Genomics 92 (2008) 457–463

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Proteome-wide prediction of PKA phosphorylation sites in eukaryotic kingdom

Xinjiao Gao, Changjiang Jin, Jian Ren, Xuebiao Yao, Yu Xue *

Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science and Technology of China, Hefei, Anhui 230027, China

ARTICLE INFO

Article history:
Received 26 March 2008
Accepted 27 August 2008
Available online 10 October 2008

Keywords:
Phosphorylation
PKA
Large-scale prediction
False positive rate

ABSTRACT

Protein phosphorylation is one of the most essential post-translational modifications (PTMs), and orchestrates a variety of cellular functions and processes. Besides experimental studies, numerous computational predictors implemented in various algorithms have been developed for phosphorylation sites prediction. However, large-scale predictions of kinase-specific phosphorylation sites have not been successfully pursued and remained to be a great challenge. In this work, we raised a "kiss farewell" model and conducted a high-throughput prediction of cAMP-dependent kinase (PKA) phosphorylation sites. Since a protein kinase (PK) should at least "kiss" its substrates and then run away, we proposed a PKA-binding protein to be a potential PKA substrate if at least one PKA site was predicted. To improve the prediction specificity, we reduced false positive rate (FPR) less than 1% when the cut-off value was set as 4. Successfully, we predicted 1387, 630, 568 and 912 potential PKA sites from 410, 217, 173 and 260 PKA-interacting proteins in *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens*, respectively. Most of these potential phosphorylation sites remained to be experimentally verified. In addition, we detected two sites in one of PKA regulatory subunits to be conserved in eukaryotes as potentially ancient regulatory signals. Our prediction results provide an excellent resource for delineating PKA-mediated signaling pathways and their system integration underlying cellular dynamics and plasticity.

Introduction

Phosphorylation is one of the most ubiquitous and important post-translational modifications (PTMs) of proteins, and implicated in almost all kinds of cellular processes and pathways [1,2]. In eukaryotes, phosphorylation is carried out by numerous protein kinases (PKs), which are members of kinome or kinase superfamily [3,4]. Each PK recognizes distinct S/T or Y residues in protein sequences and only modifies a defined subset of substrates specifically, to ensure signaling fidelity. Thus, identification of phosphorylated substrates with their kinase-specific phosphorylation sites is the foundation for understanding the molecular mechanism of phosphorylation dynamics.

Besides experimental studies, various computational approaches have been extensively employed and achieved great successes for phosphorylation sites prediction. Numerous predictors have been developed mainly for two purposes: prediction of general or kinase-specific phosphorylation sites. For the former question, NetPhos [5] and DisPhos [6] were constructed. And for the latter, we and other researchers have contributed great efforts and developed several online tools, e.g., GPS [7,8], PPSP [9], ScanSite [10], KinasePhos [11], PredPhospho [12], NetPhosK [13], Predikin [14] and pkaPS [15], etc. Recently, NetPhosYeast was constructed to predict phosphorylation sites in yeast specifically, as the first organism-specific predictor [16].

Given ~10 phosphorylation predictors in hand, a great challenge is emerging that how we can make sense for large-scale predictions of phosphorylation sites in proteome-wide level. To our knowledge, only a few articles have addressed the problem [15,17,18]. With Predikin, Brinkworth et al. predicted cognate PKs for 383 un-annotated phosphorylation sites of 216 peptide sequences in yeast [17]. Combined with a motif-based strategy and protein association information, Linding et al. developed NetworKIN to predict associated PKs for 7143 un-annotated phosphorylation sites in Phospho.ELM database, and constructed *in vivo* phosphorylation networks [18]. In addition, Neuberger et al. used pkaPS directly to predict PKA (cAMP-dependent kinase) sites in human proteome [15]. But there were only 4860 of 40,887 (11.9%) human proteins predicted not to contain a single potential PKA site [15]. Thus, large-scale predictions of kinase-specific phosphorylation sites are still very difficult and need more elaboration. For such analyses, two essential points should be addressed. The first one is control of false positive rate (FPR). It's expected that *bona fide* phosphorylation sites are only a small proportion of total S/T or Y residues, usually <1:10. Thus, even a small FPR of 10% will generate more than 50% false positive hits in total prediction results. The second one is recognition specificity between a protein kinase with its substrates. As described previously, only a short peptide flanking of a site is not sufficient for providing full specificity for a PK modification *in vivo* [19,20]. Numerous mechanisms have also been suggested to contribute additional specificities for PKs recognition, such as subcellular co-localization of PKs with their substrates, co-complex or interacting directly [19–21]. Thus, *in*

* Corresponding author.
E-mail address: xueyu@ustc.edu.cn (Y. Xue).

0888-7543/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.08.013

vivo a PK should at least "kiss" its substrates then say farewell by direct or indirect interactions. Furthermore, it was proposed that the accurate specificity of PKA substrates phosphorylation in higher eukaryotes could be achieved by compartmentalization of PKA, in great measure mediated by A-kinase anchoring proteins (AKAPs) [22–24]. However, the AKAPs are not existed in yeast [24]. In this regard, some other mechanisms of providing specificities still remain to be elucidated.

In this article, we performed a large-scale prediction of cAMP-dependent kinase (PKA) phosphorylation sites from its binding proteins in four eukaryotic proteomes. Two hypotheses were established. First, we supposed that different PKA isoforms in eukaryotes recognize similar peptide motifs/profiles for modifications. Secondly, we proposed a "kiss farewell" model that PKA should at least kiss its substrates by directly physical interaction or forming a co-complex. Although the kisses might be transient with different degrees of affinity, there are still a large proportion of the interactions between PKA with its substrates, which could be detected in standard protein–protein interaction screenings. The PKA-interacting proteins were retrieved from both of experimental verified and predicted protein–protein interaction (PPI) databases. Then the GPS software [7,8] was employed to predict PKA-specific sites in these proteins. As lack of a "gold standard" negative data set to precisely evaluate the false positive rate (FPR), we developed a simple method to estimate the theoretical maximum of FPR. And the FPR was reduced lower than 1% in this work. Successfully, we predicted 410, 217, 173 and 260 potential PKA substrates with 1387, 630, 568 and 912 potential PKA sites in *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens*, respectively. These data sets serve as a good start point for further experimentation.

