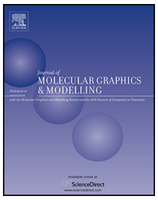
Journal of Molecular Graphics and Modelling 56 (2015) 84–90

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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



Using support vector machines to identify protein phosphorylation sites in viruses

Shu-Yun Huang[a], Shao-Ping Shi[b,c], Jian-Ding Qiu[a,b,*], Ming-Chu Liu[a,*]

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

ARTICLE INFO

Article history:
Accepted 16 December 2014
Available online 24 December 2014

Keywords:
Phosphorylation site
Virus proteins
Support vector machine
Encoding scheme based on attribute grouping
Position weight amino acid composition

ABSTRACT

Phosphorylation of viral proteins plays important roles in enhancing replication and inhibition of normal host-cell functions. Given its importance in biology, a unique opportunity has arisen to identify viral protein phosphorylation sites. However, experimental methods for identifying phosphorylation sites are resource intensive. Hence, there is significant interest in developing computational methods for reliable prediction of viral phosphorylation sites from amino acid sequences. In this study, a new method based on support vector machine is proposed to identify protein phosphorylation sites in viruses. We apply an encoding scheme based on attribute grouping and position weight amino acid composition to extract physicochemical properties and sequence information of viral proteins around phosphorylation sites. By 10-fold cross-validation, the prediction accuracies for phosphoserine, phosphothreonine and phosphotyrosine with window size of 23 are 88.8%, 95.2% and 97.1%, respectively. Furthermore, compared with the existing methods of Musite and MDD-clustered HMMs, the high sensitivity and accuracy of our presented method demonstrate the predictive effectiveness of the identified phosphorylation sites for viral proteins.

1. Introduction

Protein phosphorylation is a ubiquitous post-translational modification (PTM) that controls a number of intracellular processes. It has been estimated that at least one-third of the cellular proteins are modified by phosphorylation [1]. In eukaryotic cells, phosphorylation occurs almost exclusively on serine, threonine or tyrosine residues [2]. Also for viruses, including vesicular stomatitis virus, human immunodeficiency virus type 1 (HIV-1), mosaic virus, and H1N1 influenza virus, protein phosphorylation has been shown to regulate vital processes such as virus transcription and replication, RNA binding activity, and virus assembly [3–7]. For instance, Polo-like kinase 1 (Plk1) can phosphorylate cyclin T1 at Ser564 and inhibit the kinase activity of cyclin T1/Cdk9 complex on phosphorylation of the C-terminal domain (CTD) of RNA polymerase II [8]. Hsiang et al. demonstrated that the only serine 42 (S42) phosphorylation of the NS1 protein catalyzed by protein kinase Cα(PKCα) regulated human influenza A virus replication [9]. Cheng et al. identified membrane-associated serine/threonine kinase-like protein from Nicotiana benthamiana involved in the cell-to-cell movement of Bamboo mosaic virus (BaMV) [10]. Therefore, investigating virus phosphorylation sites can provide useful clues for drug design and the treatment of various viral infections.

Phosphorylation site identification is usually experimentally determined by mass spectrometry-based techniques [11]. This has led to the establishment of several databases of phosphorylation sites, such as 'the Phosphorylation Site Database' [12], 'Phospho.ELM' [13], 'Phosphosite' [14], and 'PhosPhAT' [15]. While useful, mass spectrometry requires very expensive instruments and specialized expertise that are not available in typical laboratories [16]. At the same time, the identification of kinase specificity rules with mass spectrometry still remains a relatively slow and often inefficacious task. Thus, various computational methods for identifying protein phosphorylation sites have been proposed, including artificial neural networks (ANNs) [17,18], hidden Markov models (HMMs) [19,20], position-specific scoring matrices (PSSMs) [21–23], support vector machines (SVMs) [24–27], and more details can be found in recent reviews [28,29].

In virus phosphorylation prediction, Schwartz and Church used the scan-*x* tool to identify 329 phosphorylation sites in proteins from 52 human viruses [30]. However, it has not investigated the

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

http://dx.doi.org/10.1016/j.jmgm.2014.12.005
1093-3263/

various substrate motifs for viral protein phosphorylation sites [31]. More recently, Bretaña et al. employed maximal dependence decomposition (MDD) to investigate kinase substrate specificities in viral protein phosphorylation sites [31]. Although, the average accuracies of serine and threonine using the MDD-clustered HMMs were 84.93% and 78.05%, respectively, the number of phosphorylated serine sites was only 233, and 54 for phosphothreonine sites. As we all know, a small number of training set may be over-fitting. Hence, there is a need to develop a computational method in identifying enormous amount of viral protein phosphorylation data by selecting more informative feature descriptors.

In this paper, we presented a new approach to predict viral phosphorylation sites based on support vector machine. Physicochemical properties of amino acids and position weight amino acid compositions were utilized to extract sequence features of virus proteins. Our current work contained the following contents: (1) two types of features were analyzed, (2) SVM was employed to deal with the problem of binary classification, (3) ten-fold cross-validation method was chosen to evaluate the performance of SVM classifier, (4) the effect of window length was investigated, and (5) the independent testing data was used to compare with the existing models.

