A Summary of Computational Resources for Protein Phosphorylation

Yu Xue[1,*], Xinjiao Gao[2], Jun Cao[2], Zexian Liu[2], Changjiang Jin[2], Longping Wen[2], Xuebiao Yao[2], and Jian Ren[3,*]

[1]Hubei Bioinformatics and Molecular Imaging Key Laboratory, Department of Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China; [2]Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science & Technology of China, Hefei, Anhui 230027, China; [3]College of Life Sciences, Sun Yat-sen University (SYSU), Guangzhou, Guangdong 510275, China

Abstract: Protein phosphorylation is the most ubiquitous post-translational modification (PTM), and plays important roles in most of biological processes. Identification of site-specific phosphorylated substrates is fundamental for understanding the molecular mechanisms of phosphorylation. Besides experimental approaches, prediction of potential candidates with computational methods has also attracted great attention for its convenience, fast-speed and low-cost. In this review, we present a comprehensive but brief summarization of computational resources of protein phosphorylation, including phosphorylation databases, prediction of non-specific or organism-specific phosphorylation sites, prediction of kinase-specific phosphorylation sites or phospho-binding motifs, and other tools. The latest compendium of computational resources for protein phosphorylation is available at: http://gps.biocuckoo.org/links.php

Keywords: Phosphorylation, post-translational modification, kinase-specific, phospho-binding motif, phosphorylation network.

INTRODUCTION

Many studies have indicated that various computational predictors developed in biology and biomedicine, such as those for predicting HIV cleavage sites in proteins [1-3], signal peptides [4], protein subcellular location sites [5-7], drug-target interaction [8], proteases and their types [9], membrane proteins and their types [10], enzyme functional class [11], enzyme specificity [12], GPCRs and their types [13], protein quaternary structural attributes [14, 15], protein folding rate [16], as well as a series of user-friendly web-servers summarized in the Table **3** of [17], can generate many useful data for which it would be time-consuming and costly to obtain by experiments alone. Actually, these data, combined with the information derived from the structural bioinformatics tools (see, e.g., [18-20]), can timely provide very useful insights for both basic research and drug development. This review is to summarize the progresses in developing phosphorylation databases and computational tools for predicting phosphorylation sites and other related features.

Phosphorylation is the most essential post-translational modification (PTM) of proteins, modulates proteins' conformation, stability, trafficking, interaction, and orchestrates cellular dynamics and plasticity. Biochemically, the catalytic domain of a protein kinase (PK) hydrolyzes adenosine triphosphates (ATPs) and transfers a phosphate moiety to the acceptor residue (serine, threonine or tyrosine in eukaryotes) in the substrate. It was estimated that there are >500 and >1000 PK genes encoded in mammalian [21, 22] and plant [23] genomes. To ensure signaling fidelity, each PK only specifically modifies a defined subset of substrates, while aberrances of PK functions often result in a variety of diseases and cancers. It was widely adopted that specific linear motifs around phosphorylation sites (p-sites) provide primary and major specificities for PK recognition [24-34]. However, numerous other mechanisms have also been proposed to contribute additional specificities for PKs modification *in vivo*, such as subcellular co-localization of PKs with their substrates, co-expression, co-complex, or physical interaction (collectively called as “context”) [35-39]. Importantly, identification of new phosphorylated substrates with PK-specific p-sites is fundamental for understanding the molecular mechanisms of phosphorylation. Although experimental researches have contributed great efforts to accumulate a large number of phosphorylated substrates with their sites, recently computational study of protein phosphorylation has also emerged as a popular approach, and provided useful information for further experimental verification. In this review, we briefly summarize more than 50 public databases and predictors of protein phosphorylation for experimental and computational researchers. We apologize that the computational studies without publicly available web links are not introduced. The softwares which detect potential p-sites from mass spectrometry data were also not included, since they were developed for special usages. For more detailed information on algorithms, model construction, and mechanisms of phosphorylation specificity, we recommend several excellent reviews [18, 22, 26-28, 30-34, 38, 39].

*Address correspondence to these authors at the Department of Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China; Tel:/Fax: 86-27-87793172; E-mail: xueyu@mail.hust.edu.cn; and College of Life Sciences, Sun Yat-sen University (SYSU), Guangzhou, Guangdong 510275, China; Tel:/Fax: 86-20-84114597; E-mail: renjian.sysu@gmail.com

1389-2037/10 $55.00+.00 © 2010 Bentham Science Publishers Ltd.

PHOSPHORYLATION DATABASES

In Table **1**, we listed 22 phosphorylation-related databases. To circumvent competitions, these databases usually focus on certain organisms. In 1998, Blom *et al.* developed the first phosphorylation database of PhosphoBase, including 156 phospho-proteins with 398 p-sites [29]. Later, Kreegipuu *et al.* released PhosphoBase 2.0, with 1,052 p-sites in 414 substrates [24]. In 2004, Diella *et al.* developed a public database of Phospho.ELM [40] (Table **1**). From scientific literature, they manually collected 556 experimentally verified phospho-proteins with 1,703 unique p-sites [40]. The full data in PhosphoBase was also integrated into Phospho.ELM [40] (Table **1**). Currently, Phospho.ELM 8.3 contains 5,115 known phosphorylated substrates (mostly in vertebrates) with 15,972 serine (S), 3,283 threonine (T) and 2,746 tyrosine (Y) p-sites [41]. Also, Hornbeck *et al.* collected 62,801 non-redundant p-sites from scientific literature and high-throughput experiments, and constructed a human- and mouse-centric database of PhosphoSitePlus [42] (Table **1**). With a similar strategy, the Phosphorylation Site Database collected known phosphorylated proteins in prokaryotic organisms [43], while the HPRD release 9 was developed with 80,751 p-sites in 8,163 human proteins [44] (Table **1**). By literature mining and data integration from other databases, PhosphoNET collected 74,473 p-sites in human. With a similar method, Li *et al.* recently collected experimental information for ~50 types of PTMs and constructed the-SysPTM 1.1 database, including 87,068 p-site in 24,705 targets [45].

Recently, phosphoproteomics with mass spectrometry (MS) techniques has generated a large number of p-sites. Gnad *et al.* collected 39,574 MS-derived p-sites from eukaryotes and prokaryotes and released PHOSIDA database [46, 47] (Table **1**). With a similar approach, Bodenmiller *et al.* also developed a similar database of PhosphoPep v2.0, containing ~25,000 p-sites in *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens* [48] (Table **1**). The LymPHOS database contains 342 MS-based p-sites for primary human T cells [49], while the PhosphoGRID collected 6,440 experimentally verified *in vivo* p-sites in *S. cerevisiae* [50].

Interestingly, studying phosphorylation in plants has also been paid much attention. The PhosPhAt 3.0 collected 12,457 MS-generated phosphopeptides in *A. thaliana* [51, 52] (Table **1**). Later, Gao *et al.* developed a more comprehensive database of P^{3}DB 1.1, with 14,670 p-sites in 6,382 plant proteins [53] (Table **1**). Recently, ProMEX contains 4,226 M MS-derived p-sites for *A. thaliana*, *C. reinhardii*, *M. truncatula*, and *S. meliloti*, etc. [54], while PlantsP collected ~300 MS-based p-sites for Arabidopsis thaliana plasma membrane proteins [55].

The Swiss-Prot knowledge base also contains experimental and predicted information for protein modifications, including phosphorylation [56] (Table **1**). By integrating Swiss-Prot information and other databases, dbPTM 2.0 contained PTMs information of proteins, including 22,363 known p-sites [57] (Table **1**). With a similar method, PhosphoPOINT database was released with 15,738 p-sites in 4,195 human substrates [58] (Table **1**). Also, the protein-protein information was integrated into PhosphoPOINT for PKs with their targets [58]. Systematic reconstruction of protein phosphorylation network was a great breakthrough in the field of computational phosphorylation [36, 37]. With motif-based predictors together with context information, Linding *et al.* developed a NetworKIN database and successfully discovered a highly potential phosphorylation network in *H. sapiens* [36, 37] (Table **1**). Importantly, it's widely adopted that proteins 3D structure determine their functions. In this regard, Phospho3D mapped Phospho.ELM data to PDB and obtained 3D structures of p-sites [59] (Table **1**). Moreover, identification of protein-protein interactions mediated by phosphoprotein-binding domains (PPBD) is also an attractive problem. Gong *et al.* constructed PepCyber :P~Pep 1.2 and collected 11,269 PPBD-mediated interactions among 387 PPBD-containing proteins and 1,471 substrates [60] (Table **1**). In addition, Ryu *et al.* developed PhosphoVariant database to identify genetic variations that potentially influence protein phosphorylation status [61] (Table **1**). We also designed a comprehensive database of PhosSNP 1.0 (Phosphorylation related SNP) to systematically detect 64,035 single nucleotide polymorphisms (SNPs) that might change phosphorylation states in 17,614 human proteins [62].

