BMC Bioinformatics

Software

PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory

Yu Xue†[1], Ao Li†[2], Lirong Wang[2], Huanqing Feng[2] and Xuebiao Yao*[1,3]

Address: [1]School of Life Science, University of Science and Technology of China, Hefei, Anhui, 230027, China, [2]Department of Electronic Science and Technology, University of Science and Technology of China, Hefei, Anhui, 230027, China and [3]Department of Physiology, Morehouse School of Medicine, Atlanta, GA 30310, USA

Email: Yu Xue - yxue@mail.ustc.edu.cn; Ao Li - liao@mail.ustc.edu.cn; Lirong Wang - lirongw@ustc.edu; Huanqing Feng - hqfeng@ustc.edu.cn; Xuebiao Yao* - yaoxb@ustc.edu.cn

* Corresponding author †Equal contributors

Published: 20 March 2006

*BMC Bioinformatics*2006, **7**:163 doi:10.1186/1471-2105-7-163

Received: 21 August 2005
Accepted: 20 March 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/163

Abstract

Background: As a reversible and dynamic post-translational modification (PTM) of proteins, phosphorylation plays essential regulatory roles in a broad spectrum of the biological processes. Although many studies have been contributed on the molecular mechanism of phosphorylation dynamics, the intrinsic feature of substrates specificity is still elusive and remains to be delineated.

Results: In this work, we present a novel, versatile and comprehensive program, PPSP (Prediction of PK-specific Phosphorylation site), deployed with approach of Bayesian decision theory (BDT). PPSP could predict the potential phosphorylation sites accurately for ~70 PK (Protein Kinase) groups. Compared with four existing tools Scansite, NetPhosK, KinasePhos and GPS, PPSP is more accurate and powerful than these tools. Moreover, PPSP also provides the prediction for many novel PKs, say, TRK, mTOR, SyK and MET/RON, etc. The accuracy of these novel PKs are also satisfying.

Conclusion: Taken together, we propose that PPSP could be a potentially powerful tool for the experimentalists who are focusing on phosphorylation substrates with their PK-specific sites identification. Moreover, the BDT strategy could also be a ubiquitous approach for PTMs, such as sumoylation and ubiquitination, etc.

Background

Protein phosphorylation, as one of the most common post-translational modifications (PTM), is reversibly and transiently performed by protein kinases (PKs). It plays crucial regulatory roles in a variety of biological cellular processes, including transcription [1], translation [2], mitosis/cell cycle [3], neurite outgrowth [4,5] and signal transductions [6], etc. Many previous researches have contributed to increase our knowledge on phosphorylation. However, the intrinsic features of phosphorylation dynamics are still cryptic and remain to be dissected. Biochemically, the catalytic site of a PK hydrolyzes adenosine triphosphate (ATP) and transfers a phosphate moiety to the acceptor residue (S/T, Y in eukaryotes) in the substrate. Each PK only modifies a defined subset of substrates specifically to ensure signaling fidelity, and defects of PK function often induce a variety of diseases and cancers [7].

There is an extensively-adopted hypothesis that PKs phosphorylate their substrates at the specific sites (consensus sequence) flanking with canonical motif [8-10]. To date, the consensus motifs of ~30 PKs have been reported [11]. However, there is still a large number of PKs with their specific target motifs remained to be identified. Therefore, elucidating PK-specific phosphorylation sites on the substrates is the foundation of understanding the molecular mechanism of substrates specificity and important for the biomedical drug design. However, it has been described that only consensus motif is not enough for providing the specificity of PK recognition *in vivo* [12]. There are numerous mechanisms have been proposed to contribute specificity for PKs, such as co-complex of PKs with their substrates, subcellular co-localization, interacting through modular docking sites, phosphopeptide-binding mechanisms, etc [12-17]. In a cell, protein kinase usually forms a tight complex with its target either through a third scaffold protein, or by recognizing and binding short sequence of the substrate, known as a docking site [12,18]. Moreover, phosphopeptide-binding domains (PBDs) are also important to achieve substrate specificity. Numerous PBDs (PTB, WW, SH2, SH3, FHA, MH2, WD40, Polo-box, and 14-3-3, etc) bind the phosphorylated forms of specific proteins, with recognizing distinct peptides surrounding the phosphorylated sites (pS/T, or pY) [14-17,19]. However, how these mechanisms achieve the additional specificity for PKs beyond phosphorylated motifs is still elusive, and there are very few computational studies published on this area [13,16,19]. In addition, many docking sites and PBDs still remain to be dissected. Thus, in this work, we focus on the prediction of PK-specific phosphorylation sites based on profiles/features of the surrounding primary sequences, as previously described [8-10].

Conventional experimental identifications of PK-specific phosphorylation sites on substrates *in vivo* and *in vitro* have provided the foundation of understanding the mechanisms of phosphorylation dynamics. However, these experiments are often time-consuming and expensive. And the enzymatic activity of the PKs are usually diminished or impeded *in vitro*, hampering on the studies of phosphorylation greatly. Recently, phospho-proteomic studies with mass spectrometry (MS) approaches have generated numerous data in yeast [20], mouse [21], and human [8], etc. But in these cases, it's still difficult to distinguish the PK-specific sites on the substrates. With regard of this, it is of note that the *in silico* prediction of PK-specific phosphorylation sites is in urgent need for the further experimental manipulation. To address this question, several excellent predictors have been implemented and reported [13,22-25]. For example, NetPhos has employed the consensus-motif-based approaches implemented in the artificial neural networks (ANNs) algorithm [22]. The enhanced version, NetPhosK can predict PK-specific phosphorylation sites for ~17 PKs [23]. Another online tool Scansite [13] has constructed the motif profiles of phosphorylation sites for ~20 PKs, and could predict their target sites, respectively. Previously, we have reported a web server GPS, which has been implemented in GPS (Group-based phosphorylation Predicting and Scoring) algorithm [26,27]. GPS could predict ~70 kinds of PK-specific phosphorylation sites, and gain excellent performance on several PK groups, especially for kinase Aurora-B. Recently, a novel and excellent web tool of KinasePhos has been incorporated with HMM (Hidden Markov Models) algorithm and constructed for phosphorylation sites predicting of 18 PK-specific groups [24,25].

In this study, we present a novel, convenient and comprehensive program, PPSP (Prediction of PK-specific Phosphorylation site), implemented in an algorithm of Bayesian decision theory (BDT). An online PPSP web service has been also constructed, accurately predicting PK-specific phosphorylation sites for 68 PK groups. The prediction performances of PPSP are satisfactory on several well-studied PKs and comparable with the other existing tools NetPhosK, Scansite, KinasePhos and GPS. Moreover, PPSP also provides the accurate prediction for many novel PKs, such as TRK, mTOR, SyK, and MET/RON, etc. Obviously, PPSP is more accurate and powerful. Therefore, we propose that PPSP could be useful and insightful for further experimental design. In addition, the prediction results of PPSP combined with delicate experiments verifications will propel our understanding of the mechanisms of phosphorylation into a new phase.

