Systems biology

PhosD: inferring kinase–substrate interactions based on protein domains

Gui-Min Qin¹,², Rui-Yi Li³ and Xing-Ming Zhao¹,*

¹School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China and ²School of Software and ³School of Computer Science and Technology, Xidian University, Xi’an 710071, China

*To whom correspondence should be addressed.

Abstract

Motivation: Identifying the kinase–substrate relationships is vital to understanding the phosphorylation events and various biological processes, especially signal transductions. Although large amount of phosphorylation sites have been detected, unfortunately, it is rarely known which kinases activate those sites. Despite distinct computational approaches have been proposed to predict the kinase–substrate interactions, the prediction accuracy still needs to be improved.

Results: In this paper, we propose a novel probabilistic model named as PhosD to predict kinase–substrate interactions based on protein domains with the assumption that kinase–substrate interactions are accomplished with kinase–domain interactions. By further taking into account protein–protein interactions, our PhosD outperforms other popular approaches on several benchmark datasets with higher precision. In addition, some of our predicted kinase–substrate relationships are validated by signaling pathways, indicating the predictive power of our approach. Furthermore, we notice that given a kinase, the more substrates are known for the kinase the more accurate its predicted substrates will be, and the domains involved in kinase–substrate interactions are found to be more conserved across proteins phosphorylated by multiple kinases. These findings can help develop more efficient computational approaches in the future.

Availability and Implementation: The data and results are available at http://comp-sysbio.org/phosd

Contact: xm_zhao@tongji.edu.cn

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Phosphorylation is one of the most important and ubiquitous post-translational-modifications (PTMs) and is involved in numerous biological processes, such as signal transduction and cell cycle (Olsen et al., 2006, 2010). In the phosphorylation process, a kinase will phosphorylate its substrates by adding a phosphate group to them. The deregulations between kinases and their substrates may lead to various diseases and cancers (Lahary et al., 2010). Therefore, dissecting the interactomes consist of kinases and their substrates can help understand various types of biological processes as well as the molecular mechanisms underlying diseases. In addition, the kinases are widely used as drug targets and the kinase–substrate relationships can help understand the mechanisms of drug actions (Brehm 2014; Naula et al., 2005).

Although thousands of phosphorylation sites can be detected in a single experiment thanks to the high-throughout technologies, most of the corresponding kinases that activate those sites remain unknown (Beausoleil et al., 2006; Han et al., 2010). For example, in the popular database Phospho.ELM that collects experimentally determined phosphorylation sites for multiple species (Diella et al., 2008), only 10% of those phosphorylation sites have the records of corresponding kinases. Similarly, less than 5% of the phosphorylation sites from PhosphoSitePlus have the records of corresponding kinases (Hornbeck et al., 2015). Therefore, different computational
tools have been developed for predicting the kinase–substrate relationships, where most of the existing approaches utilize the sequence information about the phosphorylation sites. For example, a curated database entitled as dbPAF was developed by integrating phosphorylation sites information from scientific literature and public databases, then potential kinases were predicted for phosphorylation sites based on the motifs around them (Ullah et al., 2016). Similarly, Xue et al. (2008) developed a tool named as GPS to predict the kinase–substrate interactions based on the assumption that similar kinases tend to target proteins containing similar linear sequence motifs, where the motifs contain phosphorylation sites that have been experimentally detected. By encoding the protein primary sequences, some machine learning approaches, such as support vector machines (SVMs) (Kim et al., 2004) and Bayesian decision tree (BDT) (Berger, 2013), have also been applied to predict substrates of kinases. For instance, Zou et al. developed a tool named as PKIS to identify protein kinases for known phosphorylation sites by encoding the protein sequences with composition monomer spectrum (Torii et al., 2009), which outperforms traditional binary coding approach (Zou et al., 2013). Except for sequence information, the structural information has also been explored to predict kinase-substrate relationships. For example, protein domains have been used to predict substrates of kinases (Damle and Mohanty, 2014; Liu and Tozeren, 2010). By defining an enrichment ratio between kinases and domains that measures preferential phosphorylation of the given domain by a specific kinase, PhosNetConstruct was proposed to predict substrates of kinases and showed that the prediction accuracy can be improved by protein domains compared with short sequence motifs (Damle and Mohanty, 2014). In fact, the phosphorylation event can be regarded as a specific type of protein–protein interaction (PPI). In literature, some approaches have been proposed to predict PPIs based on domains by assuming that the PPIs are achieved by domain-domain interactions (Hayashida et al., 2011; Wang et al., 2007; Zhao et al., 2010). It is not surprising about the good performance of protein domains considering they are the functional and structural units of proteins, and have been widely used in predicting drug–protein interactions (Wang et al., 2012) and microRNA–disease associations (Qin et al., 2016).

