N-Ace: Using Solvent Accessibility and Physicochemical Properties to Identify Protein N-Acetylation Sites

TZONG-YI LEE,[1]* JUSTIN BO-KAI HSU,[2]* FENG-MAO LIN,[2] WEN-CHI CHANG,[3] PO-CHIANG HSU,[4] HSIEN-DA HUANG[2,5]
[1]*Department of Computer Science and Engineering, Yuan Ze University, Chung-Li 320, Taiwan*
[2]*Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsin-Chu 300, Taiwan*
[3]*Institute of Tropical Plant Sciences, National Cheng Kung University, Tainan 701, Taiwan*
[4]*Institute of Molecular Medicine and Bioengineering, National Chiao Tung University, Hsin-Chu 300, Taiwan*
[5]*Department of Biological Science and Technology, National Chiao Tung University, Hsin-Chu 300, Taiwan*

Received 4 November 2009; Revised 21 March 2010; Accepted 29 March 2010
DOI 10.1002/jcc.21569
Published online 21 May 2010 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Protein acetylation, which is catalyzed by acetyltransferases, is a type of post-translational modification and crucial to numerous essential biological processes, including transcriptional regulation, apoptosis, and cytokine signaling. As the experimental identification of protein acetylation sites is time consuming and laboratory intensive, several computational approaches have been developed for identifying the candidates of experimental validation. In this work, solvent accessibility and the physicochemical properties of proteins are utilized to identify acetylated alanine, glycine, lysine, methionine, serine, and threonine. A two-stage support vector machine was applied to learn the computational models with combinations of amino acid sequences, and the accessible surface area and physicochemical properties of proteins. The predictive accuracy thus achieved is 5% to 14% higher than that of models trained using only amino acid sequences. Additionally, the substrate specificity of the acetylated site was investigated in detail with reference to the subcellular colocalization of acetyltransferases and acetylated proteins. The proposed method, N-Ace, is evaluated using independent test sets in various acetylated residues and predictive accuracies of 90% were achieved, indicating that the performance of N-Ace is comparable with that of other acetylation prediction methods. N-Ace not only provides a user-friendly input/output interface but also is a creative method for predicting protein acetylation sites. This novel analytical resource is now freely available at http://N-Ace.mbc.NCTU.edu.tw/.

 J Comput Chem 31: 2759–2771, 2010

Key words: protein acetylation; acetyltransferase; accessible surface area; physicochemical properties; support vector machine

Introduction

Protein acetylation is a widely studied covalent modification that affects gene regulation in eukaryotic cells. Around 50% of yeast proteins and up to 80–90% of higher eukaryotic proteins are modified by enzymatic acetylation.[1,2] Less acetylated proteins are identified in prokaryotes.[3] The two types of protein acetylation are irreversible and reversible. N^{α}-terminus acetylation is an irreversible modification that occurs cotranslationally in α-amino group, which designates the position of the central carbon atom of amino acids (AAs) and is located only on the N-terminus of the protein. However, the biological mechanism of N^{α}-terminal acetylation in eukaryotic proteins is unclear. Unlike N^{α}-terminal acetylation, N^{ε}-terminus acetylation proceeding in the ε-amino group of lysine residues designates the position of a carbon atom in the side chain. The post-translational ε-amino lysine acetylation of proteins is

Additional Supporting Information may be found in the online version of this article.

*These authors contributed equally to this work.

Authors' Contributions: H.-D.H. conceived and supervised this project. T.-Y.L. was responsible for the design, computational analyses, implementation of the databases, web interface development and drafting of the manuscript, with revisions made by H.-D.H., J.B.-K.H., and W.-C.C. participated in the design, computational analyses, web interface development, and system maintenance. F.-M.L., W.-C.C., and P.-C.H. helped with web interface development, system maintenance, and data testing. All authors read and approved the final manuscript.

Correspondence to: Wen-Chi Chang; e-mail: sarah321@mail.ncku.edu.tw or Hsien-Da Huang; e-mail: bryan@mail.nctu.edu.tw

Contract/grant sponsor: National Science Council of the Republic of China, Taiwan; contract/grant numbers: NSC 98-2627-B-009-005, NSC 98-2311-B-009-004-MY3, and NSC 99-2320-B-155-001.

highly reversible and catalyzed by many lysine acetyltransferases (KATs).[2,4] In an earlier work, the occurrence of N^{ε}-acetylation has been extensively characterized for core histones and over 60 transcription factors.[5–7] Histone acetylation has important roles in the regulation of gene expression and stabilization of the chromatin structure. Additionally, acetylation of the ε-NH_2 in lysine residues is critical to various cellular processes, such as regulation of DNA repair,[5,8,9] DNA replication and recombination,[10] and apoptosis.[11–14] Signal transduction,[5] nuclear import,[15] protein–protein interaction,[16] DNA binding,[5,10,11] and enzyme targeting[16] also involve N^{ε}-acetylation.

Various experimental methods have been used to identify N-acetylated proteins. They include mass spectrometry,[17] the radioactive chemical method,[18] and chromatin immunoprecipitation (ChIP).[19] However, most of them are time consuming and demand extensive resources. The *in silico* identification of protein acetylation sites has potential for characterizing acetylated sites before experiments are performed. This identification can support an effective analysis and efficiently reduce the number of potential targets of acetylation that require further *in vivo* and *in vitro* confirmation. Previous methods predict either irreversibly (N^{α}-terminal) or reversibly (N^{ε}-terminal) acetylated sites. Kiemer et al.[20] developed a method for predicting N-terminal acetylated alanine (A), glycine (G), serine (S), and threonine (T) residues based on a neural network. Its performance in serine acetylation prediction using data on yeast is similar to that achieved using data on mammals, but was worse when other substrates were used. Improving on Kiemer's method, Liu et al. employed the same datasets and a support vector machine (SVM) method to predict N-terminal acetylated sites, with a sensitivity and specificity as high as 86% and 97%, respectively.[21] A system named PAIL was developed for predicting N^{ε}-terminal acetylated sites in lysine using the Bayesian discriminant method (BDM).[22] The proposed accuracies of PAIL are 85.13%, 87.97%, and 89.21% at low, medium, and high thresholds, respectively. Recently, Basu et al. combined experimental methods with the clustering analysis of protein sequences to determine the local amino acid composition.[23] Their method predicts potential acetylation sites and reveals that composition of the sequence can be used to predict two independent experimental sets of data on acetylation marks. A novel method called LysAcet[24] involves protein sequence coupling patterns to improve the prediction of acetylated lysine.

Although protein acetylation is a common protein post-translational modification (PTM), the prediction of which is exceedingly difficult because of a lack of data and a clear consensus motif. Several researches have been performed on the prediction of acetylation sites in protein. However, most previous works in this field have investigated amino acid sequences that surround the acetylation sites, and their predictive performance is usually disappointing. To identify effectively various acetylation sites, this study proposes a method named N-Ace to recognize acetylated sites on alanine, glycine, lysine, methionine, serine, and threonine. Several important features, such as solvent accessibility and physicochemical properties, are considered to identify acetylated sites. A two-stage SVM is utilized to learn the computational models: the first stage of SVM is used to calculate the feature-specific probability and the second stage of SVM is used to construct the predictive models. Based on k-fold cross-validation, the model with the highest predictive accuracy is chosen to implement an effective web-based prediction system. An independent test demonstrated that the high performance of the chosen model is not the result of overfitting the training data. The performance of N-Ace is comparable with that of other methods. A user-friendly web interface is now freely available at http://N-Ace.mbc.NCTU.edu.tw/.

Materials and Methods

Supporting Information Figure S1 presents the system flow of the proposed method, N-Ace. It comprises four major analytical steps: data collection and preprocessing, feature extraction and coding (first stage of SVM), model learning (second stage of SVM) and evaluation, and independent testing. Notably, this work applies two-stage SVM to learn the models for predicting acetylation sites: the first stage of SVM calculates the feature-specific probability for each training feature and the second stage of SVM constructs the predictive model. The details of each process are as follows.

