Systematic analysis of the Plk-mediated phosphoregulation in eukaryotes

Zexian Liu, Jian Ren, Jun Cao, Jiang He, Xuebiao Yao, Changjiang Jin and Yu Xue

Submitted: 9th March 2012; Received (in revised form): 18th June 2012

Abstract
Substantial evidence has confirmed that Polo-like kinases (Plks) play a crucial role in a variety of cellular processes via phosphorylation-mediated signaling transduction. Identification of Plk phospho-binding proteins and phosphorylation substrates is fundamental for elucidating the molecular mechanisms of Plks. Here, we present an integrative approach for the analysis of Plk-specific phospho-binding and phosphorylation sites (p-sites) in proteins. From the currently available phosphoproteomic data, we predicted tens of thousands of potential Plk phospho-binding and phosphorylation sites in eukaryotes, respectively. Furthermore, statistical analysis suggested that Plk phospho-binding proteins are more closely implicated in mitosis than their phosphorylation substrates. Additional computational analysis together with in vitro and in vivo experimental assays demonstrated that human Mis18B is a novel interacting partner of Plk1, while pT14 and pS48 of Mis18B were identified as phospho-binding sites. Taken together, this systematic analysis provides a global landscape of the complexity and diversity of potential Plk-mediated phosphoregulation, and the prediction results can be helpful for further experimental investigation.

Keywords: polo-like kinase; phospho-binding; phosphorylation; PBD; GPS

INTRODUCTION
The reversible phosphorylation catalyzed by protein kinases (PKs) is one of the pivotal post-translational modifications (PTMs) of proteins and orchestrates a large number of cellular and physiological processes [1–3]. In addition, numerous phosphorylation-mediated signal transduction events are achieved through a specific recognition process, in which a variety of phosphoprotein-binding domains (PPBDs) can physically interact with short motifs in the vicinity of phosphorylated serine (pS), threonine (pT) and/or tyrosine (pY) [1, 4–6]. In the abundant regulatory proteins, kinase activity and flexible phosphoprotein-binding may be carried out in a closely correlated, ‘hand-in-glove’ manner within a single molecule; this phenomenon is well...
exemplified by the Plk family [7–12]. Previous studies have demonstrated that Plk-mediated phosphorylation play an important role in mitosis [7, 8, 12, 13]. Also, C-terminal Polo-box domains (PBDs) of Plks, acting as pS/pT-binding modules, are crucial for orchestrating the proper sub-cellular localizations and kinase activities of Plks during multiple stages of mitosis [9, 10, 12, 14, 15]. Moreover, Plk-mediated phosphoregulation can regulate a variety of biological and cellular processes beyond mitosis, such as DNA damage [16, 17], centriole duplication [18, 19], apoptosis [20, 21], transcription regulation [22] and stress-activated MAPK cascade [23]. In addition, clinical evidence suggests that aberrant Plk function is closely associated with tumorigenesis [7, 21]. In this regard, the identification of Plk phosphorylation substrates and phospho-binding proteins has emerged as an urgent topic for a better understanding of the exact Plk molecular mechanisms.

Conventionally, experimental identification of Plk phosphorylation and phospho-binding sites that is conducted solely with a site-directed mutagenesis strategy is both labor-intensive and time-consuming. Thus, the application of high-throughput techniques has attracted considerable attention. In a landmark study, Lowery et al. performed a proteome-wide detection of the Plk1-PBD interactome with high-throughput mass spectrometry (HTP-MS) in human osteosarcoma U2OS cells and identified 622 potentially phospho-binding proteins during mitosis, although the exact phospho-binding sites were not identified [15]. Recently, Santamaria et al. identified 358 potentially Plk1-mediated phosphorylation sites (p-sites) in the human mitotic spindle by comparing the phosphoproteome in the presence or absence of Plk1 activity [13]. However, only 102 (28.49%) p-sites were successfully validated using in vitro peptide arrays, while Plk phospho-binding was not demonstrated [13]. Thus, the development of more efficient approaches is still an urgent challenge at present. Recently, computational studies of the phosphoproteome have attracted great attention. In 2005, we developed an online server of GPS (Group-based Prediction System) for the prediction of kinase-specific p-sites, and firstly constructed a predictor for Plk [24]. Later, several other tools, such as KinasePhos 2.0 [25], PhoScan [26] and PostMod [27], also contained a specific predictor for identifying potential Plk1 p-sites. Recently, a PSSM (position-specific scoring matrix) model for Plk1 was generated from the results of positional scanning oriented peptide library screening (PS-OPLS) [28]. In contrast with numerous efforts on p-site prediction, few analyses have been performed for the prediction of phospho-binding sites. The Scansite 2.0 can predict specific phospho-binding sites for 14–3–3, SH2, SH3 and PDZ PPBDs [29], whereas SMALI can predict SH2-specific phospho-binding sites [30]. Based on the crystal structure of human Plk1-PBD and its phospho-peptide complex [11, 31], Huggins et al. computationally analyzed the molecular determinants of Plk1 phospho-binding with molecular dynamics simulations [32]. Since known complex structures are limited, such an approach is difficult to be used for the large-scale analysis. An applicable tool for the prediction of Plk phospho-binding sites was still not constructed.

