Protein Engineering, Design & Selection vol. 24 no. 3 pp. 255–260, 2011
Published online November 8, 2010 doi:10.1093/protein/gzq094

GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection

Yu Xue[1,2,4], Zexian Liu[2], Jun Cao[2], Qian Ma[2], Xinjiao Gao[2], Qingqi Wang[2], Changjiang Jin[2], Yanhong Zhou[1], Longping Wen[2] and Jian Ren[3,4]

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

[4]To whom correspondence should be addressed.
E-mail: xueyu@mail.hust.edu.cn (X.Y.)/renjian@ustc.edu.cn (R.J.)

Received February 11, 2010; revised August 20, 2010; accepted October 8, 2010

Edited by Joost Schymkowitz

As the most important post-translational modification of proteins, phosphorylation plays essential roles in all aspects of biological processes. Besides experimental approaches, computational prediction of phosphorylated proteins with their kinase-specific phosphorylation sites has also emerged as a popular strategy, for its low-cost, fast-speed and convenience. In this work, we developed a kinase-specific phosphorylation sites predictor of GPS 2.1 (Group-based Prediction System), with a novel but simple approach of motif length selection (MLS). By this approach, the robustness of the prediction system was greatly improved. All algorithms in GPS old versions were also reserved and integrated in GPS 2.1. The online service and local packages of GPS 2.1 were implemented in JAVA 1.5 (J2SE 5.0) and freely available for academic researches at: http://gps.biocuckoo.org.

Keywords: group-based prediction system, motif length selection/phosphorylation, post-translational modification

Introduction

Phosphorylation is the most important post-translational modification of proteins, orchestrates most of biological processes and regulates cellular dynamics and plasticity. Usually, a member of protein kinase (PK) superfamily only modifies limited substrates by mainly recognizing special sequence/structural profiles around modified residues (S/T or Y), to ensure the signaling fidelity (Kreegipuu *et al*., 1998; Blom *et al*., 2004; Kobe *et al*., 2005; Hjerrild and Gammeltoft, 2006; Ubersax and Ferrell, 2007; Miller and Blom, 2009). In this regard, identification of phosphorylation sites, especially kinase-specific phosphorylation sites, is fundamental for understanding the molecular mechanisms of phosphorylation and elucidating dynamic interactions between PKs and their substrates. Besides the experimental methods, various computational approaches have also been established to generate highly potential candidates for further experimental verification. We and other bioinformatists have developed dozens of computational programs to predict kinase-specific phosphorylation sites in proteins [reviewed in (Kobe *et al*., 2005; Hjerrild and Gammeltoft, 2006; Miller and Blom, 2009)]. The latest compendium of computational resources for protein phosphorylation including phosphorylation databases and applicable tools could be available at: http://gps.biocuckoo.org/links.php.

Previously, based on a major hypothesis of short similar peptides bearing similar biological functions, we developed a kinase-specific phosphorylation sites predictor of GPS (Group-based Phosphorylation Scoring; Xue *et al*., 2005). We defined a phosphorylation site peptide PSP(*m*, *n*) as a S/T or Y residue flanked by *m* residues upstream and *n* residues downstream. In GPS 1.10, the PSP(3, 3) was arbitrarily decided (Xue *et al*., 2005). Later, we improved the GPS algorithm and developed GPS 2.0 (renamed as Group-based Prediction System) software. In GPS 2.0, the PSP(7, 7) was arbitrarily adopted (Xue *et al*., 2008). As the first stand alone software for phosphorylation sites prediction, GPS 2.0 could predict kinase-specific phosphorylation sites for 408 human PKs in hierarchy (Xue *et al*., 2008). Recently, we revealed that different combinations of modified peptides could generate different prediction performances during studying of protein palmitoylation (Ren *et al*., 2008). In this regard, an interesting question has emerged: can we find the optimized combination of PSP(*m*, *n*) with the optimal or near-optimal performance?

