Protein Engineering, Design & Selection vol. 22 no. 11 pp. 707–712, 2009
Published online September 25, 2009 doi:10.1093/protein/gzp055

Prediction of palmitoylation sites using the composition of *k*-spaced amino acid pairs

Xiao-Bo Wang[1], Ling-Yun Wu[2], Yong-Cui Wang[1] and Nai-Yang Deng[1,3]

[1]College of Science, China Agricultural University, Beijing 100083, People's Republic of China and [2]Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, People's Republic of China

[3]To whom correspondence should be addressed.
E-mail: dengnaiyang@vip.163.com

Palmitoylation is an important hydrophobic protein modification activity that participates many cellular processes, including signaling, neuronal transmission, membrane trafficking and so on. So it is an important problem to identify palmitoylated proteins and the corresponding sites. Comparing with the expensive and time-consuming biochemical experiments, the computational methods have attracted much attention due to their good performances in predicting palmitoylation sites. In this paper, we develop a novel automated computational method to perform this work. For a sequence segment in a given protein, the encoding scheme based on the composition of *k*-spaced amino acid pairs (CKSAAP) is introduced, and then the support vector machine is used as the predictor. The proposed prediction model CKSAAP-Palm outperforms the existing method CSS-Palm2.0 on both cross-validation experiments and some independent testing data sets. These results imply that our CKSAAP-Palm is able to predict more potential palmitoylation sites and increases research productivity in palmitoylation sites discovery. The corresponding software can be freely downloaded from http://www.aporc.org/doc/wiki/CKSAAP-Palm.

Keywords: CKSAAP-Palm/palmitoylated proteins/palmitoylation/support vector machine

Introduction

Palmitoylation is a hydrophobic protein modification activity that fatty acids are covalently attached to cysteine residues of membrane proteins. In biochemistry and enzymology study, it has been observed that this modification activity uses cellular and viral membrane proteins for signal transmission (Veit and Schmidt, 2006). However, it is still unknown what the molecular signals for palmitoylation are. Although palmitoylation is known to be a reversible activity with cycles of acylation and deacylation, the relevant enzymatic mechanism has not been completely clear because some palmitoylated proteins are found without any enzyme source present. Nonetheless, palmitoylation activity has been widely studied in various areas including most signaling pathway activities (Kurayoshi *et al.*, 2006; Navarro-Lerida *et al.*, 2006). For instance, Smotrys and Linder (2004) showed that most trafficking and protein–protein interactions as well as enzyme activities depended on the existence of palmitoylated proteins. They also showed that palmitoylated proteins could enhance the membrane interactions and the reversibility of palmitoylation was an attractive mechanism for regulating protein activity and cell signaling.

As a specificity study of post-translational modifications such as phosphorylation, methylation and sumoylation, palmitoylation is very important in system biology researches for understanding how proteins are responding to extra cellular cues for information transmission along signaling pathways. One of the key points in studying palmitoylation is to identify palmitoylated proteins and the corresponding sites. However, experimental identification of modification proteins with their sites is quite difficult, time-consuming and expensive (el-Husseini Ael and Bredt, 2002; Linder and Deschenes, 2007). Therefore, *in silico* prediction of palmitoylation sites is urgent and greatly useful for further experimental verification.

In the context of predicting palmitoylation sites, Zhou *et al.* (2006), by applying a clustering and scoring strategy, built the first model named as CSS-Palm 1.0, in early 2006. Later, in late 2006, Xue *et al*., using the standard binary encoding, proposed a naive Bayes method named as NBA-Palm1.0 (Xue *et al*., 2006). Recently, the CSS-Palm1.0 is updated into the new version CSS-Palm2.0 (Ren *et al*., 2008) with the great improvement in predicting performance. By a training data set with 263 palmitoylation sites in 109 proteins and a testing data set with 56 palmitoylation sites in 29 proteins, CSS-Palm2.0 was tested. CSS-Palm2.0 was also be used to annotate novel palmitoylated proteins in budding yeast. Although they have got good results, there is still a big room for improvement.

