International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 18, 2019, pp 29-44

A Class of Computational Predictor for Prediction of Protein Phosphorylation Sites

Adiba Sultana[1, 2*], Samme Amena Tasmia[2], Md. Shahin Alam[1, 2], Alima Khanam[3], Md. Selim Reza[2], Md. Mehedi Hassan[2,4], Md. Hadiul Kabir[2] and Md. Nurul Haque Mollah[2*]

[1]School of Biology and Basic Medical Sciences, Soochow University, Suzhou, Jiangsu Province, China
[2] Laboratory of Bioinformatics, Department of Statistics, Rajshahi University, Rajshahi6205, Bangladesh
[3]Department of Biochemistry and Molecular Biology, Rajshahi University, Rajshahi-6205, Bangladesh
[4]Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan

*Correspondence should be addressed to Adiba Sultana and Md. Nurul Haque Mollah
(adibajui1992@yahoo.com, mollah.stat.bio@ru.ac.bd)

[Received March 10, 2019; Revised April 21, 2019; Accepted November 1, 2019]

Abstract

One of the most common protein post-translational modifications (PTMs) in eukaryotes is phosphorylation. Protein phosphorylation on serine (S), threonine (T) and tyrosine (Y) has appeared as a key scheme in the control of many biological processes. Identification of protein phosphorylation sites is an important and perquisite for understanding the mechanisms of phosphorylation. However, most the experimental methods for identifying phosphorylation sites are not only costly but also time consuming. Hence, sequence-based computational methods are highly desired. Feature selection plays an important role to develop the effective computational predictors of PTM sites. To select a better feature filtering approach for our current problem, we performed a comparative study on fivepopular feature selection approaches and found the non-parametric Wilcoxon-signed rank test approach as the better candidate for our dataset on phosphorylation sites with serine (S), threonine (T) and tyrosine (Y) residues. Finally, we have proposed a predictor combining CKSAAP encoding, Wilcoxon-signed rank test, 1:3 ratio of positives versus

negatives samples of windows and Support Vector Machine (SVM) classifier for prediction of phosphorylation sites of candidate proteins. Our data analysis results showed that the proposed method outperform over the existing predictors of phosphorylation sites. The proposed predictor exhibited the performance with accuracy (Ac) 98.65% (Sn =82.94%, Sp= 99.63%, MCC=0.873) for S, accuracy (Ac) 98.81% (Sn = 91.52%, Sp =99.19%, MCC= 0.877) for T and accuracy (Ac) 97.83% (Sn =97.62%, Sp =99.38%, MCC= 0.909) for Y at 10% false positive rate. Thus, the proposed method would be helpful computational method for the phosphorylation sites prediction.

Key words: Protein sequences, Protein phosphorylation Site, CKSAAP encoding, Support vector machine, Wilcoxon Signed-Rank Test, and Amino acid frequency.

AMS Classification: 92D20.

1. Introduction

Phosphorylation is one of the most common protein post-translational modifications (PTMs) in eukaryotes which plays significant roles in a wide range of cellular processes, such as DNA repair (Wood et al., 2009), regulation of transcription (Uddin et al., 2003), immune response (Kim et al., 2011), metabolism (Bu et al., 2010), cellular motility (Ressurreico et al., 2011), and environmental stress response (Wang etal., 2010). Protein phosphorylation PTM site is added to an amino acid residue (S, T, and Y) of a protein molecule. As one of the most challenging PTM site, phosphorylation is involved in many biological processes including cell cycle, apoptosis. Phosphorylation of amino acid residues serine (S), threonine (T) and tyrosine (Y) which is common in cancer-associated proteins (Iakoucheva et al., 2004) and known to be deregulated in cancer (Lim et al. 2005). Coding-region mutations in human genes are responsible for a diverse spectrum of diseases and any other phenotypes. A study (Cohen et al., 2002) indicates that 30% of proteins in the human genome can be phosphorylated, and abnormal phosphorylation is now recognized as a cause of human disease. The malfunctioning of specific chains of protein tyrosine kinases and protein tyrosine phosphatase has been linked to multiple human diseases such as obesity, insulin resistance, and type-2 diabetes mellitus. Phosphorylation of glucose is imperative in processes within the body. For example, phosphorylating glucose is necessary for insulin-dependent mechanistic target of rapamycin pathway activity within the heart. This further suggests a link between intermediary metabolism and cardiac

