Mol Divers (2010) 14:667–671
DOI 10.1007/s11030-009-9205-1

FULL-LENGTH PAPER

Prediction of subcellular location of mycobacterial protein using feature selection techniques

Hao Lin · Hui Ding · Feng-Biao Guo · Jian Huang

Received: 20 May 2009 / Accepted: 20 October 2009 / Published online: 12 November 2009

Abstract *Mycobacterium tuberculosis* is the primary pathogen causing tuberculosis, which is one of the most prevalent infectious diseases. The subcellular location of mycobacterial proteins can provide essential clues for proteins function research and drug discovery. Therefore, it is highly desirable to develop a computational method for fast and reliable prediction of subcellular location of mycobacterial proteins. In this study, we developed a support vector machine (SVM) based method to predict subcellular location of mycobacterial proteins. A total of 444 non-redundant mycobacterial proteins were used to train and test proposed model by using jackknife cross validation. By selecting traditional pseudo amino acid composition (PseAAC) as parameters, the overall accuracy of 83.3% was achieved. Moreover, a feature selection technique was developed to find out an optimal amount of PseAAC for improving predictive performance. The optimal amount of PseAAC improved overall accuracy from 83.3 to 87.2%. In addition, the reduced amino acids in N-terminus and non N-terminus of proteins were combined in models for further improving predictive successful rate. As a result, the maximum overall accuracy of 91.2% was achieved with average accuracy of 79.7%. The proposed model provides highly useful information for further experimental research. The prediction model can be accessed free of charge at http://cobi.uestc.edu.cn/cobi/people/hlin/webserver.

Keywords Protein subcellular localization · Pseudo amino acid composition · Feature selection · *Mycobacterium tuberculosis* · Reduced amino acids

H. Lin (✉) · H. Ding · F.-B. Guo · J. Huang
Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, 610054 Chengdu, China
e-mail: hlin@uestc.edu.cn

Introduction

Mycobacterium tuberculosis is the primary cause of tuberculosis (TB) in human. Although TB usually attacks the lungs it can also affect many other organs and systems including the central nervous system. Since TB is easily transmissible between persons, it causes worldwide 8–9 million cases of infection, and 1.5 million deaths every year. *Mycobacterium tuberculosis* appears to be more genetically diverse than previously recognized. This genetic diversity results in significant phenotypic differences between clinical isolates. Although drugs have been used for treatment of tuberculosis, they were ineffective against multidrug-resistant TB [1]. The appearance of vast genomic and proteomic data has provided us with great opportunity to treat this disease. The subcellular location of proteins can provide useful insights about their functions and help in understanding the intricate pathways that regulate biological processes at the cellular level. Therefore, successful prediction of subcellular location of proteins using bioinformatics method is very important for elucidating protein functions involved in various cellular processes [2].

During the past 20 years, many methods have been developed to predict subcellular location of eukaryotic and prokaryotic (Gram-positive and Gram-negative bacteria) proteins using various sequence characteristics [3–9]. However, few cases were performed to predict subcellular location of mycobacterial proteins. Some works have focused on predicting protein synthesis promoters regions in mycobacterial genome using sequence parameters as inputs for machine learning techniques such as: SVM, ANN, LDA, etc. [10–14]. These types of procedures are of general application and have been recently reviewed [15,16]. Recently, Rashid et al. [17] developed a support vector machine-based method for predicting subcellular location of mycobacterial proteins using evolutionary information and motifs. The maximum overall

accuracy of 86.8% was achieved using five-fold cross-validation. However, the database used by Rashid et al. was not non-redundant data. It has been found that a close relationship between sequence identity and predicted accuracy existed in protein subcellular location [18,19]. Using redundant data can surely lead to overestimation of the performance of the methods considered. Our recent research shows the maximum overall accuracy of 82.2% with average accuracy of 68.6% was achieved on 450 non-redundant sequences [20].

In order to improve the predictive accuracy, a new SVM-based model based on feature selection technique [8,21] and reduced amino acids [22–24] was developed to predict subcellular location of mycobacterial proteins. The performance of proposed method was compared with that of the existing methods. Results demonstrate that this model will have wide application both in the study of the functions of mycobacterial proteins and in the design of antimicrobial drugs.

