Genomics Proteomics Bioinformatics xxx (xxxx) xxx

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

WEB SERVER

GPS 5.0: An Update on the Prediction of Kinase-specific Phosphorylation Sites in Proteins

Chenwei Wang [1,#], Haodong Xu [1,#], Shaofeng Lin [1], Wankun Deng [1], Jiaqi Zhou [1], Ying Zhang [1], Ying Shi [1], Di Peng [1], Yu Xue [1,2,*]

[1] *Key Laboratory of Molecular Biophysics of Ministry of Education, Hubei Bioinformatics and Molecular Imaging Key Laboratory, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China*
[2] *Huazhong University of Science and Technology Ezhou Industrial Technology Research Institute, Ezhou 436044, China*

Received 9 August 2019; revised 18 November 2019; accepted 26 February 2020
Available online xxxx

Handled by Tzongyi Lee

KEYWORDS

Protein phosphorylation;
Protein kinase;
Group-based Prediction System;
Kinase-specific phosphorylation site;
Dual-specificity kinase

Abstract In eukaryotes, **protein phosphorylation** is specifically catalyzed by numerous **protein kinases** (PKs), faithfully orchestrates various biological processes, and reversibly determines cellular dynamics and plasticity. Here we report an updated algorithm of **Group-based Prediction System** (GPS) 5.0 to improve the performance for predicting **kinase-specific phosphorylation sites** (p-sites). Two novel methods, position weight determination (PWD) and scoring matrix optimization (SMO), were developed. Compared with other existing tools, GPS 5.0 exhibits a highly competitive accuracy. Besides serine/threonine or tyrosine kinases, GPS 5.0 also supports the prediction of **dual-specificity kinase**-specific p-sites. In the classical module of GPS 5.0, 617 individual predictors were constructed for predicting p-sites of 479 human PKs. To extend the application of GPS 5.0, a species-specific module was implemented to predict kinase-specific p-sites for 44,795 PKs in 161 eukaryotes. The online service and local packages of GPS 5.0 are freely available for academic research at http://gps.biocuckoo.cn.

Introduction

Protein phosphorylation plays a critical role in almost all of biological processes and greatly expands the proteome diversity. By covalently attaching phosphate moieties to serine, threonine, and/or tyrosine residues in a dynamic manner, phosphorylation can reversibly change the structure, enzymatic activity, and subcellular trafficking of proteins [1,2]. In eukaryotes, phosphorylation reaction is differentially and specifically catalyzed by numerous protein kinases (PKs), and each PK only modifies a limited subset of substrates to ensure the signaling fidelity [3–5]. Aberrances in either PKs or phosphorylated substrates are highly associated with human diseases such as cancer [6,7]. Therefore, the identification of kinase-specific phosphorylation sites (p-sites) is fundamental for understanding the regulatory mechanisms of phosphorylation.

Besides experiments, bioinformatics provides an alternative means for computational prediction of potential PK-specific

* Corresponding author.
E-mail: xueyu@hust.edu.cn (Xue Y).
# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2020.01.001

Please cite this article as: C. Wang, H. Xu, S. Lin et al., GPS 5.0: An Update on the Prediction of Kinase-specific Phosphorylation Sites in Proteins, Genomics Proteomics Bioinformatics, https://doi.org/10.1016/j.gpb.2020.01.001

p-sites from protein sequences [8–15] (Table S1). In 2004, we developed a novel algorithm, group-based phosphorylation site predicting and scoring (GPS) 1.0, based on the hypothesis that similar short peptides exhibit similar biological functions [8]. Accordingly, we refined the algorithm and constructed an online service of GPS 1.1, which can predict p-sites for 71 PK clusters [9]. Later, we presented GPS 2.0 and 2.1 (renamed as Group-based Prediction System), in which two methods matrix mutation (MaM) and motif length selection (MLS) were designed to improve the prediction accuracy, whereas the scoring strategy was not changed [10,11]. Using 3417 known PK-specific p-sites as a training data set, GPS 2.1 contains 213 individual predictors, and can hierarchically predict specific p-sites for 408 human PKs [11]. We also developed GPS 2.2, 3.0, and 4.0 algorithms, which are used for the prediction of post-translational modification (PTM) sites other than p-sites [16–18]. In particular, it should be noted that other bioinformaticians also put great efforts into the prediction of kinase-specific p-sites. At least 36 computational programs have been developed (Table S1).

In this study, we collected 15,194 experimentally identified PK-specific p-sites as the training data set. Updated from the GPS 2.1 algorithm, we replaced the MLS method by developing a new approach named position weight determination (PWD). PWD uses the logistic regression (LR) [19] to rapidly determine position-specific weight values of flanking sequences around p-sites. The LR algorithm is also used to modify the MaM method into scoring matrix optimization (SMO) for improving the accuracy of prediction. The leave-one-out (LOO) validation and *n*-fold cross-validations were conducted to evaluate the performance of GPS 5.0, which shows a highly competitive accuracy in comparison with other existing tools. In GPS 5.0, we separately constructed 466 and 93 individual predictors to computationally analyze phosphoserine (pS)/phosphothreonine (pT) and phosphotyrosine (pY) residues specifically modified by serine/threonine kinases and tyrosine kinases, respectively. Since a number of serine/threonine and tyrosine kinases also modify pY and pS/pT sites, respectively, we further constructed 58 additional predictors for these dual-specificity PKs. In GPS 5.0, we developed two modules including the classical module and the species-specific module. In the classical module, we constructed 617 single predictors for computationally identifying specific p-sites of 479 human PKs. The species-specific module can predict p-sites of 44,795 PKs in 161 eukaryotes. We anticipate GPS 5.0 can help to generate high-confidence candidates for the discovery of new phosphorylation events.

Method

During the past decade, the GPS algorithm has been continuously maintained and improved [8–11]. Our fundamental hypothesis is that similar short peptides bear similar biochemical properties for the modification. Thus, we defined a *phosphorylation site peptide* PSP(m, n) as a pS, pT, or pY amino acid flanked by m residues upstream and n residues downstream. Then we used an amino acid substitution matrix, *e.g.*, BLOSUM62, to calculate the similarity between two PSP(m, n) peptides. This basic scoring strategy has been reserved in all versions of GPS algorithms, although GPS 2.1 implemented two methods, MLS and MaM, for performance improvement [11]. For each PK cluster, MLS determines an optimal motif length around p-sites since different PKs recognize distinct motifs with different lengths, whereas MaM generates an optimal matrix for better estimating PSP(m, n) similarity. Since different positions around p-sites might contribute differentially to the phosphorylation specificity, GPS 2.2 added a method of weight training (WT) to determine a weight value for each position after the MLS manipulation [16]. To process large data sets, we added a k-means clustering procedure in GPS 3.0 to cluster PTM sites into multiple groups [17], whereas GPS 4.0 adopted a particle swarm optimization (PSO) to rapidly determine parameters in the steps of MLS, MaM, and/or WT [18].

