Exploiting Two-Layered Support Vector Machine to Predict Phosphorylation Sites on Virus Proteins

Cheng-Tsung Lu, Kai-Yao Huang, Neil Arvin Bretaña, Wen-Chi Chang, and Tzong-Yi Lee

***Abstract*—Protein phosphorylation in viruses plays crucial regulatory roles in enhancing progression, replication, and inhibition of host cell functions. Due to the difficulty of mass spectrometry-based identification of viral phosphorylation sites, we are motivated to develop a new method to investigate the substrate motifs and identify protein phosphorylation sites on viruses. The experimentally verified phosphorylation data were extracted from a public resource and a recursively statistical method is applied to cluster whole data set of phosphorylated sequences into subgroups containing remarkably sequence motifs around the phosphorylation sites. Two-layered Support Vector Machine (SVM) is then applied to learn a predictive model by integrating the detected sequence motifs. A five-fold cross validation evaluation on the SVM model yields an average accuracy of 0.88 for Serine and 0.83 for Threonine. Furthermore, the independent testing data collected from UniProtKB and Phospho.ELM indicates that the proposed method is comparable with three popular kinase-specific phosphorylation site prediction tools. The cross validation and independent testing demonstrated that the sequence motifs are informative for the prediction of potential kinases for virus protein phosphorylation sites. Furthermore, the proposed method is a practical means of preliminary analysis for virus phosphorylation dynamics.**

***Index Terms*—Virus, protein phosphorylation, substrate motif, support vector machine.**

I. Introduction

A virus is a biological agent that is capable of interrupting and manipulating normal cellular functions [1]. They infect humans and progress inside the body leading to various diseases and complications. Most viruses interact with its host-cell proteins in order to gain control of its cellular machinery [2]. By perturbing the cellular regulatory networks, these viruses interfere with the normal cellular processes, such as cell growth and gene expression [3]. Viruses have been reported to undergo the process of phosphorylation by host-cell kinases as a means of enhancing replication and inhibition of normal cellular functions. Protein phosphorylation is one of the well-studied post-translational modification (PTM) in eukaryotic cells [4]. The process involves the transfer of a phosphate group by a protein kinase to a target protein substrate – commonly on Serine (S), Threonine (T), and Tyrosine (Y) residues [5]. Protein kinases recognize short linear motifs for initiating phosphorylation. These linear motif signatures are shown to be vital in further investigating kinase-substrate interactions [6]. Short linear motif signatures found in phosphorylated virus proteins can be used to further elucidate interactions between host-cell kinase and virus protein substrates. Although not yet clearly elucidated, these interactions are linked to viral progression in the human body.

A better understanding of virus phosphorylation is essential due to its importance with regard to viral progression. The identification of kinases is deemed important as these are heavily pursued pharmaceutical targets due to their mechanism role in various diseases [7]. Moreover, identifying kinases responsible for phosphorylation would be beneficial for selective inhibition therapies and the development of kinase inhibitors for treatment. However, there is a great deal of difficulty in experimentally identifying virus phosphorylation sites using mass spectrometry-based techniques. Furthermore, most studies that experimentally identify virus phosphorylation sites do not include the consideration of its corresponding substrate site specificities of catalytic kinases [8]. With regard to the current state of research in virus phosphorylation, this study aims to further analyze experimentally identified virus phosphorylation sites. We present a novel method for identifying phosphorylation sites and its substrate motifs. First, experimentally verified phosphorylation sites are obtained. A statistical method is then employed to detect virus phosphorylation substrate motifs. Next, support vector machine (SVMs) are trained according to the MDD-identified substrate motifs.

II. Material and Methods

A. Data Construction

With an attempt to maintain the genuineness of the data set, only literature-based virus phosphorylation data are collected from virPTM version 1.0 which contains 329 experimentally verified phosphorylation data on 111 virus proteins. As this study aims to analyze substrate specificities of viral protein phosphorylation sites, virPTM entries annotated as phosphorylated by virus kinases were disregarded in this study. This resulted to 233, 54, and 14 phosphorylated S, T, and Y sites from 104 virus proteins as shown in Table I. In UniProtKB, the experimentally verified virus phosphorylation data are obtained by filtering out entries annotated as by similarity, potential, and probable resulting to 57 phosphorylation data on 23 virus proteins. The collected data is further refined by removing entries annotated to be phosphorylated by virally-encoded kinases

Manuscript received March 15, 2013; revised May 25, 2013. We sincerely appreciated the National Science Council of the Republic of China for financially supporting this research under Contract Number of NSC 101-2628-E-155-002-MY2.