The phosphorylation events are complex processes that are determined not only by specific binding motifs but also by various context factors of the proteins, such as PPIs (Kobe et al., 2005). With this idea, NetworKIN (Linding et al., 2008) and iGPS (Song et al., 2012) improved prediction accuracy further by taking into account PPIs in addition to the sequence information about phosphorylation sites. By integrating more context factors, including PPIs, kinase-specific phosphorylation events and cell-cycle profiles, into a model, Patrick et al. (2015) proposed a Bayesian network model named as PhosphoPICCK to predict kinase–substrate relationships with promising results.

Despite the good performance of the above mentioned computational approaches, they perform differentially on distinct datasets and there is no one single approach that always performs best in all conditions. In this paper, we propose a novel robust probabilistic model named as PhosD to predict kinase–substrate relationships based on protein domains. By assuming that kinase–substrate relationships are accomplished with kinase–domain interactions, PhosD achieves robust results on three benchmark databases and outperforms other popular approaches with higher precision. In addition, some of our novel predictions are validated by known signaling pathways and provide insights into the signal transduction processes.

## 2 Materials and methods

### 2.1 Data sources

#### 2.1.1 Gold standard kinase–substrate interactions

In this paper, the kinase–substrate interactions were obtained from multiple databases, including PhosphoSitePlus (Hornbeck et al., 2015), PhosphoELM (Diella et al., 2008), PhosphoNetworks (Hu et al., 2014), UniProtKB (Magrane and Consortium, 2011), HPRD (Keshava Prasad et al., 2009) and BioGRID (Chatr-Aryamontri et al., 2015). These databases, especially the first two experimentally verified ones, have been widely used in previous works (Narushima et al., 2016). Table 1 shows the statistics about the kinase–substrate interactions from distinct sources. Specifically, for the interactions from PhosphoELM, we only considered those interactions that clearly state which kinases phosphorylate which proteins. The interactions between kinase group/family and proteins were not considered here. As for the kinase–substrate interactions from PhosphoSitePlus, only those labeled with in vivo were taken into account, where in vivo means that the interactions are determined with in vivo experiments. In this work, only the human kinome and their substrates were considered, and a Venn diagram in Supplementary Figure S1 shows the similarity and difference of these data sources.

#### 2.1.2 Negative samples

When investigating phosphorylation events in the lab, given a kinase, those proteins phosphorylated by the kinase are generally recorded while those not phosphorylated are not annotated, which makes it difficult to get the negative samples when predicting kinase–substrate interactions (Trost and Kusalik, 2011). Here, given the kinases and their substrates, the random combinations between kinases and their substrates were generated, and these random kinase–protein pairs except determined interactions were regarded as negative samples. For example, given m kinases, n substrates and r positive interactions, the left (m*n – r) interactions were considered as negative interactions.

#### 2.1.3 Protein–protein interactions, protein expression and protein domains

The PPIs were obtained from five databases, including BioGRID (Chatr-Aryamontri et al., 2015), HPRD (Chatr-Aryamontri et al., 2007), DIP (Xenarios et al., 2002) and IntAct (Kerrien et al., 2007). After integrating the five PPI databases, we collected 277 136 interactions among 20 020 proteins.

In addition, the protein expression data from the Human Proteome Map (HPM) were also utilized to explore the co-expression between proteins. HPM is a draft map of the human proteome generated with high-resolution Fourier-transform mass spectrometry (Kim et al., 2014), and the data consists of protein expression profiling across 30 histologically normal human tissues.