PREDICTION OF NON-SPECIFIC OR ORGANISM-SPECIFIC PHOSPHORYLATION SITES

Accurate prediction of p-sites in given proteins is a major challenge in the field of computational phosphorylation. P-sites predictions could be classified into three categories, including non-specific, organism-specific and kinase-specific mode. To predict non-specific or general phosphorylation sites, Blom *et al.* prepared a training data set, including 584 phosphorylated serine sites (pS), 108 phosphorylated threonines (pT) and 210 phosphorylated tyrosines (pY) [63]. Then they developed the first online predictor of NetPhos in 1999 (current version is 2.0), with an artificial neural network (ANN) algorithm [63] (Table **2**). Later, Mackey *et al.* adopted the simple pattern recognition (SPR) method and constructed CRP to carry out an *in silico* proteolytic cleavage of the protein sequence for prediction potential non-specific p-sites [64, 65] (Table **2**). Moreover, PHOSIDA used the Support Vector Machines (SVMs) method to predict non-specific p-sites, with a training data including 4,731 pS, 664 pT and 107 pY sites [46] (Table **2**).

To improve the prediction accuracy, phosphorylation sites predictors could be designed in an organism-specific manner, since different species might have distinct patterns in substrates for PKs modification. In 2004, Iakoucheva *et al.* designed a non-specific predictor of DISPHOS, which was implemented in a position-specific scoring matrix (PSSM) strategy and protein disorder information (DI) [66]. The latest version of DISPHOS 1.3 was re-trained with 1,079 experimentally verified pS, 666 pT, and 375 pY sites, and supported species-specific p-sites prediction (Table **2**). Recently, Ingrell *et al.* collected 953 pS and 192 pT sites in 675 yeast proteins and developed the first yeast-specific p-sites predictor of NetPhosYeast 1.0 [67] (Table **2**). Later, Miller *et al.* collected 103 pS and 37 pT in bacterial proteins as the training data set, and constructed the first bacteria-specific software of NetPhosBac 1.0 [68] (Table **2**). Both of the two tools were implemented in ANN algorithms. With a training data set including 802 pS, 1,818pT and 676pY sites in

Table 1. A Summary of Phosphorylation Databases. The Data Statistics were Carried out on May 12, 2010. *a*. Method, Methods Used in Collecting the Data. SL, Manually Curated from Scientific Literature; MS, Mass Spectrometry-Derived data; PS, Predicted p-sites; OS, Orthologous Sites of Experimentally Verified p-sites; TO, Taken from Other Databases or Websites; FA, Further Computational Analysis; CI, Context Information. *b*. Ref., Whether the Information Provided in the Databases is Traceable to Origin Publications. *c*. The Reference for PhosphoNET is not Available. *d*. The Swiss-Prot Knowledge Base Contains 73,467 p-sites, Including Experimentally Verified and Predicted p-sites (Statistics on May 24, 2009). For Data Statistics, the p-sites with Annotations of "By Similarity", "Probable" and "Potential" were Removed. And only the Numbers of Experimentally Verified Phosphorylation Proteins with p-sites were Calculated. *e*. Only Experimentally Verified p-sites were Counted

Databases	Main Propose	Species	Method[a]	Ref.[b]	Numbers	
					Sub.	Sites
Phospho.ELM (PhosphoBase) 8.3	Experimentally verified p-sites in eukaryotes	Mostly in vertebrates	SL, MS	Yes	7,155	29,990
PhosphoSitePlus	Human- and mouse-centric database	Mammals	SL, MS, OS	Yes	10,708	62,801
The Phosphorylation Site Database	Experimentally verified p-sites in prokaryotic organisms	Archaea and Bacteria	SL	Yes	-	-
HPRD release 9	Mainly for human p-sites	Human	SL	Yes	8,163	80,751
PhosphoNET[c]	Mainly for human p-sites	Human	SL, TO	Yes	12,400	74,473
SysPTM 1.1	~50 types of experimentally verified PTMs	General	TO, SL	Yes	24,705	87,068
PHOSIDA	MS-based *in vivo* p-sites	Eukaryotes & prokaryotes	MS	No	12,780	39,574
PhosphoPep v2.0	MS-derived p-sites for yeast, worm, fly and human	Four species	MS	No	-	~25,000
LymPHOS	MS-based p-sites for primary human T cells	Human	MS	No	~200	342
PhosphoGRID	Experimentally verified *in vivo* p-sites in *S. cerevisiae*	Yeast	SL, MS	Yes	1,776	6,440
PhosPhAt 3.0	MS-based p-sites in Arabidopsis	Arabidopsis	MS, PS	Yes	5,170	12,457
P^{3}DB 1.1	MS-based p-sites in plants	Plants	MS	No	6,382	14,670
ProMEX	MS-derived p-sites for *A. thaliana*, *C. reinhardii*, *M. truncatula*, and *S. meliloti*	Plants and Bacteria	MS	No	1,367	4,226
PlantsP	MS-based p-sites for Arabidopsis thaliana plasma membrane proteins	Arabidopsis	MS	No	-	~300
Swiss-Prot knowledge base[d]	A catalog of proteins information	General	SL, OS	Yes	11,510	36,195
dbPTM 2.0[e]	Integration of known PTMs from other databases and prediction of PTMs	General	TO, PS	No	-	22,363
PhosphoPOINT	Human kinase interactome and phosphoprotein database	Human	TO, SL	No	4,195	15,738
NetworKIN 1.0	Human phosphorylation-modulated interaction networks	Human	TO, FA, CI	No	3,978	20,224
Phospho3D	3D structures of p-sites in Phospho.ELM	Mostly in vertebrates	TO, FA	Yes	1,219	2,726
PepCyber :P~Pep 1.2	Phospho-binding domain-mediated protein interactions	Human	SL	Yes	1,471	-

(Table 2) contd….

Databases	Main Propose	Species	Method[a]	Ref.[b]	Numbers	
					Sub.	Sites
PhosphoVariant	Genetic variations that change phosphorylation state	Human	TO, FA	Yes	-	-
PhosSNP 1.0	Genetic polymorphisms that Influence protein phosphorylation status	Human	TO, FA	No	17,614	-

Table 2. Predictors for non-Specific or Organism-Specific Phosphorylation Sites. *a*. Training Data Set, the Experimentally Verified p-sites were Taken as Positive Training Data set. *b*. Specificity, for General Propose or Organism-Specific p-sites Prediction. *c*. Method, the Computational Methods Used for Training. ANN, Artificial Neural Network; SPR, Simple Pattern Recognition; SVMs, Support Vector Machines; PSSM, Position-Specific Scoring Matrix; DI, disorder information. *d*. PTMP-UI, Whether the Predictor Follows a Unified User Interface (UI). For Example, The Input of PhosPhat 3.0 Only Allows AGI Codes from The Arabidopsis Information Resource (TAIR). *e*. N/A, Not Available

Predictors	Training Data Set[a]	Specificity[b]	Method[c]	PTMP-UI[d]			
				IN	O1	O2	O3
NetPhos 2.0	584 pS, 108 pT and 210 pY sites	General	ANN	Y	Y	Y	Y
CRP	N/A[e]	General	SPR	Y	N	N	N
PHOSIDA	4,731 pS, 664 pT and 107 pY sites	General	SVMs	Y	Y	N	N
DISPHOS 1.3	1,079 pS, 666 pT and 375 pY sites	Species-specific	PSSM, DI	Y	Y	Y	Y
NetPhosYeast 1.0	953 pS and 192 pT sites in yeast	Yeast	ANN	Y	Y	Y	Y
NetPhosBac 1.0	103 pS and 37 pT in bacterial proteins	Bacteria	ANN	Y	Y	Y	Y
PhosPhAt 3.0	802 pS, 1,818pT and 676pY sites	Arabidopsis	SVMs	N	N	Y	Y

Arabidopsis, PhosPhAt 3.0 was implemented in the SVMs algorithm as the first Arabidopsis-specific predictor [51, 52] (Table **2**).

The input and output of predictors are greatly useful for experimental researchers. Users usually regarded the complicated computational algorithms as “black boxes”. However, with a simple but straightforward user interface (UI), experimentalists can easily input their data, click on the “submit” button, and obtained the prediction results. Previously, we collected 32 PTM sites prediction tools and proposed some general rules for a unified UI [69]. The rationale post-translational modification site prediction user interface (PTMP-UI) is presented below:

1) Input (IN): protein primary sequences (usually in FASTA format)
2) Output (O1): position of the predicted PTM site
3) Output (O2): flanking peptide of the predicted PTM site
4) Output (O3): evaluation score or probability of the predicted PTM site

Most of predictors for PTMs followed this basic rationale, while some of them also provided auxiliary features [69]. In this work, we tested the UIs of all online available tools. The detailed results were shown in Table **2**.

PREDICTION OF KINASE-SPECIFIC PHOSPHORYLATION SITES OR PHOSPHO-BINDING MOTIFS

Currently, prediction of kinase-specific p-sites has emerged to be more useful for experimental researchers, since there are too many PKs in eukaryotes and each PK might recognize a distinct pattern for modification. As the demand for carrying out large-scale predictions and discovering potential phosphorylation networks evolves, accurate and robust prediction of kinase-specific p-sites has become necessary and challenging [36, 37].