Cheng-Tsung Lu, Kai-Yao Huang, Neil Arvin Bretaña, and Tzong-Yi Lee are with the Department of Computer Science and Engineering, Yuan Ze University, 135 Yuan-Tung Road, Chungli, Taoyuan 32003, Taiwan (e-mail: francis@saturn.yzu.edu.tw).

Wen-Chi Chang is with the Institute of Tropical Plant Sciences, National Cheng Kung University.

resulting to 43, and 12 phosphorylated S, and T sites from 22 virus proteins as shown in Table I. Another set of virus phosphorylation data are collected from Phospho.ELM [9] version 0910 containing 42575 phosphorylated protein entries from 47 species. Experimentally verified virus phosphorylation data in humans are obtained by extracting only entries annotated as LTP which stands for as having been identified by using low-throughput processes.

In order to avoid obtaining overlapping phosphorylation data from the three databases, each data obtained from one database is compared to the data obtained from the remaining two using the phosphorylation site position and the UniProtKB accession number utilized by all three databases. If duplicate data is found from two or more datasets, only one record is retained and the redundant data is removed. This resulted to the same number of phosphorylated virus proteins from virPTM, 12 phosphorylated proteins with 24 pSer, and 10 pThr from UniProtKB as well as 4 phosphorylated proteins with 2 pSer, and 2 pTyr from Phospho.ELM.

TABLE I: STATISTICS OF DATA USED FOR THIS STUDY

Data set			pSer	pThr	pTyr
Training set	virPTM	Positive data	233	54	14
		Negative data	2588	1170	65
		Balanced negative data	233	54	14
Independent testing set	UniProtKB	Positive data	24	10	-
		Negative data	217	159	-
		Balanced negative data	24	10	-
	Phospho.ELM	Positive data	2	-	2
		Negative data	67	-	16
		Balanced negative data	2	-	2

In order to investigate the surrounding amino acids composition, with reference to KinasePhos [10], [11], sequence fragments are extracted using a window size of 11 centered on the phosphorylated residue. Fragments centered on phosphorylated residues are obtained and regarded as positive data while fragments centered on non-phosphorylated residues are regarded as negative data. As shown in Table I, 233, 54, and 14 positive S, T, and Y fragments as well as 2588, 1170, and 65 S, T, and Y negative fragments are obtained from virPTM. From the dbPTM resource, 42, 12, and 1 positive S, T, and Y fragments are obtained as well as 679, 186, and 11 negative S, T, and Y fragments. From the UniProtKB dataset, 24, and 10 positive S and T fragments are obtained as well as 217, and 159 negative S and T fragments. Furthermore, 2 positive S and Y fragments as well as 67, and 16 negative S and Y fragments are obtained from the Phospho.ELM dataset. Since the number of negative fragments is much greater than the number of corresponding positive fragments, the data is balanced. This is done in order to avoid a biased prediction performance. With reference to previous phosphorylation prediction methods [10], [12]-[15], a smaller number of negative fragments are obtained to match the number of positive fragments. This resulted to an equal number of positive and negative S, T, and Y fragments respectively in the three data sets as shown in Table I. Finally, the balanced non-redundant data from virPTM is regarded as the training set while the balanced non-redundant data from UniProtKB and Phospho.ELM are regarded as the independent testing set.

B. Motif Detection

The phosphorylated fragments from the training set are used to investigate the motif signatures in phosphorylated virus proteins. In order to explore the conserved motifs from a large data set, maximal dependence decomposition (MDD) is applied to cluster all phosphorylated fragments into subgroups that show statistically significant motifs. This is done using the model training set acquired from virPTM. MDD is a methodology that groups a set of aligned signal sequences to moderate a large group into subgroups that capture the most significant dependencies between positions. Previous studies [12], [16] have proposed the grouping of protein sequences into smaller groups prior to computationally identifying PTM sites. For this study, MDD is applied using MDDLogo [16]. MDD adopts chi-square test to evaluate the dependence of amino acid occurrence between two positions, *Ai* and *Aj*, which surround the phosphorylation site. In order to extract motifs that have conserved biochemical property of amino acids when doing MDD, we categorize the twenty types of amino acids into five groups: neutral, acid, basic, aromatic, and imino groups.