2. Materials and methods

2.1. Data collection and statistics

All training datasets were extracted from the NCBI RefSeq and the Plant Protein Phosphorylation Database (P^3DB) [32] databases, as presented in Fig. 1. Firstly, we obtained 327 proteins with 2793 experimental phosphorylation sites by searching information containing “phosphorylation” and “virus” from the NCBI RefSeq. The P^3DB is one of the most significant in vivo data resources for studying plant phosphoproteomics. According to the keyword of virus, we obtained 363 proteins covering 1274 experimental phosphorylation sites from the P^3DB. Secondly, the sliding window strategy was used to extract positive and negative datasets from protein sequences, which were represented by peptide sequences with serine, threonine and tyrosine symmetrically surrounded by flanking residues. If the candidate phosphorylation sites were near the N- or C-terminus, we used the letter “O” instead of the absent letters. We respectively designated peptide sequences of experimentally validated phosphoserine, phosphothreonine and phosphotyrosine as positive datasets. It would be difficult to prove definitively that a particular serine/threonine/tyrosine residue is not phosphorylated under any conditions. Almost all of researchers of phosphorylation prediction made the assumption that any serine/threonine/tyrosine residue that is not marked by any phosphorylation information on the same protein is a non-phosphorylated site [25,31,33]. Besides, Radivojac et al. have concluded that the choosing of negative samples upon this assumption did not significantly influence prediction performance through comparing with that of using the validated negative samples [34]. So we adopt this assumption that negative samples were the serine/threonine/tyrosine residues that were not marked by any phosphorylation information on the same proteins, the rational of which is that the resulting negative samples are more likely to be non-phosphorylation sites than those obtained by random as these proteins were experimentally investigated. Although not all these sites are necessarily true negatives, it is reasonable to believe that a large majority of them are [35]. Moreover, the redundancy reducing process was also carried out on training datasets. For example, for two phosphorylated serine peptide sequences with 100% identity, when the phosphoserine sites in the two proteins were in the same positions, only one was kept. After strictly following the above procedures, we attained 2444 high quality positive sites for phosphoserine, 635 positive sites for phosphothreonine, and 268 positive sites for phosphotyrosine, as shown in Supplementary materials (see Tables S1–S3).

Meanwhile, in order to further evaluate the performance of our method and compare it with existing methods, an independent testing set was extracted from the viral posttranslational modification (virPTM) database (http://virptm.hms.harvard.edu/), which includes 230 phosphoserine sites and 2494 non-phosphorylated serine sites, 61 phosphothreonine sites and 1211 non-phosphorylated threonine sites, 14 phosphotyrosine sites and 57 non-phosphorylated tyrosine sites from 111 human virus proteins shown in Fig. 1. Finally, the ratio of positive and negative samples was 1:1 and three negative training sets were obtained by randomly extracting from the negative datasets, with expectation to ensure unbiased and objective results.

2.2. Feature encoding

2.2.1. Encoding based on attribute grouping

Previously, Fan and Zhang have detected that the serine and threonine acceptor site microenvironment is depleted in nonpolar and hydrophobic amino acids. Whereas the tyrosine acceptor site microenvironment is characterized by only one enriched property, namely the charge, and is depleted in cysteine (C) and proline (P), which are neutral residues [36].

Thus, we adopted an encoding scheme of protein sequences considering the hydrophobicity and charged character of amino acid residues. The encoding method based on attribute grouping (named as EBAG) divides the 20 amino acid residues into four different classes according to their physicochemical property: the hydrophobic group C1 = [A, F,G, I, L, M, P, V, W], the polar group C2 = [C, N, Q, S, T, Y], the acidic group C3 = [D, E], and the basic group C4 = [H, K, R] [37,38].

Given a protein sequence p fragment with $2L+1$ amino acid residues, we used the above classification to transform it into four binary sequences as follows:

$$\begin{array}{llll} \mathrm{H1_p}(j) = 1 & \text{if } \mathrm{p}(j) \in \mathrm{C1} & \text{else } \mathrm{H1_p}(j) = 0 & \\ \mathrm{H2_p}(j) = 1 & \text{if } \mathrm{p}(j) \in \mathrm{C2} & \text{else } \mathrm{H2_p}(j) = 0 & \\ \mathrm{H3_p}(j) = 1 & \text{if } \mathrm{p}(j) \in \mathrm{C3} & \text{else } \mathrm{H3_p}(j) = 0 & j = -L, \ldots, L \\ \mathrm{H4_p}(j) = 1 & \text{if } \mathrm{p}(j) \in \mathrm{C4} & \text{else } \mathrm{H4_p}(j) = 0 & \end{array} \tag{1}$$

2.2.2. Position weight amino acid composition

To reveal the sequence-order information around phosphorylation sites, we used position weight amino acids composition (PWAA) to extract the sequence position information of amino acid residues. Given an amino acid residue a_i ($i = 1, 2, \ldots, 20$), we can express the position information of amino acid a_i in the protein sequence fragment p with $2L+1$ amino acids by following formula:

$$C_i = \frac{1}{L(L+1)} \sum_{j=-L}^{L} x_{i,j} \left(j + \frac{|j|}{L} \right), \quad j = -L, \ldots, L \tag{2}$$

where L denotes the number of upstream residues or downstream residues from the central site in the protein sequence fragment p, $x_{i,j} = 1$ if a_i is the jth position residue in protein sequence fragment p, otherwise $x_{i,j} = 0$. Finally, a protein sequence fragment p is defined as 20 dimension feature vectors.