The methods of kinase-specific predictions could be classified into two categories: simple motif-based or complex algorithm-based. A phosphorylation motif could be represented with a pattern or a regular expression. Thus, the simple motif-based or simple pattern recognition (SPR) approach is straightforward and convenient: match or not [70, 71]. The PROSITE is the first integrated database to collect protein patterns, while its associated tool of ScanProsite could be used to search simple motifs, also including 3 kinase-specific phosphorylation motifs [70, 71] (Table **3**). With a similar approach, Puntervoll *et al.* developed a comprehensive resource of ELM to scan potential functional linear motifs in proteins [72] (Table **3**). Also, the context information was added to improve the prediction accuracy [72] (Table **3**). In 2006, Balla *et al.* collected 312 unique protein

Table 3. Predictors for Kinase-Specific Phosphorylation Sites and Phospho-Binding Motifs. *a*. Training Data set, The Experimentally Verified p-sites were Taken as Positive Training Data Set. *b*. Num. of PKs, the Number of PKs That the Predictors Could Predict for Their Specific p-sites. *c*. Method, the Computational Methods Used for Training. SPR, Simple Pattern Recognition; PSSM, Position-Specific Scoring Matrix; CI, Context Information; SA, Statistical Analysis; ANN, Artificial Neural Network; SVMs, Support Vector Machines; GPS, Group-Based Prediction System; BDT, Bayesian Decision Theory; HMM, Hidden Markov Model; LOR, Log-Odds Ratio; KSB, Simplified Kinase-Substrate Binding Model; CRF, Conditional Random Fields; WVM, Weighted Voting; SP, Sequence Patterns; EI, Evolutionary Information. *d*. PTMP-UI, Whether The Predictor Follows a Unified User Interface (UI). *e*. N/A, Not Available

Predictors	Training Data Set[a]	Num. of PKs[b]	Method[c]	PTMP-UI[d]			
				IN	O1	O2	O3
ScanProsite	N/A[e]	3	SPR, PSSM	Y	Y	Y	N
ELM	N/A	12	SPR, CI	Y	N	Y	N
Minimotif Miner 2.0	N/A	N/A	SPR, SA	Y	Y	N	Y
PhosphoMotif Finder	N/A	~90	SPR	Y	Y	Y	N
PREDIKIN 1.0	N/A	N/A	SPR	Y	N	N	N
Predikin & PredikinDB 2.0	2,335 S/T/Y PK-specific p-sites	N/A	PSSM	Y	Y	Y	Y
ScanSite 2.0	N/A	~27	PSSM	Y	Y	Y	Y
NetPhosK 1.0	N/A	17	ANN	Y	Y	N	Y
PredPhospho 1.0	~830-1071 S/T PK-specific p-sites	4 PK groups & 4 PK families	SVMs	Y	Y	N	N
PredPhospho 2.0	N/A	7 PK groups & 18 PK families	SVMs	Y	Y	N	Y
GPS 1.10	~2,060 S/T/Y PK-specific p-sites	216	GPS	Y	Y	Y	Y
GPS 2.0	3,161 S/T/Y PK-specific p-sites	408	GPS	Y	Y	Y	Y
PPSP 1.0	~2,060 S/T/Y PK-specific p-sites	~70 PK groups	BDT	Y	Y	Y	Y
KinasePhos 1.0	1,163 S/T/Y PK-specific p-sites	18 PK groups	HMM	Y	Y	Y	Y
KinasePhos 2.0	3,751 S/T/Y PK-specific p-sites	58 PK groups	SVMs	Y	Y	Y	Y
PhoScan	~400 S/T PK-specific p-sites	~48 PK families	LOR	Y	Y	Y	Y
pkaPS	239 S/T PKA-specific p-sites	1	KSB	Y	Y	Y	Y
CRPhos 0.8	2,510 S/T/Y PK-specific p-sites	34	CRF	N/A			
AutoMotif 2.0	N/A	~36	SVMs	Y	Y	Y	Y
SMALI	N/A	N/A	PSSM	Y	Y	Y	Y
MetaPredPS	N/A	N/A	WV	N/A			
NetPhorest	4,169 S/T/Y PK-specific p-sites	179	PSSM, ANN	Y	Y	Y	Y
PostMod	3,258 S/T/Y PK-specific p-sites	48 PK groups	SP, EI	Y	Y	Y	Y

motifs (with 73 phosphorylation motifs) from scientific literature, and constructed a motif-based tool of Minimotif Miner [73, 74] (Table **3**). A simple statistical enrichment ratio was calculated as the predicted score. With a similar strategy, Amanchy *et al*. designed the PhosphoMotif Finder and collected 324 p-sites motifs or phospho-binding motifs [75] (Table **3**). Also, based on the SPR strategy, PREDIKIN 1.0 was established to predict kinase-specific p-sites [76].

Although motif-based approaches were widely used, prediction of kinase-specific sites with complex algorithms was also popular for its higher accuracy. And the threshold values could be set in a more flexible manner. For example, the Predikin & PredikinDB 2.0 were implemented in a PSSM approach (Table **3**) [77, 78]. In 2004, the Scansite 2.0 was implemented in the PSSM algorithm to predict ~27 PK-specific p-sites and several phospho-binding motifs [79] (Table **3**). Also, Blom *et al*. used an ANN algorithm and desig-ned NetPhosK 1.0, which could predict kinase-specific p-sites for ~17 PKs [27] (Table **3**). Furthermore, the PredPhospho 1.0 was implemented in SVMs algorithm to predict for 4 PK groups and 4 PK families [80] (Table **3**). And its enhanced version of PredPhospho 2.0 could predict for 7 PK groups and 18 PK families (Table **3**) [61]. We also contributed great efforts on kinase-specific predictions. In 2004, we developed GPS 1.0 & 1.10 (Group-based Phosphorylation Scoring) algorithm with two hypotheses [81, 82] (Table **3**). First, we hypothesized that similar peptides might bear similar biological properties. Also, we assumed that one PK could recognize more than one motif/pattern in substrates [81, 82]. GPS 1.10 could predict kinase-specific phosphorylation sites for 71 PK groups, including 216 unique PKs [81, 82] (Table **3**). Recently, we greatly improved the GPS algorithm and released GPS 2.0 (Group-based Prediction System) software, which could predict for 408 PKs in human [83] (Table **3**). We also used the Bayesian Decision Theory (BDT) method to develop PPSP 1.0 [84] (Table **3**). And the prediction power of PPSP 1.0 was comparable with our GPS 1.10 [84]. Other researchers also constructed several predictors, including KinasePhos 1.0 (implemented in Hidden Markov Model, HMM) [85], KinasePhos 2.0 (SVMs) [86], PhoScan (Log-odds ratio, LOR) [87], pkaPS (Simplified kinase-substrate binding model, KSB) [88], CRPhos (Conditional random fields, CRF) [89], AutoMotif (SVMs) [90, 91], and PostMod (Sequence patterns and evolutionary information) [92], etc. (Table **3**). Based on the results of oriented peptide array libraries, SAMLI was constructed to predict SH2-binding peptides (usually phospho-peptides) in proteins [93, 94] (Table **3**). Furthermore, multiple complex algorithms could be combined together to improve the prediction power. For example, a recent software of MetaPred-PS, designed a weighted voting (WV) meta-predicting approach to integrate the prediction results from other programs [95] (Table **3**). Finally, simple motif-based methods could also be combined with complex algorithm-based strategies. For example, NetPhorest used the PSSMs, known patterns and machine-learning algorithms (e.g., ANN) together to predict kinase-specific phosphorylation sites or phospho-binding motifs [96] (Table **3**). Again, the input and output formats of these predictors were carefully evaluated (Table **3**).

