Protein Phosphorylation Site Prediction via Feature Discovery Support Vector Machine

Yi Shi*[1], Bo Yuan[2], Guohui Lin[1], Dale Schuurmans[1]

[1]Department of Computing Science
University of Alberta. Edmonton, Alberta, Canada
{ys3, guohui, daes}@ualberta.ca
[2]Department of Computer Science and Engineering
Shanghai Jiaotong University. Shanghai, China
yuanbo@cs.sjtu.edu.cn

Abstract—**Protein phosphorylation/dephosphorylation is the central mechanism of post-translational modification which regulates cellular responses and phenotypes. Due to the efficiency and resource constraints of the *in vivo* methods for identifying phosphorylation sites, there is a strong motivation to computationally predict potential phosphorylation sites. In this work, we propose to use a unique set of features to represent the peptides surrounding the amino acid sites of interest and use feature selection support vector machine to predict whether the serine/threonine sites are potentially phosphorylable, as well as selecting important features that may lead to phosphorylation. Experimental results indicate that the new features and the prediction method can more effectively predict protein phosphorylation sites than the existing state of the art methods. The features selected by the our prediction model provides biological insights to the *in vivo* phosphorylation.**

Index Terms—**Protein Phosphorylation, Support Vector Machine, Sparse Learning, Feature Selection, Position-Specific Scoring Matrix.**

I. Introduction

Protein phosphorylation happens in the post-translational stage of prokaryotes and eukaryotes in which a protein kinase modifies a protein by adding a covalently bound phosphate group to a residue (usually serine, threonine or tyrosine). The protein regulation by phosphorylation, also known as phosphoregulation, is one of the most common regulations of protein function. A protein switches between a phosphorylated and an unphosphorylated form in almost all cases of phosphoregulation, and if one of these two is an active form, the other one is inactive. The phosphoregulation, which is essentially the transportation of energy groups, plays a central role in the regulation of almost all cellular behaviors [29], such as apoptosis [36], regulation of transcription [30], DNA repair [34], metabolism [2], cellular differentiation [15], environmental stress response [32], cellular mobility [24], and immune response [11]. Whether a protein has phosphorylized sites is mostly addressed, while which sites on a protein that are phosphorylated still remains challenging [21].

In vivo, researchers discovered novel phosphorylation sites mostly through low-throughput biological techniques. For example, many labs used the site-directed mutagenesis based technique to characterize specific phosphorylation events [19]. Such techniques are usually time-consuming and costly. More recently, the high-throughput mass spectrometry based technique has significantly accelerated the identification of novel sites [7]. Nonetheless, this technique also has limitations, and requires very expensive instruments and specialized expertise that are not available in typical laboratories [29].

Due to the limitations of both low-throughput and high-throughput *in vivo* techniques for phosphorylation site discovery, researchers are paying much attention on the *in vitro* approaches. These *in vitro* approaches usually only require a protein sequence as input, and output some quantified measurement that indicates how likely a serine, threonine or tyrosine (S/T/Y) residue in that sequence is phosphorylated. Many computationally predicted phosphorylation sites have been experimentally validated *in vivo* [1], [4], [27].

Previous researches indicate that only when the amino acids surrounding a given S/T/Y residue fit certain patterns, the residue's phosphorylation can be catalyzed by a protein kinase [3]. However, sequence motifs that describe these patterns are neither sensitive nor specific (i.e., many patterns may occur at random) [1], [26]. Therefore, advanced machine learning methods that can identify more complex and subtle patterns are required. Many machine learning methods have been proposed which use position-specific scoring matrices (PSSMs), decision trees, genetic algorithms, artificial neural networks (ANNs), and support vector machines (SVMs). As the simplest technique, PSSM is a matrix in which rows represent amino acids and columns represent positions in a multiple sequence alignment; PSSM gives a weighted match to any given substring of fixed length. More complex PSSM variations are also developed in practice [12], [14]. However, PSSMs are not able to detect combinations of multiple amino acids patterns which may be important in practice [1]. The two most popular machine learning techniques, ANNs and SVMs, have proven to capture more complex patterns [1], thus are more widely adopted in phosphorylation site prediction, with the tradeoff of increasing computational complexity. In ANN classifiers, the classification function are usually implicit due to multiple neural layers and nonlinear weight functions [1], [9], [21]. In most SVM classifiers for protein phosphorylation site prediction, due to the adopted input usually being sparse coding of the peptides surrounding amino acid sites of interest, it is usually not clear which features essentially lead to the phosphorylation [5], [6], [10], [31], [33]. Some other methods employ the secondary structure [8] or 3D structural information [10], [16], [22], [25], [28].

