Journal of
Proteomics & Bioinformatics

Amanchy et al. J Proteomics Bioinform 2011, 4:2
DOI: 10.4172/jpb.1000163

Research Article Open Access

Identification of Novel Phosphorylation Motifs Through an Integrative Computational and Experimental Analysis of the Human Phosphoproteome

Ramars Amanchy[1#], Kumaran Kandasamy[1,2#], Suresh Mathivanan[2], Balamurugan Periaswamy[2], Raghunath Reddy[2], Wan-Hee Yoon[1], Jos Joore[3], Michael A Beer[4], Leslie Cope[5] and Akhilesh Pandey[1*]

[1]McKusick-Nathans Institute of Genetic Medicine and the Departments of Biological Chemistry and Oncology, Johns Hopkins University, Baltimore, Maryland 21205, USA
[2]Institute of Bioinformatics, International Tech Park, Bangalore 560066, India
[3]Pepscan Systems, Edelhertweg 15, 8219 PH Lelystad, The Netherlands
[4]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland 21205, USA
[5]Sidney Kimmel Comprehensive Cancer Center and the Department of Biostatistics, Bloomberg School of Public Health, and Johns Hopkins University, Baltimore, Maryland, USA
[#]These authors contributed equally

Abstract

Protein phosphorylation occurs in certain sequence/structural contexts that are still incompletely understood. The amino acids surrounding the phosphorylated residues are important in determining the binding of the kinase to the protein sequence. Upon phosphorylation these sequences also determine the binding of certain domains that specifically bind to phosphorylated sequences. Thus far, such 'motifs' have been identified through alignment of a limited number of well identified kinase substrates. Results: Experimentally determined phosphorylation sites from Human Protein Reference Database were used to identify 1,167 novel serine/threonine or tyrosine phosphorylation motifs using a computational approach. We were able to statistically validate a number of these novel motifs based on their enrichment in known phosphopeptides datasets over phosphoserine/threonine/tyrosine peptides in the human proteome. There were 299 novel serine/threonine or tyrosine phosphorylation motifs that were found to be statistically significant. Several of the novel motifs that we identified computationally have subsequently appeared in large datasets of experimentally determined phosphorylation sites since we initiated our analysis. Using a peptide microarray platform, we have experimentally evaluated the ability of casein kinase I to phosphorylate a subset of the novel motifs discovered in this study. Our results demonstrate that it is feasible to identify novel phosphorylation motifs through large phosphorylation datasets. Our study also establishes peptide microarrays as a novel platform for high throughput kinase assays and for the validation of consensus motifs. Finally, this extended catalog of phosphorylation motifs should assist in a systematic study of phosphorylation networks in signal transduction pathways.

Keywords: Phosphorylation; Motifs; Peptide array

Introduction

Protein kinases are encoded by over 500 genes in humans Manning et al. [1] and constitute the largest single enzyme family in the human genome. It has been estimated that about one-third of all proteins in mammalian cells can undergo phosphorylation. The large number of protein kinases and an even variety spectrum of their substrates involved in protein phosphorylation, reflect the true complexity of signal transduction cascades. Phosphorylation of tyrosine residues generates binding sites for modular domains such as SH2 (Src homology 2 domains) and PTB (phosphotyrosine-binding domain) in scaffolding proteins to form multi-protein complexes [2]. Similarly, serine and threonine phosphorylation is involved in formation of the multi-protein signaling complexes through interaction with phospho-serine/threonine binding domains such as 14-3-3 and WD40 [3-5]. The enzymes responsible for inducing post-translational modifications on proteins often recognize sequence patterns around the amino acid on which the modification occurs. The residues surrounding phosphorylated serine/threonine/tyrosine residues are responsible for the specific recognition by these modular phosphoprotein-binding domains. Although a number of kinases have been identified in human proteome, the exact substrate specificity is known only for a limited set of kinases. A global analysis of phosphorylation sites should aid in understanding and determining kinase specificities.

Given the large number of phosphorylation sites that are known today, it is difficult to identify phosphorylation motifs by a simple manual inspection of the sequences surrounding the phosphorylated residues. Although there are a number of programs available for identifying sequence patterns from sequence data, most of them are not specifically designed to identify phosphorylation motifs *per se*. motif-x is a web-based program that uses an iterative statistical approach to identify over-represented patterns from any sequence dataset and has recently been used for predicting phosphorylation motifs from two phosphorylation studies [6].

Some of the earlier analyses on finding tyrosine phosphorylation sequence motifs were based on the frequency of various amino acids surrounding the tyrosine residue. For example, the frequent occurrence of hydrophobic residues at positions +1 to +3 (where 0 is the position of phosphorylated tyrosine residue and + represents more C-terminal residues) has been described previously Blom et al. [7]. The occurrence of acidic residues in positions -5 to -1 upstream of tyrosine residues has been reported [8,9] and of proline in +5, +7 and +9 downstream

***Corresponding author:** Akhilesh Pandey, M.D., Ph.D., 733 N.Broadway Street, Broadway Research Building, Room 527, Baltimore, Maryland 21205, USA, Tel: 410-502-6662; Fax: (410) 502 7544; E-mail: pandey@jhmi.edu

Received December 28, 2010; **Accepted** January 28, 2011; **Published** February 10, 2011

Citation: Amanchy R, Kandasamy K, Mathivanan S, Periaswamy B, Reddy R, et al. (2011) Identification of Novel Phosphorylation Motifs Through an Integrative Computational and Experimental Analysis of the Human Phosphoproteome. J Proteomics Bioinform 4: 022-035. doi:10.4172/jpb.1000163

positions. In the case of serine/threonine phosphorylation, basic residues have been reported at positions -3 and -2 with respect to the phosphorylated serine/threonine residue Blom et al. [7]. In fact, the specificity of many kinases is dominated by the presence of acidic, basic or hydrophobic residues adjacent to the phosphorylated residue Songyang et al. [10], although the variation makes it almost impossible for manual inspection and prediction of exact phosphorylation sites. There have been some recent reports demonstrating the use of publicly available phosphorylation data for motif analyses [11-13].

Here, we have carried out a computational analysis using experimentally determined phosphorylation sites as catalogued in Human Protein Reference Database Keshava Prasad et al. [14] to detect phosphorylation motifs. Using this approach, we identified 207 phosphotyrosine motifs and 960 phosphoserine/threonine motifs. A statistical analysis showed 105 novel phosphotyrosine and 194 novel phosphoserine/threonine motifs to be statistically enriched in the phosphorylation dataset as compared to the human proteome.

We also performed experimental validation of the identified novel motifs in two different ways. First, we verified the occurrence of these novel motifs among published mass spectrometry derived phosphorylation datasets [15-19] including one from our laboratory Molina et al. [20]. One hundred and sixty five of the predicted motifs were represented >5 times among the phosphorylation sites from these studies. Second, we developed peptide microarrays as a novel platform for validation of the predicted motifs. For this, we spotted 724 peptide sequences containing novel motifs from 449 proteins which were not known to be phosphorylated and carried out kinase assays using casein kinase I (CKI). The peptides that were phosphorylated by CKI included those that were similar to known specificity of casein kinases as well as novel sequence motifs that were not suspected to be substrate motifs. Thus, peptide microarrays provide an excellent platform for the discovery of novel phosphorylation motifs and can be used to characterize kinases and their substrates in a high-throughput fashion.

Materials and Methods

Dataset of phosphorylated peptide sequences

Fifteen amino acids long peptide sequences (with the phosphorylated tyrosine/serine/threonine residue in the center) were generated from the known phosphorylation sites catalogued in the Human Protein Reference Database (http://www.hprd.org). In total, 4,810 phospho-peptides (pSer peptides – 3,184, pThr peptides - 760 and pTyr peptides - 866) were analyzed.

Distribution of amino acids surrounding phosphorylated residue

Heat maps were generated using 'R' (http://www.r-project.org) to visualize the amino acid distribution surrounding the phosphorylated tyrosine/serine/threonine in known phosphopeptides. To calculate the distribution (enrichment/depletion) of an amino acid at particular position, the local amino acid frequency was normalized with the global amino acid frequency. Local amino acid frequency was calculated using

$$L = n/(N_{sty}-n)$$

where:

n = number of a particular amino acid at a particular position with respect to phosphorylated tyrosine/serine/threonine residue

N_{sty} = Total number of tyrosine/serine/threonine residues in human proteome

Global amino acid frequency was calculated using

$$G = m/M$$

where:

m = Number of a particular amino acid in the human proteome

M = Total number of amino acids in the human proteome

Distribution value was calculated using

$$D_{dist} = L/G$$

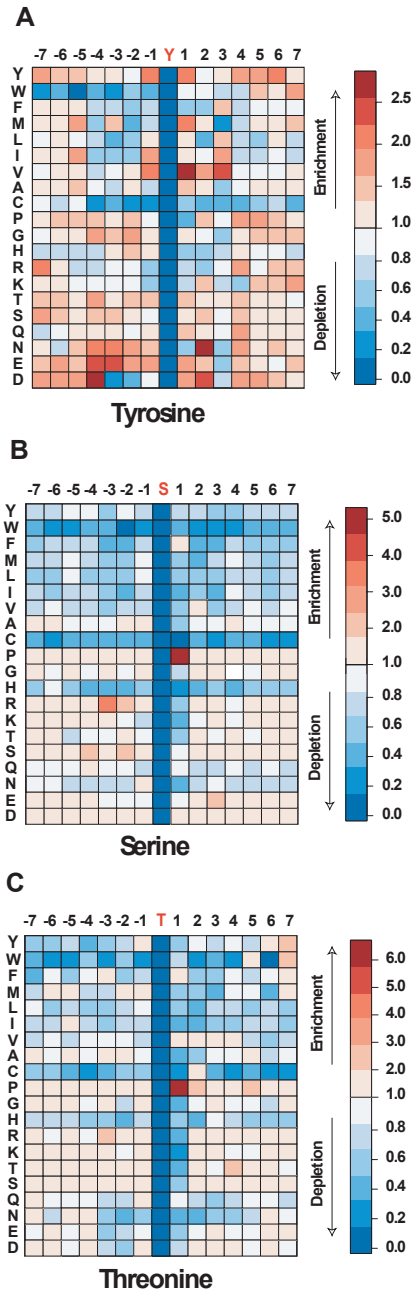

Figure 1: Heat maps for amino acid distribution surrounding known phosphorylated tyrosine (A), serine (B) and threonine (C) peptides in the human phosphoproteome, derived from Human Protein Reference Database. Enrichment or depletion of amino acids at particular positions is shown in red or blue, respectively, and was calculated as described in materials and methods.