2.3. Model learning and evaluation

SVM is a supervised learning method for classification and regression designed by Cortes and Vapnik [39]. The principle of

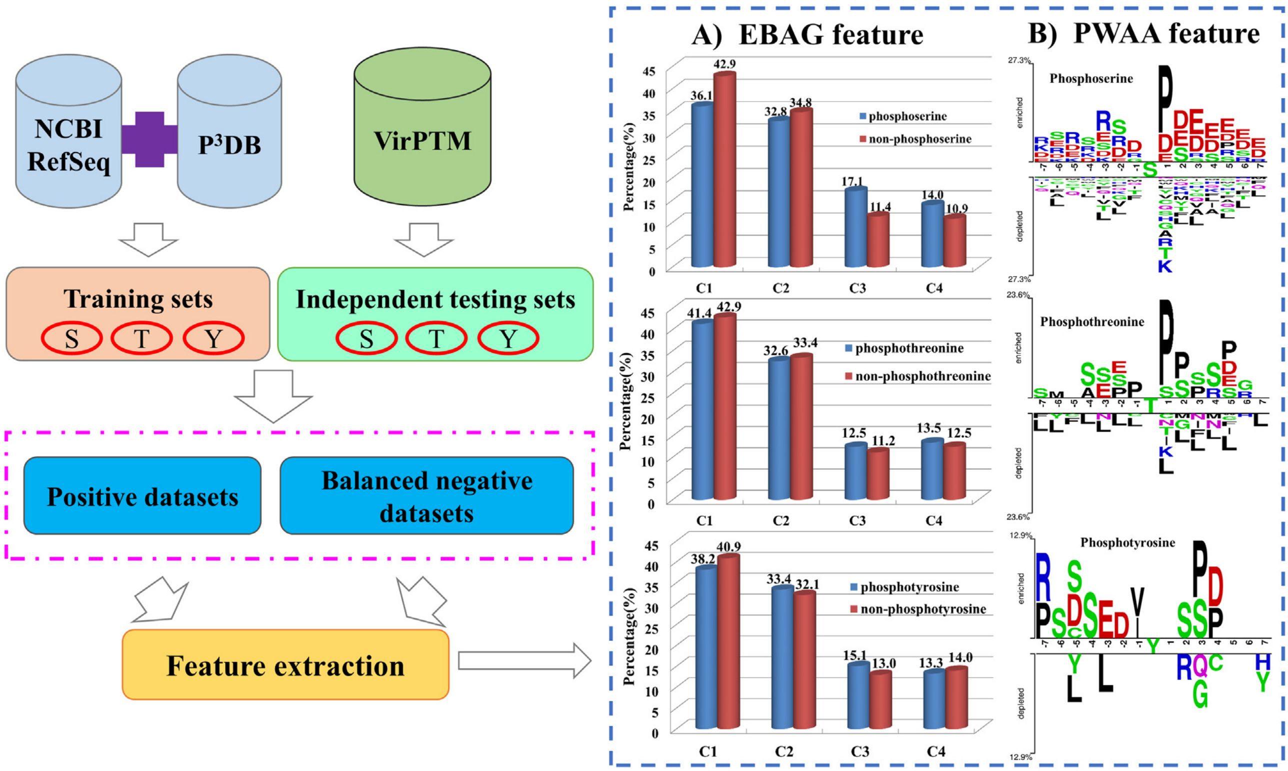


Fig. 1. The conceptual diagram of constructing the data collection and feature extraction. (A) Distribution of four types of amino acids in the EBAG feature associated with phosphorylated vs. non-phosphorylated residues surrounding 15-mer virus phosphorylation sites. C1 is hydrophobic group. C2 is polar group. C3 is acidic group. C4 is basic group. (B) Two Sample Logos of the compositional biases between phosphorylation and non-phosphorylation sites. Only amino acid residues significantly enriched or depleted (P-value <0.05; t-test) around 15-mer phosphorylation sites are shown. The position of phosphorylation sites is 0.

the SVM method is to transform the samples into a high dimension Hilbert space and seek an optimal separating hyperplane which maximizes the margin in feature space. SVM has shown successful ability to capture complex patterns without over-fitting issues, thus it is considered as one of the most popular tools for phosphorylation prediction [28]. For actual implementation, we used the LIBSVM package (version 3.1) [40]. To obtain an SVM classifier with optimal performance, radial basis function (RBF) was tested in our research, and the penalty parameter C and kernel parameter g were tuned based on the training set using the grid search strategy in LIBSVM.

Ten-fold cross-validation was applied to evaluate the powers of the prediction method. Precision (Pr), Sensitivity (Sn), specificity (Sp), accuracy (Acc) and Matthew's Correlation Coefficient (MCC) were utilized to assess the performance of prediction system. All of the above measurements were defined as follows:

$$\mathrm{Pr} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{3}$$

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{4}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{5}$$

$$\mathrm{Acc} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}} \tag{6}$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{7}$$

where TP, TN, FP and FN are the number of true positives, true negatives, false positives and false negatives, respectively. Sensitivity and specificity illustrate the correct prediction ratios of positive and negative data sets, while accuracy represents the correct ratio among both positive and negative data sets. The MCC considers both the TP and the TN as successful predictions, and it is usually regarded as a balanced measure that can be used even if the classes are of very different sizes. For these reasons, the MCC is more reliable than the Acc. The value of MCC ranges from −1 to 1, with larger values standing for better predictive performance. Next, the true positive rate (i.e. Sn) and the false positive rate (i.e. 1 − Sp) are calculated to draw the receiver operating characteristic (ROC) curves, and we use the area under the curves (AUC) to quantify the predictive quality.

