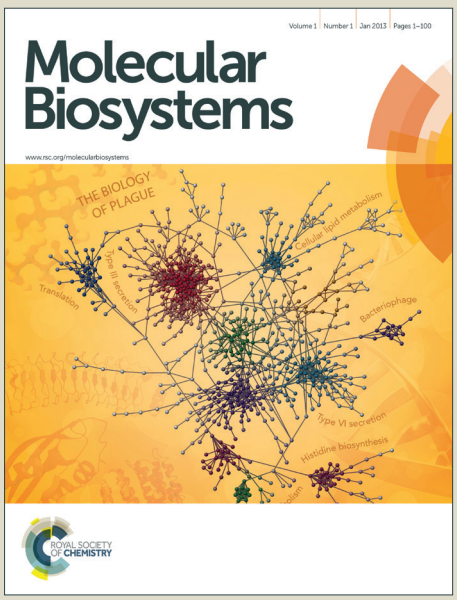
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Molecular BioSystems

Cite this: DOI: 10.1039/c0xx00000x

PAPER

www.rsc.org/molecularbiosystems

Improving the performance of protein kinase identification from high dimensional protein-protein interaction and substrate structure data

*Xiaoyi Xu[a], Ao Li[a,b], Liang Zou[a], Yi Shen[a], Wenwen Fan[a] and Minghui Wang[a, b, *]*

Received (in XXX, XXX) Xth XXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX

DOI: 10.1039/b000000x

As a crucial posttranslational modification, protein phosphorylation regulates almost all basic cellular processes. Recently, thousands of phosphorylation sites have been discovered by large-scale phospho-proteomics studies, but only about 20% of them have information of catalytic kinases, which brings a great challenge to correct identification of the protein kinases responsible for experimentally verified phosphorylation sites. In most existing identification tools, only local sequence was selected to construct predictive models, and information of protein-protein interaction (PPI) was adopted for further filtering. However, the limited information utilized by these tools is not sufficient to identify protein kinases responsible for phosphorylated proteins. In this work, a novel computational approach that fully incorporates PPI and substrate structure information is proposed to improve the performance of human protein kinase identification. To handle the issue of high-dimensional PPI and structure data, a two-step feature selection algorithm that incorporates support vector machine (SVM), is designed to detect information useful in discriminating corresponding kinase of phosphorylation sites. Benchmark datasets for kinase identification are constructed with human protein phosphorylation data extracted from the latest Phospho.ELM database. With the selected PPI and structure features the performance of kinase identification is significantly enhanced as compared with that obtain by using only sequence information. To further verify our method, we compare it with the state-of-the-art tools: NetworKIN and IGPS at two stringency levels with medium (>90.0%) and high (>99.0%) specificity. The results show that our method outperforms existing tools in identifying protein kinases. Further evaluation demonstrates our method also performs superiorly on different hierarchical levels including kinase, subfamily, family and group.

Introduction

As an important and reversible type of post-translational modification, protein phosphorylation plays an essential role in the regulation of cellular processes such as metabolism, gene expression, cell signal pathways, growth, motility, differentiation, division and membrane transport[1-5]. Phosphorylation is catalyzed by protein kinases that regulate a myriad of cellular processes and about half of them are related to cancer and other diseases [6, 7]. In this regard, identification of phosphorylation sites along with their site-specific kinase could provide more details for understanding the molecular mechanisms of various diseases and suggest potential drug targets [8]. To this end, a lot of experimental efforts have been taken to identify kinase substrates and corresponding phosphorylation sites. Although beneficial in illustrating the mechanisms of phosphorylation, these approaches are limited by the availability and optimization of enzymatic reactions [9, 10].

With the advance of mass spectrometry-based techniques [11], experimentally determined phosphorylation sites were exponentially increased and several databases [12-18] of phosphorylation substrates were built subsequently. However, the mass spectrometry experiments cannot ascertain protein kinases that phosphorylate the identified substrates, resulting in very limited kinase information in existing phosphorylation databases. For example, as a database of experimentally verified phosphorylation sites in eukaryotic proteins, Phospho.ELM [16] currently contains 37,145 human phosphorylation sites, but only 10% of them (3,599sites) are associated with corresponding kinase information. With large-scale phospho-proteomics studies, the huge gap between protein kinases and phosphorylation sites will continue to increase, which largely hampers the studies aiming to elucidate catalytic mechanism of protein phosphorylation and kinase-related components of signaling

[a] *School of Information Science and Technology, University of Science and Technology of China, Hefei AH230027, People's Republic of China; Tel: +8613866702372; E-mail: xxyyy@mail.ustc.edu.cn*

[b] *Research Centers for Biomedical Engineering, University of Science and Technology of China, Hefei AH230027, People's Republic of China; Tel: 18056082266; E-mail: mhwang@ustc.edu.cn*

†Electronic Supplementary Information (ESI) available: See DOI: 10.1039/b000000x/

pathways.

To solve this problem, a few computational approaches have been proposed recently[19-21]. For example, Linding *et al.* introduced a novel tool called NetworKIN [19] that identify kinases based on sequence similarity with known sequence motif collected from Scansite [22] and NetPhosK [23]. Song *et al.* developed IGPS [21] software that adopts the predictor in GPS 2.0 [24] to discover potential PKs for the un-annotated phosphorylation sites. Despite the success achieved by these approaches, the predictive models used by most of existing methods are mainly based on local sequence of phosphorylation sites. However, protein phosphorylation is a complicated process with various biological mechanisms involved [25], thus the sequence information cannot fully determine the corresponding protein kinase. To overcome this issue, protein-protein interactions (PPI), which reflects the potential of interactions between kinase, substrate and other proteins involved in phosphorylation process, are further adopted to filter out potentially false positives [19, 21]. However, such simple filtering strategy cannot fully utilize the PPI information and suffers from decreased prediction sensitivity, leaving large room for further improvement of protein kinase identification. Furthermore, the obstacle in incorporating PPI data lies in the fact that in an organism there are usually thousands of proteins with extremely complex interactions. For example, from STRING database [26] we can obtain 18,600 human proteins and 489,929 corresponding interactions, and such high dimensionality of input data may lead to severe problems such as heavy burden on classifier, degradation of generalization abilities and significantly decreased performance [27].