Implementation

Preparation of training data set

Firstly, we obtained the data set of phosphorylation sites from Phospho.ELM (Ver 2.0, Sep. 2004) [28] and filtered the phosphorylation sites without information of PKs. There were ~1,400 sites preserved. We also manually curated the recent literature and acquired ~660 items (Before Nov. 2004). These newly curated data has been submitted to Phospho.ELM for further integration. The two data sets were integrated, and the redundant items were removed if two items exactly pinned point to the same phosphorylation site from one protein sequence. Then the total training data set contained >2,000 non-redundant positive data with very few homologous sites (see additional file 1).

Since there were several PKs with too few known phosphorylation sites, we clustered them into distinct subgroups based on sequence homology. For example, eight ribosomal protein S6 kinases (RSK1, Q15418; RSK3, Q15349; RSK2, P51812; MSK1, O75676; MSK1, O75582; RSK4, Q9UK32; S6K1, P23443; STK14B, Q9UBS0) are

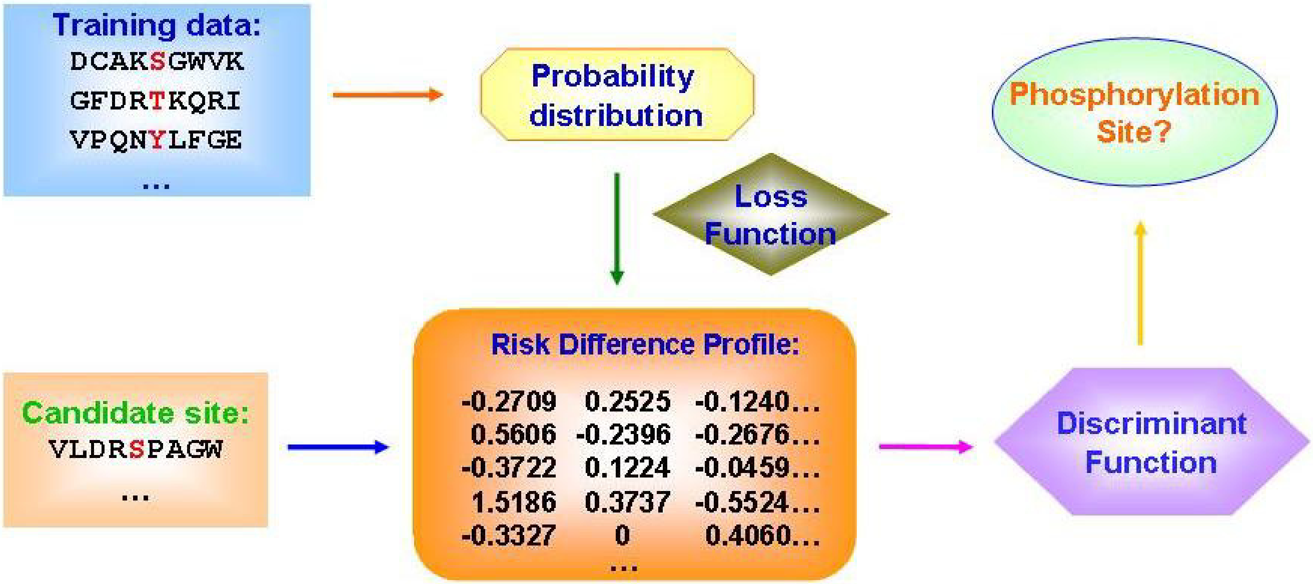


Figure 1
The outline of the training and procedure of PPSP.

homologous with high similarity, so we clustered these PKs into a unique PK group of S6K (Ribosomal protein S6 kinase, or RSK). In total, we have enabled 68 PK grouped.

Although Swiss-Prot also curates a huge amount of phosphorylation sites, we have found ~69% of the annotation to be ambiguous (7,924 of 11,520) (see additional file 2). There are only 842 items to be kinase-specific sites, and only 18 PKs with not less than ten sites (see additional file 3). Phospho.ELM has been constructed based on the rationale of allowing both experimentalists and bioinformatists to easily access extensive information of phosphoproteins with their sites, i.e., tracking the primary reference to find whether the site is really phosphorylated, identified *in vivo* or *in vitro*, and the relationship between the phosphorylation with physiological response [28]. And these data has been collected from literature manually with high quality. Taken together, although other resources also have collected some phosphorylation sites, we chose Phospho.ELM for its comprehensiveness.

Positive & negative control for evaluation

The sequence information of these phosphorylation substrates was retrieved from ExPASy. As previously described [11], we adopted the experimental phosphorylation sites as the positive control, while all other residues (S/T or Y) in the phosphorylation substrates were regarded as the negative control. The detailed statistics of the positive and negative data sets categorized by PK groups is available (see additional file 4).

Bayesian Decision Theory (BDT)

Supposed that we have an unclassified data x that belongs to one of two certain categories: *C1* (defined as phosphorylated sites in this work) and *C2* (defined as non-phosphorylated sites). In addition, suppose the posterior probability of x for these two categories can be denoted as: $p(C_1|x)$ and $p(C_2|x)$. Then the probability of wrong prediction is:

$$P(error \mid x) = \begin{matrix} p(C_1 \mid x), if\ x \in C_2 \\ p(C_2 \mid x), if\ x \in C_1 \end{matrix} \qquad (1)$$

To minimize the expectation of error probability that is defined as [29]:

$$P(error) = \int P(error|x)p(x)dx \qquad (2)$$

It is obvious that one should choose the more probable category as the prediction result, which can be formulated by the Bayesian Decision Rule [29]:

$$predict\ x\ as \begin{cases} C_1, & if\ P(C_1 \mid x) > P(C_2 \mid x) \\ C_2, & otherwise \end{cases} \qquad (3)$$

Furthermore, by definition we can introduce the loss function $\lambda(\alpha_i|C_j)$, where $\alpha_i, i = 1,2$ is the finite set of possible solution. Thus the expected loss (risk) of taking action α_i is:

Table 1: The performances of self-consistency, Jack-knife validation and *n* -fold (4-, 6-, 8-, 10-fold in this work) cross-validation for four well-studied PKs of PKA, CK2, ATM and S6K. The *n*- fold cross-validation has been performed for the large data sets (N ≥ 30).

