Incorporating Structural Characteristics for Identification of Protein Methylation Sites

DRAY-MING SHIEN,[1*] TZONG-YI LEE,[2*] WEN-CHI CHANG,[2,3] JUSTIN BO-KAI HSU,[2] JORNG-TZONG HORNG,[1,4] PO-CHIANG HSU,[5] TING-YUAN WANG,[2] HSIEN-DA HUANG[2,3]

[1]*Department of Computer Science and Information Engineering, National Central University, Chung-Li 320, Taiwan*

[2]*Institute of Bioinformatics and Systems Biology, College of Biological Science and Technology, National Chiao Tung University, Hsin-Chu 300, Taiwan*

[3]*Department of Biological Science and Technology, National Chiao Tung University, Hsin-Chu 300, Taiwan*

[4]*Department of Bioinformatics, Asia University, Taichung, Taiwan*

[5]*Institute of Biochemical Engineering, College of Biological Science and Technology, National Chiao Tung University, Hsin-Chu 300, Taiwan*

Received 2 September 2008; Revised 5 January 2009; Accepted 7 January 2009
DOI 10.1002/jcc.21232
Published online 4 March 2009 in Wiley InterScience (www.interscience.wiley.com).

Abstract: Studies over the last few years have identified protein methylation on histones and other proteins that are involved in the regulation of gene transcription. Several works have developed approaches to identify computationally the potential methylation sites on lysine and arginine. Studies of protein tertiary structure have demonstrated that the sites of protein methylation are preferentially in regions that are easily accessible. However, previous studies have not taken into account the solvent-accessible surface area (ASA) that surrounds the methylation sites. This work presents a method named MASA that combines the support vector machine with the sequence and structural characteristics of proteins to identify methylation sites on lysine, arginine, glutamate, and asparagine. Since most experimental methylation sites are not associated with corresponding protein tertiary structures in the Protein Data Bank, the effective solvent-accessible prediction tools have been adopted to determine the potential ASA values of amino acids in proteins. Evaluation of predictive performance by cross-validation indicates that the ASA values around the methylation sites can improve the accuracy of prediction. Additionally, an independent test reveals that the prediction accuracies for methylated lysine and arginine are 80.8 and 85.0%, respectively. Finally, the proposed method is implemented as an effective system for identifying protein methylation sites. The developed web server is freely available at http://MASA.mbc.nctu.edu.tw/.

© 2009 Wiley Periodicals, Inc. J Comput Chem 30: 1532–1543, 2009

Key words: protein methylation; solvent accessible surface area (ASA); support vector machine (SVM)

Introduction

Background

Protein post-translational modifications (PTMs), which influence the structural and functional diversity of proteome and determine cellular plasticity and dynamics, have critical roles in many biological processes. Protein methylation, which was discovered nearly 40 years ago,[1] is an important and reversible PTM. However, protein methylation has not been studied as much is known about the processes and implications of phosphorylation.[2] Protein methylation occur on nitrogen atoms of either the backbone or side-chain (N-methylation) of lysine, arginine, asparagine, histidine, alanine, proline, and other residues.[3–7] Methylation can also occur on the oxygen atoms of glutamate and aspartate (O-methylation)[8] and the sulfur atom of cysteine (S-methylation).[9] These modifications are carried out by a protein family called methyltransferases, which use S-adenosylmethionine as a sub-

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

*These authors contributed equally.

Correspondence to: H.-D. Huang; e-mail: bryan@mail.nctu.edu.tw

Contract/grant sponsor: National Science Council of the Republic of China, Taiwan; contract/grant numbers: NSC 96-3112-E-009-002, 97-2627-B-009-007

Contract/grant sponsor: National Research Program for Genomic Medicine (NRPGM), Taiwan

strate to transfer a methyl group.[10] Most studies of protein methylation have focused on methylated arginine and lysine residues.

Three forms of methylated arginine—mono-methylarginine, symmetric di-methylarginine and asymmetric di-methylarginine—have been produced with catalysis by eight protein arginine methyltransferases (PRMTs). Two PRMTs are applied to form mono-methylarginine. Type I PRMTs (PRMT1, PRMT3, PRMT4, and PRMT6) produce asymmetric di-methylarginine, whereas type II PRMTs (PRMT5 and PRMT7) produce symmetric dimethylarginine.[11] The methylation of arginines has been identified in transcriptional regulation, RNA processing, signal transduction, DNA repair, cell-type differentiation, genome stability, and cancer.[6,12] Lysine methylation was first identified on histone protein in the 1960s.[13] Lysine residues can be mono-, di-, or tri-methylated by histone lysine methyltransferases (HKMTs).[14] The methylation of lysine has been mostly studied in H3 and H4 histone proteins, which are critical in various biological processes, such as heterochromatin compaction, X-chromosome inactivation and transcriptional silencing or activation.[4,6] Additionally, HKMTs modify several nonhistone proteins with diverse functions.[4–6] For example, Set9 methylates a transcription factor TAF10 to increase its interacting affinity with RNA polymerase II, which is involved in the transcriptional regulation of TAF10 target genes.[15]

Structural Characteristics of Methylated Sites

A side-chain of amino acid (AA) that undergoes PTM prefers to be accessible on the surface of a protein.[16] To study the preference of the solvent accessible surface area (ASA) that surrounds methylation sites in protein tertiary structures, the experimentally identified methylation sites should be mapped to the corresponding positions of protein entries in the Protein Data Bank (PDB).[17] The preference of the secondary structure (SS) around the methylation sites is also considered. DSSP[18] is a database of SS assignments for all protein entries in the PDB. DSSP also provides a program for calculating the solvent accessibility and standardizing the SS of PDB entries. Table S1 (see Supporting Information) presents in detail the mapping hits of methylated residues between UniProtKB/Swiss-Prot[19] and PDB, which comprises seven methylarginines and 55 methyllysines. In the case of arginine, only seven methylated sites were hit, and of the three sites observed in helical region, two were in the sheet and the other was in the coil. As shown in Table S2 (see Supporting Information), the mean percentage of solvent ASA of methylarginine is 25%. For lysine, which consists of 55 sites covered by the PDB hits, of the 26 sites are observed in coil regions, 19 are in helical regions, and the others are in sheet regions. The mean percentage of solvent ASA of methyllysine is 61%, which is highly exposed to the solvent. Although the number of experimental methylated sites in the protein regions with a tertiary structure is too few to elucidate the real preferences of solvent accessibility and SS for protein methylation sites, this observation demonstrates that the methylated lysine tends to be on the exposed and coil regions. Even though protein methylated sites may not always be in solvent-accessible regions, solvent-accessible AAs are more likely to be modified than buried AAs.

Related Work

Because of the importance of protein methylation in biological mechanism, more attention is being paid to high-throughput proteomic studies, which have been identified an increasing number of experimentally verified methylation sites. However, experimental identification of methylation sites is time-consuming and lab-intensive. Computational prediction can not only identify the potential methylation sites, but also facilitate downstream functional analysis. Therefore, two works have computationally identified the potential methylation sites on lysine and arginine. Daily et al.[20] developed a method for identifying methylated arginine and methylated lysine, using a support vector machine (SVM) based on the observation that post-translation modifications (PTMs) preferentially occur in intrinsically disordered regions. Sequences are encoded by a set of features, including AA frequencies, aromatic content, a flexibility scalar, net charge, hydrophobic moment, beta entropy, disorder information as well as PSI-BLAST profiles. Chen et al.[21] constructed an effective prediction server for indentifying methylation on arginine and lysine based on SVM with positive sets of data that were experimentally confirmed methylated sites from UniProtKB/Swiss-Prot[19] (version 48) and manually collected from the literature.