To address the issue in adoption of PPI data for kinase identification, in this study a novel approach is proposed that takes full advantage of PPI information to identify protein kinases responsible for phosphorylated proteins. In addition, structures of kinase substrates are also incorporated as they are reported to be helpful in phosphorylation prediction studies[18, 28, 29]. To build predictive models from high dimensional PPI and structure data, an efficient feature selection algorithm is developed to pick up important data for kinase identification. The experimental results show that the performance of kinase identification is significantly improved by incorporating selected PPI and structure data with sequence information. Further evaluation demonstrates the proposed approach remarkably outperforms existing kinase identification tools.

Materials and Methods

Data collection and pre-processing

In this work, 37,145 experimentally verified human phosphorylated S, T, and Y sites (3,599 sites with corresponding kinase information) were derived from the latest version of Phopho.ELM (9.0) [16] as a benchmark dataset. After removing identical proteins associated with multiple PubMed IDs, 27,404 phosphorylation sites (3,151 sites with kinase information) were obtained, which represent 2,398 unique phosphorylation sites in 934 proteins with kinase information. To avoid overestimation caused by redundancy and homology bias, the protein sequences were clustered with a 70% threshold identity by Blastclust [30], and then a representative of each cluster was retained. Finally 889 proteins with 2,289 sites were extracted for further analysis, and the number of serine(S)/threonine(T) and tyrosine (Y) substrates were 1,823 and 466, respectively. For each kinase, the corresponding phosphorylation sites were regarded as positive data, while those catalyzed by other kinases were regarded as negative data. After removing kinases sets with less than 20 positive samples, 21 protein kinases were obtained for further investigation. In addition, the kinases were hierarchically organized into major groups, families, subfamilies according to the classification scheme proposed by Manning *et al.* [6]. Finally 21 datasets in kinase level, 10 datasets in subfamily level, 17 datasets in family level and 6 datasets in group level were constructed. The detailed information of positive and negative data in four hierarchical levels is summarized in Table S1.

Feature Extraction and Encoding

Protein-protein interaction and protein structures including solvent accessibility, secondary structure, disorder region, were extracted as the features in this work. The PPI data was downloaded from the STRING database (version 9.05) [26], which contains protein-protein interactions values in 18,600 human proteins. The PPI information is integrated into an adjacency matrix that contains interaction values between each pair of proteins. Higher values in the matrix indicate stronger interactions and '0' represents no interaction between two proteins. After deriving all 16,025 proteins that have interactions with the 889 non-redundancy phosphorylated proteins, finally an 889*16,025 matrix was obtained with each column representing a PPI feature.

Due to the limited number of proteins with experimentally determined structures, predicted structures of the full-length phosphoproteins were generated by SABLE [31] and VSL2 [32] and the structures of the 21-mer amino acids fragment centered on the phosphorylation site were then extracted for further analyses. For each amino acid in the query sequence, secondary structures were classified into 'coil', 'helix' or 'beta strand' by SABLE and were then encoded as '100', '010' and '001', respectively. The solvent accessibility of each amino acid was predicted as 'buried' or 'exposed' with confidence level from 0 to 9. For each amino acid the disorder score predicted by VSL2 varied from 0 to 1, with higher score indicating greater tendency to be located in disorder region. Together, the PPI and structure data finally renders 16,130 features for further analysis.

Feature selection and kinase identification

We propose an efficient wrapper feature selection algorithm including a ranking process followed by a two-step forward feature selection. As shown in the pseudocode, an efficient filtering feature selection method mRMR is first employed to rank features based on the so called "minimal-redundancy-maximal-relevance" criterion[33], which takes both the minimal information redundancy among features (*f*) and maximum information relevance with the kinase (*k*) into account based on mutual information. The mutual information between variable *x* and *y* is defined as:

$$I(x;y) = \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \qquad (1)$$

where $p(x)$ and $p(y)$ are the marginal probability distribution

functions of two variables and $p(x,y)$ is the joint probability distribution function. Specifically, for the i^{th} candidate feature f_i in feature set S, the "minimal-redundancy" and "maximal-relevance" criteria can be formulated as equation (2) and (3). Function R indicates the redundancy between two given features and function D indicates the relevance between feature and class label (here we use '1' to represent a phosphorylation site catalyzed by a specific kinase and '0' otherwise). If f_i satisfies both equations, it will then be selected by mRMR.

$$\max D(S,k), \quad D = \frac{1}{|S|}\sum_{f_i \in S} I(f_i;k) \tag{2}$$

$$\min R(S), \qquad R = \frac{1}{|S^2|}\sum_{f_i; f_j \in S} I(f_i; f_j) \tag{3}$$

Afterwards, a two-step forward feature selection is applied based on the ranked feature subsets. Features are added iteratively from higher to lower index according to the rank derived by mRMR to determine the optimal size of feature subset. First, the feature subset starts with all the sequence features and adds one structure feature each time to train a SVM model. After that, the corresponding AUC (Area under the curve, see *Performance Evaluation*) is calculated to evaluate the performance of the feature subsets. The procedure is repeated until the number of selected feature exceeds a threshold M (default value 1,000). Next, the subset with the maximal AUC is selected to be the next initial subset for PPI feature selection. Finally, an optimal subset including sequence, structure and PPI features with the best performance is returned by the procedure. The SVM models are implemented by the LIBSVM package (version 3.12) [34]. The radial basis function is chosen as the kernel function and two parameters including cost (c) and gamma (g) are optimized with the grid search strategy. To correct the imbalance between vastly outnumbered positive data and negative data, the weight parameter (w) of negative data is set as the ratio of negative data to positive data. Finally 10-fold cross validation is adopted in this work for feature selection and performance evaluation.