For the above calculations, the database considered for the human proteome was RefSeq protein database release 12.0.

Direct motif finding method

An algorithm was developed using Python scripting language which generated motif sequences up to 15 amino acids long with a Tyr/Ser/Thr residue in the center and all possible amino acid combinations for all the positions flanking the fixed residue in the center. Although 20^{14} combinations can be generated, because of the computationally exhaustive nature of this analysis, we decided to use only those instances where a given amino acid combination existed in the human proteome. The artificially generated motif sequences with all possible combinations of amino acids (where the motif length can range from 2 - 15 amino acids) were used to identify matches in the phosphopeptide dataset and the sequences that were present in ≥ 40 for Ser and ≥ 10 for Tyr/Thr (these numbers were arbitrarily chosen based on the size of the Tyr/Ser/Thr dataset that was used) phosphorylation sites were considered for further analysis. The sequences that were not described in literature Amanchy et al. [21] were considered further and p-values were calculated as described below.

Statistical validation of motifs

By definition, a phosphorylation motif is associated with phosphorylation sites in the proteome. Accordingly, candidate motifs were evaluated by measuring this association, and comparing it to a null distribution obtained by permutation. Our measure of association is relative risk for the motif, given that a site is phosphorylated and is calculated as:

$RRmotif = (N_{PM} \times N_S) / (N_M \times N_{PS})$

Where:

RRmotif = Relative risk for phosphorylation given motif

N_{PM} = Number of times the given motif is associated with a phosphorylated site.

N_S = Total number of non-phosphorylated sites

N_M = Number of times the given motif is associated with a non phosphorylated site.

N_{PS} = Total number of phosphorylated sites.

Relative risk >1 indicate good association between the motif and phosphorylation events. Significance is measured by comparing the relative risk for the candidate motif to a motif-specific null distribution of relative risk obtained by substituting random combinations of amino acids for those that define the motif.

The p-value calculated as the probability that randomly selected mock motifs has relative risk (RRmock) at least as great as the value observed for the predicted motif. Mock motifs are obtained from a predicted motif by replacing each amino acid with randomly selected ones. Because motifs are typically small, we have exhaustively checked all possible mock motifs rather than generating them randomly.

Thus the empirical p-value is calculated as p = N/D;

Where:

N = Number of mock motifs with RRmock >= RRmotif

D = Total number of possible mock motifs.

Any global dependencies in the location or occurrence of amino acids are automatically taken into account without explicit modeling. We have also made a selection criterion that all the motifs described should have at least 3 amino acids defining the motif. Motifs with a relative risk of >= 5 and p-value ><= 0.05 were considered significant.

Experimental identification of novel phosphorylation datasets

We analyzed several mass spectrometry-derived published phosphorylation datasets including Molina et al. [20] from our laboratory and other studies [15-19,22] for the presence of novel predicted motifs.

Development of peptide microarrays and casein kinase I assay

A list of individual peptide sequences corresponding to novel predicted motifs from proteins which were previously not known to be phosphorylated was prepared. These peptides were synthesized as 11-mers and spotted onto the glass slides (Pepscan Systems, Lelystad, Netherlands) as described earlier Endicott et al. [23] in triplicates. Cysteine residues were replaced by alanine residues in all peptides. Kinase assays on peptide microarrays were performed as described earlier [24,21]. Briefly, custom peptide microarrays were incubated with 50 ng of recombinant Casein Kinase I enzyme (P6030S; New England Biolabs, Beverly, MA) and 300 µCi/ml γ33P-labeled ATP (AH9968; GE Healthcare Bio-sciences Corp., Piscataway, NJ) at 25°C for 1 h in a 120 µl reaction volume containing CKI reaction buffer (New England Biolabs, Beverly, MA) supplemented with 200 µM ATP. The reaction was stopped and the following washing steps were performed; 2 washes in 2M sodium chloride containing 1% Triton X-100 followed by 3 washes in phosphate buffered saline containing 1% Triton X-100 and 1 wash in distilled water. The glass slides were then air dried and exposed to the phosphorimager screen for 12 hours and scanned using Bio-Rad Molecular Imager FX (Bio-Rad Laboratories, Inc, Hercules, CA). The image was then processed using GenePix Pro 6.0 software (Molecular Devices Corporation, Sunnyvale, CA). Phosphorylation patterns arising from radiolabeled γ33P-labeled ATP were captured on phosphorimager screen and image acquisition and analysis was performed. The assay was performed in triplicate and each peptide is spotted on a glass slide in triplicates. The intensity values obtained were transformed by applying a log base 2 transform. Spatial or positional variations arising from kinase assay on the slide were normalized and any positional effects arising from this were subtracted by applying local regression analysis (loess) with respect to chip coordinates Diks

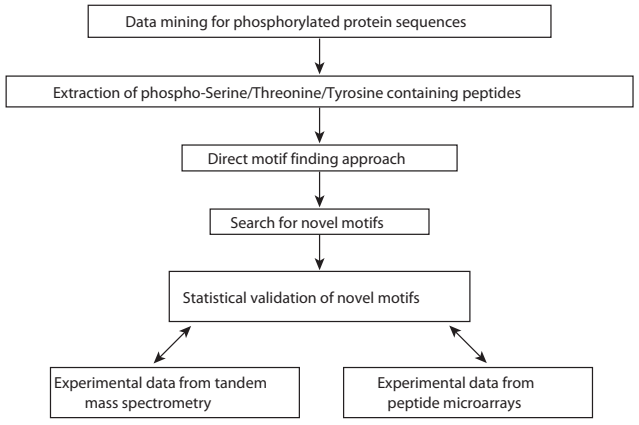

Figure 2: A schematic of the direct motif finding approach used in grouping peptide sequences into bins to derive novel phosphorylation motifs.

	Motif	Sites in the known phospho-proteome	Relative risk in tyrosine phosphorylated peptides versus entire human proteome	p-value
1	pY[D/E]XP	44	10.94	0.00251
2	[D/E]XXpYXXP	42	7.73	0.01508
3	[D/E]XXXpYXXP	40	7.72	0.00503
4	[D/E]XpYXXP	36	6.42	0.02764
5	pYX[D/E]P	26	5.79	0.01508
6	[D/E]SXpY	25	12.22	0.00251
7	pYXVP	25	8.8	0.00501
8	pY[D/E]XXP	24	6.36	0
9	SXXXXXXpYXXP	23	8.68	0.00251
10	RXXXXXXpYXXXXG	23	7.06	0.00251
11	SX[D/E]XpY	23	5.88	0.03769
12	pYXXPXS	22	6.39	0.00501
13	[D/E]XSXpY	21	10.52	0.00251
14	pYXXPP	21	5.18	0.01253
15	SXpYXXP	20	20.56	0.00251
16	pYXXPXXXE	18	6.48	0.00501
17	SXXpYXXP	16	11.27	0.00752
18	RXIXXXXpY	16	6.74	0.00501
19	SXpYXXXX[D/E]	16	6.74	0.01759
20	[K/R]XXSXXpY	16	6.42	0.01256
21	pYXXPXXA	16	5.35	0.00752
22	pYXXMXP	15	19.62	0
23	pYXXXRXY	15	9.88	0
24	DXpYXXXXXR	15	5.9	0.03008
25	pYTXXXXXK	15	5.45	0.00501
26	PXXXXpYXXXXXXG	15	5.07	0.00752
27	DXXpYXXXXXA	15	5.04	0.02506
28	pYXXTXXY	14	11.37	0.00251
29	pYVXXXXY	14	10	0
30	pYXXTXXXR	14	6.59	0.00251
31	SXXSXXpY	14	6.47	0.01253
32	SXXpYXXL	14	5.71	0.03008
33	SXXXXXXpYV	14	5.44	0.01253
34	pYVXXXP	14	5.32	0.01003
35	VXXXXpYXXP	14	5.09	0.01754
36	DSXXpY	13	9.63	0.00251
37	TXXpYXXP	13	9.56	0.01253
38	RXXXXXXpYY	13	8.13	0
39	DXXpYY	13	7.04	0.01003
40	SXXXpYXXV	13	6.27	0.01504
41	DXXpYXXXQ	13	6.15	0.01754
42	SXXXXXpYXXXXP	13	6.01	0.00752
43	IEXXXpY	13	5.9	0.00752
44	QXXpYXXP	13	5.79	0.02757
45	pYXVXP	13	5.46	0.00251
46	SXXXXXXpYXXXXP	13	5.14	0.01253
47	DXXYpY	12	44.77	0
48	TXpYXXXXXY	12	31.81	0
49	TpYXXP	12	30.99	0
50	TXpYXXXXXXR	12	21.98	0
51	pYMXXXP	12	17.02	0.00251
52	SXpYXXXE	12	11.4	0.00251
53	pYXXXXXYR	12	8.89	0.00501
54	SXXpYXXXXP	12	7.9	0.00752
55	SXXXXXXpYXN	12	6.91	0.00251
56	SXpYXS	12	6.64	0.02506
57	TXXXXXXpYXXP	12	6.01	0.00752
58	pYQXP	12	5.93	0.01003
59	SXXXXXpYXXXP	12	5.76	0.00251
60	SXXXXXpYXXP	12	5.65	0.00752
61	TXXXXXXpYV	12	5.42	0.01504
62	pYXXPXXXV	12	5.12	0.01003
63	pYAXP	12	5.12	0.01754
64	QXXXpYXXP	12	5.12	0.02256
65	SXXXpYXS	12	5.12	0.02506
66	TXpYVXXRXYR	11	1108.01	0
67	TXpYVXTXXYR	11	1108.01	0
68	TXpYVXTRXXR	11	1108.01	0
69	TXpYXXTRXYR	11	1108.01	0
70	TXpYVXTRXYR	11	1108.01	0
71	PXTpY	11	24.09	0.00251
72	SpYXXP	11	19.44	0.01253
73	GSpY	11	15.18	0.02005
74	TDpY	11	15.18	0.02005
75	SXXXXXXpYXXM	11	14.21	0
76	SXXpYXD	11	8.79	0
77	TXDXpY	11	8.72	0.01003
78	TXXXXXXpYXD	11	7.59	0
79	SXXpYXXXA	11	7.49	0.01253
80	GSXXpY	11	7.19	0.01003
81	SXXXpYXXXXP	11	6.52	0.00251
82	SDXXpY	11	6.52	0.01504
83	TXXXXpYXXXS	11	6.4	0.00251
84	pYXDXXQ	11	5.49	0.00251
85	pYEXXXXXQ	11	5.33	0.00752
86	PXpYXN	11	5.33	0.03509
87	DXXXpYXXXXP	11	5.11	0.01504
88	pYXXPQ	11	5.06	0.01504
89	TpYXD	10	33.58	0.00251
90	EXTpY	10	27.22	0
91	YpYXXXXXG	10	21.9	0.00501
92	DSpY	10	18.31	0.01003
93	DXXXTXpY	10	17.67	0
94	MXXpYXXXR	10	9.59	0.00251
95	SXpYE	10	9.16	0.01754
96	SXpYXXXS	10	7.35	0.01754
97	TXXXXpYXXI	10	6.54	0.00501
98	SXpYXXXXS	10	6.1	0.03008

