of the fusion of tripeptides and dipeptides on our dataset. Because the dimensions between dipeptide and tripeptide are different, they were independently filtered by our feature reduction technique. For dipeptides, an optimized feature set containing 42 dipeptides was obtained, which achieves a maximum overall accuracy of 79.6% in jackknife cross-validation, and we then evaluated the accuracy of the fusion of optimized dipeptides and tripeptides. Results in Table 3 show that the overall accuracy is only 91.2% which is not better

Table 3 – Comparison with PSSM and hybridizing parameters on training set.

Method	Sn (%)				OA (%)	AA (%)
	Single	Multi	Lipid	Peripheral		
SVM (optimized tripeptides)	75.0	99.5	80.0	80.0	93.1	83.6
SVM (optimized tetrapeptides)	84.4	100.0	80.0	70.0	93.4	83.6
SVM (optimized dipeptides)	18.8	97.9	55.0	43.3	79.6	53.8
SVM (PSSM)	25.0	94.8	70.0	63.3	81.4	63.3
SVM (amino acids+dipeptide)	37.5	95.8	50.0	50.0	80.7	58.3
SVM (optimized dipeptide+optimized tripeptides)	68.8	99.0	75.0	76.7	91.2	79.8

than that of optimized tripeptides. Moreover, we also listed the accuracies of the fusion of AAC and dipeptides in Table 3. Comparison demonstrates that the optimized tripeptides can represent the sequence features of mycobacterial membrane proteins better than hybridizing parameters.

3.3. *Comparison on three benchmark datasets*

For further comparison, we tested the accuracies of our method on three different benchmark datasets. The first dataset was constructed by Fan and Li [35]. This mycobacterial membrane protein dataset contains 36 single-pass membrane, 248 multi-pass membrane and 30 peripheral membrane proteins. By executing our method on their dataset, 94.6% proteins can be correctly predicted. The results in Table 4 show that both the sensitivity (Sn) and specificity (Sp) of our method are higher than those of Fan and Li's method. In particular, the overall accuracy and average accuracy of our method are 9.6% and 19.6% higher than those of Fan and Li's method. This result demonstrates that our method is superior to Fan and Li's method.

The second dataset contains 125 independent membrane proteins described in dataset section. We compared the performance of different algorithms on this independent dataset. The predictive models were constructed on our training set. Results in Table 5 demonstrate again that the proposed method is superior to other methods. Furthermore, we also used the independent dataset to examine the performance of two online tools: MemType-2L and TMHMM. The MemType-2L can distinguish eight types of membrane proteins with the overall accuracy of >85%. However, this server can only correctly recognize 61.6% cases (Table 5), suggesting that our method is more suitable to study mycobacterial membrane proteins. The TMHMM can predict the transmembrane helices in proteins. According to the number of predicted transmembrane helices in a protein, we can judge this protein belonging to single-pass, multi-pass or non-pass membrane protein. Due to its excellent performance in the prediction of transmembrane helices, it can accurately identify 72.2% single-pass membrane proteins which is better than our method (Table 5). However, its recognition capability on multi-pass membrane proteins is a little poor than our method (Table 5). Furthermore, TMHMM is helpless in the discrimination between lipid-anchor and peripheral membrane proteins. Therefore, our method has great potential in mycobacterial membrane protein type prediction.

