Journal of Molecular Graphics and Modelling 40 (2013) 125–130

Contents lists available at SciVerse ScienceDirect

Journal of Molecular Graphics and Modelling

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

The prediction of palmitoylation site locations using a multiple feature extraction method

Shao-Ping Shi[a,c], Xing-Yu Sun[a], Jian-Ding Qiu[a,b,*], Sheng-Bao Suo[a], Xiang Chen[a], Shu-Yun Huang[a], Ru-Ping Liang[a]

[a] Department of Chemistry, Nanchang University, Nanchang 330031, PR China
[b] Department of Materials and Chemical Engineering, Pingxiang College, Pingxiang 337055, PR China
[c] Department of Mathematics, Nanchang University, Nanchang 330031, PR China

ARTICLE INFO

Article history:
Accepted 20 December 2012
Available online 19 January 2013

Keywords:
Palmitoylation
Weight amino acid composition
Auto-correlation functions
Position specific scoring matrix
Support vector machine

ABSTRACT

As an extremely important and ubiquitous post-translational lipid modification, palmitoylation plays a significant role in a variety of biological and physiological processes. Unlike other lipid modifications, protein palmitoylation and depalmitoylation are highly dynamic and can regulate both protein function and localization. The dynamic nature of palmitoylation is poorly understood because of the limitations in current assay methods. The in vivo or in vitro experimental identification of palmitoylation sites is both time consuming and expensive. Due to the large volume of protein sequences generated in the post-genomic era, it is extraordinarily important in both basic research and drug discovery to rapidly identify the attributes of a new protein's palmitoylation sites. In this work, a new computational method, WAP-Palm, combining multiple feature extraction, has been developed to predict the palmitoylation sites of proteins. The performance of the WAP-Palm model is measured herein and was found to have a sensitivity of 81.53%, a specificity of 90.45%, an accuracy of 85.99% and a Matthews correlation coefficient of 72.26% in 10-fold cross-validation test. The results obtained from both the cross-validation and independent tests suggest that the WAP-Palm model might facilitate the identification and annotation of protein palmitoylation locations. The online service is available at http://bioinfo.ncu.edu.cn/WAP-Palm.aspx.

1. Introduction

Protein palmitoylation, also known as S-acylation, is one of the most ubiquitous post-translational modifications (PTM). It reversibly attaches a 16-carbon saturated fatty acid as a lipid palmitate (C16:0) to cysteine residues in protein substrates through a thioester linkage [1–6]. Biochemically, palmitoylation increases the hydrophobicity of proteins to promote protein-membrane association [1–6]. Numerous proteins are modified by palmitoylation to control protein–protein interaction [7–9], intracellular trafficking [10,11], lipid raft targeting [12,13], protein activity [8,14] and other behaviors. Palmitoylation has also been implicated in a variety of biological and physiological processes, including signal transduction [14,15], mitosis [16], neuronal development [3,6] and apoptosis [17], among others. Although protein palmitoylation has attracted extensive attention, it is still difficult to identify protein palmitoylation sites from a given protein chain.

Several conventional experimental techniques (such as mass spectrometry) and protein palmitoylation mechanisms have been previously employed to identify palmitoylation sites [2,4,18,19]. Most of the known palmitoylation sites are mapped through the mutagenesis of candidate cysteine residues with conventional biochemical methods. Although the protein palmitoylation sites can be determined through conventional experiments, the issue of palmitoylation substrate specificity is still unclear; most previous studies have shown that there is no common and canonical consensus on the palmitoylation motif [1,3–5]. Moreover, the number of protein sequences entering into databanks has increased explosively in the post-genomic era. A fast, automated and effective computational method that can predict protein palmitoylation sites is therefore highly desirable.

Currently, computational studies of post-translational modifications are attracting considerable attention. In contrast with time-consuming and expensive experimental methods, certain accurate and convenient computational approaches have been shown to rapidly generate helpful information for further experimental verification. Only a few computational methods have been developed that predict protein palmitoylation sites. Zhou et al. [20] first employed a clustering and scoring strategy (CSS-Palm 1.0) to build a model for predicting palmitoylation sites in early 2006. In

* Corresponding author at: Department of Chemistry, Nanchang University, Nanchang 330031, PR China. Tel.: +86 791 83969518.
E-mail address: jdqiu@ncu.edu.cn (J.-D. Qiu).