After reviewing most machine learning methods in protein phosphorylation prediction, Trost *et al.* raise an important question: do they model the actual biological mechanisms underlying protein kinase recognition, or do they merely recognize patterns [29]? Trost *et al.* argue that the latter is more or less the case and propose that while pattern recognition has resulted in much success, it is plausible that more closely modeling the underlying biology of substrate recognition will result in the greater gains in predictive performance.

In this paper, we use 26 explicit sequence-based features for the phosphorylation site prediction. Meanwhile, to select important features that may provide insights to the phosphorylation mechanism, we use the L_1 norm support vector machine classifier. The rest of this paper is organized as follows: In Section II the detailed features and the classification models will be introduced; in Section III, we will introduce the dataset we used, the other methods we compare our method to, the evaluation criteria, and the experimental results; Section IV concludes the paper after discussion.

II. Methods

A. *Features*

Rather than using sparse coded peptides as the input for a classifier, we use 26 explicit features for the peptides surrounding the amino acid of interest. The first 20 features are the 20 amino acid composition of the peptide, which is the occurrence of each amino acid divided by the peptide length. Although we lose some ordering information by treating each amino acid composition separately, considering the peptides we use are very short (e.g., 9-15 in length) and that the PSSM score will be used later, the information loss is not significant. Then we use the average hydrophobicity score and the average relative surface accessibility (RSA) score of the given peptide, that is, we sum up each individual amino acid's hydrophobicity (or RSA) score and divide it by the peptide length. The hydrophobicity scores and RSA scores are predicted by NetSurfP [23] using the whole protein sequence as inputs. The next 3 features are the percentage of the 3 predicted secondary structure composition, i.e., alpha helix, beta sheet, and random coil as they are believed to be important features in previous study [8]. Similar to the previous two scores, these three scores are also calculated based on NetSurfP's secondary structure predictions. The last feature we use is the position-specific scoring matrix (PSSM) score of a given peptide with respect to the positive training data [12], [14]. To calculate the PSSM score, we first construct the PSSM matrix based on the positive training data so that each row represents a symbol of the 20 amino acid alphabet and each column represents a position in the peptide. Each element of the matrix is the frequency of certain amino acid appearing in certain position. For a target peptide, its PSSM score is the sum of its amino acid in corresponding positions in the matrix. Note that the PSSM considers amino acid ordering information explicitly.

B. *L_1 norm SVM*

We use both L_2 norm SVM and L_1 norm support vector machines for predicting/classifying a given amino acid site's phosphorylation. The advantage of using L_2 norm SVM is that the original feature space can be mapped to higher dimensional space through kernelization, while the advantage of using L_1 norm SVM is that it can perform feature selection as well as classification.

The L_1 norm SVM is thus what we mainly recommend to use for this work since we aim to not only do the phosphorylation site prediction but also aim to discover important features that are biologically relevant to the protein phosphorylation.

The primal form of L_1-norm SVM is:

$$\begin{aligned}
&\min_{\mathbf{w},\mathbf{b},\xi} \beta\|\mathbf{w}\|_1 + \mathbf{1}^T\xi \\
\textbf{s.t.}:\quad &\xi \geq \mathbf{1} - \triangle(\mathbf{y})(X\mathbf{w} - \mathbf{1}b), \\
&\xi \geq \mathbf{0}.
\end{aligned} \tag{1}$$

where $\triangle(\mathbf{y})$ denotes putting the vector $\mathbf{y}$ on the main diagonal of a square matrix. $X \in \mathbb{R}^{n\times p}$, $\mathbf{y} \in \mathbb{R}^n$, n is the number of peptides, and p is the number of features. Since

$$\begin{aligned}
\|\mathbf{w}\|_1 &= \min_{\gamma\geq 0} \frac{1}{2}\sum_j \left(\frac{w_j^2}{\gamma_j} + \gamma_j\right) \\
&= \min_{\gamma\geq 0} \frac{1}{2}(\mathbf{w}^T G^{-1}\mathbf{w} + \gamma^T\mathbf{1})
\end{aligned}$$