Materials and methods

Data sets

The raw database constructed by Rashid et al. [17] contained 852 mycobacterial proteins. According to their subcellular locations, proteins were classified into four groups: 340 cytoplasmic proteins, 402 integral membrane proteins, 50 secretory proteins, and 60 proteins attached to the membrane by a lipid anchor. The following two steps were used to prepare high quality datasets. (i) The program CD-HIT [25] was used to remove the highly homologues sequences. In order to get a balance between the homologues bias and the size of the training set, the sequence identity cutoff was set to 80%. (ii) In order to utilize information of N-terminal signal peptides, some proteins with no N-terminal signal peptides were removed from non-redundant dataset. After strictly following the above procedures, we finally obtained 444 mycobacterial protein sequences: 147 cytoplasmic proteins, 238 integral membrane proteins, 23 secretory proteins, and 36 proteins attached to the membrane by a lipid anchor.

Support vector machine

Support vector machine (SVM) is a wonderful machine learning method based on statistical learning theory. It has been widely used in the field of bioinformatics. The basic idea of SVM is to transform the samples into a high dimension Hilbert space and to seek a separating hyperplane in this space. For multi-class problems, several strategies such as one-versus-rest (OVR), one-versus-one (OVO) and DAGSVM are applied to extend the traditional SVM. In this paper, OVO strategy is used for multi-class classification. The software used to implement SVM is LibSVM2.83 written by Lin's lab and can be downloaded free of charge from: http://www.csie.ntu.edu.tw/~cjlin/libsvm [26]. 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. After examining these kernel functions with various parameters, we found that RBF achieved the highest predictive accuracy.

PseAAC

The appropriate parameter is one of the most important aspects for prediction algorithms. The physicochemical properties of a protein molecular surface are adapted to the micro environment the protein localized at, and the average physicochemical properties are correlated with the amino acid composition of the sequence. The PseAAC proposed by Chou [27] describes not only the feature of amino acid composition, but also the long distance interaction of physicochemical properties between residues. Thus, we constructed a feature vector based on PseAAC.

For reader's conveniences, the concept of Chou's PseAAC was briefly introduced as follows. A protein sequence with length L amino acid residues $R_1R_2R_3 \ldots R_L$, where R_1 represents the residue at sequence position 1, R_2 represents the residue at position 2, and so forth, may be denoted as a $(20+\lambda)$-dimensional vector defined by $20+\lambda$ discrete numbers; i.e.

$$X = [x_1 \cdots x_{20} x_{20+1} \cdots x_{20+\lambda}]^T \quad (1)$$

where the first 20 numbers in the Eq. 1 represent the classic amino acid composition, and the next λ discrete numbers describe sequence correlation factor, which can be calculated according to the reported paper [27]. While using Chou's PseAAC, two parameters that are weight factor w and correlation factor λ should be optimized. For different problems, the optimal values of w and λ are different. Detailed descriptions about the PseAAC can refer to Chou's paper [27]. Recently, a PseAAC web server [28] has been developed for conveniently calculating various kinds of PseAAC.

Feature selection

Some methods like principal component analysis and genetic algorithm, etc have been developed for feature selection [8]. Other selection procedures are done according to forward and backward selection [21]. In this work, we performed forward selection (addition) with the complete set of PseAAC to find a good small feature set. It initially evaluates each PseAAC and selects the one with the best prediction rate. It then builds all the two-additional feature subsets and finds two PseAACs with highest accuracies. This process continues until increasing the size of the current feature subset leads to a lower

prediction rate. We adopted jackknife cross-validation accuracy of SVM for the selection criteria.

The reduced amino acids

Another feature vector used in this paper was the frequencies of reduced amino acids. Some residues can be clustered into groups according to their physicochemical properties. Because the residues in the same group commonly play similarly structural or functional roles in proteins, the reduced amino acids can provide a method for finding conserved regions of proteins [29]. Recently, Panek et al. [30] have classified 20 amino acids into three classifications according to their individual hydropathies. In any case, Proline (P), Glycine (G) and Cysteine (C) were grouped into a new particular classification [16,31]. Here, according to rules proposed by Chen and Li [22,23], 20 amino acids were classified into six groups: (1) strongly hydrophilic or polar (D, E, H, K, N, Q and R), (2) strongly hydrophobic (A, F, I, L, M and V), (3) weakly hydrophilic/hydrophobic (S, T, Y and W), (4) Proline (P), (5) Glycine (G) and (6) Cysteine (C). Proteins in different subcellular locations have different N-terminal residue compositions and non N-terminal residue compositions. Therefore, for an arbitrary protein sequence, it can be represented as a twelve-dimensional vector.