growth. The most commonly associated histone phosphorylation occurs during cellular responses to DNA damage, when phosphorylated histone H2A separates large chromatin domains around the site of DNA breakage. Identification of phosphorylated substrates and their corresponding sites will facilitate the understanding of the molecular mechanism of phosphorylation. So now days it is very important to study protein phosphorylation sites. Comparing with the labor-intensive and time-consuming experiment approaches, computational prediction of phosphorylation sites is much desirable due to their convenience and fast speed. Therefore, in this article, we would like to study the computational methods for prediction of phosphorylation PTM sites. The aberrances of PTMs are highly associated in diseases and cancers, while a variety of regulatory enzymes involved in PTMs have been drug targets (Lahiry et al., 2010; Norvell et al., 2010). In this regard, elucidation of PTMs regulatory roles is fundamental for understanding molecular mechanisms of diseases and cancers, and further biomedical design. It has been estimated that 30–50% of the proteome undergone phosphorylation (Pinna et al., 1996). Therefore, accurate recognition of the phosphorylation substrates and the corresponding phosphorylation sites may help fully decipher the molecular mechanisms of phosphorylation related biological processes. Conventional experimental identification of phosphorylation sites with a site-directed mutagenesis strategy is laborious, expensive, and low-throughput (Meier et al., 1997). Recently, the appearance of high-throughput mass spectrometry technique (Jensen et al. 2004) has greatly accelerated the identification of novel phosphorylation sites. Accordingly, several phosphorylation site databases have been established, such as 'Phospho.ELM' (Xue et al., 2008), 'Phosphorylation Site Database' (Gnad F et al., 2007), 'PhosPhAT'(Heazlewood JL et al., 2008), and 'Phosphosite' (Hornbeck PV et al., 2004). However, some limitations of this technique (Boersema PJ et al., 2009) make the exact prediction of phosphorylation sites difficult, and it always requires very expensive instruments and specialized expertise that are usually not available in general laboratories. With the increasing availability of protein sequence data, there is an urgent need for computational tools that can rapidly and reliably identify phosphorylation sites. In recent years, many computational predictors have been developed and applied with varying success to predict phosphorylation sites (Huang H et al., 2005). For the analysis , proteins collected from the phosphorylation site databases Phospho.ELM .There

are several proposed generalized prediction tools which used the primary sequence information for classifying phosphorylation sites, such as DISPHOS(Lakoucheva L et al., 2004), Scansite (Obenauer J et al., 2003), PPRED (Ashis KB et al., 2010), NetPhos (Blom N et al., 1999), PHOSIDA (Gnad F et al., 2007), and AutoMotif Server AMS (Plewcznski D et al., 2005). However, their performances are not yet in the satisfactory level. Therefore, in this paper, an attempt is made to propose a new predictor for phosphorylation site prediction. Feature selection plays an important role to develop the effective computational predictors of PTM sites. In order to build a more effective predictor, at first, we have to select a better feature selection approach by comparing five different popular feature selection approaches (i.e. t-test, Wilcoxon signed-rank, Kruskal-wallis, LIMMA and SAM) based on CKSAAP encoding scheme. Secondly, we haveemployed a relatively balanced ratio betweenthe positive and negative samples during the training of the classifiers (e.g. the ratio of positives versus negatives is controlled at 1:1 or 1:2 or 1:3) based on Support Vector Machine to predict phosphorylation PTM sites. We have compared three ratios because of highly unbalanced datasets of the phosphorylation and non- phosphorylation are. In order to evaluate the performance of the proposed PTM Sites predictors, four measurements will be used: sensitivity (Sn), specificity (Sp), accuracy (Ac) and Matthew correlation coefficient (MCC).

2. Materials and Methods

2.1 Datasets

The datasets used in this paper were divided into two parts: training dataset and independent testing dataset. The dataset were collected from Ashis and co-workers (Ashis KB et al., 2010). Experimentally validated phosphorylation sites were extracted from the Phospho.ELM database (version 8.1 released on August 12, 2008) (Diella F et al., 2004), which contained 837 proteins covering 1450 phosphorylated serine sites, 835 phosphorylated threonine sites and 286 phosphorylated tyrosine sites. To classify the phosphorylated proteins, the experimentally validated phosphorylated sites were detected as positive samples (i.e. phosphorylated sites) and all the other residues as negative samples (i.e. non-phosphorylated sites). In order to evaluate the prediction performance among different predictors, we take a new dataset by taking 200 protein sequences as an

independent dataset. To develop the PTM site predictors, the sliding window strategy was utilized to extract positive and negative samples. We consider the optimal window size 27 in this paper, with 13 residues located upstream and 13 residues located downstream of the phosphorylation sites in the protein sequences. Then we apply 5-fold cross-validation to evaluate and compare the performance of proposed predictors with existing predictors.

2.2 Construction of feature vectors

2.2.1 Feature encoding

In order to build an effective prediction model, we encoded each sequence fragment into a numeric vector, which was the crucial step to present the classifier and ensemble architecture. Thus, a high-quality sequence encoding method for keeping the generated code compact in dimensionality was necessary. Instead of employing a simple binary representation, three types of amino acid feature encodings were adopted, including CKSAAP, binary and AAindex encoding. In this study, the composition of *k*-spaced amino acid pairs (CKSAAP) based encoding scheme was used. CKSAAP could reflect the characteristics of the residues surrounding phosphorylation sites, and it has been successfully used for predicting palmitoylation sites (Wang XB et al., 2009) and mucin-type O-glycosylation sites (Chen YZ et al., 2008) to represent the sequence fragment. It may create (21× 21) = 441 (21means 21 types of amino acids (including the gap (O))) types of amino acid pairs (i.e. AA, AC, AD,. . ., OO) for every single k (k denotes the space between two amino acids), if window size of a fragment is 2w+ 1. For the optimal k taking k max =5, there are 21× (kmax + 1) × 21 = 2646 different amino acid pairs are created for each sequence. Then the feature vectors are calculated using the following equation:

$$\left(\frac{N_{AA}}{N_{total}},\frac{N_{AC}}{N_{total}},\frac{N_{AD}}{N_{total}},\ \ldots\ ,\frac{N_{OO}}{N_{total}}\right)441 \quad (1)$$

Where N_{total} denotes the length of the total composition residues. N_{AA}, N_{AC},..., N_{OO} are frequency of the amino acid pair within the fragment. More details are available somewhere.