### Table 1. Statistics about different kinase–substrate interaction sources

<table>
<thead>
<tr>
<th>Data sources</th>
<th>#Kinases</th>
<th>#Substrates</th>
<th>#Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phospho.ELM</td>
<td>219</td>
<td>729</td>
<td>1239</td>
</tr>
<tr>
<td>PhosphoSitePlus</td>
<td>297</td>
<td>1419</td>
<td>2922</td>
</tr>
<tr>
<td>PhosphoNetworks</td>
<td>255</td>
<td>1140</td>
<td>4375</td>
</tr>
<tr>
<td>UniProtKB</td>
<td>198</td>
<td>429</td>
<td>649</td>
</tr>
<tr>
<td>HPRD</td>
<td>251</td>
<td>735</td>
<td>1466</td>
</tr>
<tr>
<td>BioGRID</td>
<td>286</td>
<td>1685</td>
<td>3108</td>
</tr>
</tbody>
</table>
including 17 adult tissues, 7 fetal tissues and 6 purified primary haematopoietic cells.

We collected protein domain annotations from InterPro (Mitchell et al., 2015), which is a freely available resource about protein domains and functional sites. With human proteins from SwissProt, we finally obtained 38,078 associations between 3309 domains and 14,287 proteins.

2.2 Computational model

The protein domains are structural and functional units of proteins. Based on this idea, as shown in Figure 1, we proposed a new computational approach, named as PhosD, to predict kinase–substrate interactions based on protein domains. First, the kinase–domain interactions will be inferred from known kinase–substrate interactions and protein–domain associations; second, given a kinase, its potential substrates will be predicted based on the domain compositions of proteins; Finally, the predictions results will be further refined with PPIs to improve prediction accuracy and coverage. The details will be addressed below.

2.2.1 Inferring kinase–domain interactions

We assumed that each kinase–substrate interaction was accomplished through kinase–domain interactions. For a specific kinase, the domains targeted by the kinase can be inferred based on the proteins phosphorylated by the kinase. Given a pair of kinase and domain \((K_i, D_j)\), the probability of their interaction can be defined as follows.

\[
P(K_i, D_j) = \frac{\sum_{m=1}^{M} I(K_i, S_m) * I(D_j, S_m)}{\sum_{m=1}^{M} I(K_i, S_m)}
\]

(1)

where \(S_m\) is the \(m\)th protein from the set of proteins considered with size of \(M\). \(I(K_i, S_m) = 1\) if \(S_m\) is the substrate of \(K_i\) and \(I(K_i, S_m) = 0\) otherwise, and the same for \(I(D_j, S_m)\).

2.2.2 Predicting kinase–substrate interactions

With the kinase–domain interactions obtained above, given a kinase, a protein will be considered as its candidate substrate if it contains at least one domain targeted by the kinase. Furthermore, considering that a kinase phosphorylates its substrates by physically interacting with those proteins and the interacting proteins may be phosphorylated by the same kinase. The probability of kinase \(K_i\) phosphorylating protein \(S_j\) can be defined as below.

\[
P(K_i, S_j) = P_d(K_i, S_j) * CF
\]

(2)

\[
P_d(K_i, S_j) = 1 - \prod_{j \neq D_j} (1 - P(K_i, D_j))
\]

(3)

where \(D_j\) is the \(j\)th domain in protein \(S_j\), \(P_d(K_i, S_j)\) is the probability of \(K_i\) phosphorylating \(S_j\) based on the domain composition of the protein. \(CF\) is a constraint factor that was proposed to reduce false positives based on PPIs. By taking into account, the physical interactions among proteins and the protein expression, \(CF\) was defined as follows.

\[
CF = \alpha * I(K_i, S_j) + (1 - \alpha) * W(K_i, S_j)
\]

(4)

\[
W(K_i, S_j) = \sum_{n=1}^{N_{S_j \neq S_i}} P_d(K_i, S_n) * PCC(S_n, S_j) * I(S_n, S_j) / N
\]

