Hindawi
BioMed Research International
Volume 2017, Article ID 1826496, 11 pages
https://doi.org/10.1155/2017/1826496

Research Article

A Novel Phosphorylation Site-Kinase Network-Based Method for the Accurate Prediction of Kinase-Substrate Relationships

Minghui Wang,[1,2] Tao Wang,[1] Binghua Wang,[1] Yu Liu,[1] and Ao Li[1,2]

[1]*School of Information Science and Technology, University of Science and Technology of China, 443 Huangshan Road, Hefei 230027, China*
[2]*Research Centers for Biomedical Engineering, University of Science and Technology of China, 443 Huangshan Road, Hefei 230027, China*

Correspondence should be addressed to Minghui Wang; mhwang@ustc.edu.cn

Received 28 April 2017; Revised 14 August 2017; Accepted 5 September 2017; Published 12 October 2017

Academic Editor: Nasimul Noman

Protein phosphorylation is catalyzed by kinases which regulate many aspects that control death, movement, and cell growth. Identification of the phosphorylation site-specific kinase-substrate relationships (ssKSRs) is important for understanding cellular dynamics and provides a fundamental basis for further disease-related research and drug design. Although several computational methods have been developed, most of these methods mainly use local sequence of phosphorylation sites and protein-protein interactions (PPIs) to construct the prediction model. While phosphorylation presents very complicated processes and is usually involved in various biological mechanisms, the aforementioned information is not sufficient for accurate prediction. In this study, we propose a new and powerful computational approach named KSRPred for ssKSRs prediction, by introducing a novel phosphorylation site-kinase network (pSKN) profiles that can efficiently incorporate the relationships between various protein kinases and phosphorylation sites. The experimental results show that the pSKN profiles can efficiently improve the prediction performance in collaboration with local sequence and PPI information. Furthermore, we compare our method with the existing ssKSRs prediction tools and the results demonstrate that KSRPred can significantly improve the prediction performance compared with existing tools.

1. Introduction

As one of the most common posttranslational modifications (PTMs) [1, 2], phosphorylation plays an important role in the regulation of many cellular processes, such as signal transduction, translation, and transcription [3]. Phosphorylation is catalyzed by protein kinases and usually leads to a functional change, by changing cellular location, enzyme activity, or related to other proteins, of the target protein (substrate) [4, 5]. In human, nearly 75% of all proteins can be modified by protein kinases [6]. Abnormal activity of protein kinases often causes disease, especially cancer, in which protein kinases regulate many aspects that control death, movement, and cell growth [2, 7, 8]. On this point, identification of potential site-specific kinase-substrate relationships (ssKSRs) is important for understanding cellular dynamics and provides a fundamental basis for further disease-related researches and drug design.

To this end, several experimental methods, including low-throughput [9, 10] and high-throughput [11–13] biological technique, are developed to discover phosphorylation sites and corresponding kinases. However, low-throughput experimental identification employs one-by-one manner, which is not only time-consuming but also expensive. Although thousands of phosphorylation sites can be identified by high-throughput mass spectrometry (HTP-MS) techniques [13] in a single experiment [11, 12], it is still difficult to determine which of kinases is responsible for the phosphorylation of the observed site. Therefore, with large-scale phosphoproteomics studies, there is a huge gap between phosphorylation sites and protein kinases, which greatly hampers the study and elucidation of the mechanism of protein phosphorylation in signalling pathways.

So far, several computational methods [14–19] have been put forward to solve this problem during the past few decades, and most of them are mainly based on the sequence information. For example, Zou et al. [20] developed a web server, namely, PKIS, which adopts the composition of monomer spectrum (CMS) to encode the local sequence and then constructed the model with support vector machines (SVMs). Similarly, Damle and Mohanty et al. [15] develop an automated programmer called PhosNetConstruct for predicting target kinases for a substrate protein based on analysis of domain specific kinase-substrate relationships which are derived from the HMM profiles obtained from multiple sequence alignments of related proteins [15]. In addition, recently, some methods [17, 19] use protein-protein interactions (PPIs) to filter potential false positive to further improve performance. For example, Linding et al. [17] develop a web server, namely, NetworKIN, which is based on known sequence motif extracted from Scansite and NetPhosK, and the biological context of substrates is used as a filter to reduce false positives. Meanwhile, to discover the potential protein kinases of the unannotated phosphorylation sites, Song et al. develop a software package of iGPS [19], which is extended from GPS 2.0 [21] algorithm with the interaction filter.

Although these methods have achieved success, phosphorylation presents very complicated processes, it is usually involved in various biological mechanisms. In consequence, the aforementioned information adopted in the existing methods may not fully determine the corresponding protein kinase. It is well known that one protein kinase can catalyze multiple phosphorylation sites and one phosphorylation site can also be phosphorylated by multiple protein kinases [22–24]. For example, CDK2 can catalyze T8, T179, and S213 of protein SMAD3 (P84022), S567 of protein RB1 (P06400), and many other phosphorylation sites [25, 26]. Likewise, S315 of protein TP53 (P04637) can be catalyzed by AURKA, CDK1, CDK2, and so on [27, 28]. The relationships between various protein kinases and phosphorylation sites may bring valuable functional information of protein phosphorylation, which would be helpful in ssKSRs prediction in practice.

Inspired by this information, we propose a novel computational method in this study, namely, KSRPred, for ssKSRs prediction by introducing a phosphorylation site-kinase network (pSKN) profiles that can efficiently incorporate the relationships between various protein kinases and phosphorylation sites. This method is based on the framework of kernel ridge regression [29, 30], which can effectively integrate both pSKN profiles and other useful information including local sequences and PPIs. The experimental results show that the pSKN profiles can efficiently improve the prediction performance in collaboration with local sequence and PPI information. Furthermore, we compare KSRPred with the widely used ssKSRs prediction tools. The results also indicate that the proposed method has a better or comparable prediction performance compared with the existing ssKSRs prediction tools.

2. Materials and Methods

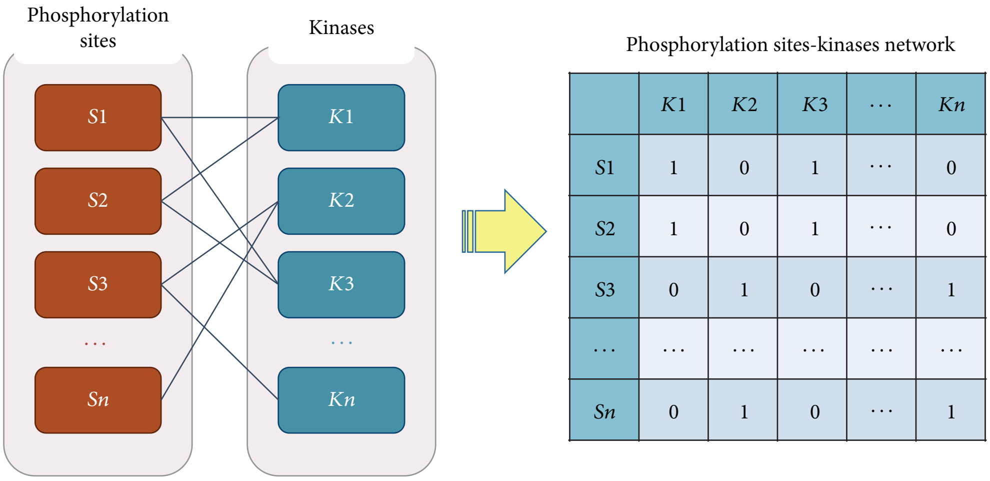
2.1. Data Collection and Preprocessing. In this study, we employ an experimentally verified human phosphorylation sites with corresponding kinases dataset, which include 6,839 verified sites and 389 kinases with 9,480 known ssKSRs extracted from Phospho.ELM [31] and the latest PhosphoSitePlus [32]. And, for this dataset, we follow Xu et al. [33] and Wang et al. [34] and use BlastClust with 70% threshold to remove substrate redundancy. Since iGPS [19], PKIS [20], and NetworKIN [17] use Phospho.ELM as training data, the phosphorylation sites existing in both training and testing data would overestimate the prediction performance. And for fair comparison with the existing tools, we extract an independent test dataset with 1,000 phosphorylation sites from the nonredundant dataset, which excludes the existing phosphorylation sites deposited in Phospho.ELM [31] and the rest as the training dataset. For a specific kinase, the verified sites modified by this kinase are taken as positive samples, and other verified sites are used as negative samples [35]. To achieve a reliable result [15, 36], here we construct models for kinases that at least 15 positive samples and finally 103 kinases are obtained. The detailed information of these kinases are summarized in Table S1 (see Supplementary Material available online at https://doi.org/10.1155/2017/1826496).