[20], where $G = \triangle(\gamma)$, (1) becomes

$$\begin{aligned}
&\min_{\mathbf{w},\mathbf{b},\xi,\gamma} \frac{\beta}{2}(\mathbf{w}^T G^{-1}\mathbf{w} + \gamma^T\mathbf{1}) + \mathbf{1}^T\xi \\
\textbf{s.t.}:\quad &\xi \geq \mathbf{1} - \triangle(\mathbf{y})(X\mathbf{w} - \mathbf{1}b), \\
&\xi \geq \mathbf{0}, \gamma \geq \mathbf{0}.
\end{aligned} \tag{2}$$

By applying Lagrangian multipliers and setting the partial derivative equal to 0, (2) becomes

$$\begin{aligned}
&\min_{\gamma}\max_{\lambda} \lambda^T\mathbf{1} - \frac{1}{2\beta}\lambda^T\triangle(\mathbf{y})XGX^T\triangle(\mathbf{y})\lambda + \frac{\beta}{2}\gamma^T\mathbf{1} \\
\textbf{s.t.}:\quad &\mathbf{0} \leq \lambda \leq \mathbf{1}, \\
&\lambda^T\mathbf{y} = 0, \\
&\gamma \geq \mathbf{0}.
\end{aligned} \tag{3}$$

Crucially, (3) is convex and can be solved globally [13]. Hence it provides an optimal form of feature selection that can be efficiently obtained in conjunction with SVM training.

III. Experimental Results

A. Dataset

We used the dataset collected and compiled by Miller *et al.* [21]. This dataset contains only serine and threonine phosphorylation sites in bacteria and the positive data points are obtained from the three sources: 14 sites from the Phosphorylation Site Database [35], 71 sites from B. subtilis [18] and 102 sites from E. coli [17]. The negative data points that do not contain serine and threonine phosphorylation are verified by the Mass Spectrometry techniques. After homology reduction using CD-HIT with default values and 90% sequence identity threshold in both full protein length and 13-mer peptide level, Miller *et al.* randomly downsampled the negative set to include about six negative examples per positive example to balance the number of positive data points and negative data points since machine learning methods work poorly on unbalanced datasets. Figure 1 demonstrates the frequency logos of the 20 amino acid composition in overall positive dataset (including both Serine and Threonine) and Serine-, Threonine-only positive dataset, while Figure 2 demonstrates logos of the negative datasets, respectively.

B. Comparison Methods and Evaluation

To compare our methods to existing methods, we choose two widely recognized and representative meth-

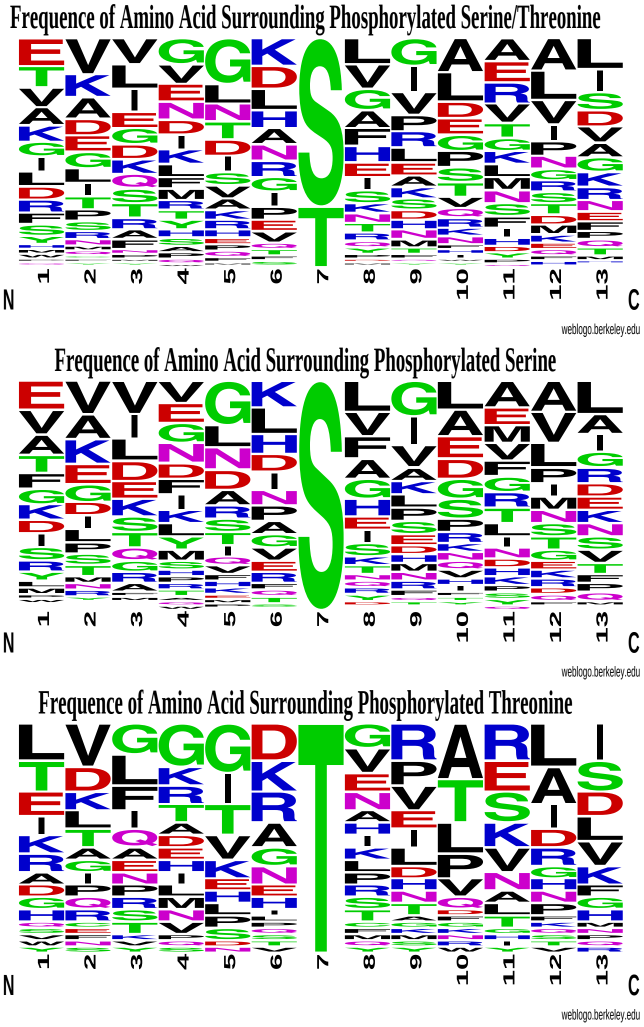

Fig. 1. The frequency logo of the 20 amino acid composition in overall positive dataset (including both Serine and Threonine) and Serine-, Threonine-only positive dataset.