PPSP		PKA		CK2		ATM		S6K	
		Sn(%)	*Sp*(%)	*Sn*(%)	*Sp*(%)	*Sn*(%)	*Sp*(%)	*Sn*(%)	*Sp*(%)
Self-consistency		90.11	91.70	83.21	90.01	93.02	94.06	92.85	97.97
Jack-knife		90.11	90.46	83.21	88.44	86.05	91.89	92.86	91.05
	4-	90.11	90.43	81.02	87.90	86.37	90.14	N/A	N/A
***n*-fold cross-validation**	6-	90.11	90.52	81.02	88.34	86.37	90.60	N/A	N/A
	8-	90.11	90.45	81.75	88.48	86.05	90.65	N/A	N/A
	10-	90.11	90.48	81.75	88.22	86.05	91.39	N/A	N/A
Data set (No.)	**Positive**	173		142		43		14	
	Negative	8, 408		5, 332		2, 048		683	

$$R(\alpha_i \mid x) = \sum_{l=1}^{2} \lambda(\alpha_i \mid C_l) P(C_l \mid x) \qquad (4)$$

In this condition, the goal of optimization becomes to minimize the overall risk for every x. Similar to the rationale of Bayesian Decision Rule, we can obtain the best performance by computing $R(\alpha_i|x)$ for each solution α_i and choose that for which has the minimal overall risk (also named as Bayes Risk) [29].

Training and prediction procedure

In this study, a local ennea-peptide (9aa) is deployed to define a candidate phosphorylation site, which has 4 upstream and 4 downstream residues of the potential phosphorylation site and can be denoted as $\vec{x}$ = $(x_1,x_2,...,x_9)'$. Given some positive (training) data, there are many ways to estimate $R(\alpha_i|x)$ (where α_1 and α_2 denote different prediction results: true and false phosphorylation sites, respectively). One simple way is to assume that all flanking residues are mutual independent, and then the Bayes Risk can be formulated as:

$$R(\alpha_i \mid \vec{x}) = \sum_{j=1}^{9} R(\alpha_i \mid x_j) \qquad (5)$$

$$R(\alpha_i \mid x_j) = E(\lambda \mid x_j, \alpha_i) = \sum_{l=1}^{2} \sum_{k=1}^{20} \lambda(j,k \mid \alpha_i, C_l) p(C_l \mid x_j) \qquad (6)$$

Here $p(C_l|x_j)$ is the posterior probability of x_j belonging to category C_l and can be further described by the Bayesian formula:

$$p(C_l \mid x_j) = \frac{p(x_j \mid C_l)p(C_l)}{p(x_j)} = \frac{p(x_j \mid C_l)p(C_l)}{\sum_{l=1}^{2} p(x_j \mid C_l)p(C_l)}, l = 1,2 \qquad (7)$$

Here $p(C_l)$ is the prior probability of category C_l and $p(x_j|C_l)$ can be estimated by observing the occurrence of each residue in training data given the hypothesis of equation (5). Although there are much more false phosphorylation site in data set, we give equal prior probability for each category (no prior information), which can avoid bias prediction results. The loss function we construct is based on BLOSUM62 matrix [30] by considering the biochemical difference of residues, which can be formulated as:

$$\lambda(j,k \mid \alpha_i, C_l) = \begin{cases} -BLOSUM62(j,k), & if\ \alpha_i \neq C_l \\ 0, & if\ \alpha_i = C_l \end{cases} \qquad (8)$$

Although other matrices could be also adopted, the BLOSUM62 matrix is chosen in this work. Moreover, we introduce a trade-off threshold b as the only parameter in this method to control the performance for different categories. Thus the final Discriminant function for prediction is:

$$predict\ \vec{x}\ as \begin{cases} C_1, & if\ R(\alpha_2 \mid \vec{x}) - R(\alpha_1 \mid \vec{x}) > b \\ C_2, & otherwise \end{cases} \qquad (9)$$

The outline of the training and procedure in this work is illustrated in Figure 1. We first estimate the probability distribution of each residue of the true/false ennea-peptide within the training data. Then the Bayes risk for either potential solution (i.e true or false phosphorylation site) is calculated, respectively. To implement the final differential function in equation (9) effectively, we built a difference profile of Bayesian decision risk for each PK family/group in prediction. In this way, a candidate site for a given protein kinase is assessed in the profile and the outcome for each residue is summed up. If the difference of risks (false prediction minus true prediction) is greater than the threshold b, it will be predicted by PPSP as a negative site that can not be phosphorylated by this PK. Otherwise, PPSP will infer the site is as a potential

Table 2: The self-consistency performance and Jack-knife validation for four novel PKs of TRK, mTOR, SyK and MET/RON.

PPSP		TRK		mTOR		SyK		MET/RON	
		Sn(%)	*Sp*(%)	*Sn*(%)	*Sp*(%)	*Sn*(%)	*Sp*(%)	*Sn*(%)	*Sp*(%)
Self-consistency		92.31	97.40	93.33	96.46	92.59	91.98	89.47	95.90
Jack-knife		84.62	96.10	93.33	91.27	77.79	86.79	73.68	91.80
Data set	**Positive**	13		14		27		19	
(No.)	**Negative**	77		433		251		122	

phosphorylation site. In this work, the threshold for each PK has been optimized automatically.

Results and discussions

Prediction performance of PPSP

Three measurements, i.e., Sensitivity (*Sn*), Specificity (*Sp*), and Mathew correlation coefficient (*MCC*) are widely employed to evaluate the performance of prediction with definitions as below:

$$Sn = \frac{TP}{TP + FN}, \quad Sp = \frac{TN}{TN + FP},$$

and

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

Among the data with positive predictions by PPSP, the real positives are regarded as *true positives* (*TP*), while the others are defined as *false positives* (*FP*). Among the data with negative predictions by PPSP, the real positives are regarded as *false negatives* (*FN*), while the others are defined as *true negatives* (*TN*). The Sensitivity (*Sn*) and Specificity (*Sp*) illustrate the correct prediction ratios of positive and negative data sets respectively. But when the number of positive data and negative data differ too much from each other, the Mathew correlation coefficient (*MCC*) should be calculated to assess the prediction performance. The value of *MCC* ranges from -1 to 1, and bigger *MCC* stands for better prediction performance.

To assess whether PPSP is unbiased and robust for prediction, we adopt the standard method "Jack-Knife" validation. We perform the Jack-Knife validation for these PKs by removing one real site from the training data set at a time and re-calculating the *Sn* &*Sp*, respectively. The final results are the average of the all *Sn* &*Sp* of the Jack-Knife validation. Although "Jack-Knife" validation does make sense when the size of the data set is small (i.e., N < 30), we have also taken an additional test with *n*-fold (4-, 6-, 8- and 10-fold in this work) cross-validation for 22 PK groups with larger positive data set (N ≥ 30). As previously proposed [25], the tests are repeated for 20 times and the *Sn* &*Sp* is computed each time. Then the average Sn & Sp are calculated as the final results (see additional file 5).