In this work, we systematically analyzed the Plk-mediated phosphoregulation in eukaryotes by constructing an integrated computational platform of GPS-Polo 1.0 for the prediction of Plk phospho-binding and phosphorylation sites. By comparison, the performance of GPS-Polo was better than other approaches. From eukaryotic phosphoproteomic data, we predicted 17 713 potential Plk phospho-binding sites in 8599 proteins and 17 232 Plk p-sites in 8712 substrates, respectively. Through a statistical comparison of the significantly differentiated GO (gene ontology) terms between the Plk phospho-binding proteins and phosphorylation substrates in *H. sapiens*, we observed that the Plk phospho-binding proteins are more closely involved in mitosis. Based on further computational analysis, we propose that human Mis18B, a critically important regulator of proper chromosome segregation during mitosis [33, 34], can be a Plk1-binding partner with high confidence. Here, we experimentally validated that Plk1 interacts with Mis18B in vitro and in vivo. Further biochemical studies demonstrated that both pT14 and pS48 in Mis18B are crucial for the interaction with Plk1. In addition, functional analysis revealed that this interaction plays a significant role in maintaining the stability of Mis18B, and probably promotes the subsequent phosphorylation of Mis18B by Plk1. Taken together, this systematic analysis provides a solid framework for a better understanding of Plk-mediated phosphoregulation, while the predictions and analyses will prove helpful for further experimental investigation. The GPS-Polo 1.0 is freely available for academic researches at: http://polo.biocuckoo.org/.
METHODS

Data collection and preparation
First, we searched PubMed with multiple key terms such as ‘Plk’, ‘Polo-like kinase’, ‘Polo-box domain’, and so on. The preponderance of the Plk-related articles was carefully curated (published before 24 September 2010). As previously described [24], the experimentally verified Plk phospho-binding or phosphorylation sites were regarded as positive data (+), while all the other noninteractive or nonphosphorylated serine/threonine (S/T) residues were taken as negative data (−), respectively. Finally, the nonredundant Plk phospho-binding data set used for training contained 56 positive sites and 5493 negative sites, whereas the Plk phosphorylation data set included 275 positive sites and 17886 negative sites. For the prediction and analysis of Plk-mediated phosphoregulation, we also collected 120877 pS/pT sites in 24739 substrates from five eukaryotic organisms (Supplementary experimental procedure).

Performance evaluation
As previously described [24], the four measurements of sensitivity (Sn), specificity (Sp), accuracy (Ac) and Mathew Correlation Coefficient (MCC) were adopted to evaluate the prediction performance. The four measurements were defined as shown below:

\[
Sn = \frac{TP}{TP + FN}, \quad Sp = \frac{TN}{TN + FP}, \\
Ac = \frac{TP + TN}{TP + FN + TN + FP}, \quad MCC = \frac{TP\times TN - FN\times FP}{\sqrt{(TP+FN)\times(TN+FP)\times(TP+FP)\times(TN+FN)}}
\]

The leave-one-out (LOO) validation and 4-, 6-, 8- and 10-fold cross-validations were performed. We also used 42 human Plk1 phospho-binding sites as the training data set and tested the performance of 14 nonhuman Plk1 binding sites in 13 proteins (Table 1 and Supplementary Table S1). For Plk phosphorylation, 262 p-sites of human Plk1 and its orthologs in other organisms (mouse and frog Plk1, fly Polo, nematode plk-1 and yeast Cdc5) were used as the training data set to test the performance of 18 non-Plk1 p-sites in 13 substrates (Table 1 and Supplementary Table S2). Moreover, we adopted 224 human Plk1 p-sites for training and tested the performance on 55 nonhuman Plk1 p-sites of 27 substrates (Table 1 and Supplementary Table S2).

Algorithm
Over the past several years, we developed a series of GPS algorithms mainly for the prediction of PTMs sites in proteins [24]. In this work, we refined the original approach and designed the GPS 2.2 algorithm, which contains two computational parts of a scoring strategy and performance improvement.

The basic hypothesis of the scoring strategy is that similarly short peptides would exhibit similar 3D structures and biochemical properties [24]. First, we defined phosphorylation/phospho-binding site peptide PSP(m, n) as a serine/threonine (S/T) amino acid flanked by m residues upstream and n residues downstream. Then we used the amino acid substitution matrix BLOSUM62 to estimate the similarity between two PSP(m, n) peptides A and B as

\[
S(A, B) = \sum_{i=1}^{n} \text{Score}(A[i], B[i])
\]

Score(A[i], B[i]) denotes the substitution score of the two amino acids of A[i] and B[i] in the BLOSUM62 at the position i. If \(S(A, B) < 0\), we redefined it as \(S(A, B) = 0\). A given PSP(m, n) is then compared with each of the experimentally verified phosphorylation/phospho-binding peptides in a pairwise manner to calculate the similarity score. The average value of the substitution scores is regarded as the final score.

Table 1: A statistical chart of the experimentally identified Plk phospho-binding and phosphorylation sites in proteins

<table>
<thead>
<tr>
<th>Organism</th>
<th>PK name</th>
<th>UniProt</th>
<th>Binding</th>
<th>p-sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>H. sapiens</td>
<td>PLK1</td>
<td>PS3350</td>
<td>42</td>
<td>224</td>
</tr>
<tr>
<td></td>
<td>PLK2</td>
<td>Q9NYY3</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>PLK3</td>
<td>Q9H4B4</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>PLK4</td>
<td>O00444</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>M. musculus</td>
<td>Pkl1</td>
<td>Q07832</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Pkl2</td>
<td>P53351</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>X. laevis</td>
<td>Pkl1</td>
<td>P70032</td>
<td>6</td>
<td>III</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>polo</td>
<td>P52304</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C. elegans</td>
<td>plk-1</td>
<td>P34331</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>CDC5</td>
<td>P32562</td>
<td>3</td>
<td>23</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>56</td>
<td>275</td>
</tr>
</tbody>
</table>

*The PK name, the Plk gene. *UniProt, the UniProt accession number. * Binding, the number of phospho-binding sites. * p-sites, the number of p-sites.

The receiver operating characteristic (ROC) curves were drawn, and AROC (area under ROC) values were calculated.
The performance improvement procedure comprises three sequential steps, including motif length selection, weight training and matrix mutation.

**Motif length selection**
In this step, the optimal combination of PSP($m$, $n$) was determined based on the highest LOO result. The $Sp$ value was fixed at 95% for phospho-binding and 90% for phosphorylation, while the combinations of PSP($m$, $n$) ($m=1, \ldots, 15$; $n=1, \ldots, 15$) were exhaustively tested. In this work, the optimal motifs for Plk phospho-binding and phosphorylation were determined as PSP(2, 4) and PSP(11, 12), respectively.