Data Collection and Preprocessing

A comprehensive PTM resource dbPTM,[25] which includes release 53 of UniProtKB/Swiss-Prot,[26,27] comprises 2062 experimentally verified acetylation sites in 1524 protein entries. Supporting Information Table S1 presented detailed statistics concerning each acetylated amino acid. After the nonexperimental sites, annotated "by similarity," "potential," and "probable," have been removed, the remaining acetylated residues for which sufficient experimentally verified data are available concerning over 50 sites were used to investigate the characteristics of substrate sites. This work focuses on acetylated alanine (A), glycine (G), lysine (K), methionine (M), serine (S), and threonine (T) residues, the numbers of which are 424, 60, 792, 240, 431, and 63, respectively. The experimental data on acetylated residues (A, G, K, M, S, or T) constitute the positive data set. The data on alanine, glycine, lysine, methionine, serine, and threonine, which are not annotated as acetylated sites in the experimentally validated acetylated proteins, constitute the negative data set. However, the positive dataset includes data on several homologous sites in orthologous proteins. To prevent any overestimation of predictive performance, homologous sequences were removed from the nonredundant positive data set using a window size of $2n + 1$ and $n + 1$ for N^{α}-terminal acetylation site and N^{ε}-terminal acetylation, respectively. With reference to the reduction of the homology of the training set in NBA-Palm,[28] as shown in Supporting Information Figure S2, two acetylated protein sequences with more than 30% identity were defined as homologous sequences. Then, two homologous sequences were specified to realign the fragment sequences using a window length of $2n + 1$, centered on the acetylated sites using BL2SEQ.[29] For two fragment sequences with 100% identity, when the acetylated sites in the two proteins are in the same positions, only one site was kept while the other was discarded. The nonhomologous positive dataset is comprised of 365 acetylalanine sites, 30 acetylglycine sites, 471 acetyllysine sites, 184 acetylmethionine sites, 343 acetylserine sites, and 57 acetylthreonine sites.

The nonhomologous negative data were generated using the same approach as positive one. To perform fivefold cross-validation, four-fifths of the nonhomologous positive data were selected as the positive training set. The balanced negative train-

ing set was extracted from the nonhomologous negative dataset. To prevent skewing the selection of the training set, given that was a nonhomologous set of N data, the N data were clustered into $N/5$ clusters by the K-mean clustering method. The test set was composed of one datum from each cluster, and the remaining data were defined as the training set. However, the negative training set, randomly selected, may be not be sufficiently randomly sampled. Therefore, 30 negative training sets are obtained by randomly extraction from the nonhomologous negative datasets. The mean predictive performance obtained using the 30 sets of training data is calculated following fivefold cross-validation. The negative test set is also randomly sampled from the nonhomologous negative datasets, which is balanced with the positive test set.

Feature Extraction and Coding

Unlike previous studies,[30,31] this work not only regards the flanking AAs as the training feature but also considers the ASA and physicochemical properties around the acetylated sites. The physicochemical properties, including absolute entropy,[32] nonbonded energy,[33] size,[34] amino acid composition,[35] steric parameter,[36] hydrophobicity,[37,38] volume,[39] mean polarity,[40] electric charge,[41] heat capacity,[32] and isoelectric point,[42] are extracted from Amino Acid index database[43] (AAindex). Fragments of AAs are extracted from positive and negative training sets using a window of length $2n + 1$ varying from 4 to 10 that is centered on N^{α}-terminal acetylation sites and a window of $n + 1$ varying from 8 to 20 for N^{ε}-terminal acetylation sites. Different values of n are used to determine the optimal window length. The positional weighted matrix (PWM) of AAs around the acetylated sites is determined for six acetylated residues (A, G, K, M, S, or T) using nonhomologous training data. The PWM specifies the relative frequency of AAs that surround the acetylated sites and is utilized in encoding the fragment sequences.

The solvent-ASA was also considered to evaluate the characteristics of acetylated residues. As most of the experimental acetylated proteins do not have corresponding protein tertiary structures in PDB, an effective tool, RVP-Net,[44,45] is applied to compute the ASA value from the protein sequence. RVP-net applied a neural network to predict the real ASA of residues based on information about their neighborhood, with a mean absolute error of 18.0–19.5%, defined as the absolute difference between the predicted and experimental values of relative ASA per residue.[45] The computed ASA is the percentage of the solvent-accessible area of each amino acid on the protein. The full-length protein sequences with experimentally identified acetylated sites are inputted to RVP-Net to compute the ASA value of all of the residues. The ASA values of AAs around the acetylated site are extracted and normalized to be between zero and one.

Version 9.1 of AAindex[43] has a total of 544 amino acid indices. It includes many published indices that specify the physicochemical properties of AAs. After the amino acid indices with the value "NA" are eliminated, the remaining 531 physicochemical properties are examined to determine the ability to distinguish the acetylation sites from the nonacetylation sites. As each physicochemical property of the AAs is specified by a set of 20 numerical values, the AAs around the acetylated sites can be encoded according to the values associated with each physicochemical property. The predictive performances obtained using the physicochemical properties are first evaluated, and the properties that are associated with a predictive accuracy of over 60% are defined as useful features for identifying acetylation sites. The AAs, accessible surface area, and useful physicochemical properties are then used to calculate feature-specific probabilities of the training data to generate an input vector use in the second stage of SVM, as displayed in Figure 1.

Model Learning and Evaluation

The SVM is applied to generate computational models that incorporate the encoded AAs, ASA, and physicochemical properties. Based on binary classification, the concept of SVM is to map the input samples into a higher dimensional space using a kernel function and then to find a hyper-plane that discriminates between the two classes with maximal margin and minimal error. A public SVM library, LibSVM,[46] is used to train the predictive model with positive and negative training sets, which are encoded with reference to various training features. The radial basis function (RBF) $K(S_i, S_j) = \exp(-\gamma \|S_i - S_j\|^2)$ is selected as the kernel function of SVM. Cross-validation is important for the application of the predictor.[47] To evaluate the predictive performance of the trained models, k-fold cross-validation is performed on acetylated alanine, lysine, methionine, serine, and threonine. The training data were divided into k groups by splitting each dataset into k approximately equally sized subgroups. In previous works, Jackknife has been demonstrated to be the most objective validation method.[47,48] Therefore, Jackknife cross-validation is adapted to acetylated glycine, for which fewer than 50 data are available. During Jackknife process, both training and testing datasets were generated, and proteins are moved sequentially from one dataset to the other.[48] The following measures of predictive performance of the trained models are defined. Precision (Pr) = TP/(TP + FP), Sensitivity (Sn) = TP/(TP + FN), Specificity (Sp) = TN/(TN + FP), Accuracy (Acc) = (TP + TN)/(TP + FP + TN + FN), and Matthews Correlation Coefficient (MCC) = $\frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP+FN) \times (TN+FP) \times (TP+FP) \times (TN+FN)}}$, where TP, TN, FP, and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively. As 30 negative training sets are used, the mean precision, sensitivity, specificity, accuracy, and MCC are determined for each model that is trained using a particular window length and features. Additionally, the parameters of the predictive models, window length, cost, and gamma value of the SVM models are optimized to maximize predictive accuracy. Finally, the window size and features that yield the highest accuracy are employed to construct predictive models for independent test.

