Journal of Theoretical Biology 304 (2012) 88–95

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

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

Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition

Guo-Liang Fan, Qian-Zhong Li*

Department of Physics, School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China

ARTICLE INFO

Article history:
Received 17 December 2011
Received in revised form
13 March 2012
Accepted 14 March 2012
Available online 23 March 2012

Keywords:
Jackknife validation
Mycobacterium tuberculosis
Reduced amino acids
Evolutionary information
Average chemical shift

ABSTRACT

Mycobacterium tuberculosis (MTB) is a pathogenic bacterial species in the genus *Mycobacterium* and the causative agent of most cases of tuberculosis (Berman et al., 2000). Knowledge of the localization of *Mycobacterial* protein may help unravel the normal function of this protein. Automated prediction of *Mycobacterial* protein subcellular localization is an important tool for genome annotation and drug discovery. In this work, a benchmark data set with 638 non-redundant *mycobacterial* proteins is constructed and an approach for predicting *Mycobacterium* subcellular localization is proposed by combining amino acid composition, dipeptide composition, reduced physicochemical property, evolutionary information, pseudo-average chemical shift. The overall prediction accuracy is 87.77% for *Mycobacterial* subcellular localizations and 85.03% for three membrane protein types in Integral membranes using the algorithm of increment of diversity combined with support vector machine. The performance of pseudo-average chemical shift is excellent. In order to check the performance of our method, the data set constructed by Rashid was also predicted and the accuracy of 98.12% was obtained. This indicates that our approach was better than other existing methods in literature.

1. Introduction

Mycobacterium tuberculosis is a gram-positive bacterium which causes tuberculosis (Berman et al., 2000), the leading cause of infectious disease mortality. About 1.5 million people die from tuberculosis each year, and it is thought that as many as 2 billion people may be infected with *M. tuberculosis* (Frieden et al., 2003). There are about 8–9 million new cases of TB annually, with much of the burden falling on the young and middle aged between 15 and 49 years old (Cegielski et al., 2002). Drug resistant tuberculosis is a significant and growing public health threat. A recent report on the Global Project on anti-tuberculosis drug resistance surveillance found a median prevalence of resistance to at least one anti-tuberculosis drug of 10.2% while multidrug-resistant (MDR) TB prevalence reached as high as 14.2%. With the appearance of vast genomic and proteomic data, there was great opportunity to treat this disease. One of the key features of a protein is cellular localization which gives important information about its functions and help in understanding the biological processes at the cellular level. Thus it is important to develop method for accurately predicting subcellular localization of a protein of a pathogenic organism like *mycobacterium*, which may assist to anti-tuberculosis drug design (Singh and Somvanshi, 2010).

Knowing protein localization is an important step to understand its function. The system for prediction protein subcellular location had been developed during the last two decades (Chou, 2001; Chou and Elrod, 1999; Chou and Cai, 2002; Nakashima and Nishikawa, 1994; Zhou and Doctor, 2003). Various features of sequence and a number of machine learning approached had been introduced for prediction protein subcellular location. Some comprehensive reviews described most of these methods in detail (Chou and Shen, 2007a; Nakai, 2000). Significant progress has been achieved in predicting protein subcellular localization with the establishment of various organism-specific benchmark data sets. Also, the predictors that can be used to deal with proteins with multiple subcellular locations have been constructed (Chou and Shen, 2010a, 2010b; Chou et al., 2011a, 2011b; Wu et al., 2011, 2012; Xiao et al., 2011a, 2011b).

Recently, some computational methods for predicting *mycobacterial* protein subcellular localizations have been proposed in literature: TBpred (Rashid et al., 2007) and MycobacSVM (Lin et al., 2010). The data set was constructed by Rashid (Rashid et al., 2007), which contained total 852 proteins with sequence identity 100% and considered four main subcellular locations: cytoplasmic, integral membranes, secretory and membrane attached proteins by a lipid anchor. TBpred used the amino acid composition

* Corresponding author. Tel.: +86 471 4993145; fax: +86 471 4993141.
E-mail address: qzli@imu.edu.cn (Q.-Z. Li).

0022-5193/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2012.03.017

(AAC), Position-Specific Scoring Matrix profiles (PSSM), Hidden Markov Model (HMM), and Multiple Motif Elicitation/Motif Alignment (Scharfe et al., 2000) and Search Tool (MEME/MAST) to construct the feature vector to train and test the redundant data set. The predictive accuracy is 87.0% for Cytoplasmic, 85.3% for integral membrane, 92.0% for secretory, 91.7% for membrane attached proteins and 86.8% for overall accuracy in using five-cross validation. MycobacSVM reduced the sequence identity to 80%, and enhanced the overall jackknife cross-validation predictive accuracy to 91.2% by using pseudo amino acid composition (PseAAC), the reduced amino acids in N-terminus and non N-terminus of proteins as information parameters. But the accuracy was 71.1% for secretory and 88.3% for membrane attached proteins, lower than those of TBpred.