(5)

where \(I(K_i, S_j) = 1\) if \(K_i\) and \(S_j\) interact and 0 otherwise, and the same for \(I(S_n, S_j)\). \(W(K_i, S_j)\) is the probability of \(S_j\) being the substrate of \(K_i\) based on the set of known substrates of \(K_i\), i.e. \(SP\). \(PCC(S_n, S_j)\) is the Pearson correlation coefficient between \(S_n\) and \(S_j\) based on their protein expression profiles across 30 tissues. It can be seen from Equation (4) that the first term describes the direct interactions between a pair of kinase and protein while the second term describes their indirect interaction based on the path \(K_i \rightarrow SP \rightarrow S_j\). If \(S_j\) interacts both directly and indirectly with \(K_i\), it is more likely that \(S_j\) is the substrate of kinase \(K_i\). The parameter \(\alpha\) was used to balance the probabilities of \(K_i \rightarrow S_j\) interaction inferred based on the direct and indirect neighbors of \(S_j\).

To determine the parameter \(\alpha\) in \(CF\) defined above, we investigated the kinase–substrate relationships in the training dataset to see how much of them have direct or indirect interactions. Basically, \(\alpha\) was determined based on the frequency of direct and indirect interactions among known kinase–substrate interactions in the training data as shown in Supplementary Figure S3 and Supplementary Methods. Here indirect interactions were considered to improve the coverage of our predictions since the kinase–substrate interactions may not be found in existing PPIs due to the incomplete interactome.

3 Results and discussion

3.1 Performance evaluation

In this paper, to evaluate the performance of PhosD, we compared it with other popular approaches, including GPS (Xue et al., 2008), iGPS (Song et al., 2012), NetworkKIN (Linding et al., 2008), PKIS (Zou et al., 2013) and PhosphoPICK (Patrick et al., 2015). GPS
predicts site specific kinase–substrate relationships based on position-specific scoring matrices inferred from sequence information. iGPS is proposed based on the results of GPS and further filters out false-positives with PPI information. NetworKIN models kinase–substrate interactions with artificial neural networks and shows high prediction accuracy. PKIS integrates the monomer spectrum (CMS) encoding strategy with SVMs to identify kinases associated with known phosphorylation sites. PhosphoPICK infers kinase–substrate interactions based on a Bayesian network model by combining PPIs and protein abundance information. All the computational approaches were compared with respect to precision, recall and F1. Precision means the ratio of predicted positives that are true positives, recall denotes the ratio of true positives that can be correctly predicted, and F1 is the harmonic mean of Precision and Recall and can reflect the overall performance. The detailed definition of the three indexes can be found in Supplementary Methods.

Among the above popular computational approaches, distinct ones are trained on different data sets since they are developed based on different datasets and have their own training data used in their tools. For example, GPS was developed based on Phospho.ELM. Table 2 shows the training data used by different methods. In order to compare in a fair way, given an approach, we trained PhosD with the same training data as used by the approach. The results of other approaches were obtained by either running corresponding softwares or submitting the data to online web servers. Considering the differences in both training and test datasets, we compared PhosD with each of the other approaches separately. For example, PKIS only tested 56 kinases by limiting the number of known phosphorylation sites (Zou et al., 2013), while PhosphoPICK performs predictions for 105 kinases (Patrick et al., 2015). In fact, we used Phospho.ELM as the training data for PhosD when compared with the first four methods, and both Phospho.ELM and HPRD were used when compared with PhosphoPICK.

To evaluate the performance of distinct approaches, four test datasets were used here, including PhosphoSitePlus, UniProtKB, PhosphoNetworks and BioGRID. Figure 2 shows the results of the six approaches to be compared. It can be seen that all the other five methods lead to a lower precision than PhosD. Compared with GPS, the performance of iGPS improves a lot by taking into account PPIs. NetworKIN works in a similar way as iGPS by considering functional associations between kinases and substrates from STRING (Szklarczyk et al., 2011). Like most of the methods based on known phosphorylation sites, PKIS yields a higher recall that is 12.5% higher than PhosD but with low precision. On the UniProtKB dataset, PhosD outperforms GPS, iGPS and NetworKIN with respect to both precision and recall. From the results on PhosphoSitePlus, we can see that PPIs can indeed help remove false positives when predicting kinase substrates with the cost of low recall. Compared with all other approaches, our proposed PhosD achieves the highest precision and has the best overall performances, i.e. F1, indicating the predictive power of our proposed PhosD. The detailed prediction results can be found in Supplementary Table S1.