Here, we hypothesized that long flanking regions around p-sites might be generally and differentially important for the recognition of PKs, which are bulky molecules to interact with phosphorylatable residues. Thus, the weight value at each position rather than the motif length could be directly and rapidly optimized by the LR algorithm [19]. Because the numbers of p-sites for most PK clusters are lower than 1000 (Tables S2 and S3), the k-means clustering is not necessary. In this regard, GPS 5.0 was updated from GPS 2.1, and comprises two parts, including the scoring strategy and performance improvement.

In the step of the scoring strategy, the average similarity score (S) between a PSP(30, 30) peptide P and peptides around all known p-sites in the training data set is defined as:

$$S = \frac{1}{N}\sum_{j=-30}^{L-31}\left(\sum_{i=1}^{N} M_{train}[P_j, T_{ij}]\right) \times W_j \quad (1)$$

where L is the length of the PSP(30, 30) peptide (L = 61 representing a relatively long flanking region). N is the number of known p-sites in the positive data set. T_{ij} is the amino acid at position j around a known p-site T_i (i = 1, 2, 3, …, N). W_j is the weight value of position j, and M_{train} denotes the optimized amino acid substitution matrix in this study.

The performance improvement procedure comprises two steps, and we updated MLS and MaM into PWD and SMO, respectively.

PWD

We first used the amino acid substitution matrix BLOSUM62 ($M_{BLOSUM62}$) to calculate an average similarity score at the position j of a PSP(30, 30) peptide P as $S_j^{'}$:

$$S_j^{'} = W_j \frac{1}{N}\sum_{i=1}^{N} M_{BLOSUM62}[P_j, T_{ij}]) \quad (2)$$

Initially, the weight value of each position W_j in the PSP(30, 30) peptide was set to 1. Then we used the one-vs-rest (OVR) classifier with the ridge (L2) penalty of the LR algorithm to optimize W_j values, by using the "newton-cg" solver in the class LogisticRegressionCV of scikit-learn v0.21.0 (https://scikit-learn.org/), an extensively used machine learning (ML) toolbox [19]. To avoid over-fitting, such a procedure was repeated for 100 times and 10-fold cross-validation was conducted to determine the inverse of regularization strength at each time. Receiver operating characteristic (ROC) curves were illustrated, and area under curve (AUC) values were calculated. The optimal W_j vectors were determined based on the highest AUC value:

$$W_j = W_{-30}, \cdots, W_{-1}, W_0, W_1, \cdots, W_{30} \quad (3)$$

Please cite this article as: C. Wang, H. Xu, S. Lin et al., GPS 5.0: An Update on the Prediction of Kinase-specific Phosphorylation Sites in Proteins, Genomics Proteomics Bioinformatics, https://doi.org/10.1016/j.gpb.2020.01.001

In order to evaluate position-specific contributions of flanking regions around p-sites for different PK clusters, the W_j vectors were normalized into −1 to 1 based on the maximum absolute value.

SMO

The average similarity score of an amino acid a in the given PSP(30, 30) peptide P and a residue b in peptides around all known p-sites is defined as S_{ab}:

$$S_{ab} = \frac{1}{N}\sum_{j=-30}^{L-31} C_j \times M_{BLOSUM62}[a,b] \times W_j \quad (4)$$

where C_j is the number of ab amino acid pairs at position j. In BLOSUM62, there are 24 types of characters including 20 types of amino acids and 4 non-canonical characters (B, aspartic acid or asparagine; Z, glutamic acid or glutamine; X, any one type of 20 amino acids; *, the ending of protein sequence). Thus, a number of $[24 \times (24 + 1)]/2 = 300$ unique S_{ab} scores ($S_{ab} = S_{ba}$) were generated. Then, the same LR algorithm was used to optimize all of S_{ab} scores to produce a new matrix M_{train}:

$$M_{train} = (S_{AA}, S_{AC}, S_{AD}, \cdots, S_{**})_{300} \quad (5)$$

Implementation

First, we took 3417 experimentally identified p-sites used in GPS 2.1 [11], and further conducted a literature curation to collect 10,225 site-specific kinase–substrate relations (ssKSRs). Also, we obtained 12,031 known PK-specific p-sites from the file "Kinase_Substrate_Dataset.gz" (Last modified on May 02, 2019) of PhosphoSitePlus (https://www.phosphosite.org/), a widely used phosphorylation database [20]. In total, our benchmark data set contained 23,195 ssKSRs for 15,194 unique p-sites (**Figure 1** and Table S4).

As previously described [10], we downloaded the hierarchical classifications of human PKs at various levels (group, family, subfamily, and single PK) from Kinase.com/KinBase (http://kinase.com/web/current/kinbase/genes/SpeciesID/9606/), thus far the best annotated resource for PKs [21]. Due to the fact that multiple aliases are present for each human PK, here we only used the standard gene names taken from iEKPD (http://iekpd.biocuckoo.org), which adopted the classification rationales of Kinase.com/KinBase to characterize and classify eukaryotic PKs at group and family levels [22]. Based on the regulatory PK information, we classified known PK-specific p-sites into different PK clusters at group, family, subfamily, and single PK levels. The PK clusters with <3 p-sites were not further considered. It is well known that serine/threonine and tyrosine kinases usually modify pS/pT and pY sites, respectively. However, we found that a considerable number of serine/threonine and tyrosine kinases could additionally phosphorylate pY and pS/pT sites with important functions, respectively. For example, interferon-induced, double-stranded RNA-activated protein kinase (EIF2AK2/PKR) in the Other/PEK family is a typical serine/threonine PK that phosphorylates a human tumor suppressor p53 on S392 through physical interaction to regulate gene expression [23].