In this article, we constructed the benchmark *Mycobacterium* data set which has 638 proteins with sequence identity 40% (denoted as M638) derived from SWISS-PROT(Release 2011_11 – Nov 16, 2011) (Wu et al., 2006) and used ID_SVM approach to predict the subcellular locations of *Mycobacterial* proteins. Five representative features are used, including Amino acid composition (AAC), dipeptide composition (DC), reduced physicochemical properties (Hn), evolutionary information (PSSM) and a novel constructed feature, pseudo-average chemical shift (PseACS). The DC was input to the ID, and then the output of ID and feature of AAC, Hn, PSSM and PseACS were selected as an input to multiclass SVM. Here, the overall predictive accuracy is 87.77% for subcellular locations, and 85.03% in three membrane types of integral membrane proteins. In order to compare the prediction performance, we get the overall predictive accuracy of 94.00% in jackknife tests for subcellular locations of the data set with 450 proteins (denoted as M450) constructed by Rashid which the sequence identity has been reduced to 80%, and 98.12% in five-cross validation tests for subcellular locations of the data set of 852 proteins (denoted as M852) constructed by Rashid.

According to a recent comprehensive review (Chou, 2011), to establish a really useful statistical predictor for a protein system, we need to consider the following procedures: (1) construct or select a valid benchmark data set to train and test the predictor; (2) formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the attribute to be predicted; (3) introduce or develop a powerful algorithm to operate the prediction; (4) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (5) establish a user-friendly Web server for the predictor that is accessible to the public. Below, we describe how to deal with these steps.

2. Materials and methods

2.1. Data sets

The *Mycobacterial* data set was constructed from SWISS-PROT (Release 2011_11 – Nov 16, 2011)(Wu et al., 2006) by searching with 'KW' containing '*Mycobacterium*', and then following steps are used to confine the benchmark data set. (1) The sequences which have any ambiguous annotation words like 'probable', 'potential', 'possible', 'by similarity' are excluded. (2) The sequences containing ambiguous residues such as 'X', 'B' 'and Z' are removed. (3) The sequences which annotated with 'fragment' are excluded. (4) The sequences' length less than 15 are dropped. (5) To remove the homologous sequences from the benchmark data set, a cutoff threshold of 25% was imposed in (Chou and Shen, 2007a, 2008; Chou and Shen, 2010a, 2010b) to exclude those proteins from the benchmark data set that have equal to or greater than 25% sequence identity to any other in a same subset. However, in this study we did not use such a stringent criterion because the currently available data do not allow us to do so. Otherwise, the numbers of proteins for some subsets would be too few to have statistical significance. we use the CD-HIT (Li et al., 2001) program to exclude the proteins with sequence identity high than 40%.

As is well known, proteins may simultaneously exist at, or move between, two or more different subcellular locations. Proteins with multiple locations or dynamic feature of this kind are particularly interesting because they may have some very special biological functions intriguing to investigators in both basic research and drug discovery. The web-servers iLoc-Euk (Chou et al., 2011a), iLoc-Hum (Chou et al., 2011b), iLoc-Plant (Wu et al., 2011), iLoc-Gpos (Wu et al., 2012), iLoc-Gneg (Xiao et al., 2011a), and iLoc-Virus (Xiao et al., 2011b) can be used to cope with the multiple location problems in eukaryotic, human, plant, Gram-positive, Gram-negative, and virus proteins, respectively.

In this study, we exclude the proteins located in more than one location, because the number of multiplex proteins in the existing *mycobacterial* protein database is not large enough to construct a statistically meaningful benchmark data set for studying the case of multiple locations.

Finally, 638 *Mycobacterial* protein sequences are obtained and listed in Table 1, which were classified into four subcellular locations according to its annotation, including 265 cytoplasmic, 314 integral membranes, 29 secretory and 30 proteins attached to the membrane by a lipid anchor. For the 314 integral membranes proteins, we divide them into 3 membrane protein types according to their annotation. The data sets are listed on our website (http://wlxy.imu.edu.cn/college/biostation/fuwu/myco/index.asp) and can be obtained from the author. The distribution of *Mycobacterium* data set constructed by Rashid et al. (2007) also are listed in Table 2.

2.2. Feature vectors

To develop a powerful predictor for a protein system, one of the keys is to formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted (Chou, 2011). To realize this, the concept of pseudo amino acid composition (PseAAC) was proposed (Chou, 2001) to replace the simple amino acid composition (AAC) for representing the sample of a protein. Ever since the concept of PseAAC was introduced, it has been widely used to study various problems in proteins and protein-related systems (see, e.g., (Chen et al., 2009; Ding et al., 2009; Esmaeili et al., 2010; Georgiou et al., 2009; Gu et al., 2010; Jiang et al., 2008; Li and Li, 2008b; Lin, 2008; Lin et al., 2008; Mohabatkar, 2010; Mohabatkar et al., 2011; Qiu et al., 2010; Yu et al., 2010; Zeng et al., 2009; Zhang et al., 2008; Zhou et al., 2007)). For various different modes of PseAAC, see (Chou and Shen, 2009). According to a recent comprehensive review (Chou, 2011), the general form of PseAAC can be formulated as (see Eq. (6)

Table 1
The distribution of *Mycobacterium* data set M638.

Label	Compartment		Sequence no.
1	Cytoplasmic		265
2	Integral membranes	Single-pass membrane protein	36
		Peripheral membrane protein	30
		Multi-pass membrane protein	248
3	Secretory		29
4	Attached to the membrane		30
Total			**638**

Table 2
The distribution of *Mycobacterium* data set constructed by Rashid.