In table 1, we list the prediction performances for four most well-studied PKs of PKA (Protein kinase A), CK2 (Casein Kinase II), ATM (Ataxia telangiectasia mutated) and S6K (Ribosomal protein S6 kinase, or RSK). The prediction performances of self-consistency, Jack-knife validation and *n*-fold cross-validation has been provided. For PKA, CK2, ATM and S6K, the *Sn* &*Sp* of the self-consistency is 92.31% & 97.40%, 93.33% & 96.46%, 92.59% & 91.98%, and 89.47% & 95.90%, while the Jack-Knife validation is 90.11% & 90.46%, 83.21% & 88.44%, 86.05% & 91.89%, 92.86% & 91.05%, respectively. Interestingly, the performances of *n*-fold cross-validation are very simi-

Table 3: With the default cut-off of PPSP, the percentile of the sites predicted to be potential true positive hits is listed. Both random ennea-peptides and data sets from human proteome have been computed, separately.

PK group	Random ennea- peptides			Ennea-peptides from human proteome		
	S	T	Y	S	T	Y
PKA	11.75%	2.20%		14.61%	3.20%	
CK2	9.18%	3.18%		12.60%	5.65%	
ATM	8.42%	1.96%		8.95%	2.13%	
S6K	14.72%	3.71%		14.90%	3.89%	
mTOR	5.95%	7.09%		8.20%	9.14%	
TRK			3.59%			3.94%
Syk			7.00%			9.74%
Met/RON			13.37%			13.77%

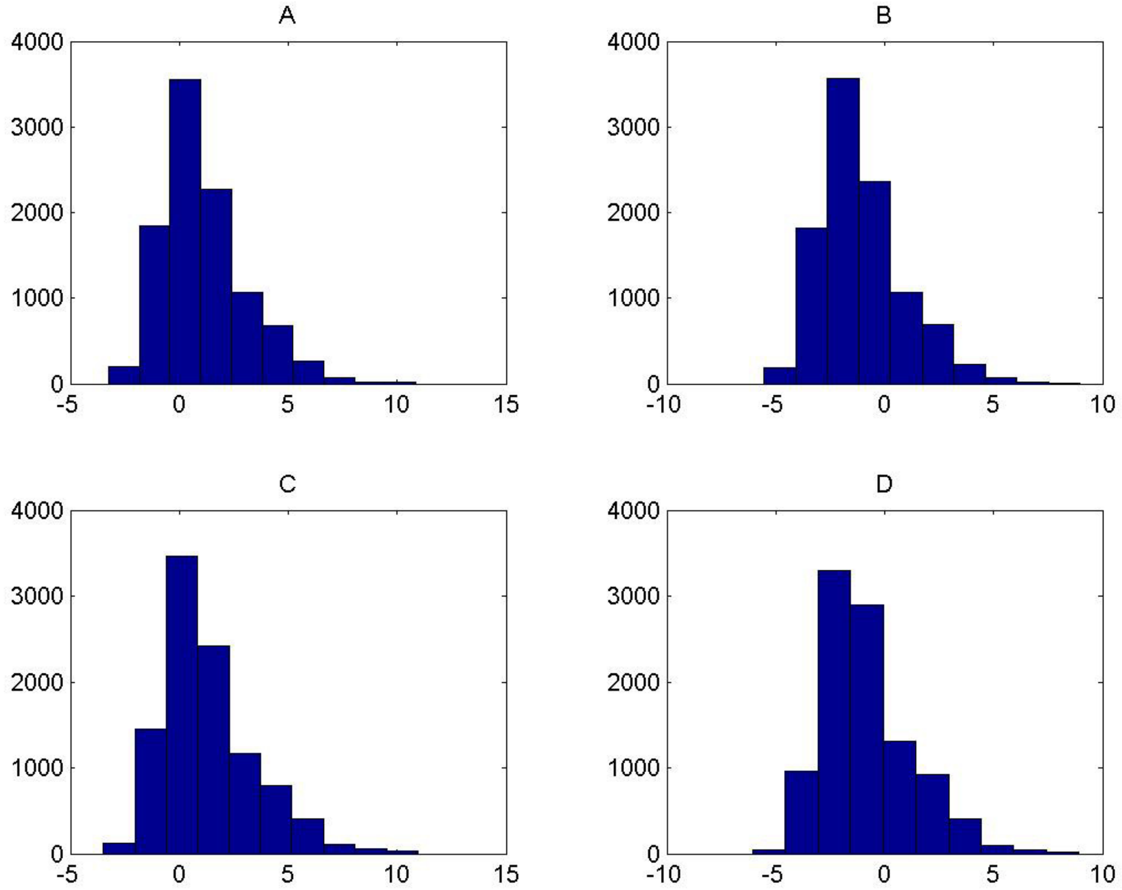


Figure 2
the distribution of risk difference of random and human proteome data set of PKA-specific site prediction is diagramed in Figure 2. A. Distribution of Risk Difference of random data set (serine) of PKA-specific site prediction. B. Distribution of Risk Difference of random data set (threonine) of PKA-specific site prediction. C. Distribution of Risk Difference of human proteome data set (serine) of PKA-specific site prediction. D. Distribution of Risk Difference of Human proteome data set (threonine) of PKA-specific site prediction.

lar and consistent with the results of the Jack-Knife validation. So the PPSP is quite robust and unbiased for these well-studied PKs. Moreover, PPSP could predict for several novel PKs (>30, see additional file 4). In Table 2, we choose four PKs, including TRK (Neurotrophic tyrosine kinase receptor), mTOR (Mammalian target of rapamycin), SyK (Spleen tyrosine kinase), and MET/RON (Hepatocyte growth factor receptor/Macrophage-stimulating protein receptor), which predictors for them are not available previously. Interestingly, the prediction performance of PPSP is also satisfying. And the Jack-knife validation proposes that the PPSP approach is also robust and unbiased for these novel PKs. The full content of the prediction performance of PPSP is available from PPSP website.

To evaluate the performance of PPSP on the signal to noise for phosphorylation sites retrieval, we also perform two additional evaluations. Firstly, we randomly generate 10, 000 serine (S) and threonine (T) ennea-peptides for serine/threonine kinases (STKs), with 10, 000 tyrosine (Y) nona-peptides for tyrosine kinases (TKs). In addition, to determine the ability of the PPSP to retrieve potential real phosphorylation sites from the full proteome, we have downloaded the protein sequences of human proteome from public database ftp://ftp.ebi.ac.uk/pub/databases/IPI/current/ipi.HUMAN.fasta.gz. Again, we randomly retrieve 10, 000 S & T and Y ennea-peptides for STKs and TKs from the human proteome, respectively. Then we compute the Risk Difference (RD) of each ennea-peptide.

Table 4: The prediction performance of Scansite, NetPhosK, KinasePhos and GPS for four well-studied PKs of PKA, CK2, ATM and S6K.