Then, a contingency table of the amino acids occurrence between two positions is constructed. The chi-square test is defined as:

$$\chi^2(A_i, A_j) = \sum_{m=1}^{5}\sum_{n=1}^{5} \frac{(X_{mn} - E_{mn})^2}{E_{mn}} \quad (1)$$

where X_{mn} represents the number of sequences that have the amino acids of group m in position Ai and have the amino acids of group n in position Aj, for each pair (Ai, Aj) with $i \neq j$. E_{mn} is calculated as $\frac{X_{mR} \cdot X_{Cn}}{X}$, where $X_{mR} = X_{m1} + \ldots + X_{m5}$, $X_{Cn} = X_{1n} + \ldots + X_{5n}$, and X denotes the total number of sequences. If a strong dependence is detected (defined as X^2 that is larger than 34.3, corresponding to a cutoff level of P=0.005 with 16 degrees of freedom) between two positions, then the process is continued as described by Burge and Karlin [17]. MDD clustering is a recursive process which divides the positive set into tree-like subgroups. When applying MDD to cluster the sequences in the positive set, a parameter, i.e., the minimum-cluster-size, should be set. If the size of a subgroup is less than the minimum-cluster-size, the subgroup will not be divided any further. The MDD process terminates until all the subgroup sizes are less than the value of the minimum-cluster-size. With reference to previous works that utilize MDD [12], [16], there exists no set values for the parameters of MDD clustering. In order to obtain an optimal minimum cluster size, MDD clustering is executed using various values. Each subgroup is represented using WebLogo [18] to graphically visualize the corresponding substrate motif. The resulting clusters are then analyzed as to whether or not they contain significant conserved motifs. Finally, resulting MDD subgroups with highly similar amino acid conservations at specific positions are further grouped together into a single cluster as shown in

the motif detection step. By combining similar clusters, the MDD clusters are further refined resulting to a non-redundant set of virus phosphorylation motifs [12].

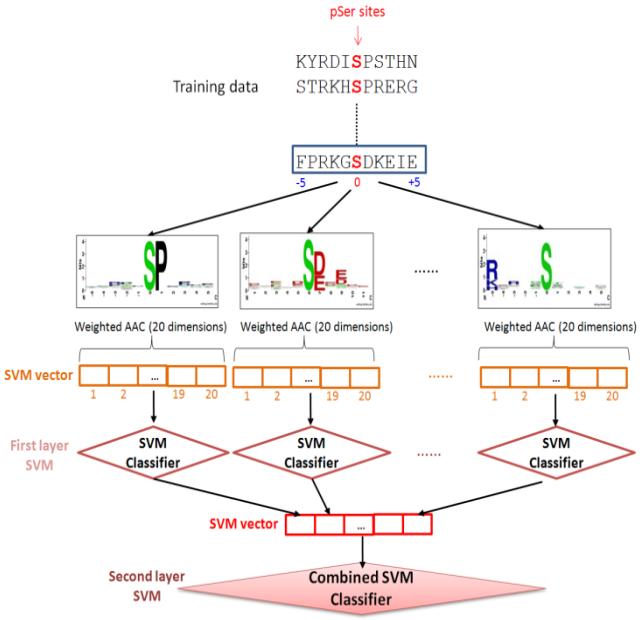


Fig. 1. The conceptual diagram of two-layered SVMs with MDD-clustered motif.

C. Model Construction and Evaluation

In this work, the support vector machine (SVM) is generated from the positive data and negative data of training set. 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 [19], is employed to train the predictive model with MDD-clustered substrate motifs which are encoded according to amino acid composition (AAC). Following, the output values of each SVM trained with the MDD-identified motif are adopted to form an input vector for second-layered SVM, as shown in Fig. 1. The radial basis function (RBF) $K(S_i, S_j) = \exp(-\gamma \|S_i - S_j\|^2)$ is adopted as the kernel function of SVM.