Input:
Structure feature set (**S**)
PPI feature set (**F**)
Maximal number of selected structure/PPI features M
Output:
Optimal feature subset (**O**)
Algorithm:
Normalize **S, F** to [0, 1]
Rank features in **S, F** with mRMR
while the number of selected features < M
Iteratively incorporate structure features in order and train a **SVM**
Calculate the AUC of the **SVM** model
end
O' ← features with max AUC value
while the number of selected features < M
Iteratively incorporate PPI features in order with **O'** and train a **SVM**
Calculate the AUC of the **SVM** model
end
O ← the subset with max AUC value

Performance Evaluation

The receiver operating characteristic (ROC) curve is the commonly used method to evaluate the overall performance of a classifier, which plots the value of (*1-Sp*, *Sn*) by using the decision values for all samples as thresholds, and the corresponding area under ROC curve (AUC) represents overall classification accuracy. In addition, accuracy (*Acc*), sensitivity (*Sn*), specificity (*Sp*), precision (*Pre*) and Matthews correlation coefficient (*MCC*) are utilized in this study to measure the identification performance at medium and high stringency levels, and the definitions are shown as below:

$$Acc = \frac{TN + TP}{TN + TP + FN + FP} \tag{4}$$

$$Sn = \frac{TP}{TP + FN} \tag{5}$$

$$Sp = \frac{TN}{TN + FP} \tag{6}$$

$$Pre = \frac{TP}{TP + FP} \tag{7}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}} \tag{8}$$

where the *TN*, *TP*, *FN* and *FP* represent the number of true negative, true positive, false negative and false positive, respectively.

Results

Analysis of structure and PPI information

Structure Information of Substrates. As protein kinases exhibit distinct recognition specificities in various structural surroundings, local structure information of the substrates, e.g. secondary structure, solvent accessibility and protein disorder region, could be helpful in recognizing corresponding kinases. In this regard, to assess the difference among secondary structures of substrates catalyzed by different kinases, the occurrence of coil, helix and beta strand in 21 positions around phosphorylation sites is calculated. For example, the substrates for two protein kinases: PKCa and PDK1, exhibit significant discrepancy in secondary structures with a *p*-value of 2.5485e-5 using Kolmogorov-Smirnov test, and the results are shown in in Fig. 1A. Compared with the substrates of PKCa, PDK1 substrates are enriched in coil but reduced in beta strand structures for the most positions. Especially, beta strands are completely missing in position -8~-3 and 6~7 but greatly enriched in position 0~1 of substrates catalyzed by PDK1, suggesting beta strands in these positions are helpful for discriminating between the substrates of PDK1 and PKCa. Moreover, the average disorder values of the 21-mer residues surrounding phosphorylation sites are also calculated. Fig.1B shows that these values range from 0.641 to 0.719 for PKCa and 0.350~0.493 for PDK1, and the *p*-value of their difference is 1.9664e-10. The results imply that substrates catalyzed by PKCa are more likely to be located in disordered regions and such information can be used for identification of PKCa substrates.

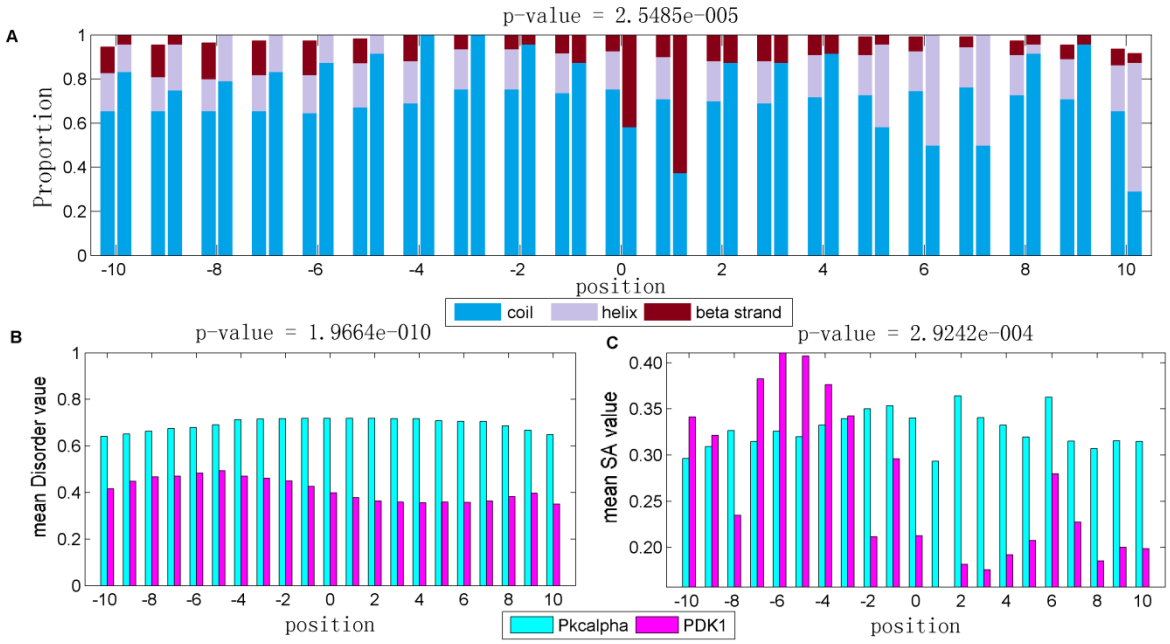


Fig. 1. Difference of structure features between kinase PKCa and PDK1. (A) Occurrence of coil, helix and beta strand in all 21 positions around phosphorylation sites. For each pair of bars, left one represent kinase PKCa and right one represent kinase PDK1 (B) Average disorder values of the 21-mer residues surrounding phosphorylation sites. (C) Average solvent accessibility values of the 21-mer residues surrounding phosphorylation sites. P-value is calculated by Kolmogorov-Smirnov test.