Independent Test

The prediction performance of the trained models may be overestimated because of the overfitting of a training set. To estimate the real prediction performance, the experimental acetylation sites of UniProtKB/Swiss-Prot release 55, which were not

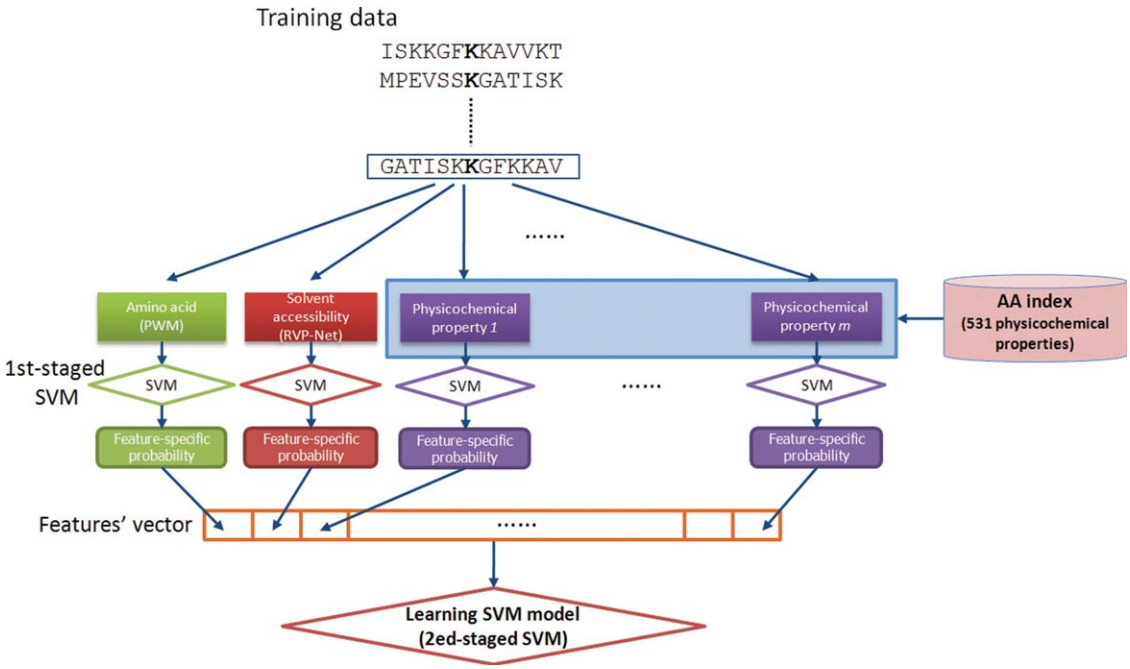


Figure 1. The conceptual diagram of two-stage Support Vector Machine.

included in dbPTM, are chosen as the independent test set. Based on the cross-validation, the trained model with the highest accuracy was used to evaluate the independent test set. As UniProtKB/Swiss-Prot release 55 has no newly identified acetylglycine or acetylmethionine, the independent test set is constructed only for lysine, alanine, serine, and threonine, which are associated with 43, 21, 8, and 2 sites, respectively. The numbers of positive samples and negative samples are equal. These negative samples are randomly selected from the nonacetylation sites. The independent test sets of data for lysine, alanine, serine, and threonine are utilized not only to test the proposed method but also to test other previously proposed protein acetylation prediction tools.

Results and Discussion

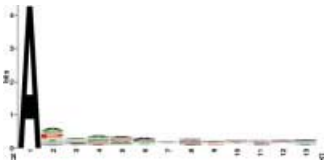
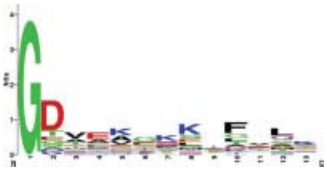
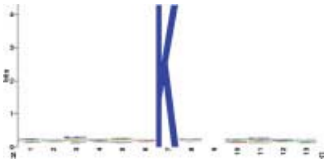
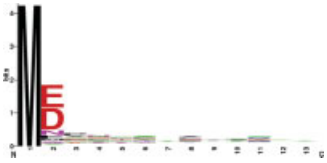
Characterization of Acetylation Sites

This study focuses on the analysis of acetylated alanine, glycine, lysine, methionine, serine, and threonine. After homologous acetylation sites have been removed, as in Table 1, the flanking AAs ($-6 \sim +6$) of the nonhomologous acetylated lysine residues (acetyllysine centered on position 0) and the downstream AAs ($0 \sim +12$) of other N-terminal acetylated residues (which are located at position 0) are graphically visualized as sequence logos. The conservation of AAs that surround the acetylation sites can then be easily explored. WebLogo[49,50] is adopted to generate the graphical sequence logo for the relative frequency of the corresponding amino acid at each position around the acetylated sites. Based on the sequence logo representation, no AAs around the modified sites is obviously conserved, but the acetylated alanine, glycine, methionine, and threonine are somewhat conserved at downstream position +1. However, the conservation of AAs in the flanking regions may be temporary because of the low abundance of experimentally confirmed data of acetylglycine and acetylthreonine.

Determination of Best Window Size Based on Sequence of Amino Acid

To determine what window lengths can be utilized to construct the model that best predicts the sites of acetylation of alanine, glycine, lysine, methionine, serine, and threonine, models that were trained with amino acid sequence are evaluated by cross-validation. Figure 2 presents the predictive performance of the validation based on various window sizes, $2n + 1$, where n is varied from 4 to 10. As different window sizes from 9 to 21 are applied to acetylated lysine, the predictive accuracy does not significantly vary. Nevertheless, the predictive specificity was improved as the window size increased from 9 to 21, while the sensitivity declined. For acetylalanine, the model that was trained with a window size of 13 or 15 was more accurate than the others. The best performance was obtained for acetylglycine and acetylmethionine using models that were trained with a window size of 13. For acetylserine and acetylthreonine, the predictive accuracies slightly improved as the window length increased. Based on computational efficiency and overall performance of the trained models, 13-mer is selected as the window length in the following implementation. Supporting Information Table S3 presents the precision (Pr), sensitivity (Sn), specificity (Sp), accuracy (Acc), and Mathew correlation coefficient (MCC) of the models that were trained with an amino acid

Table 1. The Statistics and Sequence Logos of Nonhomologous Acetylated Sites [Color Table can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Acetylated residues	Number of nonhomologous sites	Number of proteins	Window lengths	Sequence logos
Alanine (A)	356	356	0 ∼ +12	
Glycine (G)	30	30	0 ∼ +12	
Lysine (K)	471	239	−6 ∼ +6	
Methionine (M)	184	184	0 ∼ +12	
Serine (S)	343	343	0 ∼ +12	
Threonine (T)	57	57	0 ∼ +12	

sequence using a window size of 13-mer. Based on fivefold cross-validation, the predictive accuracies of alanine, glycine, lysine, methionine, serine, and threonine are 69.6%, 72.4%, 68.4%, 83.1%, 70.4%, and 73.3%, respectively, indicating that the predictive performance based only on an amino acid sequence is unsatisfactory.