In this study, we use the encoding scheme based on the composition of *k*-spaced amino acid pairs (CKSAAP) to represent the sequence fragments. Then support vector machine (SVM; Vapnik, 1995, 1998) is used as the classifier. The sensitivity of our model CKSAAP-Palm can reach 88.42% and 64.16%, respectively, in the 10-fold cross-validation on training data set and the independent testing data set constructed by Xue *et al*. For annotation of novel palmitoylated proteins in budding yeast, the sensitivity of CKSAAP-Palm is 100% which is much higher than that of CSS-Palm2.0.

Data sets and method

Data sets

The data sets used in this paper are divided into two parts: training data and testing data. The training data set comes from Xue and co-workers (Ren *et al*., 2008). They searched the scientific literatures from PubMed with keywords of 'palmitoylation' or 'palmitoylated', and manually collected 284 experimentally verified palmitoylation sites in 116 proteins published before November 2006. After deleting several homologous sites from homologous proteins (Li and Godzik, 2004) which may result in overestimation of the prediction accuracy, they used 263 positive sites in 109 proteins as their positive sites. For the negative sites, the other 1150 sits in

these 109 proteins were taken. We use the same training data set in this paper. Our testing data include three data sets. The first one also comes from Xue and co-workers (Ren *et al*., 2008). It is the 56 sites in 29 proteins published from November 2006 to 8 October 2007. The second one comes from Roth *et al.* (2006), where they identified palmitoylated proteins in budding yeast, including 16 known palmitoylated proteins and 35 novel palmitoylated proteins. In order to evaluate the false-positive rate, 120 proteins known not to be palmitoylated such as mitochondrial proteins are randomly extracted from the Swiss-Prot database (Release 53) as our third testing data set.

Method

The prediction problem is formulated as a binary classification problem, and the SVM algorithm is applied.

Construction of feature vectors To formulate a binary classification problem, the first step is to represent the object as a feature vector in the Euclidian space. We use the encoding scheme based on the CKSAAP, which has been successfully used for predicting mucin type *O*-glycosylation sites in mammalian (Chen *et al*., 2008) and flexible/rigid regions (Chen *et al*., 2007) to represent the sequence fragment. More precisely, for a site in a protein sequence, a sequence fragment with $2n + 1$ amino acids is constructed by taking n upstream residues and n downstream residues from the site, respectively. When the residues are not enough, e.g. for the sites located in N- or C-terminus, we assign a non-existing residue *O* to fill in the corresponding position. Therefore, there are totally 21 types of amino acids and 441 types of amino acid pairs in our setting. Note that these pairs are extended to the *k*-spaced amino acid pair (i.e. the pairs separated by k other amino acids). In our setting, we consider the *k*-spaced amino acid pairs with $k = 0, 1, \ldots, 10$. In fact, in order to represent a sequence fragment, we introduce a 441×11 dimensional feature vector x where each component is the frequency of the corresponding *k*-spaced amino acid pair appearing in this sequence fragment. The feature vector of a simpler example sequence fragment AAAC with $k = 0$, 1 and 2 can be found from Table I.

Support vector machine SVM is an powerful method in dealing with binary classification problem if a penalty factor C and a kernel function $K(\cdot,\cdot)$ are selected properly. Using a Euclidian vector set X_+ with positive labels and a set X_- with negative labels, it constructs a decision function:

$$f(x) = \sum_{i:x\in X_+} \alpha_i K(x, x_i) - \sum_{i:x\in X_-} \alpha_i K(x, x_i) + b,$$

Table I. An example for CKSAAP features of sequence fragment *AAAC*

K	*k*-space amino acid pairs	*k*-space features
0	$(AA,AC,AD,\ldots,OO)_{441}$	$(2, 1, 0, \ldots, 0)_{441}$
1	$(AXA,AXC,AXD,\ldots,OXO)_{441}$	$(1, 1, 0, \ldots, 0)_{441}$
2	$(AXXA,AXXC,AXXD,\ldots,OXXO)_{441}$	$(0, 1, 0, \ldots, 0)_{441}$