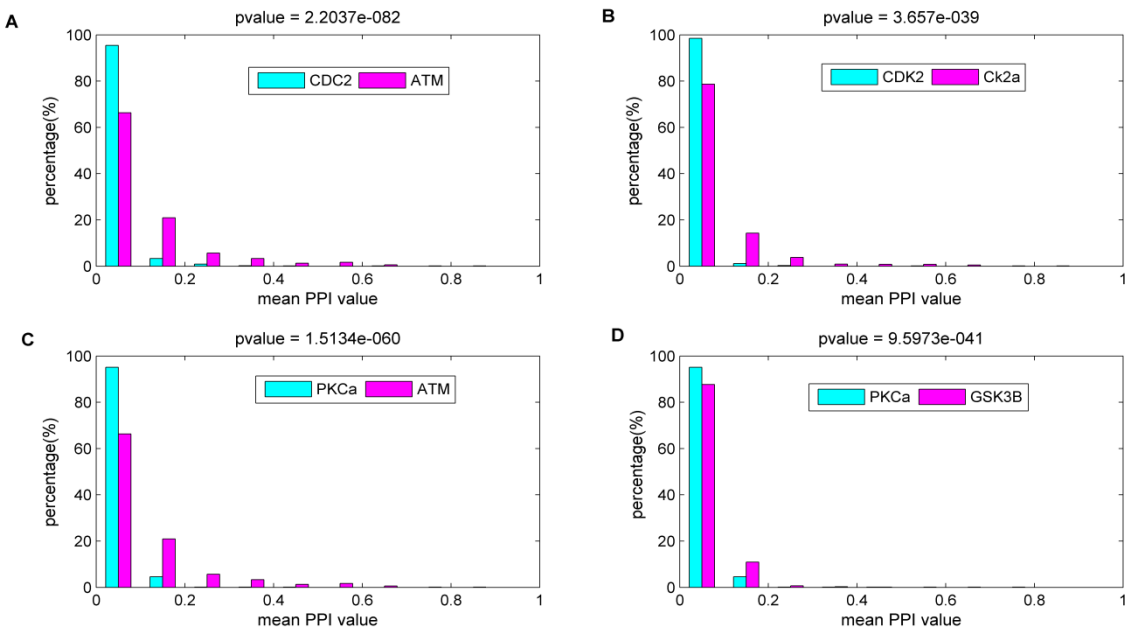


Fig. 2. Histogram of the top-ranked 1000 PPI values between different kinases. The horizontal axis represents the averaged PPI values for the substrates of a protein kinase and the vertical axis represents the percentage of average values in certain range.

Finally, our results also suggest the solvent accessibility between the substrates of PKCa and PDK1 is significantly different (Fig.1C). Therefore, all structure information is used as features for further study.

PPI Information. To evaluate PPI information, for each of the top1,000 PPI feature ranked by mRMR, the averaged PPI values for the substrates of a protein kinase are calculated. The histograms in Figure 2 show that there is significant difference between protein kinases, suggesting the PPI values vary greatly for different protein kinases. Further investigation shows there is generally little overlap between the top-ranked proteins of each two kinases, and the detailed information is shown in Table S2. There results indicate the substrates of different kinases tend to interact with divergent proteins that may be implicated in the kinase-specific phosphorylation process. We also intend to ascertain whether these top-ranking proteins of different kinases participate in disparate biological processes, and perform enrichment analysis of biological process by using the DAVID

program [35]. The results show that the top-ranked proteins of kinases own different but overlapping functions. For example, the top-ranked proteins for CDC2 and PKCa kinase are both functionally enriched in phosphoprotein. On the other hand, the proteins for CDC2 are enriched in mitotic cell cycle and cell cycle phase while proteins for PKCa are enriched in plasma membrane part and transmembrane protein. As the substrates of different kinases tend to interact with proteins involved in divergent biological processes, PPI information is useful to assign protein kinases.

Assessment of Feature selection

During the forward feature selection process, the corresponding AUC value in each loop is calculated and the results are shown in Figure 3. For comparison, a baseline SVM model is adopted which is trained with only sequence information (21-mer sequence fragment with 10 up- and down-stream peptides). For most of protein kinases, adding structure features to the baseline SVM leads to increased AUC, and the optimal prediction performance is usually obtained by using a subset of top-ranking structure features selected by the algorithm. Afterwards, with the

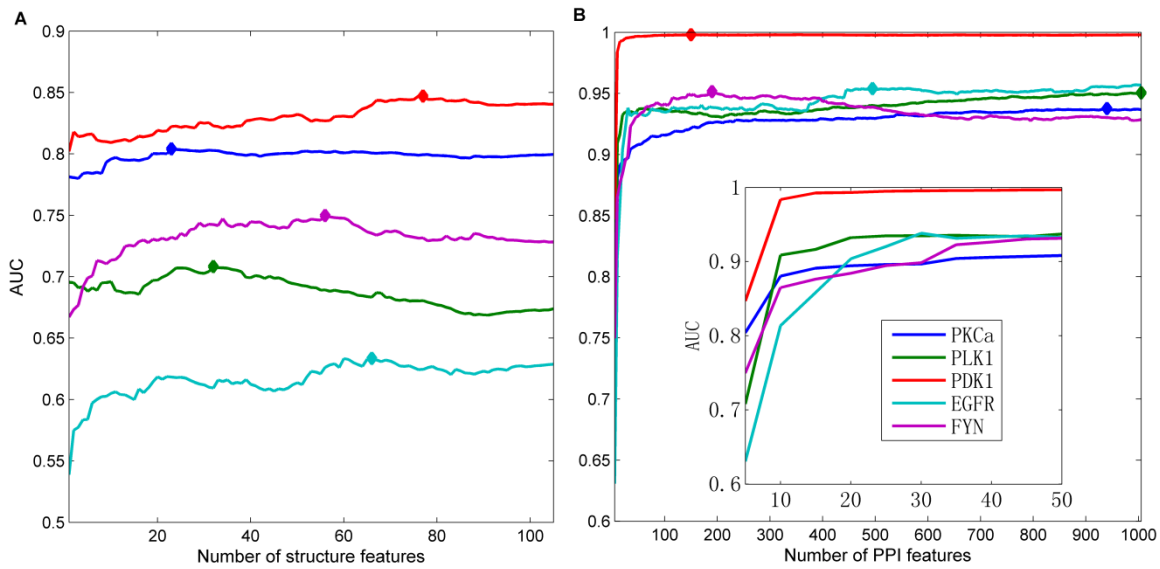


Fig. 3. AUC values obtained by feature selection process. (A) Feature selection process of structure features (B) Feature selection process of PPI features. The horizontal axis represents the number of added feature number.