By looking into the domains that have been predicted to interact with kinases, we found some specific patterns for the substrates phosphorylated by the same kinase. For example, the kinase PDGFRB phosphorylates six proteins, i.e. P12931, P09619, P06241, Q05397, P13180, Q04829, and Q05507.
Q13330 and Q06124, and the domain IPR001245 was predicted by PhosD to interact with PDGFRB. As shown in subfigure (a) of Supplementary Figure S2, IPR001245 appears in four substrates indicating that the phosphorylation events can indeed be predicted with domains. In another example shown in subfigure (b) of Supplementary Figure S2, domain IPR000719 appeared in 18 substrates for kinase PDPK1 which have 19 substrates in total, where the domain was predicted to interact with PDPK1. These two cases clearly demonstrate the conservation of domains across substrates phosphorylated by the same kinase.

What is interesting is that the phosphorylation sites are not necessarily located in the domains and even randomly distributed around the domain. This phenomenon may help explain the good performance of PhosD over other sequence-based approaches.

### 3.2 Comparison with another domain-based method

Similar to PhosD, PhosNetConstruct is another domain-based method that predicts substrates of kinases. For each kinase–domain pair, PhosNetConstruct defines an enrichment ratio (ER) that quantitatively describes the preferential phosphorylation of a given domain by a specific kinase compared to all other kinases. Based on ER, PhosNetConstruct recognizes the substrates of the kinase of interest. Here, to evaluate the performance of PhosD, we compared it against PhosNetConstruct.

For fair comparison, PhosphoSitePlus was used as the training set (the detailed predicted results can be found in Supplementary Table S2) for both PhosNetConstruct and PhosD, and the test set was obtained from the supplementary file of PhosNetConstruct (Damle and Mohanty, 2014). Since some proteins have no PPIs and protein expressions available, we only focused on those kinases and proteins that have both interaction partners and expression information. As a result, the test set used here consists of 146 substrate proteins phosphorylated by 10 kinase families. The results on these kinases and substrates by PhosNetConstruct were compared with those by PhosD as shown in Table 3. Considering the way domains used, four types of results were given by PhosNetConstruct (Damle and Mohanty, 2014): with (Phospho-Domain in Table 3) or without (Any Domain in Table 3) limiting the phosphorylation sites to protein domains together with (filter in Table 3) or without (no-filter in Table 3) requiring an ER threshold.

From the results, we can see that PhosNetConstruct yields higher recall without limiting the phosphorylation sites to protein domains, and the precision can be significantly improved with enrichment ratio. Compared with PhosNetConstruct, PhosD has the best overall performance with the highest F1 score and achieves the highest precision, implying the predictive power of our PhosD. In addition, from the results shown above, it can be seen that only part of the phosphorylation sites lie in the domain regions got better performance, which can explain the good performance of PhosD to some extent.

### 3.3 The factors affecting the performance of PhosD

As shown above, PhosD outperforms other popular approaches on multiple benchmark datasets. The good performance of PhosD is due to the assumption about kinase–domain interactions underlying kinase–substrate relationships as well as the consideration of PPIs. In this section, we further investigated possible factors affecting the performance of PhosD.

In PhosD, both direct and indirect interactions with a kinase were considered for a give protein, and these two types of interactions were combined into a constraint factor (CF). To see the effect of PPIs on the performance of PhosD, we compared the results by PhosD with or without CF as shown in Table 4. With the data from Phospho.ELM as training set, Table 4 shows the results of PhosD on the other three test sets, i.e. PhosphoNetworks, PhosphoSitePlus and UniProtKB. It can be seen from the results that by taking into account PPIs, the prediction precision of PhosD can be significantly improved. Even at the cost of low recall, the overall performance of PhosD was improved with higher F1 by taking into account the PPIs described by CF. We also investigated the effect of the parameter alpha from CF on the performance of PhosD. As shown in Supplementary Figure S4, PhosD demonstrates robust and good performance within 0.6–0.9, and the detailed values of alpha under different conditions were shown in Supplementary Table S6.