The *X* represents any kind of the 21 amino acids.

where the non-negative weights α and b are computed during training by solving a convex quadratic programming. A new vector x is then predicted to be positive or negative depending on whether the value of function $f(x)$ is greater or less than a pre-determined cutoff value. The details about SVM can be found in Vapnik (1995, 1998). Our prediction model, based on SVM with CKSAAP feature vectors, is denoted as CKSAAP-Palm. It corresponds to the following selection: the penalty factor C is selected to be 100, and the kernel function is selected to be Gaussian radial basis function:

$$K(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2),$$

with $\gamma = 0.0000015$. In addition, the cutoff value is selected to be 0.18 based on the receiver operating characteristic (ROC) curve of applying cross-validation on the training data set, which can show the trade-off between sensitivity and specificity, unless an additional introduction is attached.

Results and discussion

Performance evaluation criteria

In order to evaluate our prediction model CKSAAP-Palm, some criteria are necessary. There are four measurements: sensitivity (Sn), specificity (Sp), accuracy (Ac) and Mathew correlation coefficient (MCC). They are defined as follows:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}},$$

$$\mathrm{Ac} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}},$$

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

where TP, TN, FP and FN are the number of true positive, true negative, false positive and false negative, respectively. For a given data set, all these values can be obtained from the decision function with fixed cutoff. In addition, the prediction validity is often examined by observing its ROC curve, which shows the trade-off between sensitivity and specificity and gives a complete evaluation of the computational method. Apart from the ROC curve, the area under curve (AUC) is also an indicator, the larger, the better.

Comparison of CKSAAP-Palm with the existing tools

Now let us begin to compare our CKSAAP-Palm model with the recent works of Xue *et al*., CSS-Palm2.0 (Ren *et al*., 2008) and NBA-Palm1.0 (Xue *et al*., 2006). As mentioned above, the training and testing data sets are exactly the same with theirs. Since the developers of CSS-Palm2.0 and NBA-Palm1.0 only provide a web-based server instead of standalone software, it is impossible to obtain the results of CSS-Palm2.0 and NBA-Palm1.0 with a specified training data set. The results reported in Ren *et al.* (2008) and Xue *et al.* (2006) are used in comparison.

First, consider the training data set mentioned in 'Data sets and method'. For CKSAAP-Palm, the 10-fold cross-validation is performed and the results are listed in Table II

and Fig. 1. Table II shows that the performance of CKSAAP-Palm is significantly better than that of CSS-Palm2.0 and NBA-Palm1.0; it obtains much higher Ac than the highest value of CSS-Palm2.0 and NBA-Palm1.0 and it also receives the highest Sn, Sp and MCC values. Figure 1 shows the ROC curve of CKSAAP-Palm. Since the ROC curves of CSS-Palm2.0 and NBA-Palm1.0 are not available, we adopt a comparison method that has been used in the literature (Dang *et al*., 2008). The performance values are presented by the pairs of sensitivities/specificities reported in Ren *et al.* (2008) and Xue *et al.* (2006), which are shown as colored dots on the ROC plot (Fig. 1). These values are considered to be worse or better depending on whether these dots fall below or above the CKSAAP-Palm ROC curve. In Fig. 1 the colored dots represent the Sn and Sp pairs of the CSS-Palm2.0 with the high, medium and low thresholds, respectively. It is obvious that these colored dots fall below our ROC curve. In addition, for CKSAAP-Palm, the AUC is rather large; it reaches 0.9465, 0.9483 and 0.9593 for 6-, 8- and 10-fold cross-validation, respectively.

