A New Scheme to Predict Kinase-Specific Phosphorylation Sites on Protein Three-Dimensional Structures

Min-Gang Su, Kai-Yao Huang, Chi-Hua Tung, and Tzong-Yi Lee

***Abstract*—Due to the high-throughput of mass spectrometry-based phosphoproteomics experiment, the desire to annotate the catalytic kinases for *in vivo* phosphorylation sites has motivated. Many researches are undertaken to develop a computational method for the identification of kinase-specific phosphorylation sites using linear amino acid sequences. With an increasing interest in the structural environment of protein phosphorylation sites, herein, a new scheme has been developed for identifying kinase-specific phosphorylation sites on protein three-dimensional (3D) structures. For a large-scale investigation on 3D structures, all of the experimental phosphorylation sites are mapped to the protein entries of Protein Data Bank by sequence identity. In this work, a support vector machine (SVM) is applied to generate the predictive model learned from the information of spatial amino acid composition and structural alphabet. After the cross-validation evaluation, most of the kinase-specific models trained with the consideration of structural information outperform the models considering only the sequence information. Moreover, the independent testing set which is not included in training set has demonstrated that the proposed method could provide a stable performance. This study has demonstrated that the consideration of spatial context could improve the predictive performance compared to the model only considering the local sequence motifs.**

***Index Terms*—Phosphorylation, protein kinase, three-dimensional structure, structural alphabet, spatial amino acid composition.**

I. Introduction

Protein phosphorylation catalyzed by kinases plays crucial regulatory roles in many essential cellular processes including cellular regulation, cell death, transcriptional regulation, cellular signal pathways, metabolism, growth, differentiation, and membrane transport [1]. It has been estimated that one-third to one-half of all proteins are phosphorylated in a eukaryotic cell [2] and around half of kinome are disease- or cancer-related by chromosomal mapping [3]. Mass spectrometry-based identifications of phosphorylation sites on substrates *in vivo* and *in vitro* are the foundation of understanding the mechanisms of phosphorylation dynamics and important for the biomedical drug design [4]. However, the effort to experimentally verify the catalytic kinases remains time-consuming, labor-intensive, and expensive. Consequentially, many researches are undertaken to develop a computational method for the identification of kinase-specific phosphorylation sites, including NetPhosK [5], Scansite 2.0 [6], PredPhospho [7], GPS [8]-[10], PPSP [4], MetaPredPS [11], NetPhorest [12] and KinasePhos [13]-[15]. Particularly, Linding *et al.* [16] have proposed an excellent method, namely NetworKIN, that augments motif-based predictions with the network context of kinases and phosphoproteins. With most of the existing phosphorylation site prediction tools requiring prior knowledge of experimentally verified substrates and its kinase, a method is developed to be able to predict kinase-specific phosphorylation sites based solely on protein sequence [17].

Although over 20 methods have been developed for the accurate prediction of kinase-specific phosphorylation sites, most of them rely solely on the local amino acid sequence surrounding the phosphorylated sites. Blom *et al.* [18] were the first to propose a method with limited data for sequence and structure-based prediction of protein phosphorylation sites in eukaryotes. While one-dimensional amino acid sequence was observed to harbor most of the predictive power, Predikin [19] has proposed a method that applied the structure-based information for improving the prediction of phosphorylation sites in proteins. With an increasing interest in the structural environment of protein phosphorylation sites, Phospho3D database [20], [21] was proposed for characterizing the structural properties of phosphorylation sites on three-dimensional (3D) structures. Additionally, Phos3D [22] has extracted 3D-signature motifs from 750 experimentally verified phosphorylation sites with 3D structures available in Protein Data Back (PDB) [23] and applied them to implement a web server for structure-based detection of phosphorylation sites.

With the desire to comprehensively and accurately annotate the catalytic kinases for *in vivo* phosphorylation sites, this work has developed a new scheme for identifying kinase-specific phosphorylation sites on 3D structures. To investigate the spatial environment of phosphorylation sites, all of the experimental phosphorylation sites are mapped to the PDB protein entries using sequence identity. In this work, the linearly sequenced substrate motifs are combined with the information of spatial amino acid composition and structural alphabet, which is a new scheme for encoding a 3D structure fragment of protein backbones into 23 structural alphabets, to identify kinase-specific phosphorylation sites on 3D structures. Moreover, an independent testing set which is blind to the cross-validation process has been generated for

Manuscript received February 25, 2013; revised April 25, 2013. This work sincerely appreciates 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 and 101-2221-E-216-041.