Predictive Performance of Using Various Training Features in Cross-Validation

Most predictive models are based on the features of amino acid sequences. To determine what features can be utilized to construct models that differentiate between acetylation sites and nonacetylation sites, various features, including the sequence of AAs, the accessible surface area, and physicochemical properties are evaluated by cross-validation. The AAs and ASA around the acetylated sites are encoded using a PWM and the RVP-Net-computed ASA values, respectively. The physicochemical properties that were extracted from AAindex are used to encode the AAs that surround the acetylated sites. Table 2 shows the predictive performance achieved using the AAs, the accessible surface area (ASA) and selected physicochemical properties when the accuracy exceeds 60%, based on the fivefold cross-validation. Of the models trained using individual features, that trained with amino acid sequences slightly outperforms that trained using ASA or physicochemical properties when applied to acetylalanine, acetylglycine, acetylmethionine, and acetylthreonine, in which the AAs around the acetylated site are more conserved than are those in aceltylysine and acetylserine. In acetyllysine, the model that is trained using the ASA or hydrophobicity

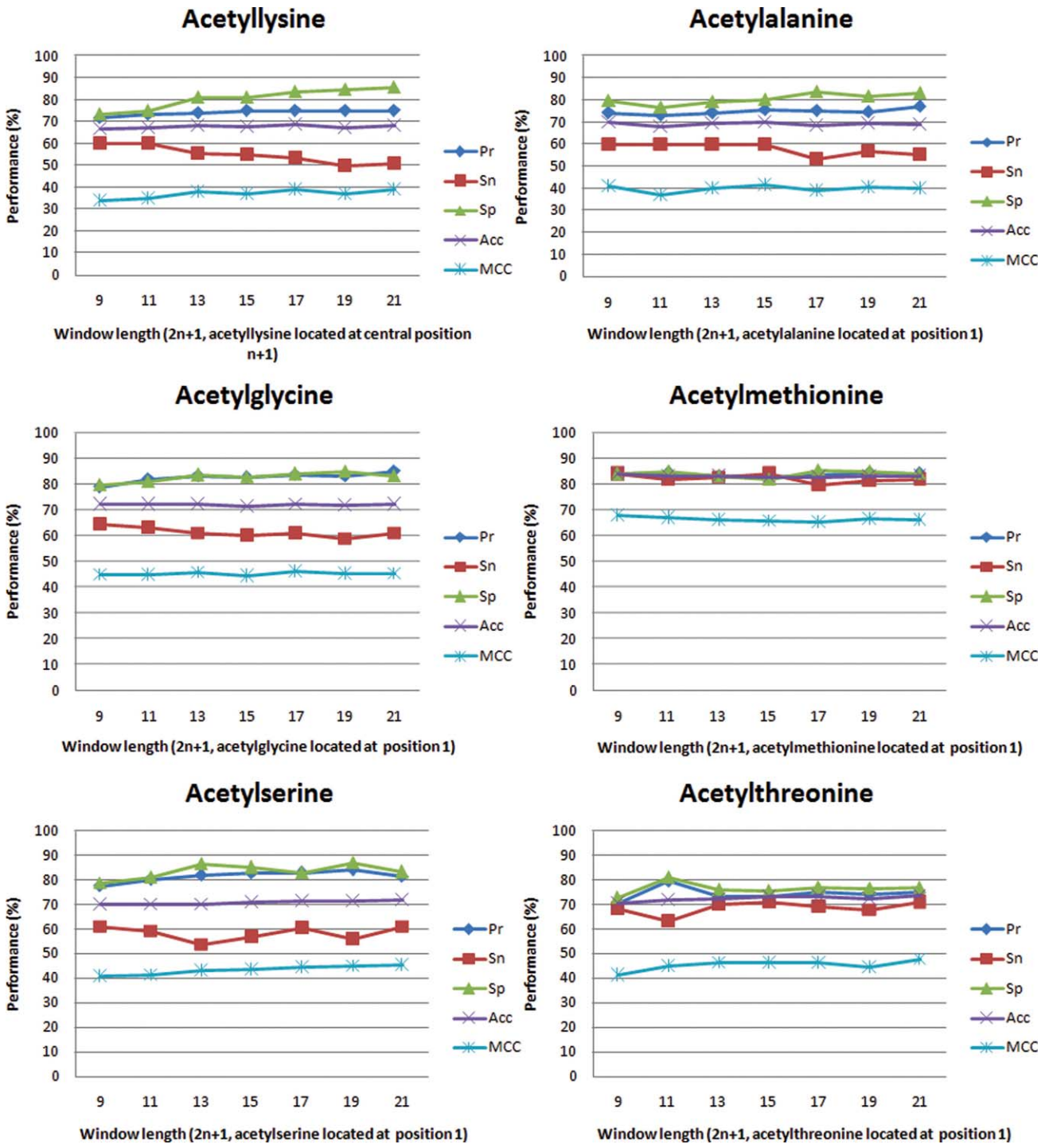


Figure 2. The predictive performance of models that are trained with various windows sizes (based on amino acid sequence). Abbreviations: Pr, precision; Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, Mathew correlation coefficient.

(AAindex: CIDH920105) outperform that trained using the sequence of AAs. For acetylserine, the model trained with hydrophobicity (AAindex: CIDH920105) is more accurate than that trained with an amino acid sequence. Although some of the selected physicochemical properties provide predictive accuracies that are less than 60% when applied to several acetylated residues, their overall performance is satisfactory. The 10 physicochemical properties are absolute entropy, nonbonded energy,

Table 2. The Predictive Powers of Amino Acid, Accessible Surface Area, and the Selected Physicochemical Properties Which Contain More Than 60% Accuracies

Features	AAindex ID	References	Predictive accuracy (%)					
			Ala	Gly	Lys	Met	Ser	Thr
Amino acid sequence	–	–	69.6	72.4	68.4	83.1	70.4	73.3
Accessible surface area	–	Ahmad et al.[44,45]	68.2	66.7	70.1	69.6	70.2	71.6
Absolute entropy	HUTJ700102	Hutchens[32]	59.1	64.3	64.6	73.5	61.8	62.0
Nonbonded energy	OOBM770104	Oobatake and Ooi[33]	65.6	63.8	58.9	69.9	63.2	67.0
Size	DAWD720101	Dawson[34]	63.2	69.7	64.3	76.9	61.3	56.2
Amino acid composition	DAYM780101	Dayhoff et al.[35]	67.6	67.1	66.2	74.9	67.5	71.3
Steric parameter	CHAM810101	Charton[36]	60.0	56.7	63.2	69.3	62.3	57.1
Hydrophobicity	CIDH920105	Jones[38] and Cid et al.[37]	69.2	64.2	69.6	72.1	71.2	67.4
Mean polarity	RADA880108	Radzicka and Wolfenden[40]	58.9	63.2	62.3	69.2	64.5	63.2
Electric charge	FAUJ880111	Fauchere et al.[41]	64.4	62.1	60.6	70.6	61.6	61.9
Heat capacity	HUTJ700101	Hutchens[32]	67.2	71.2	65.5	78.3	60.3	61.1
Isoelectric point	ZIMJ680104	Zimmerman et al.[42]	60.9	57.7	59.6	66.5	60.3	61.1

size, amino acid composition, steric parameter, hydrophobicity, volume, mean polarity, electric charge, heat capacity, and isoelectric point.

Effects of Including Accessible Surface Area and Physicochemical Properties

To improve the predictive performance of protein acetylation sites, the ASA and the selected physicochemical properties were combined with the sequence of AAs to learn the predictive models. Generally, all of the extracted features are combined as a large vector to learn a SVM classifier. As presented in Supporting Information Table S4, the predictive performances of single-stage SVM models that were trained with all features are not clearly enhanced relative to that were trained only with sequence of AAs. The performance is mostly dominated by the sequence of AAs and ASA, so that the performance is only slightly improved by the amount of information given to the SVM. This work utilizes two-stage SVM: the features of the AAs, accessible surface area, and the selected physicochemical properties are input individually to the first stage of SVM to calculate feature-specific probabilities; then, 12 feature-specific probabilities form a vector that is inputted to the second stage of SVM for learning a binary classifier. Table 3 shows that the accuracies of prediction of acetylated alanine, glycine, lysine, methionine, serine, and threonine are 84.9%, 85.1%, 74.9%, 94.0%, 81.5%, and 77.8%, respectively. Figure 3 compares the predictive performance of the models that were trained using only amino acid sequences, those that were trained using all features based on single-stage SVM, and those that were trained using two-stage SVM. The figure reveals that the performance of the prediction of acetylation sites was improved by incorporating the ASA and physicochemical properties into the model. The predictive accuracy of the models that were trained with all features based on two-stage SVM increased from 5% to 14% over those of the models that were trained with only amino acid sequences. Accordingly, the two-stage SVM models that were trained using a combination of amino acid sequences, ASA and physicochemical properties are chosen to construct the classifiers of protein acetylation sites for alanine, glycine, lysine, methionine, serine, and threonine.

