BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 12 2014, pages 1730–1738
doi:10.1093/bioinformatics/btu112

Systems biology

Advance Access publication February 25, 2014

Deciphering kinase–substrate relationships by analysis of domain-specific phosphorylation network

Nikhil Prakash Damle and Debasisa Mohanty*

Bioinformatics Centre, National Institute of Immunology, Aruna Asaf Ali Marg, New Delhi 110067, India

Associate Editor: Janet Kelso

ABSTRACT

Motivation: *In silico* prediction of site-specific kinase–substrate relationships (ssKSRs) is crucial for deciphering phosphorylation networks by linking kinomes to phosphoproteomes. However, currently available predictors for ssKSRs give rise to a large number of false-positive results because they use only a short sequence stretch around phosphosite as determinants of kinase specificity and do not consider the biological context of kinase–substrate recognition.

Results: Based on the analysis of domain-specific kinase–substrate relationships, we have constructed a domain-level phosphorylation network that implicitly incorporates various contextual factors. It reveals preferential phosphorylation of specific domains by certain kinases. These novel correlations have been implemented in PhosNetConstruct, an automated program for predicting target kinases for a substrate protein. PhosNetConstruct distinguishes cognate kinase–substrate pairs from a large number of non-cognate combinations. Benchmarking on independent datasets using various statistical measures demonstrates the superior performance of PhosNetConstruct over ssKSR-based predictors.

Availability and implementation: PhosNetConstruct is freely available at http://www.nii.ac.in/phosnetconstruct.html.

Contact: deb@nii.res.in

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 30, 2013; revised on February 3, 2014; accepted on February 19, 2014

1 INTRODUCTION

Phosphorylation of a variety of proteins by Ser/Thr or Tyr kinases is one of the major post-translational modifications, which regulates signal transduction events in different metabolic and disease-related processes within a cell. Therefore, a detailed understanding of protein phosphorylation networks is crucial for decoding the molecular basis of various cellular processes. However, an exponential increase in kinome information and identification of variety of phosphoproteins across different organisms (Bodenmiller and Aebersold, 2011; Mok *et al.*, 2010; Newman *et al.*, 2013; Prisic *et al.*, 2010) has made complete experimental characterization of phosphorylation networks enormously difficult. Even though several sequence as well as structure-based methods (Blom *et al.*, 2004; Kumar and Mohanty, 2010; Obenauer *et al.*, 2003; Saunders and Kobe, 2008; Xue *et al.*, 2006, 2008) are available for prediction of site-specific kinase–substrate relationships (ssKSRs), they are ideally designed for identification of phosphosites in a given substrate protein of a kinase. Hence, straightforward application of these methods to decipher phosphorylation networks by prediction of kinases that can potentially phosphorylate a given substrate protein can lead to the accumulation of false-positive results. This is because every protein irrespective of whether it is actually phosphorylated is treated as a putative substrate. It is well known that approximately only one-third of the entire proteome is phosphorylated (Cohen, 2000), and phosphorylation events are part of dynamic and directional protein interaction networks.

The context-dependent factors that are likely to govern the kinase–substrate recognition process are cellular localization, coregulation of interacting pairs of proteins and also the presence of other docking/interacting domains in the same polypeptide or coevolution of kinase–substrate pairs as interacting domains. However, these contextual factors remain outside the purview of the sequence motif-based methods for prediction of ssKSRs owing to the consideration of small stretches of amino acids on substrates as determinants of kinase specificity. Therefore, recently developed approaches like NetworKIN (Linding *et al.*, 2007) and iGPS (Song *et al.*, 2012) have combined motif/profile-based predictions with information about the functional association between kinases and substrates derived from STRING (Szklarczyk *et al.*, 2011) or other databases containing experimental protein–protein interaction (PPI) information. These studies have demonstrated that inclusion of interaction context leads to improved prediction of ssKSRs. However, direct information about functional association or PPIs are only available for a limited number of kinase–substrate pairs and hence it is necessary to explore alternate approaches for prediction of phosphorylation networks by inclusion of context-dependent interaction probability for a kinase–substrate pair.

Several contextual factors are likely to better correlate with sequence signatures at the globular domain level rather than signatures in short sequence motifs. Domain-level sequence signatures in polypeptide chains are assigned using Pfam (Finn *et al.*, 2010) or InterPro-type (Hunter *et al.*, 2012) functional domain definitions. These are derived from the HMM profiles obtained from multiple sequence alignments of related proteins. In fact, earlier studies by Sprinzak and Margalit (2001) have demonstrated that, correlated InterPro domain-level sequence signatures could be used as markers of PPIs. Liu and Tozeren (2010). have also attempted to predict substrates of kinases based on co-occurrence of pairs of interacting domains (Liu and Tozeren, 2010). Therefore, in this study, we have constructed a domain-level kinase–substrate network based on Pfam (Finn

*To whom correspondence should be addressed.

et al., 2010) domain analysis of the experimentally identified mammalian kinase–substrate pairs available in PhosphoSitePlus (Hornbeck *et al.*, 2012) and recently published activity-based human phosphorylation network (Hu *et al.*, 2014; Newman *et al.*, 2013). Systematic analysis of this kinase–phosphodomain network has revealed novel patterns of preferential phosphorylation of certain Pfam domains by specific kinases. We also demonstrate that these preferential domain-specific kinase–substrate relationships (dsKSRs) can be used to distinguish cognate kinase–substrate pairs from all other non-cognate combinations. Extensive benchmarking on a completely independent test set indicates that by using dsKSRs, it is possible to predict putative target kinases that are likely to phosphorylate given substrate proteins.