99	pYYXXXXXG	10	5.93	0.00251
100	PXXXXpYXN	10	5.53	0.01003
101	SXXXXpYXXP	10	5.53	0.01253
102	SXpYXXXXXS	10	5.47	0.0401
103	SXXpYXXXS	10	5.42	0.02506
104	TXXEXpY	10	5.25	0.04261
105	SXXXXXXpYXXXXQ	10	5.17	0.01003

Table 1: A list of novel phosphotyrosine motifs.

et al. [25]. The distribution of intensities of the phosphorylation of the peptides on the microarrays was smoothed by applying kernel density estimation [25]. The triplicate design of the peptide microarray was designed specifically to allow much stricter control of false positive rates. We applied high stringency criteria for further selection of peptides with high intensity values that are more likely to be true positives. If peptides for which all 3 replicate probes exceed the 99% threshold was selected, then only 1 in 1,000,000 is expected to be a false positive. Thus, we chose the top 1% of the peptides as true CKI candidate substrates.

Results and Discussion

Protein kinase substrates have traditionally been evaluated by metabolic labeling of cells using ^{32}P labeled ATP followed by 2D-gel electrophoresis of protein samples. Other methodologies for the identification of kinase substrates include screening of cDNA expression libraries or immobilization of expression clones on nitrocellulose membranes [27-29]. As discussed above, synthetic degenerate peptide libraries have also been used to determine kinase specificity *in vitro* [9,30]. Most recently, mass spectrometry based large scale proteomic studies for global analysis of phosphorylation [15,21,31,32] have generated large datasets of phosphorylation sites.

In the context of signaling pathways, the determinants of specificity of kinases and phosphatases for their substrates are not well understood. The specificity of a kinase is believed to depend both on the sequence and structure of the catalytic domain and on the sequence of residues surrounding the substrate phosphorylation site. However, identification of specificity determining residues in kinase substrates by experimental approaches can be laborious and time consuming Li et al. [33]. Establishing kinase-substrate relationships is important to study the phosphorylation networks in signal transduction pathways.

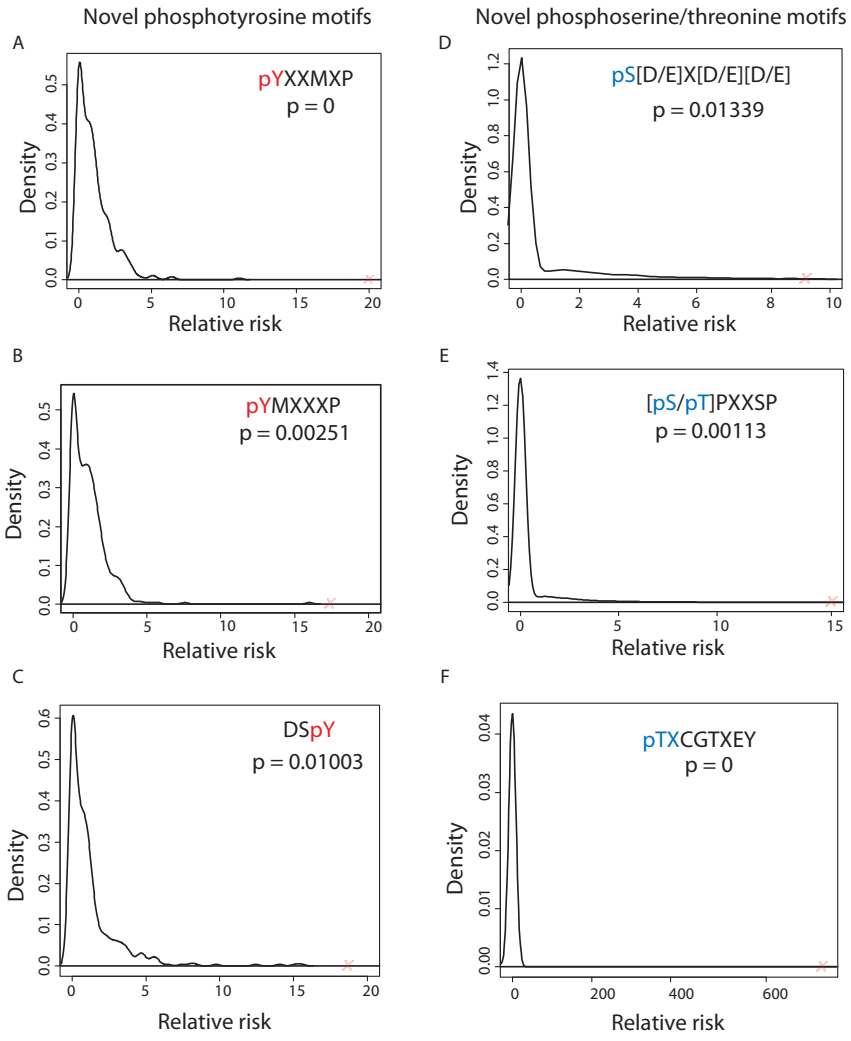

Figure 3: Distribution curves of select predicted novel phosphotyrosine (A-C) and phosphoserine/threonine (D-F) motifs. The X in red represents the relative risk for the motifs and the area under the curve to the left of X represents the number of permuted sequences with relative risk lesser than the motif under study. The phosphorylated residues are shown in red (pY) or blue (pS/pT).

Amino acid distribution surrounding phosphorylated tyrosine/serine/threonine residues in the human proteome

We first determined the abundance of certain amino acids at certain positions with respect to the phosphorylated serine/threonine/tyrosine residues by generating 'heat maps' (Figure 1). Yaffe and colleagues have previously mapped the amino acid distribution surrounding the serine/threonine/tyrosine residues in the human proteome and in residues phosphorylated by PKA, Src and EGFR kinases using this method Yaffe et al. [34]. These heat maps provide an average representation of the amino acids that are enriched at specific positions surrounding the phosphorylated tyrosine/serine/threonine residues.

We generated heat maps using HPRD phosphorylation dataset of 4,810 phosphorylation sites (pSer: 3,184; pThr: 760; pTyr: 866). Figure 1A represents the distribution of amino acids surrounding phosphorylated tyrosines. We observe an enrichment of valine at +1 and +3 positions, aspartate at -4 and +2 positions and asparagine at +2 positions as well as glutamate at -4 and -3 positions. Presence of hydrophobic amino acids at +1 and +3 positions is known to occur in motifs known to bind to SH2 domains [2,9,10]. Enrichment of asparagine at +2 position is typical of Grb2 SH2 domain binding motif YXN Yaffe [2]. Finally enrichment of acidic amino acids glutamate and aspartate N-terminal to tyrosine is typical of several tyrosine kinase substrate motifs Songyang et al. [45]. Although the heat map revealed the presence of some of the well known motifs, they failed to reveal the presence of any novel ones.

Similarly, Figure 1B and Figure 1C represent the distribution of amino acids surrounding the phosphorylated serines and threonines, respectively. Enrichment of proline at +1 and arginine at -3 positions is easily visualized from these heat maps. In phosphoserine/threonine peptides, proline at +1 position is characteristic of proline directed serine/threonine kinase substrate recognition motifs. Similarly, enrichment of arginine at -3 position is observed in Akt kinase substrate motifs. Enrichment of arginine at +2 and +3 positions is characteristic of Akt kinase substrate motifs while enrichment of glutamate at +3 position is characteristic of casein kinase substrate motifs [35,36].

Direct motif finding approach

Although many phosphorylation motifs are already known, we reasoned that additional motifs could be discovered with the availability of a larger dataset of phosphorylation sites. We decided to undertake