2.2. The Sequence Kernel Similarity. A local sequence with a length of 15 amino acids is extracted from the phosphorylation site, which contains 7 upstream and 7 downstream residues. We compute the sequence similarities of two phosphorylation sites using BLOSUM62 matrix, which is an amino acid substitution matrix that shows the similarities among 20 types of amino acids and usually used to calculate the sequence similarity [37]. The similarity between two phosphorylation sites s_i and s_j is calculated as follows:

$$S_{\text{seq}}\left(s_i, s_j\right) = \sum_{k=1}^{15} \text{BLOSUM62}\left(s_i(k), s_j(k)\right), \qquad (1)$$

where $\text{BLOSUM62}(s_i(k), s_j(k))$ is the similarity score between the kth amino acid of s_i and the kth amino acid of s_j given by BLOSUM62 matrix. Applying this operation to all phosphorylation sites pairs, we construct a similarity matrix denoted as S_{seq}. To ensure that the value of S_{seq} is distributed in the range of $[0, 1]$, normalization is performed subsequently, and the formula is defined as $K_{\text{seq}}(i, j) = (S_{\text{seq}}(i, j) - \min S_{\text{seq}})/(\max S_{\text{seq}} - \min S_{\text{seq}})$. The similarity matrix K_{seq} is considered as kernel similarity matrix of phosphorylation sites calculated from sequence level.

2.3. The PPI Kernel Similarity. The PPI information of substrates is extracted from STRING [38], which is a comprehensive, yet quality-controlled collection of protein-protein associations. Since these associations are derived from high-throughput experimental data, from the mining of database and literature and from predictions based on genomic context analysis [38], we follow Butland et al. [39] and Jafari et al. [40] and use a median (0.4) confidence cut-off value to filter the association. And 18,836 proteins that interacted with the 2,162 nonredundancy substrates are obtained. We compute the PPI similarities between two substrates using Jaccard Index [41]. The similarity between two substrates p_i and p_j is calculated as $S_{\text{ppi}}(p_i, p_j) = |J_{p_i} \cap J_{p_j}|/|J_{p_i} \cup J_{p_j}|$, where J_{p_i} and J_{p_j} represent the PPI information of corresponding substrate, respectively.



	K1	K2	K3	⋯	Kn
S1	1	0	1	⋯	0
S2	1	0	1	⋯	0
S3	0	1	0	⋯	1
⋯	⋯	⋯	⋯	⋯	⋯
Sn	0	1	0	⋯	1

(a)

(b)

Figure 1: *Construction of phosphorylation site-kinase network and extracting the pSKN profiles.* The bipartite network represents the relationships between the phosphorylation sites and kinases; the orange and blue nodes represent phosphorylation sites and kinases, respectively. The matrix represents the pSKN profiles that are extracted from the bipartite network; each row is phosphorylation site s_i and each column is kinase k_j; if s_i is catalyzed by k_j, the value is 1, otherwise 0.

Applying this operation to all substrate pairs, we construct a similarity matrix denoted as S_{ppi}. However, some substrates have more than one phosphorylation sites; these sites have same substrates and share the same PPI information [42]. The similarity matrix K_{ppi} of phosphorylation sites can be obtained by directly extracting the similarity of substrates. The similarity matrix K_{ppi} is considered as kernel similarity matrix of phosphorylation sites calculated from substrate level.

2.4. Construction of pSKN Profiles and Kernel Similarity. The relationships between various kinases and phosphorylation sites can be expressed as a bipartite network (Figure 1), from which we can extract a novel pSKN profiles. Formally, we denote the phosphorylation site set as $X_s = \{s_1, s_2, \ldots, s_{n_s}\}$ and the kinase set as $X_k = \{k_1, k_2, \ldots, k_{n_k}\}$; the relationships between various kinases and phosphorylation sites can be described as a bipartite network $G(X_s, X_k, E)$, where $E = \{e_{ij} : s_i \in X_s,\ k_j \in X_k\}$. A link is drawn between s_i and k_j when the phosphorylation site s_i has relationship with the kinase k_j. This bipartite network can be presented by an $n_s \times n_k$ adjacent matrix Y, where $y_{ij} = 1$ if s_i and k_j are linked, while all other unknown phosphorylation site-kinase pairs are labeled as 0. Afterwards, to incorporate pSKN profiles for prediction, we construct a kernel similarity matrix from the pSKN profiles using Gaussian kernel function (i.e., RBF). The similarity between two phosphorylation sites s_i and s_j is calculated as follows:

$$K_{\text{net}}\left(s_i, s_j\right) = \exp\left(-\gamma_s \left\| y_{s_i} - y_{s_j} \right\|^2\right), \tag{2}$$

where y_{s_i} and y_{s_j} represent the ith and jth row of the adjacency matrix Y, respectively. The kernel bandwidth is controlled by the parameter γ_s. It is normally defined as a new bandwidth parameter γ_s' normalized by the average number of relationships with phosphorylation site per kinase. The formula for the calculation of γ_s is

$$\gamma_s = \frac{\gamma_s'}{\left((1/n) \sum_{i=1}^{n} \left\| y_{s_i} \right\|^2\right)}. \tag{3}$$

Applying this operation to all phosphorylation site pairs, we construct a similarity matrix denoted as K_{net}. The similarity matrix K_{net} is considered as kernel similarity matrix of phosphorylation sites calculated from relationship level.

2.5. Kernel Ridge Classifier. To our knowledge, kernel ridge regression (KRR) is widely used in the field of bioinformatics [43–45], and existing studies [44] show that KRR and SVM have similar classification accuracy. In this study, we test these two algorithms on our dataset and find that KRR is comparable or slightly better than SVM. Therefore, we choose the KRR to construct the prediction model.