Label	Compartment	Sequence no. (100% identity)	Sequence no. (80% identity)	Sequence no. (30% identity)
1	Cytoplasmic	340	151	116
2	Integral membranes	402	238	178
3	Secretory	50	25	10
4	Attached to the membrane	60	36	26
Total		852 (denoted as M852)	450 (denoted as M450)	330 (denoted as M330)

of (Chou, 2011)):

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

where T is a transpose operator, while the subscript Ω is an integer and its value as well as the components ψ_1, ψ_2, … will depend on how to extract the desired information from the amino acid sequence of P. Here, we are to use a combination of the amino acid composition, dipeptide composition, reduced physicochemical property, evolutionary information, and pseudo-average chemical shift to represent the protein samples, and $\Omega=202$.

2.2.1. Amino acid composition

Some researchers have pointed out that proteins localized in same subcellular location have similar amino acid composition (AAC) which may reflect the physicochemical properties; they are adapted to the micro environment (Andrade et al., 1998). So we considered the amino acid composition. The sequence was averagely divided into three segments. The absolute occurrence frequencies of 20 amino acids from each segment were calculated, and then these vectors from each segment were merged together. Thus, the feature vector of AAC can be expressed by $20\times3=60$D coordinates.

2.2.2. Dipeptide composition

Dipeptide composition vectors contained information regarding the frequency as well as the local order of amino acid pairs in a given sequence and describe proteins using 400 features and had been extensively used to represent protein sequences (Chou, 1999; Gao et al., 2005; Idicula-Thomas et al., 2006).

We used the DC of two consecutive residues to express the sequence order information. The protein sequence was also divided into three segments, and the dimension of DC is $400\times3=1200$D. In order to reduce the dimension of DC, the ID algorithm was used and the DC of 1200D was input into ID. The ID algorithm reduced DC into 12D, and improved the accuracy at the same time.

2.2.3. Reduced physicochemical properties

Amino acid composition (AAC) reflects the occurrence frequencies of the 20 common amino acids in a protein sequence, and physicochemical properties represent the structure and function, the electronic property and the solubility of residues in this sequence, such as Strongly hydrophilic or polar (R,D,E,N,Q,K,H), strongly hydrophobic (L,I,V,A,M,F), weakly hydrophilic or weakly hydrophobic (S,T,Y,W), Proline (P), Glycine (G) and Cysteine (C) (Chen and Li, 2007; Li and Li, 2008a; Li and Li, 2008b). So we use 6 characters to represent the 20 amino acids according to following physicochemical properties: The protein sequence was divided into six regions, and then the segment numbers of successive same character composition of six characters in each segment was chosen, for example, 'T', 'TTT' or 'TTTTT' was counted as 1 hit of 'T' composition similarly.

2.2.4. Evolutionary information

In past, multiple sequence alignment information in form of position specific scoring matrix (PSSM) (Schaffer et al., 2001) had been used for developing methods (Chou and Shen, 2007b; Jones, 1999; Kaur and Raghava, 2004). In this study, PSSM has been used for predicting *Mycobacterial* proteins subcellular location. To use the evolution information, the position specific scoring matrix (PSSM) was generated by using PSI-Blast program (Schaffer et al., 2001) to search the SWISS-PROT database (released on 14 May 2011)(Wu et al., 2006) through three iterations with 0.001 as the E-value cutoff for multiple sequence alignment against the protein sequence P, then we use the standardization procedure to normalization.

$$V_{i\to j}=(V^0_{i\to j}-\overline{V^0_i})/SD(V^0_i)(i=1,2\cdots L;j=1,2\cdots20) \tag{2}$$

where $V^0_{i\to j}$ is the score directly obtained by PSI-Blast, $\overline{V^0_i}$ is the mean of $V^0_{i\to j}$ over 20 amino acids, $SD(V^0_i)$ is the standard deviation of $V^0_{i\to j}$, L is the length of protein sequence. Then the PSSM becomes:

$$P_{\text{PSSM}}=\begin{bmatrix} V_{1\to1} & V_{1\to2} & & V_{1\to20} \\ V_{2\to1} & V_{2\to2} & & V_{2\to20} \\ & & & \\ V_{i\to1} & V_{i\to2} & & V_{i\to20} \\ & & & \\ V_{L\to1} & V_{L\to2} & \cdots & V_{L\to20} \end{bmatrix} \tag{3}$$

In order to use the sequence order information, we adapt the concept of pseudo amino acid composition (Chou, 2001), and obtained the PsePSSM by the following equations:

$$P^\lambda_{\text{PsePSSM}}=[\theta^\lambda_1,\theta^\lambda_2,\ldots,\theta^\lambda_i,\ldots\theta^\lambda_{20}] \tag{4}$$

$$\theta^\lambda_i=\sum_{j=1}^{L-\lambda}[V_{j\to i}-V_{(j+\lambda)\to i}]^2/(L-\lambda)(i=1,2\cdots20;\lambda<L) \tag{5}$$

where θ^λ_i is the correlation factor of amino acid type i, whose contiguous distance is λ along the protein sequence. Especially, for $\lambda=0$, θ^0_i becomes the average score of the amino acid residues in the protein P, which is changed to amino acid type i during the evolution process.