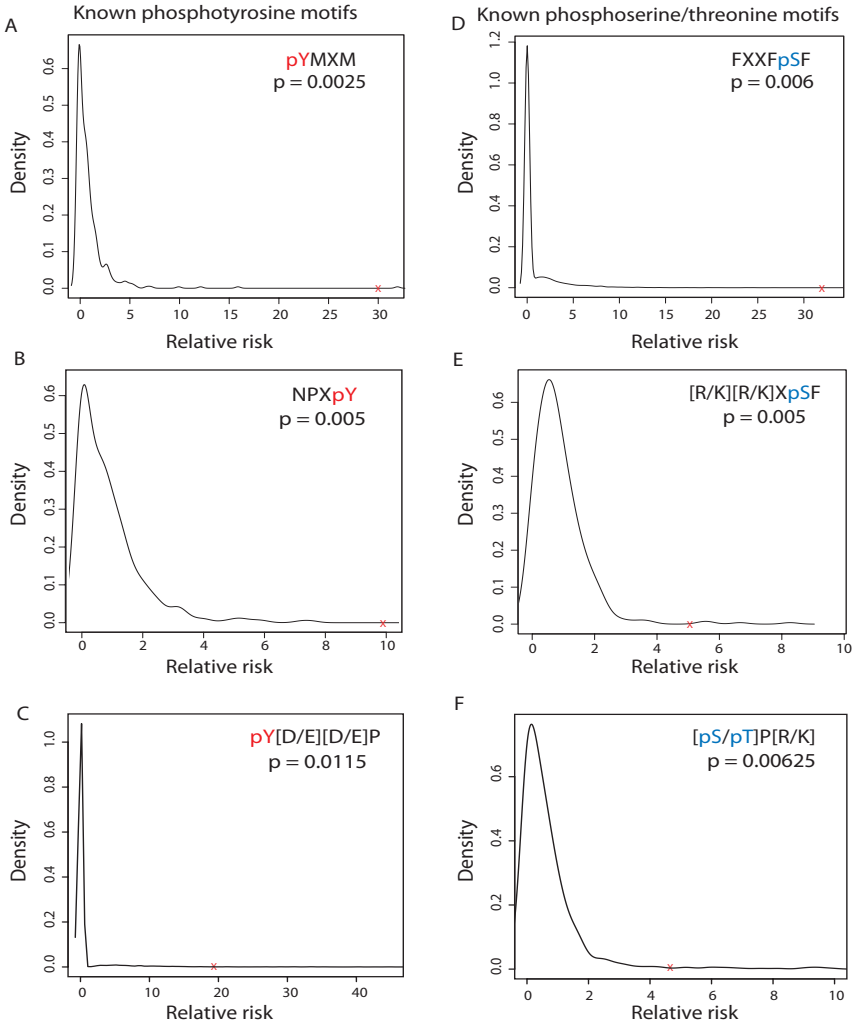

Figure 4: Distribution curves of select known phosphotyrosine (A-C) and phosphoserine/threonine (D-F) motifs. The X in red represents the relative risk for the motifs and the area under the curve to the left of X represents the number of permuted sequences with relative risk lesser than the motif under study. The phosphorylated residues are shown in red (pY) or blue (pS/pT).

a bioinformatics approach, designated direct motif finding method, which was carried out to specifically look for a minimum number of amino acids being conserved among known phosphopeptide sequences in human proteome. For this, we used peptide sequences containing the known phosphorylation sites from Human Protein Reference Database (HPRD) at the time this analysis was initiated. The initial step was to generate all phosphorylated peptide sequences from HPRD into 15 amino acid length sequences, with the phosphorylated residue in the centre (8th residue) for heat map generation and for use by the direct motif finding approach. In instances where the phosphorylated residue is close to N-terminal end or C-terminal end, the amino acid length will be less than 15 depending on the position. The overall strategy is depicted in Figure 2.

It is important to determine whether the identified motifs were previously described in the published literature before classifying them as novel. We initiated a curation effort to find all the phosphorylation motifs described in the literature. Overall, we have catalogued 122 phosphotyrosine motifs and 175 phosphoserine/threonine motifs from the literature that are available as a resource called 'PhosphoMotif Finder' in HPRD (http://hprd.org/PhosphoMotif_finder). The motifs that were identified by the computational approach were classified as novel motifs if they were not previously described in the literature. The motifs identified by computational method that were not described in the literature were further subjected to statistical validation. 94 novel phosphotyrosine motifs were found by direct motif finding approach, but when degeneracy (D/E and K/R) was allowed in this approach, 105 novel motifs were identified (Table 1). 70 novel phosphoserine/threonine motifs were identified using direct motif binding approach and 194 novel motifs were identified when degeneracy was allowed (Table 2).

Statistical significance testing of motifs

It is possible that the identified motifs occur very commonly in the human proteome and are not necessarily related to the phosphorylation event. To test whether the motifs were enriched in the known phosphopeptide dataset, we undertook a statistical validation approach where the relative risk of the each motif was calculated (see materials and methods). Though relative risk >1 indicates a good association between the motif and phosphorylation events, we had chosen a higher threshold of relative risk >5 as significant. Statistical significance was next computed by comparing the relative risk for the candidate motif to a motif-specific null distribution of relative risk obtained by substituting all combinations of amino acids for those that define the motif. Distribution curves for select predicted phosphotyrosine motifs are shown in Figure 3A-3C and for phosphoserine/threonine motifs in Figure 3D-3F.

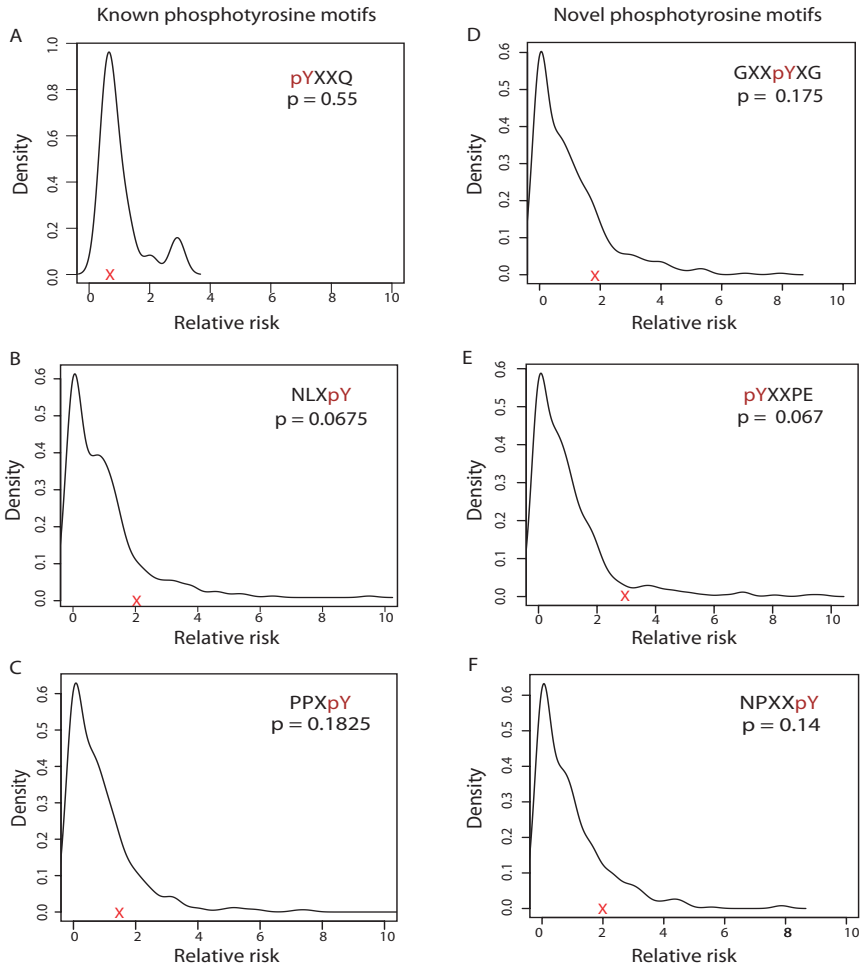

Figure 5: Distribution curves of known (A-C) and novel (D-F) phosphotyrosine motifs with insignificant p-values. The X in red represents the relative risk value for the motif, and the area under the curve to the left of X represents the number of permuted sequences with relative risk lesser than the motif under study.

The distribution curves for known phosphotyrosine and phosphoserine/threonine motifs with statistically significant p-values are represented in Figure 4. These motifs also show better relative risk than their permuted motif sequences. The distribution curves for the known and novel phosphotyrosine motifs that are not statistically significant are shown in Figure 5. In these figures, it can be observed that the majority of permuted motifs have relative risk greater than that of the candidate motif presenting statistically insignificant p-values, indicating that these motifs could have occurred by chance. The distribution curves for the known and novel phosphoserine/threonine motifs with statistically insignificant p-values are represented in Figure 6. These data illustrate the following points:

(1) The definition of a phosphorylation motif is not defined statistically, although they can indeed be tested for significance using statistical methods. A case in point is the motif, p[S/T]XX[D/E], a well known CKII substrate motif Meggio et al. [37], which has been shown to be biologically relevant although it is statistically insignificant (Figure 6B). In any case, 3 motifs related to this motif, (pS[D/E]X[D/E][D/E], p[S/T][D/E]S[D/E] and [D/E][D/E]p[S/T]X[D/E]), were identified in our analysis which were all statistically significant (Table 2). Another example of a statistically insignificant motif (Figure 6C) is the well known protein kinase C substrate motif p[S/T]X[R/K], which is optimal for phosphorylation of its substrates [38,39]. Again, 3 similar motifs identified from our analysis (PXpSPX[R/K], p[S/T]PZX[R/K], p[S/T]P[R/K][N/Q]) were statistically significant (Table 2).

(2) The converse situation, i.e. if a motif is statistically enriched then its chances being a biologically relevant motif is likely, although there is no reason to believe that this must always be the case.

(3) It is possible that as additional phosphorylation sites are identified in the future, some of the motifs that do not currently reach statistical significance might become significant.

Comparison with other algorithms

Although there are several motif discovery programs, there is only one algorithm which is specific for identification of protein phosphorylation motifs, motif-x [6]. This is a web-based program that uses an iterative statistical approach to identify phosphorylation motifs from any phosphorylated sequence dataset. We input our dataset of phosphorylated sequences into motif-x in a pre-aligned format using IPI human proteome as the background dataset (options selected: width: 15, occurrences: 10 in case of tyrosine and threonine and 40 in case of serine, significance: 0.049 as it does not allow 0.05). Unfortunately, we were not able to use the background dataset that we have used in our methods, i.e., the RefSeq. motif-x predicted a total of 110 motifs: 33 serine, 35 threonine and 42 tyrosine based motifs. We next carried