In this work, we carefully studied how different combinations of PSP(*m*, *n*) influenced prediction performance and robustness. The self-consistency validation and leave-one-out (LOO) validation were thoroughly carried out for each PK groups. We observed that the self-consistency results will be always increased with longer PSP(*m*, *n*). However, when the phosphorylated peptide was elongated, the LOO results will first reach a peak value then decrease. In this regard, we developed a novel but simple approach of motif length selection (MLS), which could automatically detect the optimal length of PSP(*m*, *n*) with the highest LOO performance. By comparing with our previous GPS 2.0 software (Xue *et al*., 2008), the average sensitivity (Sn) of the LOO was significantly increased by 15.62%, whereas the average Sn value of the self-consistency was slightly reduced by 2.28%. Importantly, it was proposed that the LOO validation might overfit in small samples, whereas the *n*-fold cross-validation (e.g. 10-fold) should do better (Shao, 1993; Dong *et al*., 2006). Thus, the 4-, 6-, 8-, 10-fold cross-validations were additionally performed. Interestingly, we observed that the LOO results were quite similar with *n*-fold cross-validations 72 PK groups with $\geq$30 phosphorylation sites. Again, for the 72 PK groups, the Sn of the LOO was averagely enhanced by 6.57%, while the Sn of the self-consistency was reduced

by 1.10%. Taken together, the newly developed MLS method could efficiently narrow down the difference between the LOO validation and self-consistency validation to improve the robustness of prediction system. The online service and local packages of GPS 2.1 were implemented in JAVA 1.5 (J2SE 5.0) and freely available at: http://gps.biocuckoo.org.

Materials and methods

Preparation of the training and testing data sets

The training data set was taken from Phospho.ELM 7.0 (Diella *et al*., 2004, 2008), including 16 462 experimentally verified phosphorylation sites. Among these sites, there were 3417 known kinase-specific phosphorylation sites reserved as the training data set in GPS 2.1. As previously described (Xue *et al*., 2008), we classified all human PKs with their verified sites into a hierarchical structure with four levels, including group, family, subfamily and single PK. The PK groups with less than three sites were not included. As previously described (Xue *et al*., 2005, 2008; Ren *et al*., 2008), we took all experimentally verified phosphorylation sites as the positive data (+), while all other S/T or Y residues in the same proteins were regarded as the negative data (−). The amino acid substitution matrix of BLOSUM62 was chosen as the initial matrix. For comparison of GPS 2.0 (Xue *et al*., 2008), we also prepared a testing data set from Phospho.ELM 6.0 (Diella *et al*., 2004, 2008), including 3157 non-redundant phosphorylation sites with kinase information.

Evaluation of prediction performance and robustness

To evaluate the prediction performances and robustness of GPS 2.1, the self-consistency validation, LOO validation and *n*-fold cross-validation were calculated. The self-consistency validation used the training positive data and negative data directly to evaluate the prediction performance, and represented the computational power of the prediction system. However, the robustness and stability of the software should be evaluated by LOO validation and *n*-fold cross-validation. In the LOO validation, which is also called as Jack–Knife validation, each sites in the data set was picked out in turn as an independent test sample, and all the remaining sites were regarded as training data. This process was repeated until each site was used as test data one time. In conventional *n*-fold cross-validation, all the (+) sites and (−) sites were combined and then divided equally into *n* parts, keeping the same distribution of (+) and (−) sites in each part. Then $n-1$ parts were merged into a training data set while the remnant part was taken as a testing data set. This process was repeated 20 times and the average performance of *n*-fold cross-validation was used to estimate the performance. In addition, a sequence-similarity-group-based *n*-fold cross-validation was also adopted that the *n* parts are not assigned randomly but based on groups of similar peptides. With the *k*-means clustering approach (Herwig *et al*., 1999; Soukas *et al*., 2000), we clustered the training data set into *n* groups ($n = 4, 6, 8, 10$ in this work). Given two PSP(m, n) peptides A and B, the similarity was measured as:

$$s(A,B) = \frac{\text{Num. of conserved substitutions}}{\text{Num. of all substitutions}}.$$

A conserved substitution is a substitution with a Score(a, b) > 0 in the BLOSUM62 matrix. The $s(A, B)$ ranges from 0 to 1. Thus, the distance between them can be defined as: $D(A, B) = 1/s(A, B)$. If $s(A, B) = 0$, $D(A, B) = \infty$.