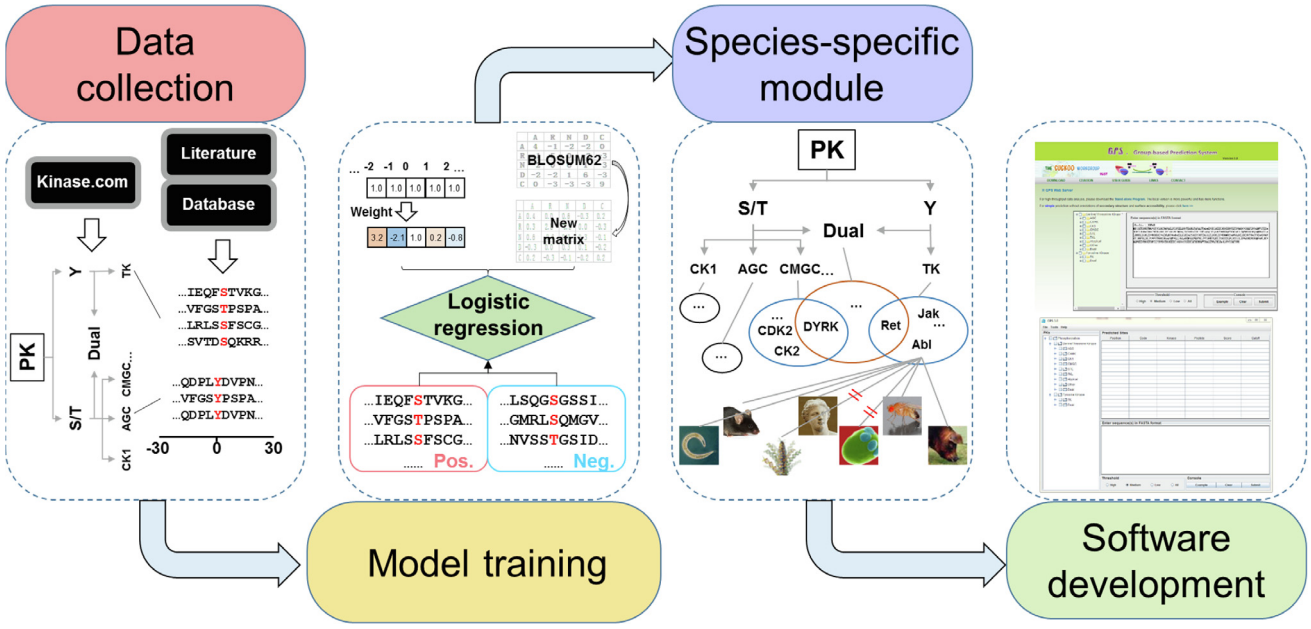

Figure 1 Experimental procedure of the study

First, experimentally identified PK-specific p-sites were collected from public databases and literature, and redundancy was cleared. Then, all known p-sites were hierarchically classified based on their upstream regulatory PKs. For model training, we updated the GPS 2.1 algorithm [11] by developing two logistic regression-based methods, PWD and SMO, which considerably improved the prediction accuracy. Besides the classical module, we also designed a species-specific module to extend the application of GPS 5.0. Finally, the online service and local packages of GPS 5.0 were implemented in PHP, JavaScript, and JAVA. PK, protein kinase; PWD, position weight determination; SMO, scoring matrix optimization. Pos., known p-sites for training; Neg., S/T or Y sites in positive phosphoproteins besides known p-sites.

EIF2AK2/PKR also exhibits tyrosine kinase activity and modifies human cyclin-dependent kinase 1 (CDK1) at Y4 to promote its ubiquitination and proteasomal degradation [24]. Moreover, human proto-oncogene tyrosine-protein kinase receptor RET in the tyrosine kinase (TK)/Ret family regulates the tyrosine kinase activity of focal adhesion kinase (FAK) through phosphorylating its Y576 and Y577 [25]. Human RET also modifies an important stress-responsive activating transcription factor 4 (ATF4) at four threonine residues, including T107, T114, T115, and T119, to inhibit ATF4-mediated apoptosis [26]. Thus, we added a class of "Dual" for the prediction of atypical p-sites of these dual-specificity PKs (Figure 1).

Then, the GPS 5.0 algorithm was adopted to individually train a computational model for each PK cluster. In the classical module, we totally constructed 617 single predictors for computationally identifying specific p-sites of 479 human PKs, including 58 predictors for the dual-specificity PKs (Figure 1). For the development of the species-specific module, eukaryotic PKs pre-classified at group and family levels were taken from iEKPD [22]. For each eukaryotic organism, PKs were reserved if their corresponding predictors at the family level could be obtained. In total, the species-specific module could predict the PK-specific p-sites of 44,795 PKs in 161 species (Figure 1).

As previously described [10], we randomly generated 10,000 PSP(30, 30) peptides based on the frequencies of the 20 amino acid residues in the training data set to estimate the false positive rate (FPR) of predictions. For each PK cluster, the process was repeated 20 times, and the average value was determined as the final FPR. The high, medium, and low thresholds were adopted with FPRs of 2%, 6%, and 10%, respectively, for serine/threonine kinases. Likewise, FPRs of 4%, 9%, and 15% were adopted for high, medium, and low thresholds for tyrosine kinases. The online service of the classical module of GPS 5.0 was implemented in PHP and JavaScript. We also integrated two web servers, IUPred [27] and NetSurfP [28], to predict surface accessibilities, disorder regions, and secondary structures of inputted proteins. The stand-alone packages of GPS 5.0 were developed in JAVA for supporting three major operation systems including Windows, Linux, and Mac OS (Figure 1).