2.3 Feature selection

Feature selection techniques have become an apparent need in many bioinformatics applications. In machine learning as the dimensionality of the data

rises, the amount of data required to provide a reliable analysis grows exponentially. Protein datasets have the high dimensionality problem. Each data point can have up to 2646 variables and processing a large number of data points involves high computational cost, and the analysis become very hard. To overcome this problem, it is necessary to find a way to reduce the number of features in construction. In this paper, we compare various feature selection techniques parametric test *t*-test (Jafari and Azuaje et al., 2006), non-parametric test Wilcoxon signed-rank, Kruskal–Wallis test. We also used some software for feature selection such as WEKA (Java), SAM (R) and Limma (R). Among them Wilcoxon signed-rank test perform better than others for sequence analysis.

2.4 SVM learning

Support vector machines (SVMs) was first developed by Vapnik (1995) described in detail by Cristianini and Shawe-Taylor (2000). Support vector machines (SVMs) are based on statistical learning theory, is a popular machine learning algorithm mainly used in dealing with binary classification problems. SVM looks for a rule that best maps each member of training set to the correct classification, and it has been widely used in bioinformatics community. Formally, given a training vector $x_i \in R_n$ and $y_i \in \{-1, +1\}$ be the corresponding class labels, i=1, …, N, SVM solves the following optimization problem:

$$\text{Minimize } \frac{1}{2}\mathrm{w}^{\mathrm{T}}.\mathrm{w} + \mathrm{c}\sum_{i=1}^{N} \xi_i$$

$$\text{Subject to } \mathrm{y_i}(\mathrm{w}^{\mathrm{T}}.\mathrm{x_i} + \mathrm{b}) y_i(w^T.x_i + b) \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0$$

Where w is a normal vector perpendicular to the hyperplane, the regularization parameter C controls the trade-off between the margin and the training error, and ξi is slake variables for allowing misclassifications.

2.5 Performance Measurements

In order to evaluate our predictor, four measurements are used: sensitivity (*Sn*), specificity (*Sp*), accuracy (*Ac*) and Matthew correlation coefficient (*MCC*). They are defined by the following formulas:

$$Sn = \frac{TP}{TP+FN}\,;\ 0 \leq Sn \leq 1 \tag{2}$$

$$Sp = \frac{TN}{TN+FN}\,;\ 0 \leq Sp \leq 1 \tag{3}$$

$$Ac = \frac{TP+TN}{TP+TN+FP+FN}\,;\, 0 \le Ac \le 1 \quad (4)$$

$$MCC = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FN)\times(TN+FP)\times(TP+FP)\times(TN+FN)}};\, -1 \le MCC \le 1 \quad (5)$$

Where, TP represents the observed positive residues predicted to be the positive sample, TN the observed negative residues predicted to be the negative sample, FP the number of the observed positive residues predicted to be the negative, and FN the number of the observed negative residues predicted to be the positive sample, respectively. The prediction validity is often examined by observing its ROC curve because they are able to show the trade-off between sensitivity and specificity and give a complete evaluation. The area under the ROC curve (AUC) is another important indicator, the larger, the better.