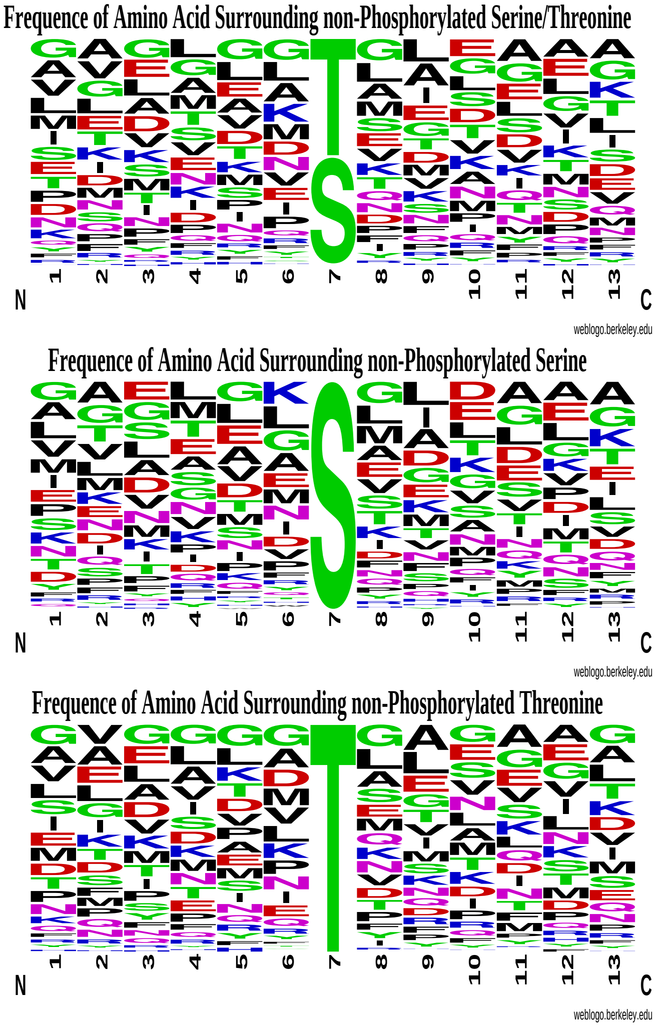

Fig. 2. The frequency logo of the 20 amino acid composition in overall negative dataset (including both Serine and Threonine) and Serine-, Threonine-only negative dataset.

ods NetPhosBac [21] and NetPhos. NetPhosBac is designed specifically for predicting phosphorylation in bacteria data, while NetPhos is designed for Eukaryotic data. All the methods are performed under the 4-fold cross validation scheme described in [21] and are repeated 100 times. We also did the experiments on leave-one-out (LOO) cross validation, but since NetPhosBac and NetPhos were originally evaluated in 4-fold cross validation and the LOO results are not available, we only compare L_1 norm SVM and L_2 norm SVM (both linear kernel and RBF kernel) under the LOO cross validation setup.

We use various criteria to evaluate the effectiveness of different methods, including area under ROC curve (AROC), area under precision-recall curve (AUPR), F-measure, precision, and recall. The AROC and AUPR are appropriate performance measures for binary classification because it is not necessary to choose an arbitrary threshold for defining if a score signifies a positive or negative prediction. The precision and recall scores of each method are obtained at the cutoff threshold where the F-measure score is maximized.

C. Results

Tables I and II demonstrate the evaluation scores AROC, AUPR, FMeasure, Precision, and Recall, of the L_1 norm SVM, linear kernel L_2 norm SVM, RBF kernel L_2 norm SVM, NetPhosBac, and NetPhos prediction methods, under 4-fold and LOO cross validation setup respectively. The corresponding ROC curve plot and precision-recall curve plot are shown in Figures 3, 4, 5, and 6. The AROC, AUPR, and FMeasure scores indicate that the L_1 norm SVM is the most effective method for classification based on the 26 features we use, followed by linear kernel L_2 norm SVM and RBF kernel L_2 norm SVM. The NetPhosBac method performed less effective than the three SVM based approaches. Usually, the RBF kernel L_2 SVM is believed to have better performance than linear kernel L_2 norm SVM, especially on datasets that have a small number of features, but it is not the case on this dataset.