Several investigations proposed links between PTMs and their associated solvent ASA. Pang et al.[16] studied the structural environment of 8378 incidences in 44 PTMs. It has been observed that protein methylation prefers to occur in regions that are intrinsically disorder and easily accessible. In previous study, solvent accessibility was incorporated into a PTM resource, dbPTM,[22] to promote the detection of phosphorylation, glycosylation, and tyrosine sulfation sites, whose residues, when solvent accessibility exceeded a threshold, were identified as surface modification sites. Arthur et al.[23] employed homology modeling of the protein tertiary structure and calculated the solvent accessibility of the predicted structure to identify phosphorylation sites. Therefore, the solvent accessibility around the protein methylated sites may be adapted to evaluate the classifying performance when differentiates the methylation site from unmethylation sites.

Motivation and Goals

Since protein methylation preferentially occurs in regions that are easily accessible, a method, named MASA, incorporates an SVM is proposed herein to identify protein methylation sites with sequenced and structural characteristics, such as the solvent ASA and SS around the methylated sites. Most of the experimentally verified methylation sites were collected from UniProtKB/Swiss-Prot[19] release 53. Additionally, various experimental methylation sites were taken from MeMo,[21] whose authors extracted many manually curated data by searching the PubMed literature database. However, most of the collected methylation sites do not have the corresponded protein tertiary structures of PDB. Because of the missing ASA and SS information for non-PDB proteins, two effective tools, RVP-Net[24,25] and PSIPRED,[26] are employed to determine the ASA value of AAs and the SSs of AAs in proteins, respectively.

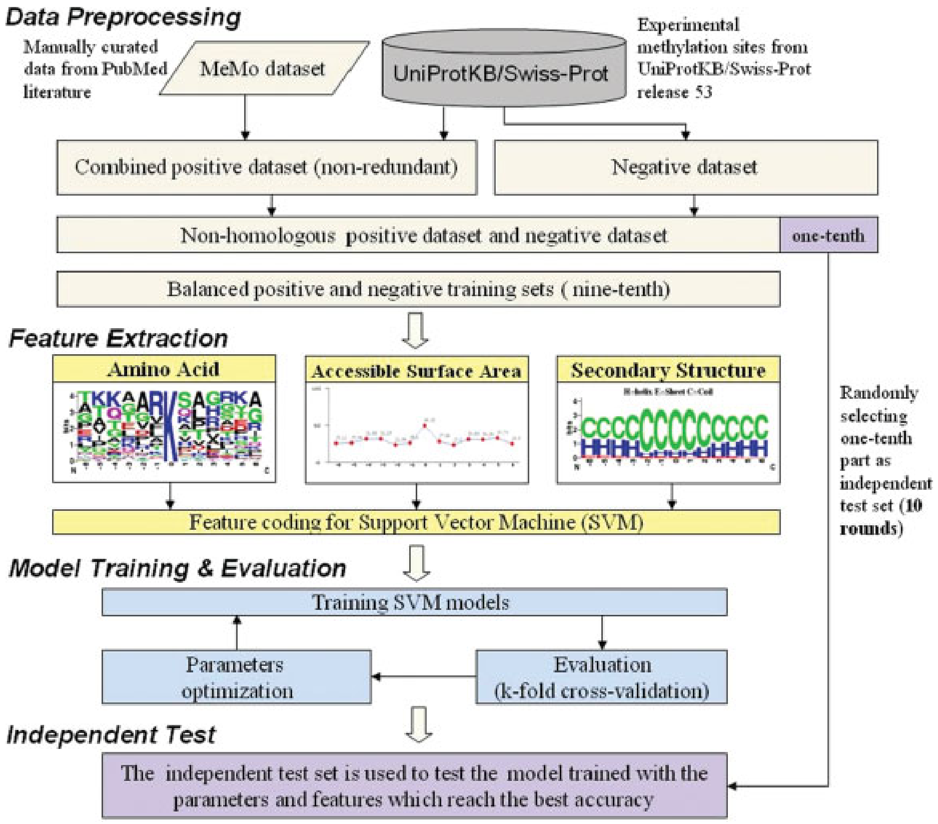

Data Preprocessing
Manually curated data from PubMed literature
MeMo dataset
UniProtKB/Swiss-Prot
Experimental methylation sites from UniProtKB/Swiss-Prot release 53
Combined positive dataset (non-redundant)
Negative dataset
Non-homologous positive dataset and negative dataset
one-tenth
Balanced positive and negative training sets (nine-tenth)
Feature Extraction
Amino Acid
Accessible Surface Area
Secondary Structure
Randomly selecting one-tenth part as independent test set (10 rounds)
Feature coding for Support Vector Machine (SVM)
Model Training & Evaluation
Training SVM models
Parameters optimization
Evaluation (k-fold cross-validation)
Independent Test
The independent test set is used to test the model trained with the parameters and features which reach the best accuracy

Figure 1. Analyzing flowchart of MASA.

This work focuses not only on methylated lysine and arginine, but also studies methylated glutamate and asparagine. To prevent overestimates of predictive performance, the homologous sequences of the nonredundant positive training set taken from UniProtKB/Swiss-Prot[19] and MeMo[21] were further removed among orthologous proteins using a given window size. The cross-validation of models trained with various features, such as AAs, SSs and solvent ASAs, indicates that the ASA value of the AAs around the methylation sites can improve the accuracy of prediction. The accuracies of predicting the methyllysine, methylarginine, methylglutamate, and methylasparagine are 74.6, 84.9, 100, and 100%, respectively. Additionally, the independent test set, which is not contained in the training set, is used to determine whether the constructed model is over fitting to the training set. The independent test shows that the proposed method does not over-fit, and the predictive accuracies of methylated lysine and arginine are 80.8 and 85.0%, respectively. Finally, the window size and training features that provide the best performance are adopted to implement an effective web-based methylation prediction system for biologists. Users can submit their uncharacterized protein sequences and select the specific residue that is to be predicted. The web server presents graphically the overall predicted methylation sites and the solvent ASA. The predicted results in tab-delimited format can be downloaded for further analysis.

Materials and Methods

Data Preprocessing

As presented in Figure 1, the proposed approach, MASA, comprises four main analytical processes—data preprocessing, feature extraction and coding, model training and evaluation, and independent testing. The experimentally verified methylation sites were taken from UniProtKB/Swiss-Prot[19] and MeMo,[21] which is a web tool for predicting protein methylation modifications on arginine (R) and lysine (K). The authors of MeMo extracted many manually curated data by surveying the literature using the keywords “methylated lysine” and “methylated arginine” for information on lysine and arginine methylation, respectively. As shown in Table 1, release 53 of UniProtKB/Swiss-Prot contains 750 experimental methylation sites, which are not annotated as “by similarity,” “potential,” or “probable,” in 352 proteins that are experimentally confirmed to be methylated protein. Because of the absence of sufficient experimental verified data (at least 20 sites) in other residues, the experimental methylated sites are categorized into four types of AAs, including 389 methylated lysines (K), 180 methylated arginines (R), 45 methylated glutamates (E), and 22 methylated asparagines (N). A total of 399 experimental methylation sites are taken from MeMo, including 264 methylated arginine and 107

Table 1. Data Statistics of Methylation Sites Obtained from UniProtKB/Swiss-Prot and MeMo.