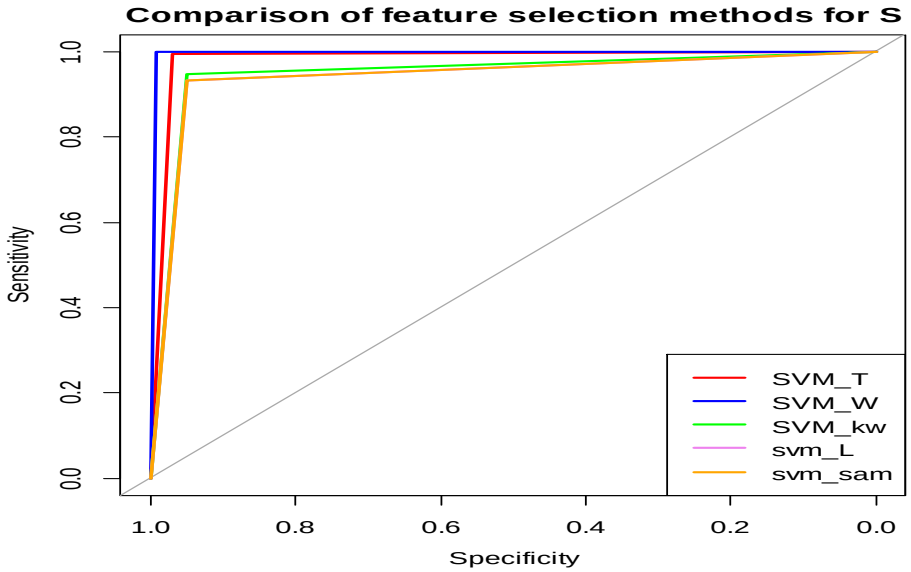
3. Results and Discussion

Firstly, the sequence fragments were encoded as numerical vectors by using CKSAAP encoding scheme, then we use feature selection technique to reduce the problem of high dimensionality and finally the predictor was established with the SVM algorithm (Table 1). The feature selection method Wilcoxon signed-rank test gives the highest performance with accuracy (Ac) reached 99.35% for S (Sn =97.45%, Sp= 99.32%, MCC=0.933), accuracy (Ac) 97.15% for T (Sn = 98.38%, Sp =97.08%, MCC= 0.782) and accuracy (Ac) 88.95% for Y (Sn =100%, Sp =88.60%, MCC= 0.443). Since the predictor is a discrete classifier, the ROC curves for each of the three residues (S, T and Y) have been plotted, as can be seen in Figure 1. By these results we can say that Wilcoxon signed-rank test is the best feature selection method among these five methods. After that we perform the average prediction results of Wilcoxon signed-rank test using different negative data sets for 5-fold cross validation test. Using 1:1, 1:2 and 1:3 shuffled protein sequences as negative data set, the average prediction under MCC value is 24.09%, 67.49% and 87.33%, respectively, when specificity controlled over 99% for Serine (S) site. The prediction MCC increases with the increase of ratio of dataset. The similar cases happened for Threonine (T) and Tyrosine (Y) sites (Table 2). Then we plot the ROC curve for comparison of the ratio for three sites (Figure 2). Finally we checked the Performance of our proposed method with different predictors in terms of serine (S), threonine (T) and tyrosine (Y) site prediction on the independent datasets (Table 3), and we found that with

CKSAAP encoding, SVM classification and for 1:3 cost validation, Wilcoxon signed-rank test is perform as the best feature selection method for phosphorylation site prediction.

Table 1: Comparison of the feature selection methods in terms of serine (S), threonine (T) and tyrosine (Y) site prediction based on test dataset

Site	Methods	Test.error	AUC	Sn (%)	Sp (%)	Ac (%)	Mcc
S	t-test	0.03088114	0.9811	96.32	99.36	96.91	0.932
	Wilcoxon signed-rank	0.006450727	0.9966	97.45	99.32	97.00	0.933
	Kruskal-wallis	0.04982158	0.9488	94.73	95.02	95.02	0.205
	LIMMA	0.05037057	0.9415	93.33	94.96	94.96	0.180
	SAM	0.04982158	0.9415	93.33	94.96	94.96	0.180
T	t-test	0.06911974	0.9238	91.66	93.10	93.08	0.312
	Wilcoxon signed-rank	0.02846107	0.9774	98.38	97.08	97.15	0.782
	Kruskal wallis	0.07725147	0.7117	50.00	92.34	92.27	0.063
	LIMMA	0.07745477	0.6759	42.85	92.32	92.25	0.049
	SAM	0.07725147	0.7117	50.00	92.34	92.27	0.063
Y	t-test	0.1245804	0.9366	1.00	87.32	87.54	0.325
	Wilcoxon signed-rank	0.1104066	0.943	1.00	88.60	88.95	0.443
	Kruskal-wallis	0.1294293	0.7344	56.84	90.04	87.05	0.384
	LIMMA	0.1346512	0.7186	53.61	90.11	86.53	0.372
	SAM	0.1294293	0.7545	61.03	89.87	87.57	0.394



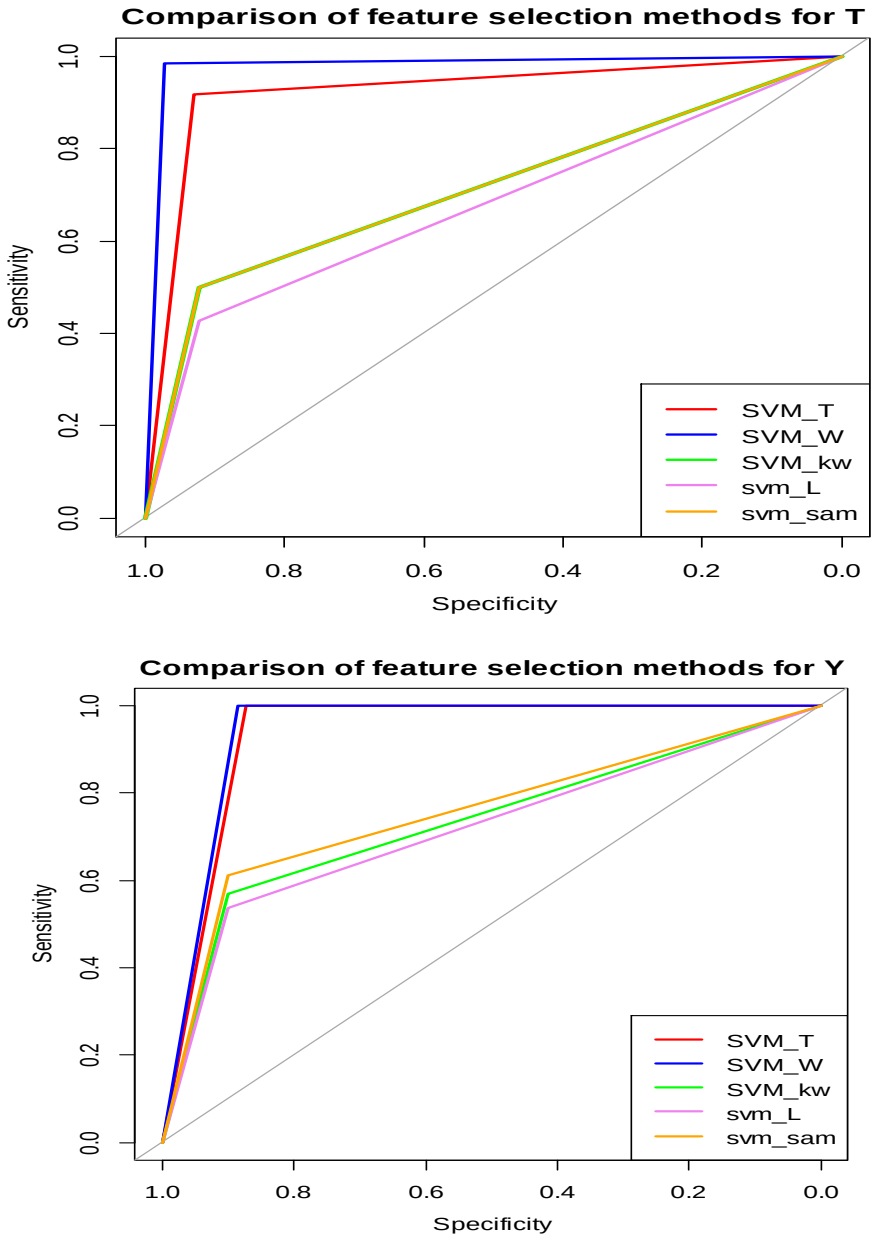
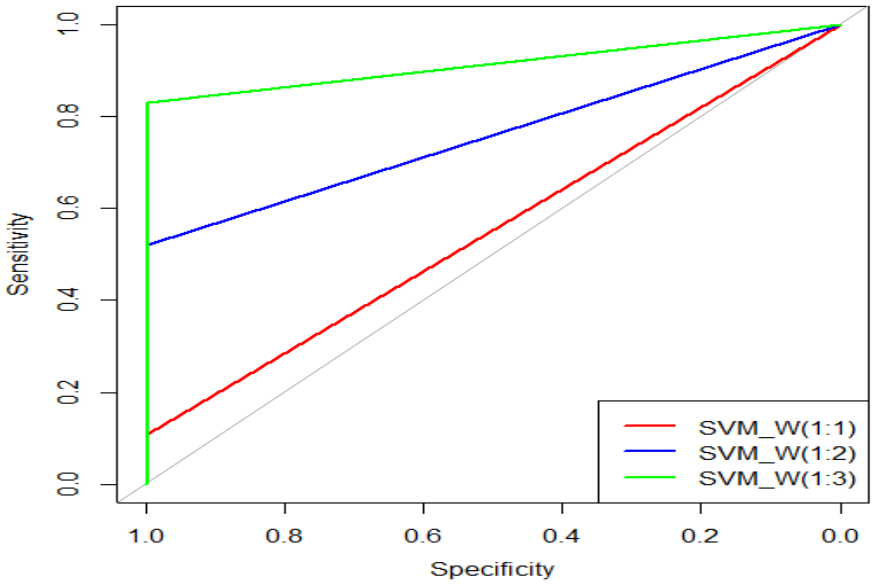