Predictive Performance of Independent Test

To evaluate whether the models are overfitted to their training data, independent sets of data concerning acetyllysine, acetylalanine, acetylserine, and acetylthreonine are constructed and used to test the two-stage SVM models, which have the highest pre-

Table 3. The Cross-Validation Performance of Two-Stage SVM Models Trained with the Combination of Amino Acid Sequences, Accessible Surface Area and the Selected Physicochemical Properties

Acetylation residue	No. of nonhomologous positive training set	Window length	Pr (%)	Sn (%)	Sp (%)	Acc (%)	MCC
Alanine	356	$0 \sim +12$	91.1	76.6	93.2	84.9	0.69
Glycine	30	$0 \sim +12$	93.0	80.0	90.1	85.1	0.74
Lysine	471	$-6 \sim +6$	84.6	63.5	86.0	74.9	0.51
Methionine	184	$0 \sim +12$	99.0	89.0	99.0	94.0	0.89
Serine	343	$0 \sim +12$	97.2	65.6	98.4	81.5	0.66
Threonine	57	$0 \sim +12$	78.7	76.0	79.6	77.8	0.56

Abbreviations: Pr, precision; Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, Mathew correlation coefficient.

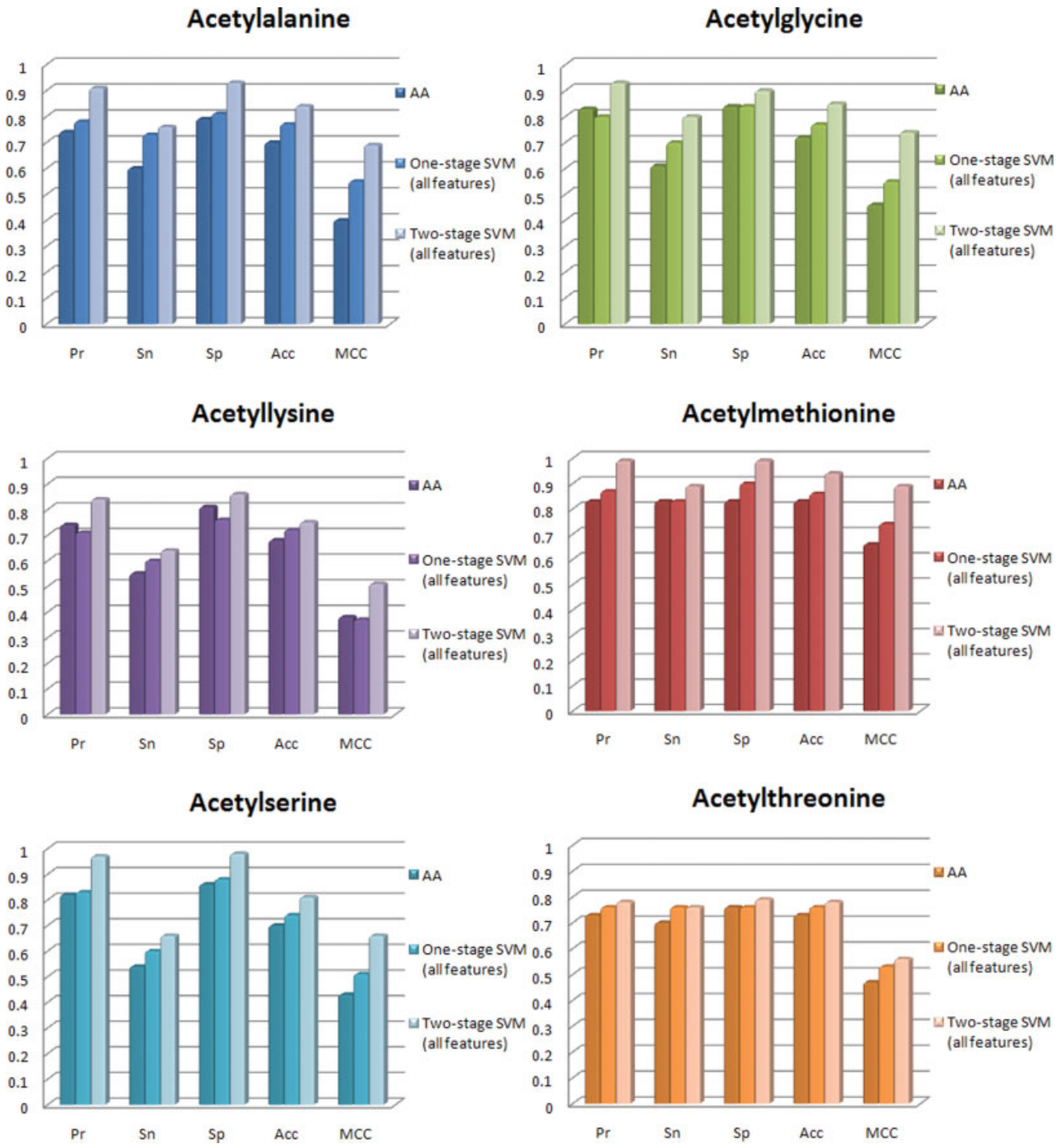


Figure 3. Effects of including accessible surface area and physicochemical properties. Abbreviations: AA, the models trained only with Amino Acid sequence; One-stage SVM, the models trained with all features based on One-stage SVM; Two-stage SVM, the models trained with all features based on Two-stage SVM. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

dictive accuracy. As given in Table 4, the predictive accuracies of the proposed method are 91.5%, 89.6%, 81.3%, and 100% when it is applied to alanine, lysine, serine, and threonine, respectively. Generally, the performance in an independent test approaches that of cross-validation. Although cross-validation outperforms independent testing, performance of the trained

Table 4. The Comparison of Predictive Performance Between N-Ace and Other Tools Based on Independent Test Sets

Tools	Acetylated residue	References	Method	Window length	No. of positive test set	No. of negative test set	Pr (%)	Sn (%)	Sp (%)	Acc (%)	MCC
LysAcet	Lysine	Li et al.[24]	Support vector machine	−6 ∼ +6	43	43	81.1	69.7	83.7	76.7	0.54
PAIL	Lysine	Li et al.[22]	Bayesian discriminant method	−6 ∼ +6	43	43	55.3	83.8	33.4	58.1	0.19
NetAcet	Alanine	Kiemer et al.[20]	Neural network	0 ∼ +12	21	21	0.0	0.0	100.0	50.0	N/A
	Serine			0 ∼ +12	8	8	60.0	75.0	50.0	62.5	0.26
	Threonine			0 ∼ +12	2	2	0.0	0.0	100.0	50.0	N/A
N-Ace	Lysine	–	Two-stage support vector machine	−6 ∼ +6	43	43	84.2	97.9	81.3	89.6	0.80
	Alanine			0 ∼ +12	21	21	88.7	94.2	88.7	91.5	0.83
	Serine			0 ∼ +12	8	8	86.4	75.0	87.5	81.3	0.63
	Threonine			0 ∼ +12	2	2	100.0	100.0	100.0	100.0	1.00

Abbreviations: Pr, precision; Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, Mathew correlation coefficient.

model may be overestimated. The independent test reveals that the constructed two-stage SVM models do not overfit the training data. The independent test sets were used to test other acetylation predictors. It indicates that PAIL[22] has high predictive sensitivity in identifying acetylated lysine but has low predictive specificity when applied to independent test sets. NetAcet[20] has poor sensitivity in identifying acetylated alanine and threonine, but has high specificity. For acetylated serine, NetAcet has satisfactory predictive sensitivity but insufficient specificity. On the prediction page of LysAcet,[24] the kernel function of SVM and feature coding scheme are chosen as polynomial and combined sequence and couple, respectively, and it can reach a good predictive accuracy, 76.7%. The independent test reveals that the two-stage SVM that was conducted herein predicted acetylated sites of both N^{α}-terminal and N^{ϵ}-terminal proteins significantly better than the others methods. This method specifying degree of significance outperforms previous approaches for predicting acetylation sites. Notably, however, the acetylated serine and threonine that were composed of only eight and two elements, respectively, did not suffice to evaluate the predictive performance.