The criteria definitions

Three test methods that are sub-sampling test, independent dataset test and jackknife cross-validation can be used to evaluate the predictive capability of an algorithm. Among these three methods, the jackknife cross-validation is deemed the most objective and rigorous one [32] that can always yield a unique outcome as demonstrated by a penetrating analysis in a recent comprehensive review [2] and has been widely and increasingly adopted by over 100 papers. For the jackknife cross-validation, each proteins in the dataset is in turn singled out as an independent test sample and all the rule parameters are calculated based on the remaining proteins without including the one being identified. In this paper, the jackknife cross-validation is adopted to evaluate proposed method.

In order to assess the accuracy of prediction methods, four parameters: sensitivity (*Sn*), specificity (*Sp*), overall accuracy (*Ac*) and average accuracy (*Aa*) are used and defined as follows:

$$Sn = TP/(TP + FN) \tag{2}$$

$$Sp = TP/(TP + FP) \tag{3}$$

$$Ac = (TP + TN)/(TP + TN + FP + FN) \tag{4}$$

$$Aa = \sum Sn/\xi \tag{5}$$

here *TP* denotes the numbers of the correctly recognized positives, *FN* denotes the numbers of the positives recognized as negatives, *FP* denotes the numbers of the negatives recognized as positives, *TN* denotes the numbers of correctly recognized negatives, ξ denotes the number of classes.

Results

SVM was initially trained and tested on 444 protein sequences by the use of traditional PseAAC. The weight factor $w$and correlation factor λ of the PseAAC must be determined in advance. Usually, the larger the λ, the more information the representation bears. However, if the PseAAC contains too many components, it would reduce the cluster-tolerant capacity [33] so as to lower down the jackknife success rate. We performed a great number of examinations to optimize w and λ of PseAAC. For the current study, the optimal values of $w = 0.05$ and $\lambda = 8$ were selected as the ones that yielded the maximum overall accuracy. The regularization parameter C and kernel parameter γ of SVM are two key parameters for SVM model selection. Here, we performed the grid-search on C and γ using jackknife cross-validation. As a result, the optimal values of $C = 10$ and $\gamma = 0.05$ were obtained. Table 1 exhibited that the overall accuracy of 83.3% was achieved with average accuracy of 69.8%.

With a view to ascertain whether there was a subset of most informative features among these 28 PseAACs, we examined predictive capability of each PseAAC to identify useful or informative features from a large collection of parameters for improving predictive accuracy. By executing forward selection technique, we found that a small feature set of 18 PseAACs achieved best predictive performance. The overall accuracy increased to 87.2% using the parameter $C = 8$ and $\gamma = 0.03$ (see in Table 1). Our simulations indicated that this method could pick out informative subsets of PseAAC and improve classification results.

Moreover, we investigated the predicted accuracy of reduced amino acids on 444 mycobacterial proteins. The 12 reduced amino acid compositions were selected from N-terminal regions with 30 residues and non N-terminal regions with remained residues of protein sequences. Table 1 showed that the overall accuracy of 79.1% with average accuracy of 61.4% was achieved. By combining 12 reduced amino acid compositions with 18 optimal PseAACs, the overall accuracy improved from 87.2 to 91.2%. The average accuracy increased to 79.7%. The performance of this model was excellent for cytoplasmic, integral membrane and membrane-attached proteins but failed to predict secretory proteins.

Discussion

Recently, Rashid et al. [17] have developed a SVM-based method to predict subcellular location of mycobacterial

Table 1 The predictive results using different parameters

Parameters	PseAAC ($C = 10$, $\gamma = 0.05$)		Reduced PseAAC ($C = 8$, $\gamma = 0.03$)		Reduced amino acids ($C = 10$, $\gamma = 0.05$)		Reduced PseAAC + Reduced amino acids ($C = 10$, $\gamma = 0.05$)	
	Sn (%)	*Sp* (%)	*Sn* (%)	*Sp* (%)	*Sn* (%)	*Sp* (%)	*Sn* (%)	*Sp* (%)
Cytoplasmic	83.0	85.3	91.2	85.4	84.4	77.0	96.6	91.0
Integral membrane	90.3	85.0	89.9	90.7	84.5	80.7	93.7	92.5
Membrane-attached	66.7	66.7	72.2	74.3	63.9	79.3	80.6	93.5
Secretory	39.1	75.0	56.5	81.3	13.0	60.0	47.8	68.8
Ac (%)	83.3		87.2		79.1		91.2	
Aa (%)	69.8		77.5		61.4		79.7	