2 METHODS

2.1 Clustering of kinases based on their domain phosphorylation profiles

Figure 1 shows various steps involved in the compilation of phosphorylation data and construction of an experimentally identified domain phosphorylation network. To have a significant number of phosphorylation events for each kinase–substrate pair in the dataset, kinase–substrate pairs were selected from PhosphoSitePlus (Hornbeck *et al.*, 2012) with the criterion that each of the protein kinases had at least 20 substrate proteins. This resulted in 2642 substrate proteins phosphorylated by 61 different protein kinases. The domain compositions for each of these substrate proteins and their target protein kinases were deciphered using Pfam (Finn *et al.*, 2010) functional domain definitions. Phosphorylation sites on each of the 2642 substrate proteins were mapped onto the corresponding Pfam functional domains resulting in 856 unique phosphodomains (Supplementary Data S1 and S2). Further details are given in Supplementary Methods. Domain phosphorylation profiles were represented by a binary matrix of dimension 61×856, where each element consisted of 1 or 0 depending on whether a selected domain is phosphorylated by a given kinase. The phosphorylation patterns of interdomain linker stretches were also represented by a similar matrix of dimensions 61×490. The distance between a pair of kinases in terms of their domain phosphorylation profiles was computed using the Jaccard coefficient between domain phosphorylation vectors followed by hierarchical clustering and the result was represented as a heat map using the pheatmap module of *R* package (www.r-project.org).

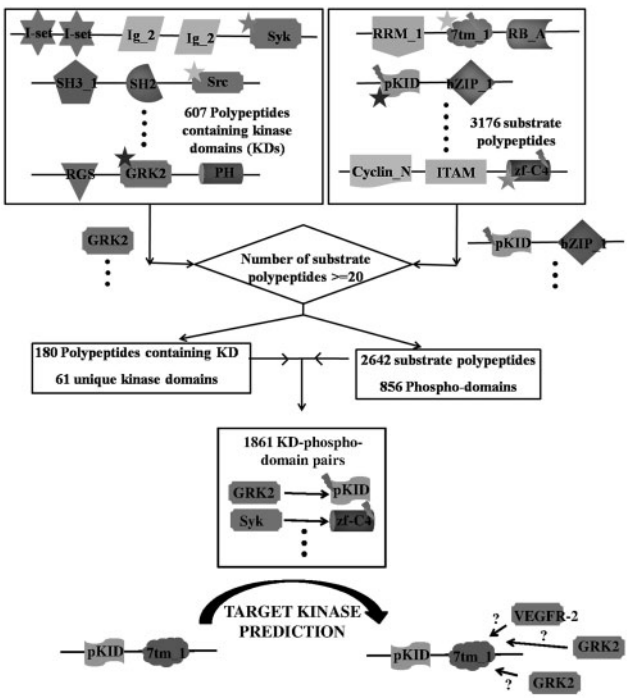

Fig. 1. Schematic representation of the protocol followed for compiling the training dataset for kinase–domain interactions. Pfam domain organization of polypeptides containing kinase domains and their substrate proteins was analyzed for those kinases having a minimum of 20 distinct substrate proteins. The cognate kinase–domain pairs obtained by this approach were analyzed to understand dsKSRs and derive predictive rules for identifying the putative target kinases for phosphoproteins based on their domain organization

2.2 Estimating relative preference of different kinases to phosphorylate a given Pfam domain

The propensities of different kinases to phosphorylate different Pfam domains were estimated based on the number of occurrences of a cognate kinase–phosphodomain pair in our training dataset (Supplementary Data S2) by using the protocol described in Supplementary Methods. In brief, target signal P(Ki,Dj) for a given kinase–phosphodomain pair (Ki,Dj) was computed as the probability of a domain Dj to be phosphorylated in the substrates of kinase Ki. The background signal for this pair is the probability of the same domain Dj being phosphorylated by all other kinases (K'i). The ratio of target signal to background signal was called the enrichment ratio (ER) for a given kinase–domain pair, i.e. $ER(Ki,Dj) = P(Ki,Dj)/P(K'i,Dj)$ (Supplementary Data S3). The ER thus gives a quantitative measure of preferential phosphorylation of the given domain by a specific kinase compared with all other kinases. For each kinase, similar ERs were also computed for all other auxiliary domains that occur on the same polypeptide as that of the substrates of kinases irrespective of their phosphorylation status.

2.3 Benchmarking of PhosNetConstruct

PhosNetConstruct was benchmarked on two different test datasets that were completely independent of the dataset used for training. The first test dataset was chosen carefully from PhosphoELM (Dinkel *et al.*, 2010), comprising substrate proteins for which target protein kinases were known from published experimental studies. This will be referred as the PhosphoELM test set. The second dataset was obtained from the recently published activity-based high-resolution human phosphorylation map (Hu *et al.*, 2014; Newman *et al.*, 2013), which became available after completion of our work. This will be referred as the Newman test set. All the substrates in these two datasets were selected with the criterion that they contained one or more of the 856 domains analyzed in our training dataset and one or more of the 61 protein kinases present in our training dataset were known to phosphorylate these substrates. However, the exact kinase–substrate pairs were not present in the training set. Thus, they had not been used in the derivation of prediction rules and hence constituted bona fide independent datasets for testing the predictive power of PhosNetConstruct. PhosphoELM test set consisted of 187 substrate polypeptides, whereas the Newman test set consisted of 229 substrate polypeptides. Each substrate protein from these test datasets was analyzed to identify its constituent Pfam domains, and putative target kinases were predicted based on the ER for each kinase–domain pair above certain cutoff. The optimum value of the cutoff was selected based on receiver operating characteristic (ROC) analysis (Fawcett, 2006). Prediction performance was tested by using six statistical parameters, namely, sensitivity, specificity, false positive rate (FPR),

positive prediction value (PPV), accuracy and Mathew's correlation coefficient (MCC) as described in Supplementary Methods.