Figure 1: ROC curves of feature selection methods in terms of serine (S), threonine (T) and tyrosine (Y) site site prediction based on test dataset.

Table 2: The prediction performance of the method based on the ratio of different positive and negative datasets of serine (S), threonine (T) and tyrosine (Y) site

Site	Methods	1:1	1:2	1:3
S	Test.error	0.4121603	0.04831183	0.01345045
	AUC	0.5537	0.7583	0.9129
	Sn	0.1100208	0.5206490	0.8294393
	Sp	0.9974509	0.9959140	0.9963546
	Ac	0.5878397	0.9516882	0.9865495
	MCC	0.2409693	0.6749445	0.8733477
T	Test.error	0.431787	0.07027175	0.01180346
	AUC	0.5485	0.7047	0.9536
	Sn	0.1028145	0.4168766	0.9152542
	Sp	0.9942166	0.9924522	0.9919215
	Ac	0.5682130	0.9297282	0.9881965
	Mcc	0.2179919	0.5736761	0.8772376
Y	Test.error	0.2857143	0.03319657	0.02163372
	AUC	0.6555	0.9355	0.985
	Sn	0.3248639	0.8921833	0.9938650
	Sp	0.9860671	0.9787879	0.9762208
	Ac	0.7142857	0.9668034	0.9783663
	Mcc	0.4386274	0.8622766	0.9090009