By exhaustively testing, PSP(15, 15) was used. First, *n* phosphorylation sites from the positive data (+) were randomly picked out as the centroids. Second, the other positive sites were pairwisely compared with the *n* centroids and clustered into the groups with highest similarity values. Third, the centroid of each cluster was updated with the highest average similarity (HAS). The second and third steps were iteratively repeated until the clusters were not changed any longer. After the *n* clusters for positive sites were determined, we put each negative site into the cluster with the HAS. Then, we used $n-1$ parts as training data set, although the remaining part was regarded as testing data set. Since the *n* groups were fixed after *k*-means clustering, the performance calculation was only repeated $n-1$ times until each part was used as test data one time, while average values were computed.

In this work, the performances of self-consistency validation and LOO validation were calculated for all PK groups (Supplementary data, Tables S2 and S3). The conventional and sequence-similarity-group-based 4-, 6-, 8-, 10-fold cross-validations were performed for 72 PK groups without <30 sites. Owing to the page limitation, the performance of conventional 4-fold cross-validation was shown (Supplementary data, Table S2).

The four standard measurements of Sn, specificity (Sp), accuracy (Ac) and Mathew's correlation coefficient (MCC) were calculated as below:

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

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

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

Implementation of the online service and local packages

The online service and local packages of GPS 2.1 were implemented in JAVA 1.5 (J2SE 5.0). For the online service, we tested GPS 2.1 on a variety of internet browsers, including Internet Explorer 6.0, Netscape Browser 8.1.3 and Firefox 2 under Windows XP Operating System (OS), Mozilla Firefox 1.5 of Fedora Core 6 OS (Linux) and Safari 3.0 of Apple Mac OS X 10.4 (Tiger) and 10.5 (Leopard). For Windows and Linux systems, a latest version of Java Runtime Environment (JRE) package (JAVA 1.4.2 or later versions) of Sun Microsystems should be pre-installed for using the GPS 2.1 program. However, for Mac OS, the GPS 2.1 could be used directly without any additional packages. For convenience, we also developed the local packages of GPS 2.1. The stand alone softwares of GPS 2.1 supported three major Operating Systems, including Windows, Linux and Mac.

Results

Different combinations of PSP(m, n) generate different prediction performances and robustnesses