We select $\lambda=2$, and then the PsePSSM would be expressed as:

$$P_{\text{PsePSSM}}=[\theta^0_1,\theta^0_2,\ldots,\theta^0_{20},\theta^1_1,\theta^1_2,\ldots,\theta^1_{20},\theta^2_1,\theta^2_2,\ldots,\theta^2_{20}] \tag{6}$$

2.2.5. Pseudo-average chemical shift

Protons are sensitive to their chemical environment, and protons in different chemical environments experience slightly different magnetic fields and absorb at different frequencies. The resonance frequencies of the different protons are expressed as chemical shifts relative to a standard value.

Chemical shifts, among the most important parameters are measured by NMR spectroscopy. They are sensitive to local environments and can be used as indicators of local conformations. As an important example, the chemical shifts of protein backbone atoms are known to correlate strongly with the backbone dihedral angles or secondary structure types (Luginbuhl et al., 1995; Spera and Bax, 1995; Wishart et al., 1991).

Several works pointed out that the averaged chemical shift (ACS) of a particular nucleus in the protein backbone correlates well to its secondary structure (Mielke and Krishnan, 2003;

Sibley et al., 2003; Zhao et al., 2010), and the protein functions are determined by its structure.

For a certain protein sequence *P*, we obtained the second structure from Porter (http://distill.ucd.ie/porter/) (Pollastri and McLysaght, 2005; Pollastri et al., 2007), which is a server for predicting the protein's second structure. As described in our previous study (Fan and Li, 2011), every amino acid in the sequence is replaced by its ACS. Then *P* is expressed as:

$$P=[C_1^i,C_2^i\cdots C_L^i](i={}^{15}N,{}^{13}C_\alpha,{}^{1}H_\alpha,{}^{1}H_N) \tag{7}$$

We select $\lambda=16$ and $i={}^1H_\alpha\ ,{}^{13}C_\alpha$, then the PseACS would be expressed as:

$$P_{\mathrm{PseACS}}=[\varphi_1^0,\varphi_1^1,\ldots,\varphi_1^{16},\varphi_2^0,\varphi_2^1,\ldots,\varphi_2^{16}] \tag{8}$$

$$\varphi_i^\lambda=\sum_{k=1}^{L-\lambda}[C_k^i-C_{k+\lambda}^i]^2/(L-\lambda)(i={}^1H_\alpha,{}^{13}C_\alpha;\lambda<L) \tag{9}$$

In order to better use the PseACS, we also established a user-friendly Web server PseACS (http://wlxy.imu.edu.cn/college/biostation/fuwu/PseACS/index.asp), which is accessible to the public.

2.3. Methods

2.3.1. Increment of diversity

In a state space of *d* dimension, n_i indicates the absolute frequency of the *i*th state. The standard diversity measure for diversity source X:{n_1, n_2, …, n_i, …, n_d} is defined as (Li and Lu, 2001):

$$D(X)=N\log N-\sum_{i=1}^{d}n_i\log_b n_i \tag{10}$$

where $N=\Sigma_{i=1}^{d}n_i$,log(0) = 0 If $n_i=0$.

In general, for two sources of diversity in the same parameter space of d dimensions X:{n_1, n_2, …, n_i, …, n_d} and Y:{m_1, m_2, …, m_i, …, m_d}, the increment of diversity (ID), denoted by *ID(X, Y)*, is defined as:

$$ID(X,Y)=D(X+Y)-D(X)-D(Y) \tag{11}$$

where $D(X+Y)$ is the measure of diversity of the sum of two diversity sources called combination diversity source space.

ID is the method for measuring the similarity level of two diversity sources. If X similar to Y, then $ID(X,Y)$ will be small, especially if $X=Y$, $ID(X,Y)=0$.

2.3.2. Support vector machine

SVM is machine learning algorithm based on statistical learning theory (Vapnik, 1998), which can be widely used for classification. In recent years, SVM-based machine learning algorithm has also been used in predicting membrane protein type (Cai et al., 2003a; Cai et al., 2004a), protein subcellular location (Chou and Cai, 2002), protein structural class (Cai et al., 2002d), specificity of GalNAc-transferase (Cai et al., 2002c), HIV protease cleavage sites in protein (Cai et al., 2002b), beta-turn types (Cai et al., 2002a), protein signal sequences and their cleavage sites (Cai et al., 2003b), alpha-turn types (Cai et al., 2003c), catalytic triads of serine hydrolases (Cai et al., 2004b), among many others.

In this work, we used the free software LIBSVM (Chang and Lin, 2011) to predict submitochodria locations. A radial basis function (RBF) was chosen as the kernel function. For multi-classification, SVM uses a one-versus-one strategy, and construct $k\times(k-1)/2$ classifiers and voting strategy to assign the class for an unknown protein.

Because of different kinds of feature vectors and large dimensions, we must first reduce the dimension. For the feature vector of DC, its dimension is 1200D, and after using the ID algorithm, the dimension was reduced to 12D. Finally, five parts of the feature vector were combined together to form a 202 dimension feature vector and inputted into SVM for training to select the best c and g for classifier of predicting *Mycobacterium* subcellular locations. If a protein predicted as an *Integral membranes* protein, then it would be send into the classifier for predicting membrane types.