		Number of methylated sites					
Data sources	Number of methylated proteins	Lysine (K)	Arginine (R)	Glutamate (E)	Asparagine (N)	Other residues	Total
UniProtKB/Swiss-Prot (Release 53)[a]	352	389	180	45	22	114	750
MeMo[b]	144	107	264	–	–	28	399
Combined experimental data (nonredundant)	446	460	303	45	22	136	966
Negative data	446	6,237	1,216	885	375	–	91,182

[a] The entries which are not annotated as "by similarity," "potential," or "probable" in the "MOD_RES" fields of UniProtKB/Swiss-Prot are the experimentally verified methylation sites.

[b] The manually curated data which is extracted by the authors of MeMo searching the PubMed literature database with the keywords of "methylation lysine" and "methylation arginine" for information on lysine and arginine methylation, respectively.

methylated lysine. The data set of UniProtKB/Swiss-Prot and MeMo are combined and removed the redundant data. The numbers of nonredundant methylated lysine, arginine, glutamate, and asparagine are 460, 303, 45, and 22, respectively.

The combined experimentally verified methylation sites (nonredundant) are defined as the positive data set. However, lysine, arginine, glutamate, and asparagine, which are not annotated as methylated sites in the experimentally validated methylated proteins, are defined as the negative data set. However, the nonredundant positive data set may contain several homologous sites in orthologous proteins. To avoid the overestimation of predictive performance, the nonredundant positive data set were further removed homologous sequences using a window size of $2n + 1$, where n varies from 4 to 10. With reference to the homology reduction of training set in MeMo,[21] as presented in Figure S1 (see Supporting Information), two methylated protein sequences with more than 30% identity were specified to realign the fragment sequences with a window length of $2n + 1$ centered on the modified sites using BL2SEQ. For two fragment sequences with 100% identity, and the methylated sites in the two proteins have the same positions, only one site was kept and the other was discarded. The process for reducing homology was applied to the negative data set.

After the homology had been reduced, nine tenths of the nonhomologous positive datasets, chosen at random, was defined as the positive training set. To prevent skewing the classification of the positive or negative set, the balanced negative training set was extracted from the nonhomologous negative datasets. However, the negative training set, if randomly selected at once, may be not be sufficiently randomly sampled. Therefore, 30 negative training sets are obtained by randomly extracting them from the nonhomologous negative datasets. The average predictive performance obtained using the 30 sets of training data is calculated following cross-validation. A randomly selected tenth of the nonhomologous positive datasets is defined as the positive independent test set. The negative independent test set is also randomly sampled from the nonhomologous negative datasets, which is balanced with the positive independent test set. Sometimes, the trained model can classify the training data effectively, but cannot classify the independent test set effectively, possibly indicating that the trained model is over-fitting to the training data. Therefore, the constructed independent test set not only can be used to evaluate the predictive performance of the trained model, but also can be used to determine whether the trained model is over fitting to the training data. To prevent skewing the sampling of the independent test set, the independent test is performed in 10 times.

Feature Extraction and Coding

This work not only takes the flanking AAs as the training feature, but also takes the solvent ASA and SS that surround the methylated sites into account. The fragment of AAs are extracted from positive and negative training sets using a window of length $2n + 1$ centered on a methylated site. An orthogonal binary coding scheme is adopted to transform AAs into numeric vectors, in the so-called 20-dimensional vector coding. For example, glycine is encoded as "10000000000000000000;" alanine is encoded as "01000000000000000000," and so on. The number of feature vectors that represent the flanking AAs that surround the methylated site is $(2n + 1) \times 20$. Different values of n from 4 to 10 are used to determine the optimized window length. The positional weighted matrix (PWM) of AAs around the methylated sites is determined for four methylated residues using nonhomologous training data. The PWM specifies the relative frequency of AAs in the methylated sites, and is used to encode the fragment sequences.

Since most of the experimental methylated proteins do not have corresponding protein tertiary structures in PDB, an effective tool, RVP-Net,[24,25] was used to compute the ASA value based on protein sequence. The computed ASA value is the percentage of the solvent-accessible area of each AA on the protein sequence. RVP-net applied a neural network to predict real value of ASAs of residues based on neighborhood information, 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.[25] The full-length protein sequences with experimental methylated sites are inputted to RVP-Net to compute the ASA value for all residues. The ASA values of AAs that surround the methylated site were extracted and scaled to zero to one.

In the investigation of SS surrounding the methylated sites, PSIPRED[26] was employed to compute the SS from the protein sequence. PSIPRED is a simple and reliable method for predicting SS, which incorporates two feed-forward neural networks to analyze the output obtained from PSI-BLAST (Position Specific Iterated-BLAST).[27] PSIPRED 2.0 achieved a mean Q_3 score of 80.6% across all 40 submitted target domains without obvious sequence similarity to structures that are present in PDB; accordingly, PSIPRED has been ranked top out of 20 evaluated methods.[28] The output of PSIPRED is given in terms of "H," "E," and "C" which stand for helix, sheet and coil, respectively. The full-length protein sequences with methylated sites are inputted to PSIPRED to determine the SS of all residues, respectively. The orthogonal binary coding scheme is used to transform the three terms that specify the SS into numeric vectors. For instance, helix is encoded as "100," sheet is encoded as "010," and coil is encoded as "001."

Model Training and Evaluation

Three main features, AA, SS and ASA, are used to discriminate between methylated and nonmethylated sites. The SVM is adopted to generate computational models that incorporate the encoded AAs and structural features, SS and ASA. Based on the 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,[29] is employed to train the predictive model with positive and negative training sets which are encoded according to different training features. The radial basis function $K(S_i, S_j) = \exp(-\gamma\|S_i - S_j\|^2)$ is adopted as the kernel function of SVM.

KinasePhos[30] incorporates profile-hidden Markov models (HMMs) to identify kinase-specific phosphorylation sites. It indicates that the HMM can accurately predict phosphorylation sites. Accordingly, HMMER[31] is applied to train the HMMs from the fragments of AAs that surround the methylated sites. An HMM describes a probability distribution over a potentially infinite number of sequences, which can be used to detect distant relationships between AA sequences. The emission and transition probabilities of HMM are generated from the positive training set to capture the characteristics of the methylated sites.

Cross-validation examination is important for practicing the application of the predictor.[32] To evaluate the predictive performance of the trained models, k-fold cross-validation is performed on methylated lysine and arginine. The dataset were divided into k groups by splitting each of their subsets into k approximately equal-sized subgroup. In previous study, Jackknife is the most objective validation method.[32,33] Therefore, Jackknife cross-validation is adapted to methylated glutamate and asparagine for which fewer than 30 data are available. During the jackknife process, both training and testing dataset were actually open, and a protein will in turn move from one dataset to the other.[33] The following measures of predictive performance of the trained models are defined. Precision (Prec) = 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) $= \frac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FN}\times\mathrm{FP})}{\sqrt{(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})}}$, where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively. Since 30 negative training sets are used, the average precision, sensitivity, specificity, accuracy, and MCC are computed for each model that is trained with a particular window length and features. Additionally, the parameters of the predictive models, window length, cost and gamma value of the SVM models, as well as the bit score of the HMM models, are optimized to maximize predictive accuracy. Finally, the window size and features that yield the highest accuracy are utilized to construct prediction models for independent test evaluation.