In this work, we carefully studied how different combinations of PSP(m, n) influenced prediction performances. The training data set was taken from Phospho.ELM 7.0 (Diella *et al.*, 2004, 2008), including 3417 known kinase-specific phosphorylation sites. To evaluate the performance and robustness of the prediction system, the self-consistency validation, LOO validation and n-fold cross-validation should be performed. The self-consistency validation uses the training data set to test the prediction performance on currently known data. However, the prediction system might be overtrained and only perfect for the training data set, with low prediction ability for new data. In this regard, the LOO validation should also be carried out. However, the LOO validation might not work well for discontinuous data sets (small samples) to be prone to overfitting (Shao, 1993; Dong *et al.*, 2006). In this regard, the n-fold cross-validation should be additionally performed (Shao, 1993; Dong *et al.*, 2006). If the self-consistency result is similar with the LOO validation and n-fold cross-validation, the prediction system is robust with less overfitting. Previously, we revealed that the results of LOO were similar with the n-fold cross-validation for PK groups with large number of phosphorylation sites ($n \geq 30$; Xue *et al.*, 2008). Thus, to test the performance and robustness of different combination of PSP(m, n), we only calculated the self-consistency and the LOO results. We exhaustively tested all combinations of PSP(m, n) ($m = 1, \ldots, 15$; $n = 1, \ldots, 15$). The Sn values were calculated under the Sp of 85, 90 and 95%. Then the average Sn was calculated as the final Sn value. From our previous experience, a higher Sp value is more important than a higher Sn to avoid too many false positive hits (Xue *et al.*, 2005, 2008; Ren *et al.*, 2008). Thus, to improve the prediction performance and robustness in the region of high Sp is more important than other regions. Owing to the page limitation, here we only presented four PK groups as instances (Number of phosphosrylation sites ≥ 30), including AGC/AKT, CMGC/CDK/CDC2/CDC2, Other/AUR and TK/InsR/IGF1R (Fig. 1). Interestingly, we observed that the self-consistency results will be always increased with longer PSP(m, n). However, when the phosphorylated peptide was elongated, the LOO results will first reach a peak value then decrease. For example, the Sn value of the self-consistency validation of AGC/AKT could reach 100% when PSP(15, 9) was chosen (the minimal length of the phosphorylated peptide with the highest Sn value; Fig. 1). However, in the LOO validation, the maximal Sn was 89.89% with PSP(5, 1) (Fig. 1). For AGC/AKT, if the PSP(15, 9) was chosen, its self-consistency performance was quite different with its LOO validation (100 versus 73.41%; Fig. 1). However, if PSP(5, 1) was selected, the self-consistency performance was very similar with its LOO validation (90.64 versus 89.89%; Fig. 1). Again, for CMGC/CDK/CDC2/CDC2, the Sn values of the self-consistency and LOO for the PSP(15, 15) were 99.76 and 53.24%, although the Sn results of the self-consistency and LOO for the optimal PSP(1, 2) were 79.86 and 78.18%, respectively (Fig. 1). Taken together, we proposed that the optimal PSP(m, n) with the highest LOO validation could efficiently narrow down the difference between the self-consistency and LOO validations to improve the robustness of the prediction system. More detailed results could be available in Supplementary data, Table S1.

A novel algorithm of MLS to improve the prediction robustness

On the basis of above analyses, here we developed GPS 2.1 with a novel approach of MLS. The methods in GPS 2.0 were also reserved in GPS 2.1. In GPS 2.0, we developed a novel approach of Matrix Mutation (MaM) to improve the prediction performance (Xue *et al.*, 2008). First, the amino