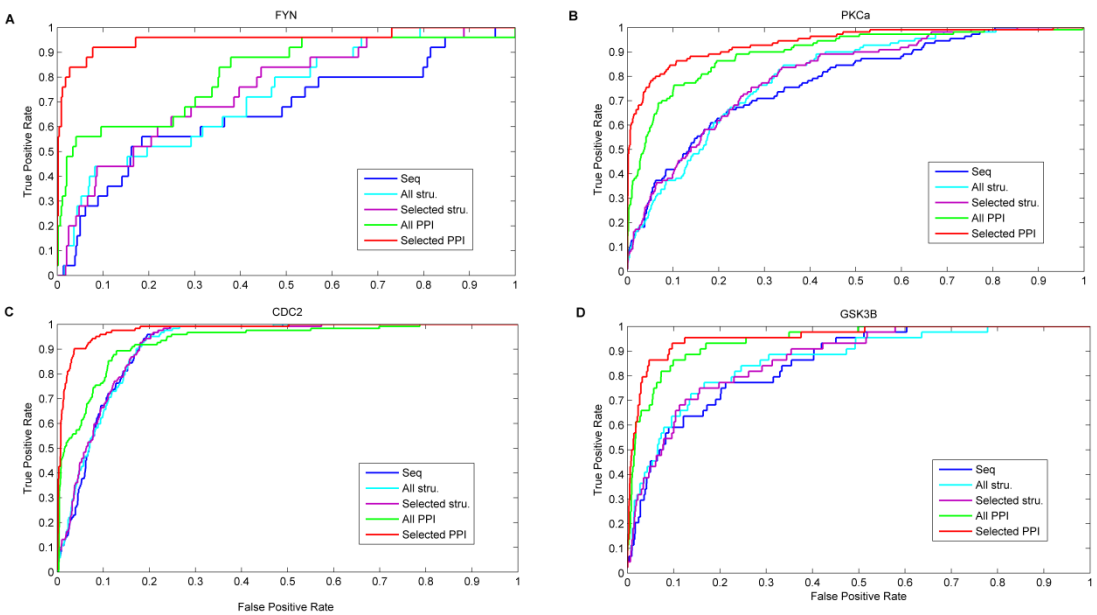


Fig. 4. ROC curves of 10-fold cross-validation performance using structure/PPI features. Seq.: local sequence features only; All stru.: primary sequence features and all structure features; Selected stru.: primary sequence features and selected structure features; All PPI: primary sequence features, selected structure features and all PPI features; Selected PPI: primary sequence features, selected structure features and selected PPI features

selected PPI features in the second step, prediction performance is further improved for all protein kinases (Fig.3). For example, with the aid of 55 selected structure features, the AUC of kinase FYN increases from 66.72% to 74.98%, and then boosts to 87.63% when 10 top-ranking PPI features are added. Finally, the optimal AUC reaches 95.15% when using 55 structure and 185 PPI features.

To further evaluate the proposed feature selection algorithm, we plot the ROC curves obtained by using selected structure/PPI features, as shown in Fig.4. In accordance with previous findings, identification performance for many protein kinases, such as FYN and PKCa, is increased by incorporating all structure features to the baseline SVM and can be further improved by feature selection. As the same time, for other kinases such as GSK3B and CDC2, it can be found that structure information is less helpful in identifying their substrates, and the reason may lie in: 1) predicted structure information for their substrates is inaccurate; 2) phosphorylation mediated by these kinases is not strongly affected by local structures of substrates. In either case, adoption of feature selection contributes little to performance. On the other hand, further incorporation of PPI features generally leads to obvious improvement in performance for all protein kinases, suggesting the importance of PPI information in determining kinases responsible for phosphorylation sites. However, it is unlikely that a great amount of proteins simultaneously participate in the phosphorylation process mediated by one kinase, and the high dimensions of PPI features may also bring trouble to classification and lead to decreased performance occasionally. Taking CDC2 kinase for instance, although adoption of all PPI features is generally beneficial for identification in other kinases, it renders even worse results at high sensitivity level than those of the baseline approach using only sequence information. Adoption of only a subset of PPI features determined by the feature selection algorithm can efficiently remedy this issue and provides remarkably enhanced performance for all protein kinases.

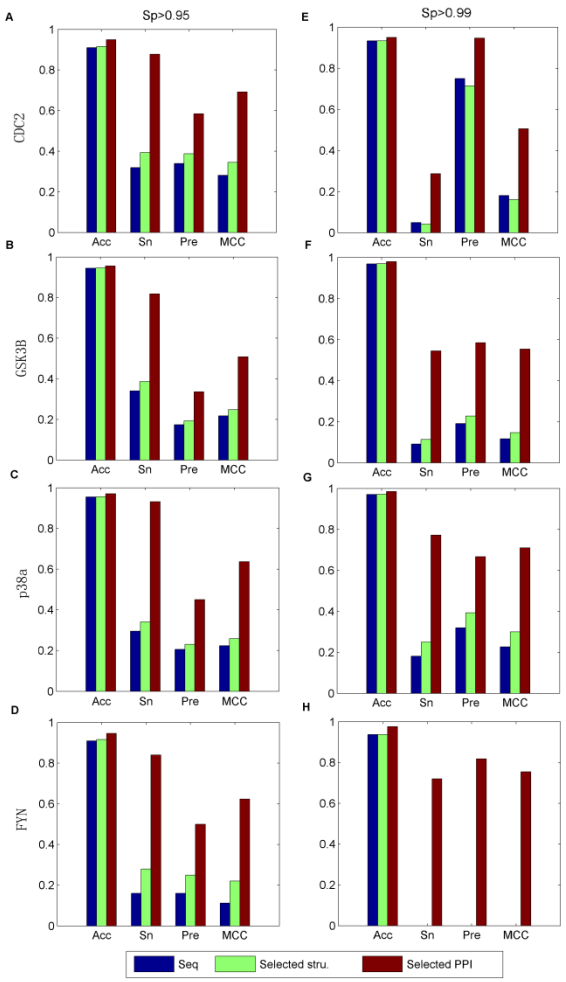


Fig.5. Identification performance of different feature sets at two stringency levels. (A-D) Measurements at specificity of 0.95. (E-H) Measurements at specificity of 0.99. The horizontal axis represents accuracy, sensitivity, precision and Matthew correlation coefficient, respectively. Seq.: local sequence features only; Selected stru.: primary sequence features and selected structure features; Selected PPI: primary sequence features, selected structure features and selected PPI features.