Formally, given a training dataset $T = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $x_i \in R^m$ and $y_i \in \{0, 1\}$, the basic idea of KRR relies on mapping the data into a higher dimensional space $\mathscr{H}$ (also called feature space) according to a mapping Φ and then finding a linear regression function with the new training set $T = \{(\Phi(x_1), y_1), \ldots, (\Phi(x_n), y_n), \}$, which represents a nonlinear regression in the original input space [46]. The linear ridge regression problem consists in minimizing the following cost:

$$L(\omega) = \sum_i \left\| y_i - \omega^T \phi\left(x_i\right) \right\|^2 + \lambda \left\| \omega \right\|^2, \tag{4}$$

where λ is a regularization parameter used to control the trade-off between the bias and variance of the estimate. By calculating the derivative of this cost function [47], we can get the optimal solution $\omega^* = \phi(\phi^T \phi + \lambda I_n)^{-1} Y$. Therefore, for a new unlabeled sample x, the predicted label y (i.e., $y = \omega^T \cdot \Phi(x)$) can be calculated by the following formula:

$$\begin{aligned} f(x) &= Y\left(\Phi^T\Phi + \lambda I_n\right)^{-1}\Phi^T\Phi(x) \\ &= Y\left(K + \lambda I_n\right)^{-1} k(x), \end{aligned} \tag{5}$$

where Y is the vector of values y_i and $K(x_i, x_j) = \Phi(x_i)^T\Phi(x_j)$ is the kernel function.

In this study, we develop three similarity kernels, namely, sequence similarity kernel, PPI similarity kernel, and pSKN similarity kernel, from different data sources. In order to make full use of these kernels, we follow van Laarhoven et al. [48] and define a custom kernel function. The formula is defined as follows:

$$K\left(x_i, x_j\right) = \sum_{\varphi\in\{\text{seq},\text{ppi},\text{net}\}} \eta_\varphi K_\varphi\left(x_i, x_j\right), \quad \eta_\varphi \geq 0. \tag{6}$$

And for the reported results of our evaluation, the unweighted average is adopted, that is, $\eta_\varphi = 1/3$, $\varphi \in \{\text{seq}, \text{ppi}, \text{net}\}$. Using (5) and (6), we can easily construct the corresponding model and make prediction for unlabeled phosphorylation sites. The model is implemented by the *scikit-learn* library (version 0.18) [49] in the *Python* environment.

2.6. Performance Evaluation. Following previous works [50, 51], we use 10-fold cross-validation to evaluate the prediction performance of classifier. The receiver operating characteristic (ROC) curve and the area under the curve (AUC) are used to calculate the average performance of 10-fold cross-validations. Meanwhile, in order to ensure the reliability, fairly, the commonly used measurement indexes are also adopted: specificity (Sp), sensitivity (Sn), Matthew's correlation coefficient (MCC), F-Measure ($F1$), and Precision (Pre). The formula is defined as follows:

$$\begin{aligned} &\text{Sp} = \frac{\text{TN}}{\text{TN} + \text{FP}} \\ &\text{Sn} = \frac{\text{TP}}{\text{FN} + \text{TP}} \\ &\text{Pre} = \frac{\text{TP}}{\text{FP} + \text{TP}} \\ &F1 = \frac{2 \times \text{Pre} \times \text{Sn}}{\text{Pre} + \text{Sn}} \\ &\text{MCC} \\ &= \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TN} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TP} + \text{FP})}}. \end{aligned} \tag{7}$$

TN and TP represent the number of positive and negative sites that are correctly predicted, commonly called true negative and true positive, respectively, while FN and FP represent the number of negative and positive sites that are wrong predicted, commonly called false negative and false positive, respectively. It is noteworthy that when the numbers of positive and negative set are significantly imbalanced, MCC can be used to obtain the balance quality.

Table 1: Compare the predictive performance of our methods using different information at medium stringency level (Sp = 90.0%).

Kinases	Methods	AUC	Sn	MCC	$F1$	Pre
CDK2	Seq	88.0%	55.9%	35.8%	40.5%	31.8%
	pSKN	91.2%	72.2%	46.7%	49.4%	37.5%
	Full	93.4%	83.1%	53.6%	54.8%	40.9%
CK2A1	Seq	93.0%	83.4%	50.1%	49.8%	35.5%
	pSKN	94.3%	86.1%	51.8%	51.0%	36.2%
	Full	94.4%	88.4%	53.1%	52.0%	36.8%
FYN	Seq	93.3%	74.1%	24.6%	17.4%	9.9%
	pSKN	94.6%	83.5%	28.1%	19.4%	11.0%
	Full	95.5%	84.7%	28.5%	19.7%	11.1%
GSK3B	Seq	82.2%	51.7%	21.0%	19.4%	11.9%
	pSKN	87.2%	68.5%	28.9%	24.9%	15.2%
	Full	89.3%	73.8%	31.4%	26.6%	16.2%
P38A	Seq	81.2%	43.2%	16.7%	16.2%	10.0%
	pSKN	87.9%	69.2%	29.0%	24.8%	15.1%
	Full	90.5%	75.3%	31.8%	26.7%	16.2%
PKACA	Seq	90.1%	70.5%	41.5%	42.5%	30.5%
	pSKN	91.9%	77.2%	45.5%	45.7%	32.4%
	Full	93.0%	81.0%	47.8%	47.4%	33.5%
PKCA	Seq	85.3%	57.9%	32.7%	34.9%	25.0%
	pSKN	90.3%	69.5%	39.8%	40.5%	28.6%
	Full	91.5%	80.2%	46.2%	45.3%	31.6%
PLK1	Seq	79.1%	48.0%	20.8%	20.7%	13.2%
	pSKN	86.3%	62.6%	28.3%	26.1%	16.5%
	Full	89.7%	80.4%	37.2%	32.4%	20.3%
SRC	Seq	94.5%	88.3%	51.1%	49.3%	34.2%
	pSKN	96.1%	86.4%	50.1%	48.5%	33.7%
	Full	97.2%	92.9%	53.8%	51.2%	35.3%

3. Results

3.1. Evaluation of pSKN Profiles. In this study, we employ a novel pSKN profiles to predict ssKSRs. To confirm the effectiveness of pSKN profiles, we compare the proposed method with and without pSKN profiles on the basis of local sequence information. The prediction performances of these two methods are evaluated on the training dataset using 10-fold cross-validation. Here, we take kinase GSK3B, PLK1, P38A (MAPK14), and CDK2 as an example to illustrate the predictive performance, as shown in Figure 2. It is indicated that the proposed method with pSKN profiles shows a higher prediction accuracy in the ssKSRs prediction. For example, for GSK3B, the AUC value of the proposed method trained with local sequences is 82.2%. After applying pSKN profiles, the AUC value is improved to 87.2%, which is 5.0% higher than the proposed method trained with local sequences only. Likewise, for PLK1, compared to the proposed method with pSKN profiles and using local sequences only, the value of AUC is increased by 7.2%. Moreover, Figure S1 also displays the ROC curves of the three most pleiotropic protein kinases (i.e., PKCA, PKACA, and CK2A1), from which we can get a consistent conclusion. Taking PKCA as an example, the AUC value of our proposed method with pSKN profiles is 90.3%, which is 5.0% higher than the method with local sequences only.