1093-3263/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmgm.2012.12.006

the same group, Xue et al. [21] applied a Naive Bayes method (NBA-Palm) to predict palmitoylation sites in late 2006. Shortly after that, Ren et al. [22] upgraded the previous CSS-Palm 1.0 to CSS-Palm 2.0 by employing an updated CSS algorithm that provided a significant performance increase. Later, Wang et al. [23] proposed a predictor algorithm, called CKSAAP-Palm, to identify the potential palmitoylation sites by using the composition of k-spaced amino acid pairs as the encoding scheme. Recently, Hu et al. [24] proposed another predictor algorithm, named IFS-Palm, for predicting the palmitoylation sites based on the features of the amino acid sequences. Among these methods, the highest Matthews correlation coefficient (*MCC*) achieved by training test was approximately 65% [23], and the highest sensitivity was only 68.60% [24]. Improving the quality of protein palmitoylation site identification is therefore a crucial issue.

In this study a novel approach, called WAP-Palm, was developed to identify palmitoylation sites based on the use of multiple feature descriptors for extracting the most informative amino acids features. The multiple feature descriptors used include the weight amino acid composition (WAAC), the auto-correlation functions (ACF), the average accessible surface area (AASA) and the position specific scoring matrix profiles (PSSM). WAAC was utilized to extract information on the amino acid sequence surrounding palmitoylation sites. ACF was applied to encode the physicochemical properties and the correlation of amino acid residues surrounding palmitoylation sites. PSSM and AASA were used to represent evolutionary information and the structural characteristics surrounding palmitoylation sites, respectively. In summary, the amino acids' composition information, sequence position information, physicochemical properties and evolutionary information were combined as inputs to multiple different classifiers for classification. Four types of features and feature analysis were considered. The influences of the feature extraction and window sizes, the ratio between the positive and negative samples and the classification algorithms on the results were all discussed. The system flow chart of the proposed method is shown in Supplementary Fig. S1, which is comprised of dataset construction and preprocessing, feature extraction, model learning and evaluation sections. The details of each process are discussed below.

2. Materials and methods

2.1. Dataset construction and preprocessing

To develop a powerful statistical predictor, the first important thing is to construct a high quality benchmark dataset [25]. The construction of our benchmark datasets was governed by the following criteria: (i) One hundred and eighty-six proteins covering experimental cysteine palmitoylation were obtained by searching for the keywords "palmitoyl cysteine" in UniProtKB/Swiss-Prot (version 2012_10, www.expasy.org) These proteins were not annotated as "by similarity", "potential" or "probable". (ii) This dataset may contain several high sequence identity protein sequences. To avoid an overestimation of the predictive performance, we clustered the protein sequences with a threshold of 40% identity by CD-HIT [26] to remove the highly homologous sequences. As a result, we obtained 137 palmitoylated proteins to constitute the benchmark dataset. (iii) For each site in a protein sequence, a sequence fragment containing $2n+1$ ($4 \le n \le 10$) amino acids was constructed by selecting n residues upstream and n residues downstream from the site. We defined palmitoylated cysteine (C) modification sites already verified through experiments as positive data, while those that were not palmitoylated were defined as negative data. (iv) Fourteen proteins were taken from the benchmark dataset at random to make up an independent test dataset which contained 31 experimental palmitoylation sites and 103 non-palmitoylation sites. The 123 remaining proteins in the benchmark dataset comprised the training dataset which contained 157 experimental palmitoylation sites and 1143 non-palmitoylation sites. Datasets can be downloaded from http://bioinfo.ncu.edu.cn/WAP-Palm.aspx. Moreover, to ensure unbiased and objective results, five negative training sets were obtained by randomly extracting from the negative training datasets. The average predictive performance obtained using the five sets of training data was calculated by the following cross-validation.

2.2. Feature extraction

After the cysteine palmitoylation training set was constructed, the issue of how to represent the protein samples had to be addressed. In this study, the average accessible surface area (AASA), weight amino acid composition (WAAC), auto-correlation functions (ACF) and position specific scoring matrix profiles (PSSM) were applied as representations.

Solvent accessibility is a key property of amino acid residues and is important for both the structure and function of proteins. Previous studies have indicated that the accessible surface area of an amino acid can promote the detection of PTM sites [27,28]. Therefore, the solvent accessibility of amino acid residues surrounding the palmitoylation sites may be adapted to evaluate the classifying performance when distinguishes between the palmitoylation sites and non-palmitoylation sites. Here, the AASA values of the amino acid residues [29] were used to represent the structural characteristics surrounding palmitoylation sites.

To avoid losing the sequence order information, the WAAC was used to extract the sequence position information of the amino acid residues. Given an amino acid type a_i ($i = 1, 2, \ldots, 20$), we can express the position information of amino acid a_i in the protein sequence fragment p with $2n+1$ amino acids by following formula:

$$C_i = \frac{1}{n(n+1)} \sum_{j=-n}^{n} x_{i,j} \left(j + \frac{|j|}{n} \right) \quad (1)$$

where n denotes the number of upstream residues or downstream residues from the central site in the protein sequence fragment p, $x_{i,j} = 1$ if a_i is the jth position residue in protein sequence fragment p, otherwise $x_{i,j} = 0$. In general, the closer residue a_i is to the central site (0 position), the smaller the absolute value of C_i is. Finally, a protein sequence fragment p is defined as 20 dimension feature vectors.