Results

Analysis of PKA proteins/isoforms in four eukaryotic organisms

PKA was firstly identified in rabbit skeletal muscle [25], conserved in eukaryotes as a serine/threonine kinase sub-family, activated by cAMP and plays important roles in a large number of cellular processes [26]. Currently, there were 337 PKA-specific phosphorylation sites from 210 substrates experimentally identified *in vivo* or *in vitro* (Phospho.ELM, ver 6.0) [27]. Despite four decades efforts on identifying PKA targets with their sites, most of the studies were focused on mammalians. For example, there were only one and three PKA substrates with sites experimentally identified in *S. cerevisiae* and *D. melanogaster*, respectively. Especially, although a PKA gene of kin-1 in *C. elegans* was discovered [28], there were no substrates of kin-1 reported. Recently, Ptacek et al. used the proteome chip technology to carry out a large-scale survey to identify *in vitro* PK-specific substrates in *S. cerevisiae* [29]. Totally, they identified 256, 29 and 79 *in vitro* substrates for TPK1, TPK2 and TPK3. Although the results still remained to be experimentally verified *in vivo*, and the exact PK-specific phosphorylation sites were not mapped, this work provided useful information for further experimental consideration. Taken together, identification of PKA substrates with their sites in eukaryotes still remain to be a great challenge.

From the annotation and classification of the kinase.com database, we obtained 3, 1, 4, 5 distinct components of PKA sub-family in *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens*, respectively (Table 1). Then these protein sequences were multi-aligned by MUSCLE 3.6 with default parameters [30]. The phylogenetic tree for PKA sub-family was constructed by MEGA 3.1 [31], with Neighbor-Joining method with bootstrap test (Fig. 1). By BLAST searching, the best hit of ScTPK1 in *E. Coli*, ORF708 (Q47592) was chosen as the outgroup. Interestingly, ScTPK1, ScTPK2 and ScTPK3 of *S. cerevisiae* and HsPRKACA, HsPRKACB and HsPRKACG of *H. sapiens* are separated after speciation, respectively. Also, HsPRKX was reported to play important roles in morphogenesis and the functions are not shared with other PKA components [32]. In our analysis, HsPRKX, HsPRKY and DmPKA-C3 were clustered into a distinct branch. Thus, proteins of PKA sub-family might have similar but distinctive functions. Despite the functional similarity and diversity of PKA sub-family, we hypothesized that all PKA proteins recognize similar peptide motifs/profiles for modification. Thus, computational predictor constructed for PKA-specific phosphorylation sites could be generally applied for eukaryotes.

Table 1
Different PKA (protein kinase A) genes in four eukaryotic organisms

	Kinase.com[a]	UniProt[b]	Exp. PPI[c]	String PPI[d]	Known sub.[e]
S. cerevisiae	TPK1	KAPA_YEAST	35	148	256
	TPK2	KAPB_YEAST	31	200	29
	TPK3	KAPC_YEAST	21	76	79
C. elegans	kin-1	KAPC_CAEEL	2	418	
D. melanogaster	Pka-C1	KAPC_DROME	12	72	
	Pka-C2	KDC1_DROME	0	57	
	PKA-C3	KDC2_DROME	4	144	
	CG12069	Q9VA47_DROME	0	20	
H. sapiens	PRKACA	Q32P54_HUMAN	154	193	1
	PRKACB	KAPCB_HUMAN	1	102	
	PRKACG	Q5VZ02_HUMAN	0	109	
	PRKX	PRKX_HUMAN	5	55	
	PRKY	PRKY_HUMAN	0	64	

[a] PKA names in kinase.com database.
[b] Accession numbers of PKA proteins in Swiss-Prot and TrEMBL (UniProt) database.
[c] The number of experimentally verified PKA-binding proteins.
[d] The number of predicted PKA-interacting proteins.
[e] The experimentally verified substrates for PKA proteins.

PKA-interacting proteins: experimental verified vs. predicted

Frankly, prediction of the exhaustive set of PKA substrates in a proteome will not make much sense, because the PKA recognition sequences may appear too frequently in proteins. For example, Neuberger et al. used pkaPS directly to predict PKA substrates in human proteome [15]. But there were 36,027 of 40,887 (88.1%) human proteins predicted as potential substrates with at least one PKA site [15]. Here, we adopted sufficient conditions for predicting a reliable data set. We proposed a simple "kiss farewell" model that the PKA should at least kiss its targets and then say farewell for modification. And the PKA must physically bind with its substrates or form a co-complex by direct or indirect interaction. Although such an interaction might be transient, and the binding affinity might also be weak, we believed that there were still a significant proportion of the interactions between PKA with its substrates, which could be detected in standard protein–protein interaction screenings. Thus, a PKA-interacting protein is highly probable to be a real substrate of PKA, if at least one PKA site is predicted with high confidence.

To obtain a comprehensive protein–protein interaction (PPI) map for each organism, several public PPI databases were employed, including DIP [33], BioGrid [34], Grid [35], MINT [36], BIND [37], Wormbase [38] and HPRD [39]. Since most of entries in BIND for *S. cerevisiae* were covered in other databases, we only used DIP, BioGrid, Grid and MINT to retrieve yeast PPI data. Also, Wormbase and HPRD only contained PPI data for *C. elegans* and *H. sapiens*, respectively. Since the protein names in the databases are different, we mapped all these proteins into Swiss-Prot and TrEMBL (UniProt) database. Then all PPI data were integrated into a non-redundant data set for each species, with the number of 52,987, 5,959, 30,558 and 51,529 pairs in *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens*, respectively (Table 2). Furthermore, a pre-calculated PPI database, STRING, was also downloaded directly [40]. Finally, the sequences of both of experimentally verified and predicted PKA-binding proteins were retrieved (Table 1).