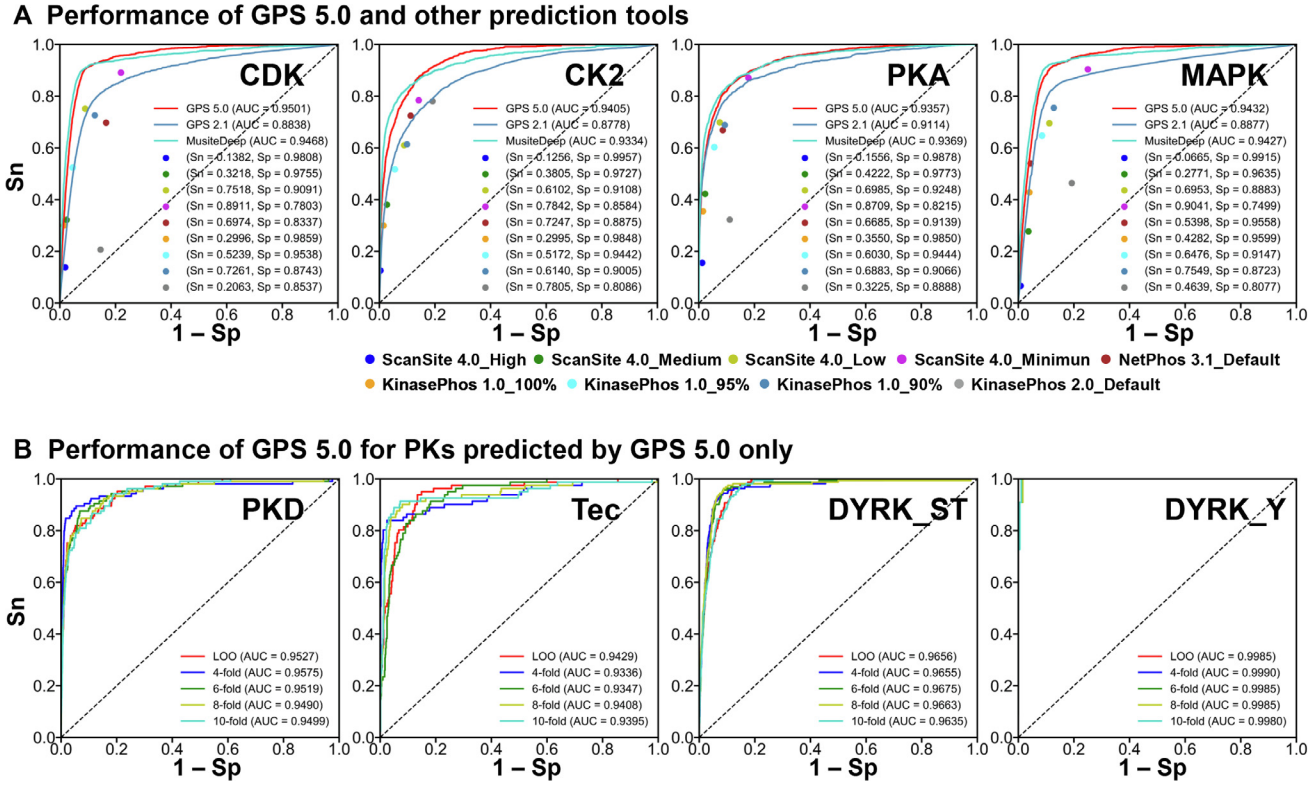

Figure 2 Performance evaluation of GPS 5.0

A. Prediction performance comparison between GPS 5.0 (red line) and other existing predictors, including GPS 2.1 [11] (cyan line), ScanSite 4.0 [12] (blue, green, yellow green, and purple dots), NetPhos3.1 [14] (brown dot), KinasePhos 1.0 [15] (yellow, light blue, and cyan dots), KinasePhos 2.0 [29] (gray dot), and MusiteDeep [13] (light blue line) for four PK families, including CDK (CMGC/CDK), CK2 (CMGC/CK2), PKA (AGC/PKA), and MAPK (CMGC/MAPK). **B.** ROC curves and AUC values of GPS 5.0 for a number of PKs that could only be predicted by GPS 5.0, including CAMK/PKD, TK/Tec, CMGC/DYRK (DYRK_ST), and dual/CMGC/DYRK (DYRK_Y). The ROC curves of LOO, 4-fold, 6-fold, 8-fold, and 10-fold cross-validations are colored in red, blue, green, yellow green, and cyan, respectively. CDK, cyclin-dependent kinase; CMGC, [CDK, mitogen-activated protein kinase (MAPK), glycogen synthase kinase (GSK) and CDC-like kinase (CLK)]; CK2, casein kinase 2; AGC, protein kinase A, G, and C; PKA, protein kinase A; CAMK, CaM kinase; PKD, protein kinase D; TK, tyrosine kinase; Tec, tyrosine kinase expressed in hepatocellular carcinoma; DYRK, dual-specificity tyrosine phosphorylation-regulated kinase; Sn, sensitivity; Sp, specificity; ROC, receiver operating characteristic; AUC; area under curve; LOO, leave-one-out.

Performance evaluation and comparison

As previously described [10], four standard measurements including accuracy (Ac), sensitivity (Sn), specificity (Sp), and Matthews correlation coefficient (MCC) were adopted to evaluate the performance and robustness of the GPS 5.0. The self-consistency validations were calculated for all PK clusters (Tables S2 and S3)). Also, 10-fold cross-validations were performed for 245 PK categories wit $\geq$30 p-sites (Table S2), and LOO validations were conducted for other PK categories (Table S3). The high congruence of different validation results indicated the promising accuracy and robustness of GPS 5.0 (Tables S2 and S3)).

To further demonstrate the superiority of GPS 5.0, we compared the prediction performance of GPS 5.0 with that of other existing predictors, such as GPS 2.1 [11], ScanSite 4.0 [12], NetPhos3.1 [14], KinasePhos 1.0 [15], KinasePhos 2.0 [29], and MusiteDeep [13] (**Figure 2**). Due to the page limitation, four typical PK families including CDK, casein kinase 2 (CK2), protein kinase A (PKA), and mitogen-activated protein kinase (MAPK) were selected for demonstration (Figure 2). For each PK family, we directly submitted its corresponding training data set into each tool to calculate the performance and compared with the 10-fold cross-validation result of GPS 5.0. The ROC curves of GPS 2.1 [11] and MusiteDeep [13] were illustrated, while the Sn and 1 − Sp values of ScanSite 4.0 [12] were calculated at high, medium, low, and minimum thresholds separately. The default cut-off scores of NetPhos3.1 [14] and KinasePhos 2.0 [29] were adopted, whereas Sn values of KinasePhos 1.0 [15] with Sp at 100%, 95%, and 90% were computed separately. As shown in Figure 2A, we found that GPS 5.0 achieved a highly competitive accuracy with MusiteDeep, a deep learning-based predictor [13]. The prediction performance of GPS 5.0 was much better than that of other tools, including GPS 2.1 [11] (Figure 2A). It should be noted that MusiteDeep only constructed 5 PK-specific predictors at the family level, whereas GPS 5.0 could predict for much more PK families, such as CaM kinase/protein kinase D (CAMK/PKD) and TK/Tec (Figure 2B). The pS/pT and pY sites differentially modified by dual-specificity tyrosine phosphorylation-regulated kinase (DYRK) could also be accurately predicted (Figure 2B).

For the four PK families of CDK, CK2, PKA, and MAPK, we further compared the performance of the two new methods in GPS 5.0 with that of previous approaches implemented in GPS 2.1. For each PK family, the AUC values of MLS, PWD, MaM, and SMO were exclusively calculated from the 10-fold cross-validations. Our results demonstrate that PWD and SMO perform better than MLS and MaM in p-site prediction as indicated by higher AUC values for all four PK families tested (**Figure 3**A). In addition, three ML algorithms in scikit-learn, including support vector machines (SVMs), random for-