The specificity and diversity of the protein's structure and function are largely attributable to the various properties of each of the 20 amino acids. Here, an auto-correlation function was used to represent the physicochemical properties of the amino acids and the correlation of amino acid residues surrounding palmitoylation sites. Because palmitoylation enhances the surface hydrophobicity of protein substrates [15], the Kyte-Doolittle hydrophobicity value [30] was used as the index. To calculate the auto-correlation functions [31], each residue in the protein sequence fragment was replaced by its hydrophobicity index. Consequently, the protein sequence fragment transforms into a numerical sequence:

$$S_h = h_1 h_2 h_3 \ldots h_i \ldots h_L \quad (2)$$

where h_i is the hydrophobicity index for the ith residue ($i = 1, 2, 3, \ldots, L$) and L is the number of residues in the protein sequence fragment. Thus, the ACF is defined as r_n:

$$r_n = \frac{\sum_{i=1}^{L-n} h_i h_{i+n}}{L-n}, \quad n = 1, 2, 3, \ldots, m \quad (3)$$

where m is an integer ($m < L$), and r_m is the correlation of two amino acid residues separated by $m-1$ other amino acids. The sequence feature S_h is defined as an m dimension feature vector:

$$R = [r_1, r_2, \ldots, r_m] \quad (4)$$

Here, preliminary tests indicated that the appropriate m was 8, as given in Supplementary Table S1.

In biological analysis, one of the most important aspects of concern is the evolutionary conservation. A more conserved residue within a protein sequence may indicate that it is more important for the protein function and hence under stronger selective pressure. We used Position Specific Iterative BLAST (PSI-BLAST) [32] to measure the conservation status for a specific residue. The PSSM conservation score were obtained by PSI-BLAST against whole Swiss-Prot protein database. In this work, the PSSM profiles were used to quantify the conservation status against 20 different amino acids of each residue in the protein sequences.

Finally, the four types of features (AASA, WAAC, ACF, PSSM) served as input for classifier system for classification.

2.3. Model learning

In this paper, we focus on the K nearest neighbor (KNN) [33] with Euclidean distance, decision tree (DT) [34,35] and support vector machine (SVM) [36] algorithms. For the detailed setup procedures of these learning algorithms, please refer to the literature [33–36]. The KNN and DT algorithms were implemented in the MATLAB programming environment. The number of nearest neighbor in the KNN method was 5, and the default parameter of the DT classifiers was used. For the SVM method, we used the LIBSVM package (version 2.81) [37], which can be freely downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm. To obtain an SVM classifier with optimal performance, a radial basis function (RBF) was tested, and the penalty parameter C and kernel parameter γ were tuned based on the training set using the grid search strategy in LIBSVM.

2.4. Model evaluation

The performance of the classifiers was evaluated through the following equations [38]:

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (5)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (6)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (7)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \quad (8)$$

where TP, TN, FP and FN denote the number of true positives, true negatives, false positives and false negatives, respectively. The sensitivity (Sn) and specificity (Sp) illustrate the correct prediction ratios of positive (palmitoylation) samples and negative (non-palmitoylation) samples, respectively. MCC is the Matthews correlation coefficient [38], which reflects both the Sn and Sp of the prediction algorithm. Accuracy (Acc) is defined as the accuracy of the protein classification by the classifier model into either positive or negative data classes. The shortcoming of this overall accuracy is that an imbalance in the data classes may result in a number of relevant problems: improper classification evaluation metrics, absolute or relative lack of data, data fragmentation, improper inductive bias and noise [39], in addition to a high overall accuracy, even if either Sn or Sp is low. Thus MCC, which is a weighted measurement, is increasingly being used to measure the predictive capability of classifier models. A MCC value of 1 indicates that the classifier model can predict the data classes of unknown compounds perfectly, a MCC value of 0 is expected for a classifier model that is no better than random guessing, and a MCC value of -1 indicates total disagreement between the predicted data classes and the actual data classes.

3. Results and discussion

3.1. Analysis of different features

As mentioned previously, the multiple feature descriptors used included four types of features: AASA, WAAC, ACF, and PSSM. Here we analyzed the distinction of AASA, WAAC, ACF, and PSSM features between palmitoylation and non-palmitoylation. Fig. 1A summarizes the AASA formed from the 19-mer palmitoylation sites and the 19-mer non-palmitoylation sites in the constructed data set. The average AASA of neighborhood residues are 37.61–51.95 for palmitoylation sites. The fluctuant range of AASA of residues surrounding palmitoylation sites is bigger than that of non-palmitoylation sites. This implies that the palmitoylation processing might have occurred where the structural surroundings are relatively large variation range. The AASA of downstream residues that surrounds palmitoylation sites exceed that around non-palmitoylation sites. Generally speaking, the AASA of residues around the palmitoylation sites and non-palmitoylation sites have a little difference.