Evaluating Predictive Models using Independent Test Sets

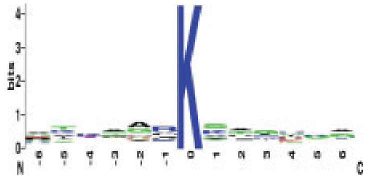
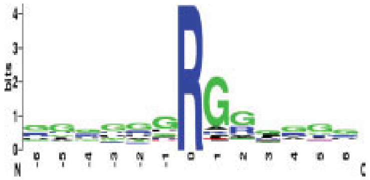
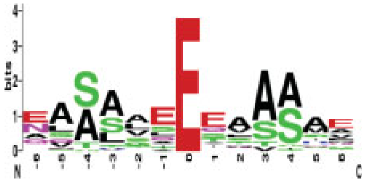
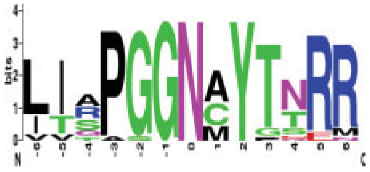
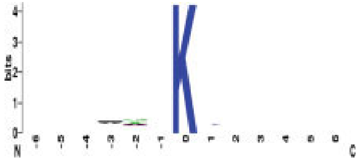
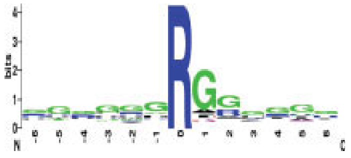
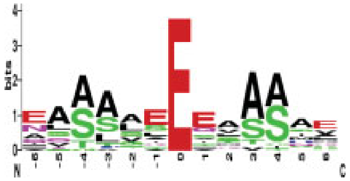
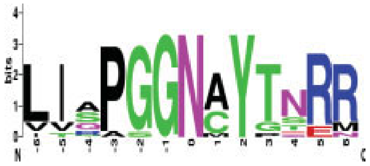
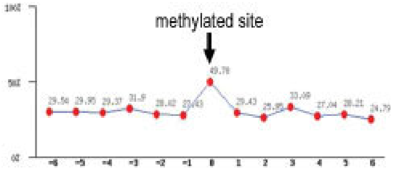
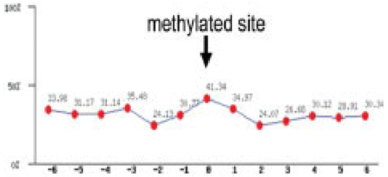
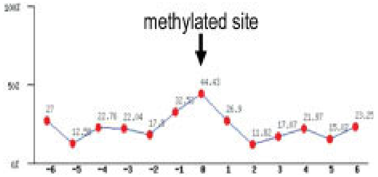
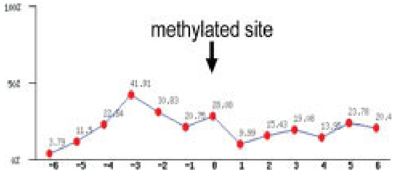
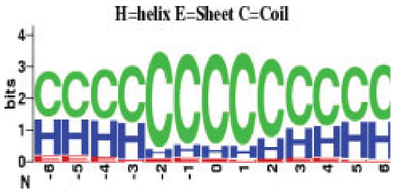
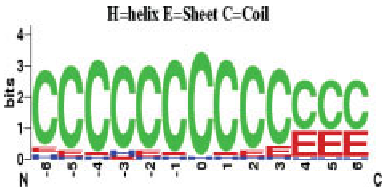
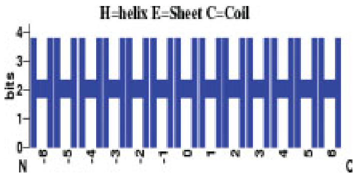
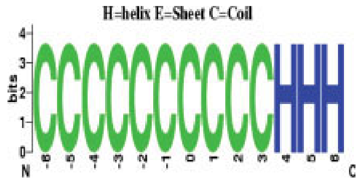
The prediction performance of the trained models may be overestimated because of the over-fitting of a training set. To estimate the real prediction performance, about one-tenth of the nonhomologous data set are randomly selected as the independent test set, which is used to evaluate the predictive performance of the trained models with the best accuracy, based on the cross-validation. Since the number of training sets in methylated glutamate and asparagine is not sufficient, the independent test set is constructed only for lysine and arginine, which are 25 and 30 sites, respectively. However, the performance of the independent test may be favorable just by chance. To avoid the skew sampling of the independent test set, the independent test is executed in 10 times. Therefore, the construction of positive and negative training sets, feature extraction, model training and evaluation and the independent test are performed over 10 rounds. The mean performance of the independent test is computed. The independent test sets of lysine and arginine are employed not only to test the proposed method but also to test other previously proposed protein methylation prediction tools.

Results and Discussion

Sequenced and Structural Features

This work is limited to the analysis of methylated lysine, arginine, glutamate, and asparagine, because of the absence of sufficient experimental verified data for other residues. As shown in Table 2, the flanking AAs of the nonredundant combined methylation sites that are categorized with reference to the modified residues are graphically displayed as a sequence logo, facilitating an investigation of the conservation of AAs around the methylated sites. WebLogo[34,35] is used to create the graphical sequence logo for the relative frequency of the corresponding AA at each position around the methylated sites, using a window −6 to +6 (where position 0 is the methylated site). The sequence logos of the experimental methylated sites which are removed form the homologous sites are also created. In the sequence logo representation, there are no significantly conserved AAs that surround the modified sites are identified. In the

Table 2. The Sequence Logo of Amino Acids, Average Accessible Surface Area, and Sequence Logo of Secondary Structure Surrounding the Experimental Methylation Sites [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Methylated residue	Lysine (K)	Arginine (R)	Glutamate (E)	Asparagine (N)
Number of experimental sites	460	303	45	22
Sequence logo of nonredundant combined experimental sites (−6 to +6)				
Number of nonhomologous experimental sites (−6 to +6)	206	276	27	14
Sequence logo of nonhomologous experimental sites				
Average accessible surface area				
Sequence logo of secondary structure				

In secondary structure, “H,” “E,” and “C” stand for helix, sheet, and coil, respectively.