Table III lists the values of the feature selection vector w obtained from training the L_1 norm SVM on the whole dataset (140 positive data points and 841 negative data points). In principle, the greater the absolute value of a feature's w value is, the more it is connected to the phosphorylation *in vivo*. For example, by investigating both Figures 1 and 2 and Table III, we can see that Arginine and Histidine have the largest w values and indeed their frequencies are very different between their positive logo profiles and negative logo profiles. These discoveries are consistent with the biochemistry principle as Arginine and Histidine both contain chemical group NH+ , which is an essential component in a phosphorylation process, because S_N2 reaction in phosphorylating an -OH group requires the help of electron-neutralizing group, i.e., NH+. Methionine has the largest negative

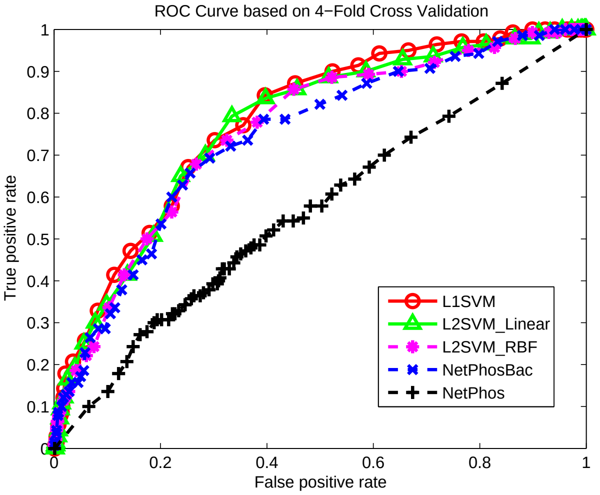

Fig. 3. The ROC curves of the prediction results of the L_1 norm SVM, linear kernel L_2 norm SVM, RBF kernel L_2 norm SVM, NetPhosBac, and NetPhos methods based on 4-fold cross validation setup.

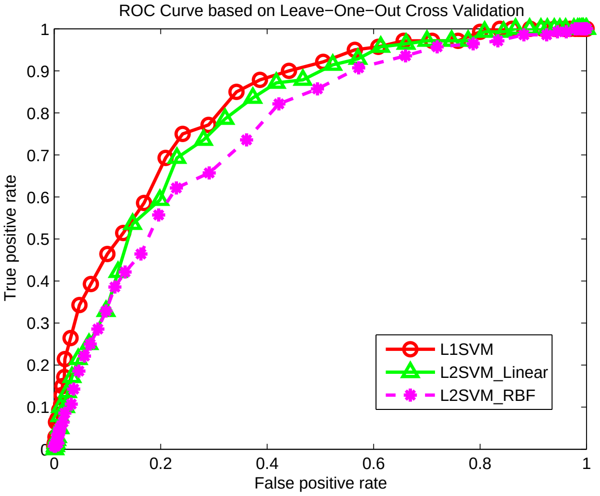

Fig. 4. The ROC curves of the prediction results of the L_1 norm SVM, linear kernel L_2 norm SVM, and RBF kernel L_2 norm SVM based on leave-one-out cross validation setup.

Methods	AROC	AUPR	FMeasure	Precision	Recall
L1SVM	**0.7833**	**0.3598**	**0.4259**	**0.3151**	0.6571
L2SVM_Linear	0.7708	0.3426	0.4152	0.3019	**0.6643**
L2SVM_RBF	0.7566	0.3401	0.4092	0.3017	0.6357
NetPhosBac	0.7416	0.3340	0.4116	0.2997	0.6571
NetPhos	0.5606	0.2369	0.2610	0.1785	0.4857

TABLE I
The evaluation scores AROC, AUPR, FMeasure, Precision, and Recall, of the $L1$ norm SVM, linear kernel $L2$ norm SVM, RBF kernel $L2$ norm SVM, NetPhosBac, and NetPhos prediction methods, under 4-fold cross validation setup.