All of the above analyses focused on mycobacterial membrane protein. To argue the feasibility of this method on non-mycobacterial membrane protein, we tested the method on Chou and Shen's training and independent datasets [12]. The dataset contains eight types of membrane proteins and can be freely downloaded from http://www.csbio.sjtu.edu.cn/bioinf/MemType/Data.htm. As it can be seen from Table 6, our method can achieve an overall accuracy of 80.5% on training dataset and an overall accuracy of 86.5% on independent dataset. Some methods did outperform our method, but the accuracies of our method are still higher than those of Least Euclidean and ProtLoc. Most of these methods such as MemType-2L and GPP&KNN used the PSSM to perform predictions. The PSSM information is very effective and powerful due to the high sequence identity criterion (<80%) in this dataset. However, if a dataset has low sequence identity, such as our training data with the <40% sequence identity, the accuracy will reduce. Moreover, the PSSM of a protein depend much on the searching dataset. If no homologous is found in the searching dataset, the PSSM will not give exact description, which results in wrong prediction. Besides, we wanted to stress again that the accuracy of MemType-2L on mycobacterial membrane protein is not better than that of our method (see in Table 5), suggesting that our method has predominance on mycobacterial membrane protein. This result also demonstrates the feasibility of our method on non-mycobacterial membrane proteins.

Table 4 – Comparison of performance on Fan and Li's dataset.

	Our method		Fan and Li's method	
	Sn (%)	Sp (%)	Sn (%)	Sp (%)
Single	72.2	100	41.7	96.8
Multi	100	82.3	95.2	54.5
Peripheral	76.7	100	53.3	97.2
OA (%)	94.6		85.0	
AA (%)	83.0		63.4	

Table 5 – The comparison of performance on independent dataset.

Method	Sn (%)				OA (%)	AA (%)
	Single	Multi	Lipid	Peripheral		
Our method	50.0	97.3	87.5	75.0	85.6	77.5
Naïve Bayes	33.3	98.7	62.5	54.2	78.4	62.2
RBF Network	16.7	92.0	12.5	58.3	69.6	44.9
Random forest	22.2	100	62.5	33.3	73.6	54.5
J48 tree	27.8	89.3	50.0	41.7	68.8	52.2
MemType-2L	5.6	92.0	25.0	20.8	61.6	35.8
TMHMM	72.2	93.3	90.6		89.6	85.4

4. Conclusion

In this work, we developed a promising method to predict the mycobacterial membrane proteins and their types. A binomial distribution-based feature selection technique was proposed to select over-represented tripeptides. In the jackknife test,

Table 6 – The comparison of performance on Chou and Shen's dataset.

Method	Overall accuracy (%)	
	Jackknife test	Independent test
Our Method	80.5	86.5
MemType-2L	85.0	91.6
Least Euclidean distance	51.7	61.4
ProtLoc	52.0	37.2
LLDA&KNN	87.2	88.7
GPP&KNN	84.0	90.2
MVP&KNN	86.1	88.4
ENS2-BORDA	Non	91.0

our proposed model achieved an overall accuracy of 93.0% for the prediction of mycobacterial membrane proteins prediction, and an overall accuracy of 93.1% for the classification of mycobacterial membrane protein types. Based on this result, we constructed an online server, called MycoMemSVM, for predicting mycobacterial membrane proteins and their types, which is freely available at http://lin.uestc.edu.cn/server/MycoMemSVM. We believe that the server will be helpful for the vast majority of experimental scientists who focus on mycobacterium and antimicrobial drugs.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable suggestions and comments, which have led to the improvement of this paper. This work was supported by the Fundamental Research Funds for the Central Universities (ZYGX2009J081), the Scientific Research Foundation of Sichuan Province (2009JY0013), the National Natural Science Foundation of China (61202256, 61100092), the Scientific Research Startup Foundation of North China Coal Medical University (No. 10101115) and the Foundation of Scientific and Technological Department of Hebei Province (No. 11275532).

REFERENCE

[1] Niederweis M, Danilchanka O, Huff J, Hoffmann C, Engelhardt H. Mycobacterial outer membranes: in search of proteins. Trends Microbiol 2010;18:109-16.

[2] Smith I. *Mycobacterium tuberculosis* pathogenesis and molecular determinants of virulence. Clin Microbiol Rev 2003;16:463-96.

[3] He Z, De Buck J. Cell wall proteome analysis of *Mycobacterium smegmatis* strain MC2 155. BMC Microbiol 2010;10:121.