	PK Group	PKA			CK2			ATM			S6K		
Predictor		*Sn*(%)	*Sp*(%)	*MCC*	*Sn*(%)	*Sp*(%)	*MCC*	*Sn*(%)	*Sp*(%)	*MCC*	*Sn*(%)	*Sp*(%)	*MCC*
PPSP	Default[a]	90.11	91.7	0.3841	83.21	90.01	0.3596	93.02	94.06	0.4627	92.85	97.97	0.6618
ScanSite	High[b]	21.98	99.96	0.4450	10.95	99.86	0.2655	18.6	99.8	0.3443	N/A	N/A	N/A
	Medium	44.51	99.39	0.5084	27.01	99.11	0.3342	25.58	98.89	0.2756	N/A	N/A	N/A
	Low	47.8	98.29	0.4041	54.02	96.34	0.3684	51.16	94.89	0.2739	N/A	N/A	N/A
NetPhosK	Default	79.12	90.65	0.3165	82.48	89.43	0.3464	86.01	98.51	0.6786	82.35	97.14	0.5404
KinasePhos	90% (Sp)[d]	90.72	91.3	0.3783	72.53	91.58	0.3384	88.37	87.8	0.3137	N/A	N/A	N/A
	95% (Sp)	89.18	94.62	0.4595	64.58	94.93	0.3806	88.37	92.14	0.3893	N/A	N/A	N/A
	100% (Sp)	76.8	98.47	0.6154	54.86	98.66	0.5222	86.05	96.89	0.5497	N/A	N/A	N/A
GPS	Default	88.88	90.57	0.3564	82.99	87.59	0.3210	90.86	89.55	0.3498	94.9	91.34	0.3964

a. The default parameters are employed for PPSP, NetPhosK and GPS.
b. ScanSite 2.0 has three thresholds for prediction, including high, medium and low stringencies.
c. N/A – not available.
d. KinasePhos has paid attention to prediction specificity with three cut-off values, as 90%, 95% and 100%.

Under the default threshold of PPSP, the percentile of the sites predicted to be potential true positive hits is listed (see in Table 3). The prediction results of random and human proteome data set are very similar. And the distribution of Risk Difference of random and human proteome data set of PKA-specific site prediction is diagramed in Figure 2. In this work, the default threshold of PKA is 3.5, and predicted Risk Differences of the most of the ennea-peptides from the two data sets are smaller than this cut-off. Based on these analyses, we propose that PPSP could efficiently predict the potential real sites with very low false positive hits. The ratio of Serine and Threonine is not exactly equal. However, we and others are unable to explain this question [25].

Comparison of PPSP with Scansite, NetPhosK, KinasePhos and GPS

With four well-studied PKs of PKA, CK2, ATM and S6K as model kinases, we compare PPSP with four previous online prediction systems: Scansite, NetPhosK, KinasePhos and GPS. In Table 4, we list the prediction performances of Scansite, NetPhosK, KinasePhos and GPS for PKA, CK2, ATM and S6K, respectively. Since we can't reperform the Jack-knife validation for the predictors, we submit the substrate sequence into these tools for prediction. And the self-consistency performance of PPSP is adopted here for comparison. Scansite has three thresholds for prediction, including high, medium and low stringency, while KinasePhos has paid attention to prediction specificity with three cut-off values, as 90%, 95% and 100%. And the default parameter is adopted for GPS. We calculate the prediction performances of Scansite and KinasePhos at three distinct thresholds, separately. As for NetPhosK, we only adopt the default cut-off value with 0.5, in mode of Prediction without filtering. Obviously, PPSP, NetPhosK, KinasePhos and GPS are better than Scansite. For PKA, the prediction performance of PPSP is 90.11% (*Sn*) and 91.70% (*Sp*), and outperforms to NetPhosK (*Sn* 79.12% &*Sp* 90.65%) with about 10% higher sensitivity and similar specificity. And for CK2, the performance of PPSP is 83.21% (*Sn*) and 90.01% (*Sp*), slightly higher than NetPhosK (*Sn* 82.48% &*Sp* 89.43%). The prediction performance of KinasePhos is similar with PPSP on PKA and CK2. However, for ATM, the NetPhosK is 86.01% (*Sn*) and 98.51% (*Sp*), whereas PPSP is 93.02% (*Sn*) and 94.06% (*Sp*). Although PPSP has a lower specificity than NetPhosK with ~4%, the sensitivity is high with ~7% enhanced. Finally, for S6K (also called as RSK in NetPhosK), although the specificity of PPSP (97.97%) and NetPhosK (97.14%) is quite similar, PPSP outperforms than NetPhosK with ~10% higher in sensitivity. With regard of this, we propose the prediction performance of PPSP could be at least comparable with the existing systems.

However, the analysis and comparison above are only in theoretical and not intuitive. Furthermore, we browse the recent literature from PubMed and randomly choose some instances for comparison. One example is Bluetongue virus (BTV) nonstructural protein 2 (NS2, P23065), a substrate of CK2 [31]. As a nonspecific single-stranded RNA (ssRNA)-binding protein, NS2 accumulates

Table 5: The experimental verified vs. predicted CK2-specific phosphorylation sites of Bluetongue virus (BTV) nonstructural protein 2 (NS2), Drosophila transcription factor GAGA and human Calmodulin protein.

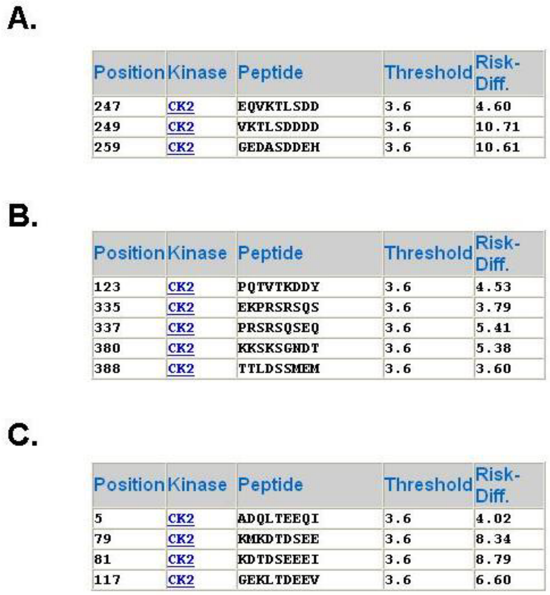
CK2		NS2 (P23065)	GAGA (Q08605)	Calmodulin (P62158)
Experimental Defined		S249, S259	S378, S388	T79, S81, S101, T117
PPSP		T247, **S249**, **S259**	T123, S335, S337, S380, **S388**	T5, **T79**, **S81**, **T117**
NetPhosK		T87, T88, S204, T247, **S249**, **S259**, T266	**S388**, T394	T5, T28, T44, T62, **T79**, **S81**, **S101**, **T117**
ScanSite	high	**S259**	N/A	N/A
	medium	**S249**, **S259**	N/A	**S81**
	low	T88, S182, T247, **S249**, **S259**	T385, T394	**T79**, **S81**, **T117**
KinasePhos	90% (Sp)	T247, **S249**, **S259**	S240, S241, S339, **T378**, S386, S389, S391, S393, S397, S518, S521, S523	T5, T44, **T79**, **S81**, **S101**, **T117**
	95% (Sp)	**S249**, **S259**	S240, S339, **T378**, S386, S391, S397, S518, S521	T5, T44, **T79**, **S81**, **S101**, **T117**
	100% (Sp)	**S259**	S339, S521	T5, **T79**, **S81**, **S101**, **T117**
GPS		**S249**, **S259**	T123, S337, S339, S385, S386, **S388**, S518, S521	T5, T44, **T79**, **S81**, **T117**