Min-Gang Su and Tzong-Yi Lee are with the Department of Computer Science and Engineering, Yuan Ze University, 135 Yuan-Tung Road, Taoyuan, Taiwan. (e-mail: francis@saturn.yzu.edu.tw).

Chi-Hua Tung is with the Department of Bioinformatics, Chung Hua University, Hsinchu 300, Taiwan (e-mail: chihua.tung@chu.edu.tw).

DOI: 10.7763/IJBBB.2013.V3.258

the evaluation of stability and reliability of our method.

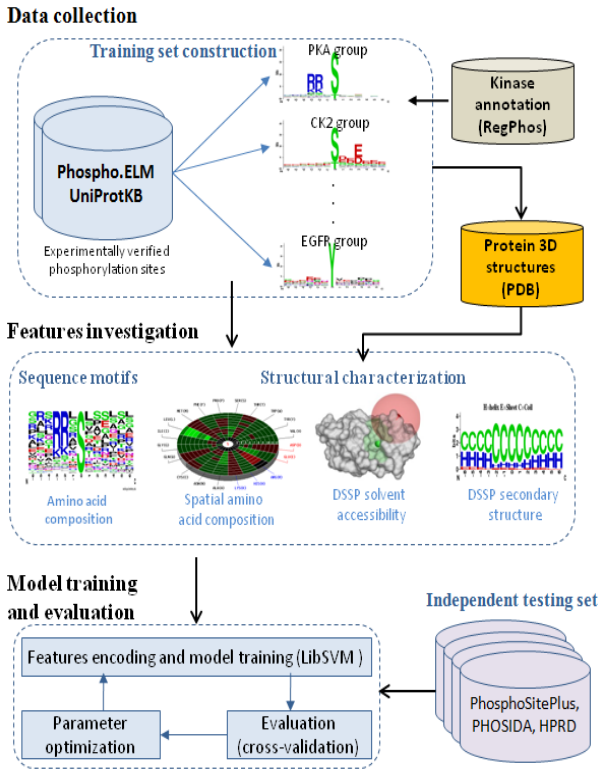


Fig. 1. The analytical flowchart of this study.

II. Material and Methods

TABLE I: Data Statistics of Experimentally Verified Phosphorylation Sites in each Resource

Data set	Data Resource	Number of phosphorylation sites		
		S	T	Y
Training set	Phospho.ELM	26,136	6,316	3,118
	UniProtKB	92,221	23,289	14,337
	Combined (NR[1])	98,376	25,269	15,188
Independent testing set	PhosphoSitePlus	73,969	19,946	14,696
	PHOSIDA	7,391	1,300	278
	HPRD	34,273	10,761	4,121
	Combined (NR[1])	97,753	27,421	16,531

Fig. 1 depicts the analytical flowchart including data collection, features investigation, model training and evaluation, and independent testing. The experimentally verified phosphorylation sites are mainly extracted from dbPTM [24], [25] which has integrated the data from version 9.0 of Phospho.ELM [26], release 20120711 of UniProtKB [27], release 20120730 of PhosphoSitePlus [28], version 1.0 of PHOSIDA [29], version 1.1 of SysPTM [30] and version 9.0 of HPRD [31]. In this work, the data set extracted from Phospho.ELM and UniProtKB is regarded as the training set for sequenced and structural investigation of phosphorylated substrate sites. After removing the redundant sites between Phospho.ELM and UniProtKB, the number of serine (S), threonine (T), and tyrosine (Y) substrate sites are 98376, 25269, and 15188, respectively, as given in Table I. According to the annotations of kinase families extracted from RegPhos [32], the substrate sites of protein phosphorylation could be further categorized into more than 200 kinase groups.

As for classification, the prediction performance of the constructed models may be overestimated owing to the over-fitting of a training set. The experimental phosphorylation sites that collected from PhosphoSitePlus, PHOSIDA and HPRD were regarded as the independent testing set.