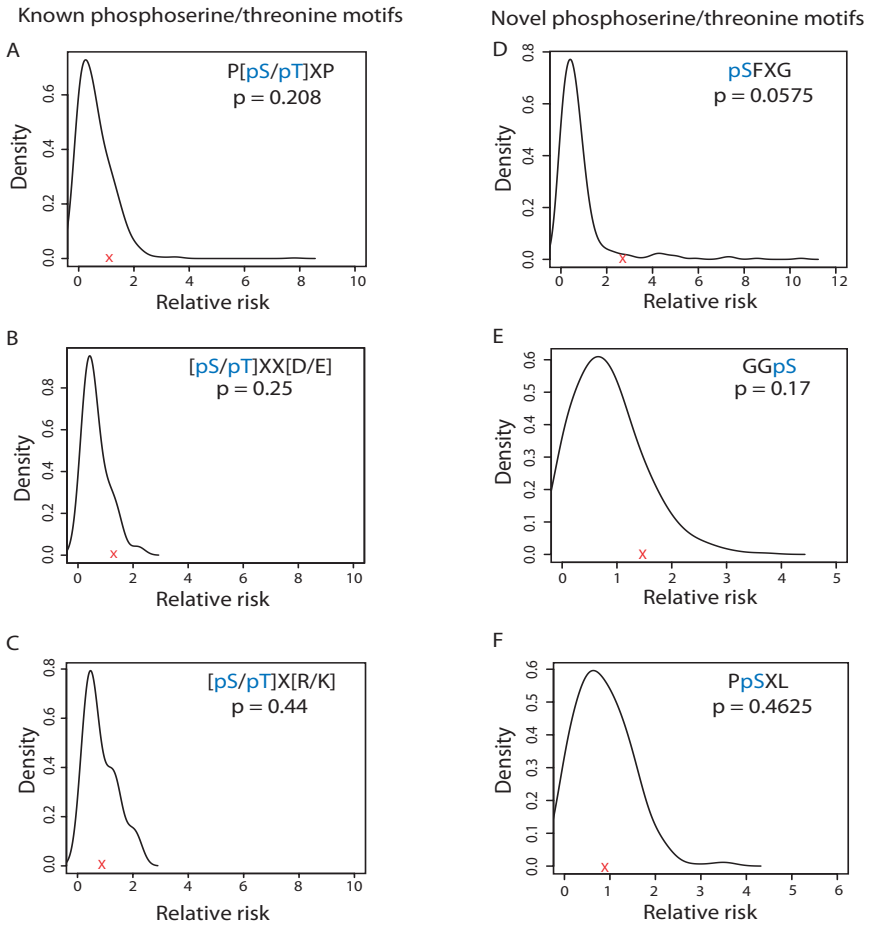

Figure 6: Distribution curves of known (A-C) and novel (D-F) phosphoserine/threonine motifs with insignificant p-values. The X in red represents the relative risk value for the motif, and the area under the curve to the left of X represents the number of permuted sequences with relative risk lesser than the motif under study.

	Motif	Sites in the known phospho-proteome	Relative risk in tyrosine phosphorylated peptides versus entire human proteome	p-value
1	[K/R]xxxxx[pS/pT]P	209	7.51	0.00251
2	[K/R]xxxx[pS/pT]P	203	7.39	0.00377
3	[K/R]xxxxxx[pS/pT]P	187	6.69	0.00503
4	[D/E]xxx[pS/pT]P	187	6.55	0.00503
5	[D/E]xxxxx[pS/pT]P	187	6.37	0.00754
6	[pS/pT]Pxxxxx[K/R]	180	7.02	0.00251
7	[K/R]xxx[pS/pT]P	174	6.17	0.00628
8	[D/E]xxxx[pS/pT]P	174	5.52	0.01131
9	[D/E]xxxxxx[pS/pT]P	171	6.15	0.00879
10	Sxxxx[pS/pT]P	170	11.45	0
11	[pS/pT]Pxx[D/E]	163	6.63	0.00251
12	[pS/pT]Pxxxxx[D/E]	161	6.4	0.00879
13	[pS/pT]Pxxx[D/E]	159	6.27	0.00628
14	[pS/pT]Pxxxx[D/E]	158	6.12	0.00754
15	[pS/pT]P[D/E]	157	5	0.01131
16	S[K/R]xx[pS/pT]	153	7.69	0
17	[pS/pT]Px[D/E]	139	5.29	0.01005
18	Axxxxxx[pS/pT]P	126	6.31	0.00752
19	[pS/pT]PxxxA	123	5.59	0.01003
20	Gxx[pS/pT]P	122	5.92	0.00627
21	[pS/pT]PxxA	119	6.1	0.00752
22	[pS/pT]PxxxxxG	116	7.33	0.00125
23	Axxxx[pS/pT]P	115	5.68	0.00877
24	Axx[pS/pT]P	115	5.3	0.01003
25	[K/R]xS[pS/pT]	112	20.92	0
26	[pS/pT]PxxxxA	110	5.39	0.01128
27	[pS/pT]PxxxxxA	108	5.9	0.01253
28	[K/R]xSxxx[pS/pT]	107	5.86	0
29	Sx[pS/pT]xS	106	6.95	0.00501
30	[pS/pT]PxA	105	5.53	0.01003
31	Gxxxx[pS/pT]P	104	5.55	0.01128
32	[pS/pT]PxxxxT	103	7.57	0.00501
33	Axxxxx[pS/pT]P	103	5.24	0.01253
34	[pS/pT]PxxxxG	100	6.58	0.00627
35	[pS/pT]PxxxG	98	5.35	0.01128
36	[pS/pT]PxxxxxT	97	6.74	0.00627
37	Gxxxxx[pS/pT]P	97	5.01	0.01378
38	Txx[pS/pT]P	96	14.99	0
39	Txxx[pS/pT]P	95	12.06	0.00125
40	pS[D/E]x[D/E][D/E]	95	8.95	0.01339
41	[pS/pT]PxxG	95	5.93	0.00877
42	Txxxxx[pS/pT]P	92	9.86	0.00125
43	Sx[pS/pT]xxxxxx[K/R]	92	7.49	0.00377
44	Sx[pS/pT]xxxxxxS	89	6.21	0.01003
45	SS[pS/pT]	88	9.94	0.00251
46	[pS/pT]PxxxT	85	6.36	0.00501
47	Sx[pS/pT]xxxS	85	5.83	0.00501
48	SxxxSx[pS/pT]	85	5.52	0.01128
49	pS[D/E]x[D/E]xx[D/E]	84	8.14	0.01664
50	V[pS/pT]P	84	5.16	0.02506
51	Vxxxxx[pS/pT]P	82	5.44	0.01128
52	RSxxx[pS/pT]	81	8.45	0
53	[pS/pT]PxT	81	5.93	0.00752
54	Vx[pS/pT]P	81	5.19	0.01253
55	Sx[pS/pT]xxP	78	9.95	0.00125
56	pSx[D/E][D/E]x[D/E]	78	6.8	0.01964
57	pSxx[D/E]x[D/E][D/E]	78	6.69	0.01814
58	Txxxx[pS/pT]P	77	9.02	0.00125
59	Qxxxxxx[pS/pT]P	77	6.22	0.00877
60	Qxxxx[pS/pT]P	77	6.19	0.00752
61	Sx[pS/pT]xxxxS	77	5.56	0.01128
62	pSx[D/E][D/E][D/E]	76	7.3	0.01601
63	[pS/pT]PxxxV	76	5.6	0.00877
64	SxxRxx[pS/pT]	75	5.83	0
65	SxxxxSx[pS/pT]	75	5.24	0.01003
66	S[pS/pT]xxS	74	8.33	0.01003
67	RxxSxxx[pS/pT]	74	7.65	0
68	pS[D/E]x[D/E]xxx[D/E]	74	7.18	0.02352
69	[pS/pT]PxxxxxQ	74	6.33	0.01003
70	[pS/pT]PxxQ	74	6.27	0.00627
71	SxxxRxx[pS/pT]	74	5.29	0
72	Qxx[pS/pT]P	73	5.92	0.00752
73	pS[D/E][D/E]xx[D/E]	72	8.05	0.01851
74	SxxpSx[D/E]	71	5.3	0.00251
75	Sx[pS/pT]xxxxxS	71	5.06	0.02005
76	pS[D/E]xxx[D/E][D/E]	70	7.9	0.01839
77	[D/E]pSx[D/E][D/E]	70	6.72	0.03953
78	Qxxx[pS/pT]P	69	5.51	0.01128
79	[pS/pT]PxV	68	5.27	0.01128
80	SpSx[D/E]	67	10.5	0.00251
81	Sx[pS/pT]xxxxxxP	67	8.73	0.00251
82	SxpSxxx[D/E]	67	6.07	0.00754
83	SxxxxxS[pS/pT]	66	8.57	0.00627
84	[D/E]xpS[D/E]x[D/E]	66	6.41	0.04116
85	[pS/pT]PQ	66	5.75	0.00752
86	PxxxxSx[pS/pT]	65	7.86	0.00125
87	pS[D/E][D/E]x[D/E]	65	6.78	0.0284
88	[D/E]xxxxpS[D/E]x[D/E]	64	6.52	0.03315
89	Sxxx[pS/pT]xP	64	5.27	0.00125
90	Sxxx[pS/pT]xxE	64	5.14	0
91	Sx[pS/pT]xxxxP	63	8.43	0.00125
92	pS[D/E]xx[D/E][D/E]	63	7.09	0.02502
93	Sxxx[pS/pT]xxxT	63	5.9	0.00125
94	[pS/pT]PxxxxQ	63	5.38	0.01253
95	SpSxxxx[D/E]	62	9.68	0.00503
96	Sx[pS/pT]xxxxxP	62	7.61	0.00125
97	[D/E]pS[D/E]xxx[D/E]	62	6.73	0.04053
98	pSx[D/E][D/E]xx[D/E]	62	6.25	0.02252