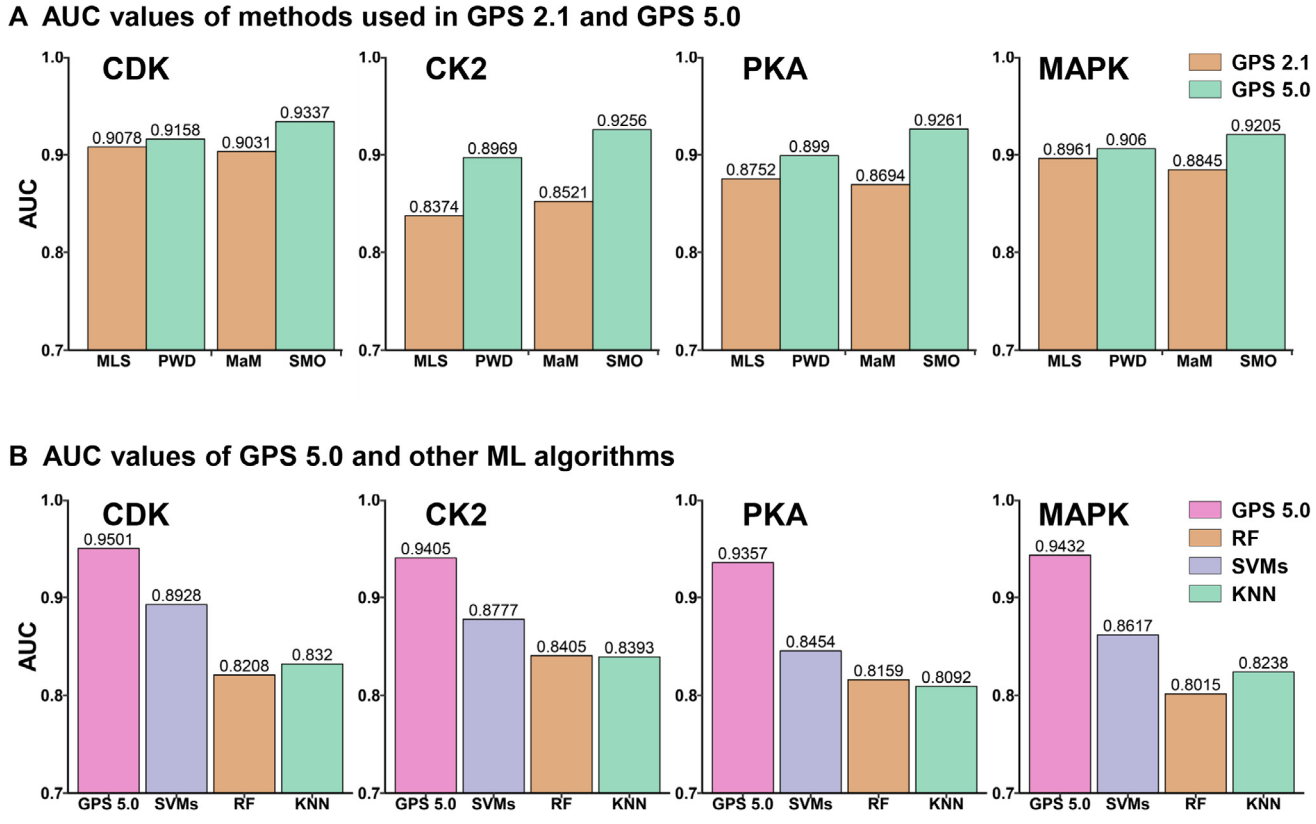

Figure 3 Performance comparison between GPS 5.0, GPS 2.1, and other ML algorithms

A. AUC values of MLS, PWD, MaM, and SMO for the four PK families including CDK, CK2, PKA, and MAPK. MLS and MaM are used in GPS 2.1 (brown), whereas PWD and SMO are developed to replace MLS and MaM in GPS 5.0 (green). **B.** AUC values of GPS 5.0 (pink) and three types of ML algorithms including SVMs (purple), RF (brown), and KNN (green) for the four PK families. AUC values were calculated from the 10-fold cross-validations. MLS, motif length selection; MaM, matrix mutation; ML, machine learning; SVM, support vector machine; RF, random forest; KNN, *k*-nearest neighbor.

est (RF), and *k*-nearest neighbor (KNN), were adopted for training models and compared with GPS 5.0. As shown in Figure 3B, GPS5.0 achieved higher AUC values of the 10-fold cross-validations than other algorithms for all four PK families tested.

Usage of GPS 5.0

For convenience, the stand-alone packages of GPS 5.0 are recommended (**Figure 4**). The main interface is the classical module of GPS 5.0, which contains three parts, including the hierarchical classification tree of PK categories shown in the left panel (Figure 4A), which enables the selection of PKs at four levels including group, family, sub-family, and single PK. In the lower-right panel, users could provide one or multiple protein sequences in FASTA format and select a threshold (Figure 4B). By left-clicking on the 'Submit' button, the prediction results will be presented in the upper-right panel as a tabular list containing position, code, PK, flanking peptide, score, and cutoff for a predicted p-site (marked in red) (Figure 4C). Alternatively, user could load a demo sequence by clicking the 'Example' button, or clear the inputs by clicking the 'Clear' button (Figure 4).

In GPS 5.0, human Beclin-1, an important autophagy-related (ATG) protein and tumor suppressor [30,31], was chosen as an example for the prediction of kinase-specific p-sites. It has been reported that the S234 and S295 of Beclin-1 are phosphorylated by Akt, which inhibits autophagy by regulating the interaction between Beclin-1 and 14-3-3 proteins [30]. The predictions of GPS 5.0 are highly consistent with experimental results. Two additional p-sites, S10 and S90, were predicted under a medium threshold. Whether the two p-sites are really phosphorylated by Akt remains to be experimentally validated.

Also, GPS 5.0 web server was developed in a user-friendly manner (**Figure 5**A, Figure S1). For each PK predictor, a sequence logo is illustrated by the R package ggseqlogo [32] with the PSP(30, 30) items of its positive data set, and a simplified logo icon is added for each prediction result (Figure S1A). A column entitled "Source" was added to denote whether a potential ssKSR was previously reported by the literature (Exp.) or just a prediction (Pred.) (Figure S1A). Besides the presentation and statistics of the predicted results, structural features such as secondary structures, surface accessibilities, and disorder regions could also be predicted and shown by IUPred and NetSurfP (Figure S1). In IUPred, the disorder propensity values range from 0 to 1, and an amino acid residue with a calculated score