Prior to the construction of a final model, the predictive performance of the models with varying parameters are evaluated by performing *k*-fold cross validation. In doing *k*-fold cross validation, the training data is divided into *k* groups by splitting each dataset into *k* approximately equal sized subgroups. In one round of cross-validation, a subgroup is regarded as the test set, and the remaining *k*-1 subgroups are regarded as the training set. The cross-validation process is repeated *k* rounds, with each of the *k* subgroups used as the test set in turn. Then, the k results are combined to produce a single estimation. The advantage of *k*-fold cross-validation is that all original data are regarded as both training set and test set, and each data is used for testing exactly once [20]. In this study, *k* is set to five. The impact of using the following features: amino acid sequence, accessible surface area, and secondary structure, is evaluated by five-fold cross-validation to determine which features are best utilized to establish models that can effectively differentiate between phosphorylation sites and non-phosphorylation sites. The following measures of predictive performance of the trained models are defined: Sensitivity (*Sn*) = *TP* / (*TP*+*FN*), Specificity (*Sp*) = *TN* / (*TN*+*FP*), Accuracy [3] = (*TP* + *TN*) / (*TP*+*FP*+*TN*+*FN*), and Matthews Correlation Coefficient

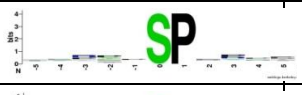
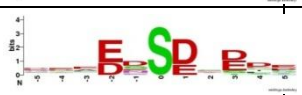
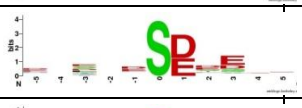
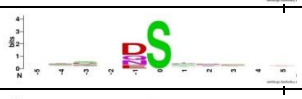
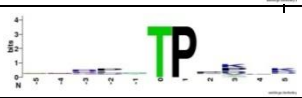
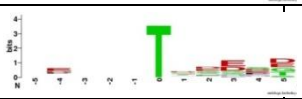
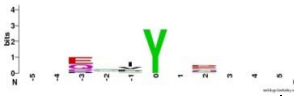
$$(MCC) = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}$$

where *TP*, *TN*, *FP* and *FN* represent the numbers of true positives, true negatives, false positives and false negatives, respectively. Subsequent to the construction of the predictive model, an independent test using the data set obtained from UniProtKB and Phospho.ELM is carried out to further evaluate the predictive performance of each SVM.

III. Results and Discussion

A. Sequence Motifs of Viral Phosphorylation Sites

TABLE II: MDD-Identified Motifs of Virus Phosphorylation Data

Residue	MDD Cluster	Motif	Fragments
pSer	S1		66
	S2		17
	S3		37
	S4		34
	S5		20
pThr	T1		19
	T2		19
pTyr	Y1		9

Phosphorylated sequences in each MDD-clustered subgroup show a conserved motif, which represents particular substrate site specificity. The flanking amino acids (-5 ~ +5) of the non-redundant phosphorylation sites, which are centered on position 0, are graphically visualized as sequence logos using WebLogo. Maximal dependence decomposition is executed multiple times with varying values in order to obtain the most optimal minimum cluster size. Setting the minimum cluster size to 50 for pSer data yielded five clusters as shown in Table II. Increasing the minimum cluster size did not result to any clusters and further lowering of the minimum cluster size resulted to several similar clusters; therefore, the minimum cluster size is set to 50. After MDD, further refinement is done by analyzing these

groups through its corresponding entropy plots. It is observed that some groups contain very similar motifs, some show no conserved motif, and some groups have too little data which makes the motif unreliable. Some of these groups are further combined together and visualized using WebLogo. For the resulting pSer MDD clusters, S2 and S3 which shows very similar motifs are combined into S2.

For virus pThr and pTyr data, the minimum cluster size was set to 10. Similar to the process of selecting the minimum cluster size for pSer, increasing the minimum cluster size did not result to any clusters and further lowering of the minimum cluster size resulted to several similar clusters. This resulted to two subgroups in pThr and one subgroup in pTyr as shown in Table II. However, due to the very low number of pTyr data, the resulting MDD clusters show no conserved motif and contain very few fragments to be considered reliable. Therefore, for this study, pTyr is not further clustered using MDD prior to training a pTyr model. MDD could identify new motifs for virus phosphorylation sites and is comparable to other methods.

B. Cross-Validation Evaluation of SVM Models

TABLE III: FIVE-FOLD CROSS VALIDATION RESULTS ON PSER MDD-CLUSTERED SVMS

SVM model	Pos.	Neg.	C	G	Sn	Sp	Acc	MCC
All data	233	233	0.5	0.125	0.76	0.72	0.74	0.49
Subgroup S1	66	66	2	0.125	0.98	0.87	0.93	0.71
Subgroup S2	54	54	8	0.03125	0.94	0.92	0.93	0.74
Subgroup S3	34	34	0.5	0.03125	0.93	0.79	0.85	0.61
Subgroup S4	20	20	2	0.125	0.9	0.84	0.88	0.63
Subgroup S5	15	15	2	0.125	0.88	0.82	0.84	0.62
Combined performance	-	-	-	-	**0.9**	**0.85**	**0.88**	**0.68**

*Pos. : Number of positive data; *Neg. : Number of negative data; *C : Cost value; *G : Gamma value.