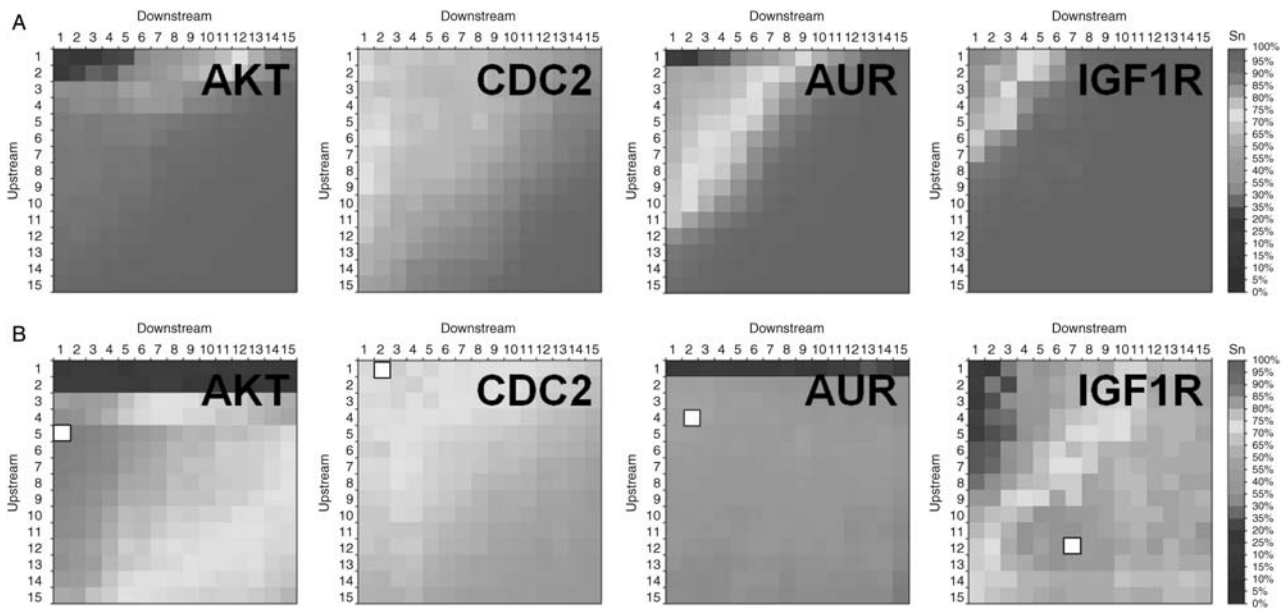

Fig. 1. The optimal PSP(m, n) with the highest LOO result (marked in white) narrows down the difference between the self-consistency and LOO results to improve the robustness of the prediction system. Four PK groups, including AGC/AKT, CMGC/CDK/CDC2/CDC2, Other/AUR and TK/InsR/IGF1R, were presented. We observed that (**A**) the self-consistency results will be always increased with longer PSP(m, n), while (**B**) the LOO results will first reach a peak value then decrease.

acids substitution matrix BLOSUM62 was selected as the initial matrix. Then the Sn and Sp values of LOO validation were calculated for each PK group. By exhaustively testing, we fixed Sp at 90% to improve Sn values by randomly mutating values of the BLOSUM62 matrix, until the Sn was no longer increased. In GPS 2.1, we first used the MaM to find the optimal matrix for each PK group. Then for each combination of PSP(m, n) ($m = 1, \ldots, 15$; $n = 1, \ldots, 15$), we calculated the average Sn values under Sp of 85, 90 and 95% with the LOO validation, respectively. And we obtained the optimal PSP(m, n) with the highest LOO validation. Owing to the training process is too time-consuming, such a procedure was not iteratively repeated. To extensively evaluate the prediction performance and robustness of GPS 2.1, the self-consistency validation and LOO validation were thoroughly carried out for all of PK groups, whereas 4-, 6-, 8-, 10-fold cross-validations were additionally calculated for 72 PK groups with $\geq$30 phosphorylation sites (Supplementary data, Tables S2 and S3). Furthermore, the sequence-similarity-group-based 4-, 6-, 8-, 10-fold cross-validations were also calculated, while several examples were shown in Supplementary data, Fig. S1. Both of conventional and sequence-similarity-group-based n-fold cross-validation results were similar with LOO validations.

Comparison of GPS 2.1 with our previous GPS 2.0

In our previous work, we extensively compared GPS 2.0 with other similar tools (Xue *et al.*, 2008). The prediction performance of GPS 2.0 is better or at least comparable with previously established programs. Thus, in this work, we only compared GPS 2.1 with GPS 2.0 to exhibit the superiority of MLS method (Table 1). To avoid any bias, we used the same data set (Phospho.ELM 6.0, including 3,161 kinase-specific phosphorylation sites) to compare the prediction performance and robustness between GPS 2.1 and GPS 2.0. For AGC/AKT, the Sn values of LOO and self-consistency were 92.06 and 97.22% in GPS 2.0, respectively. In GPS 2.1, both of the Sn values were refined as 92.46% (Table 1). And for Other/AUR, the LOO and self-consistency Sn values were 53.25 and 90.91% in GPS 2.0, while the two Sn values were 61.90 and 64.50% after MLS (Table 1). The full comparison results are available in Supplementary data, Table S4. In GPS 2.0, the average Sn values of LOO and self-consistency were calculated as 55.11 and 96.46%, while the two Sn values were 70.73 and 94.18% in GPS 2.1 (Table 1). Thus, with the MLS method, the average Sn of the LOO was significantly increased by 15.62%, while the average Sn value of the self-consistency was only slightly reduced by 2.28% (Table 1). As mentioned above, the LOO validation might be prone to overfitting for small samples, while the n-fold cross-validation should be additionally performed (Shao, 1993; Dong *et al.*, 2006). However, the results of LOO and n-fold cross-validation were quite similar for 72 PK groups with large samples ($n \geq 30$; Supplementary data, Table S2). For the 72 PK groups, the Sn of the LOO was averagely enhanced by 6.57%, while the Sn of the self-consistency was reduced by 1.10% (Supplementary data, Table S5). In this regard, the MLS method narrowed down the differences between the LOO and self-consistency to improve the robustness of GPS 2.1.