Performance Evaluation

In machine learning methods of protein phosphorylation, it is usually critical to minimize false positives, especially for proteomic-wide screening and systematic examination, and therefore stringent thresholds are commonly adopted to ensure high prediction specificity. In this regard, the influence of selected structure and PPI features on the performance of protein kinase identification is investigated at high specificity. For each kinase, two stringency levels are adopted with medium (>90.0%) and high (>99.0%) specificity, and commonly used performance measurements are calculated for comprehensive evaluation. As illustrated in Fig.5, incorporation of structure and PPI information optimized by feature selection algorithm significantly improves the performance at both stringency levels. For example, using the baseline SVM classifier with only sequence features, the performance of kinase FYN is rather low at medium stringency level (*Sp* 95.0%). Adding structure and PPI features yields consistent increase of all performance measurements and the final optimal values of *Acc*, *Sn*, *Pre* and *MCC* are 94.6%, 84.0%, 50.0% and 62.33%, respectively. Meanwhile, at high stringency level (*Sp* 99.09%) the SVM classifier cannot recognize any FYN substrates from 25 positive samples when using only sequence and structure information (*Sn* 0). After further incorporation of PPI information, most of the known FYN substrates can be identified with dramatically improved *Sn* of 72.0%.

Comparison with other existing tools. To further evaluate our identification method, the performance of SVM classifiers using selected structure and PPI features is compared with two existing state-of-the-art kinase identification tools: NetworKIN [19] and IGPS[21]. Since these tools do not provide options for cross-validation, all human phosphorylated proteins in Phospho.ELM database are submitted to these tools for identification. Furthermore, for our method 10-fold cross-validation was adopted for performance evaluation.

As shown in Figure 6 and Table S3, for most of the protein kinases our method shows significantly better performance than NetworKIN and IGPS. For example, for GSK3B kinase our

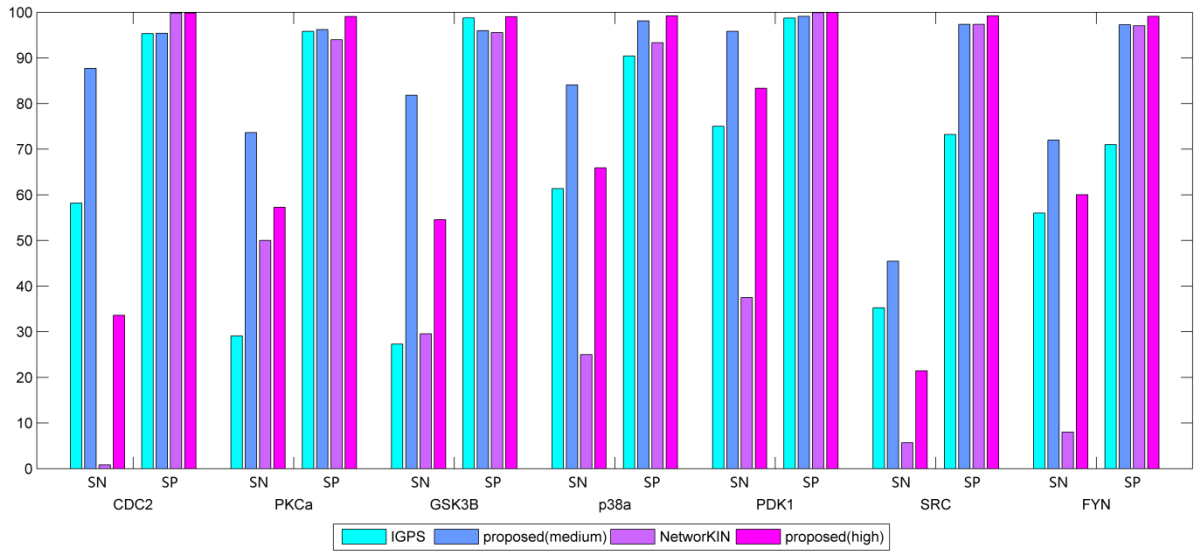


Fig.6. Performance comparison with existing tools: IGPS and NetworKIN. Two stringency level (medium: *sp*>90.0% and high: *sp*>99.0%) of the proposed method are used for comparison.

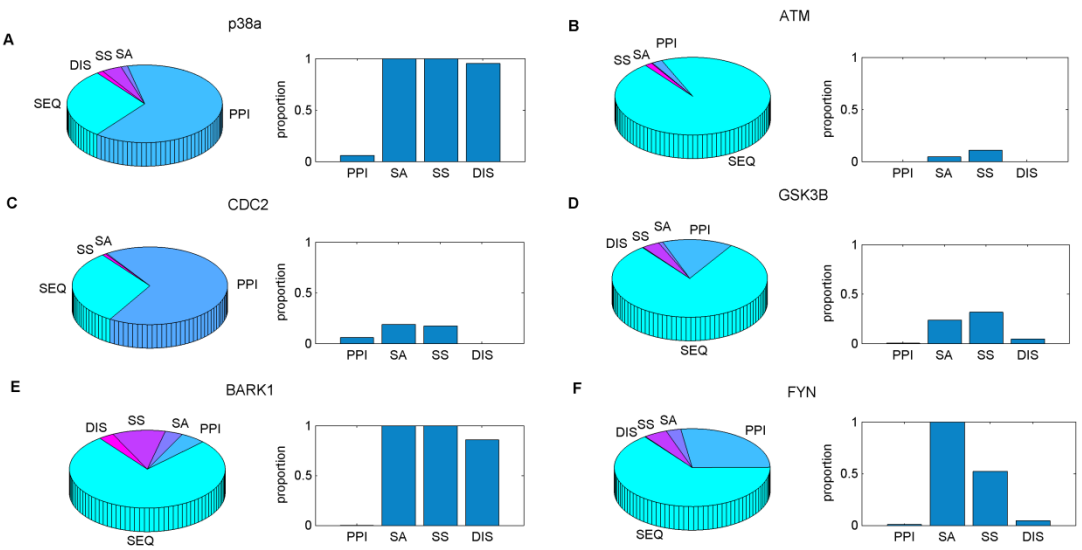


Fig.7. Analysis of selected features. For each kinase, pie graph illustrates the composition of all selected features and bar graph shows the proportion of selected features in different feature categories such as PPI, SEQ, DIS, SS and SA. (PPI: protein-protein interaction feature, SEQ: sequence feature, DIS: disorder region, SS: secondary structure feature, SA: solvent accessibility feature)