Then, consider the first testing data set which involves 53 verified palmitoylation sites in 26 proteins, the results are listed in Table III. Similarly to the training data set, CKSAAP-Palm obtains highest Ac, Sp and MCC values. The Sn of CKSAAP-Palm is smaller than that of CSS-Palm2.0 with medium and low thresholds.

In addition, the proposed method is further tested by multiple tests of randomized training and testing data. Consider the mixed data set by putting the training data set and the first testing data set together. Randomly select 25 palmitoylated proteins from the mixed data set as the new testing data set and make the remaining proteins as the new training data set. The new training data set is used to generate the decision function by our model and the new test data set is used to be tested. This process is repeated five times. The prediction results are displayed in Table IV, whereas the ROC curves are drawn in Fig. 2. The uniprot IDs of proteins in each new testing data set are also provided in Table IV so that other researchers are able to reproduce the data sets. Since the developers of CSS-Palm2.0 and NBA-Palm1.0 did not provide the standalone software and the training data set in their web-based server is fixed, we are not able to obtain the corresponding results of CSS-Palm2.0 and NBA-Palm1.0 on these testing data sets.

Table II. Comparison of CKSAAP-Palm, CSS-Palm2.0 and NBA-Palm1.0 on the training data set which includes 263 palmitoylation sites

Method	Threshold	Ac (%)	Sn (%)	Sp (%)	MCC
CKSAAP-Palm	$C = 100$, $\gamma = 0.0000015$, cutoff = 0.18	93.47	88.42	94.21	0.7536
CSS-Palm2.0	High	89.60	77.19	92.43	0.6709
	Medium	85.92	82.89	86.61	0.6142
	Low	77.00	87.83	74.52	0.5024
NBA-Palm1.0	—	86.67	67.46	92.25	0.6102

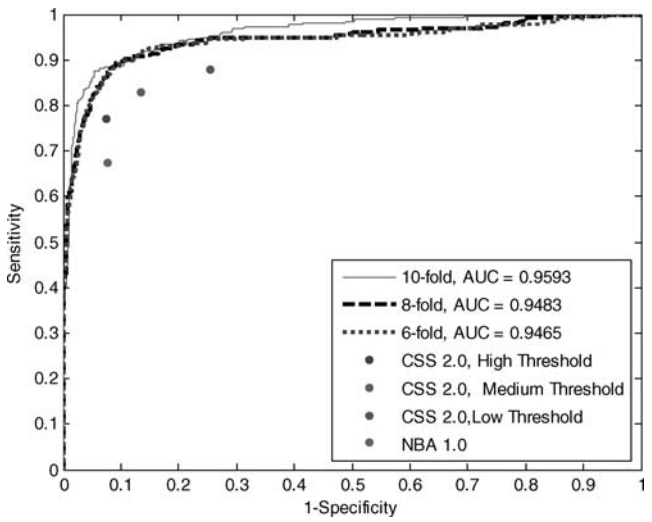

Fig. 1. The ROC curves of 6-, 8- and 10-fold cross-validations.

Table III. Comparison of CKSAAP-Palm and CSS-Palm2.0 on the first testing data set which includes 53 palmitoylation sites

Method	Threshold	Ac (%)	Sn (%)	Sp (%)	MCC
CKSAAP-Palm	$C = 100$, $\gamma = 0.0000015$, cutoff = 0.18	91.18	64.15	95.68	0.6529
CSS-Palm2.0	High	89.49	56.60	94.38	0.5227
	Medium	86.31	73.58	88.20	0.5207
	Low	76.28	81.13	75.56	0.4089

Application in annotating palmitoylated proteins of budding yeast

Our CKSAAP-Palm model is also applied to predict the potential palmitoylation sites in budding yeast proteins identified by Roth *et al.* (2006) recently, the second testing data set mentioned in 'Data sets and method'. The results are listed in Tables V and VI, where three kinds of proteins are examined: known palmitoylated proteins, novel palmitoylated proteins and amino acid permeases (AAPs). Our prediction palmitoylation sites cover most of the ones predicted by CSS-Palm2.0 with the coverage rate of 76 of 89. More interestingly, for the known and novel palmitoylated proteins, CKSAAP-Palm predicts all of them with 100% with at least one palmitoylation site in one palmitoylated protein, whereas CSS-Palm2.0 only with 75% on the known palmitoylated proteins and 74% on the novel palmitoylated proteins respectively. Considering the identified sites of known palmitoylated proteins, CKSAAP-Palm predicts them out with the level of 100%, but CSS-Palm2.0 missed one site C95 in protein Sncl.