2.4 Comparison of the performance of PhosNetConstruct with GPS, iGPS and NetPhosK

The performance of PhosNetConstruct was compared with GPS (Xue *et al.*, 2008), iGPS (Song *et al.*, 2012) and NetPhosK (Blom *et al.*, 2004), which use motif/profile or machine learning-based approaches to predict the putative sites in a substrate protein that are likely to be phosphorylated by various kinases. Based on the sites predicted on these substrate proteins for various kinases, kinase assignments were carried out for different Pfam domains in the substrate proteins. All those kinases for which sites were predicted with a score above certain cutoff values by GPS, iGPS and NetPhosK on a substrate protein were assigned as the predicted target kinases for that protein. For this purpose of comparison, a set of nine kinase classes was selected, which can be predicted by all the four methods, i.e. PhosNetConstruct, GPS, iGPS and NetPhosK (Supplementary Data S1). Of the 187 substrate proteins in the PhosphoELM test set, only 75 substrate proteins were known to be phosphorylated by these nine kinases, namely, PKA, PKC, PKG, CDK, CK1, CK2, EGFR, SRC and GSK. Therefore, these 75 substrate proteins (Supplementary Data S4) were selected for benchmarking, and performance of all four methods was compared in terms of the six statistical parameters mentioned earlier. Of the 229 substrate polypeptides in the Newman test set, 153 substrate proteins were known to be phosphorylated by 13 different kinase classes, namely, ABL, AKT, AUR, CaMK, CDK, DYRK, GRK, GSK, JAK, MAPKAPK, PLK, SGK and SYK. Because a large number of these kinases cannot be predicted by NetPhosK, benchmarking on these 153 substrate proteins from the Newman test set was carried out using PhosNetConstruct, GPS and iGPS (Supplementary Data S5).

3 RESULTS

The mammalian phosphorylation network analyzed in this study consists of 61 kinases and 856 Pfam domains as nodes, and 1861 edges that correspond to experimentally identified kinase–Pfam domain pairs (Fig. 1). Analysis of the Pfam domains phosphorylated by different kinase groups indicates that certain domains are phosphorylated exclusively by certain kinase groups (Supplementary Fig. S1). For example, 197 distinct Pfam domains are phosphorylated exclusively by the AGC group of kinases and not by any other group. A similar trend exists for all the groups, although the number of Pfam domains exclusively phosphorylated by each kinase group varies.

3.1 Analysis of known phosphorylation network

Analysis of individual kinase–domain pairs in the mammalian phosphorylation network revealed that certain kinases phosphorylate a large number of Pfam domains, whereas other kinases phosphorylate few domains (Fig. 2). For example, PKACα phosphorylates 175 distinct Pfam domains, whereas LKB1 phosphorylates only five Pfam domains. Similarly, of the 856 Pfam domains, only a small percentage of domains are phosphorylated by a large number of protein kinases, whereas the majority of the Pfam domains are phosphorylated by only six or fewer kinases (Supplementary Figs. S2 and S3). Thus, our analysis has led to the identification of certain highly connected hub kinases (Fig. 3a) and Pfam domains (Fig. 3b) in the protein phosphorylation networks. Protein kinases like PKA, PKC, CDK, CK2 and Src are crucial in multiple cellular processes and signaling pathways that are conserved within and across organisms. Therefore, they phosphorylate many different functional domains, which results in highly connected hubs in phosphorylation networks. Many different Pfam domains can potentially harbor recognition motifs for the hub kinases that phosphorylate them. However, it was intriguing how the hub Pfam domains harbor recognition motifs for so many different kinases. Therefore, we analyzed in detail the phosphorylation sites for different kinases on each of these hub domains. Figure 4 shows five phosphosites in the P53_TAD domain that are phosphorylated by 18 different kinases. As can be seen, even though

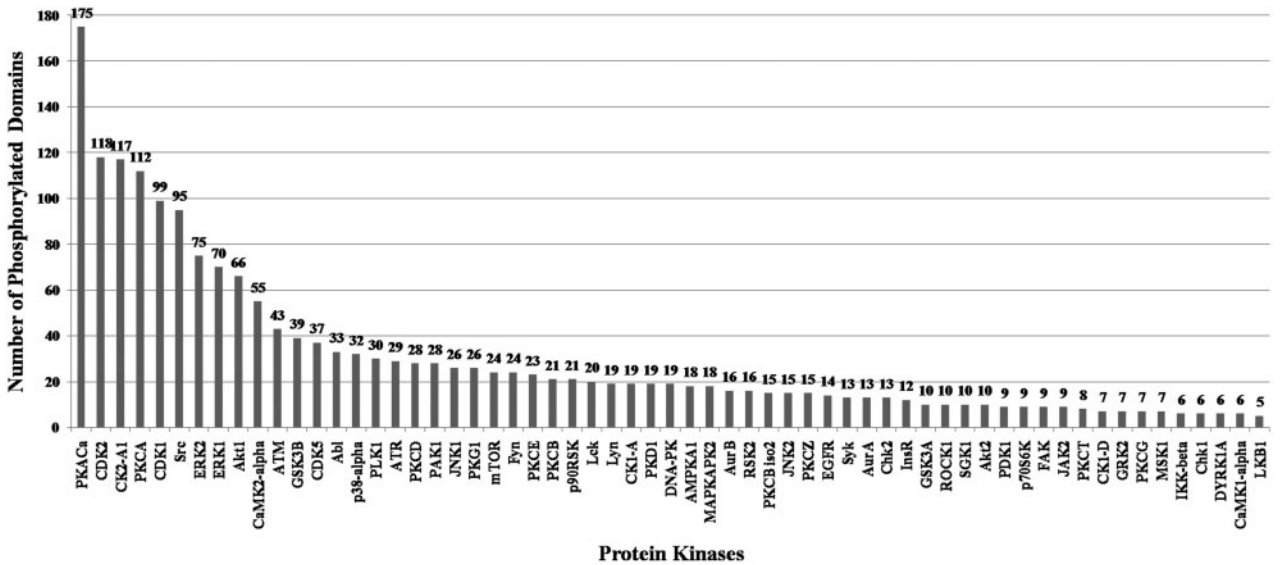

Fig. 2. Differential phosphorylation of Pfam domains by different protein kinases. Bar plot depicts the number of Pfam domains phosphorylated by each of the 61 different kinases in the dataset