Usage of GPS 2.1 software

The online service and local packages of GPS 2.1 were implemented in JAVA 1.5 (J2SE 5.0). For prediction of kianse-specific phosphorylation sites, one or multiple protein sequences must be prepared in FASTA format. And at least one PK group should be selected by left-clicking on the PK list. Also, a threshold should be chosen. Then the prediction results will appear soon by left-clicking on the 'Submit' button (Fig. 2). The online service and local packages of GPS 2.1 support this manipulation. Furthermore, the local packages have three additional tools for further analysis. If

Table I. Comparison of GPS 2.1 and GPS 2.0

Kinase family	No selection[a]				Motif length selection[b]			
	Up[c]	Down[d]	LOO[e] (%)	Self[f] (%)	Up	Down	LOO (%)	Self (%)
AGC/AKT	7	7	92.06	97.22	5	1	92.46	92.46
AGC/DMPK/ROCK	7	7	52.08	97.92	9	1	64.58	87.50
AGC/PKC/Delta	7	7	44.44	100.00	3	2	65.56	84.44
AGC/PKG	7	7	73.96	100.00	4	2	82.29	89.58
AGC/RSK	7	7	76.79	100.00	8	2	82.14	95.24
CMGC/CDK/CDC2/CDC2	7	7	83.33	86.92	2	9	83.85	85.90
CMGC/MAPK/p38/MAPK14	7	7	65.22	96.38	2	1	71.74	72.46
STE/STE20	7	7	51.67	98.89	3	3	60.00	76.11
STE/STE20/PAKA	7	7	52.94	100.00	3	3	62.75	78.43
TKL/MLK/MLK/MAP3K11	7	7	44.44	100.00	1	1	66.67	88.89
Atypical/PIKK/FRAP	7	7	88.10	100.00	1	7	92.86	97.62
Other/AUR	7	7	53.25	90.91	4	1	61.90	64.50
TK/InsR/IGF1R	7	7	49.33	100.00	1	8	58.67	86.67
TK/Tec/BTK	7	7	70.37	98.15	5	2	83.33	85.19
Average			55.11	96.46			70.73	94.18

To avoid any bias, we retrained GPS 2.1 with the data set used in GPS 2.0, with 3161 kinase-specific phosphorylation sites (Phospho.ELM 6.0).
[a]No Selection, the results in GPS 2.0.
[b]Motif length selection, the performances in GPS 2.1.
[c]Up, upstream peptides of phosphorylated residue.
[d]Down, downstream peptides of phosphorylated residue.
[e]LOO, leave-one-out validation.
[f]Self, self-consistency result.