A. Sequence-Based Investigation of Phosphorylation Sites

Since the flanking sequences of the substrate sites (position 0) are graphically visualized as the entropy plots of sequence logo [33], the conservation of amino acids surrounding the phosphorylation sites could be easily observed. The 13-mer sequences (from -6 to +6) of kinase-specific phosphorylation sites are extracted as the positive data of training sets, while all other residues (S, T and Y) in the phosphorylated proteins are regarded as the negative data. With reference to the method of SulfoSite [34], the positional weighted matrix (PWM), which specifies the relative frequency of amino acids surrounding substrate sites, was utilized in encoding the fragment sequences. A matrix of $m \times w$ elements was used to represent each residue of a training dataset, where w stands for the window size and m consists of 21 elements including 20 types of amino acids and one for terminal signal.

B. Structural Characterization of Phosphorylation Sites

With an attempt to study the spatial context of phosphorylation sites and evaluate its effectiveness for improving the predictive performance, all of the collected phosphorylation sites are mapped to the protein entries of Protein Data Bank (PDB) by sequence identity. It resulted in a total of 4508 phosphorylation sites (covering over 40 kinase groups) containing the protein 3D structures. DSSP [35] is then utilized to calculate the solvent accessibility and standardize the secondary structure of PDB entries with the mapped phosphorylation sites. Instead of the sequential amino acid composition (AAC), this work investigates the propensities for the different amino acid types to occur in the spatial vicinity of the phosphorylated sites. A spatial amino acid composition (Spatial AAC) is determined for each kinase grov ups by calculating the relative frequencies of 20 amino acid types within radial distances ranging from 3 to 12 Å from central phosphorylated amino acid residue. A radial cumulative propensity plot [22] was applied to display the spatial AAC. In order to identify the significant difference of spatial AAC between phosphorylation sites (positive data) and non-phosphorylation sites (negative data), a measurement of F-score [36] has been applied to calculate a statistical value for each radial distance. The F-score of the ith value of 11 radial distances is defined as:

$$\text{F-score}\,(i) = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n^+ - 1}\sum_{k=1}^{n^+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n^- - 1}\sum_{k=1}^{n^-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2} \quad (1)$$

where $\bar{x}_i$, $\bar{x}_i^{(+)}$ and $\bar{x}_i^{(-)}$ denote the average value of the ith distance value in whole, positive, and negative data sets, respectively; n^+ denotes the number of positive data set and n^- denotes the number of negative data set; $x_{k,i}^{(+)}$ denotes the ith distance value of the kth positive instance, and $x_{k,i}^{(-)}$ denotes the ith distance value of the kth negative instance [36].

In order to generate a predictive model that physically utilizes the local structures of phosphorylation sites, the kappa-alpha (κ, α) plot derived structural alphabet (SA) has been developed to encode the three-dimensional (3D) local structural fragments based on C_a coordinates into a one-dimensional (1D) representation based on 23 letters. These letters are used to describe three-state secondary elements, namely helix, strand, and random coils in detail, and identify protein 1D sequence that represent the local structures of protein backbones with 3D information. We defined the local structure with five residues long as a segment and various segments are encoded into 23 SAs based on their kappa (κ) and alpha (α) angles. The angle κ, ranging from 0 ° to 180 °, of residue *i* is a bond angle formed by three C_a atoms of residues *i-2*, *i* ,and *i+2*. The angle α, ranging from -180 ° to 180 °, of a residue *i* is a dihedral angle formed by the four C_a atoms of residues *i-1*, *i*, *i +1*, and *i +2*. These 23 SAs are roughly divided into five categories: Helix alphabets (A, Y, B, C, and D), helix-like alphabets (G, I, and L), strand alphabets (E, F, and H), strand-like alphabets (K and N), and others. A recent success with this approach was the development of the tool and web-site, named as 3D-BLAST [37], [38] and fastSCOP [39], for the rapid structure database searching and recognition of protein structural domains.