3. Results and discussion

3.1. Investigation of different features

3.1.1. EBAG feature analysis

As described in Section 2, the EBAG feature is mainly based on the hydrophobicity and charged character of amino acids. So we calculated statistically significant differences in the distribution of four types of residues surrounding the 15-mer phosphorylation sites to consider whether physiochemical properties had an influence on viral phosphorylation determination. Fig. 1A showed that the percentage of hydrophobic (C1) residues for phosphoserine sites reached 36.1%, which were lower than that for non-phosphorylated serine sites. On the contrary, the percentage of acidic residues (C3) for phosphoserine sites was 5.7% higher than that for non-phosphorylated serine sites. There were no significant difference in the percentage of four types of amino acids between phosphothreonine and non-phosphorylated threonine sites in Fig. 1A. For tyrosine phosphorylation, the percentage

Table 1
Prediction performance of models trained with different features in window size 15.

Residue	Feature	Performance (%)					
		Pr	Sn	Sp	Acc	MCC	AUC
Serine	EBAG	73.5	70.5	74.6	72.6	45.2	71.7
	PWAA	79.5	73.9	80.9	77.4	55.0	76.8
	EBAG + PWAA	85.2	82.9	85.6	84.2	68.5	79.5
Threonine	EBAG	66.3	71.6	63.6	67.6	35.3	68.9
	PWAA	76.7	69.6	78.9	74.2	48.7	74.6
	EBAG + PWAA	80.2	84.5	79.1	81.8	63.7	81.7
Tyrosine	EBAG	73.1	72.4	73.3	72.9	45.6	74.1
	PWAA	70.4	69.8	70.7	70.2	40.5	67.3
	EBAG + PWAA	84.5	84.9	84.4	84.7	69.3	83.6

of hydrophobic and acidic residues had significantly different between phosphotyrosine and non-phosphorylated tyrosine sites. As seen from Fig. 1A, the hydrophobic residues percentage for serine, threonine and tyrosine phosphorylation sites were lower than those for non-phosphorylation sites, whereas the acidic residues percentage were higher than those for non-phosphorylation sites. The results showed that the acidic residues contributed to the occurrence of phosphorylation sites in viruses. The accuracies of serine, threonine and tyrosine phosphorylation sites based on EBAG feature reached 72.6%, 67.6% and 72.9%, respectively, as shown in Table 1. The predictive performance of phosphothreonine sites obtained with EBAG feature was lower than that of phosphoserine and phosphotyrosine sites, which was in agreement with the above analysis. Unfortunately, the MCC of serine, threonine and tyrosine phosphorylation sites were 45.2%, 35.3% and 45.6%, respectively, indicating that just using the EBAG feature could not effectively predict the phosphorylation sites of viral proteins.

3.1.2. PWAA feature analysis

PWAA feature reflects the position information of residues surrounding phosphorylation sites. To analyze position specific properties of viral proteins, we adopted a web-based tool Two Sample Logo [41] to present the compositional biases between phosphorylation and non-phosphorylation sites. As presented in Fig. 1B, there were some significant differences between serine, threonine, tyrosine phosphorylation and their non-phosphorylation. The most pronounced feature of phosphoserine and phosphothreonine sites were proline (P) at position +1. Proline-directed kinases (PDK) have broad specificities and recognize the motif xSPx or xTPx, which are consistent with reports in the literature [25,42,43]. Arginine (R) was enriched in upstream residues of serine sites. On the contrary, aspartic acid (D) and glutamic (E) were enriched in downstream residues of serine sites in Fig. 1B. Studies show that the search pattern is xRRxSx in scanning sequences for cAMP-dependent protein kinase sites. Casein kinase II has a very widespread distribution and recognizes the motifs xSxxEx or xSxxDx [44]. Fig. 1B showed that proline was enriched around threonine phosphorylation sites. Schwartz and coworkers found three types of proline motifs surrounding phosphothreonine sites xTPx, PxTPx and xTPPx [45]. For tyrosine phosphorylation sites, aspartic acid (D) and glutamic (E) were the dominant amino acids in the flanking of phosphotyrosine sites, as presented in Fig. 1B. The above analysis suggested that there were distinct kinase-specific and residue-conservative differences for serine, threonine and tyrosine phosphorylation sites of human virus proteins. The predictive performance based on the PWAA feature had a little improvement compared with the EBAG feature, but the results could not meet the prediction goals.