Additionally, by following previous works [19, 20, 52], some measurements such as Sp, Sn, $F1$, Pre, and MCC are also adopted to ensure the reliability of performance evaluation. The measurements are evaluated at medium (Sp = 90.0%) and high (Sp = 95.0%) stringency levels, respectively. Table 1

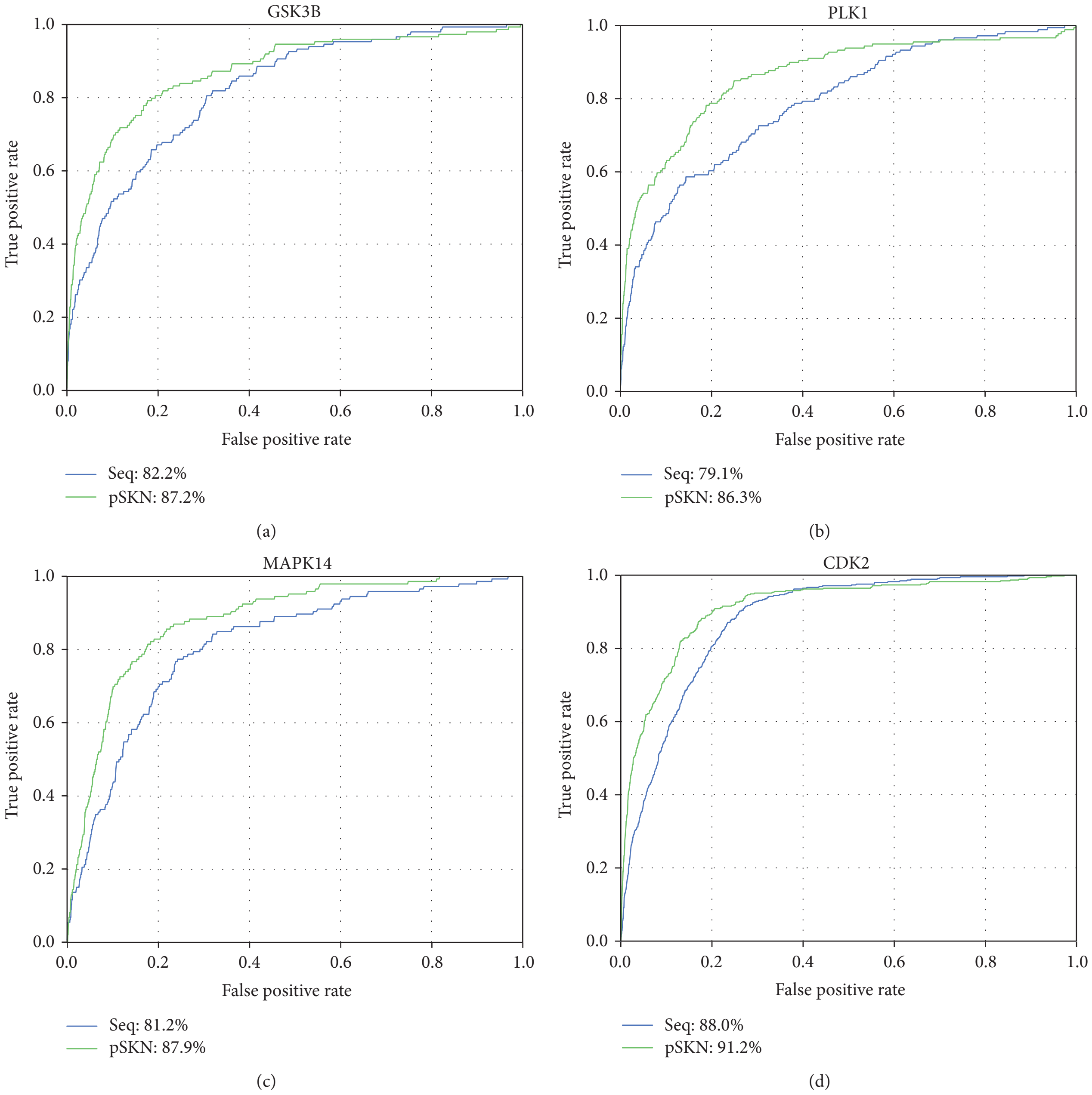
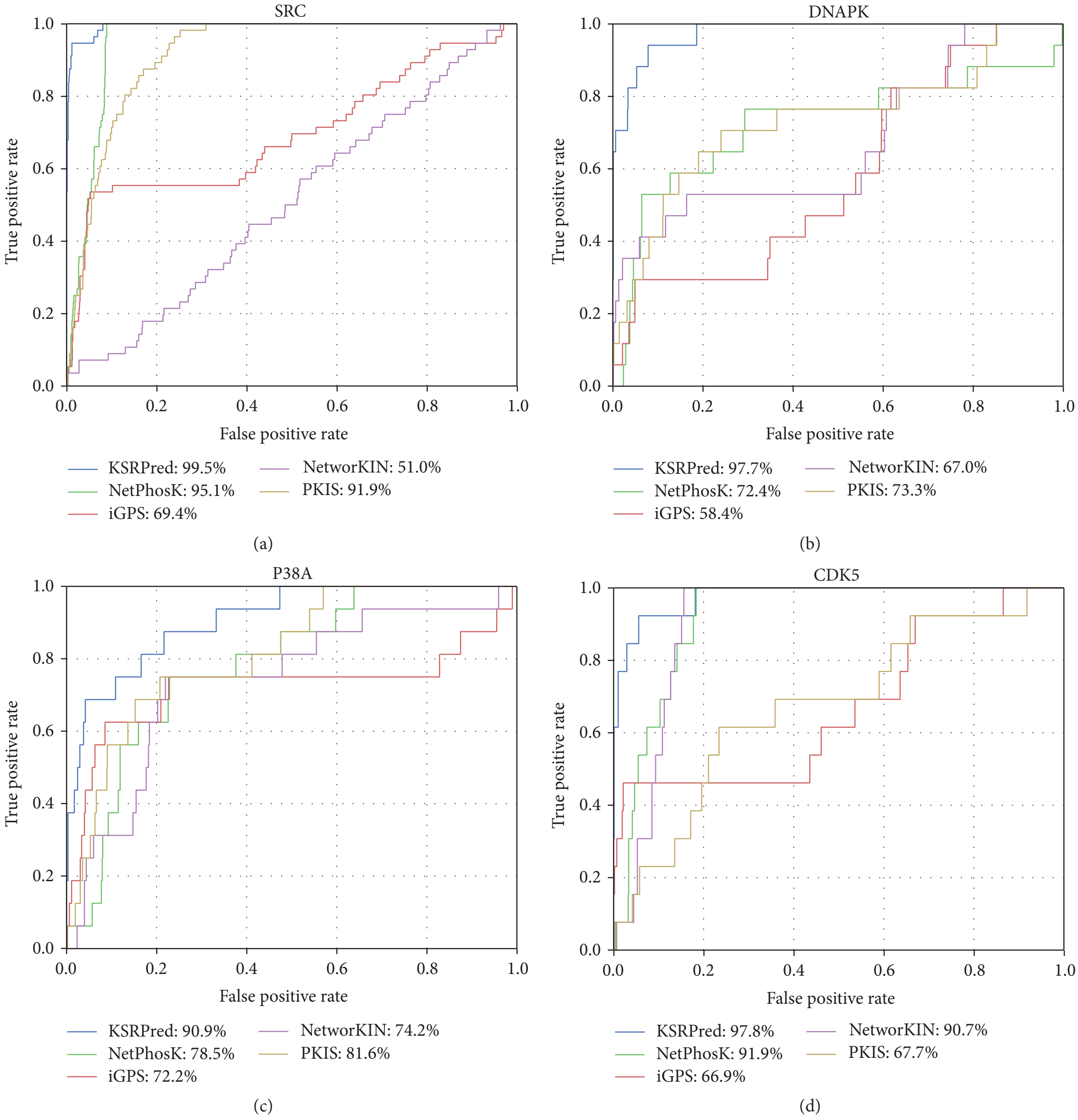


FIGURE 2: *Comparison of ROC curves using different information*. The blue lines represent our method constructed with local sequence only, and the green lines represent our method built with local sequence and pSKN profiles together.

displays the Sn, $F1$, Pre, and MCC values of different kinases at medium stringency level. It is indicated that the proposed method with pSKN profiles achieves the best predictive performance in almost all cases. For example, for PKCA, the Sn, MCC, $F1$, and Pre values are 69.5%, 39.8%, 40.5%, and 28.6%, which are improved by 11.6%, 7.1%, 5.6%, and 3.6% compared with the method using local sequences only. Moreover, Table S2 displays the high stringency level of Sn, MCC, $F1$, and Pre values, from which we can draw a consistent conclusion. In all, these results show that pSKN profiles can significantly improve the prediction performance of different kinases.