In this work, we critically evaluated and compared the prediction performances of different PK-specific predictors, including our GPS 2.0 [83], ScanSite 2.0 [79], KinasePhos 1.0 & 2.0 [85, 86], NetPhosK 1.0 [27], pkaPS [88], PPSP 1.0 [84], PhoScan [87] and NetPhorest [96]. Usually, the prediction performances could be evaluated by self-consistency validation, leave-one-out validation and *n*-fold cross-validation [83]. Since the leave-one-out validations and *n*-fold cross-validations for other tools were not available, we focused on the comparison of the self-consistency performances. From Phospho.ELM 6.0 database, we prepared a testing data set for 4 well-studied PKs, including experimentally verified p-sites for PKA, ATM, CDC2, and Src. As previously described [81-84], we took the experimentally verified phosphorylation sites as the positive data (+), while all other residues (S/T or Y) in the same substrates were regarded as the negative data (-).The data statistics were shown in Table **4** (also available at: http://gps.biocuckoo.org/links.php). Among the data with positive hits by a predictor, the real positives are defined as true positives (*TP*), while the others are defined as false positives (*FP*). Among the data with negative predictions by the predictor, the real positives are defined as false negatives (*FN*), while the others are defined as true negatives (*TN*). Then four standard performance measurements of accuracy (*Ac*), sensitivity (*Sn*), specificity (*Sp*) and Mathew correlation coefficient (*MCC*) were defined as below [81-84]:

Table 4. A Testing Data Set for PKA, ATM, CDC2 and Src. The Data Set Contains Experimentally Verified PK-Specific Phosphorylation Sites from Phospho.ELM 6.0 Database. The Data Set is Freely Available at: http://gps.biocuckoo.org /links.php

PKs	Substrates	P-Sites	
		Positive	Negative
PKA	210	337	19,091
ATM	28	55	3,712
CDC2	65	130	6,362
Src	86	136	1,758

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

$$Ac = \frac{TP + TN}{TP + FP + TN + FN}, \text{ and}$$

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

For GPS 2.0 [83], the AGC/PKA, Atypical/PIKK/ATM, CMGC/CDK/CDC2/CDC2 and TK/Src/Src were selected for PK-specific p-sites prediction. For ScanSite 2.0 [79], the “Protein kinase A”, “ATM kinase”, “Cdc2 kinase” and “Src kinase” were chosen. For KinasePhos 1.0 [85], the “cAMP-dependent protein kinase (PKA)”, “Ataxia telangiectasia mutated kinase (ATM)”, Cyclin-dependent kinase (CDK) and “Tyrosine kinase Src” were selected. For kinasePhos 2.0 [86], the “cAMP-dependent protein kinase(PKA)”, “Ataxia telangiectasia mutated(ATM)”, “Cell division cycle protein kinase(CDC2)” and “Tyrosine kinase Src(Src)” were chosen. For NetPhosK 1.0 [27], the PKA, ATM, Cdc2, and Src were selected. For pkaPS [88], only PKA was used. For PPSP 1.0 [84], the PKA, ATM, CDKs, and SRC were adopted. For PhoScan [87], the PKA, ATM_ATR_group and CDK were used. Finally, for NetPhorest [96], the PKA_group, ATM-_ATR_group, CDK1 and Src_group were chosen. Both the positive and negative data sets were submitted on these online services directly. Then the *Ac*, *Sn*, *Sp* and *MCC* values were calculated for each predictor. Due to the page limitation, the results of *Sn* and *Sp* were not shown in Table **5**. And the full performances are available in Table **S1** in Supplemental Data. For comparison, we fixed the *Sp* value of GPS 2.0 to be nearly equal with other tools and compared the *Sn* values (Table **5**). Generally, GPS 2.0 exhibits better performances than other softwares (Table **5**). In our previous

Table 5. Comparison of Several PK-Specific Predictors, Including Scansite, KinasePhos 1.0 & 2.0, NetPhosK 1.0, pkaPS, PPSP 1.0, PhoScan, NetPhorest, and GPS 2.0. We Fixed the *Sp* Value of GPS 2.0 to be Similar with That Used in Other Tools to Compare the *Sn* Values. The Performances with Better Values than Those from GPS 2.0 are Bold. The Detailed Results are Available in Table S1 in Supplementary Data

Comparison	Cut-Off	PKA		ATM		CDC2		Src	
		Sn	*Sp*	*Sn*	*Sp*	*Sn*	*Sp*	*Sn*	*Sp*
ScanSite 2.0	low	69.14%	95.02%	54.55%	93.67%	73.08%	95.13%	28.68%	95.28%
	medium	42.43%	99.17%	27.27%	98.57%	**29.23%**	**99.26%**	11.76%	99.37%
	high	**16.91%**	**99.91%**	18.18%	99.70%	8.46%	99.84%	**3.68%**	**99.94%**
KinasePhos 1.0	90%	85.16%	90.64%	89.09%	83.86%	72.31%	86.37%	47.06%	89.93%
	95%	80.12%	94.50%	87.27%	89.76%	63.08%	92.69%	38.24%	93.91%
	100%	58.46%	98.42%	81.82%	96.04%	**48.46%**	**97.99%**	25.00%	97.84%
Kinasephos 2.0	Default	55.19%	89.20%	89.09%	38.12%	**13.08%**	**99.72%**	86.76%	55.97%
NetPhosK 1.0	Default	77.74%	91.18%	85.45%	97.60%	16.92%	87.79%	33.09%	95.39%
pkaPS	Default	89.61%	90.81%						
PPSP 1.0	High Sn	97.92%	28.39%	98.18%	42.19%	93.85%	35.15%	94.85%	22.18%
	Balance	87.54%	90.58%	96.36%	91.38%	83.08%	94.11%	74.26%	74.86%
	High Sp	**1.78%**	**99.99%**	**21.82%**	**100%**	10.00%	99.80%	**7.35%**	**99.89%**
PhoScan	high	**40.95%**	**99.50%**	45.45%	99.16%	**33.85%**	**99.06%**		
	low	73.89%	91.49%	89.09%	94.77%	67.69%	94.73%		
NetPhorest	Default	94.96%	76.29%	100%	90.68%	86.92%	90.88%	54.41%	84.19%
GPS 2.0		83.09%	95.04%	100%	94.03%	82.31%	95.16%	56.62%	95.32%
		49.26%	99.17%	72.73%	98.62%	26.15%	99.27%	15.44%	99.41%
		8.61%	99.91%	32.73%	99.70%	10.77%	99.84%	3.68%	99.94%
		89.91%	90.75%	/	/	89.23%	86.52%	71.32%	89.93%
		84.27%	94.58%	/	/	86.92%	92.74%	63.97%	93.97%
		64.39%	98.43%	98.18%	96.04%	46.92%	98.00%	38.24%	98.01%
		91.69%	89.25%	/	/	11.54%	99.73%	96.32%	56.21%
		89.61%	91.26%	87.27%	97.61%	89.23%	87.90%	53.68%	95.43%
		89.91%	90.91%						
		100%	28.75%	/	/	100%	35.40%	99.26%	23.36%
		89.91%	90.63%	/	/	86.15%	94.11%	83.09%	75.00%
		0.59%	99.99%	16.36%	100%	10.77%	99.83%	4.41%	99.88%
		40.65%	99.50%	63.64%	99.16%	31.54%	99.06%		
		89.61%	91.57%	100%	94.79%	85.38%	94.73%		
		97.63%	76.31%	100%	94.03%	87.69%	90.96%	75.00%	84.31%

work [83], we observed when an extremely high *Sp* value was chosen, a predictor will only generate a few number of positive hits. In GPS 2.0, we improved the algorithm to enhance the prediction performances around *Sp* of 90% [83]. In this regard, when an extremely high *Sp* value was selected, e.g., *Sp* >99%, the GPS 2.0 was not better than other tools, including ScanSite 2.0 (PKA, CDC2, and Src), KinasePhos 2.0 (CDC2), PPSP 1.0 (PKA, ATM, and Src), and PhoScan

(PKA and CDC2) (Table **5**). The KinasePhos 1.0 with the *Sn* and *Sp* of 48.46% and 97.99% was also better than GPS 2.0 (*Sn* & *Sp* of 46.92% &98.00%) (Table **5**). However, the prediction performances of different predictors could still be comparative.

Additionally, we used the GPS 2.0 as a typical tool to compare the simple motif-based and complex algorithm-based approaches, with the same testing data set. The experimentally verified p-sites motifs for PKA, ATM, CDC2 and Src were taken from PhosphoMotif Finder website [75]. The prediction performances for several typical p-sites motifs were shown in Table **6**. More detailed information was shown in Table **S2** in Supplemental Data. Obviously, the GPS 2.0 generated better performances, when the *Sp* value was not greatly high (*Sp* < 99%) (Table **6**).

MISCELLANEOUS TOOLS

Besides phosphorylation databases and prediction of p-sites, there were several other related researches. Recently, discovery of informative phosphorylation motifs from large-scale phosphoproteomic data attracted much attention. Schwartz *et al*. developed a novel software of Motif-X with an iterative statistical algorithm, to discover potentially informative p-sites motifs from high-throughput MS-derived phosphoproteomic data [97] (Table **7**). Then they developed an associated tool of Scan-X, which used phosphorylation motifs detected from Motif-X to scan potential p-sites in proteins [98] (Table **7**). Also, Ritz *et al*. construct a similar tool of MoDL for discovery of p-sites motifs, with a Motif Description Length (MoDL) algorithm [99] (Table **7**). Interestingly, Wang *et al*. modified the classical BLAST algorithm to design the PhosphoBlast, which could detect potential p-sites by sequence similarity [100] (Table **7**). Literature mining of phosphorylation information is useful for data integration and collection. However, there was only one related software of RLIMS-P reported [101, 102] (Table **7**). For other tools, Lachmann *et al*. developed KEA (kinase enrichment analysis) to elucidate kinases-substrates relationship [103] (Table **7**). Finally, we also developed DOG 1.0, which could visualize protein functional domain and modification sites in a user-defined manner [104] (Table **7**).