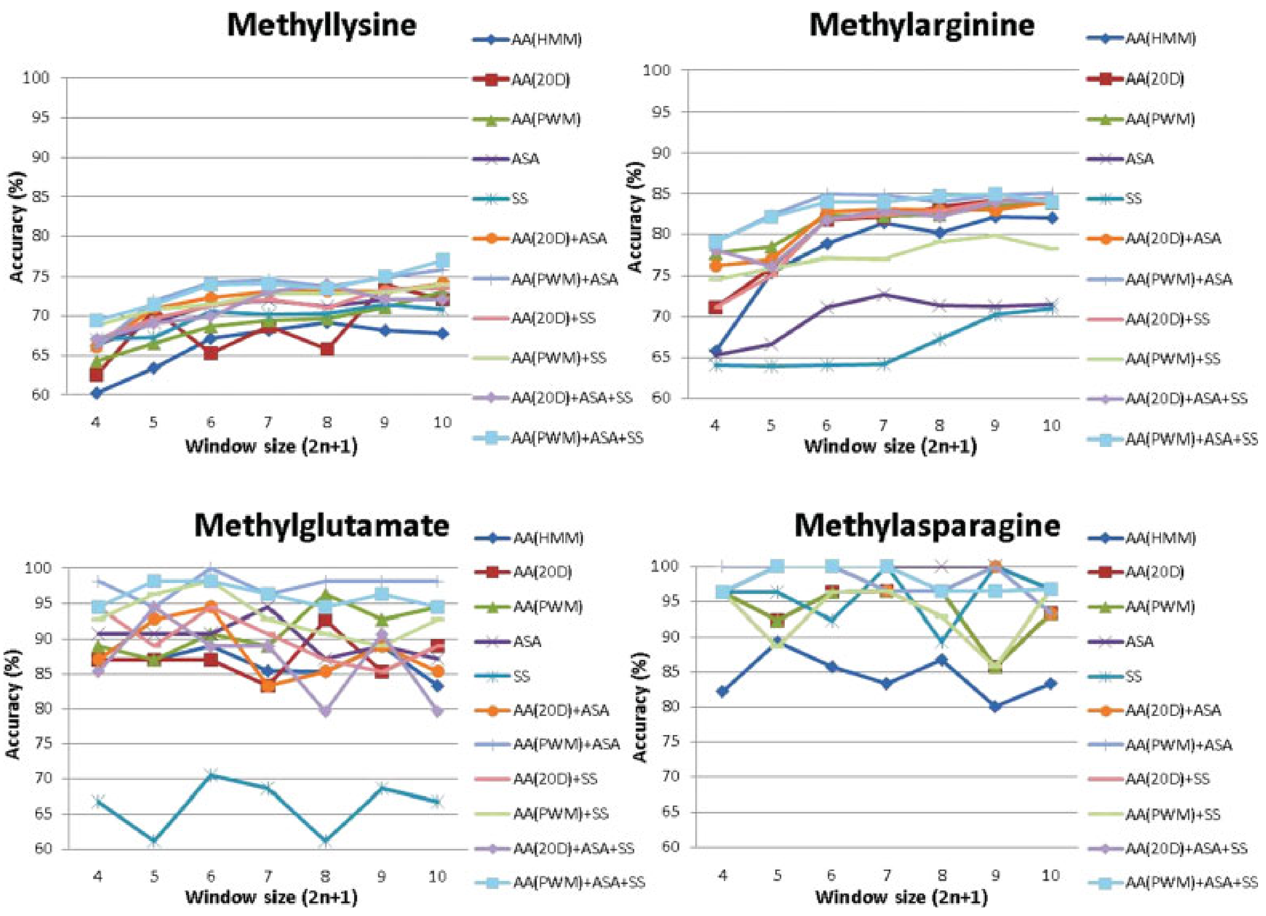


Figure 2. The predictive accuracies of the models trained with different training features based on various window sizes. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

case of methylated arginine, glycine is enriched around the modified sites, especially at positions +1 and +2. In other cases, AAs that surround the methylated glutamate and asparagine are obviously conserved. However, the conservation of AAs in flanking regions may be temporary because of the low abundance of experimentally verified methylglutamate and methylasparagine. Table 2 presents the sequence logo of the SS and the average percentage of ASA in the 13-mer window (−6 to +6). Since the number of experimental methylated sites in the PDB[17] proteins is not enough for training, RVP-Net[24,25] and PSIPRED[26] are adopted to compute the ASA value and SS from the protein sequence, respectively. Observations of SS around the methylated sites show that the methylated lysine, arginine, and asparagine are probably present on the coil (loop), but the methylated glutamate prefers the helix structure. In the study of solvent accessibility, most of the methylated sites are located in the highly ASA except for the sites of methylated asparagines. The average solvent ASA that surrounds the methylated lysine is very similar to those observed in the protein tertiary structure.

Predictive Performance

To study what window lengths and features can be adopted to construct the model that offers the best predictive performance in methylated lysine, arginine, glutamate, and asparagine, models trained with various window lengths and various features are evaluated using cross-validation. Three features—AA, ASA, and SS—are considered. The feature of AAs around the methylated sites is encoded using a 20-dimensional vector and a PWM, named “AA(20D)” and “AA(PWM),” respectively. The features of ASA and SS are encoded using the ASA values and three-dimensional vector, respectively. Figure 2 presents the predictive accuracy of the models that have been using various training features, based on various window sizes $2n + 1$, where n varies from 4 to 10. In particular, the feature of AAs around the methylated sites is also trained by a profile-HMM. Of the models trained with individual features, that trained with ASA values slightly outperforms that trained with AA or SS in methyllysine, whose AAs around the methylated site are not conserved. In methylarginine, the model trained with AA performs much better than the model trained with ASA, and the model trained with SS performs least well. In methylglutamate and methylasparagine, the model trained with AA outperforms the model trained with ASA or SS, because their flanking AAs are conserved. As can be clearly seen, the model trained with SS typically performs worst. For methyllysine, the predictive accuracy increases with window size from 4 to 10. In methylarginine, the model

Table 3. The Average Cross-Validation Performance of the Models Trained with Selected Features and Window Sizes which Achieve the Highest Accuracy.

Methylated Residue	Number of positive training set	Number of negative training set	Training features	Window size	Pr	Sn	Sp	Acc	MCC
Lysine (K)	181	181	AA(PWM) + ASA	−6 to +6	74.1	75.1	74.0	74.6	0.561
Arginine (R)	246	246	AA(PWM) + ASA	−6 to +6	86.6	82.1	87.4	84.8	0.796
Glutamate (E)	27	27	AA(PWM) + ASA	−6 to +6	100.0	100.0	100.0	100.0	1.000
Asparagine (N)	14	14	AA(PWM) + ASA	−6 to +6	100.0	100.0	100.0	100.0	1.000

AA, amino acid; ASA, accessible surface area; PWM, positional weighted matrix; Pr, precision; Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, Matthews Correlation Coefficient.

trained with a window size of six or seven performs best accuracies. The best performances for methylglutamate and methylasparagine are obtained when models trained with a window size of six.

The predictive performance of the model trained with the combination of AA, ASA and SS features is also evaluated. As described previously, the feature of ASA yields an accuracy of over 70% for methylated lysine, arginine, glutamate and asparagine. Therefore, the models trained with a combination of AA and ASA outperform the model trained with AA or ASA alone. However, the predictive accuracy of the model trained with the combination of AA(20D) and ASA is not better than that trained with ASA alone. Since the number of dimensions in AA(20D) encoding method has 20 times than in ASA, the weight of the AA feature exceeds that of ASA features in predicting methylation. Therefore, predictive performance is dominated by the AA feature. The average cross-validation performance of the models that are trained using various window sizes and features that yield the highest accuracy is presented in detail (Table S3 in Supporting Information). The training features that provide the highest accuracy are the combination of AA(PWM) and ASA. In considering the overall performance of the models trained with various window sizes, −6 to +6 are adopted as the feasible window size for the four methylated residues. Table 3 gives the average precision, sensitivity, specificity, accuracy and Matthews Correlation Coefficient of the models trained using the selected features and window sizes. The best predictive accuracies for methyllysine, methylarginine, methylglutamate, and methylasparagine are 74.6, 84.8, 100, and 100%, respectively.

Predictive Performance of Independent Test

Following evaluation by cross-validation, the independent test sets of methyllysine and methylarginine are used to evaluate the selected models with the highest predictive accuracy. To prevent skewed sampling of the independent test set, the independent test is performed in 10 times. Each time of the independent test involves balanced positive and negative sets, which comprise 25 methylated lysines and 30 methylated arginines, with a window length of −7 to +7. The mean predictive accuracies of the proposed method are 80.8 and 85.0% for lysine and arginine, respectively. The mean performance of the independent test is slightly better than that of cross-validation. If the performance of the independent test is much worse than that of cross-validation, then the trained model may be over-fitting for the training data. This independent test demonstrates that the trained model may not over-fit for methylarginine and methyllysine. The independent test sets are also applied to test other methylation predictors.