WAAC feature reflects the position information of residues surrounding palmitoylation sites and non-palmitoylation sites. To analyze position specific properties, we adopted a web-based tool Two Sample Logo [40] to present the compositional biases between 157 palmitoylation sites and 1143 non-palmitoylation sites. As we can see from Fig. 1B, no amino acids surrounding palmitoylation sites are obviously conserved, but there are some significant location-specific differences between palmitoylation sites and non-palmitoylation sites. The most pronounced feature of palmitoylation sites is the abundance of hydrophobic amino acids, such as leucine (L), glycine (G) and alanine (A) from −9 to +1 position. As reported, protein palmitoylation often occurs in close proximity of hydrophobic amino acid stretch [5,24]. Meanwhile, positively charged amino acids arginine (R) and lysine (K) are enriched at +2, +4, +5 and +8 positions for cysteine palmitoylation.

To analyze residue correlation and dependence, we calculated the average value of the ACF feature surrounding palmitoylation sites, as shown in Fig. 1C. The average ACF of neighborhood residues are 1.59–8.03 for palmitoylation sites, and 0.24–2.99 for non-palmitoylation sites. The values of the ACF feature around the palmitoylation sites are far higher than those around the non-palmitoylation sites, indicating that the residues within the palmitoylation sites have better correlation and more dependence than those within the non-palmitoylation sites. This analysis reveals that the residue interactions in the flanking sequences of the cysteine site are significantly different between the palmitoylation sites and the non-palmitoylation sites. Therefore, the ACF is suitable to be used as features for palmitoylation site prediction.

Evolutionary information is important in most of these protein structure and function prediction methods. Fig. 1D gives the mean values of PSSM scores of residues around palmitoylation sites and non-palmitoylation sites based on training data. From −9 to +4 positions, the average PSSM scores of residues surrounding palmitoylation sites are lower than those of non-palmitoylation sites, especially for −3 to +2 position. This reveals that there is significant difference of evolutionary conservation between palmitoylation and non-palmitoylation sequences and the residues directly

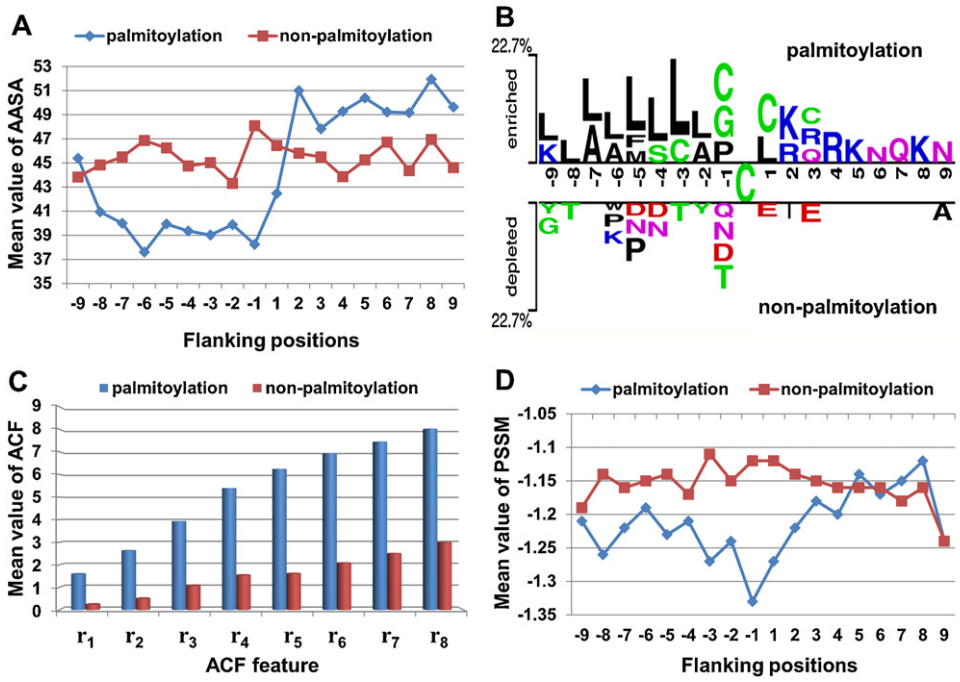

Fig. 1. Analyses of four types of features around 157 palmitoylation sites and 1143 non-palmitoylation sites, where the window size is 19 residues (−9 ~ C ~ +9). (A) Distribution of the mean value of AASA feature. (B) Two Sample Logos [40] of the compositional biases around 157 palmitoylation sites compared to 1143 non-palmitoylation sites. Only amino acid residues significantly enriched or depleted (p-value <0.05; t-test) around palmitoylation sites are shown. (C) Distribution of the mean value of ACF feature. (D) The average PSSM scores of residues around palmitoylation sites and non-palmitoylation sites.

adjacent to cysteine site have much more impact on determination of palmitoylation site.