Table 6. Comparison of Simple Motif-Based Approach to Complex Algorithm-Based Algorithm. GPS 2.0 was Chosen as An Example of Complex Algorithm-Based Predictor. The Experimentally Discovered p-sites Motifs were Taken from the PhosphoMotif Finder Website. Again, We Fixed the *Sp* Value of GPS 2.0 to be Similar with SPR Performance to Compare the *Sn* Values. The Performances with Better Values Than Those From GPS 2.0 are Bold. The Full Comparisons are Available in Table S2 in Supplementary Data

Simple Motifs	SPR Performance				GPS 2.0 Performance			
	Ac	*Sn*	*Sp*	*MCC*	*Ac*	*Sn*	*Sp*	*MCC*
PKA								
RRXpS[M/I/L/V/F/Y]	**98.41%**	**11.31%**	**99.96%**	**0.3022**	98.24%	1.48%	99.96%	0.0728
RXpS	96.11%	50.30%	96.92%	0.3189	96.53%	75.07%	96.92%	0.4627
RXXpS	95.83%	54.17%	96.57%	0.3267	96.22%	76.26%	96.58%	0.4511
[R/K]X[pS/pT]	89.69%	74.70%	89.96%	0.2685	90.05%	91.10%	90.03%	0.3346
KXX[pS/pT]	93.04%	13.99%	94.44%	0.0475	94.29%	85.16%	94.45%	0.4102
ATM								
[P/L/I/M]X[L/I/D/E]pSQ	98.59%	18.52%	99.78%	0.3154	98.72%	27.27%	99.78%	0.4166
LpSQE	98.57%	3.70%	99.97%	0.1549	98.77%	18.18%	99.97%	0.4035
pSQ	96.08%	92.59%	96.13%	0.4809	96.20%	98.18%	96.17%	0.5109
CDC2								
[R/K]pSP[R/P][R/K/H]	**98.00%**	**0.78%**	**100%**	**0.0872**	97.95%	0.77%	99.95%	0.0407
[pS/pT]PX[R/K]	**98.16%**	**32.56%**	**99.51%**	**0.4244**	97.88%	19.23%	99.51%	0.2838
HHH[R/K]pSPR[R/K]R	97.99%	0	100%	N/A	N/A			
SRC								
pYMXM	92.69%	2.22%	99.88%	0.1054	92.90%	5.15%	99.88%	0.1886
EEEIpY[G/E]EFD	92.64%	0	100%	N/A	N/A			
pY[A/G/S/T/D/E]	63.28%	60.00%	63.55%	0.1266	65.78%	89.71%	63.88%	0.2858

Table 7. Miscellaneous Tools that Were Not Classified. *a*. Method. IS, Iterative Statistical Approach; SPR, Simple Pattern Recognition; MoDL, Motif Description Length; TM, Text-Mining; SA, Statistical Analysis

Tools	Main Propose	Method^a
Motif-X	Identification of phosphorylation motifs from large-scale data	IS
Scan-X	Prediction of potential p-sites in yeast, fly, mouse and human	SPR
MoDL	Discovery of phosphorylation motifs from phosphorylated peptides	MoDL
Phos-phoBlast	For searching homologus phosphorylated peptides	BLAST
RLIMS-P	Extract protein phosphorylation information from the abstracts	TM
KEA	Prediction of kinase-substrate association	SA
DOG 1.0	Visualization of protein functional domain and PTM sites	JAVA

DISCUSSION

In this review, we briefly summarize the current progress of most aspects of computational resources for protein phosphorylation. The computational studies without web links were not introduced, because it's not convenient to be used by experimental researchers. Totally, there are 16 phosphorylation databases and 36 computational programs listed. The web links and references for these resources are available in Table **S3** in Supplemental Data. We believe that more and more related studies will be carried out, and more and more databases and softwares will be constructed and released in the near future. For further computational studies, we give several personal perspectives on the computational phosphorylation:

(1) Integration of experimentally verified phosphorylation information. As we described above, there were more than ten phosphorylation databases constructed (Table **1**). However, no one contains the full data set. And the data qualities are heterogeneous in different databases. In this regard, we expected that some efforts should be carried out to integrate phosphorylation information from different resources, with careful curation.

(2) Standardization of the input and output format. Currently, the input and output formats of existed databases and predictors are still not unified, which might be difficult for users. For example, Phospho.ELM database allows the protein/gene name, public database accession, or primary protein sequences as input [24, 29, 40, 41], while only protein/gene names are permitted in PhosphoSitePlus [42]. A unified input and output rationale should be established for users. And most of predictors have already followed a unified user interface (PTMP-UI) [69]. In addition, we suggest that the data storage in phosphorylation databases could also be organized in a unified format, which might be useful for data sharing, distribution and integration.

(3) Improvement of existed approaches and development of novel methods. Although several simple motif-based or complex algorithm-based algorithms were adopted for p-sites prediction, the performance could still be improved. The existed approaches could be improved, e.g., GPS 2.0 [83]; Different existed algorithms could be combined together, e.g., MetaPredPS [95] and NetPhorest [96]; New algorithms could be developed, e.g., CRPhos 0.8 [89]. We and other researchers are still working on development of more efficient and accurate algorithms.

(4) Combining protein 3D structures and evolutionary information. Most of researchers believed that protein 3D information will be useful for p-sites prediction [26-28, 33]. However, the 3D structure information of proteins is still very limited compared to the huge number of proteins in the public databases. And structural computational is time-consuming and slow-speed. The evolutionary information was also proposed to be useful for performance improvement, e.g., NetPhosK 1.0 [27]. However, this additional procedure will also slow down the prediction process. How to include 3D and evolutionary information without slowing down the prediction speed is still a great challenge.

(5) Construction of more organism-specific predictors. Prediction of p-sites in a species- specific mode will be more accurate than the non-specific manner, since different organisms might have different patterns in substrates for PKs modification. Currently, there were four organism-specific predictors developed (Table **2**). And we believe that more and more species-specific predictors will be released in the near future.

(6) Analysis of large-scale phosphoproteomic data. Recently, large-scale phosphoproteomic studies with high-throughput MS-based techniques have been widely carried out to generate a large number of p-sites. Usually, the cognate PKs for these p-sites were not known. In this regard, annotation of PK information for large-scale phosphoproteomic data will be helpful for further experimental consideration. Previously, we directly used GPS 2.0 to annotate PK information for ~12,000 non-specific p-sites in Phospho.ELM database [83]. Also, discovery of potential p-sites motifs from phosphoproteomic data is also helpful for prediction, e.g., Motif-X [97] and MoDL [99].

(7) Re-construction of phosphorylation pathways and networks. Systematically re-construction of potential phosphorylation pathways and networks will be useful for further experimental design. For example, Linding *et al.* developed a NetworKIN database and successfully discovered a highly potential phosphorylation network in *H. sapiens* [36, 37] (Table **1**). Re-construction of phosphorylation networks beyond human will be a great challenge for computational researchers.

(8) From prediction to drug design. Aberrances of phosphorylation system are frequently involved in various diseases and cancers [105]. The deleterious variations, e.g., non-synonymous single nucleotide polymorphisms (SNPs) and somatic mutations in kinases or substrates could change their original functions and properties [62, 105]. Currently, it

was estimated that ~20% of all potential drug targets are PKs [105]. In this regard, further computational studies on regulatory roles of phosphorylation will be helpful for drug design. For example, structural modeling analyses revealed the interacting mechanisms of CDK5 and its activators [106, 107]. If genetic variations occur at residues located in binding interface, they might disrupt the kinase-regulator interaction and rewire signaling pathways. Analogously, structural modeling of kinase-substrate interaction should also be carried out in the near future.

For other personal suggestions, we propose that the online services or downloadable packages should be prepared at least for academic usages. And either a phosphorylation database or a predictor should be designed in an easy-to-use manner. Finally, the version number should be added, if the database or software will be updated later. Again, we believe that computational studies together with experimental verifications will propel the phosphorylation research into a new phase.

ACKNOWLWDGEMENTS

The authors are grateful for the three anonymous reviewers, whose suggestions have greatly improved the presentation of this manuscript. We thank Dr. Martin Miller (New York, USA) for his helpful suggestions. We are also grateful that Dr. Lorenzo A. Pinna (Padua, Italy) kindly sent us a hard-copy reprint of his article. This work was supported by grants from the National Basic Research Program (973 project) (2006CB933300, 2007CB947401, 2010CB945400), National Natural Science Foundation of China (90919001, 30700138, 30830036, 30900835), Chinese Academy of Sciences (KSCX2-YW-R-139, INFO-115-C01-SDB4-36), and National Science Foundation for Post-doctoral Scientists (20080430100).

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publishers Web site along with the published article.

REFERENCES

[1] Chou, K.C. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. *J. Biol. Chem.*, **1993**, *268(23)*, 16938-16948.