The cross-validation process includes the selection of the threshold parameter for each model. The threshold is tuned to a specific value which allows an SVM to yield a high and balanced Specificity and Sensitivity for a specific classification. Table III shows the threshold score selected for each model of pSer together with its individual predictive performance and the predictive performance of all MDD-clustered models. It can be observed that MDD clusters featuring an obvious conserved motif are able to yield high predictive accuracies. For instance, cluster S1 containing a conserved Proline residue in positions +1 yields an accuracy of 0.93 when used individually. On the other hand, MDD clusters that do not seem to have an obvious conserved motif yield a significantly lower predictive performance. For instance, cluster S6 which does not show a strongly conserved motif based on its WebLogo only yields an accuracy of a 0.68 when used individually. According to a five-fold cross-validation evaluation, the predictive performance of MDD-clustered SVMs performs significantly better than non-MDD clustered SVM in overall. As shown in Table III, the SVM model trained with the combined MDD-clustered motifs yields a higher performance with a sensitivity of 0.90, a specificity of 0.85, an accuracy of 0.88, and a MCC of 0.68 as compared to the SVM with all pSer data which yields a sensitivity of 0.76, a specificity of 0.72, an accuracy of 0.74, and a MCC of 0.49.

Table IV shows the threshold score selected for each model of pThr together with its individual predictive performance and the predictive performance of using all models together. The pThr SVM model trained with the combined MDD-clustered motifs yields a higher performance with a sensitivity of 0.82, a specificity of 0.84, an accuracy of 0.83, and a MCC of 0.65 as compared to the SVM model with all pThr data which yields a sensitivity of 0.71, a specificity of 0.71, an accuracy of 0.71, and a MCC of 0.47. Due to a lack of virus pTyr data, MDD clustering could not be performed to form SVM model for computationally identifying pTyr sites; thus, a single SVM is used for pTyr until sufficient experimentally-verified virus pTyr sites are acquired. The SVM models containing the best predictive performance could be used to construct the prediction tool of virus phosphorylation sites.

TABLE IV: FIVE-FOLD CROSS VALIDATION RESULTS ON PTHR MDD-CLUSTERED SVMS

SVM model	Pos.	Neg.	C	G	Sn	Sp	Acc	MCC
All data	54	54	2	0.125	0.71	0.71	0.71	0.47
Subgroup T1	19	19	2	0.125	0.95	0.91	0.94	0.75
Subgroup T2	19	19	2	0.03125	0.98	0.96	0.97	0.81
Combined performance	-	-	-	-	**0.82**	**0.84**	**0.83**	**0.65**

*Pos. : Number of positive data; *Neg. : Number of negative data; *C : Cost value; *G : Gamma value.

C. Independent Testing

The data set obtained from UniProtKB and Phospho.ELM which do not have overlapping data with the training set is utilized for further evaluating the MDD-clustered SVMs. A total of 36 viral protein phosphorylation sites (in 23 viral protein sequences), which are not included in the training data, are regarded as the positive set of the independent test data. In order to evaluate the predictive specificity, the S and T residues, which are not annotated as the phosphorylation sites in the 23 viral protein sequences, are regarded as the negative set of the independent testing data. As a result, the independent testing data consisting of 36 positive sites and 474 negative sites are used to compare the predictive sensitivity, specificity and accuracy between the Single SVM and MDD-clustered SVMs. As shown in Fig. 2A, the SVM model trained with all pSer data (Single SVM) yields a sensitivity of 0.54, a specificity of 0.66, an accuracy of 0.60, and the MCC of 0.29. Additionally, using all the pSer MDD-clustered SVMs altogether yields a sensitivity of 0.92, a specificity of 0.79, an accuracy of 0.86, and the MCC of 0.61. On the other hand, Fig. 2B shows that using the independent data on Single pThr SVM model yields a sensitivity of 0.64, a specificity of 0.82, an accuracy of 0.73, and the MCC of 0.38. Furthermore, the combined model using all pThr MDD-clustered SVMs yields a sensitivity of 0.95, a specificity of 0.90, an accuracy of 0.93, and the MCC of 0.73.