Subcellular Localization of Acetyltransferases and Acetylated Proteins

High-throughput mass spectrometry-based proteomics have led to a rapid increase in the number of experimentally verified acetylated sites, motivating an investigation of the substrates (acetylated proteins) for specific acetyltransferases that are associated with subcellular localization. Based on the annotations of UniProtKB/Swiss-Prot, a total of 317 collected acetyltransferases are categorized according to the localization of nucleus, cytoplasm, membrane, mitochondrion, and others. As presented in Table 5, the subcellular localization of acetyltransferases is mostly in the nucleus, where they are involved in DNA replication, DNA repair, and transcriptional regulation. However, some acetylated proteins are located in various cellular components and participate in different functions. The acetylated proteins may be supposed to be catalyzed by specific acetyltransferases, as determined by the cellular colocalization. Hence, the subcellular localization of acetylated proteins can be used to elucidate the specificity of the substrate for acetylated sites. Supporting Information Table S5 shows various sequence logos of acetylated sites associated with the subcellular localization of acetylated proteins. Figure 4 presents the concept of various acetyllysine site specificities among the acetylated proteins that are localized in different cellular components. The sequence logo of nonhomologous acetylated lysine that is localized in the nucleus has more conserved motifs than the acetylated proteins that are localized in other cellular components.

Table 6 shows the sequence logos of nonhomologous acetylalanine, acetyllysine, acetylmethionine, and acetylserine sites, categorized by the subcellular localization of acetylated proteins. The sequences that flank the acetylation site are clustered into several subgroups, according to the subcellular localization of acetylated proteins. The clustered acetylation sites are separately regarded as training sets to construct the localization-specific SVM models for evaluating the ability to differentiate between acetylated sites and nonacetylated sites. As given in Table 7, the models that are learned from localization-clustered data sets are more sensitive than those to which localization-specific clustering is not applied for both acetyllysine and, acetylated proteins that are localized in the nucleus. However, the specificity of localization-specific models in predicting the sites of the acetylation of lysine is slightly decreased. Additionally, the models that were learned from localization-specific data sets do not outper-

Table 5. Subcellular Localizations of Acetyltransferases Based on the Annotations of UniProtKB/Swiss-Prot Release 53

Subcellular localization	Number of acetyltransferase
Nucleus	145
Cytoplasm	74
Membrane	27
Mitochondrion	0
Others	26
Without annotation	45

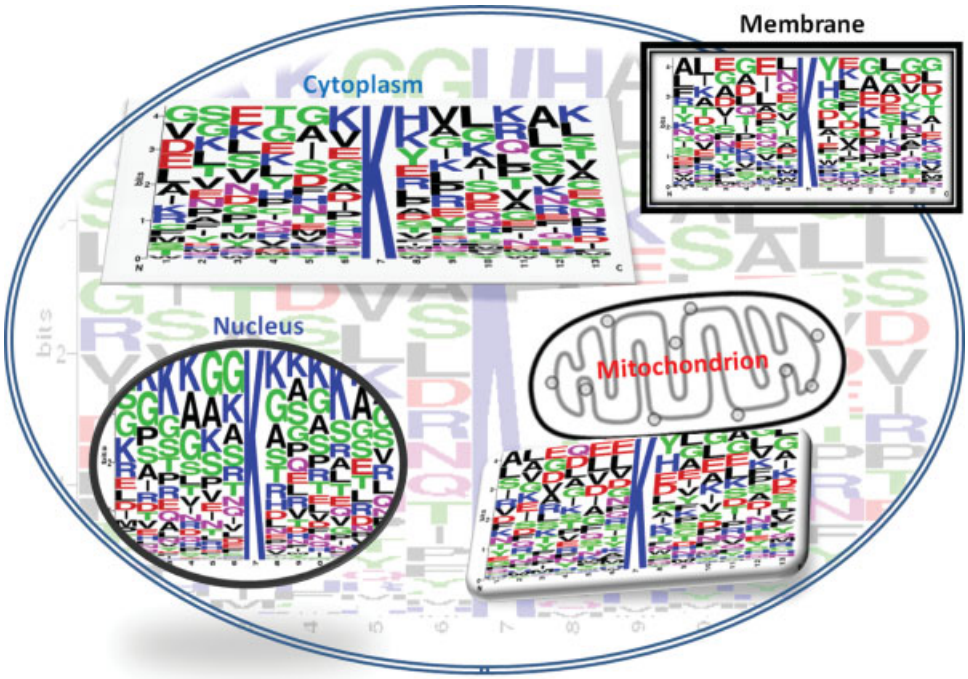


Figure 4. The sequence logos of acetyllysine sites among the acetylated proteins that are localized in different cellular components. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

form those without localization-specific clustering for acetylalanine, acetylmethionine, and acetylserine.

Determination of Data Size for Implementing Web-Based Prediction Models

In the construction of web-based prediction models, the negative set can be argued to be much larger than the positive set in public databases; the negative set may be an unfair sample for training models. This problem also exists for other methods. In this work, the positive and negative training sets were balanced during cross-validation. Thus, 30 sets of negative training data are randomly extracted and used to evaluate predictive performance. However, extracting 30 negative sets to construct 30 predictive models is impossible when a web server is being implemented. Therefore, a larger negative set should be constructed. Unfortunately, using a larger negative set will cause the trained model to prefer to classify negative data correctly, to maximize accuracy. Supporting Information Figure S3 shows the performance of the acetyllysine models, which are trained using different ratios of positive to negative sets. Comprehensively considering the sensitivity, specificity, accuracy, and size of a negative set yields a preferred ratio of the numbers of positive to negative sets of 1:2. This ratio is proposed for use in the model for predicting protein acetylation on the authors' web server.

Implementation of Web-Based Tool for Identifying Protein Acetylation Sites

In the time consuming and laboratory-intensive experimental identification of protein acetylation sites, even though a protein can be acetylated, precisely identifying the acetylated sites on the substrate is difficult. Therefore, an effective prediction tool should be developed to efficiently identify potential acetylation sites. Following evaluation by cross-validation and an independent test, amino acid sequences, the ASA, and 10 useful physicochemical properties are utilized in the construction of two-stage SVM models for predicting the acetylation of alanine, glycine, lysine, methionine, serine, and threonine. As presented in Supporting Information Figure S4, users can submit their uncharacterized protein sequences and select the specific residue whose characteristics are to be predicted. The system efficiently returns the predictions, including acetylated position and the flanking AAs. In particular, users can select various localization-specific models for predicting the acetylation of lysine.

To demonstrate the performance of N-Ace, a case study was presented. Lysine-acetylated proteins are well known to have critical roles in regulating transcription and other DNA-dependent nuclear processes. The FK506 binding protein 4 (FKBP52 protein) is known as a steroid receptor-associated protein. Previous studies have suggested that N6-acetylated Lys274 of FKBP52 is associated with the motor protein dynein and with the cytoskeleton during mitosis.[51] The acetylated Lys274 on FKBP52 is easily retrieved by N-Ace (Supporting Information Fig. S5).