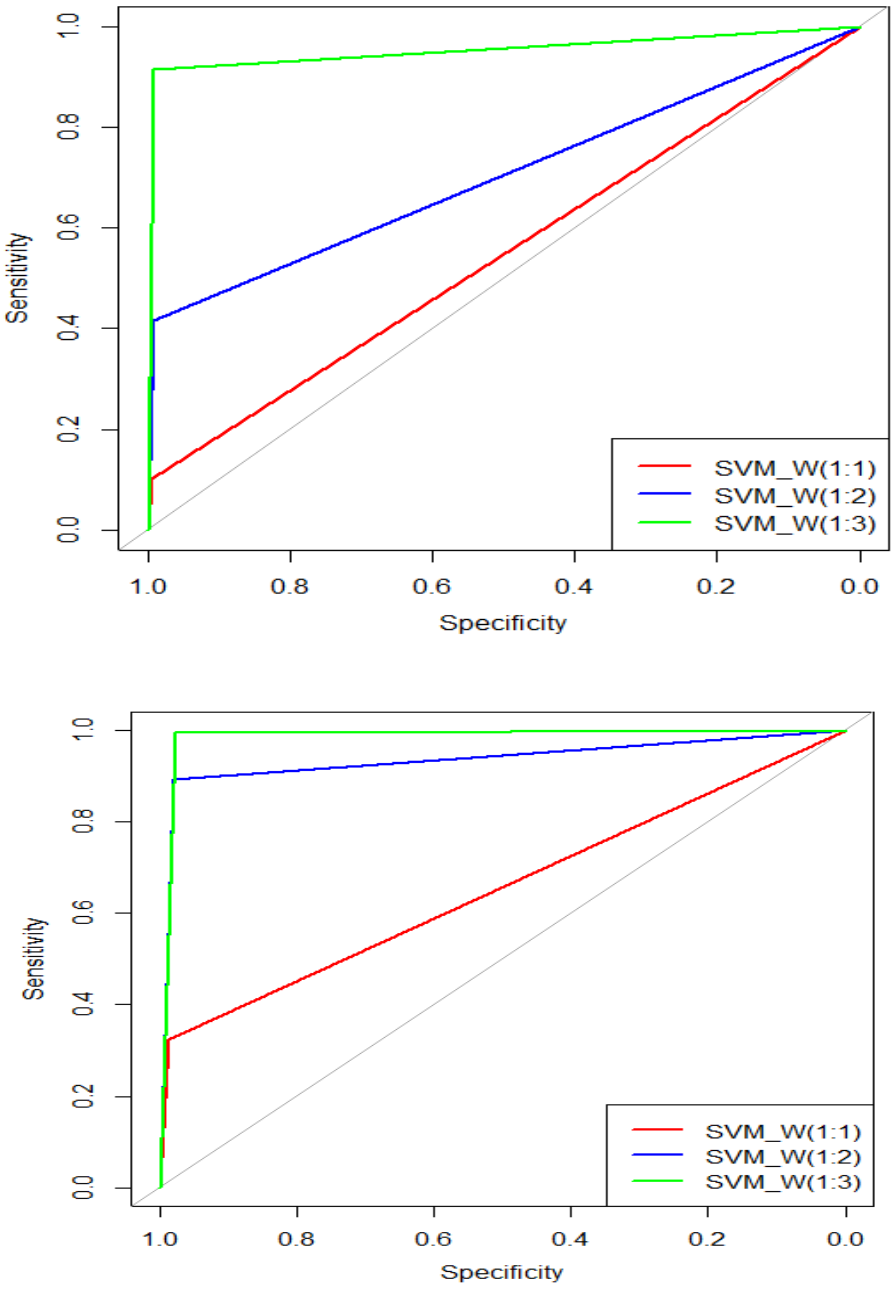
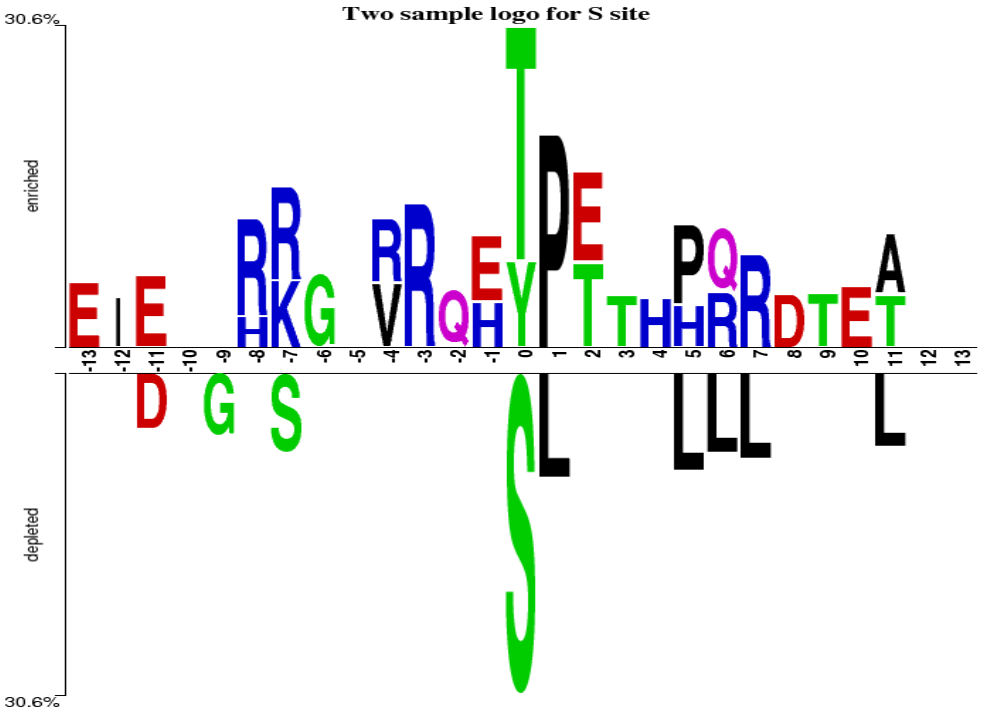


Figure 2: ROC curve for comparison of ratio for Serine (S), Threonine (T), Tyrosine (Y) site.