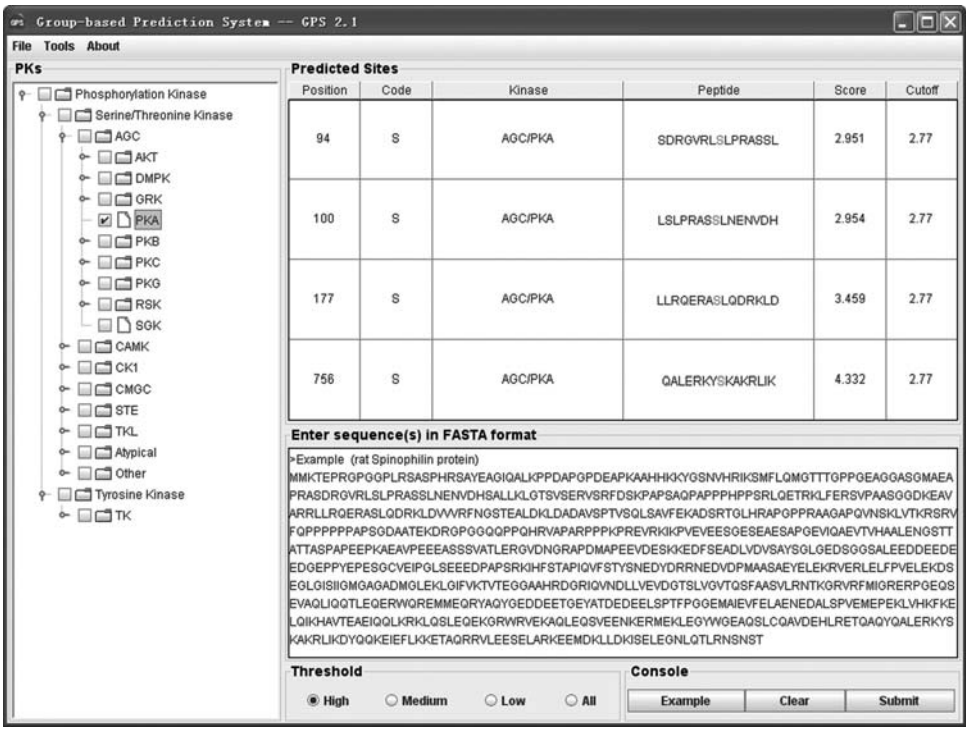

Fig. 2. The screen snapshot of GPS 2.1 software. One or multiple protein sequences should be prepared in FASTA format as the input data.

users have too many proteins for prediction (e.g. >1500 proteins), we recommend that users could install the local packages of GPS 2.1 and use the 'Batch Predictor' program in the Tools menu (Supplementary data, Fig. S2). Also, users could use the 'Performance' program in the Tools menu to view the self-consistency performances of GPS 2.1 (Supplementary data, Fig. S3). Previously, we designed a software of DOG 1.0 (Domain Graph) to visualize the organizations of protein domains, motifs and functional sites (Ren *et al*., 2009). This program was also integrated in GPS 2.1 as an auxiliary tool. Users could drag the mouse cursor in the field of prediction form, right-click to open a menu, and then left-click on the 'Visualization' button to visualize the prediction results (Supplementary data, Fig. S4). Furthermore, users could use the 'Domain Graph' tool in the Tools menu to diagram protein domain structures, motifs and phosphorylation sites directly. Finally, a detailed manual could be either directly downloaded from GPS 2.1 website or obtained in local packages.

Discussion

In the post-genomic era, dissection of gene/protein functions, regulatory roles and interactions has generated a new field of functional genomics and emerged as a great challenge. Identification of phosphorylation sites and elucidation of synergetic associations between PKs with their substrates are also important for understanding the genomic dynamics and plasticity (Kreegipuu *et al*., 1998; Blom *et al*., 2004; Kobe *et al*., 2005; Hjerrild and Gammeltoft, 2006; Ubersax and Ferrell, 2007; Miller and Blom, 2009). Contrasting to labor-intensive and time-consuming experimental approaches, numerous computational methods have also been developed as useful tools for their fast-speed, low-cost and convenience (Kobe *et al*., 2005; Hjerrild and Gammeltoft, 2006; Miller and Blom, 2009). Although a dozen of kinase-specific predictors were developed (http://gps.biocuckoo.org/links.php), the mathematical models and algorithms still remained to be improved.

To predict non-specific or kinase-specific phosphorylation sites, a widely adopted hypothesis is that a PK could recognize distinct sequence patterns/motifs of substrates by its kinase domain for modification (Kreegipuu *et al*., 1998; Blom *et al*., 2004; Kobe *et al*., 2005; Hjerrild and Gammeltoft, 2006; Ubersax and Ferrell, 2007; Miller and Blom, 2009). Thus, an informative phosphorylated motif for modification should be decided before training. For example, we defined a phosphorylation site peptide PSP(m, n) as a S/T or Y residue flanked by m residues upstream and n residues downstream. And similar or analogous terms were used in other researches (Kobe *et al*., 2005; Hjerrild and Gammeltoft, 2006; Miller and Blom, 2009). Previously, we and other researchers casually and arbitrarily defined the PSP(m, n) (Kobe *et al*., 2005; Xue *et al*., 2005; Hjerrild and Gammeltoft, 2006; Xue *et al*., 2008; Miller and Blom, 2009). However, in our previous studies, we observed that different combinations of modified peptides could generate different performance and robustness (Ren *et al*., 2008). Furthermore, in this work, we revealed that the self-consistency result will be always enhanced with longer PSP(m, n), although the LOO value will first reach a peak value then decrease.