[2] Chou, K.C. Prediction of human immunodeficiency virus protease cleavage sites in proteins. *Anal. Biochem*., **1996**, *233(1)*, 1-14.

[3] Shen, H.B.; Chou, K.C. HIVcleave: a web-server for predicting human immunodeficiency virus protease cleavage sites in proteins. *Anal. Biochem.*, **2008**, *375(2)*, 388-390.

[4] Chou, K.C.; Shen, H.B. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem. Biophys. Res. Commun.*, **2007**, *357(3)*, 633-640.

[5] Chou, K.C.; Shen, H.B. Recent progress in protein subcellular location prediction. *Anal. Biochem.*, **2007**, *370(1)*, 1-16.

[6] Chou, K.C.; Shen, H.B. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.*, **2008**, *3(2)*, 153-162.

[7] Chou, K.C.; Shen, H.B. A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. *PLoS One*, **2010**, *5(4)*, e9931.

[8] He, Z.; Zhang, J.; Shi, X.H.; Hu, L.L.; Kong, X.; Cai, Y.D.; Chou, K.C. Predicting drug-target interaction networks based on functional groups and biological features. *PLoS One*, **2010**, *5(3)*, e9603.

[9] Chou, K.C.; Shen, H.B. ProtIdent: a web server for identifying proteases and their types by fusing functional domain and sequential evolution information. *Biochem. Biophys. Res. Commun.*, **2008**, *376(2)*, 321-325.

[10] Chou, K.C.; Shen, H.B. MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem. Biophys. Res. Commun*., **2007**, *360(2)*, 339-345.

[11] Shen, H.B.; Chou, K.C. EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. *Biochem. Biophys. Res. Commun*., **2007**, *364(1)*, 53-59.

[12] Chou, K.C. A sequence-coupled vector-projection model for predicting the specificity of GalNAc-transferase. *Protein Sci.*, **1995**, *4(7)*, 1365-1383.

[13] Xiao, X.; Wang, P.; Chou, K.C. GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes. *J. Comput. Chem.*, **2009**, *30(9)*, 1414-1423.

[14] Shen, H.B.; Chou, K.C. QuatIdent: a web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. *J. Proteome Res.*, **2009**, *8(3)*, 1577-1584.

[15] Xiao, X.; Wang, P.; Chou, K.C. Predicting the quaternary structure attribute of a protein by hybridizing functional domain composition and pseudo amino acid composition. *J. Appl. Cryst*., **2009**, *42*, 169-173.

[16] Shen, H.B.; Song, J.N.; Chou, K.C. Prediction of protein folding rates from primary sequence by fusing multiple sequential features. *J. Biomed. Sci. Eng*., **2009**, *2*, 136-143.

[17] Chou, K.C.; Shen, H.B. REVIEW: Recent advances in developing web-servers for predicting protein attributes. *Nat. Sci.*, **2009**, *2*, 63-92.

[18] Chou, K.C. Structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem.*, **2004**, *11(16)*, 2105-2134.

[19] Chou, K.C. Modelling extracellular domains of GABA-A receptors: subtypes 1, 2, 3, and 5. *Biochem. Biophys. Res. Commun.*, **2004**, *316(3)*, 636-642.

[20] Chou, K.C. Molecular therapeutic target for type-2 diabetes. *J. Proteome Res*., **2004**, *3(6)*, 1284-1288.

[21] Caenepeel, S.; Charydczak, G.; Sudarsanam, S.; Hunter, T.; Manning, G. The mouse kinome: discovery and comparative genomics of all mouse protein kinases. *Proc. Natl. Acad. Sci. USA*, **2004**, *101(32)*, 11707-11712.

[22] Manning, G.; Whyte, D.B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the human genome. *Science*, **2002**, *298(5600)*, 1912-1934.

[23] Vlad, F.; Turk, B.E.; Peynot, P.; Leung, J.; Merlot, S. A versatile strategy to define the phosphorylation preferences of plant protein kinases and screen for putative substrates. *Plant J.*, **2008**, *55(1)*, 104-117.

[24] Kreegipuu, A.; Blom, N.; Brunak, S. PhosphoBase, a database of phosphorylation sites: release 2.0. *Nucleic Acids Res.*, **1999**, *27(1)*, 237-239.

[25] Kreegipuu, A.; Blom, N.; Brunak, S.; Jarv, J. Statistical analysis of protein kinase specificity determinants. *FEBS lett.*, **1998**, *430(1-2)*, 45-50.

[26] Pinna, L.A.; Ruzzene, M. How do protein kinases recognize their substrates? *Biochim. Biophys. Acta*, **1996**, *1314(3)*, 191-225.

[27] 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.

[28] Kobe, B.; Kampmann, T.; Forwood, J.K.; Listwan, P.; Brinkworth, R.I. Substrate specificity of protein kinases and computational prediction of substrates. *Biochim. Biophys. Acta*, **2005**, *1754(1-2)*, 200-209.

[29] Blom, N.; Kreegipuu, A.; Brunak, S. PhosphoBase: a database of phosphorylation sites. *Nucleic Acids Res.*, **1998**, *26(1)*, 382-386.

[30] Diella, F.; Haslam, N.; Chica, C.; Budd, A.; Michael, S.; Brown, N.P.; Trave, G.; Gibson, T.J. Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front. Biosci.*, **2008**, *13*, 6580-6603.

[31] Hjerrild, M.; Gammeltoft, S. Phosphoproteomics toolbox: computational biology, protein chemistry and mass spectrometry. *FEBS Lett.*, **2006**, *580(20)*, 4764-4770.

[32] Miller, M.L.; Blom, N. Kinase-specific prediction of protein phosphorylation sites. *Methods Mol. Biol.*, **2009**, *527*, 299-310.

[33] Ubersax, J.A.; Ferrell, J.E., Jr. Mechanisms of specificity in protein phosphorylation. *Nat. Rev. Mol. Cell. Biol*., **2007**, *8(7)*, 530-541.

[34] Zhu, G.; Liu, Y.; Shaw, S. Protein kinase specificity. A strategic collaboration between kinase peptide specificity and substrate recruitment. *Cell Cycle*, **2005**, *4(1)*, 52-56.

[35] Yaffe, M.B.; Leparc, G.G.; Lai, J.; Obata, T.; Volinia, S.; Cantley, L.C. A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nat. Biotechnol*., **2001**, *19(4)*, 348-353.

[36] Linding, R.; Jensen, L.J.; Ostheimer, G.J., van Vugt, M.A.; Jorgensen, C.; Miron, I.M.; Diella, F.; Colwill, K.; Taylor, L.; Elder, K.; Metalnikov, P.; Nguyen, V.; Pasculescu, A.; Jin, J.; Park, J.G.; Samson, L.D.; Woodgett, J.R.; Russell, R.B.; Bork, P.; Yaffe, M.B.; Pawson, T. Systematic discovery of *in vivo* phosphorylation networks. *Cell*, **2007**, *129(7)*, 1415-1426.

[37] Linding, R.; Jensen, L.J.; Pasculescu, A.; Olhovsky, M.; Colwill, K.; Bork, P.; Yaffe, M.B.; Pawson, T. NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Res*., **2008**, *36(Database issue)*, D695-699.

[38] Biondi, R.M.; Nebreda, A.R. Signalling specificity of Ser/Thr protein kinases through docking-site-mediated interactions. *Biochem. J*., **2003**, *372(Pt 1)*, 1-13.

[39] Holland, P.M.; Cooper, J.A. Protein modification: docking sites for kinases. *Curr. Biol.*, **1999**, *9(9)*, R329-331.

[40] Diella, F.; Cameron, S.; Gemund, C.; Linding, R.; Via, A.; Kuster, B.; Sicheritz-Ponten, T.; Blom, N.; Gibson, T.J. Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, **2004**, *5*, 79.

[41] Diella, F.; Gould, C.M.; Chica, C.; Via, A.; Gibson, T.J. Phospho.ELM: a database of phosphorylation sites--update 2008. *Nucleic Acids Res.*, **2008**, *36(Database issue)*, D240-244.

[42] Hornbeck, P.V., Chabra, I.; Kornhauser, J.M., Skrzypek, E.; Zhang, B. PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, **2004**, *4(6)*, 1551-1561.

[43] Wurgler-Murphy, S.M.; King, D.M.; Kennelly, P.J. The Phosphorylation Site Database: A guide to the serine-, threonine-, and/or tyrosine-phosphorylated proteins in prokaryotic organisms. *Proteomics*, **2004**, *4(6)*, 1562-1570.

[44] Keshava Prasad, T.S.; Goel, R.; Kandasamy, K.; Keerthikumar, S.; Kumar, S.; Mathivanan, S.; Telikicherla, D.; Raju, R.; Shafreen, B.; Venugopal, A.; Balakrishnan, L.; Marimuthu, A.; Banerjee, S.; Somanathan, D.S.; Sebastian, A.; Rani, S.; Ray, S.; Harrys Kishore, C.J.; Kanth, S.; Ahmed, M.; Kashyap, M.K.; Mohmood, R.; Ramachandra, Y.L.; Krishna, V.; Rahiman, B.A.; Mohan, S.; Ranganathan, P.; Ramabadran, S.; Chaerkady, R.; Pandey, A. Human Protein Reference Database--2009 update. *Nucleic Acids Res*., **2009**, *37(Database issue)*, D767-772.