**Weight training**
The substitution score between two PSP($m$, $n$) peptides $A$ and $B$ was refined as

$$S'(A,B) = \sum_{m\leq i \leq n} w_i \text{Score}(A[i],B[i])$$

Initially, the weight of each position in PSP($m$, $n$) was defined as 1. The $w_i$ value is the weight of position $i$. Again, if $S'(A, B)<0$, we redefined it as $S'(A, B)=0$. Then we randomly picked out the weight of any position for +1 or −1 and recomputed the LOO result. The $Sp$ value was fixed at 95% for phospho-binding and 90% for phosphorylation. The manipulation was adopted when the $Sn$ value was increased. This process was continued until the $Sn$ value was not increased any further.

**Matrix mutation**
The aim of this step is to generate an optimal or near-optimal scoring matrix. BLOSUM62 was chosen as the initial matrix, and the LOO performance was calculated. Again, we fixed the $Sp$ value at 95% and 90% for the prediction of phospho-binding and phosphorylation, respectively, to improve the $Sn$ though randomly picking out an element of the BLOSUM62 matrix to be used as +1 or −1. This process was repeated until convergence was reached. Choosing a different initial matrix, e.g. BLOSUM45, will generate a convergent result if the training time is sufficient (Data not shown).

**Biochemical and cellular assays**
Plasmids were transfected using Lipofectamine 2000 (Invitrogen) for HeLa cells according to the manufacturer’s instructions or Ca$_3$(PO$_4$)$_2$ methods for 293T cells. For the in vitro pull-down assay, GST fusion protein-bound Sepharose beads were used as an affinity matrix to isolate proteins interacting with PBD by incubation with 293T cell lysate containing ectopically expressed FLAG-Mis18B wild type (WT) and specific site mutations. Briefly, they were incubated in pull-down buffer (PBS, 1 mM DTT, 1 mM PMSF, 1 mM DNAase, 1 mM RNAase) containing 0.5% Triton X-100 and a protease inhibitor cocktail for 3 h at 4°C. After the incubation, the beads were extensively washed with PBS containing 0.5% Triton X-100 and boiled in protein sample buffer, followed by fractionation of the bound proteins on 10% SDS-PAGE gel. Proteins were then transferred onto a nitrocellulose membrane for western blotting with the appropriate antibodies. Immunoprecipitation was carried out using lysis buffer (50 mM HEPES pH 7.4, 100 mM NaCl, 2 mM EGTA, 1 mM MgCl$_2$, 1 mM DTT) containing 0.25% Triton X-100, 1 mM phenylmethylsulfony fluoride (PMSF) and a protease inhibitor cocktail. For protein degradation assay, 24 h after transfection, cells were treated with cycloheximide (20 μg/ml) and harvested at the indicated time points. In some cases, 20 μM MG132 were added at 4 h prior to the harvesting. The following materials and methods can be obtained in the Supplementary experimental procedure: reagents, plasmid construction, recombinant protein expression, cell culture, synchronization, immunoprecipitation and immunofluorescence.

**RESULTS**

**Analysis of the Plk family and development of GPS-Polo**
From the scientific literature, we collected 56 experimentally identified Plk phospho-binding sites in 47 proteins (Supplementary Table S1) and 124 Plk substrates with 275 known p-sites (Supplementary Table S2). The statistical results of the data collection are shown in Table 1. Currently, most efforts have been undertaken for human Plk1, while only a few phospho-binding or phosphorylation sites have been experimentally identified for other Plks (Table 1). The phylogenetical relationship of Plk family members in various organisms were analyzed (Supplementary experimental procedure). All of the Plks have a conserved kinese domain in the N-terminus, with a less conserved tandem set or at least one PBD domain in the C-terminus (Figure 1). Plk1 in vertebrates, and Plk2 and Plk3 in H. sapiens, M. musculus and X. laevis, were originated from...
yeast Cdc5 (ScCdc5) and became diversified before speciation (Figure 1). In particular, a newly reported mouse Plk5 (MmPlk5) also originated from ScCdc5 was found to be close to Plk2 and Plk3. Although its biological functions are not yet fully elucidated [17], Plk5 together with Plk1, 2 and 3 might recognize similar phospho-binding or phosphorylation motifs such as ScCdc5. In our results, a fourth family member (Plk4) in a separate branch demonstrates that there is a certain degree of homology with other components, which is presumably ascribable to their sharing of a particular structure (the single Polo-box domain). However, their kinase domains are closely related to the other members. Moreover, additional evidence has confirmed that they share certain overlapping functions [18, 19]. For example, both Plk2 and Plk4 can phosphorylate human CENPJ at S595 [18] (Supplementary Table S2). Since no phospho-binding site and only four p-sites were experimentally identified for Plk4 (Table 1), we hypothesized that Plk4 might recognize distinct but nevertheless similar motifs as the other Plks.

Taken together, based on the hypothesis that members of the Plk family can recognize similar phospho-binding and phosphorylation motifs in proteins, we developed the general predictor GPS-Polo 1.0, in which all of the experimental phospho-binding and phosphorylation sites were regarded as sets of training data. More specific predictors of individual Plks were not designed due to the limitation of the available data. For performance improvement, a refined GPS 2.2 algorithm was adopted. The differences among the different versions of the GPS algorithms are described in the Supplementary experimental procedure and Supplementary Table S3. In GPS-Polo 1.0, the three thresholds of high, medium and low were chosen, while the medium cut-off was selected as the default configuration. The usage of GPS-Polo 1.0 is described in the Supplementary results.