Recently, several studies [17, 19] use the PPI information to filter false positive predictions, which can improve the precision of prediction results with the cost of reduced sensitivity [19]. Subsequently, we test the full method that integrates pSKN profile, local sequence, and PPI information to examine the ability of KSRPred in incorporating PPI information. The performance of AUC values and other measurements at high and medium stringency levels is listed in Table 1 and Table S2. As can be seen, for most of kinases, the proposed method can not only improve the precision of prediction results but also enhance the corresponding sensitivity, which indicates that the proposed method can make



(a)

(b)

(c)

(d)

FIGURE 3: *Compare the ROC curves with different methods on the independent dataset*. The blue lines represent the ROC curve of KSRPred, and the green, red, purple, and yellow lines represent the ROC curves of NetPhosK, iGPS, NetworKIN, and PKIS, respectively.

better use of PPI information in comparison with the existing methods [17, 19]. Taking P38A as an example, the AUC value of this full method is increased to 90.5%, which is 2.6% higher than the method with pSKN profiles. Besides, the Sn, MCC, $F1$, and Pre values at medium stringency level (Sp = 90.0%) are improved by 6.1%, 2.8%, 1.9%, and 1.1%, respectively. We also display the performance of other kinases in Table S3.

3.2. Comparison with the Existing ssKSRs Prediction Tools. In the previous section, we have verified the effectiveness of pSKN profiles. In this section, we use the independent test dataset to compare KSRPred with four widely used ssKSRs prediction tools, namely, NetPhosK [53], iGPS [19], NetworKIN [17], and PKIS [20], to evaluate the power of the proposed method. Here, we take four kinases that could be predicted by these tools as an example, and the corresponding ROC curves are displayed in Figure 3. It is indicated that the proposed method is generally superior to the existing tools. For example, for P38A, the AUC value of KSRPred is 90.9%, which is 12.4%, 18.7%, 16.7%, and 9.3% higher than those of NetPhosK, iGPS, NetworKIN, and PKIS, respectively. Likewise, for SRC, the AUC value of KSRPred is 4.40%,

Table 2: Information of top 20 potential phosphorylation sites for CDK2 kinase.

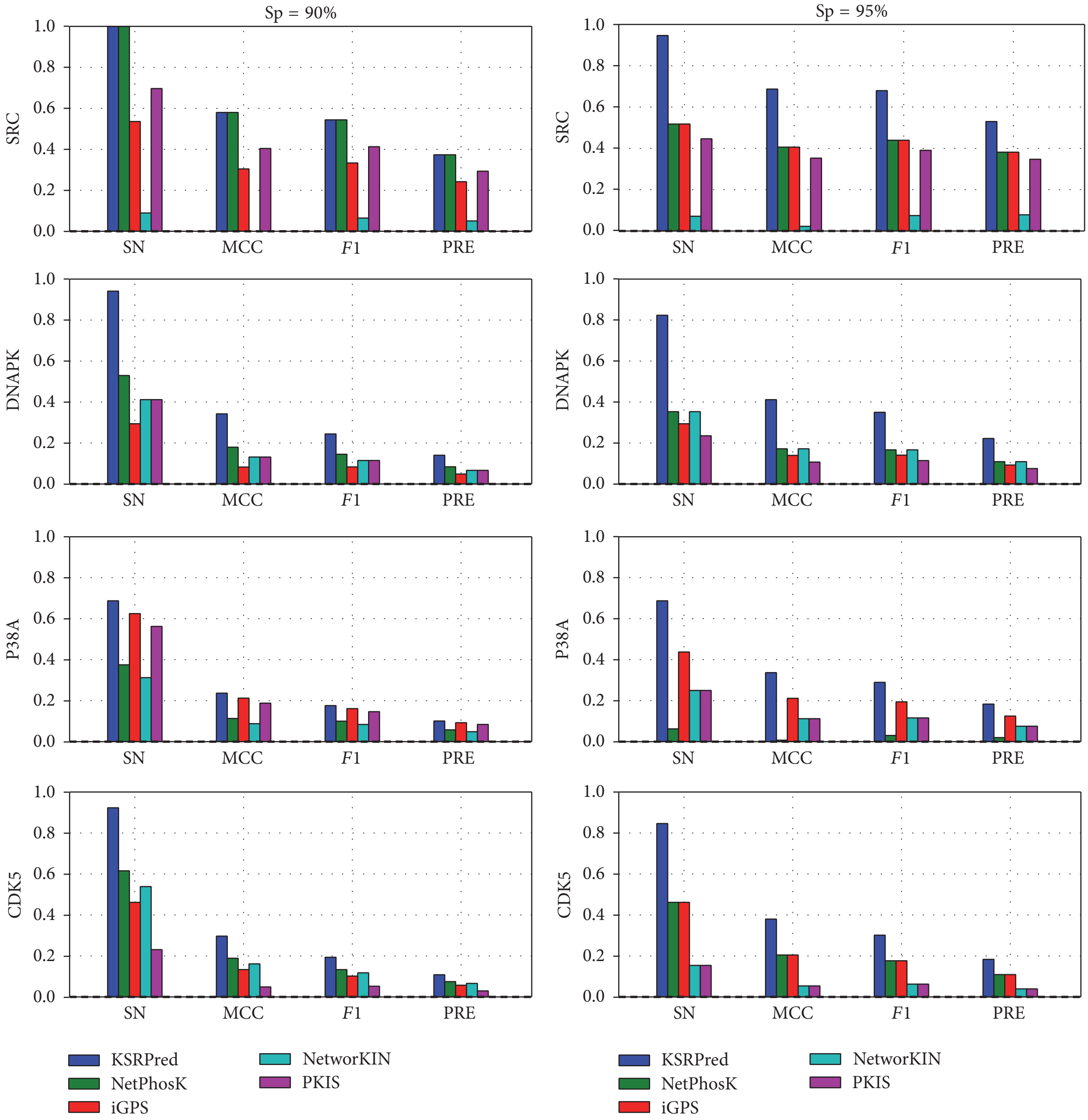
Ranking	UniProtKB	Protein name	Site	Score
1	Q08999	RBL2	S1035	0.4707
2	P28749	RBL1	S964	0.4672
3	P28749	RBL1	T369	0.4481
4	Q08999	RBL2	S672	0.4403
5	P28749	RBL1	S975	0.4389
6	Q08999	RBL2	T401	0.4313
7	Q9UQ35	SRRM2	T1413	0.3952
8	P49736	MCM2	S31	0.3736
9	Q9Y5N6	ORC6	T195	0.3717
10	P24928	POLR2A	S1878	0.3663
11	Q15910	EZH2	T487	0.3653
12	Q9UQ35	SRRM2	T866	0.3553
13	O15446	CD3EAP	S285	0.3505
14	P24928	POLR2A	S1920	0.3495
15	P24928	POLR2A	S1934	0.3492
16	Q02539	HIST1H1A	S183	0.3488
17	P10276	RARA	S77	0.3425
18	Q5TKA1	LIN9	T96	0.3412
19	Q9P1Z0	ZBTB4	T983	0.3347
20	P49736	MCM2	T59	0.3338

30.10%, 48.50%, and 7.60% larger than those of NetPhosK, iGPS, NetworKIN, and PKIS, respectively.