Moreover, Roth *et al.* (2006) proposed a novel sequence pattern for palmitoylation recognition which was suggested to be potentially palmitoylated at cysteines cytoplasmically adjacent to their single transmembrane domains. On the basis of this hypothesis, Roth *et al*. gave several suggested palmitoylation sites, including 13 palmitoylated proteins: Snc1, Snc2, Tlg1, Syn8, Sso1, Sso2, Vam3, Tlg2, Mnn10, Mnn11, Pin2, Mnn1 and Ylr001c. CKSAAP-Palm finds almost all of these sites (see embolden text in Table V), whereas CSS-Palm2.0 can only find two of them. This seems not only to support Roth's hypothesis but also show the capability of CKSAAP-Palm to find novel palmitoylation sites, although these sites are still not verified by experiments currently. The eight AAPs, including Agp1, Bap2, Gap1,

Table IV. The performances of CKSAAP-Palm on random select testing data set

Data set	Testing proteins	Sn (%)	Sp (%)	Ac (%)	MCC
1	Q9JJY3, P49802, P03372, Q5ZL02, P35228, P27701, Q777H4, O15498, O35049, P22723, P01730, P04578, P01732, P23292, Q15465, P24668, P11801, P02730, P10966, P25101, P24530, P10300, Q62108, P16144, P04114	61.97	95.09	89.42	0.6312
2	P06401, Q9JIB2, P11686, Q9DDC9, P10275, P13726, P30518, P17677, Q04573, P17501, Q92731, P02699, P07550, P08908, P03522, Q87046, P29474, Q02936, Q9H3Z4, P01889, P49795, O75976, P59768, P10235, O15162	71.66	92.59	88.45	0.6386
3	Q8BQP9, P43119, P28647, P14013, Q13393, Q7ZL00, Q05329, P49817, O75228, P97711, Q99MG9, P13595, P70498, O43665, Q8MMP4, P97288-2, P43250, P36149, P56817, Q91VB2, P43146, P43220, P34981, Q09470, P10235	66.27	90.17	87.51	0.5801
4	P46096, Q9R0N7, P07307, P07306, Q9JI22, Q9JM47, P49798, P42858, P39968, Q16518, P41220, O43617, P16284, Q9H9H5, P26664, P28230, P42261, P16795, P10966, P25101, P24530, P10300, Q62108, P16144, P04114	70.00	90.22	87.38	0.5975
5	Q02936, Q9H3Z4, P01889, P97711, Q99MG9, P13595, P70498, P48509, P04289, P25445, P60033, P21926, Q91X05, O00161, P60880, P42261, P22888, P06241, P06240, P22909, P22725, Q64264, P46096, Q9R0N7, P07307	73.68	93.62	87.87	0.6668

Each testing data set included 25 palmitoylated proteins.