Conclusion

Acetylation prediction methods in previous studies, such as NetAcet[20] and PAIL,[22] have focused only on protein sequence characteristics. However, the scheme in this work incorporates more critical protein features to improve the prediction of pro-

Table 6. The Sequence Logos of Nonhomologous Acetylation Sites Categorized by the Subcellular Localization of Acetylated Proteins. [Color Table can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

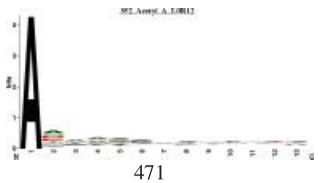
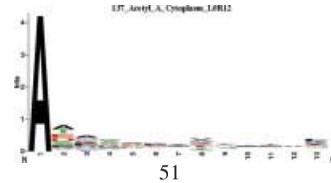
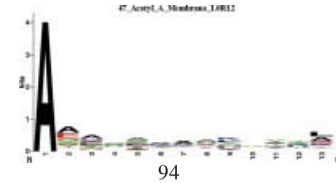
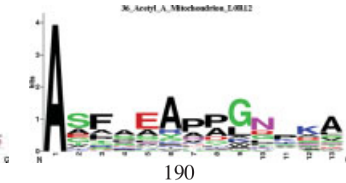
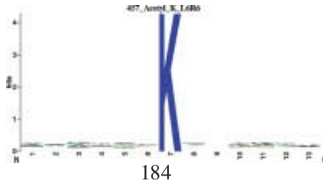
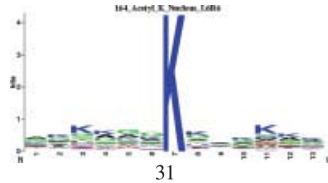
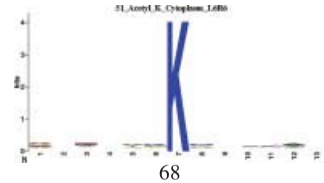
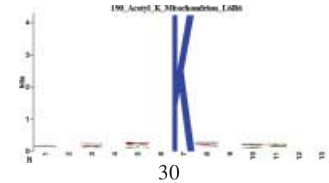
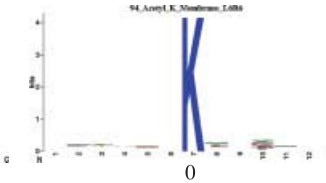
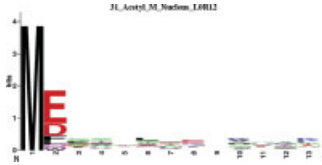
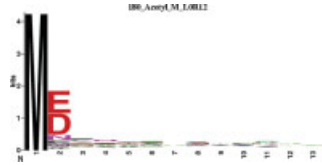
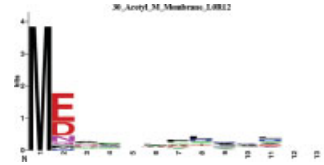
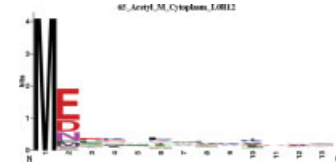
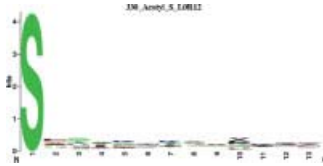
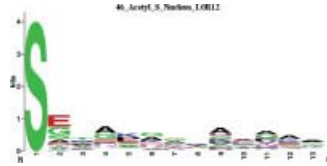
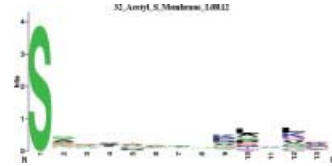
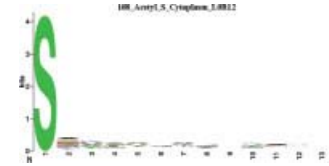
Acetylated residue	Window length	Subcellular localization				
		Total	Nucleus	Cytoplasm	Membrane	Mitochondrion
Alanine	0 ∼ +12	356	48	139	48	36
						
Lysine	−6 ∼ +6	471	178	51	94	190
						
Methionine	0 ∼ +12	184	31	68	30	0
						N/A
Serine	0 ∼ +12	343	52	110	32	0
						N/A

Table 7. The Predictive Performance of the Localization-Specific Models Trained with the Combination of Amino Acid Sequences, Accessible Surface Area, and Physicochemical Properties Based on Two-Stage SVM

Acetylated residue	Subcellular localization	No. of nonhomologous training set	Window length	Pr	Sn	Sp	Acc	MCC
Alanine	Nucleus	48	0 ~ +12	0.77	0.78	0.77	0.77	0.56
	Cytoplasm	139	0 ~ +12	0.93	0.73	0.94	0.84	0.70
	Mitochondrion	36	0 ~ +12	0.93	0.61	0.95	0.78	0.60
	Membrane	48	0 ~ +12	0.89	0.71	0.90	0.80	0.63
	Average	–	0 ~ +12	0.90	0.72	0.90	0.81	0.65
	Total	356	0 ~ +12	0.91	0.76	0.93	0.84	0.69
Lysine	Nucleus	178	−6 ~ +6	0.97	0.81	0.98	0.90	0.80
	Cytoplasm	51	−6 ~ +6	0.77	0.73	0.79	0.76	0.52
	Mitochondrion	190	−6 ~ +6	0.70	0.67	0.72	0.69	0.39
	Membrane	94	−6 ~ +6	0.66	0.68	0.65	0.67	0.34
	Average	–	−6 ~ +6	0.80	0.74	0.80	0.77	0.53
	Total	471	−6 ~ +6	0.84	0.64	0.86	0.75	0.51
Methionine	Nucleus	31	0 ~ +12	0.90	0.93	0.88	0.90	0.83
	Cytoplasm	68	0 ~ +12	0.90	0.87	0.89	0.88	0.77
	Membrane	30	0 ~ +12	0.89	0.80	0.90	0.85	0.71
	Average	–	0 ~ +12	0.90	0.87	0.89	0.88	0.77
	Total	184	0 ~ +12	0.99	0.89	0.99	0.94	0.89
Serine	Nucleus	52	0 ~ +12	0.97	0.65	0.98	0.82	0.67
	Cytoplasm	110	0 ~ +12	0.79	0.60	0.84	0.72	0.45
	Membrane	32	0 ~ +12	0.66	0.71	0.63	0.67	0.35
	Average	–	0 ~ +12	0.82	0.63	0.84	0.74	0.49
	Total	343	0 ~ +12	0.97	0.65	0.98	0.81	0.66

Abbreviations: Pr, precision; Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, Mathew correlation coefficient.

tein acetylation sites. These features include the amino acid sequence, accessible surface area, absolute entropy, non-(bonded TRY bonding) energy, size, amino acid composition, steric parameter, hydrophobicity, volume, mean polarity, electric charge, heat capacity, and isoelectric point. Based on two-stage SVM, the predictive accuracies of acetyllysine, acetylalanine, acetylglycine, acetylmethionine, acetylserine, and acetylthreonine are 84.9%, 85.1%, 74.9%, 94.0%, 81.5%, and 77.8%, respectively. A comparison of the performance of our approach and previous methods[20,22] reveals that N-Ace has a much higher predictive accuracy than the other methods according to independent testing. Additionally, the models for predicting the acetyllysine, which learn from localization-specific datasets, outperform those in which localization-specific clustering is not applied, especially for acetylated proteins that are localized in the nucleus.

Although the proposed method can perform accurately and robustly, according to independent tests, some issues must still be addressed in future work. First, the structural preferences of acetylated sites should be investigated in greater detail, especially in acetylated lysine and serine, whose flanking residues are not conserved. As well as the solvent accessible surface area, secondary structure, B-factor, intrinsic disordered region, protein linker region, and other factors at experimental acetylation sites that are located in the protein regions with PDB entries, should be studied. Second, the independent test sets that are proposed herein are really blind to the trained model during cross-validation, but may not be to previously proposed predictors. Hence, a benchmark for constructing test sets that are truly independent of each predictor is important. Finally, about 100 acetylated lysine sites were annotated as methylation sites, based on the statistics in UniProtKB/SwissProt release 53.0. Therefore, N-Ace might not effectively distinguish the acetylated lysine from the methylated lysine because the methyllysine and acetyllysine alternate in many locations of acetylated proteins. Acetyllysine and methyllysine should be examined in detail, not only with reference to AAs.