Since we supposed that the proteins containing the same domain(s) will be phosphorylated by the same kinase, given a kinase, we expected that the more its substrates are known the more accurate our predictions will be. Therefore, we investigated the influence of the number of known substrates on the performance of PhosD. For this, based on the number of their known substrates as annotated in Phosphor.ELM database, we classified the kinases into two groups, i.e. the class of ‘SNum < 10’ and the class of ‘SNum > 10’, where SNum means the number of known substrates. Figure 3 shows the performance of PhosD on two different types of kinases. From the results on three different test sets, including PhosphoSitePlus, UniProtKB and PhosphoNetworks, we can see that with more substrates available, the prediction precision of PhosD is significantly improved. The better performance of PhosD on kinases with more than 10 substrates implies that the kinase–domain interaction patterns can be better discovered with more substrates are known for each kinase, which also indicates that kinase–substrate interactions are indeed accomplished with kinase–domain interactions.

On the other hand, one protein may be phosphorylated by one specific kinase. It is expected that PhosD should perform better on the ‘one protein targeted by one kinase’ phosphorylation events.
Surprisingly, we found that PhosD performs much better on those proteins phosphorylated by multiple kinases with leave-one-out cross-validations as shown in Figure 4, where the three most reliable and experimentally determined phosphorylation datasets were used. This phenomenon indicates that domains are more conserved in those proteins phosphorylated by multiple kinases. To test this hypothesis, we investigated the domain compositions of those phosphorylated proteins and found the domains are indeed more conserved for those proteins targeted by multiple kinases (detailed results can be found in Supplementary Table S3). In addition, we investigated the interaction partners of the two types of phosphorylated proteins for each kinase based on the score defined by Equation (6), which reflects the interaction intensity among substrates targeted by a kinase. We noticed that the proteins phosphorylated by multiple kinases tend to interact with each other compared with those targeted by single kinases (P-value 3.02E−03 for Phospho.ELM, P-value 2.13E−07 for PhosphoSitePlus, P-value 3.02E−03 for HPRD). The detailed results can be found in Supplementary Table S4. The detailed results can be found in Supplementary Table S4.

\[
\text{Clust}_j(K) = 2 \sum_{i=1}^{n-1} E_{ij}/(n(n-1))
\]

where \(n\) is the size of known substrate set (i.e. \(K\)) for a kinase of interest, and \(E_{ij} = 1\) if protein \(i\) interact with protein \(j\) and \(E_{ij} = 0\) otherwise.

From the results shown above, we can see that the more substrates are known for a kinase, the more accurate its predicted substrates will be. Furthermore, the domains involved in kinase-domain interactions are more conserved across proteins targeted by multiple kinases. These findings can help predict potential substrates for a kinase in the future.

3.4 Understanding signal transductions with new phosphorylation events

It is well known that phosphorylation plays important roles in signal transduction (Zhao et al., 2008). With our predicted kinase-substrate interactions, it is expected that more details about signaling pathways can be uncovered. We first retrieved 39 signaling pathways from KEGG (Kanehisa and Goto, 2000) with cyKEGGparser (Nersisyan et al., 2014). By limiting the kinases and their substrates to those only occurring in signaling pathways, we investigated how many of those kinase–substrate interactions can be predicted by PhosD. With all the five databases as training set, we found that 87 kinase–substrate interactions from our novel predictions occurred in the signaling pathways while these interactions cannot be found in the five databases (Supplementary Table S5).