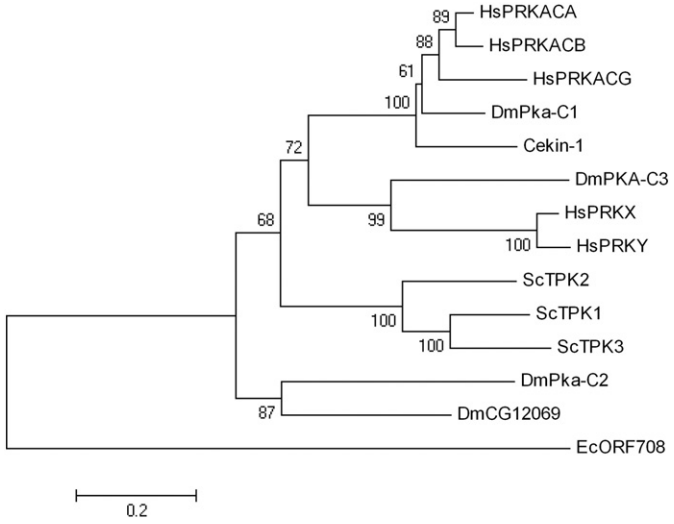

Fig. 1. Phylogenetic analysis of PKA sub-family. By BLAST searching, the best hit of ScTPK1 in *E. Coli*, ORF708 (Q47592) was chosen as the outgroup.

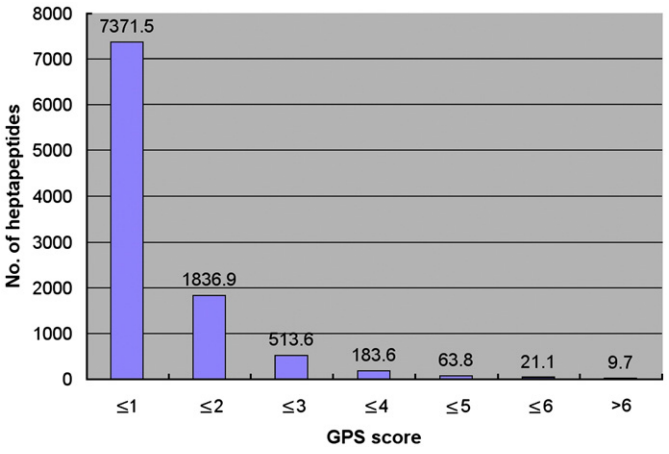

Fig. 2. The distributions of GPS scores of PKA on randomly generated heptapeptides (XXX-S/T-XXX, X is a random residue). Totally 10,000 random heptapeptides were generated and the calculation was repeated for twenty times. Then the average numbers were computed and diagramed below.

Clearly, most of PKA-binding proteins were experimentally identified in *H. sapiens*. Especially, the alpha-catalytic subunit of PKA in human, PRKACA, attracted the most studies. Totally, there were 154 PRKACA-binding proteins identified, while only one PRKACB-interacting protein was discovered (Table 1). And PRKACG-binding proteins still remain to be reported. Most of PKA-binding proteins in *S. cerevisiae* (TPK1, 2, 3) were identified from large-scale experiments. And only three substrates of PKA were identified in *S. cerevisiae*. Thus, these yeast PKA-binding proteins would be an excellent data set for PKA substrates identification. Again, since PRKACA was proved to interact with more than 150 proteins, it could be estimated that PRKACB and PRKACG also have numerous binding proteins. In this regard, even predicted PPIs will be also important for large-scale prediction of PKA substrates.

Control and reduction of false positive rate (FPR)

Control of false positive prediction is the key point in large-scale predictions of kinase-specific phosphorylation sites. The false positive rate (FPR) is the proportion of negative sites that are erroneously predicted as positive hits. Actually, the real PKA phosphorylation sites are only a very small part of all S/T residues in proteins. For PKA, the positive sites vs. the negative sites in current data set is <1:56 (337 vs. 18,964). Thus, even a predictor with sensitivity (*Sn*) of 100% with a FPR of 10% will generate ~85% of false positive hits in final predictions. In this regard, the FPR should be controlled and reduced greatly.

Given a data set containing all non-phosphorylation sites, the FPR could be easily computed. However, precise calculation of FPR is unavailable due to lack of a "gold standard" negative data set. Previously, we and other researchers simply took all verified sites as positive data and regarded all un-identified sites as negative sites [7–9,11,12]. However, this procedure is too stringent since many real sites are not discovered rather than negative hits. Thus, the prediction performances of most of predictors were under-estimated.

In this article, we developed an alternatively method to estimate the FPR theoretically. We randomly generated 10,000 heptapeptides as XXX-S/T-XXX, which X is a random amino acid and the centered residue is S or T. Then we used GPS to score these peptides. The distribution of GPS score of PKA on these peptides were shown in Fig. 2. Although there were a few sites to be real hits, its proportion would be very small. With this approach, we estimated the theoretical maximum of FPR for PKA. The procedure was repeated twenty times and the average FPR value was calculated. The theoretical FPRs with cut-off values of 4, 2.4 and 1 were estimated as 0.94%, 5.18% and 26.09%, respectively. Thus, we selected the cut-off value of 4 for large-scale predictions.

Large-scale prediction of PKA phosphorylation sites for its binding proteins

In this work, we directly employed GPS 1.10 with the cut-off value of 4 (high threshold) to predict PKA sites in its interacting proteins. For *S. cerevisiae*, we also included a data set from a high-throughput experiment [29], to make the prediction more integrated.

Then the theoretically maximal false positive hits for each data set was calculated (Table 3). Firstly, all serine (S) and threonine (T) residues in PKA-binding proteins were accounted. If all of these residues are real negative sites, the false positive hits could be

Table 2
Experimentally verified vs. predicted PPI (protein–protein interaction) data sets in four species

	S. cerevisiae	*C. elegans*	*D. melanogaster*	*H. sapiens*
DIP	44,156	4028	22,819	1397
BioGrid	81,775	4433	32,817	38,217
Grid	23,952	4453	28,406	
MINT	14,408	4710	20,780	8127
BIND[a]		4919	22,462	10,756
Wormbase		4090		
HPRD				33,412
Non-redundant[b]	52,987	5959	30,558	51,529
STRING	200,974	230,509	132,923	690,143

[a] PPI data of *S. cerevisiae* in BIND database were not used.
[b] Non-redundant data for experimentally verified PPIs.

Table 3
Theoretically minimal precisions (*Pr*) of GPS (Group-based Phosphorylation Scoring) on PKA-binding proteins in four organisms

	Total[a]	Predicted[b]	Est. false[c]	Pr[d]
S. cerevisiae	47,755	1387	449	67.63%
C. elegans	28,467	630	268	57.46%
D. melanogaster	24,179	568	227	60.04%
H. sapiens	31,093	912	292	67.98%

The false positive rate (FPR) of GPS 1.10 is 0.94% under the high threshold (cut-off value of 4). Then given total S and T residues, the theoretically maximal false positive hit and minimal precision (Pr) for each organism was calculated, respectively.
[a] Total S and T residues.
[b] S/T residues predicted as positive hits.
[c] Estimated false positive hits.
[d] Theoretically minimal precisions.