method achieves a sensitivity and specificity of 81.82% and 95.98% at medium stringency level, which outperforms NetworKIN (*Sn* 29.55% & *Sp* 95.53%) with about 52% higher sensitivity and similar specificity. Likewise, it also outperforms IGPS by an improvement of 27.28% in sensitivity at a high specificity of 99.03%. For only three protein kinases: ATM, PLK1 and SYK, IGPS shows better sensitivity with comparable or slightly lower specificity (Table S3). It is noteworthy that the results of IGPS are biased because the phosphorylation data in the Phospho.ELM database are also used by IGPS to generate prediction models (Song, *et al*., 2012), which inevitably leads to over-estimations of performance. On the contrary, the performance of our method is examined by 10-fold cross-validation, which can accurately reflect the true performance. In addition, we also evaluate the performance on different hierarchical levels of protein kinases, and Table S4 shows our method yields better overall performance in identifying kinase subfamilies, families and groups.

Interpretation of selected features

The distributions of selected features for different protein kinases are investigated for further interpretation, and the results are

shown in Fig.7. Intriguingly, although structure and PPI features are picked up by the feature selection algorithm, distinct patterns of feature composition for various kinases are found in the selected feature. For example, for protein kinase P38a, the majority of selected features are extracted from structure and PPI information. On the contrary, for some other kinases such as GSK3B, only a small proportion of structure and PPI features are selected. Given their prominent contribution to the performance of kinase identification (especially the selected PPI features), the results indicate that these features may represent key factors that are involved in the phosphorylation process. Further analysis of the selected features shows that beta strand structure and solvent accessibility of the phosphorylation site and its flanking position (-1) play an important role in determining the substrates of GSK3B kinase.

To better understand the biological significance of selected PPI features, functional enrichment analysis of the proteins associated with PPI features selected for GSK3B kinase is performed. Interestingly, it can be found that these proteins are enriched in glycogen metabolic process, glycogen biosynthesis, and glucan metabolic process (Table S5), which are in accordance with the well known regulatory function of GSK3B kinase in glycogen synthase. We further explore the literature and find one selected protein named Frat2 was reported to significantly increase GSK3B mediated phosphorylation of protein Tau by binding to kinase GSK3B [36]. Since the high PPI value for Frat2 and Tau (528) also indicates Frat2 can specifically interact with Tau, it is hypothesized that Frat2 can act as a bridge between substrate (Tau) and protein kinase (GSK3B) and then promote the phosphorylation progress through spatial location approximation (Fig. S1), by which the selected protein bring together the kinase and substrate to increase the rate of reaction [36]. Therefore, the PPI feature that shows selective binding of substrate protein to Frat2 is helpful in determining whether the corresponding kinase is GSK3B.

Discussions and Conclusions

With the increasing amount of phosphorylation sites discovered by high-throughput technologies, identification of corresponding protein kinase is attracting significant attentions. Although a few computational approaches have been proposed to this end, most of them are mainly focused on local sequence information, which are inadequate for accurate protein kinase identification. As various biological mechanisms are involved in protein phosphorylation process, the contributions of these factors may be very important and cannot be neglected. In addition, these approaches simply use PPI in post-processing filtering could lead to decreased prediction sensitivity, and leave a large room for fully utilizing the PPI information to improve the performance of protein kinase identification. In this work, a novel kinase identification approach is proposed, which incorporates an efficient two-step feature selection algorithm to handle tremendous protein-protein interaction and substrate structure information. The comprehensive analysis suggests that selected PPI and structure information are useful in discriminating corresponding kinase of phosphorylation sites at all hierarchical levels and the feature selection process is indispensable for decreasing high dimensionality of the input data.

Instead of using a binary coding scheme indicating the existence of interaction between two proteins, the PPI values from STRING database are adopted in this study that indicate the confidence score. We find that the feature selection algorithm tends to choose proteins that have strong interactions with both protein kinase and substrate, which may provide insights into the underlying mechanism of protein phosphorylation. For example, the PPI values for the interactions of Frat2 to GSK3B and Tau are 996 and 528, which are significantly larger than the average PPI values for these proteins. This indicates that the proteins associated with the selected PPI features may act as bridges between substrates and protein kinases to promote the phosphorylation progress. It is noteworthy that although there might be noise in the PPI data extracted from STRING database from predicted interactions, the results of performance evaluation suggest that the feature selection algorithm and the SVM models have good robustness.

The results of this study further confirms the conclusion in previous studies that protein structure information promotes the prediction performance of phosphorylation site, even though substrate structure information is not as helpful as PPI information in general. Besides the reason of inaccuracy in predicted structure information, it is also speculated that the weak effect of structure information in our study may be due to the distinction between prediction of phosphorylation site and identification of protein kinases. In other words, structural surroundings of substrates for different kinases are similar while the structural surroundings of phosphorylation sites and non-phosphorylation sites are various.

Although the approach proposed in this study demonstrates strong ability of kinase identification, the performance of a few kinases is still unsatisfied and therefore further improvement is needed. As a complex biological process, protein phosphorylation is affected by diverse biological mechanisms. Thus incorporating more biological information such as subcellular localization and evolutionary information may enhance the performance of protein kinase identification and help to reveal the intrinsic mechanism among protein substrate and kinases. Besides, more databases of experimental verified phosphorylation sites deposited in other bioinformatic resources such as PhosphoSitePlus could also be incorporated to construct better prediction models as more training data usually leads to improved classification performance. Furthermore, our study only focuses on identifying protein kinase of phosphorylation in human, but the proposed method that depends on PPI and structure information is generally applicable to other organisms and other kinds of post-translational modifications.