[45] Li, H.; Xing, X.; Ding, G.; Li, Q.; Wang, C.; Xie, L.; Zeng, R.; Li, Y. SysPTM: a systematic resource for proteomic research on post-translational modifications. *Mol. Cell. Proteomics*, **2009**, *8(8)*, 1839-1849.

[46] Gnad, F.; Ren, S.; Cox, J.; Olsen, J.V.; Macek, B.; Oroshi, M.; Mann, M. PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.*, **2007**, *8(11)*, R250.

[47] Gnad, F.; de Godoy, L.M.; Cox, J.; Neuhauser, N.; Ren, S.; Olsen, J.V.; Mann, M. High-accuracy identification and bioinformatic analysis of *in vivo* protein phosphorylation sites in yeast. *Proteomics*, **2009**, *9(20)*, 4642-4652.

[48] Bodenmiller, B.; Campbell, D.; Gerrits, B.; Lam, H.; Jovanovic, M.; Picotti, P.; Schlapbach, R.; Aebersold, R. PhosphoPep--a database of protein phosphorylation sites in model organisms. *Nat. Biotechnol*., **2008**, *26(12)*, 1339-1340.

[49] Ovelleiro, D.; Carrascal, M.; Casas, V.; Abian, J. LymPHOS: design of a phosphosite database of primary human T cells. *Proteomics*, **2009**, *9(14)*, 3741-3751.

[50] Stark, C.; Su, T.C.; Breitkreutz, A.; Lourenco, P.; Dahabieh, M.; Breitkreutz, B.J.; Tyers, M.; Sadowski, I. PhosphoGRID: a database of experimentally verified *in vivo* protein phosphorylation sites from the budding yeast Saccharomyces cerevisiae. *Database (Oxford)*, **2010**, *2010*, bap026.

[51] Heazlewood, J.L.; Durek, P.; Hummel, J.; Selbig, J.; Weckwerth, W.; Walther, D.; Schulze, W.X. PhosPhAt: a database of phosphorylation sites in Arabidopsis thaliana and a plant-specific phosphorylation site predictor. *Nucleic Acids Res.*, **2008**, *36(Database issue)*, D1015-1021.

[52] Durek, P.; Schmidt, R.; Heazlewood, J.L.; Jones, A.; MacLean, D.; Nagel, A.; Kersten, B.; Schulze, W.X. PhosPhAt: the Arabidopsis thaliana phosphorylation site database. An update. *Nucleic Acids Res.*, **2010**, *38(Database issue)*, D828-834.

[53] Gao, J.; Agrawal, G.K.; Thelen, J.J.; Xu, D. P3DB: a plant protein phosphorylation database. *Nucleic Acids Res.*, **2009**, *37(Database issue)*, D960-962.

[54] Hummel, J.; Niemann, M.; Wienkoop, S.; Schulze, W.; Steinhauser, D.; Selbig, J.; Walther, D.; Weckwerth, W. ProMEX: a mass spectral reference database for proteins and protein phosphorylation sites. *BMC Bioinformatics*, **2007**, *8*, 216.

[55] Nuhse, T.S.; Stensballe, A.; Jensen, O.N.; Peck, S.C. Phosphoproteomics of the Arabidopsis plasma membrane and a new phosphorylation site database. *Plant Cell*, **2004**, *16(9)*, 2394-2405.

[56] Farriol-Mathis, N.; Garavelli, J.S.; Boeckmann, B.; Duvaud, S.; Gasteiger, E.; Gateau, A.; Veuthey, A.L.; Bairoch, A. Annotation of post-translational modifications in the Swiss-Prot knowledge base. *Proteomics*, **2004**, *4(6)*, 1537-1550.

[57] Lee, T.Y.; Huang, H.D.; Hung, J.H.; Huang, H.Y.; Yang, Y.S., Wang, T.H. dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res*., **2006**, *34(Database issue)*, D622-627.

[58] Yang, C.Y.; Chang, C.H.; Yu, Y.L.; Lin, T.C.; Lee, S.A.; Yen, C.C.; Yang, J.M.; Lai, J.M.; Hong, Y.R.; Tseng, T.L. Chao, K.M.; Huang, C.Y. PhosphoPOINT: a comprehensive human kinase interactome and phospho-protein database. *Bioinformatics*, **2008**, *24(16)*, i14-20.

[59] Zanzoni, A.; Ausiello, G.; Via, A.; Gherardini, P.F.; Helmer-Citterich, M. Phospho3D: a database of three-dimensional structures of protein phosphorylation sites. *Nucleic Acids Res*., **2007**, *35(Database issue)*, D229-231.

[60] Gong, W.; Zhou, D.; Ren, Y.; Wang, Y.; Zuo, Z.; Shen, Y.; Xiao, F.; Zhu, Q.; Hong, A.; Zhou, X.; Gao, X.; Li, T. PepCyber:P~PEP: a database of human protein protein interactions mediated by phosphoprotein-binding domains. *Nucleic Acids Res.*, **2008**, *36(Database issue)*, D679-683.

[61] Ryu, G.M.; Song, P.; Kim, K.W.; Oh, K.S.; Park, K.J.; Kim, J.H. Genome-wide analysis to predict protein sequence variations that change phosphorylation sites or their corresponding kinases. *Nucleic Acids Res*., **2009**, *37(4)*, 1297-1307.

[62] Ren, J.; Jiang, C.; Gao, X.; Liu, Z.; Yuan, Z.; Jin, C.; Wen, L.; Zhang, Z.; Xue, Y.; Yao, X. PhosSNP for systematic analysis of genetic polymorphisms that influence protein phosphorylation. *Mol. Cell. Proteomics*, **2010**, *9(4)*, 623-634.

[63] Blom, N.; Gammeltoft, S.; Brunak, S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.*, **1999**, *294(5)*, 1351-1362.

[64] MacDonald, J.A.; Mackey, A.J.; Pearson, W.R.; Haystead, T.A. A strategy for the rapid identification of phosphorylation sites in the phosphoproteome. *Mol. Cell. Proteomics*, 2002, *1(4)*, 314-322.

[65] Mackey, A.J.; Haystead, T.A.; Pearson, W.R. CRP: Cleavage of radiolabeled phosphoproteins. *Nucleic Acids Res.*, **2003**, *31(13)*, 3859-3861.

[66] Iakoucheva, L.M.; Radivojac, P.; Brown, C.J.; O'Connor, T.R.; Sikes, J.G.; Obradovic, Z.; Dunker, A.K. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res*., **2004**, *32(3)*, 1037-1049.

[67] Ingrell, C.R.; Miller, M.L.; Jensen, O.N.; Blom, N. NetPhosYeast: prediction of protein phosphorylation sites in yeast. *Bioinformatics*, **2007**, *23(7)*, 895-897.

[68] Miller, M.L.; Soufi, B.; Jers, C.; Blom, N.; Macek, B.; Mijakovic, I. NetPhosBac - a predictor for Ser/Thr phosphorylation sites in bacterial proteins. *Proteomics*, **2009**, *9(1)*, 116-125.

[69] Zhou, F.; Xue, Y.; Yao, X.; Xu, Y. A general user interface for prediction servers of proteins' post-translational modification sites. *Nat. Protoc*., **2006**, *1(3)*, 1318-1321.

[70] de Castro, E.; Sigrist, C.J.; Gattiker, A.; Bulliard, V.; Langendijk-Genevaux, P.S.; Gasteiger, E.; Bairoch, A.; Hulo, N. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res*., **2006**, *34(Web Server issue)*, W362-365.

[71] Hulo, N.; Bairoch, A.; Bulliard, V.; Cerutti, L.; Cuche, B.A.; de Castro, E.; Lachaize, C.; Langendijk-Genevaux, P.S.; Sigrist, C.J. The 20 years of PROSITE. *Nucleic Acids Res.*, **2008**, *36(Database issue)*, D245-249.

[72] Puntervoll, P.; Linding, R.; Gemund, C.; Chabanis-Davidson, S.; Mattingsdal, M.; Cameron, S.; Martin, D.M.; Ausiello, G.; Brannetti, B.; Costantini, A.; Ferre, F.; Maselli, V.; Via, A.; Cesareni, G.; Diella, F.; Superti-Furga, G.; Wyrwicz, L.; Ramu, C.; McGuigan, C.; Gudavalli, R.; Letunic, I.; Bork, P.; Rychlewski, L.; Kuster, B.; Helmer-Citterich, M.; Hunter, W.N.; Aasland, R.; Gibson, T.J. ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.*, **2003**, *31(13)*, 3625-3630.