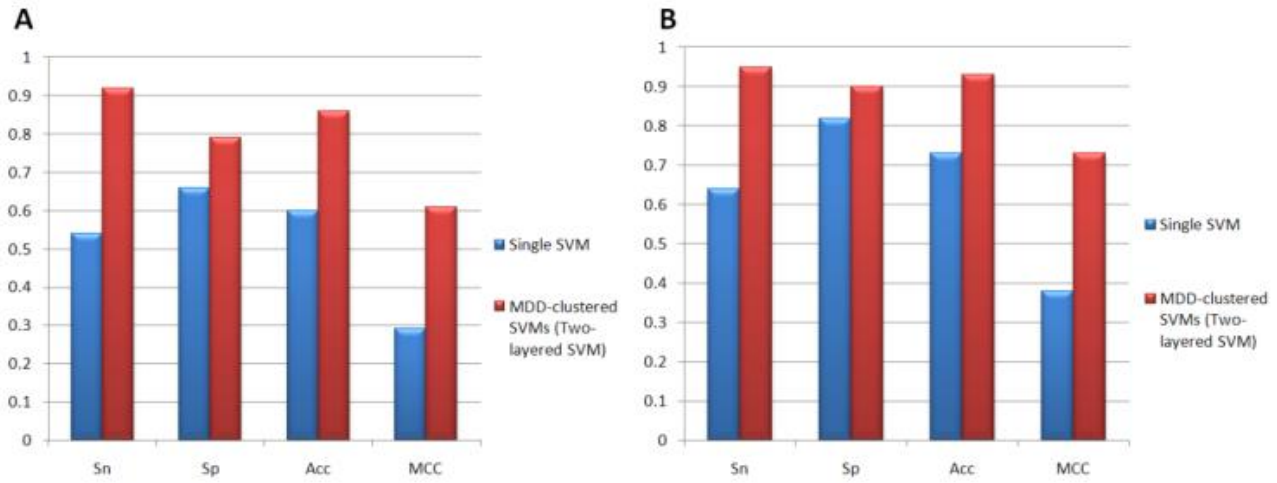


Fig. 2. Comparison of independent testing performance.

To further demonstrate the effectiveness of the proposed method, the independent testing set is used to make a comparison between the performances of three popular kinase-specific phosphorylation site prediction tools, PPSP [15], KinasePhos 2.0 [10], and GPS 2.1 [21]. Without any pr ior information of catalytic kinases for the testing data, all of the kinase-specific models in the prediction tools are chosen for predicting the phosphorylation sites. The independent testing indicates that all of the prediction tools containing multiple models have a high predictive sensitivity. However, it is notable that the proposed method is able to yield a higher specificity compared to the other tools. Since potential kinase information for viral protein phosphorylation sites are still unknown, PPSP yields a higher specificity than KinasePhos and GPS. Overall, the proposed method outperforms the other three tools. With reference to the comparison of independent testing, the high sensitivity and specificity of MDD-clustered SVMs show that the substrate site motifs are effective for the identification of viral protein phosphorylation sites.

D. Motifs Comparison

In order to identify potential host kinases for virus substrates, the motif of each MDD-generated virus phosphorylation cluster is compared with the well-discovered kinase substrate motifs of Phospho.ELM. Cluster S1 is matched to be potentially phosphorylated by CDK group and MAPK group due to a strong similarity with regard to the conserved Proline in positions +1. CK2 group is also matched to be a potential host kinase that phosphorylated virus substrates in cluster S2 due to a similarly conserved Aspartic acid and Glutamic acid residues at position +3. Furthermore, cluster S4 is matched to be potentially phosphorylated by PKB group due to a conserved Arginine in position -5 as shown in its respective motifs. In terms of pThr, cluster T1 is matched to be potentially phosphorylated by CDK group and MAPK group due to a conserved Proline in position +1. Cluster T2 is then matched to be potentially phosphorylated by CK2 group due to a similarly conserved Aspartic acid and Glutamic acid residues in position +3.