Table 4 compares the proposed method (MASA) with that of Daily et al.[20] and MeMo.[21] It presents the methods, materials, training features, selected window lengths, cross-validation performance and functions of the web server. The proposed predictive specificities of Daily's predictor and MeMo exceed their sensitivities, whereas the specificity and sensitivity are close in MASA. In predicting methylated lysine, the proposed method slightly outperforms that of Daily et al. and MeMo. Since the training data are not identical among the methods, comparison of cross-validation performance may be unreasonable. However, the predictive performance for methylated lysine is improved using the ASA information in the proposed method, in which users can choose different thresholds for methylation prediction based on predictive sensitivity. The predicted results in tab-delimited format can be downloaded. As given in Table 5, the independent test sets are also used to test other methylation predictors. Since the web site developed by Daily et al.[20] is unavailable, the independent test sets are applied only in MeMo. The independent test demonstrates that MeMo has high predictive specificity in identifying methylated lysine. However, the trained models of MeMo are not sufficiently sensitive to the positive datasets used in independent test sets, especially for lysine.

Evaluation of Methylated Sites with Available 3D Structure

To confirm the quality of the predicted ASA data is accessible for training model. The system is trained only on predicted data and tested on methylated proteins with available 3D structure. Because only four methylated arginines have enough window length of surrounding ASA and SS, this test focuses on methylated lysine (the detailed information is shown in Table S4 of Supporting Information). Forty-one methylated lysines with available 3D structure are used as the independent test set, and the remainder with predicted ASA and SS is adapted to train the SVM model. As shown in Table S5 (see Supporting Information), the balanced positive and negative training sets are applied to construct the SVM model, and 41 positive test data with ex-

Table 4. Comparison of Our Method with Previous Works.

Tools	Methylation Predictor[20]	MeMo[21]	MASA
Material	UniProtKB/Swiss-Prot version 45	UniProtKB/Swiss-Prot version 48 + PubMed literatures	UniProtKB/Swiss-Prot version 53 + MeMo (PubMed literatures)
Method	SVM	SVM	SVM
Training features	Amino acid + intrinsic disorder	Amino acid	Amino acid + accessible surface area
Selected window length ($2n + 1$)	Not specific	15	13
Methylated LYSINE (K)	Sn = 65.9%, Sp = 60.4%	Sn = 69.2%, Sp = 66.7%	Sn = 75.1%, Sp = 74.0%
Methylated arginine (R)	Sn = 73.6%, Sp = 82.2%	Sn = 69.6%, Sp = 89.2%	Sn = 82.1%, Sp = 87.4%
Methylated glutamate (E)	–	–	Sn = 100%, Sp = 100%
Methylated asparagine (N)	–	–	Sn = 100%, Sp = 100%
Web server	Yes (not available)	Yes	Yes
Various prediction threshold	–	–	Selecting different threshold based on predictive sensitivity

SVM, support vector machine; Sn, sensitivity; Sp, specificity.

perimental ASA value and SS are used to evaluate the predictive performance of the constructed model, according to the various training features. This independent test shows that the SVM model trained with AA sequence and predicted ASA can correctly identify 33 positive test data (80.5% sensitivity), which is slightly better than the performance of cross-validation. This test shows that the SVM model trained with predicted ASA or SS could perform effective prediction.

Unbalanced Positive Training Set and Negative Training Set

In this work, positive and negative training sets are balanced during cross-validation. Since the negative dataset is much larger than the positive data, the sampling of negative set may not always be sufficiently random. Accordingly, 30 sets of negative training data are randomly extracted and used to evaluate the prediction performance. In the prediction of protein methylation, the profile-HMM can be constructed using only the positive set, but the SVM model must be constructed from the positive and negative sets, based on binary classification.[29] Based on the construction of the binary SVM classifier with balanced positive and negative sets, the extraction of the negative set may be skewed when the original negative data set is much larger than the extracted negative set. Extracting 30 negative sets to construct 30 predictive models is impossible when the web server is implemented. To prevent the skew sampling of a negative set, a larger negative set should be constructed. Unfortunately, a larger negative set will cause the trained model preferentially to classify negative data correctly, driven by the requirement to maximize accuracy. As displayed in Figure 3, the predictive specificity of the methyllysine model, which is trained using different ratios of positive and negative sets, increases with the relative size of the negative set. When both the sensitivity and the size of the negative set are taken into account, the optimal ratio of the data sizes of positive to negative sets is 1:5 (the detailed information is shown in Table S6 of Supporting Information). Consequently, this suitable ratio of positive to negative sets is used to construct the prediction model used in the protein methylation web server.

Alternative Methylated and Acetylated Lysine

Histone acetylation and methylation are the two major modifications that regulate specific transcription in response to various cellular signals. Their combinatorial effects in transcriptional control are particularly important.[36] Although the mechanism of action of these modifications in transcription is not well understood, recent discoveries of histone acetyltransferase and methyltransferase activity in transcriptional regulators have important implications for histone modification as key to the precise regulation of transcription processes.[36] However, specific lysine residues in H3 histone protein tails appear to be targeted for either methylation or acetylation.[37] As shown in Table S7 (see

Table 5. The Average Performances of Our Method and MeMo Evaluated by Independent Test.

Tools	Methylated residue	Number of positive test set	Number of negative test set	Pr	Sn	Sp	Acc
MASA	Lysine (K)	25	25	82.7	78.8	82.8	80.8
	Arginine (R)	30	30	86.5	82.9	87.2	85.0
MeMo[21]	Lysine (K)	25	25	93.0	74.2	94.2	84.2
	Arginine (R)	30	30	78.8	77.6	79.1	78.3

Pr, precision; Sn, sensitivity; Sp, specificity; Acc, accuracy.

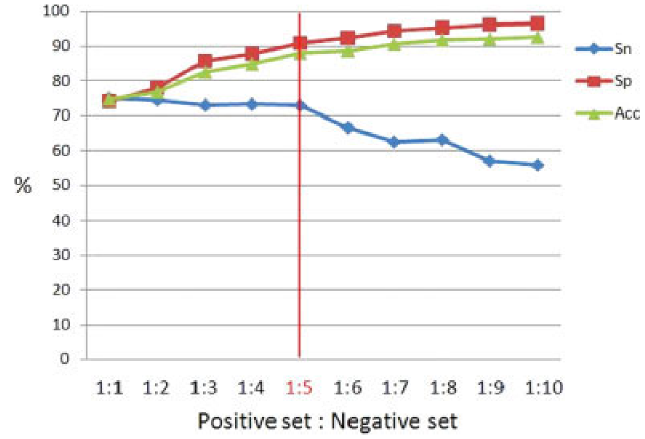
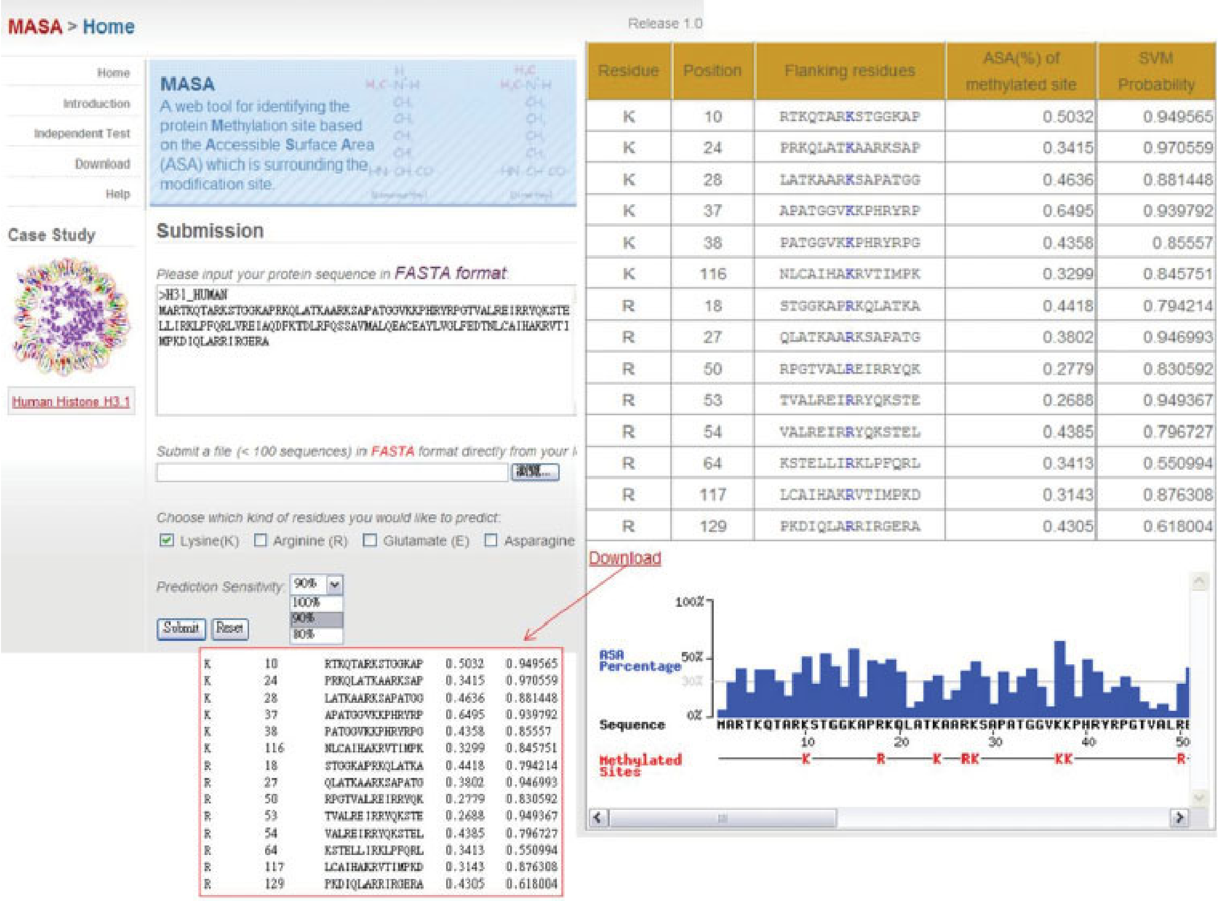