C. Model Training and Evaluation

This work incorporates support vector machine (SVM) with the sequenced and structural features to generate the predictive models for the identification of kinase-specific phosphorylation sites. Based on binary classification, the concept of SVM is to map the input samples into a higher dimensional space using a kernel function, and then to find a hyper-plane that discriminates between the two classes with maximal margin and minimal error. A public SVM library, namely LIBSVM [40], is applied for training the predictive models. The radial basis function (RBF) $K(S_i, S_j) = \exp(-\gamma \left\| S_i - S_j \right\|^2)$ is selected as the kernel function of SVM. Five-fold cross-validation is used to evaluate the predictive performance of the models trained from the large data sets such as PKA, PKC, CK2, and MAPK groups, while Jackknife cross-validation is applied for models trained from the data size smaller than 30 substrate sites. We balance the positive set and negative set and the sizes of positive data and negative data are equal during the cross-validation processes. The cross-validation is performed for ten times to obtain an average accuracy for each kinase group. The following measures of predictive performance of the trained models are defined: Precision (Pre) = TP / (TP+FP), Sensitivity (Sn) = TP / (TP+FN), Specificity (Sp) = TN / (TN+FP) and Accuracy (Acc) = (TP + TN) / (TP+FP+TN+FN), where TP, TN, FP and FN are true positive, true negative, false positive and false negative predictions, respectively. The models trained with various features that yield the highest accuracy in each kinase group are utilized to implement the prediction system and are further evaluated by independent testing set. For a meaningful comparison with other published tools, the ratio of data size between positive set and negative set is 1:2 [22].

III. Results and Discussions

A. Sequenced and Structural Amino Acid Composition of Phosphorylation Sites Involved in Different Kinases

Most of the kinase groups have conserved amino acids surrounding the phosphorylation sites. The solvent accessibility and secondary structure computed from a full-length protein sequence are also presented. With the comprehensive mapping between the collected phosphorylation data and PDB protein 3D structures, the spatial environment of phosphorylation sites was investigated in detail, as well as the sequential neighborhood. Fig. 2 shows the sequence logos (sequential neighborhood) and radial cumulative propensity plots (spatial neighborhood) of six well-known kinase-specific substrate groups.

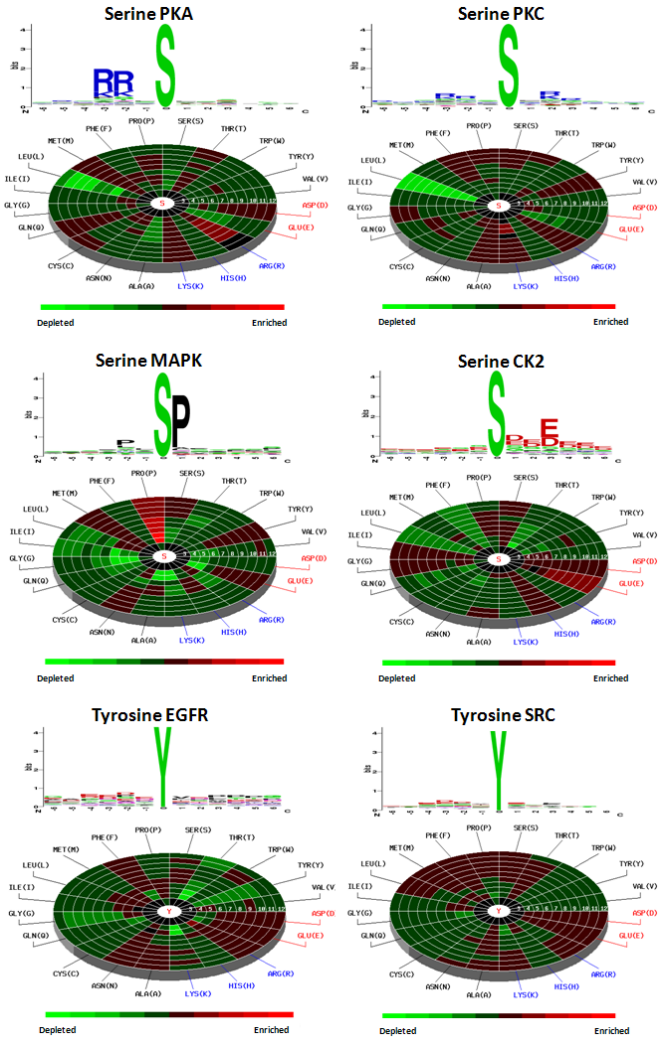


Fig. 2. Sequence logos and radial cumulative propensity plots of six kinase-specific substrate groups.