Table 2 Compared results with other methods

	Success rate (sensitivity) (%)				
	Cytoplasmic	Integral membrane	Secretory	Membrane-attached	Overall accuracy
SVM + PSSM profile [17]	94.7	87.8	44.0	68.3	86.6
Hybrid model (10) [17]	87.0	85.3	92.0	91.7	86.8
SVM + 18 PseAAC + 12 reduced amino acids ($C = 128$, $\gamma = 0.125$)	96.6	94.3	71.1	88.3	93.5

proteins. A maximum overall accuracy of 86.8% was achieved on 852 redundant datasets using five-fold cross-validation. It is important to compare proposed model with Rashid's method using same benchmark dataset. However, there are 18 proteins not contained N-terminal peptides. Therefore, our model was evaluated by only 834 proteins. The predicted results of five-fold cross-validation were recorded in Table 2. As it can be seen from Table 2, the overall accuracy of our method was 93.5% which was higher than that of Rashid methods. Results reveal that the feature selection technique was effective methods for improving predictive performance.

We also checked the performance of proposed method for low identity datasets using jackknife cross-validation. By use of 30% sequence identity as the cutoff, we obtained 330 mycobacterial protein sequences included 116 cytoplasmic proteins, 178 integral membrane proteins, 10 secretory proteins, and 26 proteins attached to the membrane by a lipid anchor. The sensitivities of cytoplasmic proteins, integral membrane proteins, secretory proteins and membrane-attached proteins were 96.4, 93.8, 31.3 and 74.1%, respectively. The overall accuracy of 90% with average accuracy of 73.9% was achieved. The overall accuracy just decreased by 1.2% with the sequence identity decreasing from 80 to 30%. These results demonstrated that the proposed model is robust.

The investigation by Rashid et al. [17] showed that cytoplasmic proteins are too different to have any specific motifs. Membrane proteins maintained certain type of secondary structure, so there may be few motifs in these proteins. Our study exhibited that the correlation of physicochemical properties of amino acids is conservation for cytoplasmic proteins and integral membrane proteins. The signal peptides contain important information for subcellular localization. This conclusion is consistent with other investigations [22,23]. The reason of low accuracy for secretory proteins may be due to lack enough sequences in this location for training the proposed model.

Conclusion

We have developed a SVM-based method for subcellular location prediction of mycobacterial proteins using feature selection techniques. Feature selection techniques can decrease data dimensionality and find out an optimal amount of features, leading to a better performance of predictive model. The good results confirmed that amino acids and physicochemical characteristics contain sufficient information for predicting subcellular location of mycobacterial proteins. This promising method will have broad applications ranging from protein function research to antimicrobial drugs design.

Acknowledgements This study was supported in part by Scientific Research Startup Foundation of UESTC and Scientific Research Foundation of UESTC (JX0769).