In addition to the AUC values, the measurements (i.e., Sn, $F1$, Pre, and MCC) at medium and high stringency levels are also adopted to evaluate the performance. We draw the Sn-MCC-$F1$-Pre bar chart of the five methods based on the high and medium stringency levels, as shown in Figure 4 and the details are listed in Table S4. The experimental results show that KSRPred achieves the best performance in almost all circumstances in comparison with the existing tools. For example, for SRC, at the high stringency level, the Sn, MCC, $F1$, and Pre values of KSRPred are increased by 42.9%, 28.1%, 24.0%, and 14.8% compared with iGPS and have an improvement of 50.0%, 33.4%, 28.9%, and 18.3% compared with PKIS, respectively. Similarly, compared with NetPhosK and NetworKIN, the Sn, MCC, $F1$, and Pre values of KSRPred are also improved 42.9%, 28.1%, 24.0%, and 14.8% and 87.5%, 66.5%, 60.5%, and 45.2%, respectively. We further analyze the results of this kinase and find that at the high stringency level some phosphorylation sites can be correctly assigned by KSRPred, yet not by the existing tools. For example, Y53 of AKAP8 (O43823) is catalyzed by SRC and can be correctly assigned by our method but cannot be predicted by the existing tools. In summary, these results suggest that KSRPred achieves a better or comparable performance as compared with the existing ssKSRs prediction tools. In addition, in Figure S2, we also compare the performance of the proposed method without pSKN profile with NetPhosK and iGPS. The result shows that, compared with these two tools, KSRPred without pSKN profile can also get a better performance. Taken P38A as an example, the AUC achieved by KSRPred without pSKN profile is 7.8% and 14.1% higher than NetPhosK and iGPS, respectively.

3.3. Detailed Analysis of the Prediction Results. After confirming the advantages of the proposed method, we conduct a detailed analysis on the prediction results. It is known that the predicted top-ranked results are more important in practice, which are utilized for proteomic-wide screening and systematic examination [42]. This requires the computational method with low false positive rate. Hence, we compare the numbers of correctly retrieved ssKSRs according to different percentiles. For each percentile $p\%$, we count the number of true ssKSRs in the top-ranked $p\% * 1{,}000$ predictions. Taking P38A as an example, results of five percentiles 1%, 2%, 5%, 10%, and 15% of the total phosphorylation sites number are compared, as shown in Figure 5. It is indicated that at all percentiles KSRPred can retrieve a more true positive prediction compared with NetPhosK, NetworKIN, iGPS, and PKIS.

In addition, due to the difficulty of experimental verification, computational method is also required to have the ability to detect unknown ssKSRs [42]. In view of this, we analyze the prediction result of top 20 potential phosphorylation sites. Taking CDK2 as an example, the detailed information of these phosphorylation sites is listed in Table 2. By mining of the literature, we find that some results have been confirmed as the phosphorylation sites catalyzed by this kinase. For example, Leng et al. [54] have reported that CDK2 can catalyze the S964 site of protein RBL1 (P28749). Likewise, from the UniProtKB database, we find that this kinase can catalyze the S975 site of protein RBL1 (P28749) (http://www.uniprot.org/uniprot/P28749#ptm_processing). These discoveries suggest that KSRPred has not only a lower false positive rate but also the ability to discover unknown ssKSRs, which could be helpful for the subsequent experimental verification.



(a)

(b)

Figure 4: *Compare the Sn, MCC, F*1*, and Pre values of different methods on the independent dataset*. (a) represents the performance at specificity of 90.0%, and (b) represents the performance at specificity of 95.0%. The horizontal axis represents sensitivity, Matthew correlation coefficient, $F1$-measure, and precision, respectively.

4. Discussions and Conclusions

Phosphorylation plays a significant role in a wide range of cellular processes, which is catalyzed by protein kinases and many phosphorylation-related diseases are closely related to kinases. Prediction of ssKSRs is important for understanding phosphorylation process and provides a fundamental basis for further cell dynamics studies and drug design. However, traditional experimental methods are high-cost and time-consuming, and it is important to develop effective computational methods to predict ssKSRs. Although several computational methods for ssKSRs prediction have been proposed, these methods usually use the local sequence and PPI information, which are not sufficient for accurate prediction. In this study, we present the pSKN profiles that can efficiently incorporate the relationships between various kinases and phosphorylation sites. Using these pSKN profiles, the performance of our proposed method has been

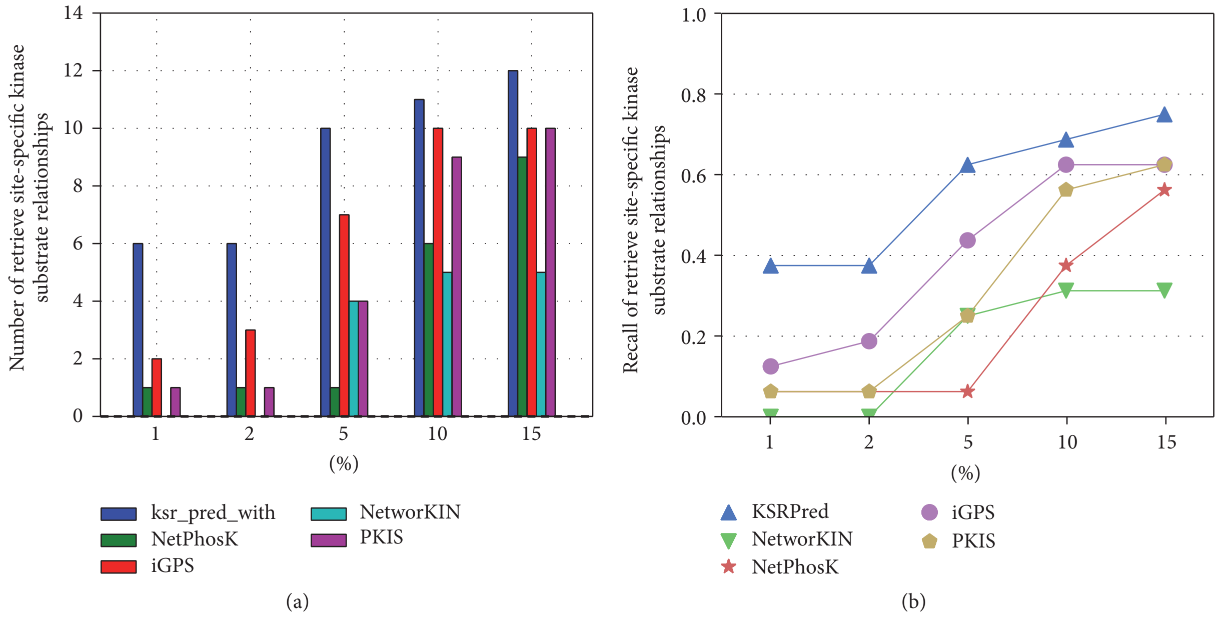


FIGURE 5: *Compare the ability of different methods in retrieve site-specific kinase-substrate relationships.* (a) represents the number of retrieved site-specific kinase-substrate relationships at the different percentiles, and (b) represents the fraction of retrieved site-specific kinase-substrate (recall).

significantly improved. Meanwhile, we use PPIs extracted from STRING database as the substrate feature, and the experimental results show that our proposed method could make better use of this information compared with the existing method (e.g., iGPS and NetworKIN). Furthermore, through the analysis of potential phosphorylation sites, we find that some highly ranked results have been confirmed as phosphorylation sites catalyzed by kinases, suggesting its efficiency in discovering new potential ssKSRs for experimental validations and elucidating the molecular mechanism of protein phosphorylation.