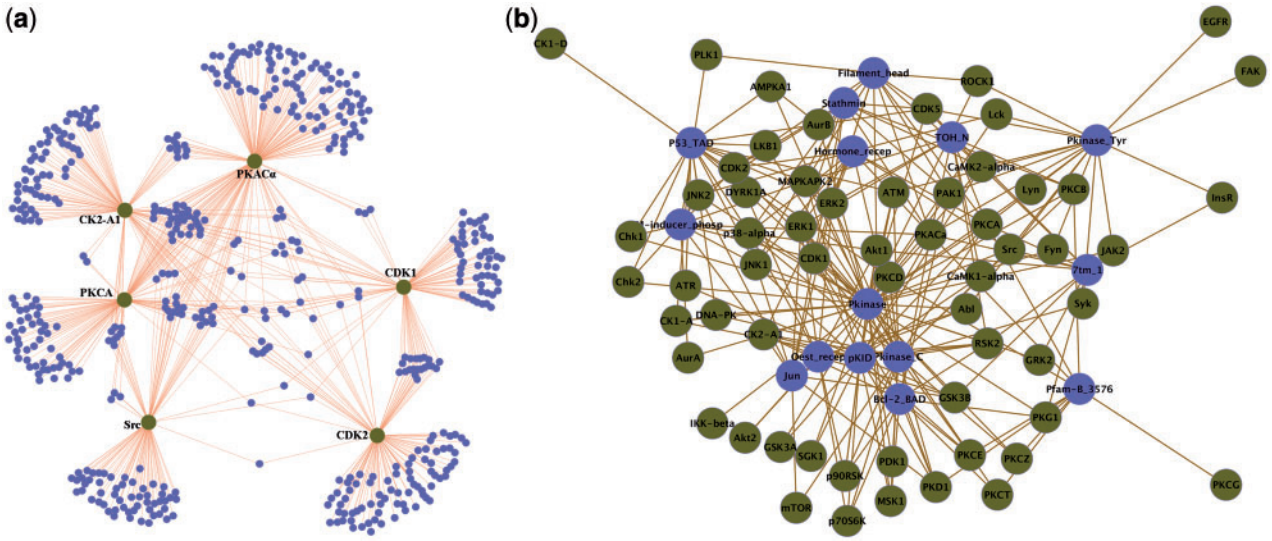

Fig. 3. Spring-embedded layouts of kinase–domain phosphorylation networks depicted using Cytoscape 2.7. Kinases are shown as green circles, while the Pfam domains are shown as blue circles. Each edge in the network represents phosphorylation of a Pfam domain by a kinase. (**a**) Only the top six hub kinases that phosphorylate the maximum number of domains are depicted and labeled. (**b**) The top 15 hub domains that are phosphorylated by multiple kinases are denoted. As can be seen, a large number of Pfam domains are exclusively phosphorylated by each of these hub kinases, while relatively fewer Pfam domains are phosphorylated by more than one of these hub kinases

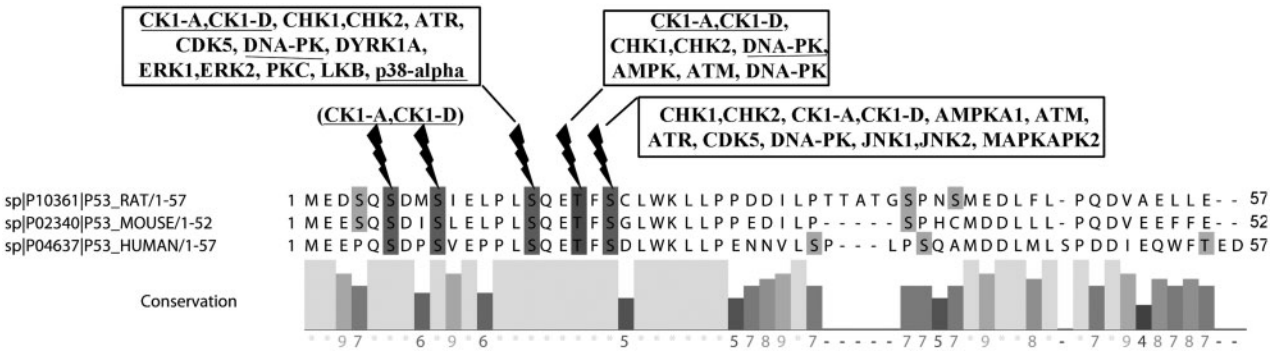

Fig. 4. Multiple sequence alignment of the P53_TAD domain of P53 protein from rat, mouse and human. The Ser and Thr residues are shaded. The sites that are phosphorylated are indicated with a lightening mark, and the kinases that phosphorylate these sites are shown within the boxes. Kinases for which a substrate recognition motif (as compiled by NetPhorest) exists in the sequence stretch surrounding the site are underlined

these phosphosites on P53_TAD contain recognition motifs [as defined by NetPhorest (Miller *et al*., 2008)] for only four kinases, i.e. DNA-PK, p38-alpha, CK1-A and CK1-D, they are phosphorylated by 14 other kinases despite absence of their cognate recognition motifs. It is interesting to note that proline-directed kinases like CDKs also phosphorylate non-proline sites on P53_TAD (Lee *et al*., 2007). These results suggest that, apart from the sequence motifs, other context-dependent features as described earlier also play a significant role in the kinase–substrate recognition process.

To estimate the propensity of different kinases to phosphorylate different Pfam domains, we analyzed the number of occurrences of each kinase–phosphodomain pair in this network (Supplementary Fig. S4). Supplementary Figure S5 shows the number of occurrences of the kinase–phosphodomain pairs and the total number of corresponding phosphosites for the pKID domain to highlight the differential phosphorylation of pKID by different kinases.