3.1.3. Optimal feature set

As described previously, the models trained with individual EBAG and PWAA feature could not effectively distinguish viral phosphorylation and non-phosphorylation sites. However, for serine, threonine and tyrosine phosphorylation, the predictive performance of the model trained with the combination of EBAG and PWAA features (EBAG + PWAA) had been remarkably enhanced, as presented in Table 1. The Pr, Sn, Sp, Acc, Mcc and AUC for viral phosphoserine sites were 85.2%, 82.9%, 85.6%, 84.2%, 68.5% and 79.5%, respectively, and Pr of 80.2%, Sn of 84.5%, Sp of 79.1%, Acc of 81.8%, Mcc of 63.7% and AUC of 81.7% for viral phosphothreonine sites. For viral tyrosine phosphorylation sites, the predictive performance of EBAG + PWAA features outperformed those of the models trained with EBAG or PWAA features. The ROCs of serine, threonine and tyrosine phosphorylation sites trained with various features are presented in Fig. S1. Those demonstrated that two kinds of features contributed to distinguish serine, threonine and tyrosine phosphorylation from their non-phosphorylation. There was a strong complementary effect between the two features. Therefore, the combination of EBAG + PWAA, which considered not only physicochemical properties of amino acids but also residue sequence order information, was selected as the optimal feature set to predict protein phosphorylation sites in viruses.

3.2. Investigation of window sizes

The number of residues surrounding the phosphorylation site that are taken into account is important because too few means information useful for making predictions gets ignored, while too many will decrease the signal-to-noise ratio [28]. Several strategies have been used to determine the optimum number of residues. First, it has been argued that the optimum should be consistent with the number of residues in physical contact with the kinase [17]. However, the residues contacted by the kinase may not be the same as the residues surrounding the phosphorylation site in the linear sequence [28]. Second, Neuberger et al. examined how residues around phosphorylation sites compare with residues in general proteins with respect to two properties—hydrophobicity and flexibility [46]. Because this experiment was done only for protein kinase A, it is not known whether its results generalize to other protein kinases. Third, some authors have empirically tested various numbers of residues, and then chosen the number that gives the best predictive performance. The authors of PostMod tried between 7 and 101 residues, and found that 41 resulted in the best accuracy [47]. Other authors reported much smaller optima, with Blom et al. suggesting between 9 and 11 [17] and Biswas et al. reporting 15 [48]. Given that reported optima are inconsistent, we should investigate the most appropriate number of residues for developing new method. In this work, we considered between 15 and 27 residues for SVM prediction.

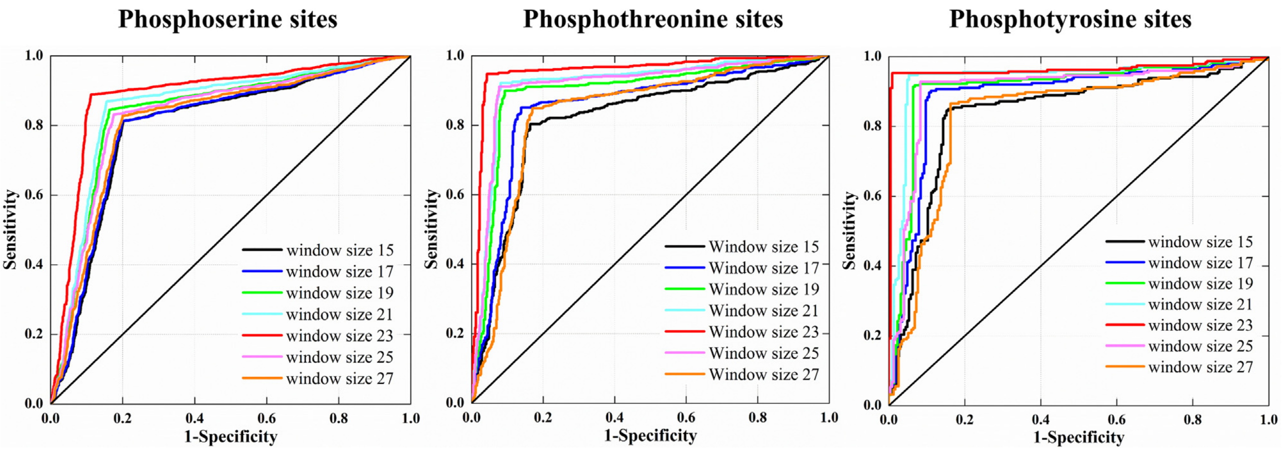


Fig. 2. The ROC curves of models trained with various window sizes for viral phosphorylation sites based on EBAG + PWAA features.

The predictive performance of models trained with different window sizes (15–27) for viral serine, threonine and tyrosine phosphorylation sites were illustrated in Tables S4–S6, respectively. As seen from Table S4, the prediction performance for phosphoserine sites increased with the increasing of window size until the window size reached 23. The Pr, Sn, Sp, Acc, Mcc and AUC for phosphoserine sites with window size of 23 achieved 88.9%, 88.7%, 88.9%, 88.8%, 77.7% and 88.6%, respectively, which were higher than those with other window sizes. Similarly, for threonine, tyrosine phosphorylation sites, the predictive performance of models trained with window size of 23 outperformed other window sizes in Tables S5 and S6. Especially, the Mcc of serine phosphorylation sites with window size of 23 compared with other window sizes increased from 6.4% to 11.4%, 6.5% to 26.8% for phosphothreonine sites and 5.0% to 25.0% for phosphotyrosine sites. From the receiver operating characteristic (ROC) evaluation, we could find there are 23 residues (between −11 and +11) of serine, threonine and tyrosine phosphorylation with areas under ROC curves (AUCs) larger than other number of residues, as shown in Fig. 2. Based on the computational efficiency and overall performance of the models trained with different window length, 23-mer was adopted as the feasible window size for the three phosphorylation residues site identification of viral proteins in this study.