in BTV-infected cells, and is functional in viral replication and morphogenesis [31-34]. NS2 could hydrolyze both ATP and GTP with high affinity, showing strong enzymatic activity [32]. Using mutagenesis analysis, CK2 was demonstrated to phosphorylate NS2 in two serine sites S249 and S259, probably modulating its RNA binding properties, enzymatic activity or influencing its ability to interact with other viral proteins [31]. For CK2-specific phosphorylation sites prediction, all of the four programs can detect them successfully (see in Table 5). In this case, the Scansite with medium stringency get the best hits. PPSP predict three sites as positive hits (T247, S249, and S259), but NetPhosK provide too much results with seven positive hits. Two additional instances are also provided in Table 5. One is *Drosophila* transcription factor protein GAGA (Q08605), regulating gene transcription and chromatin remodeling, etc [35]. The other is human Calmodulin protein (P62158) [36]. The prediction results of the four programs are shown in Table 5. And the online prediction of PPSP is diagramed in Figure 3. Obviously, for the well-studied PKs, i.e. CK2, PPSP is accurate and comparable with the existing tools.

Application of PPSP to the novel PKs

For application of PPSP to the novel PKs, here we employ PPSP to predict the phosphorylation sites of TRK. TRK is a sub-family of receptor tyrosine kinases (RTK), consisting three highly similar homologs, TRK-A, -B, and -C [37]. TRK-A, -B, and -C could be activated specifically by nerve growth factor (NGF), brain-derived neurotrophic factor (BDNF) and NT-4/-5, and NT-3, respectively. Under activated state, TRK could regulate a variety of biological processes including cell survival, embryo, differentiation, proliferation, axon and dendrite growth and patterning, and apoptosis, etc [37].

Recently, protein Ras guanine-releasing factor 1 (RasGrf1, Q13972), a GTPase of the Ras and Rho family, has been proposed to be phosphorylated and interact with TRK-A, -B, -C in co-transfection experiments [5]. However, the exact phosphorylation sites of RasGrf1 by TRK remain to be identified. PPSP has predicted that there are totally two potential phosphorylation sites on RasGrf1 (Y94 & Y1209) (see in Figure 4). Moreover, the human tumorous imaginal disc 1 (TID1, Q96EY1) was verified as a substrate



A.

Position	Kinase	Peptide	Threshold	Risk-Diff.
247	CK2	EQVKTLSDD	3.6	4.60
249	CK2	VKTLSDDDD	3.6	10.71
259	CK2	GEDASDDEH	3.6	10.61

B.

Position	Kinase	Peptide	Threshold	Risk-Diff.
123	CK2	PQTVTKDDY	3.6	4.53
335	CK2	EKPRSRSQS	3.6	3.79
337	CK2	PRSRSQSEQ	3.6	5.41
380	CK2	KKSKSGNDT	3.6	5.38
388	CK2	TTLDSSMEM	3.6	3.60

C.

Position	Kinase	Peptide	Threshold	Risk-Diff.
5	CK2	ADQLTEEQI	3.6	4.02
79	CK2	KMKDTDSEE	3.6	8.34
81	CK2	KDTDSEEEI	3.6	8.79
117	CK2	GEKLTDEEV	3.6	6.60

Figure 3
The prediction results of Bluetongue virus (BTV) nonstructural protein 2 (NS2), Drosophila transcription factor GAGA and human Calmodulin protein with PPSP. Figure 3A – prediction results of NS2; Figure 3B – prediction results of GAGA; Figure 3C – prediction results of Calmodulin.