3.2 Classification of kinases based on the domains they phosphorylate

Supplementary Figure S6 shows the heat map depicting the clustering of the 61 kinases based on their domain phosphorylation profiles. Interestingly, the domain phosphorylation profile-based clustering closely resembles Manning's sequence-based clustering (Supplementary Fig. S7) of kinase catalytic domains in general (Manning *et al*., 2002). It may be noted that no sequence information of kinase catalytic domains has been used in domain phosphorylation-based clustering. However, it uses substrate sequence information in the form of Pfam functional domain annotations. This result suggests that evolutionarily related kinases

have similar domain phosphorylation profiles. However, there are subtle differences that seem to have interesting functional implications. Unlike sequence-based clustering, in domain phosphorylation profile-based clustering, the AGC group of kinases forms three distinct clusters. PKC isoforms—β, δ, γ and ε—phosphorylate a different set of Pfam domains than other PKC isoforms—ζ, τ and α. Similarly, the CMGC group has also split into two different clusters. Such subtle alterations in the substrate specificity of homologous kinases indicate the involvement of certain kinase–substrate pairs in a specific biological context. This contextual specificity of phosphorylation networks is not apparent from the sequence-based classification. Although the current study takes into consideration only 61 kinases, the conservation patterns in specificity of domain phosphorylation by different groups of protein kinases are likely to prevail even when other kinases are included. Therefore, domain phosphorylation patterns might help to construct protein phosphorylation networks in a more refined way.

3.3 Estimating relative preference of different kinases to phosphorylate a given Pfam domain

Analysis of kinase–domain phosphorylation network and classification of kinases based on their domain phosphorylation profiles revealed that different kinases have distinct preferences to phosphorylate certain Pfam domains with higher probabilities than others. Therefore, we attempted to estimate the propensities of different kinases to phosphorylate different Pfam domains based on the number of occurrences of kinase–phosphodomain pairs in our dataset. As mentioned in Section 2.2, we computed ERs that give a quantitative measure of preferential phosphorylation of a given domain by a specific kinase compared with all other kinases. The value of ER would range from $ER = 0$ for a non-cognate kinase–domain pair to $ER = \infty$ for those pairs where a given kinase phosphorylates its cognate domain exclusively and no other kinase phosphorylates that domain. In cases where $ER = \infty$, ER was assigned the value of 999 for convenience. Using this ER matrix, it will be possible to predict kinases that are likely to phosphorylate a given Pfam domain. Supplementary Figure S8 shows a heat map of ER values for the 61 kinases and the top 24 hub domains in terms of $\log_{10}$ scale. For example, of nine cognate kinases, PKCα phosphorylates the Integrin_b_cyt domain with the highest probability and PDK1 with the least. Thus, an ER matrix consisting of 61 kinases and 856 phosphodomains was computed for all 52 216 possible kinase–domain pairs. ERs were also obtained for all co-occurring domains, and similar to phosphodomain enrichment, the ER matrix was computed for 61 kinases and 2166 domains irrespective of their phosphorylation status. These results indicate that potential target kinases for a phosphoprotein can also be predicted based on the enrichment of phospho as well as co-occurring auxiliary domains. It may be noted that, in contrast to the phosphodomain enrichment that can predict domain-specific phosphorylation, enrichment of co-occurring auxiliary domains represents the phosphorylation preference of a complete substrate protein within the domain as well as interdomain linker regions. Because interdomain linker regions are also extensively phosphorylated (Iakoucheva *et al*., 2004), we also analyzed the phosphorylation patterns in these linker regions. As can be seen from Supplementary Figure S9, a large number of protein kinases phosphorylate only few linker regions, whereas a small set of kinases phosphorylates a large number of interdomain linkers. Interestingly, the kinases that phosphorylate the flanking domains also phosphorylate the intervening linker stretches. Therefore, for prediction of phosphorylations on domain as well as linker regions of a substrate protein, it might be appropriate to consider any domain enrichment, i.e. enrichment of phosphodomains as well as co-occurring auxiliary domains.

3.4 Development of PhosNetConstruct server for prediction of protein phosphorylation networks

We have developed an automated program PhosNetConstruct for predicting kinases that are likely to phosphorylate a given substrate protein. The results of our analysis of a kinase–phosphodomain network and the enrichment matrix derived from it constitute the knowledge base for kinase predictions by PhosNetConstruct. Given a FASTA sequence of a putative substrate protein as input, PhosNetConstruct identifies its Pfam domain organization and then predicts target kinases for each of its constituent Pfam domains based on the ERs for respective kinase–phosphodomain pairs. For each of the predicted kinases, PhosNetConstruct also allows visualization of the experimentally identified substrates, their constituent Pfam domains and sites of phosphorylation. PhosNetConstruct also provides known PPI networks of kinases based on the information from BioGRID (Stark *et al*., 2006). The visualization of the phosphorylation networks is facilitated using WebCytoscape (Lopes *et al*., 2010). It also provides information about 3D structures of interacting domains by linking to 3DID (Stein *et al*., 2011).

3.5 Benchmarking of PhosNetConstruct

3.5.1 Prediction of KSR using PhosNetConstruct, GPS, iGPS and NetPhosK We wanted to compare the performance of PhosNetConstruct, GPS, iGPS and NetPhosK by predicting the target kinases for the 75 substrate proteins in PhosphoELM test set. GPS (Xue *et al*., 2008) and NetPhosK (Blom *et al*., 2004) predict ssKSRs. iGPS (Song *et al*., 2012) also predicts ssKSRs similar to GPS, but filters the predictions by combining kinase–substrate interaction information from experimental PPI databases (Szklarczyk *et al*., 2011). This test set consisted of 75 substrate proteins phosphorylated by nine different kinases. Thus, 675 kinase–substrate pairs are possible, of which only 86 are experimentally known cognate pairs, whereas the remaining 589 constitute non-cognate combinations. We wanted to investigate whether for each of these 75 substrate proteins, these four prediction tools can correctly identify the cognate kinases as true-positive (TP) results as well as non-cognate kinases as true-negative (TN) results. Supplementary Figure S10 shows the results of prediction for a typical phosphoprotein containing pKID and bZIP_1 Pfam domains, and kinases predicted by each of the tools are listed as TP, FP, FN and TN values. A similar prediction approach was followed for all 75 substrates (Supplementary Data S4 and S6).