**Performance evaluation and comparison**

To evaluate the performance and robustness of GPS-Polo 1.0, the LOO validation and 4-, 6-, 8-, 10-fold cross-validations were performed for Plk phospho-binding (Figure 2A) and phosphorylation (Figure 2B), respectively. To further evaluate the limitation and bias of the training data sets, we also tested the performance of nonhuman Plk1 binding proteins by using human Plk1 phospho-binding sites for training (Figure 2A). Moreover, the performance values of non-Plk1 and nonhuman Plk1 substrates were calculated with the training data sets of Plk1.
orthologs and human Plk1 p-sites, respectively (Figure 2B). Since the results of different validations are quite similar, it is evident that GPS-Polo 1.0 is a stable and robust predictor, while phospho-binding and phosphorylation sites of different Plk members follow similar sequence patterns. The LOO results were adopted as the major indicators of GPS-Polo 1.0 for the purposes of further comparison.

To exhibit the superiority of GPS-Polo, we compared it to several other reported approaches. Previously, Elia et al. proposed that Plk1-PBD recognizes a core consensus motif S-(pT/pS)-(P/X) (Motif 1; X, any amino acid except Cys) for phospho-binding [14]. To avoid any bias, we calculated the prediction performance of the Motif 1 with the same training data set used in GPS-Polo. Since no other predictors of Plk phospho-binding have been developed, we directly compared the GPS-Polo to Motif 1 (Table 2). When the Sp value was \( \approx 87.9\% \), the Sn value of GPS-Polo (85.71\%) was better than Motif 1 (80.36\%) (Table 2). The interactions of Plks and their phospho-binding proteins can be mediated through the priming phosphorylation by CDK/MAPK [9, 14], which also recognize the S/T-P motif in substrates for phosphorylation [35]. Thus, we speculated whether the prediction of Plk phospho-binding sites is identical to the predictions of CDK or MAPK p-sites. To address this problem, we chose two PK groups of CMGC/CDK and CMGC/MAPK from GPS 2.1 [24], and calculated the prediction performances with our training data set for Plk phospho-binding, respectively (Table 2). When the Sp value was \( \approx 95.5\% \), the Sn value of GPS-Polo (82.14\%) was much better than GPS 2.1 (CMGC/CDK: 63.64\%; CMGC/MAPK: 37.50\%) (Table 2). In this regard, using known Plk phospho-binding sites for training can significantly improve the prediction accuracy since other PKs such as CaMKII [36] and Nek2 [37] can also act as priming kinases for Plks.

For Plk phosphorylation, we also compared GPS-Polo to a variety of predictors, including GPS [24], PostMod [27], KinasePhos 2.0 [25] and PhoScan [26]. Previously, Nakajima et al. suggested that D/E-X-S/T-Φ-X-D/E (Motif 2; X, any amino acid; Φ, a hydrophobic amino acid) is a consensus phosphorylation motif modified by Plk1 [8]. The prediction performance of Motif 2 was also calculated. The same training data set used in GPS-Polo was directly used for evaluating the performance of these approaches (Table 2). We fixed the Sp values of GPS-Polo 1.0 so as to be similar to these methods, and then compared the Sn values. With the similar
Sp values, the corresponding Sn value of GPS-Polo 1.0 was found to be much better than the other approaches (Table 2).

Systematic prediction and analysis of the Plk-mediated phosphoregulation

Since the real p-sites are only a very small part of all of the Ser/Thr residues in proteins [24], ab initio prediction of Plk phospho-binding and phosphorylation proteins with their sites merely from primary protein sequences will generate an excess of false positive hits. Recently, rapid advances in HTP-MS techniques have resulted in an explosive increase of phosphoproteomic data, with the systematic detection of thousands of p-sites in just a single experiment. However, the exact regulatory PKs for most of these p-sites have not been experimentally elucidated. To reduce the false positive rate, here we predicted Plk phospho-binding and phosphorylation of proteins from the available phosphoproteomic data.

From the public databases and several large-scale experiments, we collected 120,877 pS/pT sites in 24,739 proteins in eukaryotes, including S. cerevisiae, C. elegans, D. melanogaster, M. musculus and H. sapiens (more detailed information is available in the Supplementary experimental procedure and Supplementary Table S4). In total, there were 17,713 potential Plk phospho-binding sites predicted in 8,599 proteins and 17,232 p-sites in 8,712 substrates, using the default threshold (Table 3). From the results, we estimated that on average 14.65% of the total proteins in a eukaryotic organism are potential Plk phospho-binding proteins, while 14.26% of the total proteins are potentially phosphorylated by the Plks (Table 3). Although these potential Plk phospho-binding and phosphorylation of proteins are abundant, they have a relatively small overlap rate of 31.62% (Table 3). Only 755 (5.73%) proteins appear to interact with and are phosphorylated by Plks at the same p-sites, while 3,405 (25.89%) proteins interact with and are phosphorylated by Plks at different sites in the same protein (Table 3). In this regard, phospho-binding and phosphorylation may play distinct regulatory roles in a variety of biological processes.

With the GO annotations, we statistically analyzed the functional abundance and diversity of Plk-mediated phosphoregulation in H. sapiens with the