99	Sxxx[pS/pT]xxxxP	62	5.05	0.00125
100	RxxxSx[pS/pT]	61	9.94	0
101	S[pS/pT]xS	61	7.13	0.02005
102	Sxx[pS/pT]xxE	61	6.17	0
103	[D/E]xxxpS[D/E]x[D/E]	61	6.13	0.03553
104	[K/R]xxxSpS	60	11.01	0
105	SpSxxxxx[D/E]	60	10.33	0
106	S[pS/pT]xxxxxS	60	7.63	0.02005
107	LxxSx[pS/pT]	60	6.17	0.00501
108	pSxx[D/E][D/E]x[D/E]	60	5.78	0.02677
109	SxpSxxxx[D/E]	60	5.6	0.02764
110	SxSx[pS/pT]P	59	42.43	0.00081
111	SxxxxS[pS/pT]	58	6.93	0.02506
112	pSxxE[D/E][D/E]	57	6.76	0.01913
113	pSx[D/E]x[D/E]x[D/E]	56	6.02	0.02777
114	SxS[pS/pT]	56	5.97	0.02632
115	[pS/pT]PxxSP	55	14.68	0.00113
116	S[pS/pT]xxxxxxP	55	13.01	0.00125
117	SxRx[pS/pT]	55	6.05	0.01003
118	pSx[D/E]x[D/E][D/E]	55	5.86	0.02627
119	SxRxx[pS/pT]	55	5.42	0.00251
120	PxxPx[pS/pT]P	54	9.73	0.01275
121	Sx[pS/pT]xxxP	54	6.91	0.00251
122	Sxxx[pS/pT]xxxxR	54	5.39	0
123	pSxx[D/E]xx[D/E][D/E]	54	5.35	0.03503
124	S[pS/pT]xP	53	11.16	0.00251
125	[K/R]xxxxxSpS	53	8.97	0.00251
126	PxxxSx[pS/pT]	53	6.97	0.00376
127	RSx[pS/pT]P	52	77.43	0.00025
128	Sx[pS/pT]xxxxR	52	8.88	0
129	GSx[pS/pT]	52	6.56	0.00251
130	pS[D/E]xxx[D/E]x[D/E]	52	5.51	0.04478
131	SxPx[pS/pT]P	51	23.34	0.00388
132	SPxx[pS/pT]P	51	22.75	0.00113
133	pSPxxx[K/R][K/R]	51	14	0.0045
134	pS[D/E]xx[D/E]x[D/E]	51	6	0.03828
135	[K/R]xx[pS/pT]PP	50	18.42	0.00213
136	[K/R]xxpSPxP	50	17.69	0.0045
137	S[pS/pT]xxxxP	50	11.62	0.00125
138	SpSxxxx[K/R]	50	8.76	0.01005
139	SpSxxx[K/R]	50	8.71	0.01005
140	Nxxx[pS/pT]P	50	6.2	0.00627
141	RxxSxx[pS/pT]	50	5.7	0.00251
142	SpSxxx[D/E]	49	7.82	0.02513
143	[K/R]xxSpS	48	7.82	0.0201
144	SpSxxxxS	48	6.27	0.04762
145	SpSxxxS	48	5.89	0.04762
146	SxpSPxP	47	49.88	0.001
147	[D/E]xxpS[D/E]xE	47	7.02	0.03064
148	SxpS[D/E]x[D/E]	46	21.5	0.00675
149	SpSxx[K/R]	46	8.79	0.00503
150	[D/E]xxSpS	46	8.19	0.01759
151	pSxx[D/E]xEx[D/E]	46	6.54	0.02501
152	[D/E]xSpS	45	7.82	0.01256
153	SxpSxxxG	45	6.96	0
154	pSxxE[D/E]xx[D/E]	45	6.76	0.01888
155	SxpSxxxL	45	5.36	0.01754
156	SpSxxxxxx[D/E]	44	6.58	0.03518
157	[D/E]pS[D/E]xxE	43	8.42	0.03064
158	SxpSxA	42	6.84	0.01253
159	pSxS[D/E]x[D/E]	42	6.7	0.01988
160	[D/E]pSxxEx[D/E]	42	6.64	0.04252
161	SxpSxxxxxD	41	10.06	0
162	SpSxxxxx[K/R]	41	7.37	0.02261
163	PxxSxpS	41	6.16	0.01253
164	[D/E]pSxxE[D/E]	41	5.98	0.04865
165	pSPxxI	41	5.29	0.02506
166	GxxxxpTxxxxxxY	16	11.59	0
167	GxxxxpTxCxxxxY	14	305.14	0.00025
168	GxxxxpTxCGxxxY	13	1416.74	0.00001
169	GxxxxpTxxGxxxY	13	94.45	0.00125
170	pTxxVxxxW	13	27.25	0
171	pTxxVxTxW	12	435.92	0.00013
172	TxpTxxxxxxY	12	38.46	0
173	pTxxxxTxW	12	24.91	0
174	TxpTxxGxxxY	11	342.51	0.00025
175	SxxxpTxxxTP	11	72.65	0.00163
176	pTxCGTxEY	10	726.54	0
177	GxxxxpTxCxTxxY	10	726.54	0.00001
178	pTxCGxxEY	10	726.54	0.00001
179	pTxxGTxEY	10	726.54	0.00001
180	pTxCxTxEY	10	726.54	0.00001
181	TxpTxxxTxxY	10	435.92	0.00025
182	pTxCxxxEY	10	363.27	0.0005
183	TxpTFxG	10	311.37	0.00013
184	TxpTxCxT	10	242.18	0.00025
185	SpTxxGxP	10	198.15	0.00163
186	GxxxxpTxxxTxxY	10	155.69	0.0005
187	pTxxxTxEY	10	145.31	0.00063
188	SxxxpTxVxT	10	128.21	0.00038
189	pTxxGxxEY	10	128.21	0.00075
190	GxxxxpTxxxxPxY	10	128.21	0.001
191	SpTxxxTP	10	114.72	0.00175
192	TxxpTxxxxxxY	10	23.69	0
193	GpTxxxxxPE	10	18.95	0.02113
194	LxxxxTxpT	10	8.48	0.00501

Table 2: A list of novel phosphoserine/threonine motifs.

out a comparison of these 110 motifs predicted by motif-x against the motifs predicted by our methods. Eleven motifs predicted by motif-x were identical to those predicted by us. Further 34 of the motifs (31%) were part of the longer motifs predicted by our method. Conversely, the large majority of the motifs predicted by us were not identified by motif-x. One possible reason for this might be that ~43% of their motifs

are smaller than 3 amino acids where as we had a cut off of minimum 3 amino acids for any given motif. Overall, the motifs generated by these approaches are mostly non-overlapping and hence the strategies can complement each other to achieve a more consolidated list of predicted motifs for the research community.

Validation of predicted motifs by mass spectrometry derived phosphorylation sites

We further investigated to see if the newly discovered motifs were present in more recent mass spectrometry based experimental phosphorylation datasets [15-20,22]. We expected that the predicted motifs that we have described might be present in this dataset if they were indeed commonly occurring motifs. We found that there were 244 of 1,435 (17%) cases where the novel phosphorylation site was in the context of the novel motifs predicted by us (Table 3). In particular, several of the motifs that we had previously established as statistically significant (pS[D/E]X[D/E][D/E], pSPXXXP, GGpS, pSPXXXT, DXXXp[S/T]P) were found. From the phosphorylation sites catalogued in the above mentioned large-scale studies, we found 16,535 instances where the novel phosphorylation site was in the context of the novel motifs predicted by us (Table 3). Again, pS[D/E]X[D/E][D/E], pSPXXXP, DXXXp[S/T]P, pSPXXXT and QXp[S/T]P were the most enriched motifs from the catalog. The identification of novel motifs from these large scale phosphorylation datasets acts as validation of newly discovered motifs.

	Novel motif	Instances in the known phosphoproteome	Instances in phosphorylation dataset identified by Molina et al., [20]	Instances in phosphorylation dataset identified by five other mass spectromtery studies
1	pS[E/D]X[E/D][E/D]	99	55	105
2	pSPXXXP	155	31	1146
3	GGpS	29	13	33
4	pSPXXXT	58	27	1045
5	DXXXp[S/T]P	82	14	1086
6	QXp[S/T]P	47	12	1034
7	p[S/T]PPP	56	12	1056
8	PXpSPX[R/K]	29	8	1231
9	PSp[S/T]P	37	11	1035
10	EXSXp[S/T]P	20	9	1283
11	PPp[S/T]P	36	9	1035
12	PPXp[S/T]P	46	9	1032
13	PpSXL	17	7	15
14	PLp[S/T]P	32	6	1027
15	TpTP	31	5	75
16	QXPp[S/T]P	20	5	1006
17	GXpYG	11	3	2
18	GXXpYX[G/S]	20	2	11
19	Mp[S/T]P	26	2	1006
20	p[S/T]PGG	15	2	1009
21	pYXXPE	14	1	8
22	QPXp[S/T]P	18	1	1011

Table 3: Incidence of predicted motifs in experimental data sets obtained by recently published large scale phosphoproteome analyses.

Testing of predicted motifs using peptide microarrays

Identification of phosphopeptides is a challenging task in proteomics. Earlier we have demonstrated that peptide microarrays are a good platform for high-throughput identification of phosphopeptides Amanchy et al. [40]. In this study, we sought to test whether a subset of the novel motifs can indeed be experimentally phosphorylated using a platform that allows testing of individual peptide sequences in a high-throughput fashion. For this purpose, peptides containing predicted motifs that were derived from proteins not previously known to be phosphorylated were synthesized and spotted on peptide microarrays. 724 peptide sequences from 449 proteins containing novel serine/threonine phosphorylation motifs were spotted as triplicates on the glass slide along with 768 empty spots as negative controls.

Although numerous studies pertaining to this CKI have been carried out, there is disagreement about the exact consensus sequence recognized by CKI. We sought to explore the CKI motifs using our predictions by using an experimental platform. Some of the earlier studies have pointed that efficient activity of CKI requires an already phosphorylated residue Joseph et al. [24]. In vitro studies using synthetic peptide libraries showed that CKI displays substrate specificity much more readily for peptides with large acidic clusters on their N-terminal region Marin et al. [36] and that phosphorylated residues are not required for the phosphorylation of proteins by CKI Flotow et al. [41]. We performed CKI kinase assay on peptides containing novel motifs that have been spotted on peptide microarrays. Peptide microarrays were exposed to a phosphor imager screen and the image was processed to obtain intensity values that are proportional to the phosphorylation of the peptides (Figure 7A, 7B). The intensity values obtained were then transformed by applying a log base 2 transform and spatial or positional variations were normalized. The distribution of intensities of the phosphorylation of the peptides on the microarrays was plotted using kernel density estimation (Figure 7C). We looked for novel motifs in peptides showing a high degree of phosphorylation above threshold (Figure 4C) in all three replicates. Table 4 lists all the peptides that were phosphorylated by CKI. Some of the predicted motifs containing peptides were significantly phosphorylated while many motifs were not phosphorylated at all (Figure 7D). Motifs containing acidic residue clusters on N-or C-terminus of the phosphoserine/threonine were more commonly phosphorylated by CKI. Interestingly, we found that a number of motifs containing a proline at + 1 position were phosphorylated by CKI. This is significant, because even though the occurrence of a proline residue in + 1 position followed by basic residues in + 3 position has been described as a phosphorylation substrate motif for CDK family of kinases Marin et al. [42], this motif has not been reported for casein kinases.