In this regard, here we developed a novel but simple approach of MLS to detect optimal or near-optimal PSP(m, n) with the highest LOO validation. Although the LOO validation might be prone to overfitting for small samples, our results suggested that at least the performances of the LOO validation and n-fold cross-validation were quite similar for 72 PK groups with large samples ($n \geq 30$). The n-fold cross-validation was not chosen during training, because its result fluctuates due to resampling randomicity. With this newly

developed method, the current release of GPS 2.1 exhibit much higher robustness to narrow down the difference between the LOO and self-consistency to improve the robustness. Taken together, we proposed that GPS 2.1 will be a helpful tool for experimental researchers.

Supplementary data

Supplementary data are available at *PEDS* online.

Acknowledgments

The authors thank two anonymous reviewers, whose suggestions have greatly improved the presentation of this manuscript. 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 6 years. The authors are also grateful for helpful suggestions from Dr Zhaolei Zhang (U. Toronto).

Funding

This work was supported by grants from the National Basic Research Program (973 project) (2010CB945400, 2007CB947401), National Natural Science Foundation of China (90919001, 30700138, 30900835, 30830036, 31071154), and Chinese Academy of Sciences (INFO-115-C01-SDB4-36).

References

Blom,N., Sicheritz-Ponten,T., Gupta,R., Gammeltoft,S. and Brunak,S. (2004) *Proteomics*, **4**, 1633–1649.

Diella,F., Cameron,S., Gemund,C., Linding,R., Via,A., Kuster,B., Sicheritz-Ponten,T., Blom,N. and Gibson,T.J. (2004) *BMC Bioinformatics*, **5**, 79.

Diella,F., Gould,C.M., Chica,C., Via,A. and Gibson,T.J. (2008) *Nucleic Acids Res.*, **36**, D240–D244.

Dong,L., Yuan,Y. and Cai,Y. (2006) *J. Biomol. Struct. Dyn.*, **24**, 239–242.

Herwig,R., Poustka,A.J., Muller,C., Bull,C., Lehrach,H. and O'Brien,J. (1999) *Genome Res.*, **9**, 1093–1105.

Hjerrild,M. and Gammeltoft,S. (2006) *FEBS Lett.*, **580**, 4764–4770.

Kobe,B., Kampmann,T., Forwood,J.K., Listwan,P. and Brinkworth,R.I. (2005) *Biochim. Biophys. Acta*, **1754**, 200–209.

Kreegipuu,A., Blom,N., Brunak,S. and Jarv,J. (1998) *FEBS Lett.*, **430**, 45–50.

Miller,M.L. and Blom,N. (2009) *Methods Mol. Biol.*, **527**, 299–310.

Ren,J., Wen,L., Gao,X., Jin,C., Xue,Y. and Yao,X. (2008) *Protein Eng. Des. Sel.*, **21**, 639–644.

Ren,J., Wen,L., Gao,X., Jin,C., Xue,Y. and Yao,X. (2009) *Cell Res.*, **19**, 271–273.

Shao,J. (1993) *J. Am. Stat. Assoc.*, **88**, 486–494.

Soukas,A., Cohen,P., Socci,N.D. and Friedman,J.M. (2000) *Genes Dev.*, **14**, 963–980.

Ubersax,J.A. and Ferrell,J.E., Jr. (2007) *Nat. Rev. Mol. Cell Biol.*, **8**, 530–541.

Xue,Y., Zhou,F., Zhu,M., Ahmed,K., Chen,G. and Yao,X. (2005) *Nucleic Acids Res.*, **33**, W184–W187.

Xue,Y., Ren,J., Gao,X., Jin,C., Wen,L. and Yao,X. (2008) *Mol. Cell. Proteomics*, **7**, 1598–1608.