Methods	AROC	AUPR	FMeasure	Precision	Recall
L1SVM	**0.8242**	**0.4353**	**0.4651**	**0.3448**	0.7143
L2SVM_Linear	0.7965	0.3606	0.4289	0.2981	**0.7643**
L2SVM_RBF	0.7571	0.3174	0.4027	0.2932	0.6429

TABLE II
The evaluation scores AROC, AUPR, FMeasure, Precision, and Recall, of the $L1$ norm SVM, linear kernel $L2$ norm SVM, RBF kernel $L2$ norm SVM, NetPhosBac, and NetPhos prediction methods, under leave-one-out cross validation setup.

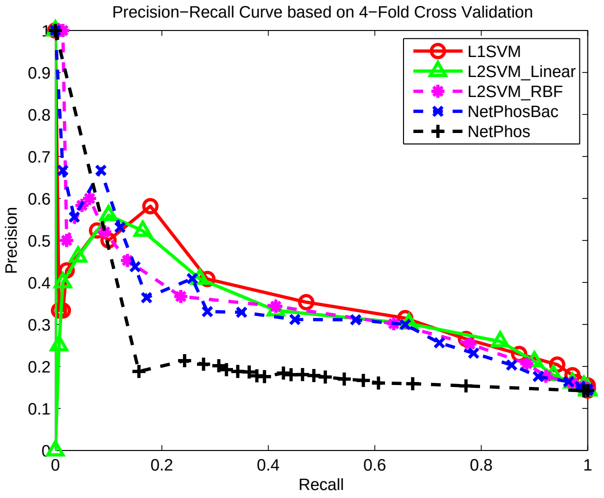

Fig. 5. The precision-recall curves of the prediction results of the L_1 norm SVM, linear kernel L_2 norm SVM, RBF kernel L_2 norm SVM, NetPhosBac, and NetPhos methods based on 4-fold cross validation setup.

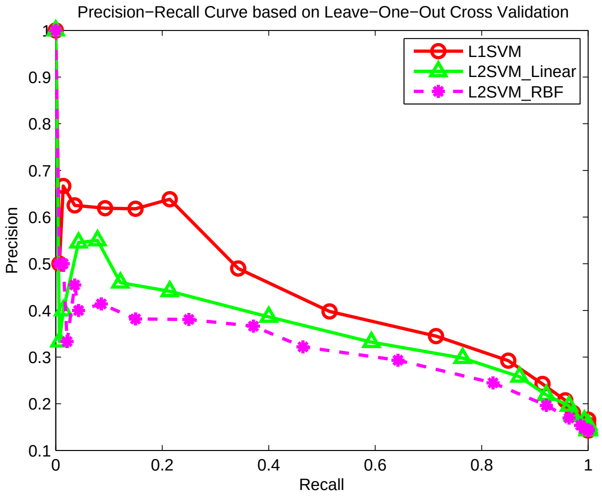

Fig. 6. The precision-recall curves of the prediction results of the L_1 norm SVM, linear kernel L_2 norm SVM, and RBF kernel L_2 norm SVM methods based on leave-one-out cross validation setup.

value and indeed it appears much more frequently in negative data points than in positive data points. Also, according to the w values, hydrophobicity score isn't a very significant feature; relative surface accessibility (RSA) score is fairly important; Alpha helix and random coil appear more in negative data points than in positive data points; the PSSM score is an important feature which is in accord with common sense. Table III also lists the correlation coefficient between each individual feature and the training label. We can see that the correlation coefficient scores associated with the features are strongly connected to the trained w values. We calculated the correlation coefficient score between the w vector and $coef.$ vector, which is 0.75 and re-affirms their connections. Based on these, we may argue that our L_1 norm SVM can not only predict the protein phosphorylation sites more accurately, but also select the important features that are connected to the underlining biological reasons that cause protein phosphorylation.

IV. Discussion and Conclusion

In this work, we propose to use a unique set of features to represent the peptides surrounding the amino acid sites that may be potentially phosphorylated. We use L_1 norm feature selection support vector machine to predict whether the sites are phosphorylable, as well as selecting important features that may cause the phosphorylation. The feature selection functions in our prediction model were never investigated before to the best of our knowledge, making our method a novel approach in protein phosphorylation prediction area. Experimental results indicate that the features and the prediction method can more effectively predict protein phosphorylation than existing state of the art methods. The features selected by the our prediction model provide insights into the biochemical cause of protein phosphorylation *in vivo*.