Figure 3. The cross-validation sensitivity, specificity, and accuracy of the methyllysine model trained with different ratio of positive and negative sets.

Supporting Information), 92 lysine residues, which are alternative methylation and acetylation sites, were extracted from UniProtKB/Swiss-Prot release 53. For example, Lys5, Lys10, Lys19, Lys24, Lys28, Lys37, Lys57, and Lys80 of human H3 histone protein (UniProtKB/Swiss-Prot ID: H31_HUMAN) are alternative methylated or acetylated sites. Hence, the classifying ability of the trained methyllysine model in distinguishing methyllysine from acetyllysine was tested. The homologous sites are removed, based on a window size of −6 to +6, from a total of 792 experimentally verified acetylated lysines, extracted from UniProtKB/Swiss-Prot release 53. The 459 non-homologous acetylated sites are inputted to the methyllysine model that was trained with AA and ASA, and 143 acetylated sites (∼31%) were predicted as methylation sites. This result indicates that the constructed methyllysine model cannot effectively differentiate acetyllysine from methyllysine, probably because of the alternative methylated and acetylated sites. To test the distinguishing between methylated and acetylated lysines, the experimental methylated lysine and acetylated lysine are defined as positive and negative sets, respectively, and are adopted to train a SVM model with AA and ASA features based on a binary classifier. According to evaluation using five-fold cross-validation, the corrected classification between methyllysine and acetyllysine is 78.2% accuracy, revealing that the methylated and acetylated lysines could be distinguished by the model trained with AA and ASA features. The incorrect classifications were mostly made by the alternative methylated and acetylated lysines.


MASA > Home
Release 1.0
Home
Introduction
Independent Test
Download
Help
MASA
A web tool for identifying the protein Methylation site based on the Accessible Surface Area (ASA) which is surrounding the modification site.
Case Study
Human Histone H3.1
Submission
Download
ASA Percentage
100%
50%
30%
0%
Sequence
MARTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPGTVALRE
10
20
30
40
50
Methylated Sites
K
R
K
RK
KK
R

Please input your protein sequence in FASTA format

```
>H31_HUMAN
MARTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPGTVALREIRRYQKSTE
LLIRKLPFQRLVREIAQDFKTDLRFQSSAVMALQEACEAYLVGLFEDTNLCAIHAKRVTI
MPKDIQLARRIRGERA
```

Submit a file (< 100 sequences) in FASTA format directly from your

Choose which kind of residues you would like to predict:
☑ Lysine(K) □ Arginine (R) □ Glutamate (E) □ Asparagine

Prediction Sensitivity: 90% (100% / 90% / 80%)

Submit Reset

Residue	Position	Flanking residues	ASA(%) of methylated site	SVM Probability
K	10	RTKQTARKSTGGKAP	0.5032	0.949565
K	24	PRKQLATKAARKSAP	0.3415	0.970559
K	28	LATKAARKSAPATGG	0.4636	0.881448
K	37	APATGGVKKPHRYRP	0.6495	0.939792
K	38	PATGGVKKPHRYRPG	0.4358	0.85557
K	116	NLCAIHAKRVTIMPK	0.3299	0.845751
R	18	STGGKAPRKQLATKA	0.4418	0.794214
R	27	QLATKAARKSAPATG	0.3802	0.946993
R	50	RPGTVALREIRRYQK	0.2779	0.830592
R	53	TVALREIRRYQKSTE	0.2688	0.949367
R	54	VALREIRRYQKSTEL	0.4385	0.796727
R	64	KSTELLIRKLPFQRL	0.3413	0.550994
R	117	LCAIHAKRVTIMPKD	0.3143	0.876308
R	129	PKDIQLARRIRGERA	0.4305	0.618004

```
K    10    RTKQTARKSTGGKAP    0.5032    0.949565
K    24    PRKQLATKAARKSAP    0.3415    0.970559
K    28    LATKAARKSAPATGG    0.4636    0.881448
K    37    APATGGVKKPHRYRP    0.6495    0.939792
K    38    PATGGVKKPHRYRPG    0.4358    0.85557
K    116   NLCAIHAKRVTIMPK    0.3299    0.845751
R    18    STGGKAPRKQLATKA    0.4418    0.794214
R    27    QLATKAARKSAPATG    0.3802    0.946993
R    50    RPGTVALREIRRYQK    0.2779    0.830592
R    53    TVALREIRRYQKSTE    0.2688    0.949367
R    54    VALREIRRYQKSTEL    0.4385    0.796727
R    64    KSTELLIRKLPFQRL    0.3413    0.550994
R    117   LCAIHAKRVTIMPKD    0.3143    0.876308
R    129   PKDIQLARRIRGERA    0.4305    0.618004
```

Figure 4. Web interface of MASA.

Web Interface

With the time-consuming and lab-intensive experimental identification of protein methylation sites, a biologist may understand only that a protein can be methylated: the precise identification of the methylated sites on the substrate remains unknown. Therefore, an effective prediction server can help to focus efficiently on potential sites. After evaluation by cross-validation and the independent test, the combination of ASA, AA and window size −6 to +6 is employed to construct the models for predicting methylated lysine, arginine, glutamate, and asparagine. Based on a binary SVM classifier with balanced positive and negative sets, the negative set may be skew-sampled when the original negative data set is much larger than the extracted negative set. To avoid the skew sampling of the negative set, a negative set whose size is five times that of the positive set is randomly selected as a target to implement the methylation prediction sever based on AA and ASA. As displayed in Figure 4, 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 methylated position, flanking AAs, and ASA values, which are predicted by RVP-Net. Additionally, users can choose various thresholds for predicting methylation based on predictive sensitivity. The overall identified methylation sites and ASAs of AAs can be graphically presented. Users can download the predicted results in tab-delimited format for further analysis, and the independent test sets can also be downloaded from the proposed web site.