Table 2: Comparison of the GPS 2.2 algorithm with other approaches

<table>
<thead>
<tr>
<th>Method</th>
<th>Threshold</th>
<th>Acc (%)</th>
<th>Sn (%)</th>
<th>Sp (%)</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phospho-binding</td>
<td>High</td>
<td>97.56</td>
<td>57.14</td>
<td>97.97</td>
<td>0.3477</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>95.27</td>
<td>82.14</td>
<td>95.41</td>
<td>0.3444</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>89.99</td>
<td>85.71</td>
<td>90.03</td>
<td>0.2451</td>
</tr>
<tr>
<td></td>
<td></td>
<td>87.89</td>
<td>85.71</td>
<td>87.91</td>
<td>0.2205</td>
</tr>
<tr>
<td>Phosphorylation</td>
<td>High</td>
<td>93.89</td>
<td>28.99</td>
<td>94.89</td>
<td>0.1283</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>87.20</td>
<td>50.72</td>
<td>87.76</td>
<td>0.1408</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>81.52</td>
<td>57.61</td>
<td>81.89</td>
<td>0.1238</td>
</tr>
<tr>
<td></td>
<td></td>
<td>98.16</td>
<td>2.90</td>
<td>99.63</td>
<td>0.0482</td>
</tr>
<tr>
<td></td>
<td></td>
<td>91.92</td>
<td>38.41</td>
<td>92.74</td>
<td>0.1426</td>
</tr>
<tr>
<td></td>
<td></td>
<td>66.08</td>
<td>76.81</td>
<td>65.91</td>
<td>0.1097</td>
</tr>
<tr>
<td></td>
<td></td>
<td>97.93</td>
<td>4.35</td>
<td>99.37</td>
<td>0.0533</td>
</tr>
<tr>
<td>Motif 1e</td>
<td></td>
<td>97.93</td>
<td>0.36</td>
<td>87.93</td>
<td>0.2046</td>
</tr>
<tr>
<td>Motif 2e</td>
<td></td>
<td>98.14</td>
<td>1.82</td>
<td>99.63</td>
<td>0.0281</td>
</tr>
<tr>
<td>CMGC/CDK</td>
<td></td>
<td>95.15</td>
<td>63.64</td>
<td>95.47</td>
<td>0.2657</td>
</tr>
<tr>
<td>CMGC/MAPK</td>
<td></td>
<td>94.88</td>
<td>37.50</td>
<td>95.47</td>
<td>0.1532</td>
</tr>
<tr>
<td>GPS 2.1</td>
<td>High</td>
<td>92.64</td>
<td>18.25</td>
<td>93.78</td>
<td>0.0599</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>84.30</td>
<td>28.83</td>
<td>85.15</td>
<td>0.0476</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>76.82</td>
<td>39.05</td>
<td>77.40</td>
<td>0.0478</td>
</tr>
<tr>
<td>PostMod</td>
<td></td>
<td>91.67</td>
<td>21.17</td>
<td>92.75</td>
<td>0.0646</td>
</tr>
<tr>
<td>KinasePhos 2.0</td>
<td></td>
<td>65.64</td>
<td>52.73</td>
<td>65.84</td>
<td>0.0477</td>
</tr>
<tr>
<td>PhoScan</td>
<td></td>
<td>9793</td>
<td>3.64</td>
<td>99.38</td>
<td>0.0453</td>
</tr>
</tbody>
</table>

Note. For the construction of the GPS-Polo 1.0 software, the three thresholds of high, medium and low were chosen for Plk phospho-binding and phosphorylation, respectively. We fixed the Sp values of GPS-Polo 1.0 to be identical or similar to other methods and compared the Sn values. A Plk-PBD phospho-binding motif S-(pT/pS)-(P/X) (X, any amino acid except Cys) [4]. A Plk phosphorylation motif D/E-X-S/T-P-X-D/E (X, any amino acid; X, a hydrophobic amino acid) [8]. The Sn values of GPS-Polo are higher than the simple motifs. For the prediction of the Plk p-sites, the Sn values of GPS-Polo are evidently much better than the other computational approaches. Two PK groups of CMGC/CDK and CMGC/MAPK were selected from GPS 2.1 [24].
The over-represented GO terms of biological processes, molecular functions and cellular components were calculated for Plk phospho-binding (Supplementary Table S5) and phosphorylation (Supplementary Table S6) proteins, respectively. In particular, we also compared the differentiated GO terms between the phospho-binding and phosphorylation targets with Yates’ Chi-square ($\chi^2$) test (Table 4 and Supplementary experimental procedure). Interestingly, although both Plk phospho-binding and phosphorylation of proteins are significantly involved in cell cycle and mitosis processes (Supplementary Tables S5 and S6), the former is more heavily implicated in mitosis, since the top two most differentiated GO terms are mitosis (GO:0007067) and the mitotic cell cycle (GO:0000278) (Table 4). Moreover, although Plk-mediated phosphoregulation was proposed to regulate apoptosis/cell death [20, 21] and stress-activated MAPK cascade [23], our results suggested that Plk phospho-binding targets more proteins than its phosphorylation (Table 4). Furthermore, we observed that both Plk phospho-binding and phosphorylation are implicated in DNA damage (GO:0006974, Supplementary Table S5) or double-strand break repair (GO:0006302, Supplementary Table S6), which are consistent with previous studies [16, 17]. In addition, Plk phospho-binding but not phosphorylation is significantly implicated in transcription regulation (GO:0045449, Supplementary Table S5).

### Accurate prediction of the in vivo Plk-mediated phosphoregulation

In the above section, we predicted several tens of thousands of potential Plk phospho-binding and phosphorylation sites from the phosphoproteomic data in eukaryotes. However, most of these sites may only interact or be phosphorylated by Plks in vitro but not in vivo because Plks must co-localize and ‘kiss’ their partners for either interaction or phosphorylation to take place in a cell [24]. In this regard, accurate prediction of in vivo Plk-mediated phosphoregulation is still a great challenge.

The accumulated evidence suggests that Plks play an essential role during mitotic progression and can localize in various distinct regions, such as the midbody, centrosome and kinetochore [10, 14, 38, 39]. Previously, we developed the MiCroKit 3.0 database that contains proteins that localize in the midbody, centrosome and/or kinetochore (microkit proteins) [39]. All of the microkit proteins were experimentally identified with directly corroborating evidence for subcellular localization under fluorescent microscopy [39]. Given the functional importance of the midbody, centrosome and kinetochore in mitosis and co-localization, we hypothesized that microkit proteins are likely to heavily interact with or be phosphorylated by Plks.