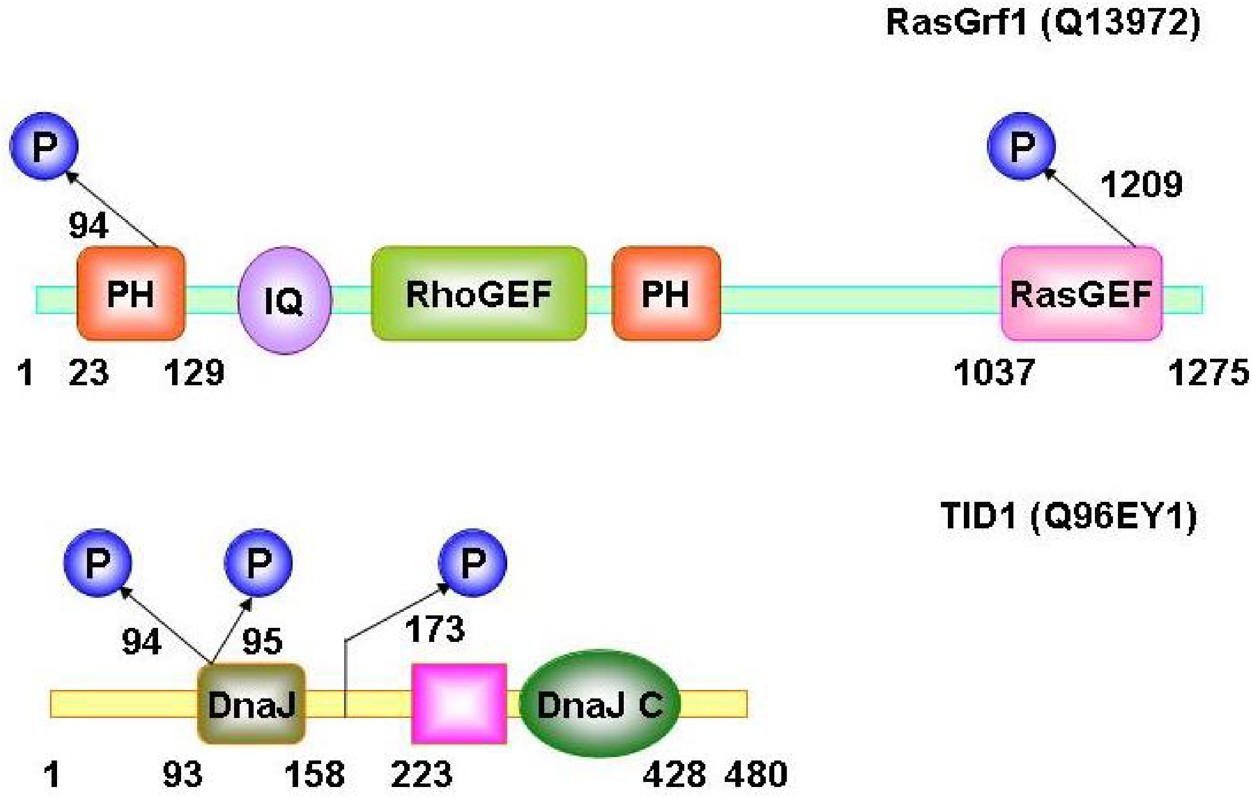


Figure 4
The diagram of potential phosphorylation sites of human RasGrf1 (Q13972) and TID1 (Q96EY1) by TRK.

of TRK with co-immunoprecipitation (Co-IP) [4] and the phosphorylation sites were not elucidated. PPSP could predict that there are three candidate sites with Y94, Y95 and Y173 (see in Figure 4). These prediction results would be very useful for the further experimentation and elucidation phospho-regulation underlying cellular dynamics.

Conclusion

In this work, we present a novel computational program– PPSP (prediction of PK-specific phosphorylation sites) based on Bayesian decision theory (BDT). We model a candidate phosphorylation motif as an ennea/nona-peptide (9aa) flanking with 4 upstream and 4 downstream residues of a potential phosphorylation site (S/T, or Y). With the BDT algorithm, we estimate the probability distributions of true and false phosphorylation sites and make prediction based on a loss function constructed with BLOSUM62 matrix [30]. We have evaluated the sensitivity and specificity of PPSP by "Jack-knife" validation. An online PPSP web service has been also constructed, accurately predicting PK-specific phosphorylation sites for 68 PK groups. For comparison with four reported systems Scansite, NetPhosK, KinasePhos and GPS, we take four well-studied PKs of PKA, CK2, ATM and S6K as model kinases. The prediction performances of PPSP are satisfactory judged using these well-studied PKs and comparable with the other existing tools. Moreover, PPSP also provides the accurate prediction for many novel PKs, such as TRK, mTOR, SyK, and MET/RON, etc. Thus, comparison with the previous work, PPSP provides more accurate and powerful ability. Moreover, the BDT approach could also be an extensive method for PTMs prediction, such as sumoylation and ubiquitination, etc. In addition, although many phospho-proteomic researches have generated numerous data [8,20,21], however, the up-regulated PKs still remain to be dissected. Despite the demonstration of phosphor-regulation of protein kinases and their respective substrates, the exact phosphorylation sites are unclear [4,5]. Taken together, the prediction results of PPSP should be insightful and important for further experiments. The combination of computational and experimental identifications will propel our understanding of phosphorylation dynamics into a new phase.

Availability and requirements

PPSP has been implemented in Linux + Apache + PHP, and is freely available at: http://bioinformatics.lcd-ustc.org/PPSP. A latest web browser (eg. Internet Explorer, Netscape, or Mozilla, etc) is required.

Authors' contributions

YX and AL should be regarded as joint First Authors. YX and AL designed the methodology, carried out the analysis and drafted the manuscript. LW developed the web service, contributed several insightful opinions and improved manuscript greatly. XY coordinated the research and finalized the manuscript.

Additional material

Additional file 1

To test whether the training data sets are highly redundant, we retrieve all protein sequences for each PK-specific substrate. Then we use CD-HIT to check whether many protein sequences are highly homologous. The result of the eight PK groups employed in this work is listed. However, most of the PK-specific substrates are shown with low similarity. For CK2 and PKA, we carefully check each pairs of the homologous protein sequences. However, most of the phosphorylation sites are not homologous sites. Thus, we propose the training data set is proper for this work with low redundant.

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-7-163-S1.xls]

Additional file 2

The statistics of the annotations of the phosphorylation information from Swiss-Prot database. The entries annotated with "by similarity", "potential", "probable" or "partial" are regarded as ambiguous annotations. There are only 842 annotations of the kinase-specific phosphorylation sites provided.

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-7-163-S2.xls]

Additional file 3

The statistics of the annotations of the kinase-specific phosphorylation sites from Swiss-Prot database. There are only 18 PK groups with not less than ten sites.

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-7-163-S3.xls]

Additional file 4

Data set description for each protein kinase.

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-7-163-S4.xls]

Additional file 5

The prediction performance of PPSP (self-consistency, Jack-Knife validation and n*-fold cross-validation) for 22 PK groups with large data set (N ≥ 30).*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-7-163-S5.xls]

Acknowledgements

We thank Dr. Fengfeng Zhou (UGA) for helpful discussions and critical reading of this manuscript. The authors thank Drs T.J. Gibson and F. Diella for providing the data set of Phospho.ELM for this work. We thank the anonymous referees for their many helpful comments. The work is supported by grants from Chinese Natural Science Foundation (39925018 and 30121001), Chinese Academy of Science (KSCX2-2-01), Chinese 973 project (2002CB713700), Beijing Office for Science (H020220020220) and American Cancer Society (RPG-99-173-01) to X. Yao. X. Yao is a GCC Distinguished Cancer Research Scholar.