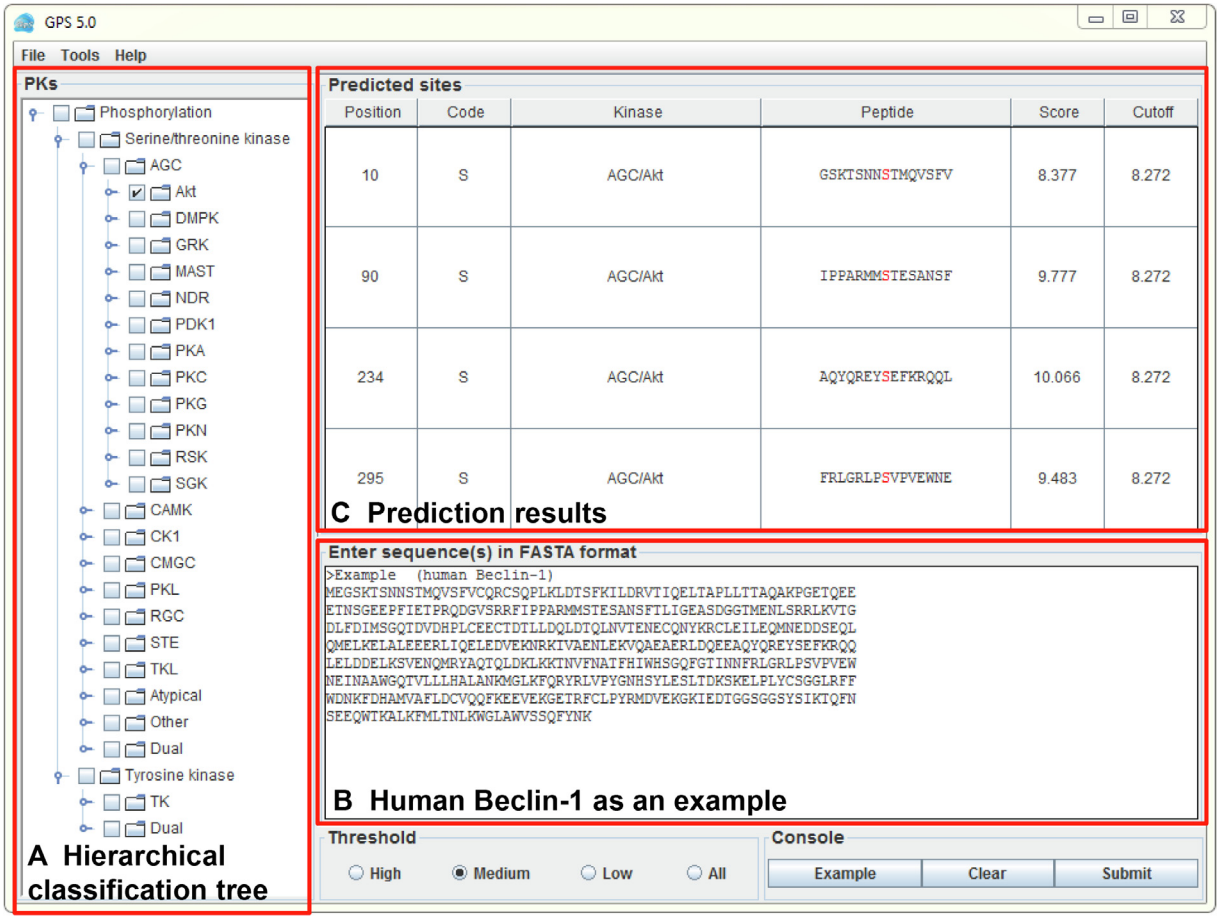

Figure 4 Interface of the classical module in GPS 5.0 software package

A. The hierarchical classification tree of individual PK predictors. **B.** The protein sequence of human Beclin-1 is presented as an example. **C.** The prediction results are shown in a tabular format, including positions, amino acid types, regulatory PKs, flanking regions, predicted scores, and pre-defined cutoff values for predicted kinase-specific p-sites.

> 0.5 would be considered as disordered [27]. In NetSurfP, the relative surface area (RSA) was calculated for measuring the surface accessibility, and an amino acid with an RSA value > 0.25 would be taken as an exposed residue [28]. From protein sequences, NetSurfP could also predict three types of potential secondary structures, including α-helix, β-strand, and coil, for each amino acid residue [28].

To further exhibit the superiority of GPS 5.0, other known PKs that phosphorylate human Beclin-1 were collected from the literature [31,33–36]. Unc-51-like kinase 1 (ULK1) has been reported to induce autophagy through phosphorylating Beclin-1 at S15 [33], while S90 is phosphorylated by CAMK2 to promote the ubiquitination of Beclin-1 for the activation of autophagy [34]. Also, it is known that two pY sites (Y229 and Y233) in Beclin-1 are phosphorylated by epidermal growth factor receptor (EGFR), which is primarily responsible for the suppression of autophagy [31]. The prediction results of GPS 5.0 covered most of the known kinase-specific p-sites. Moreover, two additional p-sites, S64 and S177 of Beclin-1, were predicted by GPS5.0 to be specifically modified by ULK1 (Figures 5B and S1).

Future developments

In this study, we updated a highly useful tool named GPS 5.0 for the prediction of PK-specific p-sites, including a classical module (Figure 4) and a species-specific module (Figure S2). In the former, there were 617 individual predictors constructed for predicting p-sites of 479 human PKs, whereas kinase-specific p-sites of 44,795 PKs could be predicted for 161 eukaryotes in the latter. In GPS 5.0, two novel methods, PWD and SMO, were developed to improve the training efficiency and performance of the previous developed GPS 2.1 algorithm [11].

For each PK predictor, the information content R_i was calculated in bits for the position i in the alignment as previously described [37]:

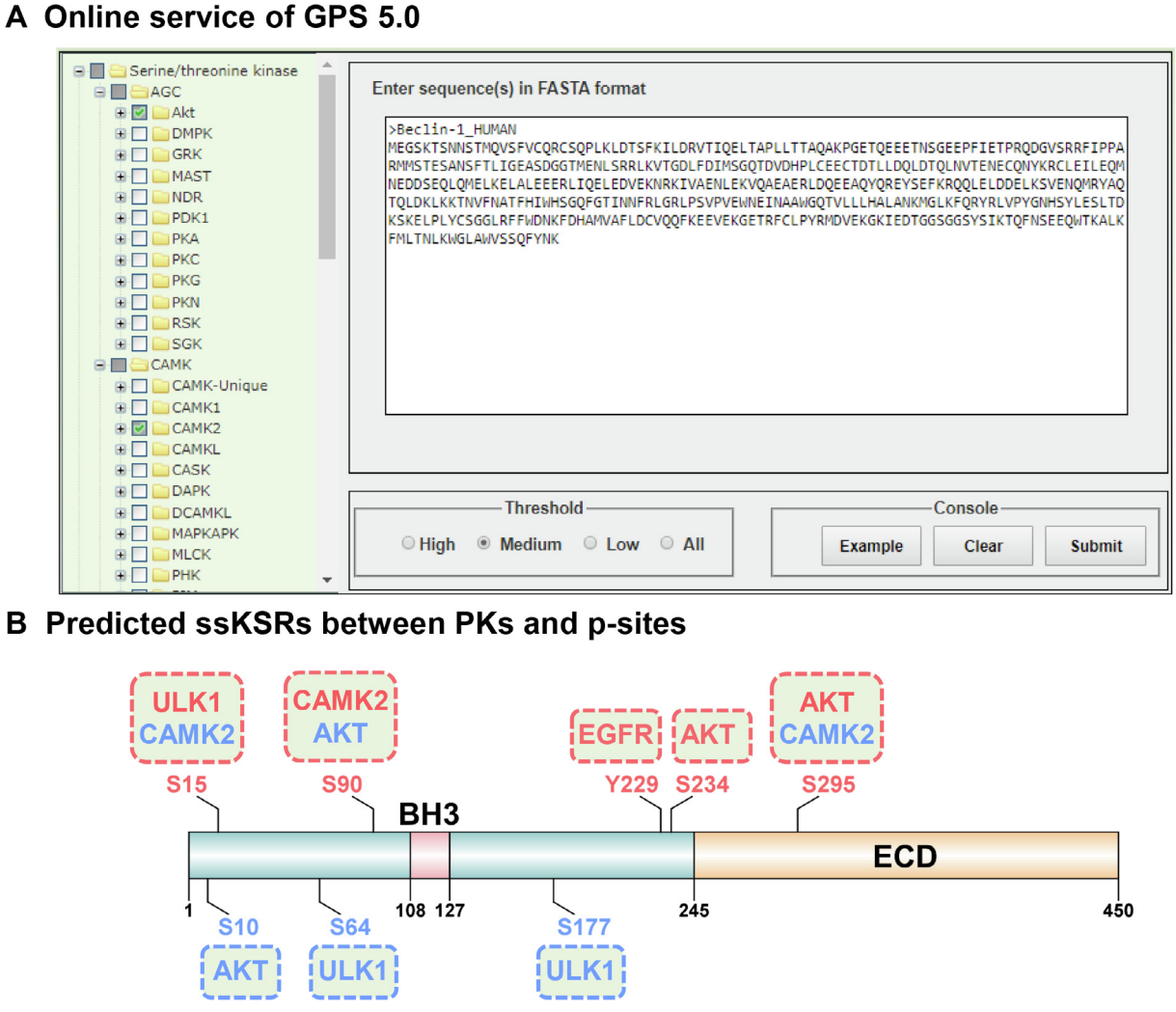

Figure 5 Prediction of kinase-specific p-sites in human Beclin-1

A. Multiple PK predictors including AGC/Akt, CAMK/CAMK2, TK/EGFR, and Other/ULK/ULK/ULK1 were selected in the online service of GPS 5.0. Such a manipulation could also be carried out in GPS local packages. **B.** Predicted ssKSRs between PKs and p-sites. Known PKs and p-sites are shown in red, while newly predicted PKs and p-sites are marked in blue. ssKSR, site-specific kinase–substrate relations; EGFR, epidermal growth factor receptor; ULK, Unc-51-like kinase 1.

$$R_i = E_{All} - E_{Pos} = -\sum_{n=1}^{20} p_n log_2 p_n - (-\sum_{n=1}^{20} q_n log_2 q_n) \quad (6)$$

where E_{All} and E_{Pos} indicate Shannon entropies measured from the PSP(30, 30) items in all phosphorylated proteins and in the PK-specific positive data set, respectively. The symbol n denotes one of the 20 types of amino acid residues, whereas p_n and q_n indicate the observed frequencies of n estimated from the background and foreground data sets, respectively. The R_i values were separately calculated for serine/threonine PKs and tyrosine PKs, while the middle p-sites were not included for the computation. Then, Pearson correlation coefficient (PCC) values are pairwisely calculated between the *Ri* scores and the outputs of PWD training processes for 617 individual PK predictors (Table S5). The average PCC value was 0.606, which was increased to 0.656 if only PK predictors with ≥30 p-sites were considered. Several instances for the correlation between the information content and the PWD output were shown (Figure S3). For example, weight values for AGC/Akt at positions −5 and − 3 were determined as 0.8454 and 1.0000. Such a result follows the canonical motif R-X-R-X-X-S/T of the Akt family [38] (Figure S3). Also, only the position +1 for Atypical/PIKK/ATM was determined as 1.0000, which is consistent with the S/T-Q motif of ATM/ATR [39]. Moreover, the weight values of 0.5677 and 1.0000 at positions −2 and +1 are consistent with the P-X-S/T-P motif of CMGC/MAPK [40], and the weight value of 1.0000 at the +3 position supports a known motif Y-X-X-P of TK/Abl [41] (Table S5 and Figure S3). In this regard, higher position weights derived from PWD are generally consistent with information contents of known PK consensus motifs.

Since December 2004, the online service of GPS 1.1 and local packages of GPS 2.0 as well as 2.1 have been freely accessible to academic use for nearly 15 years [9–11]. In future, GPS 5.0 will be continuously maintained and improved. The computational models will be updated if new experimentally characterized kinase-specific p-sites are available. In addition, we are currently testing various types of methods including both traditional ML algorithms and deep-learning algorithms, which will hopefully further improve the accuracy of GPS. It is also worth mentioning that only sequence information has been considered at the current stage, and we will test structural features and further integrate both sequence and structural features to improve the performance. We believe that GPS 5.0 could serve as a high-profile tool and provide useful information for further studies of phosphorylation.

Availability

The online service and local packages of GPS 5.0 were freely available for academic use at http://gps.biocuckoo.cn. Species-specific predictions are available either from the web server at http://gps.biocuckoo.cn/online_species.php, or in GPS 5.0 software packages by clicking 'Tools' in the menu bar. The benchmark data set for training was provided in Table S4.

Authors' contributions

YX conceived, designed, and supervised the study. CW and HX collected the data, designed the algorithm, performed the analysis, as well as constructed the online service and local packages. SL, WD, JZ, YZ, YS, and DP contributed to data analysis. YX, CW, and HX wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

Funding for open access charge: Special Project on Precision Medicine under the National Key R&D Program of China (Grant Nos. 2017YFC0906600 and 2018YFC0910500), National Natural Science Foundation of China (Grant Nos. 31671360, 81701567, and 31801095), National Program for Support of Top-Notch Young Professionals, Changjiang Scholars Program of China. This study is also supported by the program for HUST Academic Frontier Youth Team, Fundamental Research Funds for the Central Universities, China (Grant Nos. 2017KFXKJC001 and 2019kfyRCPY043), and China Postdoctoral Science Foundation (Grant Nos. 2018M642816 and 2018M632870).

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2020.01.001.