According to the observation from sequence logos, PKA and PKC have the significant enrichments of Arginine (R) and Lysine (K) in the sequential neighborhood of substrate sites, which is the hallmark sequence motif for AGC kinase families. However, the radial cumulative propensity plots present that there is an additional enrichment of amino acid residues in the spatial neighborhood. For instance, PKA exhibits the enrichments of Methionine (M), Glutamine (Q) and Aspartic acid (D) in the spatial neighborhood, accompanied by a remarkable depletion of Leucine (L) residue. For tyrosine kinase families, both EGFR and SRC

have the enrichments of Aspartic acid (D) and Glutamic acid (E) in the sequential and spatial neighborhood. In particular, EGFR has a significant depletion of Threonine (T) according to the radial cumulative propensity plot, but SRC is enriched in T residue instead. In summary, the radial cumulative propensity plot reveals spatial preferences of amino acids composition which cannot be identified by inspecting the sequence logo alone.

TABLE II: CROSS-VALIDATION EVALUATION OF SEQUENCE AND STRUCTURE-BASED PHOSPHORYLATION SITE PREDICTIONS ON 3D STRUCTURES

Kinase group	Number of positive data	Number of negative data	Sequence-only			Structural information			Combination of sequence and structural information		
			Sn	Sp	Acc	Sn	Sp	Acc	Sn	Sp	Acc
Phosphorylated Serine (pSer)											
All serine data	1554	3108	61.4%	62.0%	61.8%	66.9%	68.1%	67.7%	**72.9%**	**71.1%**	**71.7%**
CDK	11	22	72.7%	81.8%	78.8%	**90.9%**	**86.8%**	**87.9%**	90.9%	86.8%	87.9%
CK1	10	20	20.0%	90.0%	66.7%	**100%**	**95.0%**	**96.7%**	100%	95.0%	96.7%
CK2	24	48	66.7%	87.5%	80.6%	87.5%	87.5%	87.5%	**91.7%**	**89.6%**	**90.3%**
MAPK	17	34	52.9%	94.1%	80.4%	76.5%	97.1%	90.2%	**82.4%**	**97.1%**	**92.2%**
PIKK	15	30	26.7%	83.3%	64.4%	**80.0%**	**86.7%**	**84.4%**	73.3%	83.3%	80.0%
PKA	56	112	79.1%	78.8%	78.9%	83.6%	84.3%	84.1%	**89.1%**	**91.4%**	**90.7%**
PKB	12	24	75.0%	66.7%	69.4%	75.0%	83.3%	80.6%	**83.3%**	**83.3%**	**83.3%**
PKC	50	100	77.3%	78.0%	77.8%	81.2%	80.0%	80.4%	**85.3%**	**86.0%**	**85.8%**
PKG	10	20	80.0%	80.0%	80.0%	**80.0%**	**85.0%**	**83.3%**	80.0%	85.0%	83.3%
PLK	10	20	60.0%	80.0%	73.3%	**70.0%**	**90.0%**	**83.3%**	70.0%	90.0%	83.3%
STE20	10	20	70.0%	75.0%	73.3%	**80.0%**	**90.0%**	**86.7%**	80.0%	90.0%	86.7%
Phosphorylated Threonine (pThr)											
All Threonine data	603	1206	60.9%	59.7%	60.1%	67.8%	67.2%	67.4%	**70.1%**	**72.5%**	**71.3%**
MAPK	13	26	**69.2%**	**76.9%**	**74.3%**	69.2%	76.9%	74.3%	69.2%	76.9%	74.3%
PKA	10	20	70.0%	90.0%	83.3%	80.0%	85.0%	83.3%	**80.0%**	**95.0%**	**90.0%**
PKC	13	26	61.5%	76.9%	71.8%	**69.2%**	**88.5%**	**82.1%**	69.2%	88.5%	82.1%
STE20	10	20	40.0%	95.0%	76.7%	70.0%	70.0%	70.0%	**70.0%**	**90.0%**	**80.0%**
Phosphorylated Tyrosine (pTyr)											
All tyrosine data	629	1258	62.0%	63.3%	62.8%	64.1%	63.4%	63.8%	**67.6%**	**68.6%**	**68.3%**
Abl	18	36	50.0%	88.9%	75.9%	**66.7%**	**80.6%**	**75.9%**	66.7%	80.6%	75.9%
EGFR	10	20	60.0%	80.0%	73.3%	**60.0%**	**95.0%**	**83.3%**	60.0%	95.0%	83.3%
InsR	15	30	73.3%	83.3%	80.0%	80.0%	80.0%	80.0%	**80.0%**	**90.0%**	**86.7%**
Src	57	114	77.2%	75.4%	76.0%	79.1%	83.3%	81.9%	**79.1%**	**84.9%**	**82.9%**
Syk	11	22	63.6%	90.9%	81.8%	72.7%	86.4%	81.8%	**72.7%**	**95.5%**	**87.9%**