3. Results and discussion

3.1. Evaluation methods

In statistical prediction, the following three testing methods are often used to examine a predictor for its effectiveness in practical application: independent data set test, sub-sampling test, and jackknife test (Chou and Zhang, 1995). However, as elucidated by (Chou and Shen, 2008) and demonstrated in (Chou and Shen, 2007a), among the three testing methods, the jackknife test is deemed the most objective one (Feng, 2002) and can always yield a unique result for a given benchmark data set; hence, it has been increasingly used by investigators to examine the accuracy of various predictors (Bi et al., 2011; Hayat and Khan, 2011; Kandaswamy et al., 2011; Lin, 2008; Lin and Ding, 2011; Lin et al., 2008; Liu et al., 2010; Zakeri et al., 2011; Zhang and Fang, 2008; Zhou and Assa-Munt, 2001; Zhou and Doctor, 2003; Zhou et al., 2007). During the jackknife test process, each protein is singled out in turn as a test sample; the remaining proteins are used as training set to calculate test sample's membership and predict the class.

The prediction performance was evaluated by the sensitivity (*Sn*), specificity (*Sp*) (Schaffer et al., 2001), positive predictive value (*PPV*), accuracy (Scharfe et al., 2000), Mathew's correlation coefficient (*MCC*) (Matthews, 1975) and average accuracy (*Aa*), which defined as follows:

$$Sn=TP/(TP+FN) \tag{12}$$

$$Sp=TN/(TN+FP) \tag{13}$$

$$PPV=TP/(TP+FP) \tag{14}$$

$$Acc=(TP+TN)/(TP+FN+TN+FP) \tag{15}$$

$$MCC=[(TP\times TN)-(FP\times FN)]/\sqrt{(TP+FN)\times(TN+FN)\times(TP+FP)\times(TN+FP)} \tag{16}$$

$$Aa=\sum Sn/\xi \tag{17}$$

where, *TP* denotes the numbers of the correctly predicted positives, *FN* denotes the numbers of the positives predicted as negatives, *FP* denotes the numbers of the negatives predicted as positives, and *TN* denotes the numbers of correctly predicted negatives, ξ denotes the number of classes.

3.2. Results on leave-one-out tests for M638

Two kinds of SVMs are constructed for four *Mycobacterium* subcellular locations and three types of *Mycobacterium Integral membranes* protein using five kinds feature vectors, which reduced dimension by ID algorithm. The RBF kernel function and the grid-search approach was used to find the best parameters of *C* and γ for SVMs. Finally, the $C=32$, and $\gamma=0.03125$ for SVM of four *Mycobacterium* subcellular locations, and the $C=8$, and $\gamma=0.0078125$ for SVM of three types of *Mycobacterium Integral membranes* protein were obtained. The prediction results for *Mycobacterium* subcellular locations of the data set M638 are shown in Table 3, and the predictive accuracy for three membrane protein types is also shown in Table 4.

From Tables 3 and 4, we can see that the prediction performance is quite good; although the identity and homology of sequences had been reduced to 40%, the total jackknife validation accuracy of 87.77% for *Mycobacterium* subcellular locations and 85.03% for three membrane protein types were achieved.

3.3. Comparison with other methods

In order to assess the performance of our predictor, we applied our method to the data set M330, M450 and M852 which constructed by Rashid. In Table 5, the results which predicted by MycobacSVM (Lin et al., 2010) was compared with our method for the data set M450 with 80% cutoff (Lin et al., 2010). Using the ID and SVM algorithm and combined several feature vectors, 94.00% accuracy was obtained in the jackknife validation. It was 2.8% higher than MycobacSVM (Lin et al., 2010). We also compared the performance of method for on the jackknife validation of data set M330 with 30% cutoff (Lin et al., 2010). The sensitivities of cytoplasmic proteins, integral membrane proteins, secretory proteins and membrane attached proteins were 95.69%, 96.63%, 40.0% and 80.77%, respectively. The overall accuracy of 93.33% with average accuracy of 78.27% was achieved. The overall accuracy just lost 0.67% with the sequence identity decreasing from 80% to 30%, and 3.33% higher than MycobacSVM (Lin et al., 2010). In Table 6, for the data set M852 (Rashid et al., 2007), the overall accuracy reached 98.12% when the five-cross validation used. When using AAC, DC-ID, Hn, PseACS and PSSM as feature vector independently, the prediction results for *Mycobacterium*

Table 3
The predictive accuracy for *Mycobacterium* locations in data set M638.

Submitochondria locations	TP	TN	FP	FN	ACC(%)	MCC
Cytoplasmic	247	343	30	18	92.5	0.847
Integral membranes	284	285	39	30	89.2	0.784
Secretory	10	605	4	19	96.4	0.481
Attached to the membrane	19	603	5	11	97.5	0.695
Overall accuracy(%)	87.77					
Average accuracy(%)	70.37					

Table 4
The predictive accuracy for three membrane protein types in *Integral membranes* of M638.

Membrane protein types	TP	TN	FP	FN	ACC(%)	MCC
Single-pass membrane	15	269	9	21	90.5	0.461
Peripheral membrane	16	276	8	14	93.0	0.559
Multi-pass membrane	236	36	30	12	86.6	0.563
Overall accuracy(%)	85.03					
Average accuracy(%)	63.39					

Table 5
Comparison of predictive accuracy for *Mycobacterium* locations of M450 with other methods.