Table 4
There were 410, 217, 173 and 260 potential PKA substrates predicted in *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens*, respectively

	Proteome-wide[a]		Exp. PPI		STRING PPI		Total		*p*-value[b]
	Total[c]	*Pre.*[d]	*Total*	*Pre.*	*Total*	*Pre.*	*Total*	*Pre.*	
S. cerevisiae	7565	4104	58	48	267	182	569	410	2.9E−17
C. elegans	23,283	12,854	2	2	418	217	418	217	0.9
D. melanogaster	27,867	16,325	16	12	241	172	242	173	2.3E−05
H. sapiens	70,646	34,819	157	130	299	223	348	260	3.8E−22

[a] Proteome-wide, we carried out genome-wide predictions of PKA sites for the four proteomes directly.
[b] *p*-value, we used the Fisher's Exact Test to evaluate that the predicted PKA targets were significantly enriched in *S. cerevisiae*, *D. melanogaster* and *H. sapiens* (p-value $\ll 0.05$) but not in *C. elegans* (p-value > 0.05).
[c] *Total*, Total proteins in a data set.
[d] *Pre.*, the number of predicted substrates.

calculated (0.94%*Number of all S/T residues). Since not all of S/T residues in PKA-interacting are negative sites, the false positive hits will be over-estimated. However, although the false positive rate was over-estimated, our analysis still generated satisfying results. For example, there were 31,093 S/T residues contained in human PKA-binding proteins. And GPS predicted out 912 sites as positive hits. With the maximal FPR of 0.94% at cut-off value of 4, there were at most 292 (31,093×0.94%) sites to be false positive hits. Then we could calculate the minimal precision in this instance as 67.98% (620/912). Also, we defined theoretically minimal precision (Pr) as below:

$$\mathrm{Pr} = \frac{M - N^{*}FPR}{M}$$

Here, N is the total number of S/T sites for prediction; M is the number of predicted sites by GPS. When the cut-off value was taken as 4, the precision (Pr) values were 67.63%, 57.46%, 60.04% and 67.98% for *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens*, respectively (Table 3).

Successfully, we predicted 410, 217, 173 and 260 potential PKA substrates with 1387, 630, 568 and 912 potential PKA sites in *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens*, respectively (Table 4). And the detailed results of the large-scale predictions were shown in Supplementary Table1–4. In this analysis, two questions should be addressed. The first problem was whether our approach could significantly enrich the prediction set of proteins with PKA sites. Firstly, we carried out proteome-wide predictions of PKA sites in the four proteomes directly (Table 4). Then we used the Fisher's Exact Test (http://www.langsrud.com/fisher.htm) to evaluate whether the predicted PKA substrates would be statistically present in PKA-binding proteins. In the results, we found that the predicted PKA targets were significantly enriched in *S. cerevisiae*, *D. melanogaster* and *H. sapiens* (p-value $\ll 0.05$) but not in *C. elegans* (p-value > 0.05). The results could be attributed to the different qualities of PPI data sets. Obviously, the quality of experimental PPI data was much better than predicted data. In *H. sapiens*, there were 157 of 348 potential PKA-binding proteins to be experimentally verified (p-value = 3.8E−22). But in *C. elegans*, there were only 2 proteins experimentally identified (p-value = 0.9) (Table 4). In this regard, further experimental progresses on PKA-interacting proteins verification will be a great help for PKA substrates identification. The second question was how many known PKA substrates had been determined as PKA-interacting proteins by experiments and predictions. Since most of PKA substrates were identified in mammalians, especially in human, we carried out a simple analysis to address the problem. In Phospho.ELM (ver 6.0) [27], there were 140 PKA substrates identified in *H. sapiens*, while 88 (62.86%) of them were PKA-binding proteins. And GPS could predict 80 (57.14%) of 140 human PKA targets as positive hits with at least one PKA site. Taken together, although several real PKA substrates were missed, our method could still generate satisfying results for further experimental verifications.

Table 5
Several predicted PKA substrates with their sites are shown

UniProt (SRS)	String PKA[a]	Exp. PKA[b]	GPS (PKA)[c]	Exp. site[d]	PMID
Experimentally verified PKA substrates[e]					
GMFB_HUMAN	PRKACA	PRKACA	T26	T26;S82	9030586
S4A4_HUMAN	PRKACA	PRKACA	S177;T254;S1026	S1026	11744745
HDAC8_HUMAN	PRKACA	PRKACA	S39;S83	S39	14701748
NEB2_HUMAN	PRKACA	PRKACA	S100;S754	S94;S100	12417592
KAPR_YEAST	TPK1;TPK2;TPK3	TPK1;TPK2;TPK3	S73;T143;S144;T379	S144	92065884
Experimentally verified PKA-interacting proteins[f]					
GFAP_HUMAN	PRKACA	PRKACA	T7;S38;S68;S393		
BCA3_HUMAN	PRKACA	PRKACA	T143;S144		
PHX2A_HUMAN	PRKACA	PRKACA	S153		
KAPR_CAEEL	Kin-1	Kin-1	S122		
CHK1_YEAST	TPK1;TPK2	TPK2	T271;S286		
Predicted PKA-binding proteins[g]					
RASK_HUMAN	PRKACG		S172		
CNR1_HUMAN	PRKACA		S185		
KAPR2_DROME	Pka-C1;PKA-C3		S51;S84;S323		
GPA16_CAEEL	Kin-1		S15		
BUB2_YEAST	TPK2		S98		

[a] The PKA isoform interacts the protein in STRING database.
[b] The PKA isoform experimentally interacts the protein.
[c] Predicted PKA sites.
[d] Experimentally verified PKA sites.
[e] The five predicted PKA substrates were verified as real substrates previously, and these proteins also were experimentally verified and predicted as PKA-binding proteins.
[f] The five proteins were both of experimentally verified and predicted as PKA-binding proteins, but not verified as PKA substrates.
[g] The five proteins were only predicted as PKA-binding proteins.