Although the proposed method has shown the good ability for ssKSRs prediction, there is still much room for improvement. It is well known that the quantity of training data plays crucial roles in mastering the performance of machine learning methods [55, 56], and when more training data is available, the performance would be further improved. Additionally, kinases have corresponding family information and there are studies [33, 36] showing that this information is useful for ssKSRs prediction. In this study, we do not consider the influence of kinase family information, which can be integrated into the proposed method in further work. Moreover, the PPI dataset used in this study is from STRING database, and there are many other PPI databases that are publicly available, for example, MINT [57] and I2D [58], which can be included to further improve the performance of the proposed method. Furthermore, as kinase catalyzed phosphorylation site is a complex biological process affected by various mechanisms, incorporating more relevant functional information may also enhance the performance of ssKSRs prediction. Finally, the pSKN profiles are extracted from the relationships between kinases and phosphorylation sits, and the experimental results show that this information can effectively improve the prediction performance. However, available experimentally verified relationships between kinases and phosphorylation sits are still comparatively rare. Hence, it is expected that the performance of KSRPred will be further improved when more relationships can be obtained.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by National Natural Science Foundation of China (no. 61471331, no. 61571414, and no. 61101061).

References

[1] Y. Lou, J. Yao, and A. Zereshki, "NEK2A interacts with MAD1 and possibly functions as a novel integrator of the spindle checkpoint signaling," *The Journal of Biological Chemistry*, vol. 279, no. 19, pp. 20049–20057, 2004.

[2] C. R. Singh, C. Curtis, Y. Yamamoto et al., "Eukaryotic translation initiation factor 5 is critical for integrity of the scanning preinitiation complex and accurate control of GCN4 translation," *Molecular and Cellular Biology*, vol. 25, no. 13, pp. 5480–5491, 2005.

[3] P. Cohen, "The origins of protein phosphorylation," *Nature Cell Biology*, vol. 4, no. 5, pp. E127–E130, 2002.

[4] T. Hunter, "Signaling—2000 and beyond," *Cell*, vol. 100, no. 1, pp. 113–127, 2000.

[5] F.-F. Zhou, Y. Xue, G.-L. Chen, and X. Yao, "GPS: A novel group-based phosphorylation predicting and scoring method,"

Biochemical and Biophysical Research Communications, vol. 325, no. 4, pp. 1443–1448, 2004.

[6] K. Sharma, R. C. J. D'Souza, S. Tyanova et al., "Ultra-deep Human Phosphoproteome Reveals a Distinct Regulatory Nature of Tyr and Ser/Thr-Based Signaling," *Cell Reports*, vol. 8, no. 5, pp. 1583–1594, 2014.

[7] M. Bajpai, "Fostamatinib, a Syk inhibitor prodrug for the treatment of inflammatory diseases," *IDrugs*, vol. 12, no. 3, pp. 174–185, 2009.

[8] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam, "The protein kinase complement of the human genome," *Science*, vol. 298, no. 5600, pp. 1912–1934, 2002.

[9] Z. Lin, P.-W. Zhang, X. Zhu et al., "Phosphatidylinositol 3-kinase, protein kinase C, and MEK1/2 kinase regulation of dopamine transporters (DAT) require N-terminal DAT phosphoacceptor sites," *Journal of Biological Chemistry*, vol. 278, no. 22, pp. 20162–20170, 2003.

[10] M. Salinas, J. Wang, M. Rosa De Sagarra et al., "Protein kinase Akt/PKB phosphorylates heme oxygenase-1 in vitro and in vivo," *FEBS Letters*, vol. 578, no. 1-2, pp. 90–94, 2004.

[11] G. Han, M. Ye, H. Liu et al., "Phosphoproteome analysis of human liver tissue by long-gradient nanoflow LC coupled with multiple stage MS analysis," *Electrophoresis*, vol. 31, no. 6, pp. 1080–1089, 2010.

[12] C. Song, M. Ye, G. Han et al., "Reversed-phase-reversed-phase liquid chromatography approach with high orthogonality for multidimensional separation of phosphopeptides," *Analytical Chemistry*, vol. 82, no. 1, pp. 53–56, 2010.

[13] J. Villén, S. A. Beausoleil, S. A. Gerber, and S. P. Gygi, "Large-scale phosphorylation analysis of mouse liver," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 5, pp. 1488–1493, 2007.

[14] N. Blom, T. Sicheritz-Pontén, R. Gupta, S. Gammeltoft, and S. Brunak, "Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence," *Proteomics*, vol. 4, no. 6, pp. 1633–1649, 2004.

[15] N. P. Damle and D. Mohanty, "Deciphering kinase-substrate relationships by analysis of domain-specific phosphorylation network," *Bioinformatics*, vol. 30, no. 12, pp. 1730–1738, 2014.

[16] N. Kumar and D. Mohanty, "Identification of substrates for Ser/Thr kinases using residue-based statistical pair potentials," *Bioinformatics*, vol. 26, no. 2, Article ID btp633, pp. 189–197, 2010.

[17] R. Linding, L. J. Jensen, A. Pasculescu et al., "NetworKIN: A resource for exploring cellular phosphorylation networks," *Nucleic Acids Research*, vol. 36, no. 1, pp. D695–D699, 2008.

[18] N. F. W. Saunders and B. Kobe, "The Predikin webserver: improved prediction of protein kinase peptide specificity using structural information.," *Nucleic acids research*, vol. 36, pp. W286–290, 2008.

[19] C. Song et al., "Systematic analysis of protein phosphorylation networks from phosphoproteomic data," *MCP*, vol. 11, pp. 1070–1083, 2012.

[20] L. Zou, M. Wang, Y. Shen, J. Liao, A. Li, and M. Wang, "PKIS: Computational identification of protein kinases for experimentally discovered protein phosphorylation sites," *BMC Bioinformatics*, vol. 14, no. 1, article 247, 2013.

[21] Y. Xue, J. Ren, X. Gao, C. Jin, L. Wen, and X. Yao, "GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy," *MCP*, vol. 7, pp. 1598–1608, 2008.

[22] D. Alessi, M. T. Kozlowski, Q.-P. Weng, N. Morrice, and J. Avruch, "3-phosphoinositide-dependent protein kinase 1 (PDK1) phosphorylates and activates the p70 S6 kinase in vivo and in vitro," *Current Biology*, vol. 8, no. 2, pp. 69–81, 1998.

[23] M. P. Coba et al., "Neurotransmitters drive combinatorial multistate postsynaptic density networks Science signaling," 2009.

[24] A. Venerando, L. Cesaro, and L. A. Pinna, "From phosphoproteins to phosphoproteomes: A historical account," *FEBS Journal*, vol. 284, pp. 1936–1951, 2017.