	Our's		MycobacSVM (Lin et al., 2010)	
	Sn (%)	Sp (%)	Sn (%)	Sp (%)
Cytoplasmic	96.7	96.7	96.6	91.0
Integral membranes	97.1	94.3	93.7	92.5
Secretory	64.0	99.8	47.8	68.8
Attached to the membrane	83.3	99.0	80.6	93.5
Overall accuracy (%)	94.00		91.2	
Average accuracy (%)	85.27		79.7	

Table 6
Comparison of predictive accuracy (sensitivity) for *Mycobacterium* locations of M852 with other methods.

	Our's	TBpred (Rashid et al., 2007)		MycobacSVM (Lin et al., 2010)
		Hybrid model	SVM+PSSM	
Cytoplasmic	99.1	87.0	94.7	96.6
Integral membranes	98.8	85.3	87.8	94.3
Secretory	90.0	92.0	44.0	71.1
Attached to the membrane	95.0	91.7	68.3	88.3
Overall accuracy (%)	98.12	86.8	86.6	93.5

Table 7
The contribution (sensitivity (%)) of each feature vector to *Mycobacterium* locations of M638, M852 and M450.

	Success rate (sensitivity) (%)						
	AAC	DC-ID	Hn	PseACS	PSSM	Ac	Aa
M638	76.96	76.65	77.12	75.55	86.05	87.77	70.37
M852	93.54	91.08	87.79	84.27	96.48	98.12	95.72
M450	85.78	82.0	81.78	79.56	92.66	94.0	85.27

Table 8
Comparison of predictive accuracy for *Mycobacterium* locations of M638 with other methods.

	Ours		TBpred (Rashid et al., 2007) Hybrid model (E value=10)		MycobacSVM (Lin et al., 2010)	
	ACC **(%)**	MCC	ACC **(%)**	MCC	ACC **(%)**	MCC
Cytoplasmic	92.48	0.847	62.70	0.193	83.54	0.674
Integral membranes	89.18	0.784	69.75	0.409	79.94	0.599
Secretory	96.39	0.481	86.21	0.272	95.61	0.318
Attached to the membrane	97.49	0.695	72.57	0.268	96.08	0.513
Overall accuracy (%)	87.77		45.61		77.59	
Average accuracy (%)	70.37		57.19		58.09	

subcellular locations of the data set M638, M852, M450 are listed in Table 7, in which the independent contribution from each feature vector to predictive results are shown for *Mycobacterium* locations of M638, M852 and M450. We also compared the performance of our method with Rashid's and Lin's in the data set of ours' M638 with 40% cutoff, and the results are listed in Table 8. From the results, we can see that performance of our predictor is best because of its high accuracy and strong robustness.

4. Conclusions

In this work, a benchmark data set of *Mycobacterial* proteins was constructed, which the protein identity was reduced to 40%. The data set has high quality than the data set constructed by Rashid, and can be used to perform much detailed research about *Mycobacterium*.

The various features of *Mycobacterial protein* are considered, and an ID algorithm and SVM to construct the classifier. By using this method, we obtained 87.77% on the jackknife validation of our data set, and 98.12% with the data set of Rashid, which is better than the best approach literature.

Studying (Chothia and Lesk, 1986; Dickerson et al., 1976; Pastore and Lesk, 1990) reveals that structure is more conserved than sequence. In other words, proteins with different sequences can adopt the same 3D structure. So the structure information was more important in predicting subcellular locations. In our work, a novel constructed feature PseACS was proposed, and the performance of PseACS is also excellent for predicting accuracy of *Mycobacterium* subcellular locations, it represents the more structure information of a protein. Therefore, PseACS can be an effective tool for future proteomics studies.

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

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgements

The authors would like to thank the reviewers for their helpful comments on our manuscript. This work was supported by a grant from National Natural Science Foundation of China (61063016, 31160188), The Research Fund for the Doctoral Program of Higher Education of China (no. 20101501110004), The Science and Technology Foundation of Ministry of Education of China (211030).

References

Andrade, M.A., O'Donoghue, S.I., Rost, B., 1998. Adaption of protein surface to subcellular location. J. Mol. Biol., 517–525.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucleic Acids Res. 28, 235–242.

Bi, J., Yang, H., Yan, H., Song, R., Fan, J., 2011. Knowledge-based virtual screening of HLA-A*0201-restricted CD8+ T-cell epitope peptides from herpes simplex virus genome. J. Theor Biol. 281, 133–139.

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

Cai, Y.D., Lin, S.L., Chou, K.C., 2003b. Support vector machines for prediction of protein signal sequences and their cleavage sites. Peptides 24, 159–161.

Cai, Y.D., Liu, X.J., Xu, X.B., Chou, K.C., 2002a. Support vector machines for the classification and prediction of beta-turn types. J. Pept. Sci. 8, 297–301.

Cai, Y.D., Liu, X.J., Xu, X.B., Chou, K.C., 2002b. Support vector machines for predicting HIV protease cleavage sites in protein. J. Comput. Chem. 23, 267–274.

Cai, Y.D., Liu, X.J., Xu, X.B., Chou, K.C., 2002c. Support vector machines for predicting the specificity of GalNAc-transferase. Peptides 23, 205–208.

Cai, Y.D., Liu, X.J., Xu, X.B., Chou, K.C., 2002d. Prediction of protein structural classes by support vector machines. Comput. Chem. 26, 293–296.