In a further investigation of the matched motifs, a literature survey was done in order to find studies that experimentally identify host kinases which phosphorylate specific virus protein substrates. Reports have been published that CDK group, especially the CDK2, is involved in the transcription and replication of Human Immunodeficiency Virus - 1 by means of phosphorylation [22], [23]. Previous studies [8], [24] also show that CK2 group phosphorylates Hepatitis C Virus NS5A proteins and Human Immunodeficiency Virus - 1 gp120, gp41, p27, and p17 proteins to name a few, on both S and T residues. These findings support the matching of MDD groups S2 and T2 with CK2 group. With regard to PKB which is matched with cluster S4, it is reported to be involved in the regulation of the Herpes Simplex virus - 1 [25]. Additionally, experimental research also claims that PKB signaling benefits coxsackie virus B3 replication [26].

IV. Conclusion

In this study, virus phosphorylation sites catalyzed by host kinases are further elucidated by means of identifying its potential substrate site specificities. According the motif comparison, this study has identified the informative motifs that matched to several well-studied kinase groups including CDK, MAPK, CK2 and PKB as potential catalytic kinases for virus protein substrates. The identified substrate motifs are further exploited to identify virus phosphorylation sites. A five-fold cross validation evaluation shows that the proposed method can identify virus phosphorylation sites based on the MDD-identified motifs. Furthermore, an independent test done using data not included in the model training confirms the ability of our MDD-clustered SVMs.

The proposed approach offers the scientific community some clues regarding host kinases that may be responsible for the phosphorylation of human virus proteins. It is important to note, however, that the further acquisition of experimentally verified virus phosphorylation sites is required to identify more meaningful virus phosphorylation motifs. Also, a more abundant set of experimentally verified kinase-catalyzed virus phosphorylation sites could be extracted by literature survey. These developments could benefit this work by allowing a more accurate identification of phosphorylation sites on virus proteins.

References

[1] D. Schwartz and G. M. Church, "Collection and motif-based prediction of phosphorylation sites in human viruses," *Sci Signal*, vol. 3, no. 137, pp. rs2, 2010.

[2] R. Zell, A. Krumbholz, and P. Wutzler, "Impact of global warming on viral diseases: what is the evidence?" *Curr Opin Biotechnol*, vol. 19, no. 6, pp. 652-60, 2008.
[3] A. Chatr-aryamontri *et al.*, "VirusMINT: A viral protein interaction database," *Nucleic Acids Res*, vol. 37 (Database issue), pp. D669-73, 2009.
[4] H. Steen *et al.*, "Phosphorylation analysis by mass spectrometry: myths, facts, and the consequences for qualitative and quantitative measurements," *Mol Cell Proteomics*, vol. 5, no. 1, pp. 172-81, 2006.
[5] D. Secko. (2003). Protein phosphorylation: A global regulator of cellular activity. *The Science Creative Quarterly*. [Online]. pp. 271. Available: http://www.scq.ubc.ca/protein-phosphorylation-a-global-regulator-of-cellular-activity/.
[6] V. Neduva and R. B. Russell, "Peptides mediating interaction networks: new leads at last," *Curr Opin Biotechnol*, vol. 17, no. 5, pp. 465-71, 2006.
[7] N. G. A. J. Olaharski, H. Bitter, D. Goldstein, S. Kirchner, H. Uppal, and K. Kolaja, "Identification of a kinase profile that predicts chromosome damage induced by small molecule kinase inhibitors," *PLoS Computational Biology*, pp. e1000446, 2009.
[8] C. Coito *et al.*, "High-throughput screening of the yeast kinome: identification of human serine/threonine protein kinases that phosphorylate the hepatitis C virus NS5A protein," *J Virol*, vol. 78, no. 7, pp. 3502-13, 2004.
[9] F. Diella *et al.*, "Phospho.ELM: A database of experimentally verified phosphorylation sites in eukaryotic proteins," *BMC Bioinformatics*, vol. 5, pp. 79, 2004.
[10] Y. H. Wong *et al.*, "KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns," *Nucleic Acids Res*, vol. 35 (Web Server issue), pp. W588-94, 2007.
[11] H. D. Huang *et al.*, "KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites," *Nucleic Acids Res*, vol. 33 (Web Server issue), pp. W226-9, 2005.
[12] T. Y. Lee, N. A. Bretana, and C. T. Lu, "PlantPhos: using maximal dependence decomposition to identify plant phosphorylation sites with substrate site specificity," *BMC Bioinformatics*, vol. 12, pp. 261, 2011.
[13] T. Y. Lee *et al.*, "RegPhos: a system to explore the protein kinase-substrate phosphorylation network in humans," *Nucleic Acids Res*, vol. 39 (Database issue), pp. D777-87, 2011.
[14] Y. Xue *et al.*, "GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy," *Mol Cell Proteomics*, vol. 7, no. 9, pp. 1598-608, 2008.
[15] Y. Xue *et al.*, "PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory," *BMC Bioinformatics*, vol. 7, pp. 163, 2006.
[16] T. Y. Lee *et al.*, "Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences," *Bioinformatics*, vol. 27, no. 13, pp. 1780-7, 2011.
[17] C. Burge and S. Karlin, "Prediction of complete gene structures in human genomic *DNA*," *J Mol Biol*, vol. 268, no. 1, pp. 78-94, 1997.
[18] G. E. Crooks *et al.*, "WebLogo: a sequence logo generator," *Genome Res*, vol. 14, no. 6, pp. 1188-90, 2004.
[19] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 27, pp. 1-27, 2001.
[20] C. T. Lu *et al.*, "Carboxylator: incorporating solvent-accessible surface area for identifying protein carboxylation sites," *J Comput Aided Mol Des*, vol. 25, no. 10, pp. 987-95, 2011.
[21] Y. Xue *et al.*, "GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection," *Protein Eng Des Sel*, vol. 24, no. 3, pp. 255-60, 2010.
[22] T. Ammosova *et al.*, "RNA interference directed to CDK2 inhibits HIV-1 transcription," *Virology*, vol. 341, no. 2, pp. 171-8, 2005.
[23] L. Deng *et al.*, "HIV-1 Tat interaction with RNA polymerase II C-terminal domain (CTD) and a dynamic association with CDK2 induce CTD phosphorylation and transcription from HIV-1 promoter," *J Biol Chem*, vol. 277, no. 37, pp. 33922-9, 2002.
[24] F. Meggio and L. A. Pinna, "One-thousand-and-one substrates of protein kinase CK2?" *FASEB J*, vol. 17, no. 3, pp. 349-68, 2003.
[25] L. Benetti and B. Roizman, "Protein kinase B/Akt is present in activated form throughout the entire replicative cycle of deltaU(S)3 mutant virus but only at early times after infection with wild-type herpes simplex virus 1," *J Virol*, vol. 80, no. 7, pp. 3341-8, 2006.
[26] M. Esfandiarei *et al.*, "Protein kinase B/Akt regulates coxsackievirus B3 replication through a mechanism which is not caspase dependent," *J Virol*, vol. 78, no. 8, pp. 4289-98, 2004.