Using the MiCroKit database as a reference, we mapped the annotated phosphoproteomic data in the above section to microkit proteins. With the hypergeometric test, our statistical results clearly indicated that Plk phospho-binding or phosphorylation proteins are significantly enriched in microkit proteins, at least in S. cerevisiae and H. sapiens (P < 0.01) because the information for the other species is far from integrated (Table 5). In this regard, the midbody, centrosome and kinetochore are hotspots of the in vivo Plk-mediated phosphoregulation. Several examples were randomly selected, and their prediction results are shown in Figure 3. Previous studies...
reported that human Mis18B (O43482) is an important regulator for proper chromosome segregation during mitosis, while its centromeric localization is crucial for recruiting de novo-synthesized CENP-A to the kinetochore [33, 34]. However, the precise molecular mechanisms underlying Mis18B regulation during mitosis remain elusive. Here, we predicted that its pS48 and pT221 might be potential Plk phospho-binding sites, while pS225 might be phosphorylated by Plks (Figure 3A). The tumor suppressor TP53/p53 (P04637), which is positioned at centrosomes during mitosis, was identified as a critical phosphorylation target of Plk1, whereas the kinase domain region of Plk1 (99–329aa) was suggested to be sufficient for its interaction with the DNA-binding domain of TP53 [20]. However, no Plk1 p-sites were experimentally identified in TP53. Our predictions revealed that the single residue S9 might be modified by Plks (Figure 3B). It was also confirmed that the physical interaction between Plk1 and TP53 inhibits the pro-apoptotic function of the latter, and the findings suggest that pT81, pT150 and pS315 are potential Plk phospho-binding sites (Figure 3B). Moreover, extensive research has suggested that ZW10 (O43264) associates specifically with kinetochores and acts as a mitotic checkpoint protein involved in chromosome segregation [40]. However, it is still unclear whether PBDs participate in the Plk–ZW10 interactions. With GPS-Polo, pT438 was predicted as a potential Plk phospho-binding site (Figure 3C). In addition, the results confirmed that NDC80 (O14777) is an evolutionarily conserved kinetochore-associated protein that plays a significant role in chromosome alignment and spindle organization [41]. Here, we predicted S5, S15 and S44 to be putative Plk p-sites (Figure 3D).
Human Plk-mediated phosphoregulation favors the distributive phosphorylation model

The large-scale analyses provided a great opportunity to dissect how Plk phospho-binding and phosphorylation synergistically regulate proteins. Previously, two hypothetical models, a processive phosphorylation model and distributive phosphorylation model, were proposed for PBD-mediated substrate targeting [9, 10] (Figure 7). In the former model, the Plks first bind with a phosphorylated protein by PBDs and further phosphorylate the protein at one or multiple sites by their kinase domains (Figure 7). In the latter model, a phosphorylated protein first interacts with the PBDs serving as a scaffold and recruits additional substrates to be further phosphorylated by the Plks (Figure 7). In both models, the protein must be first phosphorylated by Plks (self-priming) or other PKs (nonself-priming) (Figure 7). If self-priming is the predominant mode, most of the annotated p-sites would be expected to be both Plk phospho-binding and phosphorylation sites. However, only 859 p-sites in 755 (5.73%) proteins interact with and are phosphorylated by Plks (Table 3). Furthermore, upon analyzing the mitosis-specific Plk-mediated phosphoregulation, this proportion is unchanged (Supplementary Table S7). Thus, the proteins favor a nonself-priming pattern. Again, if the processive phosphorylation model were predominant, most of the proteins would contain both Plk phospho-binding and phosphorylation sites. In our analysis of the eukaryotic phosphoproteomic data, only 4159 (31.62%) of the total annotated proteins were found to both interact with and be phosphorylated by Plks (Table 3). Although this proportion is moderately enhanced to 48.90% (289 proteins) in the mitosis-specific Plk-mediated phosphoregulation (Supplementary Table S7), our results suggest that Plk regulation favors the distributive model.

### Table 5: The statistics of the prediction results for the in vivo Plk-mediated phosphoregulation

<table>
<thead>
<tr>
<th>Organism</th>
<th>Num.</th>
<th>Phospho-binding Pro. Sites</th>
<th>P value</th>
<th>Phosphorylation Sub. Sites</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. cerevisiae</td>
<td>191</td>
<td>97</td>
<td>206</td>
<td>1.33 × 10⁻⁵</td>
<td>114</td>
</tr>
<tr>
<td>C. elegans</td>
<td>29</td>
<td>8</td>
<td>8</td>
<td>1.37 × 10⁻¹</td>
<td>9</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>61</td>
<td>28</td>
<td>54</td>
<td>1.26 × 10⁻³</td>
<td>26</td>
</tr>
<tr>
<td>M. musculus</td>
<td>89</td>
<td>40</td>
<td>108</td>
<td>8.95 × 10⁻²</td>
<td>34</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>500</td>
<td>283</td>
<td>830</td>
<td>2.11 × 10⁻⁷</td>
<td>241</td>
</tr>
<tr>
<td>Total</td>
<td>870</td>
<td>456</td>
<td>1206</td>
<td>1.45 × 10⁻⁷</td>
<td>424</td>
</tr>
</tbody>
</table>

Note. The protein localization information from the MiCroKit database [39] was utilized as an efficient filter to remove potentially false positive hits. Num., the number of proteins collected in the MiCroKit database. Pro., the number of annotated Plk phospho-binding proteins. P value, the result of the hypergeometric test. Sub., the number of annotated Plk phosphorylation substrates.