B. Predictive Performance of Kinase-Specific SVM Models

For finding the best predictive performance of SVM models in each kinase-specific group, the SVM models trained with linear sequence motifs or structural characteristics are evaluated based on cross-validation. To obtain a stable performance for each kinase-specific prediction models, the cross-validation process is performed for ten times and the average sensitivity (Sn), specificity (Sp), and accuracy (Acc) of the SVM models. As given in Table II, the overall cross-validation performance of SVM models trained with the hybrid combination of sequenced and structural characteristics, whose average accuracy is close to 90.0%, is performing better than the SVM models trained with only amino acid composition. Most of the SVM models have a predictive accuracy approaching to their cross-validation performance, while several kinase-specific SVM models trained with small data size of training set have an unstable predictive accuracy.

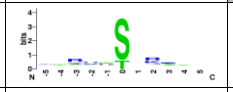
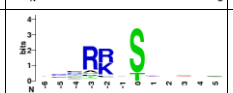
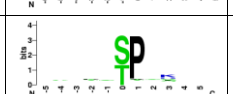
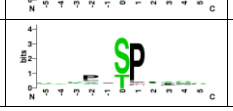
With the consideration of data sufficiency in structural investigation, the kinase-specific groups containing more than ten phosphorylation sites on 3D structures are studied in this work. In general, the kinase-specific SVM models trained with structural information yield a better predictive accuracy than the SVM models trained with only sequence information. Additionally, the SVM models trained with the combination of sequence and structural characteristics were observed to perform at comparable or even slightly better performance levels compared to the SVM models trained with structural information. In summary, for all kinase-specific phosphorylation sites prediction, a consistent increase in performance was obtained suggesting that including 3D structural information does indeed improve the sensitivity and specificity.

C. Effect of Including Structural Information for Identifying Kinase-Specific Phosphorylation Sites with Similar Sequence Motifs

It would be noticed that some of kinase groups have similar substrate motifs. For instance, several kinases (PKA, PKB, PKC, PKG, GRK, RSK,) of AGC family prefer to recognize the substrate sites with basic amino acids (Arginine, Lysine or Histidine) at positions of -2 or -3 relative to the phosphorylation sites (position 0). Assessing the cross classifying specificities among the kinase-specific models containing the similar substrate site motifs, a particular group is regarded as the positive set and the other groups are regarded as the negative sets one by one. As given in Table III, in the first row the classifying specificity (Sp) of PKA model corresponding to the PKC, PKB and PKG data sets are 51.4%, 27.5% and 38.6%, respectively. This investigation indicates the cross classifying specificities are relatively lower among the kinases PKA, PKC, PKB, and PKG in basophilic group. Similarly, the Sp values marked in blue are relatively lower between the kinases CDK and MAPK in proline-directed group. We observe that the cross classifying specificities corresponding to the kinase-specific models in the same kinase group, such as basophilic, acidophilic, and proline-directed groups, are relatively lower than the specificities corresponding to the kinase-specific models in different groups. To investigate the effect of including

structural characteristics for identifying kinase-specific phosphorylation sites with similar substrate motifs, the cross classifying specificities among the kinase-specific models trained with the combination of sequence and structural information are evaluated. Almost all of the Sp values are increased, especially for the Sp values marked in red, green, and blue. This investigation demonstrates that the consideration of structural information could improve the predictive specificity when identifying the kinase-specific phosphorylation sites with similar sequence motifs.

TABLE IIII: CROSS CLASSIFYING SPECIFICITY AMONG THE KINASE-SPECIFIC MODELS TRAINED WITH SEQUENCE-BASED CHARACTERISTICS

Positive set \ Negative set		PKA (66)	PKC (63)	PKB (18)	PKG (15)	CDK (16)	MAPK (30)
PKA (66)		Sn = 89.1%	51.4%	27.5%	38.6%	68.7%	80.0%
PKC (63)		34.3%	Sn = 85.3%	36.2%	62.1%	87.5%	93.3%
PKB (18)		47.8%	83.2%	Sn = 83.3%	81.2%	100%	100%
PKG (15)		41.8%	71.3%	70.2%	Sn = 80.0%	87.5%	80.0%
CDK (16)		100%	96.2%	100%	100%	Sn = 93.7%	57.6%
MAPK (30)		97.0%	98.6%	94.4%	93.3%	46.2%	Sn = 90.0%