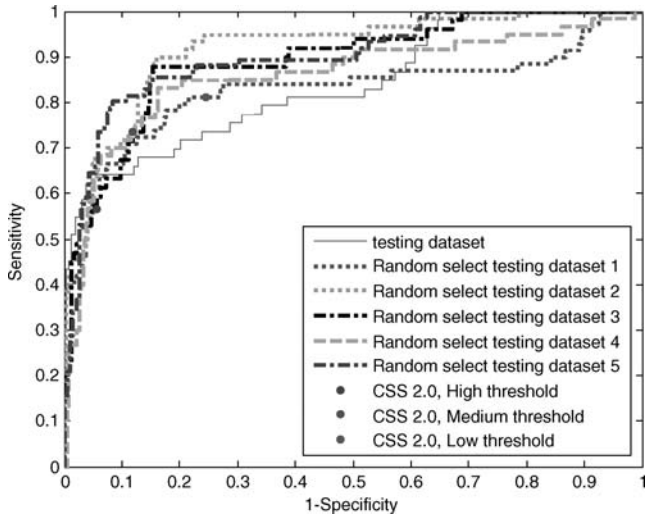

Fig. 2. The ROC curves of the first testing data set and randomly selected testing data sets.

Gnp1, Hip1, Sam3, Tat1 and Tat2, that were suggested to be palmitoylated at C-terminal cysteines (Roth *et al*., 2006). CKSAAP-Palm finds all of them out, whereas CSS-Palm2.0 missed one.

In order to give some instruments for predicting palmitoylation sites, we count the frequencies of features (the *k*-spaced amino acid pairs) which appear in the positive training sequence segments and list the top 20 frequent features in Table VII. These may imply that a new sequence segment containing these features in rich would have palmitoylation sites with high probability. These top 20 features are also used to train the SVM model and the results are shown in Table VIII, which demonstrate that these 20 features may reflect the characteristic of the palmitoylated proteins to some extent.

False-positive prediction

Since the program may be used for whole genome annotation, low false-positive rate is one of the most important characteristics of a good palmitoylation prediction model. In order to evaluate the false-positive rate of CKSAAP-Palm, consider the third testing data set, which includes 120 proteins known not to be palmitoylated, randomly extracted from the Swiss-Prot database (Release 53) with three times. The CKSAAP-Palm model is repeated three times and the average results are shown in Table IX. In site level, the average accuracy is 0.905, which means the false positive rate is $<10\%$. The average accuracy in protein level is 0.767.

Conclusion

In this paper, we have developed an automated computational method to predict palmitoylation sites from protein sequences. The proposed CKSAAP-Palm demonstrates higher prediction accuracy than the existing predictors. The encoding scheme based on the CKSAAP has been proved to be particularly suitable for the prediction of palmitoylation sites. By performing 10-fold cross-validation on the same training data set, CKSAAP-Palm outperforms CSS-Palm2.0 not only in high Sn but also in high Sp and MCC values. On the same testing data set used by CSS-Palm2.0,

Table V. The prediction results for 16 known palmitoylated proteins and 35 novel palmitoylated proteins in budding yeast