The thresholds or significance values for predictions by GPS, iGPS and NetPhosK were chosen as per the optimum values recommended by the respective developers. However, for

PhosNetConstruct, the threshold for ER was chosen based on ROC analysis. As can be seen from Figure 5, the ROC curves have a convex shape above the diagonal for the ER values in the range of 0–999. Increasing the ER cutoff reduces the false-positive predictions. However, in our model, all the kinase–domain pairs that are absent in the training set are assigned an ER value of 0. Hence, an ER value of 0 implies that the corresponding kinase–substrate pair will always be predicted by PhosNetConstruct as negative. Hence, the ROC curves do not cover the full range of FPR up to 1.0 but reach optimum values of FPR and TPR at the ER cutoff of 0.2. Therefore, for comparison with other tools, predictions were carried out with an ER cutoff of 0.2 (Table 1). For the purpose of comparison, predictions by PhosNetConstruct were also carried out without any ER cutoff, i.e. based on the presence or absence of a particular kinase–domain pair. It is interesting to note that phosphodomain enrichment without the ER filter achieves a TPR of 62.8% at an FPR of 23%. However, when the ER filter is used along with phosphodomain enrichment, the FPR reduces to 16.6%, but this also lowers the TPR to ~59%. In contrast to the results of our domain-based KSR prediction approach, the motif/profile-based ssKSR predictors GPS and NetPhosK give a much larger number of false-positive predictions (Supplementary Table S1), which results in FPRs of 66.38 and 71.82%, respectively. The performance of iGPS is comparable with PhosNetConstruct with relaxed enrichment criterion because of the inclusion of results from PPI data. However, the specificity, TPP and accuracy for PhosNetConstruct are better than iGPS, indicating its higher efficiency at distinguishing cognate kinase–substrate pairs from non-cognate combinations. Interestingly, any domain enrichment with the ER filter has best performance among all seven prediction approaches compared here, as it achieves a TPR of 81.4% at an FPR of 30.4%. It also has the highest accuracy and MCC values. These results indicate that performance of PhosNetConstruct is superior to ssKSR predictors like GPS, NetPhosK and iGPS. The higher performance of any domain-enrichment approach over phosphodomain enrichment suggests that certain domains preferentially occur in substrates of specific kinases, even if they are not phosphorylated. It is possible that they mediate interactions between the substrate protein and the proteins containing kinase domain by domain–domain interactions. In fact, Liu and Tozeren (2010) have also demonstrated that kinase–substrate pairs can be predicted based on the enrichment of co-occurring domains.

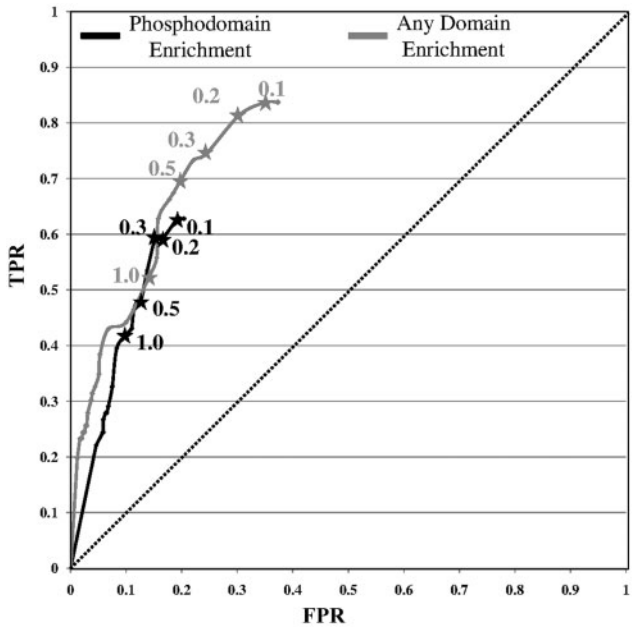

Fig. 5. ROC curves indicating the performance of PhosNetConstruct on a comparative benchmarking test set of 75 substrates phosphorylated by nine kinases with phosphodomain enrichment (black) and any domain enrichment (gray)

We also analyzed the performance of PhosNetConstruct, GPS and iGPS on a Newman test set consisting of 153 substrates and 13 kinases (Supplementary Data S5 and S7). The results from different predictors are shown in Table 2, and Supplementary Figure S11 shows ROC curves for predictions by PhosNetConstruct. Phosphodomain enrichment without ER cutoff achieves a sensitivity of 37.1% and an FPR of 19.8%, which are less than iGPS (48 and 22%, respectively). However, any domain enrichment without ER cutoff achieves a sensitivity of 71.5% with an FPR of 37%. On the other hand, because of increased false-positive results (Supplementary Table S2), GPS

Table 1. Statistical parameters showing comparative benchmarking results of PhosNetConstruct on a completely independent dataset of 75 substrates and nine kinases from PhosphoELM such that these kinase–substrate pairs could be predicted by all four approaches—PhosNetConstruct, GPS, iGPS and NetPhosK

Parameter	GPS	iGPS	NetPhosK	PhosNetConstruct (Phosphodomain enrichment)		PhosNetConstruct (Any domain enrichment)	
Threshold	High	Low	0.5	No filter	ER ≥ 0.2	No filter	ER ≥ 0.2
Sensitivity/TPR (%)	70.93	73.26	93.02	62.79	59.30	83.72	81.39
Specificity (%)	33.62	74.53	28.18	76.91	83.36	54.50	69.61
FPR (%)	66.38	25.47	71.82	23.09	16.64	45.50	30.39
TPP (%)	13.5	29.58	15.90	28.42	34.23	21.18	28.11
Accuracy (%)	38.37	74.37	36.44	75.11	80.30	58.22	71.11
MCC	0.032	0.343	0.162	0.294	0.343	0.255	0.352

Table 2. Statistical parameters showing comparative benchmarking results of PhosNetConstruct, GPS and iGPS on completely independent kinase–substrate pairs consisting of 13 kinases and 153 substrates from Newman test dataset