Cai, Y.D., Feng, K.Y., Li, Y.X., Chou, K.C., 2003c. Support vector machine for predicting alpha-turn types. Peptides 24, 629–630.

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

Cai, Y.D., Zhou, G.P., Jen, C.H., Lin, S.L., Chou, K.C., 2004b. Identify catalytic triads of serine hydrolases by support vector machines. J. Theor. Biol. 228, 551–557.

Cegielski, J.P., Chin, D.P., Espinal, M.A., Frieden, T.R., Rodriquez Cruz, R., Talbot, E.A., Weil, D.E., Zaleskis, R., Raviglione, M.C., 2002. The global tuberculosis situation. Progress and problems in the 20th century, prospects for the 21st century. Infect Dis. Clin. North Am. 16, 1–58.

Chang, C.C., Lin, C.J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2. 27:1–27:27.

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

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

Chothia, C., Lesk, A.M., 1986. The relation between the divergence of sequence and structure in proteins. EMBO J. 5, 823–826.

Chou, K.C., 1999. Using pair-coupled amino acid composition to predict protein secondary structure content. J. Protein Chem. 18, 473–480.

Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43, 246–255.

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

Chou, K.C., Zhang, C.T., 1995. Prediction of protein structural classes. Crit Rev. Biochem. Mol. Biol. 30, 275–349.

Chou, K.C., Elrod, D.W., 1999. Protein subcellular location prediction. Protein Eng. 12, 107–118.

Chou, K.C., Cai, Y.D., 2002. Using functional domain composition and support vector machines for prediction of protein subcellular location. J. Biol. Chem. 277, 45765–45769.

Chou, K.C., Shen, H.B., 2007a. Recent progress in protein subcellular location prediction. Anal. Biochem. 370, 1–16.

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

Chou, K.C., Shen, H.B., 2008. Cell-PLoc: a package of web servers for predicting subcellular localization of proteins in various organisms. Nat. Protoc. 3, 153–162.

Chou, K.C., and Shen, H.B., 2009. Review: recent advances in developing web-servers for predicting protein attributes. Natural science 2, 63–92(openly accessible at http://www.scirp.org/journal/NS/).

Chou, K.C., Shen, H.B., 2010a. A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. PLoS One 5, e9931.

Chou, K.C., Shen, H.B., 2010b. Cell-PLoc2.: a improved package of web servers for predicting subcellular localization of proteins in various organisms. Nat. Sci. 2, 1090–1103.

Chou, K.C., Wu, Z.C., Xiao, X., 2011a. 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., and Xiao, X., 2011b. iLoc-Hum: using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol. Biosyst. 10.1039/C1MB05420a.

Dickerson, R.E., Timkovich, R., Almassy, R.J., 1976. The cytochrome fold and the evolution of bacterial energy metabolism. J. Mol. Biol. 100, 473–491.

Ding, H., Luo, L., Lin, H., 2009. Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. Protein Pept. Lett. 16, 351–355.

Esmaeili, M., Mohabatkar, H., Mohsenzadeh, S., 2010. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. J. Theor. Biol. 263, 203–209.

Fan, G.L., and Li, Q.Z., 2011. Predicting protein submitochondria locations by combining different descriptors into the general form of Chou's pseudo amino acid composition. Amino Acids. 10.1007/s00726-011-1143-4.

Feng, Z.P., 2002. An overview on predicting the subcellular location of a protein. In Silicon Biol. 2, 291–303.

Frieden, T.R., Sterling, T.R., Munsiff, S.S., Watt, C.J., Dye, C., 2003. Tuberculosis. Lancet 362, 887–899.

Gao, Q.B., Wang, Z.Z., Yan, C., Du, Y.H., 2005. Prediction of protein subcellular location using a combined feature of sequence. FEBS Lett. 579, 3444–3448.

Georgiou, D.N., Karakasidis, T.E., Nieto, J.J., Torres, A., 2009. Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. J. Theor. Biol. 257, 17–26.

Gu, Q., Ding, Y.S., Zhang, T.L., 2010. Prediction of G-protein-coupled receptor classes in low homology using Chou's Pseudo amino acid composition with approximate entropy and hydrophobicity patterns. Protein Pept. Lett. 17, 559–567.

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.

Idicula-Thomas, S., Kulkarni, A.J., Kulkarni, B.D., Jayaraman, V.K., Balaji, P.V., 2006. A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in escherichia coli. Bioinformatics 22, 278–284.

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

Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195–202.

Kandaswamy, K.K., Chou, K.C., Martinetz, T., Moller, S., Suganthan, P.N., Sridharan, S., Pugalenthi, G., 2011. AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties. J. Theor. Biol. 270, 56–62.

Kaur, H., Raghava, G.P., 2004. Prediction of alpha-turns in proteins using PSI-BLAST profiles and secondary structure information. Proteins 55, 83–90.

Li, F.M., Li, Q.Z., 2008a. Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. Amino Acids 34, 119–125.

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

Li, Q.Z., Lu, Z.Q., 2001. The prediction of the structural class of protein: application of the measure of diversity. J. Theor. Biol. 213, 493–502.

Li, W., Jaroszewski, L., Godzik, A., 2001. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 17, 282–283.