Table 3: Performance of different predictor and proposed method in terms of serine (S), threonine (T) and tyrosine (Y) site prediction on the independent datasets

Site	Methods	Sn (%)	Sp (%)	Ac (%)	MCC (%)
S	CKSAAP_PhSite	84.81	86.07	85.43	70.90
	DISPHOS	38.88	98.90	96.11	44.03
	PPRED	72.62	56.54	62.87	28.60
	NetPhos	47.78	74.75	64.70	23.10
	Proposed Method	82.94	99.63	98.65	87.3
T	CKSAAP_PhSite	78.59	82.26	80.31	59.90
	DISPHOS	22.22	94.19	88.24	21.11
	PPRED	48.26	70.34	62.12	18.7
	NetPhos	47.78	74.75	64.70	23.10
	Proposed Method	91.52	99.19	98.81	87.7
Y	CKSAAP_PhSite	74.44	78.03	76.21	52.40
	DISPHOS	58.33	97.91	78.30	59.21
	PPRED	43.01	65.35	56.42	8.40
	NetPhos	45.80	69.30	69.92	15.40
	Proposed Method	97.62	99.38	97.83	90.9



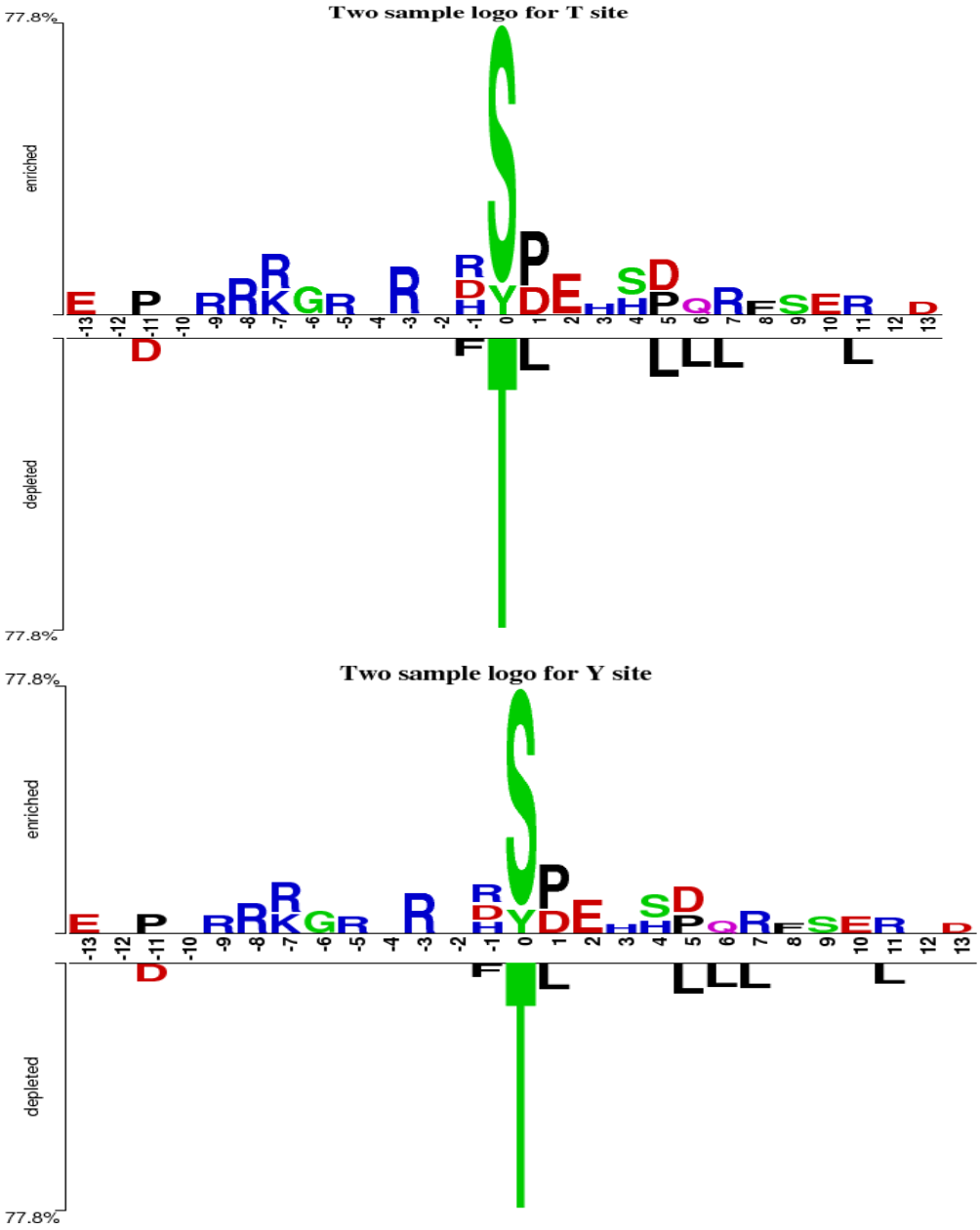


Figure 3: Three Two-Sample-Logos of the position-specific residue composition surrounding the phosphorylated site and nonphosphorylated sites. (A) serine site logo, (B) threonine site logo, (C) tyrosine site logo. These three logos were generated using the web server http://www.twosamplelogo.org/ and only residues significantly enriched and depleted surrounding phosphorylated sites (t-test, P,0.05) are shown.