	GPS	iGPS	PhosNetConstruct (Phosphodomain enrichment)		PhosNetConstruct (Any domain enrichment)	
Threshold	High	Low	No filter	ER ≥0.2	No filter	ER ≥0.2
Sensitivity/TPR (%)	86.42	47.96	37.10	27.60	71.49	61.99
Specificity (%)	31.73	77.60	80.20	89.48	63.00	76.13
FPR (%)	68.27	22.40	19.80	10.52	36.99	23.87
TPP (%)	13.66	21.12	18.98	24.69	19.46	24.51
Accuracy (%)	37.81	74.31	75.41	82.60	63.95	74.56
MCC	0.125	0.185	0.132	0.163	0.221	0.267

shows a high FPR of 68.3%, although it achieves an increased sensitivity of 86.4%. Thus, even on this larger test set, PhosNetConstruct has higher specificity compared with other ssKSR predictors. However, both PhosNetConstruct and iGPS have lower sensitivities compared with their performance on the PhosphoELM test set (Table 1), even though specificities and TPP values are similar for both test sets. Detailed analysis revealed that the reduced sensitivities arise from an increase in false-negative predictions (Supplementary Table S2) in the Newman test set. This is primarily because the Newman test set consists of many novel kinase–phosphodomain pairs that were absent in the training sets of both PhosNetConstruct as well as iGPS. Therefore, we decided to include recently released data from a high-resolution map of the human phosphorylation network (Newman *et al.*, 2013) for training PhosNetConstruct.

3.6 Performance of PhosNetConstruct using enhanced training data

The activity-based human phosphorylation network has provided huge volumes of data on novel kinase–substrate pairs. However, dsKSRs can only be derived for those pairs for which phosphosite information is available. This high-resolution human phosphorylation map consisted of 230 kinases and 648 substrates (Newman *et al.*, 2013). However, there were only 21 kinases that had >20 substrates. To have significant numbers of kinase–phosphodomain pairs, only these 21 kinases and 391 substrates phosphorylated by them were combined with our PhosphoSitePlus training data, and ERs for preferential domain phosphorylation were recalculated with an increased number of kinases and substrates (Supplementary Data S8 and S9). Many of the substrates in the high-resolution human phosphorylation map, though overlapped with earlier PhosphoSitePlus training data, were found to be phosphorylated by different kinases. Thus, our new training data consisted of 956 phosphodomains from 2851 distinct substrate polypeptides that are known to be phosphorylated by 80 kinases. This resulted in 2361 kinase–phosphodomain pairs compared with 1861 kinase–phosphodomain pairs in our older training set.

We selected 168 substrates from the PhosphoELM test set of 187 substrates by eliminating those that overlap with our new

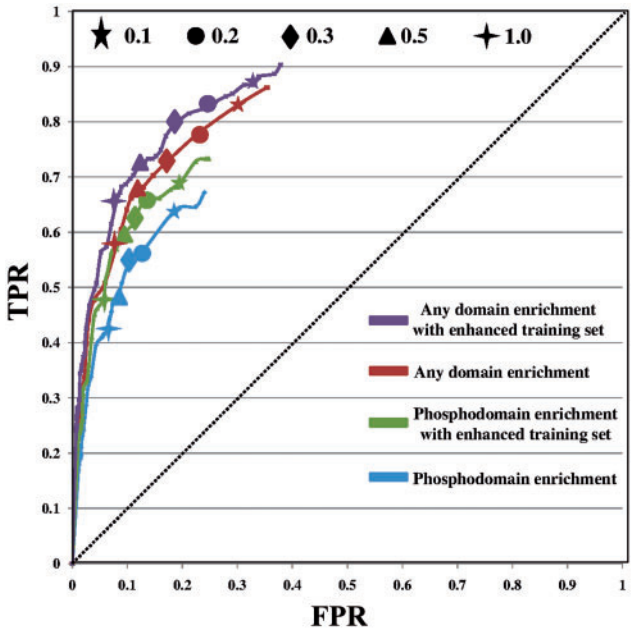

Fig. 6. ROC curves indicating the performance of PhosNetConstruct on the PhosphoELM test set of 168 substrates phosphorylated by 30 kinases. Blue and brown curves indicate prediction performance of phosphodomain and any domain enrichment, respectively, using PhosphoSitePlus training data. Green and violet curves indicate prediction performance of phosphodomain and any domain enrichment, respectively, using enhanced training data as described in the text

training data. Using this completely independent test set, we compared the performance of PhosNetConstruct trained on enhanced data with that of an older version that used only PhosphoSitePlus data for training (Supplementary Data S10 and S11). Although predictions by PhosNetConstruct were carried out at an individual kinase level, statistical parameters were computed by merging closely related individual family members into a single family, thereby giving rise to 30 kinase families as highlighted in Supplementary Data S1. The ROC curves (Fig. 6) indicate distinct improvement in performance on inclusion of Newman's data in our training set. Supplementary Tables S3

and S4 show performance of PhosNetConstruct in terms of six statistical parameters mentioned earlier with an ER cutoff of 0.2 and without any ER filter. As can be seen from Figure 6 and Supplementary Table S3, when PhosphoSitePlus data alone were used for training and predictions were carried out using phosphodomain enrichment with an ER cutoff of 0.2, experimentally known kinase–substrate pairs could be identified with a TPR of 56% at an FPR of 12%. On inclusion of high-resolution KSRs in training data, the corresponding value of TPR increased to 66% at an FPR of 13%. Instead of phosphodomain enrichment, if the criterion of any domain enrichment with ER filter is used, PhosNetConstruct trained using enhanced data achieves a TPR of 82% at an FPR of 25%. Thus, it can be seen that inclusion of high-confidence KSRs from Newman *et al*. (2013) is advantageous in terms of prediction of dsKSRs. Currently, PhosNetConstruct is the only program that covers so many different kinases and can still predict target kinases for substrate proteins with such high sensitivity and specificity values. Thus, our benchmarking on completely independent test sets indicates that, using the dsKSRs derived in this work, it will be possible to predict phosphorylation networks with reasonable accuracy. After identifying kinase–substrate pairs using a dsKSR-based approach implemented in PhosNetConstruct, in the next stage individual phosphosites can be identified using ssKSR-based methods like NetPhosK and GPS.