Also, we randomly selected several predicted PKA substrates to depict our analysis (Table 5). Five experimentally verified PKA substrates were picked out, including GMFB (GMFB_HUMAN), SLC4A4 (S4A4_HUMAN), HDAC8 (HDAC8_HUMAN), PPP1R9B (NEB2_HUMAN) and BCY1 (KAPR_YEAST). These PKA substrates were both of experimentally verified and predicted as PKA-interacting proteins, with diverse functions. For example, HDAC8 is a member of human class I Histone deacetylases, and its enzymatic activity is negatively regulated by PKA phosphorylation [41]. Also, we listed five predicted PKA substrates for experimentally verified or predicted PKA-binding proteins, separately. The GFAP (GFAP_HUMAN), C11orf17 (BCA3_HUMAN), PHOX2A (PHX2A_HUMAN), kin-2 (KAPR_CAEEL) and CHK1 (CHK1_YEAST) were experimentally verified and also predicted as PKA-interacting proteins (Table 5). Mutations of GFAP, a class-III intermediate filament, are involved in Alexander disease [42]. And CHK1, as a protein kinase, plays an important role as a checkpoint of DNA damage to arrest cell cycle progression [43]. These proteins physically interact with PKA and were predicted as potential PKA substrates. And experimentally verification of PKA sites in these proteins will be important to elucidate their biological dynamics and functions under phosphorylation regulation. Again, five predicted PKA-binding proteins were also shown, including KRAS (RASK_HUMAN), CNR1 (CNR1_HUMAN), Pka-R2 (KAPR2_DROME), gpa-16 (GPA16_CAEEL) and Bub2 (BUB2_YEAST) (Table 5). These proteins are also important in various cellular processes. For example, Bub2 is a TBC (Tre-2/Bub2/Cdc16) domain-containing protein, as a GTPase activator and spindle checkpoint during mitosis [44]. And our predictions might be useful for further experimental design.

Two conserved PKA phosphorylation signals across eukaryotes

An interesting question is emerging: Are there any potential phosphorylation signals conserved across eukaryotes? From conserved PKA-binding proteins, the conserved and predicted PKA substrates were retrieved and shown in Supplementary Table 5. Only one of PKA regulatory subunits is conserved in four organisms, with the name of ScBCY1 (KAPR_YEAST), Cekin-2 (KAPR_CAEEL), DmPka-R2 (KAPR2_DROME), and HsPRKAR2A (KAP2_HUMAN) in *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens*, respectively.

In *S. cerevisiae*, ScBCY1 was proved to be a PKA substrate with the verified site of S145. Also, ScBCY1 was experimentally verified and predicted as an interacting protein of TPK1, TPK2 and TPK3. And in *C. elegans*, Cekin-2 was also proved to interact with PKA/kin-1. Again, in *H. sapiens*, HsPRKAR2A was also verified as a PKA target with the site of S98. It was experimentally verified to interact with PRKX, but prediction PPI data proposed it could interact with all members of PKA sub-family. Finally, in *D. melanogaster*, only prediction PPI data suggest it could interact with fly Pka-C1 and PKA-C3.

The protein sequences of four PKA regulatory subunits were retrieved and predicted by GPS, with the cut-off value of 4. Successfully, several potential PKA sites were predicted, including S74, T144, S145 and T380 of ScBCY1, S122 of Cekin-2, S51, S84 and S323 of DmPka-R2, and S98 and S349 of HsPRKAR2A. Both of two experimental verified sites were predicted correctly by GPS. And newly predicted sites are useful for further experimental verifications.

We aligned the four proteins with MUSCLE 3.6 [30]. And conserved PKA sites were shown in Table 6 and Fig. 3. The first potentially conserved phosphorylation signal was S145, S84 and S99 of yeast, fruit fly and human separately, while it was not conserved in nematode. We also checked the PKA regulatory subunit in other numerous species, while the site was conserved and predicted as a PKA site (GPS score>4) (data not shown). Thus the phosphorylation signal is potentially conserved across eukaryotes expect nematodes. And the second potential phosphorylation site is T380, T323, S323 and S350 in *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens*, respectively. Although T323 of Cekin-2 was predicted with score of 3.111, which was lower than the threshold, we proposed it to be a potentially conserved PKA site (Fig. 3). Again, we also found the site was conserved in other species with the GPS score greater than 4. Thus, the PKA regulatory subunit might be co-evolved with PKA as its conserved substrate and interacting protein. Our analysis proposed that PKA sites on the PKA regulatory subunit were not well conserved in nematodes. However, there was only one PKA gene (kin-1) in *C. elegans*, while other organisms have multi-PKA genes. Thus, the recognition profile of nematode PKA/kin-1 might be slightly different against other organisms. Also, several real PKA sites in nematode may diminish during evolution, and several newly generated sites still retain original functions. In this regard, experimental identification of PKA/kin-1 sites in nematodes will be a great help for understanding similarity and specificity of PKA recognition and modification during evolution.

Table 6
Two potentially conserved phosphorylation sites with their GPS scores

	UniProt (SRS)	Position 1	GPS score	Position 2	GPS score
ScBCY1	KAPR_YEAST	S145	8.672	T380	4.706
Cekin-2	KAPR_CAEEL			T323	3.111
DmPka-R2	KAPR2_DROME	S84	7.150	S323	4.300
HsPRKAR2A	KAP2_HUMAN	S99	7.372	S350	4.156

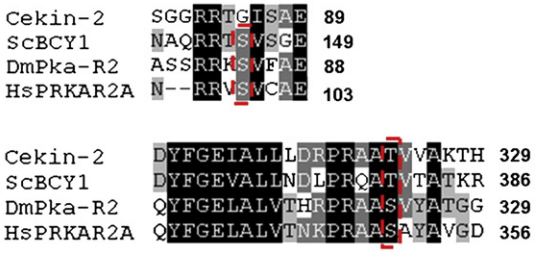

Fig. 3. Two potentially conserved phosphorylation signals are shown. The first one is S145 of *S. cerevisiae*, S84 of *D. melanogaster* and S99 of *H. sapien*, but not conserved in *C. elegans*. The second one is T380, T323, S323 and S350 in *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapien*, respectively. The yeast S145 and human S98 were experimentally verified to be phosphorylated by ScBCY1 and HsPRKAR2A, respectively.