Protein	Uniprot	Experimental Site	CSS-Palm2.0 predicted sites	CKSAAP-Palm predicted sites	Roth *et al*. suggested sites
Known palm-proteins					
Ras1	P01119	305	305, 306, 309	305, 306, 309	
Ras2	P01120	318	318, 319	318, 319	
Ste18	P18852	106	56, 106, 107	106, 107	
Gpa1	P08539	3	3	3	
Vac8	P39968	4, 5, 7	4, 5, 7, 106, 149	4, 5, 7	
Gpa2	P10823		4	4	
Yck1	P23291		537, 538	537, 538	
Yck2	P23292	545, 546	545, 546	545, 546	
Yck3	P39962	517, 518, 519, 520	510, 517, 518, 519		
		522, 523, 524	520, 522, 523, 524		
Bet3	P36149	80	80	65, 80, 93	
Lcb4	Q12246	43, 46	43, 46, 358, 359	43, 46, 607	
Akr1	P39010		443, 598	598	
Snc1	**P31109**	95	—	**95**	**95**
Snc2	**P33328**		—	**94**	**94**
Tlg1	**Q03322**		—	**205, 206**	**205, 206**
Syn8	**P31377**		—	**238**	**238**
Novel palm-proteins					
Rho2	P06781	188	188, 189	18, 188, 189	
Rho3	Q00245	5	5, 228	5, 228	
Ycp4	P25349		243, 244	129, 243, 244	
Psr1	Q07800		9, 10	9, 10	
Psr2	Q0794		9, 10	9, 10	
Meh1	Q02205		7, 8	7, 8	
Ygl108c	P53139		4	4	
Ypl236c	Q12003		13, 14, 15	13, 14, 15, 341	
Lsb6	P42951		607	607	
Ypl199c	Q08954		235	233, 235	
Ykl047w	P36090		511, 516	511, 516	
Ybr016w	P38216		119, 122	110, 114, 119, 121,122	
Pin2	Q12057		4, 66, 79, 81, 82	4, **53**, 66, 79, 81	35, 41, **53**
			84	82,84	
Sna4	Q07549		2, 3, 5, 7, 8	2, 3, 5, 7, 8	
Mnn1	P39106		17	17	17
Ylr001c	Q07895		780	334, 780	780
Mlf3	P32047		2, 450	2, 450	
Mse1	P48525		12, 169	12, 140, 142	
Nuc1	P08466		2	2	
Sso1	P32867		—	**266**	**266**
Sso2	P39926		—	**270, 274**	**270, 274**
Vam3	Q12241		—	**262, 274**	**262, 274**
Tlg2	Q08144		—	**325**	317, **325**
Mnn10	P50108		—	**44**	**44**
Mnn11	P46985		—	402, 408	35
Tvp18	A6ZMD0		—	38, 78, 98	
Ylr326w	Q06170		—	154	
AAPs					
Agp1	P25376		—	265, 266, 633	
Bap2	P38084	609	435, 609	241, 609	
Gap1	P19145		397, 602	253, 602	
Gnp1	P48813	663	193, 663	193, 295, 663	
Hip1	P06775	603	339, 397, 400, 603	252, 397, 400, 603	
Sam3	Q08986		123, 377, 587	398, 440, 487, 587	
Tat1	P38085	619	619	239, 619	
Tat2	P38967		289, 592	443, 592	

AAP was suggested to be palmitoylated at C-terminal cysteines Roth *et al.* (2006).

Table VI. Comparison of CKSAAP-Palm and CSS-Palm2.0 on annotation for 16 known palmitoylated proteins and 35 novel palmitoylated proteins in budding yeast

Method	Threshold	Ac (%)	Sn (%)	Sp (%)	MCC
CKSAAP-Palm	$C = 100$, $\gamma = 0.0000015$, cutoff = 0.18	94.85	100.00	93.82	0.8466
CSS-Palm2.0	High	89.60	77.19	92.43	0.6709

CKSAAP-Palm achieves a comparative Sn, but much higher Sp and MCC. Moreover, CKSAAP-Palm predicts more novel palmitoylated proteins and sites suggested by Roth *et al*. (2006). These results show that CKSAAP-Palm is able to predict potential palmitoylation sites and increases research productivity in palmitoylation sites discovery.

It should be pointed out that unlike most of the existing methods for palmitoylation prediction which only provide web-based server, our model CKSAAP-Palm is implemented

Table VII. The top 20 amino acid pairs extracted from training palmitoylated data set

Top 20 features	Amino acid pairs
1	CA
2	LK
3	CXXA
4	LXXK
5	CXXXA
6	CXXXXA
7	LXXXXXA
8	CXA
9	LXXXXXXK
10	LXXXK
11	LXXXXXXXK
12	LXK
13	LXXXXXXXXK
14	LXXXXXA
15	LXXXXXXA
16	LXXXXK
17	LXXA
18	LXXXXXXXXXK
19	LXT
20	VXXXK

For example, CXXA represents a 2-spaced amino acid pair of CA, where X stands for any amino acid.

Table VIII. Comparison of the results using all features and top 20 features

Features	Parameters	Ac (%)	Sn (%)	Sp (%)	MCC
All	$C = 100$, $\gamma = 0.0000015$, cutoff = 0.18	93.47	88.42	94.21	0.7536
Top 20	$C = 100$, $\gamma = 0.15$, cutoff = 0.18	87.03	61.68	92.77	0.5496

The experiments use 10-fold cross-validation on training data set.