3.3. Comparisons with existing methods

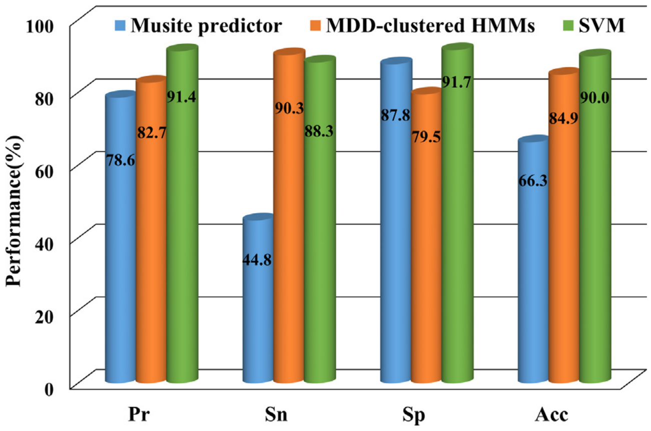


Fig. 3. Comparison of predictive performance for viral phosphoserine sites in Musite predictor, MDD-clustered HMMs and SVM methods on the independent testing data.

Moreover, in order to further evaluate the prediction performance of our SVM method objectively, the independent testing data is used to make comparisons with the existing method of MDD-clustered HMMs method [31] which had the same testing sets. Besides, we put our independent test sets into previously developed Musite predictor [35] to make phosphorylation site identification by using our training datasets. The comparisons of predictive performance for serine and threonine phosphorylation in the SVM method, MDD-clustered HMMs method and Musite predictor were shown in Figs. 3 and 4, respectively. As seen from Figs. 3 and 4, the Sn of the Musite predictor for serine and threonine phosphorylation sites only reached 44.8% and 37.7%, respectively, which were significantly lower than that of SVM and MDD-clustered HMMs methods. Though the Sn of MDD-clustered HMMs method for phosphoserine sites was 2.0% higher than that of SVM method, the Pr, Sp and Acc of SVM method achieved 91.4%, 91.7% and 90.0%, respectively, which were higher than those of MDD-clustered HMMs method in Fig. 3. As shown in Fig. 4, for phosphothreonine sites, the Pr, Sn, Sp and Acc of SVM method were 12.9%, 5.3%, 14.1% and 9.7% higher than those of MDD-clustered HMMs method, respectively. Due to a lack of virus phosphotyrosine data, the MDD-clustered HMMs method could not predict the results of tyrosine phosphorylation sites. The predictive performance for viral phosphotyrosine sites of SVM method and Musite predictor on the independent testing data was shown in Table S7. The Sn of the SVM method for tyrosine phosphorylation sites was 42.9% higher than that of Musite predictor. Moreover, the area under the curves of SVM method for serine, threonine and tyrosine phosphorylation sites were larger

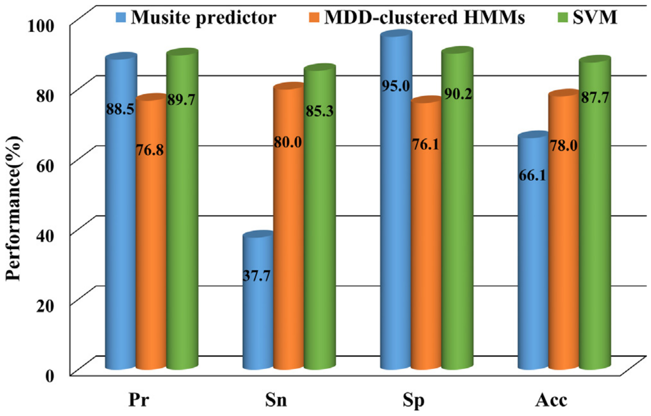


Fig. 4. Comparison of predictive performance for viral phosphothreonine sites in Musite predictor, MDD-clustered HMMs and SVM methods on the independent testing data.

than that of Musite predictor in Fig. S2, indicating that the SVM method could effectively predict viral phosphotyrosine sites. There are two possible reasons for our significant improvements: first, using a large number of the training data from NCBI and P^3DB databases means more information to the SVM, resulting in a more accurate model. Second, the combination of EBAG + PWAA feature is effective in identifying phosphorylation status. The analysis result revealed that the SVM model incorporated EBAG + PWAA feature was effective and feasible in identifying phosphorylation sites of viral proteins.

4. Conclusion

In this work, we developed a method to identify protein phosphorylation sites on viruses. Our approach considered not only protein sequence information but also physicochemical properties of amino acids within the serine, threonine and tyrosine regions. The prediction model achieved a promising performance, and an independent testing set also demonstrated that our proposed method outperformed MDD-clustered HMMs and Musite methods. Feature analyses revealed that acidic residues contributed to the occurrence of viral phosphorylation sites, and there were distinct kinase-specific and residue-conservative differences for serine, threonine, and tyrosine phosphorylation sites of virus proteins. The detailed feature analysis in this study might help understand the viral phosphorylation mechanism and guide the related experimental validation.

Acknowledgements

This work was supported by Program for New Century Excellent Talents in University (NCET-11-1002) and the National Natural Science Foundation of China (21175064 and 21305062).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jmgm.2014.12.005.

References

[1] P. Blume-Jensen, T. Hunter, Oncogenic kinase signalling, Nature 411 (2001) 355–365.