References

1. Schafmeier T, Haase A, Kaldi K, Scholz J, Fuchs M, Brunner M: **Transcriptional feedback of neurospora circadian clock gene by phosphorylation-dependent inactivation of its transcription factor.** *Cell* 2005, **122(2):**235-246.
2. Singh CR, Curtis C, Yamamoto Y, Hall NS, Kruse DS, He H, Hannig EM, Asano K: **Eukaryotic translation initiation factor 5 is critical for integrity of the scanning preinitiation complex and accurate control of GCN4 translation.** *Mol Cell Biol* 2005, **25(13):**5480-5491.
3. Lou Y, Yao J, Zereshki A, Dou Z, Ahmed K, Wang H, Hu J, Wang Y, Yao X: **NEK2A interacts with MAD1 and possibly functions as a novel integrator of the spindle checkpoint signaling.** *J Biol Chem* 2004, **279(19):**20049-20057.
4. Liu HY, MacDonald JI, Hryciw T, Li C, Meakin SO: **Human tumorous imaginal disc 1 (TID1) associates with Trk receptor tyrosine kinases and regulates neurite outgrowth in nnr5-TrkA cells.** *J Biol Chem* 2005, **280(20):**19461-19471.
5. Robinson KN, Manto K, Buchsbaum RJ, MacDonald JI, Meakin SO: **Neurotrophin-dependent tyrosine phosphorylation of Ras guanine-releasing factor 1 and associated neurite outgrowth is dependent on the HIKE domain of TrkA.** *J Biol Chem* 2005, **280(1):**225-235.
6. Pawson T: **Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems.** *Cell* 2004, **116(2):**191-203.
7. Ma L, Chen Z, Erdjument-Bromage H, Tempst P, Pandolfi PP: **Phosphorylation and functional inactivation of TSC2 by Erk implications for tuberous sclerosis and cancer pathogenesis.** *Cell* 2005, **121(2):**179-193.
8. Beausoleil SA, Jedrychowski M, Schwartz D, Elias JE, Villen J, Li J, Cohn MA, Cantley LC, Gygi SP: **Large-scale characterization of HeLa cell nuclear phosphoproteins.** *Proc Natl Acad Sci U S A* 2004, **101(33):**12130-12135.
9. Kreegipuu A, Blom N, Brunak S: **PhosphoBase, a database of phosphorylation sites: release 2.0.** *Nucleic Acids Res* 1999, **27(1):**237-239.
10. Manning BD, Cantley LC: **Hitting the target: emerging technologies in the search for kinase substrates.** *Sci STKE* 2002, **2002(162):**PE49.
11. Kim JH, Lee J, Oh B, Kimm K, Koh I: **Prediction of phosphorylation sites using SVMs.** *Bioinformatics* 2004, **20(17):**3179-3184.
12. Biondi RM, Nebreda AR: **Signalling specificity of Ser/Thr protein kinases through docking-site-mediated interactions.** *Biochem J* 2003, **372(Pt 1):**1-13.
13. Obenauer JC, Cantley LC, Yaffe MB: **Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs.** *Nucleic Acids Res* 2003, **31(13):**3635-3641.

14. Uhlik MT, Temple B, Bencharit S, Kimple AJ, Siderovski DP, Johnson GL: **Structural and evolutionary division of phosphotyrosine binding (PTB) domains.** *J Mol Biol* 2005, **345(1):**1-20.
15. Yaffe MB, Elia AE: **Phosphoserine/threonine-binding domains.** *Curr Opin Cell Biol* 2001, **13(2):**131-138.
16. Yaffe MB, Leparc GG, Lai J, Obata T, Volinia S, Cantley LC: **A motif-based profile scanning approach for genome-wide prediction of signaling pathways.** *Nat Biotechnol* 2001, **19(4):**348-353.
17. Yaffe MB, Smerdon SJ: **The use of in vitro peptide-library screens in the analysis of phosphoserine/threonine-binding domain structure and function.** *Annu Rev Biophys Biomol Struct* 2004, **33:**225-244.
18. Holland PM, Cooper JA: **Protein modification: docking sites for kinases.** *Curr Biol* 1999, **9(9):**R329-31.
19. Joughin BA, Tidor B, Yaffe MB: **A computational method for the analysis and prediction of protein:phosphopeptide-binding sites.** *Protein Sci* 2005, **14(1):**131-139.
20. Ficarro SB, McCleland ML, Stukenberg PT, Burke DJ, Ross MM, Shabanowitz J, Hunt DF, White FM: **Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae.** *Nat Biotechnol* 2002, **20(3):**301-305.
21. Ballif BA, Villen J, Beausoleil SA, Schwartz D, Gygi SP: **Phosphoproteomic analysis of the developing mouse brain.** *Mol Cell Proteomics* 2004, **3(11):**1093-1101.
22. Blom N, Gammeltoft S, Brunak S: **Sequence and structure-based prediction of eukaryotic protein phosphorylation sites.** *J Mol Biol* 1999, **294(5):**1351-1362.
23. Blom N, Sicheritz-Ponten T, Gupta R, Gammeltoft S, Brunak S: **Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence.** *Proteomics* 2004, **4(6):**1633-1649.
24. Huang HD, Lee TY, Tzeng SW, Horng JT: **KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites.** *Nucleic Acids Res* 2005, **33(Web Server issue):**W226-9.
25. Huang HD, Lee TY, Tzeng SW, Wu LC, Horng JT, Tsou AP, Huang KT: **Incorporating hidden Markov models for identifying protein kinase-specific phosphorylation sites.** *J Comput Chem* 2005, **26(10):**1032-1041.
26. Xue Y, Zhou F, Zhu M, Ahmed K, Chen G, Yao X: **GPS: a comprehensive www server for phosphorylation sites prediction.** *Nucleic Acids Res* 2005, **33(Web Server issue):**W184-7.
27. Zhou FF, Xue Y, Chen GL, Yao X: **GPS: a novel group-based phosphorylation predicting and scoring method.** *Biochem Biophys Res Commun* 2004, **325(4):**1443-1448.
28. Diella F, Cameron S, Gemund C, Linding R, Via A, Kuster B, Sicheritz-Ponten T, Blom N, Gibson TJ: **Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins.** *BMC Bioinformatics* 2004, **5(1):**79.
29. Duda RO, Hart PE, Stork DG: **Pattern classification.** 2nd edition. Beijing , China Machine Press; 2004:680.
30. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci U S A* 1992, **89(22):**10915-10919.
31. Modrof J, Lymperopoulos K, Roy P: **Phosphorylation of bluetongue virus nonstructural protein 2 is essential for formation of viral inclusion bodies.** *J Virol* 2005, **79(15):**10023-10031.
32. Horscroft NJ, Roy P: **NTP binding and phosphohydrolase activity associated with purified bluetongue virus non-structural protein NS2.** *J Gen Virol* 2000, **81(Pt 8):**1961-1965.
33. Lymperopoulos K, Wirblich C, Brierley I, Roy P: **Sequence specificity in the interaction of Bluetongue virus non-structural protein 2 (NS2) with viral RNA.** *J Biol Chem* 2003, **278(34):**31722-31730.
34. Taraporewala ZF, Chen D, Patton JT: **Multimers of the bluetongue virus nonstructural protein, NS2, possess nucleotidyl phosphatase activity: similarities between NS2 and rotavirus NSP2.** *Virology* 2001, **280(2):**221-231.
35. Bonet C, Fernandez I, Aran X, Bernues J, Giralt E, Azorin F: **The GAGA Protein of Drosophila is Phosphorylated by CK2.** *J Mol Biol* 2005, **351(3):**562-572.
36. Arrigoni G, Marin O, Pagano MA, Settimo L, Paolin B, Meggio F, Pinna LA: **Phosphorylation of calmodulin fragments by protein kinase CK2. Mechanistic aspects and structural consequences.** *Biochemistry* 2004, **43(40):**12788-12798.
37. Huang EJ, Reichardt LF: **Trk receptors: roles in neuronal signal transduction.** *Annu Rev Biochem* 2003, **72:**609-642.