3.2. Optimal feature set

We constructed nine prediction models composed of AASA, WAAC, ACF, and PSSM to investigate the influences of different features. When the window size was 19 and the ratio between positive and negative samples was 1:1, the predictive performance of models trained with various features using the SVM algorithm was shown in Table 1. The model trained with PSSM outperformed the models trained with AASA, WAAC or ACF. This may draw that evolution information acts an irreplaceable role for the prediction of palmitoylation site. But in general, the models trained using individual features did not effectively predict cysteine palmitoylation. However, the predictive performance of the model trained with combination of the three features or four features showed an improvement of 1.08–10% over the most accurate of the single feature extraction models. The model trained with WAAC + ACF + PSSM obtained the best performance with a *Sn* of 81.53%, a *Sp* of 90.45%, an *Acc* of 85.99% and a *MCC* of 72.26%. The results show that the multiple feature extraction method can incorporate more information on palmitoylation and accurately classify this information.

Table 1
The performance of models trained with various features using the SVM algorithm.

Training features	*Sn* (%)	*Sp* (%)	*Acc* (%)	*MCC* (%)
AASA	71.21	70.19	70.70	41.43
WAAC	73.25	72.61	72.93	45.86
ACF	67.52	79.62	73.57	47.48
PSSM	74.27	77.71	75.99	52.02
AASA + ACF + WAAC	75.16	78.98	77.07	54.18
AASA + ACF + PSSM	75.80	82.80	79.30	58.74
AASA + WAAC + PSSM	75.16	81.53	78.34	56.80
WAAC + ACF + PSSM	81.53	90.45	85.99	72.26
AASA + ACF + WAAC + PSSM	79.62	88.54	84.08	68.43

Henceforth, the combination of WAAC + ACF + PSSM was selected as an optimal feature set to learn the predictive model.

3.3. Window size

For each palmitoylation or non-palmitoylation sites, its profile features were taken from a sequence fragment containing the n nearest residues (spatially); thus, it is crucial to confirm the appropriate window size and to realize its effects on the prediction performance. The predictive performance of models trained with different window sizes (9–21) are illustrated in Fig. 2 and Supplementary Table S2, where training feature was WAAC + ACF + PSSM and the ratio between positive and negative samples was 1:1. With the increase of window size, the predictive performance was increased. When the window size increased to 19, the prediction performance reached their peak. However, with the window size continues to increase, the performance showed

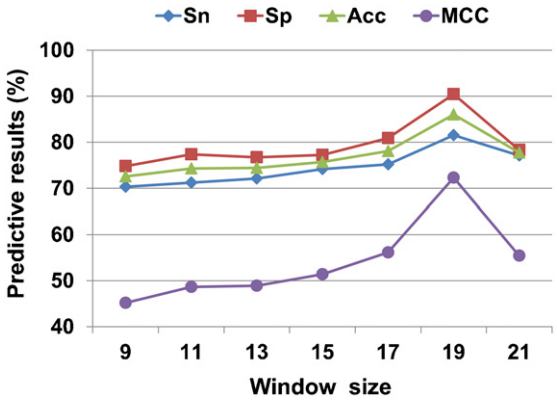

Fig. 2. The 10-fold cross validation performance of models trained with different window size. The ratio between positive and negative samples is 1:1 and training feature is WAAC + ACF + PSSM.

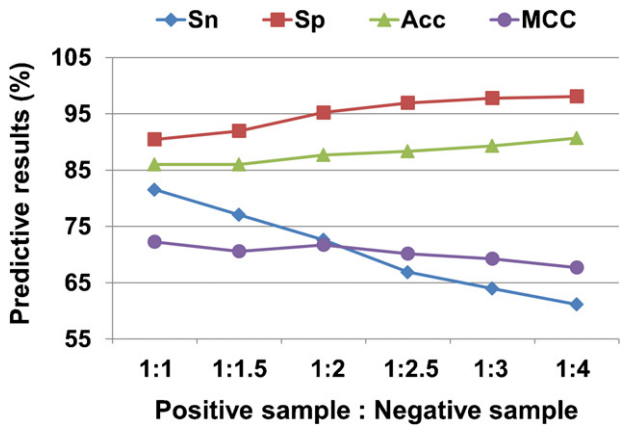

Fig. 3. The performance of models trained with different positive to negative sample ratios. The window size is 19 and training feature is WAAC + ACF + PSSM.

a decrease tendency. Based on the computational efficiency and overall performance of the trained models, 19-mer was adopted as the feasible window size in this study.