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, H., Ding, H., 2011. Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. J. Theor. Biol. 269, 64–69.

Lin, H., Ding, H., Guo, F.B., Huang, J., 2010. Prediction of subcellular location of mycobacterial protein using feature selection techniques. Mol. Diversity 14, 667–671.

Lin, H., Ding, H., Guo, F.B., Zhang, A.Y., Huang, J., 2008. Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein Pept. Lett. 15, 739–744.

Liu, T., Zheng, X., Wang, C., Wang, J., 2010. Prediction of subcellular location of apoptosis proteins using pseudo amino acid composition: an approach from auto covariance transformation. Protein Pept. Lett. 17, 1263–1269.

Luginbuhl, P., Szyperski, T., Wuthrich, K., 1995. Statistical basis for the use of ^{13}C a chemical shifts in protein structure determination. J. Magn. Reson. B 109, 229–233.

Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 405, 442–451.

Mielke, S.P., Krishnan, V.V., 2003. Protein structural class identification directly from NMR spectra using averaged chemical shifts. Bioinformatics 19, 2054–2064.

Mohabatkar, H., 2010. Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein Pept. Lett. 17, 1207–1214.

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

Nakai, K., 2000. Protein sorting signals and prediction of subcellular localization. Adv. Protein Chem. 54, 277–344.

Nakashima, H., Nishikawa, K., 1994. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J. Mol. Biol. 238, 54–61.

Pastore, A., Lesk, A.M., 1990. Comparison of the structures of globins and phycocyanins: evidence for evolutionary relationship. Proteins 8, 133–155.

Pollastri, G., McLysaght, A., 2005. Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 21, 1719–1720.

Pollastri, G., Martin, A.J., Mooney, C., Vullo, A., 2007. Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. BMC Bioinf. 8, 201.

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

Rashid, M., Saha, S., Raghava, G.P., 2007. Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. BMC Bioinf. 8, 337.

Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V., Altschul, S.F., 2001. Improving the accuracy of PSI–BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. 29, 2994–3005.

Scharfe, C., Zaccaria, P., Hoertnagel, K., Jaksch, M., Klopstock, T., Dembowski, M., Lill, R., Prokisch, H., Gerbitz, K.D., Neupert, W., Mewes, H.W., Meitinger, T., 2000. MITOP, the mitochondrial proteome database: 2000 update. Nucleic Acids Res. 28, 155–158.

Sibley, A.B., Cosman, M., Krishnan, V.V., 2003. An empirical correlation between secondary structure content and averaged chemical shifts in proteins. Biophys J 84, 1223–1227.

Singh, V., Somvanshi, P., 2010. Toward the virtual screening of potential drugs in the homology modeled NAD+ dependent DNA ligase from *Mycobacterium tuberculosis*. Protein Pept. Lett. 17, 269–276.

Spera, S., Bax, A., 1995. Empirical correlation between protein backbone conformation and C_α and C_β ^{13}C Nuclear Magnetic Resonance chemical shifts. J. Am. Chem. Soc. 113, 5490–5492.

Vapnik, V., 1998. Statistical Learning Theory. Wiley-interscience, New York.

Wishart, D.S., Sykes, B.D., Richards, F.M., 1991. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. J. Mol. Biol. 222, 311–333.

Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Mazumder, R., O'Donovan, C., Redaschi, N., Suzek, B., 2006. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res. 34, D187–D191.

Wu, Z.C., Xiao, X., Chou, K.C., 2011. iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. Mol. Biosyst. 7, 3287–3297.

Wu, Z.C., Xiao, X., Chou, K.C., 2012. iLoc-Gpos: A Multi-Layer Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Gram-Positive Bacterial Proteins. Protein Pept. Lett. 19, 4–14.

Xiao, X., Wu, Z.C., Chou, K.C., 2011a. A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites. PLoS One 6, e20592.

Xiao, X., Wu, Z.C., Chou, K.C., 2011b. iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. J. Theor. Biol. 284, 42–51.

Yu, L., Guo, Y., Li, Y., Li, G., Li, M., Luo, J., Xiong, W., Qin, W., 2010. SecretP: identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition. J. Theor. Biol. 267, 1–6.

Zakeri, P., Moshiri, B., Sadeghi, M., 2011. Prediction of protein submitochondria locations based on data fusion of various features of sequences. J. Theor. Biol. 269, 208–216.

Zeng, Y.H., Guo, Y.Z., Xiao, R.Q., Yang, L., Yu, L.Z., Li, M.L., 2009. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. J. Theor. Biol. 259, 366–372.

Zhang, G.Y., Fang, B.S., 2008. Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo-amino acid composition. J. Theor. Biol. 253, 310–315.

Zhang, G.Y., Li, H.C., Gao, J.Q., Fang, B.S., 2008. Predicting lipase types by improved Chou's pseudo-amino acid composition. Protein Pept. Lett. 15, 1132–1137.

Zhao, Y., Alipanahi, B., Li, S.C., Li, M., 2010. Protein secondary structure prediction using NMR chemical shift data. J. Bioinf. Comput. Biol. 8, 867–884.

Zhou, G.P., Assa-Munt, N., 2001. Some insights into protein structural class prediction. Proteins 44, 57–59.

Zhou, G.P., Doctor, K., 2003. Subcellular location prediction of apoptosis proteins. Proteins 50, 44–48.

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