IV. CONCLUSION

The aim of this work is to develop a computational method for effectively identifying the kinase-specific phosphorylation sites on protein three-dimensional structures. With the high-throughput mass spectrometry (MS)-based experiment, the desire to comprehensively annotate the catalytic kinases for *in vivo* phosphorylation sites has been highly motivated. Herein, the proposed method could provide a large-scale prediction of kinase-specific phosphorylation sites with reliable accuracy and stable performance. This study has demonstrated that the kinase-specific models trained with the consideration of 3D structural information could perform better than the models trained with only the sequence information, especially improving the cross classifying specificities among the kinase groups containing similar sequence motifs. Additionally, the proposed method was compared with several popular phosphorylation prediction tools, including GPS 2.0, PPSP, and KinasePhos 2.0. The number of kinase groups, sensitivity and specificity of four well-known kinase groups (PKA, PKC, CK2 and SRC) are compared. Particularly, GPS 2.0 and our method could provide more than 100 kinase-specific groups for phosphorylation sites prediction. In the independent testing performance of PKA, PKC, CK2 and SRC groups, our method is comparable with other tools.

REFERENCES

[1] M. Steffen *et al.*, "Automated modelling of signal transduction networks," *BMC Bioinformatics*, vol. 3, pp. 34, 2002.

[2] M. J. Hubbard and P. Cohen, "On target with a new mechanism for the regulation of protein phosphorylation," *Trends Biochem Sci*, vol. 18, no. 5, pp. 172-7, 1993.

[3] G. Manning *et al.*, "The protein kinase complement of the human genome," *Science*, vol. 298, no. 5600, pp. 1912-34, 2002.

[4] Y. Xue *et al.*, "PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory," *BMC Bioinformatics*, vol. 7, pp. 163, 2006.

[5] N. Blom *et al.*, "Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence," *Proteomics*, vol. 4, no. 6, pp. 1633-49, 2004.

[6] J. C. Obenauer, L. C. Cantley, and M. B. Yaffe, "Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs," *Nucleic Acids Res*, vol. 31, no. 13, pp. 3635-41, 2003.

[7] J. H. Kim *et al.*, "Prediction of phosphorylation sites using SVMs," *Bioinformatics*, vol. 20, no. 17, pp. 3179-84, 2004.

[8] 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.

[9] 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.

[10] Y. Xue *et al.*, "GPS: a comprehensive www server for phosphorylation sites prediction," *Nucleic Acids Res*, vol. 33, pp. W184-7, 2005.

[11] J. Wan *et al.*, "Meta-prediction of phosphorylation sites with weighted voting and restricted grid search parameter selection," *Nucleic Acids Res*, vol. 36, no. 4, pp. e22, 2008.

[12] M. L. Miller *et al.*, "Linear motif atlas for phosphorylation-dependent signaling," *Sci Signal*, vol. 1, no. 35, pp. ra2, 2008.

[13] H. D. Huang *et al.*, "KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites," *Nucleic Acids Res*, vol. 33, pp. W226-9, 2005.

[14] H. D. Huang *et al.*, "Incorporating hidden Markov models for identifying protein kinase-specific phosphorylation sites," *J Comput Chem*, vol. 26, no. 10, pp. 1032-41, 2005.

[15] 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, pp. W588-94, 2007.

[16] R. Linding *et al.*, "Systematic discovery of in vivo phosphorylation networks," *Cell*, vol. 129, no. 7, pp. 1415-26, 2007.

[17] B. Kobe *et al.*, "Substrate specificity of protein kinases and computational prediction of substrates," *Biochim Biophys Acta*, vol. 1754, no. 1-2, pp. 200-9, 2005.

[18] N. Blom, S. Gammeltoft, and S. Brunak, "Sequence and structure-based prediction of eukaryotic protein phosphorylation sites," *J Mol Biol*, vol. 294, no. 5, pp. 1351-62, 1999.

[19] N. F. Saunders and B. Kobe, "The Predikin webserver: improved prediction of protein kinase peptide specificity using structural information," *Nucleic Acids Res*, vol. 36, pp. W286-90, 2008.

[20] A. Zanzoni *et al.*, "Phospho3D 2.0: an enhanced database of three-dimensional structures of phosphorylation sites," *Nucleic Acids Res*, vol. 39, pp. D268-71, 2011.