3.4. The impact of the ratio between positive and negative samples

Based on the optimal feature (WAAC + ACF + PSSM) and the window size 19, the performance of models trained with different positive to negative sample ratios is shown in Fig. 3 and Supplementary Table S3. With the increase of relative size of the negative set, the *Sp* of the predictive models increases, instead the *Sn* keeps decreasing. This is because the number of negative samples was larger than that of positive samples during the 10-fold cross-validation, and a larger negative set will cause the trained model to preferentially predict negative data correctly to maximize accuracy. To avoid these tendencies in the negative set extraction and reduce the false positive rate, a balanced dataset should be randomly collected. Hence, 1:1 was as the suitable ratio between positive samples and negative samples to construct the optimal predictive model WAP-Palm.

3.5. Comparison of the different classifiers

To further optimize the classification algorithms, the performance of the following three classifiers were compared: *K* nearest neighbor (KNN), decision trees (DT) and SVM. These three classifiers are state-of-the-art methods within the field of protein prediction. SVM yielded the best performance with respect to *Sn*, *Acc* and *MCC* (Supplementary Table S4), which indicated that SVM automatically obtains a better trade-off between specificity and sensitivity. Moreover, the under receiver operating characteristic curves (ROC) [41] used for the assessment of the performance of the three classifiers, were plotted in Fig. 4, with different curves denoting the prediction performances of DT, SVM and KNN. Larger values indicate better

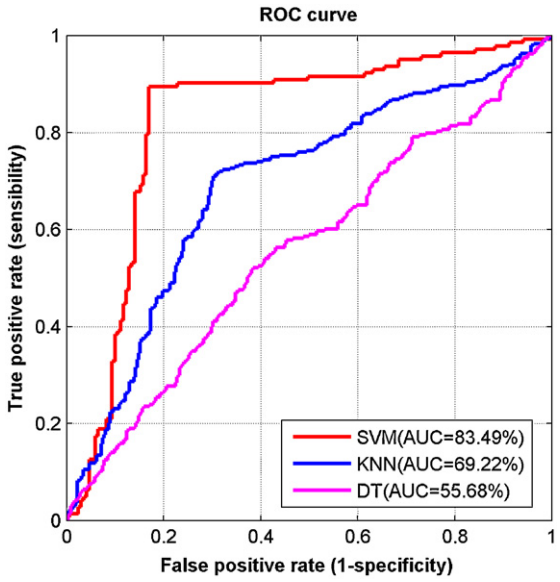

Fig. 4. The ROC curve of different classifiers.

overall performances of the models. Comparing the results from Fig. 4, we found that the SVM classifier was superior to the other classifiers. Therefore, SVM was selected as the appropriate classifier to construct the optimal model for predicting palmitoylation sites in this study.

3.6. Comparisons with existing methods

In order to further evaluate the prediction performance of the WAP-Palm method objectively, we made comparisons with other palmitoylation predictor. Since the training data are not identical among the methods, comparison of cross-validation performance might be unreasonable. The computational predictors CSS-Palm 1.0 [20], NBA-Palm [21] and CSS-Palm 2.0 [22] were unavailable on the net provided by these papers. IFS-Palm [24] did not support web service. Therefore, the comparison was carried out between CKSAAP-Palm [23] and WAP-Palm on the same training and independent test datasets used in this study. CKSAAP-Palm was retrained on the training dataset, and then the prediction was performed on the 14 proteins in independent test dataset. The predictive performance of CKSAAP-Palm and WAP-Palm is shown in Table 2 and Supplementary Table S5. Based on threshold (SVM probability) 0.6, the *Sn* in WAP-Palm reached 80.65%, which was about 41.94% higher than that in CKSAAP-Palm. Based on threshold 0.8, the *Sn* and *MCC* in WAP-Palm were increased by 29.03% and 39.17% in comparison with the results in CKSAAP-Palm. It is worth noting that CKSAAP-Palm has high *Sp* and low *Sn*, whereas WAP-Palm can obtain better trade-off between *Sp* and *Sn* automatically. In summary, the WAP-Palm outperformed CKSAAP-Palm. The results reveal that our improvements can be attributed to the

Table 2
Comparison of CKSAAP-Palm [23] and WAP-Palm on the independent test set which includes 31 palmitoylation sites.