Table IX. Accuracy of CKSAAP-Palm for predicting not palmitoylated proteins

Accuracy	Data set 1	Data set 2	Data set 3	Average
Site level	0.836	0.971	0.909	0.905
Protein level	0.675	0.875	0.750	0.767

Each data set contains 40 not palmitoylated proteins randomly selected from the Swiss-Prot database. The parameters are $C = 100$, $\gamma = 0.0000015$ and cutoff = 0.18. Site level accuracy is the percentage of sites that are predicted as non-plamitoylated sites, whereas protein level accuracy is the percentage of proteins whose sites are all predicted as non-palmitoylated sites.

as the standalone software. Taking into account the performance of prediction method depends on the available gold-standard data, the CKSAAP-Palm allows the users to input their training data set. It also allows users to select all parameters such as C, γ and cutoff value. The software can be freely downloaded from http://www.aporc.org/doc/wiki/CKSAAP-Palm.

Acknowledgement

We thank Dr Xue for providing the data set used in the paper. We also thank Professor Daniel Otzen and anonymous reviewers for valuable suggestions.

Funding

This work is supported by National Natural Science Foundation of China (Grant No. 10631070 and No. 60970091), and the Knowledge Innovation Program of the Chinese Academy of Sciences (Grant No. kjcx-yw-s7).

References

Chen,K., Kurgan,L.A. and Ruan,J. (2007) *BMC Struct. Biol.*, **7**, 101–112.
Chen,Y.Z., Tang,Y.R., Sheng,Z.Y. and Zhang,Z.D. (2008) *BMC Bioinformatics*, **9**, 101–112.
Dang,T.H., Leemput,K.V., Verschoren,A. and Laukens,K. (2008) *Bioinformatics*, **9**, 1–21.
el-Husseini Ael,D. and Bredt,D.S. (2002) *Nat. Rev. Neurosci.*, **3**, 792–802.
Kurayoshi,M., Yamamoto,H., Izumi,S. and Kikuchi,A. (2006) *Biochem. J.*, **3**, 101–112.
Li,W.Z. and Godzik,A. (2004) *Bioinformatics*, **22**, 1658–1659.
Linder,M.E. and Deschenes,R.J. (2007) *Nat. Rev. Mol. Cell Biol.*, **8**, 74–84.
Navarro-Lerida,I., Alvarez-Barrientos,A. and Rodrguez-Crespo,I. (2006) *J. Cell Sci.*, **119**, 1558–1596.
Ren,J., Wen,L.P., Gao,X.J., Jin,C.J., Xue,Y. and Yao,X.B. (2008) *Protein Eng. Des. Sel.*, **21**, 639–644.
Roth,A.F., Wan,J., Bailey,A.O., Sun,B., Kuchar,J.A., Green,W.N., Phinney,B.S., Yates,J.R. and Davis,N.G. (2006) *Cell*, **125**, 1003–1013.
Smotrys,J.E. and Linder,M.E. (2004) *Annu. Rev. Biochem.*, **73**, 559–587.
Vapnik,V. (1995) *The Nature of Statistical Learning Theory.* Springer, New York.
Vapnik,V. (1998) *Statistical Learning Theory.* John Wiley and Sons, New York.
Veit,M. and Schmidt,M.F.G. (1998) *Methods Mol. Biol.*, **88**, 227–239.
Xue,Y., Chen,H., Jin,C.J., Sun,Z.R. and Yao,X.B. (2006) *BMC Bioinformatics*, **7**, 458–467.
Zhou,F., Xue,Y., Yao,X. and Xu,Y. (2006) *Bioinformatics*, **22**, 894–896.

Received April 29, 2009; revised August 19, 2009; accepted August 20, 2009

Edited by Daniel Otzen