Cheng-Tsung Lu obtained his master degree of Biomedical Informatics from Yuan Ze University. He is currently a PhD student in Department of Computer Science and Engineering, Yuan Ze University. His research interests include protein phosphorylation, protein s-nitrosylation, data mining, database system, machine learning, and bioinformatics.

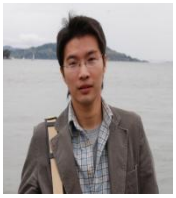
Kai-Yao Huang received his B.S. in 2010 in Department of Computer Science (CS), Chinese Culture University (PCCU), Republic of China.

Start from Summer 2011, he study for a PhD in Bioinformatic from the School of Yuan Ze University (YZU) at the Department of Computer Science & Engineering and Graduate Program in Biomedical Informatics.

His current research interests are data mining and machine learning applied to biological and medical data, molecular networks related to diseases and gene transcriptional regulation in particular.

Neil Arvin Bretaña was born in Philippine. He obtained his Master degree in graduate program of Biomedical Informatics from Yuan Ze University. He is currently a PhD student in UNSW under Fabio Luciani. His research interest is focusing on HCV pathology.

Wen-Chi Chang obtained her master and PhD at Institute of Bioinformatics and Structural Biology, National Tsing Hua University (NTHU). She is currently a professor at Institute of Tropical Plant Sciences, National Cheng Kung University (NCKU). Her research interests include plant science, gene transcriptional regulation, promoter analysis, microRNA regulatory network, next generation sequencing, and systems biology.



Tzong-Yi Lee received his B.E. and M.E. in Department of Computer Science and Information Engineering, National Central University (NCU). He obtained the Ph.D. in Institute of Bioinformatics, National Chiao Tung University (NCTU). Currently he is a professor in the Department of Computer Science and Engineering, Graduate Program in Biomedical Informatics, Yu- an Ze University. His research interests include bioinformatics, computational proteomics, systems biology, data mining and machine learning.