Discussion

For prediction of kinase-specific phosphorylation sites, a widely-adopted hypothesis is that a kinase could recognize specific sequence profiles/motifs/patterns around phosphorylation sites for modification in a substrate [7,8,13]. If a given protein is really phosphorylated by a kinase, the current predicters, including GPS [7,8], PPSP [9], ScanSite [10], KinasePhos [11], PredPhospho [12], NetPhosK [13], Predikin [14] and pkaPS [15], could predict potential kinase-specific sites in the protein. However, only the short phospho-peptide could not provide full specificity for a kinase recognition *in vivo*. Previously, researches proposed various mechanisms, such as subcellular co-localization of kinases with their substrates, co-complex or physical interaction [19–21]. However, these additional mechanisms were not included in the current predictors. In those cases where known interacting partners are not PKA substrates, the current predictors will still predict potential PKA-specific phosphorylation sites in these proteins. In this regard, directly prediction of PKA substrates in a proteome will not make much sense, because the PKA recognition sequences may appear too frequently in proteins.

In this article, we performed a large-scale prediction of PKA phosphorylation sites from its interacting proteins in four eukaryotes. We hypothesized that different PKA genes conserved in eukaryotes recognize similar consensus motif/profile for modification. Also, we raised a simple “kiss farewell” model that the PKA should at least kiss its targets and then say farewell for modification, by direct physical binding or indirectly forming a co-complex mediated by other linker

proteins. Although the interactions between PKA and its substrates might be transient or weak, our approach could still significantly enrich the prediction set of proteins with PKA sites, at least in *S. cerevisiae*, *D. melanogaster* and *H. sapiens*. Based on these two hypotheses, we employed GPS 1.10 directly. The theoretically maximal false positive rate (FPR) was controlled and reduced as 0.94%, when the cut-off value is taken as 4. Successfully, we predicted 410, 217, 173 and 260 potential PKA substrates with 1387, 630, 568 and 912 potential PKA sites in *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens*, respectively. Most of the potential substrates with their sites still remain to be experimentally verified. Our prediction results provide a useful data set for further experimental verification. Furthermore, our analysis also provides a linkage between current studies of PPI and PTM to be useful for further PTM researches.

Materials and methods

Preparation of benchmark sequences for four eukaryotic proteomes

There are many public databases of protein sequences, while the protein names or accession numbers are various without standardization. Although other databases could be chosen, in this work, we used Swiss-Prot and TrEMBL (UniProt) database as the benchmark (May, 2006). From Sequence Retrieval System (SRS5) at ExPASy website (http://www.expasy.ch/srs5/), we retrieved protein sequences for four eukaryotic organisms. Totally, there were 7565, 23,283, 27,867 and 70,646 protein sequences obtained from *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens*, separately.

Collecting all members of PKA sub-family with their protein sequences

The kinase.com database identified and collected most of the protein kinases in human, mouse and other eukaryotic organisms [3,4]. Then these PKs were classified into many sub-families [3,4]. The PKA sub-family is composed of several PKA paralogs/isoforms to be conserved and important in various cellular processes. Totally, there were 3, 1, 4 and 5 PKA members found in *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens*, respectively (Table 1).

Retrieving PKA-interacting proteins

Both of experimentally verified and predicted protein–protein interaction (PPI) databases were employed for retrieving of the PKA-binding proteins. Currently, large-scale PPI data is only available for *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens*. Thus we focused on the four organisms in the study. For experimentally verified PPI data, we took out PPI data from several public databases, including DIP [33], BioGrid [34], Grid [35], MINT [36], BIND [37], Wormbase [38] and HPRD [39]. All these PPI data were combined into an integrated and non-redundant data set, with the number of PPI pairs is 52,987, 5959, 30,558 and 51,529 in *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens*, respectively (Table 2). For predicted PPI data, we simply used the STRING database, which is an excellent pre-calculated PPI database for numerous species [40]. Both of predicted PPI data and their corresponding sequences were downloaded. The detailed statistics of PPI data in four species are shown in Table 2. In addition, Ptacek et al. carried out a large-scale experiment to identify PK-specific substrates in *S. cerevisiae* [29]. Totally, there were 256, 29, and 79 targets identified for TPK1, TPK2, and TPK3, respectively (Table 1). However, these substrates were not experimentally verified *in vivo*. And the exact PKA phosphorylation sites in these substrates were not mapped. Thus for *S. cerevisiae*, we also adopted this data set to make the analysis more integrated.

The experimentally verified and predicted PPI data were mapped to Swiss-Prot and TrEMBL (UniProt) database by BLAST for normalization of protein names (Table 1). Also, the experimentally verified substrates for eukaryotic PKA proteins were also listed (Table 1). Although there were 140 human proteins identified as PKA substrates, however, most of the annotation information in Phospho.ELM was "PKA Group". Thus, the information of which human PKA isoform will phosphorylate the proteins was not known. There was only one protein clearly annotated as "PKA alpha" (Table 1).

Prediction of PKA-specific phosphorylation sites with GPS

Previously, the GPS (Group-based Phosphorylation Scoring method, ver 1.10) software was developed for general purpose to predict PK-specific phosphorylation sites, including PKA [7–8]. PKA group is a Serine/Threonine kinase sub-family. And the training data set was taken from Phospho.ELM (Ver 2.0) with 180 verified PKA sites [27]. Three cut-off values of 1, 2.4 and 4 were established for low, medium and high threshold, respectively. And the prediction performance under high threshold was *Sn* 67.80% and *Sp* 97.88%. As the database updated, now there were 337 PKA sites identified (Ver 6.0). To test whether these new collected data will reduce the prediction performance of GPS, we re-calculated the sensitivity (*Sn*) and specificity (*Sp*) of GPS with the newly generated data set. In this study, the threshold was chosen as 4, and the prediction performance of GPS on current data set was *Sn* 64.99% and *Sp* 97.83%. Thus, the prediction performance of GPS 1.10 is still satisfying and slightly reduced for the new data. In this regard, we used GPS 1.10 directly for PKA phosphorylation sites prediction.

Orthologs detection for PKA-interacting proteins

To detect conserved PKA-binding proteins in the four species, we downloaded stand-alone InParanoid program 1.0 (http://inparanoid.sbc.su.se/) [45]. With the default parameters, we calculated orthologs pairwisely from protein sequences of four eukaryotic proteomes (Supplementary Table 5).