Method	Training features	Threshold	*Sn* (%)	*Sp* (%)	*Acc* (%)	*MCC* (%)
CKSAAP-Palm	*k*-Spaced amino acid pairs	$C = 100$, $\gamma = 0.0000015$, cutoff = 0.18	38.71	84.47	73.88	24.04
WAP-Palm	WAAC + ACF + PSSM	0.6 0.8	80.65 67.74	79.61 93.20	79.85 87.31	53.52 63.21

Where C and γ are the penalty parameter and kernel parameter of SVM, respectively.

adoption of evolutionary information and auto-correlation function, as elucidated in the above feature analysis, the combination of WAAC + ACF + PSSM is effective for palmitoylated protein prediction.

4. Conclusions

Palmitoylation prediction methods used in previous studies, such as CCS-Palm 1.0, NBA-Palm, CSS-Palm 2.0 and CKSAAP-Palm have focused only on protein sequence characteristics. However, the multiple feature descriptors in the WAP-Palm model described in our work incorporate amino acid composition, sequence position information, physicochemical properties, and evolutionary information to improve the prediction of protein palmitoylation sites. The cross-validation results demonstrated that WAP-Palm model performs promisingly. Additionally, the WAP-Palm model performs accurately and robustly in an independent test. In conclusion, our model performs well not only at detecting palmitoylation sites but also in helping biologists identify the potential protein cysteine palmitoylation sites.

Acknowledgments

This work was supported by the Program for New Century Excellent Talents in University (NCET-11-1002) and the National Natural Science Foundation of China (21175064 and 20605010).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jmgm.2012.12.006.