[2] L.N. Johnson, D. Barford, The effects of phosphorylation on the structure and function of proteins, Annu. Rev. Biophys. Biomol. Struct. 22 (1993) 199–232.

[3] L.M.J. Law, J.C. Everitt, M.D. Beatch, C.F.B. Holmes, T.C. Hobman, Phosphorylation of rubella virus capsid regulates its RNA binding activity and virus replication, J. Virol. 77 (2003) 1764–1771.

[4] S.C. Das, A.K. Pattnaik, Phosphorylation of vesicular stomatitis virus phosphoprotein P is indispensable for virus growth, J. Virol. 78 (2004) 6420–6430.

[5] B. Hemonnot, The host cell MAP kinase ERK-2 regulates viral assembly and release by phosphorylating the p6gag protein of HIV-1, J. Biol. Chem. 279 (2004) 32426–32434.

[6] S. Wang, Z. Zhao, Y. Bi, L. Sun, X. Liu, W. Liu, Tyrosine 132 phosphorylation of influenza a virus M1 protein is crucial for virus replication by controlling the nuclear import of M1, J. Virol. 87 (2013) 6182–6191.

[7] T. Kleinow, M. Nischang, A. Beck, U. Kratzer, F. Tanwir, W. Preiss, G. Kepp, H. Jeske, Three C-terminal phosphorylation sites in the Abutilon mosaic virus movement protein affect symptom development and viral DNA accumulation, Virology 390 (2009) 89–101.

[8] Y. Wang, L. Jiang, Y. Huang, M. Deng, T. Liu, W. Lai, X. Ye, Polo-like kinase 1 inhibits the activity of positive transcription elongation factor of RNA pol II b (P-TEFb), PLOS ONE 8 (2013) e72289.

[9] T.Y. Hsiang, L. Zhou, R.M. Krug, Roles of the phosphorylation of specific serines and threonines in the NS1 protein of human influenza A viruses, J. Virol. 86 (2012) 10370–10376.

[10] S.F. Cheng, M.S. Tsai, C.L. Huang, Y.P. Huang, I.H. Chen, N.S. Lin, Y.H. Hsu, C.H. Tsai, C.P. Cheng, Ser/Thr kinase-like protein of Nicotiana benthamiana is involved in the cell-to-cell movement of Bamboo mosaic virus, PLOS ONE 8 (2013) e62907.

[11] R. Aebersold, M. Mann, Mass spectrometry-based proteomics, Nature 422 (2003) 198–207.

[12] S.M. Wurgler-Murphy, D.M. King, P.J. Kennelly, The phosphorylation site database: a guide to the serine-, threonine-, and/or tyrosine-phosphorylated proteins in prokaryotic organisms, Proteomics 4 (2004) 1562–1570.

[13] F. Diella, C.M. Gould, C. Chica, A. Via, T.J. Gibson, Phospho.ELM: a database of phosphorylation sites update 2008, Nucleic Acids Res. 36 (2007) D240–D244.

[14] P.V. Hornbeck, I. Chabra, J.M. Kornhauser, E. Skrzypek, B. Zhang, PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation, Proteomics 4 (2004) 1551–1561.

[15] P. Durek, R. Schmidt, J.L. Heazlewood, A. Jones, D. MacLean, A. Nagel, B. Kersten, W.X. Schulze, PhosPhAt: the *Arabidopsis thaliana* phosphorylation site database. An update, Nucleic Acids Res. 38 (2010) D828–D834.

[16] P.J. Boersema, S. Mohammed, A.J. Heck, Phosphopeptide fragmentation and analysis by mass spectrometry, J. Mass Spectrom. 44 (2009) 861–878.

[17] N. Blom, S. Gammeltoft, S. Brunak, Sequence and structure-based prediction of eukaryotic protein phosphorylation sites, J. Mol. Biol. 294 (1999) 1351–1362.

[18] C.R. Ingrell, M.L. Miller, O.N. Jensen, N. Blom, NetPhosYeast: prediction of protein phosphorylation sites in yeast, Bioinformatics 23 (2007) 895–897.

[19] H.D. Huang, T.Y. Lee, S.W. Tzeng, L.C. Wu, J.T. Horng, A.P. Tsou, K.T. Huang, Incorporating hidden Markov models for identifying protein kinase-specific phosphorylation sites, J. Comput. Chem. 26 (2005) 1032–1041.

[20] N.F. Saunders, R.I. Brinkworth, T. Huber, B.E. Kemp, B. Kobe, Predikin and PredikinDB: a computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites, BMC Bioinform. 9 (2008) 245.

[21] M.B. Yaffe, G.G. Leparc, J. Lai, T. Obata, S. Volinia, L.C. Cantley, A motif-based profile scanning approach for genome-wide prediction of signaling pathways, Nat. Biotechnol. 19 (2001) 348–353.

[22] L. Li, C. Wu, H. Huang, K. Zhang, J. Gan, S.S. Li, Prediction of phosphotyrosine signaling networks using a scoring matrix-assisted ligand identification approach, Nucleic Acids Res. 36 (2008) 3263–3273.

[23] B. Sobolev, D. Filimonov, A. Lagunin, A. Zakharov, O. Koborova, A. Kel, V. Poroikov, Functional classification of proteins based on projection of amino acid sequences: application for prediction of protein kinase substrates, BMC Bioinform. 11 (2010) 313.