References

[1] N. Blom, S. Gammeltoft, and S. Brunak1. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *Journal of Molecular Biology*, 294:1356–1362, 1999.

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

[3] S. H. Diks, K. Parikh, M. V. D. Sijde, J. Joore, T. Ritsema, and M. P. Peppelenbosch. Evidence for a minimal eukaryotic phosphoproteome. *PLoS One*, 2:e777, 2007.

[4] H. C. Fan, X. Zhang, and P. A. McNaughton. Activation of the TRPV4 ion channel is enhanced by phosphorylation. *J. Biol. Chem.*, 284:27884–27891, 2009.

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

[6] J. L. Heazlewood, P. Durek, J. Hummel, J. Selbig, W. Weckwerth, D. Walther, and W. X. Schulze. PhosPhAt: a database of phosphorylation sites in *arabidopsis thaliana* and a plant-specific phosphorylation site predictor. *Nucleic Acids Res.*, 36:D1015–D1021, 2008.

[7] E. L. Huttlin, M. P. Jedrychowski, J. E. Elias, T. Goswami, R. Rad, S. A. Beausoleil, J. Villen, W. Haas, M. E. Sowa, and S. P. Gygi. A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell*, 143:1174–1189, 2010.

[8] L. M. Iakoucheva, P. Radivojac, C. J. Brown, T. R. O'connor, J. G. Sikes, Z. Obradovic, and A. K. Dunker. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.*, 32:1037–1049, 2004.

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

[10] J. H. Kim, J. Lee, B. Oh, K. Kimm, and I. Koh. Prediction of phosphorylation sites using SVMs. *Bioinformatics*, 20:3179–3184, 2004.

[11] S. H. Kim and C. E. Lee. Counter-regulation mechanism of IL-4 and IFN-signal transduction through cytosolic retention of the pY-STAT6:pY-STAT2:p48 complex. *Eur. J. Immunol.*, 41:461–472, 2011.

[12] M. Koenig and N. Grabe. Highly specific prediction of phos-

Feature	W	Coef.	Feature	W	Coef.
Alanine	-0.783	-0.030	*Phenylalanine*	0.00	0.006
Arginine	2.444	0.194	*Proline*	0.00	-0.020
Asparagine	-0.087	-0.030	*Serine*	0.00	0.080
Aspartic	-0.057	0.001	*Threonine*	-0.011	-0.111
Cysteine	0.00	0.094	*Tryptophan*	0.00	0.038
Glutamic	-0.482	-0.031	*Tyrosine*	-0.537	-0.07
Glutamine	-0.232	-0.063	*Valine*	0.034	0.088
Glycine	-0.551	0.00	Hydrophob.	0.049	0.005
Histidine	2.453	0.129	RSA	0.179	-0.009
Isoleucine	0.00	0.055	Alpha Helix	-0.610	0.055
Leucine	-0.503	0.031	Beta Sheet	0.00	-0.002
Lysine	0.061	0.003	Coil	-1.670	-0.070
Methionine	-2.965	-0.188	PSSM Score	0.738	0.237

TABLE III
The values of the feature selection vector W, and the correlation coefficient values between each feature vector and the training label vector.

phorylation sites in proteins. *Bioinformatics*, 20:3620–3627, 2004.

[13] G. Lanckriet, M. Cristianini, P. Bartlett, and et al. Learning the kernel matrix with semi-definite programming. *J. Mach. Learn. Res.*, 5:27–72, 2004.

[14] T. Li, F. Li, and X. Zhang. Prediction of kinase-specific phosphorylation sites with sequence features by a log-odds ratio approach. *Proteins*, 70:404–414, 2008.

[15] I. Lian, J. Kim, H. Okazawa1, J. Zhao, B. Zhao, J. Yu, A. Chinnaiyan, M. A. Israel, L. S. B. Goldstein, R. Abujarour, S. Ding, and K. L. Guan. The role of YAP transcription coactivator in regulating stem cell self-renewal and differentiation. *Genes Dev.*, 24:1106–1118, 2010.