[73] Balla, S.; Thapar, V.; Verma, S.; Luong, T.; Faghri, T.; Huang, C.H.; Rajasekaran, S.; del Campo, J.J.; Shinn, J.H.; Mohler, W.A.; Maciejewski, M.W.; Gryk, M.R.; Piccirillo, B., Schiller, S.R.; Schiller, M.R. Minimotif Miner: a tool for investigating protein function. *Nat. Methods*, **2006**, *3(3)*, 175-177.

[74] Rajasekaran, S.; Balla, S.; Gradie, P.; Gryk, M.R.; Kadaveru, K.; Kundeti, V.; Maciejewski, M.W.; Mi, T.; Rubino, N.; Vyas, J.; Schiller, M.R. Minimotif miner 2nd release: a database and web system for motif search. *Nucleic Acids Res.*, **2009**, *37(Database issue)*, D185-190.

[75] Amanchy, R.; Periaswamy, B.; Mathivanan, S.; Reddy, R.; Tattikota, S.G.; Pandey, A. A curated compendium of phosphorylation motifs. *Nat. Biotechnol.*, **2007**, *25(3)*, 285-286.

[76] Brinkworth, R.I.; Breinl, R.A.; Kobe, B. Structural basis and prediction of substrate specificity in protein serine/threonine kinases. *Proc. Natl. Acad. Sci. U. S. A.*, **2003**, *100(1)*, 74-79.

[77] Saunders, N.F.; Brinkworth, R.I.; Huber, T.; Kemp, B.E.; Kobe, B. Predikin and PredikinDB: a computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites. *BMC Bioinformatics*, **2008**, *9*, 245.

[78] Saunders, N.F.; Kobe, B. The Predikin webserver: improved prediction of protein kinase peptide specificity using structural information. *Nucleic Acids Res.*, **2008**, *36(Web Server issue)*, W286-290.

[79] Obenauer, J.C.; Cantley, L.C.; Yaffe, M.B. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.*, **2003**, *31(13)*, 3635-3641.

[80] Kim, J.H.; Lee, J.; Oh, B.; Kimm, K.; Koh, I. Prediction of phosphorylation sites using SVMs. *Bioinformatics*, **2004**, *20(17)*, 3179-3184.

[81] 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)*, W184-187.

[82] Zhou, F.F.; Xue, Y.; Chen, G.L.; Yao, X. GPS: a novel group-based phosphorylation predicting and scoring method. *Biochem. Biophys. Res. Commun.*, **2004**, *325(4)*, 1443-1448.

[83] Xue, Y.; Ren, J.; Gao, X.; Jin, C.; Wen, L.; Yao, X. GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Mol. Cell. Proteomics*, **2008**, *7(9)*, 1598-1608.

[84] Xue, Y.; Li, A.; Wang, L.; Feng, H.; Yao, X. PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinformatics*, **2006**, *7*, 163.

[85] Huang, H.D.; Lee, T.Y.; Tzeng, S.W.; Horng, J.T. KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res.*, **2005**, *33(Web Server issue)*, W226-229.

[86] Wong, Y.H.; Lee, T.Y.; Liang, H.K.; Huang, C.M.; Wang, T.Y.; Yang, Y.H.; Chu, C.H.; Huang H.D.; Ko, M.T.; Hwang, J.K. KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res.*, **2007**, *35(Web Server issue)*, W588-594.

[87] Li, T.; Li, F.; Zhang, X. Prediction of kinase-specific phosphorylation sites with sequence features by a log-odds ratio approach. *Proteins*, **2008**, *70(2)*, 404-414.

[88] Neuberger, G.; Schneider, G.; Eisenhaber, F. pkaPS: prediction of protein kinase A phosphorylation sites with the simplified kinase-substrate binding model. *Biol. Direct*, **2007**, *2*, 1.

[89] Dang, T.H.; Van Leemput, K.; Verschoren, A.; Laukens, K. Prediction of kinase-specific phosphorylation sites using conditional random fields. *Bioinformatics*, **2008**, *24(24)*, 2857-2864.

[90] Plewczynski, D.; Tkacz, A.; Wyrwicz, L.S.; Rychlewski, L. AutoMotif server: prediction of single residue post-translational modifications in proteins. *Bioinformatics*, **2005**, *21(10)*, 2525-2527.

[91] Plewczynski, D.; Tkacz A., Wyrwicz L. S., Rychlewski L., Ginalski K. AutoMotif Server for prediction of phosphorylation sites in proteins using support vector machine: 2007 update. *J. Mol. Model.*, **2008**, *14(1)*, 69-76.

[92] Jung, I.; Matsuyama, A.; Yoshida, M.; Kim, D. PostMod: sequence based prediction of kinase-specific phosphorylation sites with indirect relationship. *BMC Bioinformatics*, **2010**, *11 (Suppl 1)*, S10.

[93] Huang, H.; Li, L.; Wu, C.; Schibli, D.; Colwill, K.; Ma, S.; Li, C.; Roy, P.; Ho, K.; Songyang, Z.; Pawson, T.; Gao, Y.; Li, S.S. Defining the specificity space of the human SRC homology 2 domain. *Mol. Cell. Proteomics*, **2008**, *7(4)*, 768-784.

[94] Li, L.; Wu, C.; Huang, H.; Zhang, K.; Gan, J.; Li, S.S. Prediction of phosphotyrosine signaling networks using a scoring matrix-assisted ligand identification approach. *Nucleic Acids Res.*, **2008**, *36(10)*, 3263-3273.

[95] Wan, J.; Kang, S.; Tang, C.; Yan, J.; Ren, Y.; Liu, J.; Gao, X.; Banerjee, A.; Ellis, L.B.; Li, T. Meta-prediction of phosphorylation sites with weighted voting and restricted grid search parameter selection. *Nucleic Acids Res.*, **2008**, *36(4)*, e22.

[96] Miller, M.L.; Jensen, L.J.; Diella, F.; Jorgensen, C.; Tinti, M.; Li, L.; Hsiung, M.; Parker, S.A.; Bordeaux, J.; Sicheritz-Ponten, T.; Olhovsky, M.; Pasculescu, A.; Alexander, J.; Knapp, S.; Blom, N.; Bork, P.; Li, S.; Cesareni, G.; Pawson, T.; Turk, B.E.; Yaffe, M.B.; Brunak, S.; Linding, R. Linear motif atlas for phosphorylation-dependent signaling. *Sci. Signal.*, **2008**, *1(35)*, ra2.

[97] Schwartz, D.; Gygi, S.P. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat. Biotechnol.*, **2005**, *23(11)*, 1391-1398.

[98] Schwartz, D.; Chou, M.F.; Church, G.M. Predicting protein post-translational modifications using meta-analysis of proteome scale data sets. *Mol. Cell. Proteomics*, **2009**, *8(2)*, 365-379.

[99] Ritz, A.; Shakhnarovich, G.; Salomon, A.R.; Raphael, B.J. Discovery of phosphorylation motif mixtures in phosphoproteomics data. *Bioinformatics*, **2009**, *25(1)*, 14-21.

[100] Wang, Y.; Klemke, R.L. PhosphoBlast, a computational tool for comparing phosphoprotein signatures among large datasets. *Mol. Cell. Proteomics*, **2008**, *7(1)*, 145-162.

[101] Hu, Z.Z.; Narayanaswamy, M.; Ravikumar, K.E.; Vijay-Shanker, K.; Wu, C.H. Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics*, **2005**, *21(11)*, 2759-2765.

[102] Yuan, X.; Hu, Z.Z.; Wu, H.T.; Torii, M.; Narayanaswamy, M.; Ravikumar, K.E.; Vijay-Shanker, K.; Wu, C.H. An online literature mining tool for protein phosphorylation. *Bioinformatics*, **2006**, *22(13)*, 1668-1669.

[103] Lachmann, A.; Ma'ayan, A. KEA: kinase enrichment analysis. *Bioinformatics*, **2009**, *25(5)*, 684-686.

[104] Ren, J.; Wen, L.; Gao, X.; Jin, C.; Xue, Y.; Yao, X. DOG 1.0: illustrator of protein domain structures. *Cell Res.*, **2009**, *19(2)*, 271-273.

[105] Lahiry, P.; Torkamani, A.; Schork, N.J.; Hegele, R.A. Kinase mutations in human disease: interpreting genotype-phenotype relationships. *Nat. Rev. Genet.*, **2010**, *11(1)*, 60-74.

[106] Chou, K.C.; Watenpaugh, K.D.; Heinrikson, R.L. A model of the complex between cyclin-dependent kinase 5 and the activation domain of neuronal Cdk5 activator. *Biochem. Biophys. Res. Commun.*, **1999**, *259(2)*, 420-428.

[107] Zhang, J.; Luan, C.H.; Chou, K.C.; Johnson, G.V. Identification of the N-terminal functional domains of Cdk5 by molecular truncation and computer modeling. *Proteins*, **2002**, *48(3)*, 447-453.

Received: December 08, 2009 **Revised: May 17, 2010** **Accepted: May 17, 2010**

PMID: 20491621