Acknowledgments

The authors thank two anonymous reviewers, whose suggestions have greatly improved the presentation of this manuscript. This work was supported by grants from Chinese Academy of Science (KSCX2-YW-H 10, KSCX2-YW-R65), Chinese Natural Science Foundation (39925018, 30270654, 30270293, 90508002, 30500183, 30700138 and 60533020), Chinese 973 Project (2002CB713700), Chinese 863 Project (2001AA215331), Chinese Minister of Education (20020358051) and the National Institutes of Health (DK56292; CA89019).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2008.08.013.

References

[1] J. Ptacek, M. Snyder, Charging it up: global analysis of protein phosphorylation, Trends Genet. 22 (2006) 545–554.
[2] B. Kobe, T. Kampmann, J.K. Forwood, P. Listwan, R.I. Brinkworth, Substrate specificity of protein kinases and computational prediction of substrates, Biochim. Biophys. Acta 1754 (2005) 200–209.
[3] G. Manning, D.B. Whyte, R. Martinez, T. Hunter, S. Sudarsanam, The protein kinase complement of the human genome, Science 298 (2002) 1912–1934.
[4] S. Caenepeel, G. Charydczak, S. Sudarsanam, T. Hunter, G. Manning, The mouse kinome: discovery and comparative genomics of all mouse protein kinases, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 11707–11712.
[5] N. Blom, S. Gammeltoft, S. Brunak, Sequence and structure-based prediction of eukaryotic protein phosphorylation sites, J. Mol. Biol. 294 (1999) 1351–1362.
[6] L.M. Iakoucheva, P. Radivojac, C.J. Brown, T.R. O'Connor, J.G. Sikes, Z. Obradovic, A.K. Dunker, The importance of intrinsic disorder for protein phosphorylation, Nucleic Acids Res. 32 (2004) 1037–1049.