For example, the two pairs of AKT3–GS3KB and PDPK1–PRKCI were identified by PhosD but missed in the five popular phosphorylation databases. These two interactions play important roles in the insulin signaling pathway that is closely related to
diabetes and aging (Krüger et al., 2008). With the two newly detected phosphorylation events, the signal transduction within the insulin signaling pathway can be better understood. As shown in Figure 5a, 3-phosphoinositide-dependent protein kinase 1 (PDPK1) activates Akt, a serine kinase, which in turn deactivates glycogen synthase kinase 3 (GSK-3), leading to the activation of glycogen synthase (GYS) and glycogen synthesis. As shown in Figure 5b, the two phosphorylation events of AKT3–GSK3B and AKT3–RAF1 involved in PI3K/Akt signaling pathway that is implicated in a number of human diseases including cancer, diabetes, cardiovascular disease and neurological diseases (Hers et al., 2011), were identified by PhosD. In fact, Zimmermann et al. have shown that Akt can phosphorylate Raf in vitro (Zimmermann and Moelling, 1999). The Neurotrophin signaling pathway plays important roles in neural development and higher-order activities such as learning and memory (Huang and Reichardt, 2003). As shown in Figure 5c, with our predicted NTRK1 (Trk A)–SHC2, NTRK1–IRS1, NTRK1–SH2B2, AKT3–GSK3B, MAPK7–RPS6KA6 and MAP3K3–MAP2K5 interactions, we can learn how the signals flow from the Trk receptors (Trk A, Trk B, Trk C) to GSK3β. The inhibitors of GSK3 have been clinically proven effective for antidepression, which indicates the important role of Akt in synaptic transmission, memory and psychosis. From the findings shown above, we can see that the newly detected kinase–substrate interactions can help us better understand the signal transduction processes. In addition, the newly discovered kinase–substrate interactions unveils how the aberrant signaling pathways lead to diseases, which is important for designing effective target therapies.

4 Conclusion

Protein phosphorylation plays important roles in various biological processes, such as sugar metabolism, growth and signal transduction, and so on. However, the lack of information about which kinases phosphorylate which proteins makes it difficult to understand the phosphorylation events as well as the biological processes in which they are involved. In this paper, we presented a novel probabilistic model, namely, PhosD, to predict the kinase–substrate relationships. By assuming that the kinase–substrate interactions are achieved by kinase–domain interactions, PhosD identifies the kinases that phosphorylate corresponding proteins by further considering PPIs and outperforms other popular approaches on three benchmark datasets. Some of our novel predictions were validated by signaling pathways in which they were involved, indicating the predictive power of our proposed PhosD. In addition, it was found that the number of known substrates will affect the prediction accuracy when predicting new substrates for a kinase and the domains in proteins targeted by multiple kinases are more conserved. These findings may help design more accurate predictive models in the future.

We also noticed that there is much space to improve our approach. For example, the domains are not independent of each other within a protein, the synergistic interactions among domains may help improve the prediction accuracy. Furthermore, with more phosphorylation sites within its substrate, a kinase can phosphorylate the substrate with higher affinity. Therefore, the number of phosphorylations sites may also help improve the prediction results. In addition, the kinase–substrate interactions may be predicted with higher precision if direct association approaches, e.g. PMI (Zhao et al., 2016), are employed in the future.

Funding

This work was supported by the National Natural Science Foundation of China [61572363, 91530321, 61602347], and the Fundamental Research Funds for the Central Universities [BDZ011401, JB151005].

Conflict of Interest: none declared.

References

Huang, E.J. and Reichardt, L.F. (2003) Trk receptors: roles in neuronal signal transduction, and so on. However, the lack of information about which kinases phosphorylate which proteins makes it difficult to understand the phosphorylation events as well as the biological processes in which they are involved. In this paper, we presented a novel probabilistic model, namely, PhosD, to predict the kinase–substrate relationships. By assuming that the kinase–substrate interactions are achieved by kinase–domain interactions, PhosD identifies the kinases that phosphorylate corresponding proteins by further considering PPIs and outperforms other popular approaches on three benchmark datasets. Some of our novel predictions were validated by signaling pathways in which they were involved, indicating the predictive power of our proposed PhosD. In addition, it was found that the number of known substrates will affect the prediction accuracy when predicting new substrates for a kinase and the domains in proteins targeted by multiple kinases are more conserved. These findings may help design more accurate predictive models in the future.

We also noticed that there is much space to improve our approach. For example, the domains are not independent of each other within a protein, the synergistic interactions among domains may help improve the prediction accuracy. Furthermore, with more phosphorylation sites within its substrate, a kinase can phosphorylate the substrate with higher affinity. Therefore, the number of phosphorylations sites may also help improve the prediction results. In addition, the kinase–substrate interactions may be predicted with higher precision if direct association approaches, e.g. PMI (Zhao et al., 2016), are employed in the future.