[16] R. Linding, L. J. Jensen, G. J. Ostheimer, M. A. van Vugt, C. Jorgensen, I. M. Miron, F. Diella, K. Colwill, L. Taylor, K. Elder, P. Metalnikov, V. Nguyen, A. Pasculescu, J. Jin, J. G. Park, L. D. Samson, J. R. Woodgett, R. B. Russell, P. Bork, M. B. Yaffe, and T. Pawson. Systematic discovery of *in vivo* phosphorylation networks. *Cell*, 129:1415–1426, 2007.

[17] B. Macek, F. Gnad, B. Soufi, C. Kumar, J. V. Olsen, I. Mijakovic, and M. Mann. Phosphoproteome analysis of *e. coli* reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation. *Mol. Cell. Proteomics*, 7:299–307, 2008.

[18] B. Macek, I. Mijakovic, J. V. Olsen, F. Gnad, C. Kumar, P. R. Jensen, and M. Mann. The serine/threonine/tyrosine phosphoproteome of the model bacterium *bacillus subtilis*. *Mol. Cell. Proteomics*, 6:697–707, 2007.

[19] R. Meier, D. R. Alessi, P. Cron, M. Andjelkovic, and B. A. Hemmings. Mitogenic activation, phosphorylation, and nuclear translocation of protein kinase beta. *J. Biol. Chem.*, 272:30491–30497, 1997.

[20] C. A. Micchelli and M. Pontil. Learning the kernel function via regularization. *J. Mach. Learn. Res.*, 6:1099–1125, 2006.

[21] M. L. Miller, B. Soufi, C. Jers, N. Blom, B. Macek, and I. Mijakovic. Netphosbac – a predictor for ser/thr phosphorylation sites in bacterial proteins. *Proteomics*, 9:116–125, 2009.

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

[23] B. Petersen, T. N. Petersen, P. Andersen, M. Nielsen, and C. Lundegaard. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct Biol*, 9:51, 2009.

[24] M. Ressurrecao, D. Rollinson, A. M. Emery, and A. J. Walker. A role for p38MAPK in the regulation of ciliary motion in a

eukaryote. *BMC Cell Biol.*, 12:6, 2011.

[25] N. F. W. Saunders and B. Kobe. The Predikin webserver: improved prediction of protein kinase peptide specificity using structural information. *Nucleic Acids Res.*, 36:W286–W290, 2008.

[26] C. J. Sigrist, L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, and P. Bucher. PROSITE: a documented database using patterns and profiles as motif descriptors. *Bioinformatics*, 18:265–274, 2002.

[27] S. A. Slaugenhaupt, A. Blumenfeld, S. P. Gill, M. Leyne, J. Mull, M. P. Cuajungco, C. B. Liebert, B. Chadwick, M. Idelson, L. Reznik, C. M. Robbins, . M. J. B. Izabela Makalowska, D. Krappmann, C. Scheidereit, C. Maayan, F. B. Axelrod, and J. F. Gusella. Tissue-specific expression of a splicing mutation in the IKBKAP gene causes familial dysautonomia. *Am. J. Hum. Genet.*, 68:598–605, 2001.

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

[29] B. Trost and A. Kusalik. Computational prediction of eukaryotic phosphorylation sites. *Bioinformatics*, 27:2927–2935, 2011.

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

[31] M. Wang, C. Li, W. Chen, and C. Wang. Prediction of PK-specific phosphorylation site based on information entropy. *Sci. China C Life Sci.*, 51:12–20, 2008.

[32] Y. Y. Wang, S. M. Chen, and H. Li. Hydrogen peroxide stress stimulates phosphorylation of foxo1 in rat aortic endothelial cells. *Acta Pharmacol. Sin.*, 31:160–164, 2010.

[33] Y. H. Wong, T. Y. Lee, H. K. Liang, C. M. Huang, T. Y. Wang, Y. H. Yang, C. H. Chu, H. D. Huang, M. T. Ko, and J. K. Hwang. KinasePhos 2.0: a web server for identifying protein kinasespecific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res.*, 35:W588–W594, 2007.

[34] C. D. Wood, T. M. Thornton, G. Sabio, R. A. Davis, and M. Rincon. Nuclear localization of p38MAPK in response to DNA damage. *Int. J. Biol. Sci.*, 5:428–437, 2009.

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

[36] J. Zhang and G. V. Johnson. Tau protein is hyperphosphorylated in a site-specific manner in apoptotic neuronal PC12 cells. *J. Neurochem.*, 75:2346–2357, 2000.