[4] Adams R, Worth CL, Guenther S, Dunkel M, Lehmann R, Preissner R. Binding sites in membrane proteins—diversity, druggability and prospects. Eur J Cell Biol 2012;91:326-39.

[5] Chen W, Luo L. Classification of antimicrobial peptide using diversity measure with quadratic discriminant analysis. J Microbiol Methods 2009;78:94-6.

[6] Hayat M, Khan A, Yeasin M. Prediction of membrane proteins using split amino acid and ensemble classification. Amino Acids 2012;42(6):2447-60.

[7] Chou KC, Elrod DW. Prediction of membrane protein types and subcellular locations. Proteins 1999;34:137-53.

[8] Cai YD, Zhou GP, Chou KC. Support vector machines for predicting membrane protein types by using functional domain composition. Biophys J 2003;84:3257-63.

[9] Wang M, Yang J, Liu GP, Xu ZJ, Chou KC. Weighted-support vector machines for predicting membrane protein types based on pseudo-amino acid composition. Protein Eng Des Sel 2004;17:509-16.

[10] Shen H, Chou KC. Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types. Biochem Biophys Res Commun 2005;334:288-92.

[11] Wang SQ, Yang J, Chou KC. Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition. J Theor Biol 2006;242:941-6.

[12] Chou KC, Shen HB. MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Commun 2007;360:339-45.

[13] Lin H. The modified Mahalanobis Discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J Theor Biol 2008;252:350-6.

[14] Walzer G, Rosenberg E, Ron EZ. Identification of outer membrane proteins with emulsifying activity by prediction of beta-barrel regions. J Microbiol Methods 2009;76:52-7.

[15] Yang XG, Luo RY, Feng ZP. Using amino acid and peptide composition to predict membrane protein types. Biochem Biophys Res Commun 2007;353:164-9.

[16] Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 2001;43:246-55.

[17] Wang J, Li Y, Wang Q, You X, Man J, Wang C, et al. ProClusEnsem: predicting membrane protein types by fusing different modes of pseudo amino acid composition. Comput Biol Med 2012;42:564-74.

[18] Mahdavi A, Jahandideh S. Application of density similarities to predict membrane protein types based on pseudo-amino acid composition. J Theor Biol 2011;276:132-7.

[19] Hayat M, Khan A. Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. J Theor Biol 2011;271(1):10-7.

[20] Wang M, Yang J, Xu ZJ, Chou KC. SLLE for predicting membrane protein types. J Theor Biol 2005;232:7–15.

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

[22] Wang T, Xia T, Hu XM. Geometry preserving projections algorithm for predicting membrane protein types. J Theor Biol 2010;262:208-13.

[23] Hayat M, Khan A. Mem-PHybrid: hybrid features-based prediction system for classifying membrane protein types. Anal Biochem 2012;424:35-44.

[24] Rezaei MA, Abdolmaleki P, Karami Z, Asadabadi EB, Sherafat MA, Abrishami-Moghaddam H, et al. Prediction of membrane protein types by means of wavelet analysis and cascaded neural networks. J Theor Biol 2008;254:817-20.

[25] Cai YD, Ricardo PW, Jen CH, Chou KC. Application of SVM to predict membrane protein types. J Theor Biol 2004;226:373-6.

[26] Liu H, Yang J, Wang M, Xue L, Chou KC. Using Fourier spectrum analysis and pseudo amino acid composition for prediction of membrane protein types. Protein J 2005;24: 385-9.

[27] Cai YD, Liu XJ, Chou KC. Artificial neural network model for predicting membrane protein types. J Biomol Struct Dyn 2001;18:607-10.

[28] Shen HB, Yang J, Chou KC. Fuzzy KNN for predicting membrane protein types from pseudo-amino acid composition. J Theor Biol 2006;240:9–13.

[29] Jia P, Qian Z, Feng K, Lu W, Li Y, Cai Y. Prediction of membrane protein types in a hybrid space. J Proteome Res 2008;7:1131-7.