References

[1] M.J. Bijlmakers, M. Marsh, The on-off story of protein palmitoylation, Trends in Cell Biology 13 (2003) 32–42.
[2] L.E. Dietrich, C. Ungermann, On the mechanism of protein palmitoylation, EMBO Reports 5 (2004) 1053–1057.
[3] A.D. El-Husseini, D.S. Bredt, Protein palmitoylation: a regulator of neuronal development and function, Nature Reviews Neuroscience 3 (2002) 791–802.
[4] M.E. Linder, R.J. Deschenes, New insights into the mechanisms of protein palmitoylation, Biochemistry 42 (2003) 4311–4320.
[5] J.E. Smotrys, M.E. Linder, Palmitoylation of intracellular signaling proteins: regulation and function, Annual Review of Biochemistry 73 (2004) 559–587.
[6] K. Huang, A. El-Husseini, Modulation of neuronal protein trafficking and function by palmitoylation, Current Opinion in Neurobiology 15 (2005) 527–535.
[7] X. Yang, O.V. Kovalenko, W. Tang, C. Claas, C.S. Stipp, M.E. Hemler, Palmitoylation supports assembly and function of integrintetraspanin complexes, Journal of Cell Biology 167 (2004) 1231–1240.
[8] B. Zhou, L. Liu, M. Reddivari, X.A. Zhang, The palmitoylation of metastasis suppressor KAI1/CD82 is important for its motility and invasiveness-inhibitory activity, Cancer Research 64 (2004) 7455–7463.
[9] K.L. Clark, A. Oelke, M.E. Johnson, K.D. Eilert, P.C. Simpson, S.C. Todd, CD81 associates with 14-3-3 in a redox-regulated palmitoylation dependent manner, Journal of Biological Chemistry 279 (2004) 19401–19406.
[10] E.V. Kalinina, L.D. Fricker, Palmitoylation of carboxypeptidase D. Implications for intracellular trafficking, Journal of Biological Chemistry 278 (2003) 9244–9249.
[11] I. Navarro-Lerida, M.M. Corvi, A.A. Barrientos, F. Gavilanes, L.G. Berthiaume, I. Rodriguez-Crespo, Palmitoylation of inducible nitricoxide synthase at Cys-3 is required for proper intracellular traffic and nitric oxide synthesis, Journal of Biological Chemistry 279 (2004) 55682–55689.
[12] C. Salaun, G.W. Gould, L.H. Chamberlain, The SNARE proteins SNAP-25 and SNAP-23 display different affinities for lipid rafts in PC12 cells, regulation by distinct cysteine-rich domains, Journal of Biological Chemistry 280 (2005) 1236–1240.
[13] W. Wong, L.C. Schlichter, Differential recruitment of Kv1.4 and Kv4.2 to lipid rafts by PSD-95, Journal of Biological Chemistry 279 (2004) 444–452.
[14] P. Vazquez, I. Roncero, E. Blazquez, E. Alvarez, Substitution of the cysteine 438 residue in the cytoplasmic tail of the glucagonlike peptide-1 receptor alters signal transduction activity, Journal of Endocrinology 185 (2005) 35–44.
[15] C. Kleuss, E. Krause, Galpha(s) is palmitoylated at the N-terminal glycine, EMBO Journal 22 (2003) 826–832.
[16] J.M. Caron, L.R. Vega, J. Fleming, R. Bishop, F. Solomon, Single site alpha-tubulin mutation affects astral microtubules and nuclear positioning during anaphase in *Saccharomyces cerevisiae*: possible role for palmitoylation of alpha-tubulin, Molecular Biology of the Cell 12 (2001) 2672–2687.
[17] D.A. Wang, S.M. Sebti, Palmitoylated cysteine 192 is required for RhoB tumor-suppressive and apoptotic activities, Journal of Biological Chemistry 280 (2005) 19243–19249.
[18] M. Fukata, Y. Fukata, H. Adesnik, R.A. Nicoll, D.S. Bredt, Identification of PSD-95 palmitoylating enzymes, Neuron 44 (2004) 987–996.
[19] K. Huang, A. Yanai, R. Kang, P. Arstikaitis, R.R. Singaraja, M. Metzler, A. Mullard, B. Haigh, C. Gauthier-Campbell, C.A. Gutekunst, M.R. Hayden, A. El-Husseini, Huntingtin-interacting protein HIP14 is a palmitoyl transferase involved in palmitoylation and trafficking of multiple neuronal proteins, Neuron 44 (2004) 977–986.
[20] F.F. Zhou, Y. Xue, X.B. Yao, Y. Xu, CSS-Palm: palmitoylation site prediction with a clustering and scoring strategy (CSS), Bioinformatics 22 (2006) 894–896.
[21] Y. Xue, H. Chen, C.J. Jin, Z.R. Sun, X.B. Yao, NBA-Palm: prediction of palmitoylation site implemented in Naïve Bayes algorithm, BMC Bioinformatics 7 (2006) 458–468.
[22] J. Ren, L. Wen, X. Gao, C. Jin, Y. Xue, X.B. Yao, CSS-Palm 2.0: an updated software for palmitoylation sites prediction, Protein Engineering Design and Selection 21 (2008) 639–644.
[23] X.B. Wang, L.Y. Wu, Y.C. Wang, N.Y. Deng, Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs, Protein Engineering Design and Selection 22 (2009) 707–712.
[24] L.L. Hu, S.B. Wan, S. Niu, X.H. Shi, H.P. Li, Y.D. Cai, K.C. Chou, Prediction and analysis of protein palmitoylation sites, Biochimie 93 (2011) 489–496.
[25] K.C. Chou, H.B. Shen, Review: Recent progresses in protein subcellular location prediction, Analytical Biochemistry 370 (2007) 1–16.
[26] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (2006) 1658–1659.
[27] D.M. Shien, T.Y. Lee, W.C. Chang, J.B.K. Hsu, J.T. Horng, P.C. Hsu, T.Y. Wang, H.D. Huang, Incorporating structural characteristics for identification of protein methylation sites, Journal of Computational Chemistry 30 (2009) 1532–1543.
[28] S. Ahmad, M.M. Gromiha, A. Sarai, Real value prediction of solvent accessibility from amino acid, Proteins 50 (2003) 629–635.
[29] J. Janin, S. Wodak, Conformation of amino acid side-chains in proteins, Journal of Molecular Biology 125 (1978) 357–386.
[30] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, Journal of Molecular Biology 157 (1982) 105–132.
[31] J.L. Cornette, K.B. Cease, H. Margalit, J.L. Spouge, J.A. Berzofsky, C. DeLisi, Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins, Journal of Molecular Biology 195 (1987) 659–685.
[32] S.F. Altschul, T.L. Madden, A.A. Schaffer, J.H. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Research 25 (1997) 3389–3402.
[33] T. Denoeux, A k-nearest neighbor classification rule based on Dempster-Shafer theory, IEEE Transactions on Systems, Man and Cybernetics 25 (1995) 804–813.
[34] J.R. Quinlan, Induction of decision trees, Machine Learning 1 (1986) 81–106.
[35] J.R. Quinlan, C4.5 Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA, 1993.
[36] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273–297.
[37] C.C. Chang, C.J. Lin, LIBSVM: a library for support machines [software], 2001. www.csie.ntu.edu.tw/∼cjlin/libsvm
[38] J. Davis, M. Goadrich, The relationship between Precision-Recall and ROC curves, in: ICML 06: Proceedings of the 23rd International Conference on Machine Learning, New York, NY, USA, 2006, pp. 233–240.
[39] T.M. Mitchell, Machine Learning, McGraw Hill, New York, 1997.
[40] V. Vacic, L.M. Iakoucheva, P. Radivojac, Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments, Bioinformatics 22 (2006) 1536–1537.
[41] A. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition 30 (1997) 1145–1159.