4. Conclusion

Accurate identification of the phosphorylation substrates and the corresponding phosphorylation sites could helpfully decipher the molecular mechanisms of phosphorylation related biological processes. Though some researchers have focused on this problem, the overall accuracy of their predictionsare still not yet satisfied. Therefore, in this paper, an attempt is made to propose a new predictor for phosphorylation site prediction. Feature selection plays an important role to develop the effective computational predictors of PTM sites. To select a better feature filtering approach for our current problem, we performed a comparative study on five popular feature selection approaches (i.e. t-test, Wilcoxon signed-rank, Kruskal-wallis, LIMMA and SAM) and found the non-parametric Wilcoxon-signed rank test approach as the better candidate for our dataset on phosphorylation sites with serine (S), threonine (T) and tyrosine (Y) residues. We have employed a relatively balanced ratio between the positive and negative samples during the training of the classifiers (e.g. the ratio of positives versus negatives is controlled at 1:1 or 1:2 or 1:3) based on Support Vector Machine to predict phosphorylation PTM sites. Finally, we have proposed a predictor combining CKSAAP encoding, Wilcoxon-signed rank test, 1:3 ratioof positives versus negatives samples of windows and Support Vector Machine (SVM) classifier for prediction of phosphorylation sites of candidate proteins. Our data analysis results showed that the proposed method outperform over the existing predictors (such as, CKSAAP_PhSite, DISPHOS, PPRED, NetPhos) of phosphorylation sites. In order to evaluate the performance of the proposed PTM Sites predictors, four measurements will be used: sensitivity (Sn), specificity (Sp), accuracy (Ac) and Matthew correlation coefficient (MCC). The proposed predictor exhibited the performance with accuracy (Ac) 98.65% (Sn =82.94%, Sp= 99.63%, MCC=0.873) for S, accuracy (Ac) 98.81% (Sn = 91.52%, Sp =99.19%, MCC= 0.877) for T and accuracy (Ac) 97.83% (Sn =97.62%, Sp =99.38%, MCC= 0.909) for Y at 10% false positive rate.Thus the conclusion derived from this paper might help to understand the phosphorylation mechanism more accurately. It would be help to decrease in the overall cost and time period for disease diagnosis and drug or vaccine discovery.

Reference

[1] Ashis, K. B., Nasimul, N. and Abdur, R. S. (2010). Machine learning approach to predict protein phosphorylation sites by incorporating evolutionary information. BMC Bioinformatics 11: 273.

[2] Blom, N., Gammetltoft, S. and Brunak, S. (1999). Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. Journal of Molecular Biology 294: 1351–1362.

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

[4] Bu, Y. H., He, Y. L., Zhou, H. D., Liu, W., Peng, D., Tang, A. G., Tang, L. L., Xie, H., Huang, Q. X., Luo, X. H. and Liao, E. Y. (2010). Insulin receptor substrate 1 regulates the cellular differentiation and the matrix metallopeptidase expression of preosteoblastic cells. J Endocrinol 206: 271–277.

[5] Chen, Y. Z., Tang, Y. R., Sheng, Z.Y. and Zhang. Z. D. (2008). Prediction of mucin-type O-glycosylation sites using the composition of k-spaced amino acid pairs. BMC bioinformatics 9: 101.

[6] Diella, F., Cameron, S., Gemund, C., Linding, R., Via A, et al. (2004). Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins.

[7] Gnad, F., Ren, S., Cox, J., Olsen, J. V., Macek, B., et al. (2007). PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biol 8: R250.

[8] Heazlewood, J. L., Durek, P., Hummel, J., Selbig, J., Weckwerth, W., et al. (2008). PhosPhAt: a database of phosphorylation sites in Arabidopsis thaliana and a plant-specific phosphorylation site predictor. Nucleic Acids Res 36: D1015–D1021

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

[10] Jafari, P. and Azuaje, F. (2006). An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. BMC Med. Inform. Decis. Mak., 6, 27.

[11] Kim, S. H. and Lee, C. E. (2011). Counter-regulation mechanism of IL-4 and IFN-a signal transduction through cytosolic retention of the pY-STAT6: pYSTAT2: p48 complex. Eur J Immunol 41: 461–472.

[12] Lakoucheva, L., Radivojac, P., Brown, C., Oconnor, T., Sikes, J., et al. (2004). The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Research 32: 1037.

[13] Lakoucheva, L., Radivojac, P., Brown, C., Oconnor, T., Sikes, J., et al. (2004). The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Research 32: 1037.

[14] Obenauer, J., Cantley, L. and Yaffe, M. (2003). Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Research 31: 3635–3641

[15] Plewcznski, D., Tkacz, A., Wyrwicz, L. and Rychlewsk, L. (2005). AutoMotif server: prediction of single residue post-translational modifications in proteins. Bioinformatics 21: 2525.

[16] Ressurreico, M., Rollinson, D., Emery, A. M. and Walker, A. J. (2011). A role for p38 MAPK in the regulation of ciliary motion in a eukaryote. BMC Cell Biol 12: 6.

[17] Uddin, S., Lekmine, F., Sassano, A., Rui, H., Fish, E. N. and Platanias, L. C. (2003). Role of Stat5 in type I interferon-signaling and transcriptional regulation. BiochemBiophys Res Commun 308: 325–330.

[18] Wang, Y. Y., Chen, S. M. and Li, H. (2010). Hydrogen peroxide stress stimulates phosphorylation of FoxO1 in rat aortic endothelial cells. ActaPharmacol Sin 31:160–164.

[19] Wood, C. D., Tina, M. T., Guadalupe, S., Roger, A. D. and Mercedes, R. (2009). Nuclear localization of p38MAPKin response to DNA damage. Int J Biol Sci 5: 428–437.

[20] XB, Wu LY, Wang, Y. C. and Deng, N. Y. (2009). Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs. Protein Engineering Design and Selection 22: 707–712

[21] Zhao, X., Zhang, W, Xu, X., Ma, Z. and Yin, M. (2012). Prediction of Protein Phosphorylation Sites by Using the Composition of k-Spaced Amino Acid Pairs. PLoS ONE 7(10):e46302. doi:10.1371/journal.pone.0046302