References

1. Yeh JI, Mao L (2006) Prediction of membrane proteins in *Mycobacterium tuberculosis* using a support vector machine algorithm. J Comput Biol 13:126–129. doi:10.1089cmb.2006.13.126
2. Chou KC, Shen HB (2007) Review: recent progresses in protein subcellular location prediction. Anal Biochem 370:1–16. doi:10.1016/j.ab.2007.07.006
3. Chou KC, Shen HB (2008) Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nat Protoc 3:153–162. doi:10.1038/nprot.2007.494
4. Shen HB, Chou KC (2007) Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem Biophys Res Commun 355:1006–1011. doi:10.1016/j.bbrc.2007.02.071
5. Shen HB, Chou KC (2007) Gpos-Ploc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. Protein Eng Des Sel 20:39–46. doi:10.1093/protein/gzl053
6. Shen HB, Chou KC (2007) Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers 85:233–240. doi:10.1002/bip.20640
7. Shen HB, Yang J, Chou KC (2007) Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. Amino Acids 33:57–61. doi:10.1007/s00726-006-0478-8
8. Wang T, Yang J (2009) Using the nonlinear dimensionality reduction method for the prediction of subcellular localization of Gram-negative bacterial proteins. Mol Divers. doi:10.1007/s11030-009-9134-z
9. Niu B, Jian YH, Feng KY, Lu WC, Cai YD, Li GZ (2008) Using AdaBoost for the prediction of subcellular location of prokaryotic and eukaryotic proteins. Mol Divers 12:41–45. doi:10.1007/s11030-008-9073-0
10. Kalate RN, Tambe SS, Kulkarni BD (2003) Artificial neural networks for prediction of mycobacterial promoter sequences. Comput Biol Chem 27:555–564. doi:10.1016/j.compbiolchem.2003.09.004
11. González-Díaz H, Pérez-Bello A, Uriarte E, González-Díaz Y (2006) QSAR study for mycobacterial promoters with low sequence homology. Bioorg Med Chem Lett 16:547–553. doi:10.1016/j.bmcl.2005.10.057
12. González-Díaz H, Pérez-Bello A, Uriarte E (2005) Stochastic molecular descriptors for polymers. 3. Markov electrostatic moments as polymer 2D-folding descriptors: RNA-QSAR for mycobacterial promoters. Polymer 46:6461–6473. doi:10.1016/j.polymer.2005.04.104
13. González-Díaz H, Pérez-Bello A, Cruz-Monteagudo M, González-Díaz Y, Santana L, Uriarte E (2007) Chemometrics for QSAR with low sequence homology: mycobacterial promoter sequences recognition with 2D-RNA entropies. Chemom Intell Lab Syst 85:20–26. doi:10.1016/j.chemolab.2006.03.005
14. Perez-Bello A, Munteanu CR, Ubeira FM, De Magalhães AL, Uriarte E, González-Díaz H (2009) Alignment-free prediction of mycobacterial DNA promoters based on pseudo-folding lattice network or star-graph topological indices. J Theor Biol 256:458–466. doi:10.1016/j.jtbi.2008.09.035
15. González-Díaz H, Prado-Prado F, Ubeira FM (2008) Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach. Curr Top Med Chem 8:1676–1690. doi:10.2174/156802608786786543
16. González-Díaz H, González-Díaz Y, Santana L, Ubeira FM, Uriarte E (2008) Proteomics, networks and connectivity indices. Proteomics 8:750–778. doi:10.1002/pmic.200700638
17. Rashid M, Saha S, Raghava GPS (2007) Support vector machine-based method for predicting subcellular localization of mycobacterial proteins using evolutional information and motifs. BMC Bioinformatics 8:337. doi:10.1186/1471-2105-8-337
18. Nair R, Rost B (2002) Sequence conserved for subcellular localization. Protein Sci 11:2836–2847. doi:10.1110/ps.0207402
19. Yu CS, Chen YC, Lu CH, Hwang JK (2006) Prediction of protein subcellular localization. Proteins 64:643–651. doi:10.1002/prot.21018
20. Lin H, Ding H, Guo FB, Zhang AY, Huang J (2008) Predicting subcellular localization of Mycobacterial proteins by using Chou's pseudo amino acid composition. Protein Pept Lett 15:739–744. doi:10.2174/092986608785133681
21. Park KJ, Gromiha MM, Horton P, Suwa M (2005) Discrimination of outer membrane proteins using support vector machines. Bioinformatics 21:4223–4229. doi:10.1093/bioinformatics/bti697
22. Chen YL, Li QZ (2007) Prediction of the subcellular location of apoptosis proteins. J Theor Biol 245:775–783. doi:10.1016/j.jtbi.2006.11.010
23. Chen YL, Li QZ (2007) Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo-amino acid composition. J Theor Biol 248:377–381. doi:10.1016/j.jtbi.2007.05.019
24. Emanuelsson O, Nielsen H, Brunak S, Heijine G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300:1005–1016. doi:10.1006/jmbi.2000.3903
25. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659. doi:10.1093/bioinformatics/btl158
26. Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/_cjlin/libsvm
27. Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins 43:246–255. doi:10.1002/prot.1035
28. Shen HB, Chou KC (2008) PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Anal Biochem 373:386–388. doi:10.1016/j.ab.2007.10.012
29. Russell RB, Saqi MA, Sayle RA, Bates PA, Sternberg MJ (1997) Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. J Mol Biol 269:423–439. doi:10.1006/jmbi.1997.1019
30. Pánek J, Eidhammer I, Aasland R (2005) A new method for identification of protein (Sub)families in a set of proteins based on hydropathy distribution in proteins. Proteins 58:923–934. doi:10.1002/prot.20356
31. Agüero-Chapin G, González-Díaz H, Molina R, Varona-Santos J, Uriarte E, González-Díaz Y (2006) Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from *Psidium guajava* L. FEBS Lett 580:723–730. doi:10.1016/j.febslet.2005.12.072
32. Chou KC, Zhang CT (1995) Review: prediction of protein structural classes. Crit Rev Biochem Mol Biol 30:275–349. doi:10.3109/10409239509083488
33. Chou KC (1999) A key driving force in determination of protein structural classes. Biochem Biophys Res Commun 264:216–224. doi:10.1006/bbrc.1999.1325