4 DISCUSSION

In this work, we have used experimentally identified kinase–substrate pairs and their Pfam domain composition to analyze dsKSRs and constructed a domain-based kinase–substrate interaction network. Experimentally identified phosphosites on substrate proteins have been mapped onto Pfam domains or interdomain linker regions. Interestingly, degree distribution of kinases, as well as domains of this network, shows a power law-type behavior, and our analysis reveals presence of a small number of highly connected hub kinases and hub domains. This suggests that only a few kinases can phosphorylate a large number of Pfam domains, while most kinases phosphorylate only a few Pfam domains and vice versa. The differential specificities of various kinases for different Pfam domains have been quantified in the form of ER of kinase–domain pairs. Our analysis of dsKSRs indicates that different Pfam domains are preferentially phosphorylated by a limited number of specific kinases only. Similarly, we have also computed ER values for kinase–auxiliary domain pairs, which represent preference of a kinase to phosphorylate a substrate protein because of enriched occurrences of the given auxiliary domain, even if the auxiliary domain is not phosphorylated by the kinase.

Short sequence motif-based approaches for deciphering ssKSRs have their origin in the use of peptide library data, which obviously lacks contextual information. Hence, prediction of phosphorylation networks based on ssKSRs results in a large number of false-positive KSRs. On the other hand, dsKSRs derived in this work implicitly incorporate various context-dependent features. Second, Pfam domains are essentially longer signature motifs, and thus they become more specific. This helps in eliminating false-positive kinase–substrate pairs that arise in ssKSR-based predictions. Therefore, we have incorporated these dsKSRs into a web-based automated computational tool, PhosNetConstruct, for genome-wide identification of protein phosphorylation networks.

Using these ER values, PhosNetConstruct distinguishes cognate kinase–domain pairs from a large number of non-cognate combinations. Benchmarking on completely independent test datasets indicates that performance of PhosNetConstruct is superior to ssKSR-based predictors like GPS and NetPhosK. However, iGPS that incorporates PPI information is the only approach that attains a comparable trade-off between sensitivity and specificity. Thus, our benchmarking results also suggest that dsKSR derived by us has implicitly incorporated context-dependent information, which methods like iGPS and NetworKIN incorporate by including direct PPIs. Therefore, our work demonstrates that even in cases where information about PPI is not available, PhosNetConstruct can generate a potential list of cognate kinase–substrate pairs by filtering out a large number of non-cognate combinations and subsequently exact phosphosites can be identified using ssKSR predictors like GPS and NetPhosK.

Funding: Department of Biotechnology, Government of India, grant to National Institute of Immunology, New Delhi. D.M. also acknowledges financial support from DBT, India under BTIS project and National Bioscience Career Development award. N.P.D. is grateful to DBT, India, for the BINC fellowship.

Conflict of Interest: none declared.

REFERENCES

Blom,N. *et al.* (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, **4**, 1633–1649.

Bodenmiller,B. and Aebersold,R. (2011) Phosphoproteome resource for systems biology research. *Methods Mol. Biol.*, **694**, 307–322.

Cohen,P. (2000) The regulation of protein function by multisite phosphorylation-a 25 year update. *Trends Biochem. Sci.*, **25**, 596–601.

Dinkel,H. *et al.* (2010) Phospho.ELM: a database of phosphorylation sites-update 2011. *Nucleic Acids Res.*, **39**, D261–D267.

Fawcett,T. (2006) An introduction to ROC analysis. *Pattern Recognit. Lett.*, **27**, 861–874.

Finn,R.D. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.

Hornbeck,P.V. *et al.* (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.*, **40**, D261–D270.

Hu,J. *et al.* (2014) PhosphoNetworks: a database for human phosphorylation networks. *Bioinformatics*, **30**, 141–142.

Hunter,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.

Iakoucheva,L.M. *et al.* (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.*, **32**, 1037–1049.

Kumar,N. and Mohanty,D. (2010) Identification of substrates for Ser/Thr kinases using residue-based statistical pair potentials. *Bioinformatics*, **26**, 189–197.

Lee,J.H. *et al.* (2007) Stabilization and activation of p53 induced by Cdk5 contributes to neuronal cell death. *J. Cell Sci.*, **120**, 2259–2271.

Linding,R. *et al.* (2007) Systematic discovery of *in vivo* phosphorylation networks. *Cell*, **129**, 1415–1426.

Liu,Y. and Tozeren,A. (2010) Modular composition predicts kinase/substrate interactions. *BMC Bioinformatics*, **11**, 349.

Lopes,C.T. *et al.* (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.

Manning,G. *et al.* (2002) The protein kinase complement of the human genome. *Science*, **298**, 1912–1934.
Miller,M.L. *et al.* (2008) Linear motif atlas for phosphorylation-dependent signaling. *Sci. Signal.*, **1**, ra2.
Mok,J. *et al.* (2010) Deciphering protein kinase specificity through large-scale analysis of yeast phosphorylation site motifs. *Sci. Signal.*, **3**, ra12.
Newman,R.H. *et al.* (2013) Construction of human activity-based phosphorylation networks. *Mol. Syst. Biol.*, **9**, 655.
Obenauer,J.C. *et al.* (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.*, **31**, 3635–3641.
Prisic,S. *et al.* (2010) Extensive phosphorylation with overlapping specificity by *Mycobacterium tuberculosis* serine/threonine protein kinases. *Proc. Natl Acad. Sci. USA*, **107**, 7521–7526.
Saunders,N.F. and Kobe,B. (2008) The Predikin webserver: improved prediction of protein kinase peptide specificity using structural information. *Nucleic Acids Res.*, **36**, W286–W290.
Song,C. *et al.* (2012) Systematic analysis of protein phosphorylation networks from phosphoproteomic data. *Mol. Cell Proteomics*, **11**, 1070–1083.
Sprinzak,E. and Margalit,H. (2001) Correlated sequence-signatures as markers of protein-protein interaction. *J. Mol. Biol.*, **311**, 681–692.
Stark,C. *et al.* (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.
Stein,A. *et al.* (2011) 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.*, **39**, D718–D723.
Szklarczyk,D. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.
Xue,Y. *et al.* (2006) PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinformatics*, **7**, 163.
Xue,Y. *et al.* (2008) GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Mol. Cell. Proteomics*, **7**, 1598–1608.