Human Plk1 interacts with and regulates phosphorylated Mis18B

Although a number of computational analyses were performed, it was still not clear whether the prediction results would be reliable with high confidence. To further evaluate the prediction accuracy, ‘wet’ experiment was needed. Here we mainly focused on validating the phosphorylation-dependent interaction between Plk1 and Mis18B. We first explored the spatiotemporal patterns of Plk1 (Figure 4A) and its co-localization profiles together with Mis18B (Figure 4B). Immunofluorescence showed that both Mis18B and Plk1 readily appear at the centrosomes in prophase HeLa cells.
In this regard, we hypothesized that Plk1 might physically interact with Mis18B in vivo. The phosphoproteomic data showed Mis18B to be a phosphorylated protein, with pS48 and pT221 predicted as potential Plk phospho-binding sites (Figure 3A). We directly examined whether the PBD of Plk1 is responsible for its interaction with phosphorylated Mis18B. The human plasmid Plk1-PBD (Plk1-C, 305-603) was constructed (Figure 5A). Recombinant GST-tagged PBD and GST were incubated with lysate of 293T cells transiently transfected with FLAG-Mis18B, which ensured the correct phosphorylation of Mis18B. Preliminary experiments revealed that pT221 is not responsible for interacting with Plk1-PBD (data not shown), while Mis18B can still interact with Plk1-PBD even when pS48 was mutated. In this regard, additional there must be phospho-binding sites present in Mis18B. To avoid overlooking any positive hits, we input the full-length sequence of Mis18B into GPS-Polo and predicted the additional phospho-binding site of pT14 (Figure 4B). Interestingly, although this site was not included in our phosphoproteomic data sets, a recent large-scale experiment demonstrated that this site is phosphorylated in the Akt-RSK-S6 kinase signaling networks [42].

Thus, we further investigated whether the two p-sites are responsible for interacting with Plk1-PBD. As illustrated in Figure 5C, GST-PBD, but not GST, could pull down the FLAG-Mis18B (lane 5 and 6), demonstrating that the C-terminal PBD domain is required for the Plk1-Mis18B interaction. Furthermore, GST pull-down assays demonstrated that double-site mutations almost abolished the binding affinity between Mis18B and PBD, while a single-site mutation (T14A or S48A) diminished nearly half of the binding affinity with PBD, respectively (Figure 5C, lanes 7–9). Thus, we concluded that pT14 and pS48 of Mis18B are Plk1 phospho-binding sites. In addition, we employed immunoprecipitation assay to investigate whether Plk1 forms a complex with Mis18B in vivo. As shown in Figure 5D, GFP-Mis18B immunoprecipitated with FLAG-Plk1 (lanes 3 and 4) indicating that Mis18B is a bona fide Plk1-interacting partner. This in vivo physical interaction was also confirmed in a recent experiment by a similar strategy [43].

Previously, extensive experimental studies showed that Plk1 plays a significant role in regulating protein stability [21]. Thus, we suspected that this interaction modulated the stability of Mis18B. When the cells were treated with the protein biosynthesis inhibitor cycloheximide (CHX), the stability of
FLAG-Mis18B WT was almost identical with FLAG-Mis18B T14A/S48A, without the presence of Plk1 (Figure 6A). When GFP-Plk1 was co-transfected, the degradation rate of FLAG-Mis18B T14A/S48A was unchanged, while the stability of FLAG-Mis18B WT was greatly enhanced. Also, the protein level of Mis18B remained constant after treatment with the proteasome inhibitor MG132 (Figure 6A, lanes 1, 6, 11 and 12; Figure 6B, lanes 1, 6, 11 and 12). In this regard, the degradation of Mis18B is mediated through the proteosomal pathway. These existing data amply confirmed that the phospho-binding between Plk1 and Mis18B is crucial for maintaining the stability of Mis18B. Moreover, the migratory bands were present during the co-existence of FLAG-Mis18B WT and GFP-Plk1 (Figure 6B, lanes 1–4 and 11), indicating that the Plk1-Mis18B interaction probably promotes the subsequent phosphorylation of Mis18B.

**DISCUSSION**

As a multi-faceted kinase family, Plks have been extensively characterized by their specific phosphorylation-dependent signaling transduction, orchestrate
a variety of biological and cellular processes especially mitosis, and are closely associated with tumorigenesis [7, 9, 10]. Thus, the identification of the Plk-mediated phosphoregulation by the detection of Plk phospho-binding and phosphorylation sites is fundamental to an understanding of the molecular mechanisms and regulatory roles of Plks. Although an optimal phosphorylation motif (Motif 2) of Plk1 was in vitro verified [8], only 1.82% of all the known Plk p-sites follow this pattern (Table 2). Moreover, although a consensus motif S-(pT/pS)-(P/X) (Motif 1) was proposed for Plk1 phospho-binding [14], we observed that the newly identified pT14 (AT14P) of human Mis18B does not follow this pattern (Figure 5B). Thus, accurate prediction of Plk phospho-binding and phosphorylation sites is still a great challenge. In this work, we developed a novel software package termed GPS-Polo. From the scientific literature, we extensively collected the known Plk phospho-binding and phosphorylation sites as the training data sets, while a previously designed GPS algorithm [24] was greatly refined for the purposes of this study. By direct comparison, the prediction performance is much better than the other available approaches (Table 2).

Ab initio prediction of Plk phospho-binding and phosphorylation sites merely from protein primary sequences will inevitably generate a huge amount of false positive hits because most of the Ser/Thr residues are not modified at all. In this regard, efficient filters need to be added to reduce false positive predictions. Recent progress in phosphoproteomic studies have identified over one hundred thousand p-sites in eukaryotic cells, and these can be used as source material from which the number of potential candidates is ultimately narrowed down. Using GPS-Polo 1.0, we predicted 17 713 potential Plk phospho-binding sites in 8599 proteins and 17 232 p-sites in 8712 substrates from eukaryotic phosphoproteomic data (Table 3). However, the
results are still of limited utility for further experiments because most of these sites may only interact with or be phosphorylated by Plks in vitro not in vivo since a number of contextual factors such as physical interaction and subcellular co-localization can provide additional specificity for the phosphoregulation [44].

Recently, large-scale identification of potential Plk p-sites has attracted much attention by quantitatively comparing the human phosphoproteomes in the presence or absence of Plk activity with small-molecule inhibitors [13, 43, 45, 46]. Since 102 in vitro validated Plk1 p-sites was included in our training data set [13], here we focused on analyzing the other three studies by integrating 1979 potential Plk p-sites in 810 proteins (Supplementary Table S8) [43, 45, 46]. From the results, we observed that the overlapping rate among the three analyses is limited, and only 205 (10.36%) p-sites occur in more than one experiment (Supplementary Table S8).