[30] Bagos PG, Liakopoulos TD, Spyropoulos IC, Hamodrakas SJ. A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins. BMC Bioinformatics 2004;5:29.

[31] Ding H, Liu L, Guo FB, Huang J, Lin H. Identify Golgi protein types with modified Mahalanobis discriminant algorithm and pseudo amino acid composition. Protein Pept Lett 2011;18:58-63.

[32] Yeh JI, Mao L. Prediction of membrane proteins in *Mycobacterium tuberculosis* using a support vector machine algorithm. J Comput Biol 2006;13:126-9.

[33] Pajon R, Yero D, Lage A, Llanes A, Borroto CJ. Computational identification of beta-barrel outer-membrane proteins in *Mycobacterium tuberculosis* predicted proteomes as putative vaccine candidates. Tuberculosis 2006;86:290-302.

[34] Song H, Sandie R, Wang Y, Andrade-Navarro MA, Niederweis M. Identification of outer membrane proteins of *Mycobacterium tuberculosis*. Tuberculosis 2008;88:526-44.

[35] Fan GL, Li QZ. Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition. J Theor Biol 2012;304C:88-95.

[36] Magrane M, Consortium U. UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011 http://dx.doi.org/10.1093/database/bar009.

[37] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658-9.

[38] Ung P, Winkler DA. Tripeptide motifs in biology: targets for peptidomimetic design. J Med Chem 2011;54:1111-25.

[39] Anishetty S, Pennathur G, Anishetty R. Tripeptide analysis of protein structures. BMC Struct Biol 2002;2:9.

[40] Lin H, Ding H, Guo FB, Huang J. Prediction of subcellular location of mycobacterial protein using feature selection techniques. Mol Divers 2010;14:667-71.

[41] Ma J, Gu H. A novel method for predicting protein subcellular localization based on pseudo amino acid composition. BMB Rep 2010;43:670-6.

[42] Olivier I, Loots du T. A metabolomics approach to characterise and identify various *Mycobacterium* species. J Microbiol Methods 2012;88:419-26.

[43] Yin JB, Fan YX, Shen HB. Conotoxin superfamily prediction using diffusion maps dimensionality reduction and subspace classifier. Curr Protein Pept Sci 2011;12:580-8.

[44] Huang T, Xu Z, Chen L, Cai YD, Kong X. Computational analysis of HIV-1 resistance based on gene expression profiles and the virus–host interaction network. PLoS One 2011;6:e17291.

[45] Lin H, Ding H. Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. J Theor Biol 2011;269:64-9.

[46] Lin H, Chen W. Prediction of thermophilic proteins using feature selection technique. J Microbiol Methods 2011;84:67-70.

[47] Wang T, Yang J, Shen HB, Chou KC. Predicting membrane protein types by the LLDA algorithm. Protein Pept Lett 2008;15:915-21.

[48] Feng Y, Luo L. Use of tetrapeptide signals for protein secondary-structure prediction. Amino Acids 2008;35:607-14.

[49] Ding CH, Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 2001;17:349-58.

[50] Lin H, Ding H, Guo FB, Zhang AY, Huang J. Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein Pept Lett 2008;15:739-44.

[51] Kumar M, Gromiha MM, Raghava GP. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. J Mol Recognit 2011;24:303-13.

[52] Chen C, Chen L, Zou X, Cai P. Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. Protein Pept Lett 2009;16:27-31.

[53] Chou KC, Cai YD. Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 2002;277:45765-9.

[54] Fan RE, Chen PH, Lin CJ. Working set selection using second order information for training support vector machines. J Mach Learn Res 2005;6:1889-918.

[55] Chou KC, Zhang CT. Prediction of protein structural classes. Crit Rev Biochem Mol Biol 1995;30:275-349.

[56] Rashid M, Saha S, Raghava GP. Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. BMC Bioinformatics 2007;8:337.