Because the peptides that were spotted on peptide microarrays were derived from proteins that were not known to be phosphorylated, we wanted to know if any particular class of proteins from whom the phosphorylated peptides were derived was enriched using the "Molecule class" in HPRD. We found 57% proteins were DNA binding proteins. In particular, the majority of the pS[E/D]X[E/D][E/D] motif-containing peptide (78%) and [pT/pS]PXXXP motif-containing peptides (73%) that were phosphorylated were from DNA binding proteins or transcription factors. Nearly half of the peptides containing the pS[D/E]S[D/E] motif that were phosphorylated were from membrane proteins or transcription factors. Additional experiments need to be carried out to confirm that the proteins from which these peptides are derived are potential substrates of CKI *in vivo*.

It is important to point out that we did not observe phosphorylation

of all peptides containing any particular motif. For instance, one of the peptides from Methyl CpG binding domain protein 1 (RLLPSVWSESEDGAG) contains the same motif (pS[D/E]S[D/E]) as a peptide from Histone deacetylase 1 (IACEEEFSDSEEEGE) although only the latter peptide was phosphorylated by CKI. This could occur because of various reasons: i) The structure of one peptide is favorable while that of a related peptide might not be so; ii) Although different peptides might contain the same motif, other amino acid residues that are not shared between peptides still play a role in determining the extent of phosphorylation; iii) We have used stringent criteria to label a peptide as 'phosphorylated,' thus it is possible that some of the

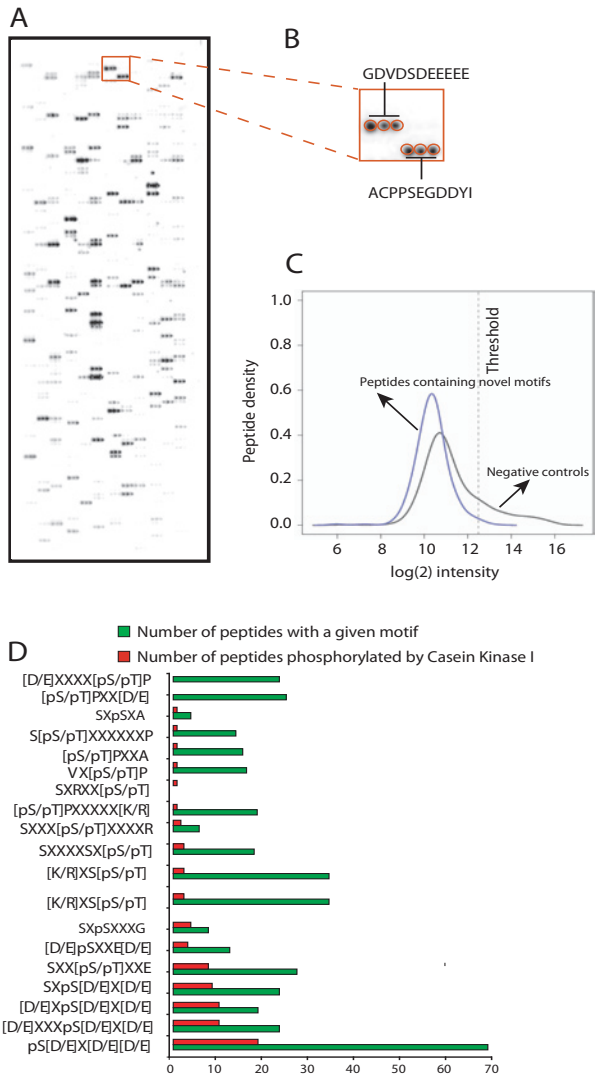

Figure 7: Peptide microarrays for validation of novel motifs (A) Casein Kinase I (CKI) assay was performed on custom peptide microarrays spotted with peptides from sequences not known to be phosphorylated but containing the predicted motifs. (B) Magnification of a section of the peptide microarray showing two peptides that were significantly phosphorylated. The peptides were spotted in triplicate and are indicated by circles. (C) A kernel distribution curve showing the intensity values calculated from CKI assays from exposure of peptide microarrays to a phosphorimager screen. The blue line represents intensity values of negative controls and the black line represents intensity values of peptides containing novel motifs (D) A bar graph showing the number of peptides containing novel predicted motifs that were spotted on the peptide microarrays and the number of peptides containing novel motifs that were phosphorylated by CKI.

	Peptide	Protein
1	LSAGSDSEEGL	Aristaless related homeobox, X-linked
2	LFPPSPQPPDS	Atrophin 1
3	MSGRSPLLPRE	C terminal binding protein 2
4	VAPPTPLTPTS	Casein kinase 1, delta
5	LKRYSEMDDHY	Centromeric protein E
6	SPLSSPVFPRA	Desmin
7	PGPPSPTPPAP	Dopamine receptor D4
8	EVDESDVEEDI	Double strand break repair protein MRE11A
9	SGRSSAKKKGG	Dual specificity phosphatase 11
10	ELYESARIGGA	Dual specificity phosphatase 9
11	LYLGSARDSAN	Dual specificity phosphatase 9
12	WSTLSPIAPRS	ELK 1
13	GRRSSPAQTRE	FAS associated factor 1
14	RRRSSPSKNRG	Fibroblast Growth Factor 14
15	GPPPSPGKGGG	Forkhead box protein L2
16	LKLSSDEDDLE	Fragile X mental retardation 2
17	GFRSSPKTPPR	GAB1
18	SRRSSLASPFP	GLI1
19	TPPGTPRPPDT	Grb7
20	WKTMSAKEKSK	High mobility group box 2
21	EEEFSDSEEEG	Histone deacetylase 1
22	EEEFSDSEEEG	Histone deacetylase 1
23	DEEFSDSEDEG	Histone deacetylase 2
24	VKAESPEKPER	Homeo box 7
25	ASASSEPEEAA	HOX B5
26	PKAISEEEEEV	Huntingtin
27	SAESSDSEGFV	IGF-I receptor
28	PRRSSSDEQGL	IL1 receptor accessory protein
29	PDPETPLKARK	INCENP
30	IQGDSESEAHL	Integrin beta 4
31	GVPDTPTRLVF	Integrin beta 4
32	ERRPSQGAAGS	Interleukin 3 receptor, beta
33	VGGTSENDDPS	Katanin p60 subunit A1
34	DSVKSEDEDGD	LIM homeobox transcription factor 1, beta
35	LDLNSDSDPYP	LDL receptor related protein 5
36	GDPGSPPPARG	MAP3K12
37	SHPWTPDDSTD	MAP3K7
38	SDGESDEEEFQ	MLL4
39	PPAPSPPPPLP	MLL4
40	PPLPSPPPPPA	MLL4
41	SRRPSPLAPRP	MLL4
42	SYDGSDREDHR	Myocyte specific enhancer factor 2C
43	KRRSSSNTKAV	N-myc
44	VGYSSDSETLD	Nuclear factor erythroid 2 like 1
45	GQEHSDSEKAS	Pituitary homeobox 3
46	TAEISARTMQS	Protein tyrosine phosphatase, receptor type, delta
47	VFRLSARNKVG	Protein tyrosine phosphatase, receptor type, delta
48	QAPSSPPRRVQ	Protein tyrosine phosphatase, receptor type, F
49	SSGGSDSDESV	RAS related associated with diabetes
50	APWDSAKKDEN	Receptor protein tyrosine phosphatase mu

51	NPEESESESEG	Retinal degeneration slow protein
52	SRRPSDSGPPA	Rhotekin
53	GASFSDSEDES	Ron
54	GKFSSDSDIWS	ROR1 receptor tyrosine kinase
55	VINESFEGEDG	ROS1
56	SLPSSPLRLGS	Serine/threonine protein phosphatase with EF-hands 2
57	TYQVSESDSSG	Serum response factor
58	SPVLSPTLPAE	SMAD4 interacting transcription factor
59	YIESSDSEEIE	SMARCA 3
60	DSEESDSDYEE	SMARCA2
61	VAPRSDSEESG	SMARCA4
62	TGRPSPAPPAV	SMARCA4
63	LLKASEVEEIL	SMARCB1
64	GGSGSDSEPDS	SREBP1
65	SLLASDSEPLK	Thrombopoietin receptor
66	AQALSDSEIQL	TIE1 receptor tyrosine kinase
67	TAYPSAKTPSS	Transcription factor 3
68	VETNSDSDDED	Transcription factor 8
69	SRRSSFSMEEG	Transcription factor EB
70	DENISDSEIEQ	Transcription factor IIB related factor
71	AMERSETEEKF	Transcriptional coactivator CRSP130
72	QMGVSAKRRPK	Transcriptional co-activator CRSP34
73	DLSTSDEDDLY	Trithorax like protein
74	ELLKSDSDNNN	Trithorax like protein
75	SVKTSPRKPRG	Trithorax like protein
76	QTSSSPPPPLL	Trithorax like protein
77	KRRSSRRSAGG	TWIST
78	QRHGSDSEYTE	VEGF receptor 3
79	KGKQSPPGPGK	Zinc finger protein 217
80	RDYLSDSELNQ	Zinc finger protein 231
81	EDDWSDWEEHP	Zinc finger protein 277
82	GLDGSEEEEKG	Zinc finger protein 35
83	KDEYSERDENV	Zinc finger protein, subfamily 1A, member 3

Table 4: A list of peptides spotted on peptide microarrays that were phosphorylated by Casein Kinase I.

peptides are weakly phosphorylated and thus ignored in our analysis, iv) certain amino acids may negatively affect kinase binding e.g. incompatible charges, v) the peptide may not be properly displayed in the experiment. Overall, these results not only confirm the validity of the new motifs identified but also redefine the CKI substrate specificity that can be investigated further.