[7] F.F. Zhou, Y. Xue, G.L. Chen, X. Yao, GPS: a novel group-based phosphorylation predicting and scoring method, Biochem. Biophys. Res. Commun. 325 (2004) 1443–1448.
[8] Y. Xue, F. Zhou, M. Zhu, K. Ahmed, G. Chen, X. Yao, GPS: a comprehensive www server for phosphorylation sites prediction, Nucleic Acids Res. 33 (2005) W184–W187.
[9] Y. Xue, A. Li, L. Wang, H. Feng, X. Yao, PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory, BMC Bioinformatics 7 (2006) 163.
[10] J.C. Obenauer, L.C. Cantley, M.B. Yaffe, Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs, Nucleic Acids Res. 31 (2003) 3635–3641.
[11] H.D. Huang, T.Y. Lee, S.W. Tzeng, J.T. Horng, KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites, Nucleic Acids Res. 33 (2005) W226–W229.
[12] J.H. Kim, J. Lee, B. Oh, K. Kimm, I. Koh, Prediction of phosphorylation sites using SVMs, Bioinformatics 20 (2004) 3179–3184.
[13] N. Blom, T. Sicheritz-Ponten, R. Gupta, S. Gammeltoft, S. Brunak, Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence, Proteomics 4 (2004) 1633–1649.
[14] R.I. Brinkworth, R.A. Breinl, B. Kobe, Structural basis and prediction of substrate specificity in protein serine/threonine kinases, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 74–79.
[15] G. Neuberger, G. Schneider, F. Eisenhaber, pkaPS: prediction of protein kinase A phosphorylation sites with the simplified kinase-substrate binding model, Biol. Direct. 2 (2007) 1.
[16] C.R. Ingrell, M.L. Miller, O.N. Jensen, N. Blom, NetPhosYeast: prediction of protein phosphorylation sites in yeast, Bioinformatics 23 (2007) 895–897.
[17] R.I. Brinkworth, A.L. Munn, B. Kobe, Protein kinases associated with the yeast phosphoproteome, BMC Bioinformatics 7 (2006) 47.
[18] R. Linding, L.J. Jensen, G.J. Ostheimer, M.A. van Vugt, C. Jorgensen, I.M. Miron, F. Diella, K. Colwill, L. Taylor, K. Elder, P. Metalnikov, V. Nguyen, A. Pasculescu, J. Jin, J.G. Park, L.D. Samson, J.R. Woodgett, R.B. Russell, P. Bork, M.B. Yaffe, T. Pawson, Systematic discovery of in vivo phosphorylation networks, Cell 129 (2007) 1415–1426.
[19] R.M. Biondi, A.R. Nebreda, Signalling specificity of Ser/Thr protein kinases through docking-site-mediated interactions, Biochem. J. 372 (2003) 1–13.
[20] P.M. Holland, J.A. Cooper, Protein modification: docking sites for kinases, Curr. Biol. 9 (1999) R329–R331.
[21] M.B. Yaffe, G.G. Leparc, J. Lai, T. Obata, S. Volinia, L.C. Cantley, A motif-based profile scanning approach for genome-wide prediction of signaling pathways, Nat. Biotechnol. 19 (2001) 348–353.
[22] G. McConnachie, L.K. Langeberg, J.D. Scott, AKAP signaling complexes: getting to the heart of the matter, Trends. Mol. Med. 12 (2006) 317–323.
[23] D.L. Beene, J.D. Scott, A-kinase anchoring proteins take shape, Curr. Opin. Cell. Biol. 19 (2007) 192–198.
[24] J.J. Michel, J.D. Scott, AKAP mediated signal transduction, Annu. Rev. Pharmacol. Toxicol. 42 (2002) 235–257.
[25] D.A. Walsh, J.P. Perkins, E.G. Krebs, An adenosine 3′,5′-monophosphate-dependant protein kinase from rabbit skeletal muscle, J. Biol. Chem. 243 (1968) 3763–3765.
[26] K. Tasken, E.M. Aandahl, Localized effects of cAMP mediated by distinct routes of protein kinase A, Physiol. Rev. 84 (2004) 137–167.
[27] F. Diella, S. Cameron, C. Gemund, R. Linding, A. Via, B. Kuster, T. Sicheritz-Ponten, N. Blom, T.J. Gibson, Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins, BMC Bioinformatics 5 (2004) 79.
[28] M. Tabish, R.A. Clegg, H.H. Rees, M.J. Fisher, Organization and alternative splicing of the *Caenorhabditis elegans* cAMP-dependent protein kinase catalytic-subunit gene (kin-1), Biochem. J. 339 (Pt 1) (1999) 209–216.
[29] J. Ptacek, G. Devgan, G. Michaud, H. Zhu, X. Zhu, J. Fasolo, H. Guo, G. Jona, A. Breitkreutz, R. Sopko, R.R. McCartney, M.C. Schmidt, N. Rachidi, S.J. Lee, A.S. Mah, L. Meng, M.J. Stark, D.F. Stern, C. De Virgilio, M. Tyers, B. Andrews, M. Gerstein, B. Schweitzer, P.F. Predki, M. Snyder, Global analysis of protein phosphorylation in yeast, Nature 438 (2005) 679–684.
[30] R.C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity, BMC Bioinformatics 5 (2004) 113.
[31] S. Kumar, K. Tamura, M. Nei, MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment, Brief. Bioinform. 5 (2004) 150–163.
[32] X. Li, H.P. Li, K. Amsler, D. Hyink, P.D. Wilson, C.R. Burrow, PRKX, a phylogenetically and functionally distinct cAMP-dependent protein kinase, activates renal epithelial cell migration and morphogenesis, Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 9260–9265.
[33] L. Salwinski, C.S. Miller, A.J. Smith, F.K. Pettit, J.U. Bowie, D. Eisenberg, The Database of Interacting Proteins: 2004 update, Nucleic Acids Res. 32 (2004) D449–D451.
[34] C. Stark, B.J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, M. Tyers, BioGRID: a general repository for interaction datasets, Nucleic Acids Res. 34 (2006) D535–D539.
[35] B.J. Breitkreutz, C. Stark, M. Tyers, The GRID: the General Repository for Interaction Datasets, Genome Biol. 4 (2003) R23.
[36] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, G. Cesareni, MINT: a Molecular INTeraction database, FEBS Lett. 513 (2002) 135–140.
[37] C. Alfarano, C.E. Andrade, K. Anthony, N. Bahroos, M. Bajec, K. Bantoft, D. Betel, B. Bobechko, K. Boutilier, E. Burgess, K. Buzadzija, R. Cavero, C. D'Abreo, I. Donaldson, D. Dorairajoo, M.J. Dumontier, M.R. Dumontier, V. Earles, R. Farrall, H. Feldman, E. Garderman, Y. Gong, R. Gonzaga, V. Grytsan, E. Gryz, V. Gu, E. Haldorsen, A. Halupa, R. Haw, A. Hrvojic, L. Hurrell, R. Isserlin, F. Jack, F. Juma, A. Khan, T. Kon, S. Konopinsky, V. Le, E. Lee, S. Ling, M. Magidin, J. Moniakis, J. Montojo, S. Moore, B. Muskat, I. Ng, J.P. Paraiso, B. Parker, G. Pintilie, R. Pirone, J.J. Salama, S. Sgro, T. Shan, Y. Shu, J. Siew, D. Skinner, K. Snyder, R. Stasiuk, D. Strumpf, B. Tuekam, S. Tao, Z. Wang, M. White, R. Willis, C. Wolting, S. Wong, A. Wrong, C. Xin, R. Yao, B. Yates, S. Zhang, K. Zheng, T. Pawson, B.F. Ouellette, C.W. Hogue, The Biomolecular Interaction Network Database and related tools 2005 update, Nucleic Acids Res. 33 (2005) D418–D424.
[38] T. Bieri, D. Blasiar, P. Ozersky, I. Antoshechkin, C. Bastiani, P. Canaran, J. Chan, N. Chen, W.J. Chen, P. Davis, T.J. Fiedler, L. Girard, M. Han, T.W. Harris, R. Kishore, R. Lee, S. McKay, H.M. Muller, C. Nakamura, A. Petcherski, A. Rangarajan, A. Rogers, G. Schindelman, E.M. Schwarz, W. Spooner, M.A. Tuli, K. Van Auken, D. Wang, X. Wang, G. Williams, R. Durbin, L.D. Stein, P.W. Sternberg, J. Spieth, WormBase: new content and better access, Nucleic Acids Res. 35 (2007) D506–D510.
[39] G.R. Mishra, M. Suresh, K. Kumaran, N. Kannabiran, S. Suresh, P. Bala, K. Shivakumar, N. Anuradha, R. Reddy, T.M. Raghavan, S. Menon, G. Hanumanthu, M. Gupta, S. Upendran, S. Gupta, M. Mahesh, B. Jacob, P. Mathew, P. Chatterjee, K.S. Arun, S. Sharma, K.N. Chandrika, N. Deshpande, K. Palvankar, R. Raghavnath, R. Krishnakanth, H. Karathia, B. Rekha, R. Nayak, G. Vishnupriya, H.G. Kumar, M. Nagini, G.S. Kumar, R. Jose, P. Deepthi, S.S. Mohan, T.K. Gandhi, H.C. Harsha, K.S. Deshpande, M. Sarker, T.S. Prasad, A. Pandey, Human protein reference database—2006 update, Nucleic Acids Res. 34 (2006) D411–D414.
[40] C. von Mering, L.J. Jensen, B. Snel, S.D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M.A. Huynen, P. Bork, STRING: known and predicted protein-protein associations, integrated and transferred across organisms, Nucleic Acids Res. 33 (2005) D433–D437.
[41] H. Lee, N. Rezai-Zadeh, E. Seto, Negative regulation of histone deacetylase 8 activity by cyclic AMP-dependent protein kinase A, Mol. Cell. Biol. 24 (2004) 765–773.
[42] M. Brenner, A.B. Johnson, O. Boespflug-Tanguy, D. Rodriguez, J.E. Goldman, A. Messing, Mutations in GFAP, encoding glial fibrillary acidic protein, are associated with Alexander disease, Nat. Genet. 27 (2001) 117–120.
[43] Y. Sanchez, J. Bachant, H. Wang, F. Hu, D. Liu, M. Tetzlaff, S.J. Elledge, Control of the DNA damage checkpoint by chk1 and rad53 protein kinases through distinct mechanisms, Science 286 (1999) 1166–1171.
[44] M.A. Hoyt, L. Totis, B.T. Roberts, *S. cerevisiae* genes required for cell cycle arrest in response to loss of microtubule function, Cell 66 (1991) 507–517.
[45] M. Remm, C.E. Storm, E.L. Sonnhammer, Automatic clustering of orthologs and in-paralogs from pairwise species comparisons, J. Mol. Biol. 314 (2001) 1041–1052.