References

[1] Swaffer MP, Jones AW, Flynn HR, Snijders AP, Nurse P. CDK substrate phosphorylation and ordering the cell cycle. Cell 2016;167:1750–61.
[2] Xu XQ, Xu J, Wu JC, Hu Y, Han YM, Gu Y, et al. Phosphorylation-mediated IFN-gamma R2 membrane translocation is required to activate macrophage innate response. Cell 2018;175:1336–51.
[3] Cohen P. The origins of protein phosphorylation. Nat Cell Biol 2002;4:E127–30.
[4] Hunter T. Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. Cell 1995;80:225–36.
[5] Ubersax JA, Ferrell Jr JE. Mechanisms of specificity in protein phosphorylation. Nat Rev Mol Cell Biol 2007;8:530–41.
[6] Casado P, Rodriguez-Prados JC, Cosulich SC, Guichard S, Vanhaesebroeck B, Joel S, et al. Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. Sci Signaling 2013;6:rs6.
[7] Drake JM, Paull EO, Graham NA, Lee JK, Smith BA, Titz B, et al. Phosphoproteome integration reveals patient-specific networks in prostate cancer. Cell 2016;166:1041–54.
[8] Zhou FF, Xue Y, Chen GL, Yao X. GPS: a novel group-based phosphorylation predicting and scoring method. Biochem Biophys Res Commun 2004;325:1443–8.
[9] Xue Y, Zhou F, Zhu M, Ahmed K, Chen G, Yao X. GPS: a comprehensive www server for phosphorylation sites prediction. Nucleic Acids Res 2005;33:W184–7.
[10] Xue Y, Ren J, Gao X, Jin C, Wen L, Yao X. GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. Mol Cell Proteomics 2008;7:1598–608.
[11] Xue Y, Liu Z, Cao J, Ma Q, Gao X, Wang Q, et al. GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with

an algorithm of motif length selection. Protein Eng Des Sel 2011;24:255–60.
[12] Obenauer JC, Cantley LC, Yaffe MB. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res 2003;31:3635–41.
[13] Wang D, Zeng S, Xu C, Qiu W, Liang Y, Joshi T, et al. MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. Bioinformatics 2017;33:3909–16.
[14] Blom N, Sicheritz-Ponten T, Gupta R, Gammeltoft S, Brunak S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics 2004;4:1633–49.
[15] Huang HD, Lee TY, Tzeng SW, Horng JT. KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. Nucleic Acids Res 2005;33:W226–9.
[16] Liu Z, Yuan F, Ren J, Cao J, Zhou Y, Yang Q, et al. GPS-ARM: computational analysis of the APC/C recognition motif by predicting D-boxes and KEN-boxes. PLoS One 2012;7:e34370.
[17] Deng W, Wang Y, Ma L, Zhang Y, Ullah S, Xue Y. Computational prediction of methylation types of covalently modified lysine and arginine residues in proteins. Brief Bioinform 2017;18:647–58.
[18] Zhao Q, Xie Y, Zheng Y, Jiang S, Liu W, Mu W, et al. GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs. Nucleic Acids Res 2014;42:W325–30.
[19] Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. Front Neuroinform 2014:8–14.
[20] Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res 2015;43:D512–20.
[21] Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science 2002;298:1912–34.
[22] Guo Y, Peng D, Zhou J, Lin S, Wang C, Ning W, et al. iEKPD 2.0: an update with rich annotations for eukaryotic protein kinases, protein phosphatases and proteins containing phospho-protein-binding domains. Nucleic Acids Res 2019;47:D344–50.
[23] Cuddihy AR, Wong AHT, Tam NWN, Li SY, Koromilas AE. The double-stranded RNA activated protein kinase PKR physically associates with the tumor suppressor p53 protein and phosphorylates human p53 on serine 392 *in vitro*. Oncogene 1999;18:2690–702.
[24] Yoon CH, Miah MA, Kim KP, Bae YS. New Cdc2 Tyr 4 phosphorylation by dsRNA-activated protein kinase triggers Cdc2 polyubiquitination and G2 arrest under genotoxic stresses. EMBO Rep 2010;11:393–9.
[25] Plaza-Menacho I, Morandi A, Mologni L, Boender P, Gambacorti-Passerini C, Magee AI, et al. Focal Adhesion Kinase (FAK) binds RET kinase via its FERM domain, priming a direct and reciprocal RET-FAK transactivation mechanism. J Biol Chem 2011;286:17292–302.
[26] Bagheri-Yarmand R, Sinha KM, Gururaj AE, Ahmed Z, Rizvi YQ, Huang SC, et al. A novel dual kinase function of the RET proto-oncogene negatively regulates activating transcription factor 4-mediated apoptosis. J Biol Chem 2015;290:11749–61.
[27] Dosztanyi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 2005;21:3433–4.
[28] Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct Biol 2009;9:51.
[29] Wong YH, Lee TY, Liang HK, Huang CM, Wang TY, Yang YH, et al. KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. Nucleic Acids Res 2007;35:W588–94.
[30] Wang RC, Wei Y, An Z, Zou Z, Xiao G, Bhagat G, et al. Akt-mediated regulation of autophagy and tumorigenesis through Beclin 1 phosphorylation. Science 2012;338:956–9.
[31] Wei Y, Zou Z, Becker N, Anderson M, Sumpter R, Xiao G, et al. EGFR-mediated Beclin 1 phosphorylation in autophagy suppression, tumor progression, and tumor chemoresistance. Cell 2013;154:1269–84.
[32] Wagih O. ggseqlogo: a versatile R package for drawing sequence logos. Bioinformatics 2017;33:3645–7.
[33] Russell RC, Tian Y, Yuan H, Park HW, Chang YY, Kim J, et al. ULK1 induces autophagy by phosphorylating Beclin-1 and activating VPS34 lipid kinase. Nat Cell Biol 2013;15:741–50.
[34] Li X, Wu XQ, Deng R, Li DD, Tang J, Chen WD, et al. CaMKII-mediated Beclin 1 phosphorylation regulates autophagy that promotes degradation of Id and neuroblastoma cell differentiation. Nat Commun 2017;8:1159.
[35] Zalckvar E, Berissi H, Mizrachy L, Idelchuk Y, Koren I, Eisenstein M, et al. DAP-kinase-mediated phosphorylation on the BH3 domain of beclin 1 promotes dissociation of beclin 1 from Bcl-XL and induction of autophagy. EMBO Rep 2009;10:285–92.
[36] Fujiwara N, Usui T, Ohama T, Sato K. Regulation of Beclin 1 protein phosphorylation and autophagy by protein phosphatase 2A (PP2A) and death-associated protein kinase 3 (DAPK3). J Biol Chem 2016;291:10858–66.
[37] Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res 2004;14:1188–90.
[38] Manning BD, Cantley LC. AKT/PKB signaling: navigating downstream. Cell 2007;129:1261–74.
[39] Matsuoka S, Ballif BA, Smogorzewska A, McDonald 3rd ER, Hurov KE, Luo J, et al. ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. Science 2007;316:1160–6.
[40] Carlson SM, Chouinard CR, Labadorf A, Lam CJ, Schmelzle K, Fraenkel E, et al. Large-scale discovery of ERK2 substrates identifies ERK-mediated transcriptional regulation by ETV3. Sci Signal 2011;4:rs11.
[41] Andoniou CE, Thien CB, Langdon WY. The two major sites of cbl tyrosine phosphorylation in abl-transformed cells select the crkL SH2 domain. Oncogene 1996;12:1981–9.

Please cite this article as: C. Wang, H. Xu, S. Lin et al., GPS 5.0: An Update on the Prediction of Kinase-specific Phosphorylation Sites in Proteins, Genomics Proteomics Bioinformatics, https://doi.org/10.1016/j.gpb.2020.01.001