[21] A. Zanzoni *et al.*, "Phospho3D: a database of three-dimensional structures of protein phosphorylation sites," *Nucleic Acids Res*, vol. 35, pp. D229-31, 2007.
[22] P. Durek *et al.*, "Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins," *BMC Bioinformatics*, vol. 10, pp. 117, 2009.
[23] H. M. Berman *et al.*, "The protein data bank," *Nucleic Acids Res*, vol. 28, no. 1, pp. 235-42, 2000.
[24] T. Y. Lee *et al.*, "dbPTM: an information repository of protein post-translational modification," *Nucleic Acids Res*, vol. 34, pp. D622-7, 2006.
[25] C. T. Lu *et al.*, "dbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications," *Nucleic Acids Res*, vol. 41, no. D1, pp. D295-305, 2013.
[26] H. Dinkel *et al.*, "Phospho.ELM: a database of phosphorylation sites--update 2011," *Nucleic Acids Res*, vol. 39, pp. D261-7, 2011.
[27] N. Farriol-Mathis *et al.*, "Annotation of post-translational modifications in the Swiss-Prot knowledge base," *Proteomics*, vol. 4, no. 6, pp. 1537-50, 2004.
[28] P. V. Hornbeck *et al.*, "PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse," *Nucleic Acids Res*, vol. 40, pp. D261-70, 2012.
[29] F. Gnad *et al.*, "PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites," *Genome Biol*, vol. 8, no. 11, pp. R250, 2007.
[30] H. Li, *et al.*, "SysPTM: a systematic resource for proteomic research on post-translational modifications," *Mol Cell Proteomics*, vol. 8, no. 8, pp. 1839-49, 2009.
[31] G. R. Mishra *et al.*, "Human protein reference database--2006 update," *Nucleic Acids Res*, vol. 34, pp. D411-4, 2006.
[32] T. Y. Lee *et al.*, "RegPhos: a system to explore the protein kinase-substrate phosphorylation network in humans," *Nucleic Acids Res*, vol. 39, pp. D777-87, 2011.
[33] T. D. Schneider and R. M. Stephens, "Sequence logos: a new way to display consensus sequences," *Nucleic Acids Res*, vol. 18, no. 20, pp. 6097-100, 1990.
[34] W. C. Chang *et al.*, "Incorporating support vector machine for identifying protein tyrosine sulfation sites," *J Comput Chem*, vol. 30, no. 15, pp. 2526-37, 2009.
[35] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577-637, 1983.
[36] C. J. Lin and Y. W. Chen, "Combining SVMs with various feature selection strategies," *NIPS 2003 feature Selection Challenge*, pp. 1-10, 2003.
[37] C. H. Tung, J. W. Huang, and J. M. Yang, "Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for rapid search of protein structure database," *Genome Biol*, vol. 8, no. 3. pp. R31, 2007.
[38] J. M. Yang and C. H. Tung, "Protein structure database search and evolutionary classification," *Nucleic Acids Res*, vol. 34, no. 13, pp. 3646-59, 2006.
[39] C. H. Tung and J. M. Yang, "fastSCOP: a fast web server for recognizing protein structural domains and SCOP superfamilies," *Nucleic Acids Res*, vol. 35, pp. W438-43, 2007.
[40] C. C. Lin and C. J. Ca. LIBSVM: A library for support vector machines. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

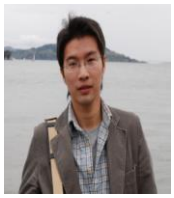
Min-Gang Su obtained his Bachelor of Engineering from Chung Hua University in 2009 and obtained his Master from Yuan Ze University in 2011. He is currently a student of the PhD course at Graduate School of Computational Systems Biology, Yuan Ze University (Yuan Ze, Taiwan). His research interests include intelligent data processing and bioinformatics.

Kai-Yao Huang received his B.S. (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.

Chi-Hua Tung obtained his B.E. in Department of Biological Sciences, National Sun Yat-Sen University. He obtained the M.E. and Ph.D. in Institute of Bioinformatics and Systems Biology, National Chiao Tung University (NCTU). Currently he is a professor in the Department of Bioinformatics, Chung-Hua University. His research interests include bioinformatics, protein structure, structural biology, systems biology, data mining and machine learning.



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.