Acknowledgments

Ted Knoy is appreciated for his editorial assistance. N-Ace can be accessed via a web interface and is freely available to all interested users at http://N-Ace.mbc.nctu.edu.tw. The authors would like to thank the National Science Council of the Republic of China, Taiwan, for financially supporting this research under Contract No. NSC 98-2627-B-009-005, NSC 98-2311-B-009-004-MY3, and NSC 99-2320-B-155-001.

References

1. Polevoda, B.; Sherman, F. J Biol Chem 2000, 275, 36479.
2. Polevoda, B.; Sherman, F. Genome Biol 2002, 3, reviews 0006.1–0006.6.
3. Polevoda, B.; Sherman, F. J Mol Biol 2003, 325, 595.
4. Yang, X. J. Bioessays 2004, 26, 1076.
5. Yang, X. J.; Gregoire, S. EMBO Rep 2007, 8, 556.
6. Glozak, M. A.; Sengupta, N.; Zhang, X.; Seto, E. Gene 2005, 363, 15.

7. Nightingale, K. P.; O'Neill, L. P.; Turner, B. M. Curr Opin Genet Dev 2006, 16, 125.
8. Ramanathan, B.; Smerdon, M. J. J Biol Chem 1989, 264, 11026.
9. Murr, R.; Loizou, J. I.; Yang, Y. G.; Cuenin, C.; Li, H.; Wang, Z. Q.; Herceg, Z. Nat Cell Biol 2006, 8, 91.
10. Groth, A.; Rocha, W.; Verreault, A.; Almouzni, G. Cell 2007, 128, 721.
11. Cohen, H. Y.; Lavu, S.; Bitterman, K. J.; Hekking, B.; Imahiyerobo, T. A.; Miller, C.; Frye, R.; Ploegh, H.; Kessler, B. M.; Sinclair, D. A. Mol Cell 2004, 13, 627.
12. Luo, J.; Su, F.; Chen, D.; Shiloh, A.; Gu, W. Nature 2000, 408, 377.
13. Subramanian, C.; Opipari, A. W., Jr.; Bian, X.; Castle, V. P.; Kwok, R. P. Proc Natl Acad Sci USA 2005, 102, 4842.
14. Tang, Y.; Luo, J.; Zhang, W.; Gu, W. Mol Cell 2006, 24, 827.
15. Bannister, A. J.; Miska, E. A.; Gorlich, D.; Kouzarides, T. Curr Biol 2000, 10, 467.
16. Kouzarides, T. EMBO J 2000, 19, 1176.
17. Medzihradszky, K. F. Methods Enzymol 2005, 402, 209.
18. Welsch, D. J.; Nelsestuen, G. L. Biochemistry 1988, 27, 4939.
19. Umlauf, D.; Goto, Y.; Feil, R. Methods Mol Biol 2004, 287, 99.
20. Kiemer, L.; Bendtsen, J. D.; Blom, N. Bioinformatics 2005, 21, 1269.
21. Liu, Y.; Lin, Y. Genomics Proteomics Bioinformatics 2004, 2, 253.
22. Li, A.; Xue, Y.; Jin, C.; Wang, M.; Yao, X. Biochem Biophys Res Commun 2006, 350, 818.
23. Basu, A.; Rose, K. L.; Zhang, J.; Beavis, R. C.; Ueberheide, B.; Garcia, B. A.; Chait, B.; Zhao, Y.; Hunt, D. F.; Segal, E.; Allis, C. D.; Hake, S. B. Proc Natl Acad Sci USA 2009, 106, 13785.
24. Li, S.; Li, H.; Li, M.; Shyr, Y.; Xie, L.; Li, Y. Protein Pept Lett 2009, 16, 977.
25. Lee, T. Y.; Huang, H. D.; Hung, J. H.; Huang, H. Y.; Yang, Y. S.; Wang, T. H. Nucleic Acids Res 2006, 34(Database issue), D622.
26. Bairoch, A.; Apweiler, R. Nucleic Acids Res 1998, 26, 38.
27. Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M. C.; Estreicher, A.; Gasteiger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I.; Pilbout, S.; Schneider, M. Nucleic Acids Res 2003, 31, 365.
28. Xue, Y.; Chen, H.; Jin, C.; Sun, Z.; Yao, X. BMC Bioinformatics 2006, 7, 458.
29. Tatusova, T. A.; Madden, T. L. FEMS Microbiol Lett 1999, 174, 247.
30. Shien, D. M.; Lee, T. Y.; Chang, W. C.; Hsu, J. B.; Horng, J. T.; Hsu, P. C.; Wang, T. Y.; Huang, H. D. J Comput Chem 2009, 30, 1532.
31. Chang, W. C.; Lee, T. Y.; Shien, D. M.; Hsu, J. B.; Horng, J. T.; Hsu, P. C.; Wang, T. Y.; Huang, H. D.; Pan, R. L. J Comput Chem 2009, 30, 2526–2537.
32. Hutchens, J. O. In Handbook of Biochemistry, 2nd ed.;Sober, H. A., Ed.; Chemical Rubber Co: Cleveland, Ohio, 1970; pp. B60–B61.
33. Oobatake, M.; Ooi, T. J Theor Biol 1977, 67, 567.
34. Dawson, D. M.; Brock, D. J. H.; Mayo, O., Eds.; The Biochemical Genetics of Man, Academic Press: New York, 1972; pp. 1–38.
35. Dayhoff, M. O.; Hunt, L. T.; Hurst-Calderone, S. In Atlas of Protein Sequence and Structure, Vol 5, Suppl 3; Dayhoff, M. O., Ed.; National Biomedical Research Foundation: Washington, DC, 1978; p 363.
36. Charton, M. J Theor Biol 1981, 91, 115.
37. Cid, H.; Bunster, M.; Canales, M.; Gazitua, F. Protein Eng 1992, 5, 373.
38. Jones, D. D. J Theor Biol 1975, 50, 167.
39. Pontius, J.; Richelle, J.; Wodak, S. J. J Mol Biol 1996, 264, 121.
40. Radzicka, A.; Wolfenden, R. Biochemistry 1988, 27, 1664.
41. Fauchere, J. L.; Charton, M.; Kier, L. B.; Verloop, A.; Pliska, V. Int J Pept Protein Res 1988, 32, 269.
42. Zimmerman, J. M.; Eliezer, N.; Simha, R. J Theor Biol 1968, 21, 170.
43. Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. Nucleic Acids Res 2008, 36(Database issue), D202–205.
44. Ahmad, S.; Gromiha, M. M.; Sarai, A. Bioinformatics 2003, 19, 1849.
45. Ahmad, S.; Gromiha, M. M.; Sarai, A. Proteins 2003, 50, 629.
46. Chang, C.-C.; Lin, C.-J. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.
47. Chou, K. C.; Shen, H. B. Anal Biochem 2007, 370, 1.
48. Chou, K. C.; Zhang, C. T. Crit Rev Biochem Mol Biol 1995, 30, 275.
49. Crooks, G. E.; Hon, G.; Chandonia, J. M.; Brenner, S. E. Genome Res 2004, 14, 1188.
50. Schneider, T. D.; Stephens, R. M. Nucleic Acids Res 1990, 18, 6097.
51. Wochnik, G. M.; Ruegg, J.; Abel, G. A.; Schmidt, U.; Holsboer, F.; Rein, T. J Biol Chem 2005, 280, 4609.