We used GPS-Polo to predict potential Plk phospho-binding and phosphorylation sites in these proteins and performed the Fisher’s Exact Test (http://www.langsrud.com/fisher.htm, 2-Tail) (Table 6). Interestingly, the statistical results suggested that Plk phosphorylation substrates were significantly enriched by ~2-fold ($P<0.01$), while Plk phospho-binding proteins were significantly under-represented (Table 6). Thus, the quantitative approach can be an efficient filter for Plk phosphorylation but not phospho-binding. Also, Kettenbach et al. experimentally identified 584 Plk1-interacting partners by immunoprecipitation [43]. We mapped all known p-sites to these proteins and performed the prediction with GPS-Polo (Supplementary Table S9). Again, the Fisher’s exact test was carried out, while the statistical results suggested that both Plk phospho-binding and phosphorylation proteins were significantly over-represented (Table 6). Thus, the interaction information can be an efficient filter for both Plk phospho-binding and phosphorylation. In addition, the co-localization of Plks and their

Figure 7: Two hypothetical models for phosphorylation-mediated Plk regulation [9, 10]. In both models, a protein is first phosphorylated by a priming PK, such as Plks (self-priming) or other PKs (non-self-priming). Then in the processive phosphorylation model, Plks interact with the phosphorylated protein by means of the PBDs, and further phosphorylate it at one or multiple sites via the kinase domains. In the distributive model, the phosphorylated protein can act as a scaffold by binding PBDs, and it recruits more proteins to be phosphorylated by Plks.
partners is a prerequisite for either interaction or modification within a cell [24]. By comparison, the co-localization information can be a more efficient filter for Plk phospho-binding (Table 5 and 6). Previous studies and our own Immunofluorescent assays demonstrated that Plks localize at the midbody, centrosome and kinetochore during mitosis [10, 14, 38, 39] (Figure 4A). Therefore, we used the protein localization information from the MiCroKit database to remove potentially false positive predictions from the large-scale analysis (Table 5).

Based on the prediction results (Figure 3A), we experimentally confirmed that human Mis18B is a novel interacting partner of Plk1 both in vitro and in vivo. Together with additional computational analysis, we further confirmed that pT14 and pS48 of Mis18B are responsible for its interaction with Plk1-PBD (Figure 5). In addition, the association between Plk1 and Mis18B is crucial for maintaining the stability of Mis18B and probably promotes the subsequently phosphorylation of Mis18B at S225 by Plk1 (Figure 3A). Interestingly, previous reports showed that Mis18B is essential for recruiting de novo-synthesized CENP-A to the centromeres [33, 34]. Thus, we speculated that the interaction between Plk1 and Mis18B may participate in regulating CENP-A dynamics during mitosis, and this hypothesis will be tested in future experimental work.

In our data set, there are 42 (75.0%) phospho-binding and 224 (81.5%) phosphorylation sites for the human Plk1 (Table 1). Thus, the intrinsic limitation and bias of the training data set cannot be avoided. For example, a previous experiment indicated that Plk3 can phosphorylate p53 at S20 [20], which was not predicted by GPS-Polo (Figure 3B). We believed that the prediction performance can be further improved with a greater supply of available data and additional refinement of the GPS algorithm. Furthermore, although we collected 120 877 pS/pT sites in 24 739 substrates in eukaryotes, the phosphoproteomic data set is still far from fully integrated. For example, the pT14 of Mis18B was recently identified as a real p-site but is not included in our data set [42]. In this regard, we will keep collecting and integrating the emerging phosphoproteomic data from the latest literature. In addition, more experimental studies will be carried out to further validate the prediction results.

Although a number of analyses remain to be performed, the systematic analysis reported here clearly exhibited the functional complexity and diversity of the Plk-mediated phosphoregulation in eukaryotes. The extensive computational evaluation and experimental confirmation suggest that our results are accurate and thus useful for further investigation. We believe that computational prediction followed by experimental verification will help advance the understanding of the molecular mechanisms and dynamics of Plk-centered regulation.

### SUPPLEMENTARY DATA

Supplementary data are available online at [http://bib.oxfordjournals.org/](http://bib.oxfordjournals.org/).

### Key Points
- The GPS-Polo is a software package for the prediction of Plk phospho-binding and phosphorylation sites.
- Systematic analysis suggests that Plk phospho-binding proteins are more closely implicated in mitosis than their phosphorylation substrates.
- The subcellular co-localization information can be used to reduce potentially false positive hits.
- Human Mis18B was predicted and experimentally validated as a Plk1 phospho-binding protein.
ACKNOWLEDGEMENTS

The authors thank Dr. Francesca Diella and Dr. Toby J. Gibson (EMBL) for always providing the newly data set of Phospho.Elm database during the past seven years. They thank Dr. Peter Hornbeck (Cell Signaling Technology, USA) for providing the PhosphoSitePlus data set on 14 July 2009, and Drs. Hong Li and Guohui Ding (SHS, China) for providing the SysPTM data set. Pacific Edit reviewed the manuscript prior to submission.

FUNDING

This work was supported, in whole or in part, by the National Basic Research Program (973 project) (2012CB911201, 2012CB910101, 2010CB945401, 2010CB912103 and 2011CB910400), Natural Science Foundation of China (9059002, 90913016, 31071154, 30900835, 30830036, 91019020, 31171300, 31171263 and 21075045), Chinese Academy of Sciences (KSCX1-YW-R-65, KSCX2-YW-H-10, KSCX2-YW-R-195 and KSCX2-YW-R-140), Fundamental Research Funds for the Central Universities (HUST: 2011TS085; SYSU: 11lgzd11) and Anhui Province Key Project (08040102005).

References


36. Hansen DV, Tung JJ, Jackson PK. CaMKII and polo-like kinase 1 sequentially phosphorylate the cytostatic factor Emi2/XErp1 to trigger its destruction and meiotic exit. *Proc Natl Acad Sci USA* 2006;103:608–13.