Conclusions

This work presents a method that combines the SVM with AAs and solvent ASA to identify protein methylation sites. To prevent any overestimation of predictive performance, the homologous sequences were removed using a given window size from the collected data sets. Although the ASA value is predicted using RVP-Net,[24,25] the cross-validation results demonstrated that the integration of the ASA value around the methylation sites can improve the prediction performance of protein methylation sites, especially for lysine. The independent test also shows that the proposed method performs very well in differentiating methylated sites from unmethylated sites. In this effective prediction system, the selected window size −6 to +6 provides the best overall accuracy, and is used to implement the methylation prediction server based on AA and solvent accessibility. Although the training data for methylglutamate and methylasparagine are very few, the trained models may help biologists efficiently to discover the novel methylation sites. The overall predicted methylation sites and ASA values can be graphically presented. Furthermore, users can download the predicted results with tab-delimited format for further analysis.

Although the proposed method can perform accurately and robustly according to independent tests, some issues should still be addressed in future work. First, the structural preferences of methylated sites should be investigated in greater detail—especially in methylated lysine whose flanking residues are not conserved. In addition to the solvent ASA and SS, the B-factor, intrinsic disordered region, protein linker region, and other factors should be examined at experimental methylation sites which are located in the protein regions with PDB entries. With reference to another study of phosphorylation,[38] the local 3D structure of methylated sites may be extracted for further analysis. Second, the independent test sets proposed herein are really blind to the trained model during cross-validation, but may be not to other previously proposed predictors. Hence, a benchmark for constructing test sets that are really independent of each predictor is important. Finally, the trained methyllysine model, in this work, cannot distinguish the methylation site from the acetylation site effectively, because the methyllysine and acetyllysine alternate in several locations of protein. Methyllysine and acetyllysine should be investigated in detail—not only with reference to AAs and ASA.

Availability

The web server of MASA will be continuously maintained and updated. The web server is now freely available at http://MASA.mbc.nctu.edu.tw/.

Acknowledgments

The authors thank the developers of MeMo[21] for providing the manually curated data from their survey of the relevant literature. Furthermore, they sincerely appreciate the National Science Council of the Republic of China for financially supporting this research under Contract Numbers of NSC 96-3112-E-009-002, NSC 97-2811-B-009-001 and NSC 97-2627-B-009-007. Special thanks for financial support go to the National Research Program for Genomic Medicine (NRPGM), Taiwan. They also thank Ted Knoy for his editorial assistance.

References

1. Paik, W. K.; Kim, S. Biochem Biophys Res Commun 1967, 29, 14.
2. Beausoleil, S. A.; Jedrychowski, M.; Schwartz, D.; Elias, J. E.; Villen, J.; Li, J.; Cohn, M. A.; Cantley, L. C.; Gygi, S. P. Proc Natl Acad Sci USA 2004, 101, 12130.
3. Sayegh, J.; Webb, K.; Cheng, D.; Bedford, M. T.; Clarke, S. G. J Biol Chem 2007, 282, 36444.
4. Martin, C.; Zhang, Y. Nat Rev Mol Cell Biol 2005, 6, 838.
5. Bannister, A. J.; Kouzarides, T. Nature 2005, 436, 1103.
6. Lee, D. Y.; Teyssier, C.; Strahl, B. D.; Stallcup, M. R. Endocr Rev 2005, 26, 147.
7. Paik, W. K.; DiMaria, P. Methods Enzymol 1984, 106, 274.
8. Predel, R.; Brandt, W.; Kellner, R.; Rapus, J.; Nachman, R. J.; Gade, G. Eur J Biochem 1999, 263, 552.
9. Lapko, V. N.; Cerny, R. L.; Smith, D. L.; Smith, J. B. Protein Sci 2005, 14, 45.
10. Aletta, J. M.; Cimato, T. R.; Ettinger, M. J. Trends Biochem Sci 1998, 23, 89.
11. Bedford, M. T.; Richard, S. Mol Cell 2005, 18, 263.
12. Stallcup, M. R. Oncogene 2001, 20, 3014.
13. Murray, K. Biochemistry 1964, 3, 10.
14. Paik, W. K.; Kim, S. Yonsei Med J 1986, 27, 159.

15. Kouskouti, A.; Scheer, E.; Staub, A.; Tora, L.; Talianidis, I. Mol Cell 2004, 14, 175.
16. Pang, C. N.; Hayen, A.; Wilkins, M. R. J Proteome Res 2007, 6, 1833.
17. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucleic Acids Res 2000, 28, 235.
18. Kabsch, W.; Sander, C. Biopolymers 1983, 22, 2577.
19. Farriol-Mathis, N.; Garavelli, J. S.; Boeckmann, B.; Duvaud, S.; Gasteiger, E.; Gateau, A.; Veuthey, A. L.; Bairoch, A. Proteomics 2004, 4, 1537–1550.
20. Daily, K. M.; Radivojac, P.; Dunker A. K. In IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, November 2005; San Diego, CA, pp. 475–481.
21. Chen, H.; Xue, Y.; Huang, N.; Yao, X.; Sun, Z. Nucleic Acids Res 2006, 34 (Web Server issue), W249.
22. 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.
23. Arthur, J. W.; Sanchez-Perez, A.; Cook, D. I. Bioinformatics 2006, 22, 2192.
24. Ahmad, S.; Gromiha, M. M.; Sarai, A. Bioinformatics 2003, 19, 1849.
25. Ahmad, S.; Gromiha, M. M.; Sarai, A. Proteins 2003, 50, 629.
26. McGuffin, L. J.; Bryson, K.; Jones, D. T. Bioinformatics 2000, 16, 404.
27. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Nucleic Acids Res 1997, 25, 3389.
28. Bryson, K.; McGuffin, L. J.; Marsden, R. L.; Ward, J. J.; Sodhi, J. S.; Jones, D. T. Nucleic Acids Res 2005, 33 (Web Server issue), W36.
29. Chang, C.-C.; Lin, C.-J. Available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001, October 2007.
30. Huang, H. D.; Lee, T. Y.; Tzeng, S. W.; Wu, L. C.; Horng, J. T.; Tsou, A. P.; Huang, K. T. J Comput Chem 2005, 26, 1032.
31. Eddy, S. R. Bioinformatics 1998, 14, 755.
32. Chou, K. C.; Shen, H. B. Anal Biochem 2007, 370, 1.
33. Chou, K. C.; Zhang, C. T. Crit Rev Biochem Mol Biol 1995, 30, 275.
34. Crooks, G. E.; Hon, G.; Chandonia, J. M.; Brenner, S. E. Genome Res 2004, 14, 1188.
35. Schneider, T. D.; Stephens, R. M. Nucleic Acids Res 1990, 18, 6097.
36. An, W. Subcell Biochem 2007, 41, 351.
37. Rice, J. C.; Allis, C. D. Curr Opin Cell Biol 2001, 13, 263.
38. Zanzoni, A.; Ausiello, G.; Via, A.; Gherardini, P. F.; Helmer-Citterich, M. Nucleic Acids Res 2007, 35 (Database issue), D229.