[25] J. W. Harbour, R. X. Luo, A. Dei Santi, A. A. Postigo, and D. C. Dean, "Cdk phosphorylation triggers sequential intramolecular interactions that progressively block Rb functions as cells move through G1," *Cell*, vol. 98, no. 6, pp. 859–869, 1999.

[26] I. Matsuura, N. G. Denissova, G. Wang, D. He, J. Long, and F. Liu, "Cyclin-dependent kinases regulate the antiproliferative function of Smads," *Nature*, vol. 430, no. 6996, pp. 226–231, 2004.

[27] H. Katayama, K. Sasai, H. Kawai et al., "Phosphorylation by aurora kinase A induces Mdm2-mediated destabilization and inhibition of p53," *Nature Genetics*, vol. 36, no. 1, pp. 55–62, 2004.

[28] M. G. Luciani, J. R. A. Hutchins, D. Zheleva, and T. R. Hupp, "The C-terminal regulatory domain of p53 contains a functional docking site for cyclin A," *Journal of Molecular Biology*, vol. 300, no. 3, pp. 503–518, 2000.

[29] J. He, L. Ding, L. Jiang, and L. Ma, "Kernel ridge regression classification," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '14)*, pp. 2263–2267, July 2014.

[30] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, The MIT Press, 2012.

[31] H. Dinkel, C. Chica, A. Via et al., "Phospho.ELM: a database of phosphorylation sites-update 2011," *Nucleic Acids Research*, vol. 39, no. 1, pp. D261–D267, 2011.

[32] P. V. Hornbeck, J. M. Kornhauser, S. Tkachev et al., "PhosphoSitePlus: A comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse," *Nucleic Acids Research*, vol. 40, no. 1, pp. D261–D270, 2012.

[33] X. Xu, A. Li, L. Zou, Y. Shen, W. Fan, and M. Wang, "Improving the performance of protein kinase identification via high dimensional protein-protein interactions and substrate structure data," *Molecular BioSystems*, vol. 10, no. 3, pp. 694–702, 2014.

[34] M. Wang, C. Li, W. Chen, and C. Wang, "Prediction of PK-specific phosphorylation site based on information entropy," *Science in China, Series C: Life Sciences*, vol. 51, no. 1, pp. 12–20, 2008.

[35] M. Wang, Y. Jiang, and X. Xu, "A novel method for predicting post-translational modifications on serine and threonine sites by using site-modification network profiles," *Molecular BioSystems*, vol. 11, no. 11, pp. 3092–3100, 2015.

[36] A. Li, X. Xu, H. Zhang, and M. Wang, "Kinase identification with supervised laplacian regularized least squares," *PLoS ONE*, vol. 10, no. 10, Article ID 139676, 2015.

[37] A. Li, L. Wang, Y. Shi, M. Wang, Z. Jiang, and H. Feng, "Phosphorylation site prediction with a modified k-Nearest Neighbor algorithm and BLOSUM62 matrix," in *Proceedings of the 27th Annual International Conference of the Engineering in Medicine and Biology Society (IEEE-EMBS '05)*, pp. 6075–6078, September 2005.

[38] D. Szklarczyk, A. Franceschini, M. Kuhn et al., "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Research*, vol. 39, no. 1, pp. D561–D568, 2011.

[39] G. Butland, M. Babu, J. Greenblatt, and A. Emili, "eSGA: E. coli Synthetic Genetic Array analysis," *Protocol Exchange*, vol. 5, pp. 789–795, 2008.

[40] M. Jafari, P. Nickchi, A. Safari, S. J. Tazehkand, and M. Mirzaie, "IMAN: Interlog protein network reconstruction," *Matching and ANalysis*, 2016.

[41] J. I. F. Bass, A. Diallo, J. Nelson, J. M. Soto, C. L. Myers, and A. J. M. Walhout, "Using networks to measure similarity between genes: association index selection," *Nature Methods*, vol. 10, no. 12, pp. 1169–1176, 2013.

[42] X. Xu and M. Wang, "Inferring disease associated phosphorylation sites via random walk on multi-layer heterogeneous network," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 5, pp. 836–844, 2016.

[43] S. Giguère, M. Marchand, F. Laviolette, A. Drouin, and J. Corbeil, "Learning a peptide-protein binding affinity predictor with kernel ridge regression," *BMC Bioinformatics*, vol. 14, article 82, 2013.

[44] A. Statnikov, M. Henaff, V. Narendra et al., "A comprehensive evaluation of multicategory classification methods for microbiomic data," *Microbiome*, vol. 1, no. 1, article 11, 2013.

[45] J. Tamez-Pena, P. Gonzalez, E. Schreyer, and S. Totterman, "Structural biomarkers predict onset of knee pain: data from the osteoarthritis initiative," *Osteoarthritis and Cartilage*, vol. 20, p. S34, 2012.

[46] D. H. Trinh, M. Wong, J.-M. Rocchisani, C. D. Pham, and F. Dibos, "Medical image denoising using Kernel ridge regression," in *Proceedings of the 18th IEEE International Conference on Image Processing (ICIP '11)*, pp. 1597–1600, 2011.

[47] C. Saunders, A. Gammerman, and V. Vovk, "Ridge Regression Learning Algorithm in Dual Variables," in *Proceedings of the Paper presented at the Proceedings of the Fifteenth International Conference on Machine Learning*, 1998.

[48] T. van Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug-target interaction," *Bioinformatics*, vol. 27, no. 21, pp. 3036–3043, 2011.

[49] F. Pedregosa et al., "Scikit-learn: Machine Learning," *Python J Mach Learn Res*, vol. 12, pp. 2825–2830, 2011.

[50] N. Yang, *Systems and Computational Biology - Molecular and Cellular Experimental Systems*, 2011.

[51] Z. R. Yang, "Predicting sulfotyrosine sites using the random forest algorithm with significantly improved prediction accuracy," *BMC Bioinformatics*, vol. 10, article 1471, p. 361, 2009.

[52] Y. Xue, A. Li, L. Wang, H. Feng, and X. Yao, "PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory," *BMC Bioinformatics*, vol. 7, article 163, 2006.

[53] M. L. Miller, T. S. Ponten, T. N. Petersen, and N. Blom, "NetPhosK - Prediction of kinase-specific phosphorylation from sequence and sequence-derived features," *FEBS Journal*, pp. 272-111, 2005.

[54] X. Leng, M. Noble, P. D. Adams, J. Qin, and J. W. Harper, "Reversal of growth suppression by p107 via direct phosphorylation by cyclin D1/cyclin-dependent kinase 4," *Molecular and Cellular Biology*, vol. 22, no. 7, pp. 2242–2254, 2002.

[55] S. J. Burian, S. R. Durrans, S. J. Nix, and R. E. Pitt, "Training artificial neural networks to perform rainfall disaggregation," *Journal of Hydrologic Engineering*, vol. 6, no. 1, pp. 43–51, 2001.

[56] E. J. Hughes, M. Lewis, C. M. Alabaster, and L. F. Soldani, "Automatic target recognition: Problems of data separability and decision making," in *Proceedings of the IET Seminar on High Resolution Imaging and Target Classification*, pp. 29–37, November 2005.

[57] L. Licata, L. Briganti, D. Peluso et al., "MINT, the molecular interaction database: 2012 update," *Nucleic Acids Research*, vol. 40, no. 1, pp. D857–D861, 2012.

[58] K. R. Brown and I. Jurisica, "Unequal evolutionary conservation of human protein interactions in interologous networks," *Genome Biology*, vol. 8, no. 5, article R95, 2007.