References

1. B. Trost and A. Kusalik, *Bioinformatics*, 2011, 27, 2927-2935.
2. J. Schlessinger, *Cell*, 2000, 103, 211-225.
3. D. M. Clifford, S. M. Marinco and G. S. Brush, *Journal of Biological Chemistry*, 2004, 279, 6163-6170.
4. T. Hunter, *Cell*, 2000, 100, 113-127.
5. F.-F. Zhou, Y. Xue, G.-L. Chen and X. Yao, *Biochemical and biophysical research communications*, 2004, 325, 1443-1448.
6. G. Manning, D. B. Whyte, R. Martinez, T. Hunter and S. Sudarsanam, *Science*, 2002, 298, 1912-1934.
7. B. Trost and A. Kusalik, *Bioinformatics*, 2013, 29, 686-694.

8. R. I. Brinkworth, R. A. Breinl and B. Kobe, *Proceedings of the National Academy of Sciences*, 2003, 100, 74-79.
9. M. Salinas, J. Wang, M. Rosa de Sagarra, D. Martín, A. I. Rojo, J. Martin-Perez, P. R. Ortiz de Montellano and A. Cuadrado, *FEBS letters*, 2004, 578, 90-94.
10. Z. Lin, P.-W. Zhang, X. Zhu, J.-M. Melgari, R. Huff, R. L. Spieldoch and G. R. Uhl, *Journal of Biological Chemistry*, 2003, 278, 20162-20170.
11. O. Nørregaard Jensen, *Current opinion in chemical biology*, 2004, 8, 33-41.
12. P. V. Hornbeck, I. Chabra, J. M. Kornhauser, E. Skrzypek and B. Zhang, *Proteomics*, 2004, 4, 1551-1561.
13. J. L. Heazlewood, P. Durek, J. Hummel, J. Selbig, W. Weckwerth, D. Walther and W. X. Schulze, *Nucleic Acids Research*, 2008, 36, D1015-D1021.
14. P. V. Hornbeck, J. M. Kornhauser, S. Tkachev, B. Zhang, E. Skrzypek, B. Murray, V. Latham and M. Sullivan, *Nucleic Acids Research*, 2012, 40, D261-D270.
15. A. Kreegipuu, N. Blom and S. Brunak, *Nucleic Acids Research*, 1999, 27, 237-239.
16. H. Dinkel, C. Chica, A. Via, C. M. Gould, L. J. Jensen, T. J. Gibson and F. Diella, *Nucleic Acids Research*, 2011, 39, D261-D267.
17. F. Diella, C. M. Gould, C. Chica, A. Via and T. J. Gibson, *Nucleic Acids Research*, 2008, 36, D240-D244.
18. F. Gnad, S. Ren, J. Cox, J. V. Olsen, B. Macek, M. Oroshi and M. Mann, *Genome biology*, 2007, 8, R250.
19. R. Linding, L. J. Jensen, A. Pasculescu, M. Olhovsky, K. Colwill, P. Bork, M. B. Yaffe and T. Pawson, *Nucleic Acids Research*, 2008, 36, D695-D699.
20. T.-Y. Lee, J. B.-K. Hsu, W.-C. Chang and H.-D. Huang, *Nucleic Acids Research*, 2011, 39, D777-D787.
21. C. Song, M. Ye, Z. Liu, H. Cheng, X. Jiang, G. Han, Z. Songyang, Y. Tan, H. Wang and J. Ren, *Molecular & Cellular Proteomics*, 2012, 11, 1070-1083.
22. J. C. Obenauer, L. C. Cantley and M. B. Yaffe, *Nucleic Acids Research*, 2003, 31, 3635-3641.
23. N. Blom, T. Sicheritz‐Pontén, R. Gupta, S. Gammeltoft and S. Brunak, *Proteomics*, 2004, 4, 1633-1649.
24. Y. Xue, J. Ren, X. Gao, C. Jin, L. Wen and X. Yao, *Molecular & Cellular Proteomics*, 2008, 7, 1598-1608.
25. J. A. Ubersax and J. E. Ferrell Jr, *Nature Reviews Molecular Cell Biology*, 2007, 8, 530-541.
26. D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller and P. Bork, *Nucleic Acids Research*, 2011, 39, D561-D568.
27. Y. Saeys, I. Inza and P. Larrañaga, *Bioinformatics*, 2007, 23, 2507-2517.
28. L. M. Iakoucheva, P. Radivojac, C. J. Brown, T. R. O'Connor, J. G. Sikes, Z. Obradovic and A. K. Dunker, *Nucleic Acids Research*, 2004, 32, 1037-1049.
29. H.-D. Huang, T.-Y. Lee, S.-W. Tzeng and J.-T. Horng, *Nucleic Acids Research*, 2005, 33, W226-W229.
30. I. Dondoshansky and Y. Wolf, *NCBI, Bethesda, Md*, 2002.
31. R. Adamczak, A. Porollo and J. Meller, *Proteins: Structure, Function, and Bioinformatics*, 2005, 59, 467-475.
32. S. Vucetic, C. J. Brown, A. K. Dunker and Z. Obradovic, *Proteins: Structure, Function, and Bioinformatics*, 2003, 52, 573-584.
33. H. Peng, F. Long and C. Ding, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2005, 27, 1226-1238.
34. C. Cortes and V. Vapnik, *Machine learning*, 1995, 20, 273-297.
35. B. T. S. Da Wei Huang and R. A. Lempicki, *Nature protocols*, 2008, 4, 44-57.
36. W. H. Stoothoff, J.-H. Cho, R. P. McDonald and G. V. Johnson, *Journal of Biological Chemistry*, 2004.

PPI and structure features extracted by a two-step feature selection algorithm can significantly enhanced the performance of kinase identification