[24] D. Plewczynski, A. Tkacz, L.S. Wyrwicz, L. Rychlewski, K. Ginalski, AutoMotif Server for prediction of phosphorylation sites in proteins using support vector machine: 2007 update, J. Mol. Model. 14 (2008) 69–76.

[25] P. Durek, C. Schudoma, W. Weckwerth, J. Selbig, D. Walther, Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins, BMC Bioinform. 10 (2009) 1–17.

[26] J. Gao, D. Xu, The Musite open-source framework for phosphorylation-site prediction, BMC Bioinform. 11 (Suppl. 12) (2010) S9.

[27] K. Swaminathan, R. Adamczak, A. Porollo, J. Meller, Enhanced prediction of conformational flexibility and phosphorylation in proteins, Adv. Exp. Med. Biol. 680 (2010) 307–319.

[28] B. Trost, A. Kusalik, Computational prediction of eukaryotic phosphorylation sites, Bioinformatics 27 (2011) 2927–2935.

[29] A. Via, F. Diella, T.J. Gibson, M. Helmer-Citterich, From sequence to structural analysis in protein phosphorylation motifs, Front. Biosci. 16 (2011) 1261–1275.

[30] D. Schwartz, G.M. Church, Collection and motif-based prediction of phosphorylation sites in human viruses, Sci. Signal. 3 (2010) rs2.

[31] N.A. Bretaña, C.T. Lu, C.Y. Chiang, M.G. Su, K.i. Huang, T.Y. Lee, S.L. Weng, Identifying protein phosphorylation sites with kinase substrate specificity on human viruses, PLOS ONE 7 (2012) e40694.

[32] Q. Yao, H. Ge, S. Wu, N. Zhang, W. Chen, C. Xu, J. Gao, J.J. Thelen, D. Xu, P^3DB 3.0: from plant phosphorylation sites to protein networks, Nucleic Acids Res. 42 (2013) D1206–D1213.

[33] T.-Y. Lee, N. Bretaña, C.-T. Lu, PlantPhos: using maximal dependence decomposition to identify plant phosphorylation sites with substrate site specificity, BMC Bioinform. 12 (2011) 261.

[34] P. Radivojac, V. Vacic, C. Haynes, R.R. Cocklin, A. Mohan, J.W. Heyen, M.G. Goebl, L.M. Iakoucheva, Identification, analysis, and prediction of protein ubiquitination sites, Proteins 78 (2010) 365–380.

[35] J. Gao, J.J. Thelen, A.K. Dunker, X. Dong, Musite, a tool for global prediction of general and kinase-specific phosphorylation sites, Mol. Cell Proteomics 9 (2010) 2586–2600.

[36] S.C. Fan, X.G. Zhang, Characterizing the microenvironment surrounding phosphorylated protein sites, Genomics Proteomics Bioinform. 3 (2005) 213–217.

[37] Z.H. Zhang, Z.H. Wang, Z.R. Zhang, Y.X. Wang, A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine, FEBS Lett. 580 (2006) 6169–6174.

[38] L. Nanni, A. Lumini, An ensemble of reduced alphabets with protein encoding based on grouped weight for predicting DNA-binding proteins, Amino Acids 36 (2009) 167–175.

[39] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.

[40] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 1–27, Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm

[41] V. Vacic, L.M. Iakoucheva, P. Radivojac, Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments, Bioinformatics 22 (2006) 1536–1537.

[42] M.L. Miller, L.J. Jensen, F. Diella, C. Jorgensen, M. Tinti, L. Li, M. Hsiung, S.A. Parker, J. Bordeaux, T. Sicheritz-Ponten, M. Olhovsky, A. Pasculescu, J. Alexander, S. Knapp, N. Blom, P. Bork, S. Li, G. Cesareni, T. Pawson, B.E. Turk, M.B.

Yaffe, S. Brunak, R. Linding, Linear motif atlas for phosphorylation-dependent signaling, Sci. Signal. 1 (2008) ra2.

[43] R. Amanchy, K. Kandasamy, S. Mathivanan, B. Periaswamy, R. Reddy, Identification of novel phosphorylation motifs through an integrative computational and experimental analysis of the human phosphoproteome, J. Proteomics Bioinform. 04 (2011) 022–035.

[44] B.E. Kemp, R.B. Pearson, Protein kinase recognition sequence motifs, Trends Biochem. Sci. 15 (1990) 342–346.

[45] D. Schwartz, M.F. Chou, G.M. Church, Predicting protein post-translational modifications using meta-analysis of proteome scale data sets, Mol. Cell Proteomics 8 (2009) 365–379.

[46] G. Neuberger, G. Schneider, F. Eisenhaber, pkaPS: prediction of protein kinase A phosphorylation sites with the simplified kinase-substrate binding model, Biol. Direct 2 (2007) 1.

[47] I. Jung, A. Matsuyama, M. Yoshida, D. Kim, PostMod: sequence based prediction of kinase-specific phosphorylation sites with indirect relationship, BMC Bioinform. 11 (Suppl. 1) (2010) S10.

[48] A. Biswas, N. Noman, S. Abdur, Machine learning approach to predict protein phosphorylation sites by incorporating evolutionary information, BMC Bioinform. 11 (2010) 273.