Discussion

Our analysis further extends the study of kinase-substrate interactions in humans by deriving a catalog of several novel phosphorylation motifs. The bioinformatics approach presented here highlights the importance of refining our biological predictions when larger datasets become available. The current knowledge about the known phosphorylation motifs along with novel ones could lead us to better understand kinase-substrate interactions and kinase specificities. The list of all the literature derived known phosphorylation motifs published previously by our group is available in "PhosphoMotif Finder" as part of Human Protein reference database (HPRD), where it is possible to look for a specific phosphorylation motif in a given sequence and also know whether such a motif binds to a kinase, a phosphatase or an interaction domain. The significance of this study is that the novel motifs identified were predicted based on experimentally determined phosphorylation sites in the literature. Peptide microarray technology is a high-throughput platform, which could be used for such studies using a single kinase at a time. This strategy could be extended to identification of phosphorylation sites and kinase specificity *in vitro*. Any potential peptide substrate that is positive in these assays could be subsequently tested *in vivo*. Because of the extensive crosstalk in signaling pathways and the overlap in the specificity of kinases, a systematic identification of motifs and assignment of each motif to one or more kinases should permit a better understanding of how signals are transmitted within cells.

Acknowledgements

We thank the Department of Biotechnology of the Government of India for research support to the Institute of Bioinformatics, Bangalore. Akhilesh Pandey is supported by grants from the National Institutes of Health (CA106424 and U54 RR020839), National Heart, Lung, and Blood Institute, National Institutes of Health, under contract number HV-28180 and a Department of Defense Era of Hope Scholar award (W81XWH-06-1-0428).

Conflict of Interest

Jos Joore is VP of Array Technology at PepScan Systems.

References

1. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. Science 298: 1912-1934.
2. Yaffe MB (2002) Phosphotyrosine-binding domains in signal transduction. Nat Rev Mol Cell Biol 3: 177-186.
3. Tzivion G, Shen YH, Zhu J (2001) 14-3-3 proteins; bringing new definitions to scaffolding. Oncogene 20: 6331-6338.
4. Yaffe MB, Elia AE (2001) Phosphoserine/threonine-binding domains. Curr Opin Cell Biol 13: 131-138.
5. Yaffe MB, Smerdon SJ (2001) PhosphoSerine/threonine binding domains: you can't pSERious? Structure 9: R33-38.
6. Schwartz D, Gygi SP (2005) An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. Nat Biotechnol 23: 1391-1398.
7. Blom N, Gammeltoft S, Brunak S (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 294: 1351-1362.
8. Patschinsky T, Hunter T, Esch FS, Cooper JA, Sefton BM (1982) Analysis of the sequence of amino acids surrounding sites of tyrosine phosphorylation. Proc Natl Acad Sci U S A 79: 973-977.
9. Songyang Z, Cantley LC (1995) Recognition and specificity in protein tyrosine kinase-mediated signalling. Trends Biochem Sci 20: 470-475.
10. Songyang Z, Lu KP, Kwon YT, Tsai LH, Filhol O, et al. (1996) A structural basis for substrate specificities of protein Ser/Thr kinases: primary sequence preference of casein kinases I and II, NIMA, phosphorylase kinase, calmodulin-dependent kinase II, CDK5, and Erk1. Mol Cell Biol 16: 6486-6493.
11. Miller ML, Hanke S, Hinsby AM, Friis C, Brunak S, et al. (2008) Motif decomposition of the phosphotyrosine proteome reveals a new N-terminal binding motif for SHIP2. Mol Cell Proteomics 7: 181-192.
12. Ritz A, Shakhnarovich G, Salomon AR, Raphael BJ (2009) Discovery of phosphorylation motif mixtures in phosphoproteomics data. Bioinformatics 25: 14-21.
13. Schwartz D, Chou MF, Church GM (2009) Predicting protein post-translational modifications using meta-analysis of proteome scale data sets. Mol Cell Proteomics 8: 365-379.
14. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, et al. (2009) Human Protein Reference Database--2009 update. Nucleic Acids Res 37: D767-772.

15. Beausoleil SA, Jedrychowski M, Schwartz D, Elias JE, Villen J, et al. (2004) Large-scale characterization of HeLa cell nuclear phosphoproteins. Proc Natl Acad Sci U S A 101: 12130-12135.

16. Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol 24: 1285-1292.

17. Kim JE, Tannenbaum SR, White FM (2005) Global phosphoproteome of HT-29 human colon adenocarcinoma cells. J Proteome Res 4: 1339-1346.

18. Nousiainen M, Sillje HH, Sauer G, Nigg EA, Korner R (2006) Phosphoproteome analysis of the human mitotic spindle. Proc Natl Acad Sci U S A 103: 5391-5396.

19. Swaney DL, Wenger CD, Thomson JA, Coon JJ (2009) Human embryonic stem cell phosphoproteome revealed by electron transfer dissociation tandem mass spectrometry. Proc Natl Acad Sci U S A 106: 995-1000.

20. Molina H, Horn DM, Tang N, Mathivanan S, Pandey A (2007) Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry. Proc Natl Acad Sci U S A 104: 2199-2204.

21. Amanchy R, Periaswamy B, Mathivanan S, Reddy R, Tattikota SG, et al. (2007) A curated compendium of phosphorylation motifs. Nat Biotechnol 25: 285-286.

22. Carrascal M, Ovelleiro D, Casas V, Gay M, Abian J (2008) Phosphorylation analysis of primary human T lymphocytes using sequential IMAC and titanium oxide enrichment. J Proteome Res 7: 5167-5176.

23. Endicott JA, Noble ME, Tucker JA (1999) Cyclin-dependent kinases: inhibition and substrate recognition. Curr Opin Struct Biol 9: 738-744.

24. Joseph CK, Byun HS, Bittman R, Kolesnick RN (1993) Substrate recognition by ceramide-activated protein kinase. Evidence that kinase activity is proline-directed. J Biol Chem 268: 20002-20006.

25. Diks SH, Kok K, O'Toole T, Hommes DW, van Dijken P, et al. (2004) Kinome profiling for studying lipopolysaccharide signal transduction in human peripheral blood mononuclear cells. J Biol Chem 279: 49206-49213.

26. Cleveland WS (1995) Loader CR: Smoothing by local regression: Principles and methods. NY: Murray Hill.

27. Al-Obeidi FA, Wu JJ, Lam KS (1998) Protein tyrosine kinases: structure, substrate specificity, and drug discovery. Biopolymers 47: 197-223.

28. Fukunaga R, Hunter T (1997) MNK1, a new MAP kinase-activated protein kinase, isolated by a novel expression screening method for identifying protein kinase substrates. Embo J 16: 1921-1933.

29. Valtorta F, Schiebler W, Jahn R, Ceccarelli B, Greengard P (1986) A solid-phase assay for the phosphorylation of proteins blotted on nitrocellulose membrane filters. Anal Biochem 158: 130-137.

30. Songyang Z, Shoelson SE, McGlade J, Olivier P, Pawson T, et al. (1994) Specific motifs recognized by the SH2 domains of Csk, 3BP2, fps/fes, GRB-2, HCP, SHC, Syk, and Vav. Mol Cell Biol 14: 2777-2785.

31. Salomon AR, Ficarro SB, Brill LM, Brinker A, Phung QT, et al. (2003) Profiling of tyrosine phosphorylation pathways in human cells using mass spectrometry. Proc Natl Acad Sci U S A 100: 443-448.

32. Zhang Y, Wolf-Yadlin A, Ross PL, Pappin DJ, Rush J, et al. (2005) Time-resolved mass spectrometry of tyrosine phosphorylation sites in the epidermal growth factor receptor signaling network reveals dynamic modules. Mol Cell Proteomics 4: 1240-1250.

33. Li L, Shakhnovich EI, Mirny LA (2003) Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases. Proc Natl Acad Sci U S A 100: 4463-4468.

34. Yaffe MB, Leparc GG, Lai J, Obata T, Volinia S, et al. (2001) A motif-based profile scanning approach for genome-wide prediction of signaling pathways. Nat Biotechnol 19: 348-353.

35. Blatch GL, Lassle M, Zetter BR, Kundra V (1997) Isolation of a mouse cDNA encoding mSTI1, a stress-inducible protein containing the TPR motif. Gene 194: 277-282.

36. Marin O, Bustos VH, Cesaro L, Meggio F, Pagano MA, et al. (2003) A noncanonical sequence phosphorylated by casein kinase 1 in beta-catenin may play a role in casein kinase 1 targeting of important signaling proteins. Proc Natl Acad Sci U S A 100: 10193-10200.

37. Meggio F, Perich JW, Reynolds EC, Pinna LA (1991) A synthetic beta-casein phosphopeptide and analogues as model substrates for casein kinase-1, a ubiquitous, phosphate directed protein kinase. FEBS Lett 283: 303-306.

38. Adams JC (2002) Characterization of a Drosophila melanogaster orthologue of muskelin. Gene 297: 69-78.

39. Pearson RB, Kemp BE (1991) Protein kinase phosphorylation site sequences and consensus specificity motifs: tabulations. Methods Enzymol 200: 62-81.

40. Amanchy R, Zhong J, Molina H, Chaerkady R, Iwahori A, et al. (2008) Identification of c-Src tyrosine kinase substrates using mass spectrometry and peptide microarrays. J Proteome Res 7: 3900-3910.

41. Flotow H, Graves PR, Wang AQ, Fiol CJ, Roeske RW, et al. (1990) Phosphate groups as substrate determinants for casein kinase I action. J Biol Chem 265: 14264-14269.

42. Marin O, Meggio F, Sarno S, Andretta M, Pinna LA (1994) Phosphorylation of synthetic fragments of inhibitor-2 of protein phosphatase-1 by casein kinase-1 and -2. Evidence that phosphorylated residues are not strictly required for efficient targeting by casein kinase-1. Eur J Biochem 223: 647-653.

43. Amanchy R, Kalume DE, Iwahori A, Zhong J, Pandey A (2005) Phosphoproteome analysis of HeLa cells using stable isotope labeling with amino acids in cell culture (SILAC). J Proteome Res 4: 1661-1671.

44. Songyang Z, Blechner S, Hoagland N, Hoekstra MF, Piwnica-Worms H, et al. (1994) Use of an oriented peptide library to determine the optimal substrates of protein kinases. Curr Biol 4: 973-982.

45. Songyang Z, Carraway KL 3rd, Eck MJ, Harrison SC, Feldman RA, et al. (1995) Catalytic specificity of protein-tyrosine kinases is critical for selective signalling. Nature 373: 536-539.

46. Songyang Z, Shoelson SE, Chaudhuri M, Gish G, Pawson T, et al. (1993) SH2 domains recognize specific phosphopeptide sequences. Cell 72: 767-778.