GPS Manual

Group-based Prediction System

Version 5.0
20/7/2019

Author: Chenwei Wang, Haodong Xu and Yu Xue*

Contact: Chenwei Wang, wangchenwei@hust.edu.cn, Yu Xue, xueyu@hust.edu.cn

Index

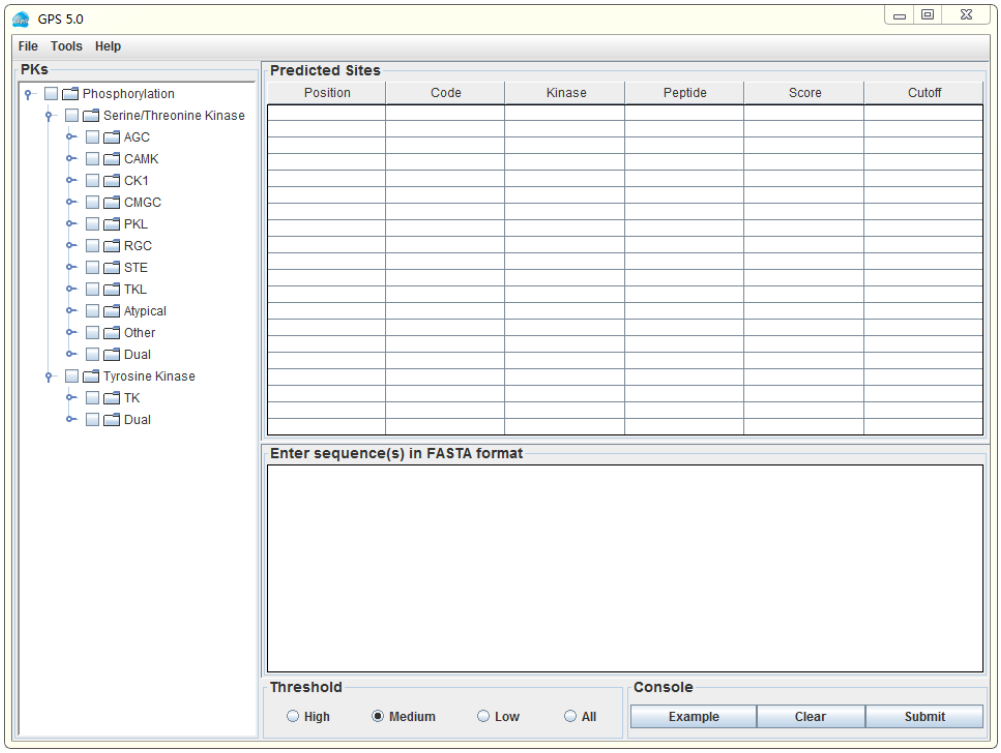
Statement

1. **Implementation**. The softwares of the CUCKOO Workgroup are implemented in JAVA (J2SE). Usually, both of online service and local stand-alone packages will be provided.
2. **Availability**. Our softwares are freely available for academic researches. For non-profit users, you can copy, distribute and use the softwares for your scientific studies. Our softwares are not free for commercial usage.
3. **GPS**. Previously, we used the GPS to denote our Group-based Phosphorylation Scoring algorithm. Currently, we are developing an integrated computational platform for post-translational modifications (PTMs) of proteins. We re-denote the GPS as Group-based Prediction Systems. This software is an indispensable part of GPS.
4. **Usage**. Our softwares are designed in an easy-to-use manner. Also, we invite you to read the manual before using the softwares.
5. **Updation**. Our softwares will be updated routinely based on users' suggestions and advices. Thus, your feedback is greatly important for our future updation. Please do not hesitate to contact with us if you have any concerns.
6. **Citation**. Usually, the latest published articles will be shown on the software websites. We wish you could cite the article if the software has been helpful for your work.
7. **Acknowledgements.** Funding for open access charge: Special Project on Precision Medicine under the National Key R&D Program [2017YFC0906600, 2018YFC0910500]; Natural Science Foundation of China [31671360, 81701567, 31801095]; National Program for Support of Top-Notch Young Professionals; Changjiang Scholars Program of China; The program for HUST Academic Frontier Youth Team; China Postdoctoral Science Foundation [2018M642816, 2018M632870].

Introduction

Identification of phosphorylation sites with their cognate protein kinases (PKs) is the foundation for understanding the functional dynamics and plasticity of various cellular processes. Although nearly 10 kinase-specific predictors were developed, numerous PKs were casually classified into sub-groups without a standard rule. And for large-scale predictions, the false positive rate (FPR) was also never addressed. Here we adopted a well-established rule to classify PKs with their verified sites into a hierarchical structure with four levels, including group, family, subfamily and single PK [1]. Then we constructed the **GPS** (**Group-based Prediction System**, ver 2.0) software, with a modified version of GPS (Group-based Phosphorylation Scoring) algorithm [2, 3]. As the first stand-alone software for computational phosphorylation, GPS 2.0 was implemented in JAVA and could predict kinase-specific phosphorylation sites for **408** human PKs in hierarchy.

Currently, as more and more phosphorylation sites have been experimentally validated, a more accurate and reliable kinases-specific phosphorylation sites prediction tool is in urgent need. At the same time, several advanced GPS algorithm have been developed and achieved higher performance, including GPS 2.1 and GPS 2.2. To improve the performance of phosphorylation sites prediction, we applied the GPS 2.2 algorithm and enlarged the training set. In addition, we included the phosphorylation sites prediction in 161 species. For dul-specific kinases, we also provided the prediction for these kinases. Finally, we constructed the **GPS** (**Group-based Prediction System**, ver 5.0) software package.

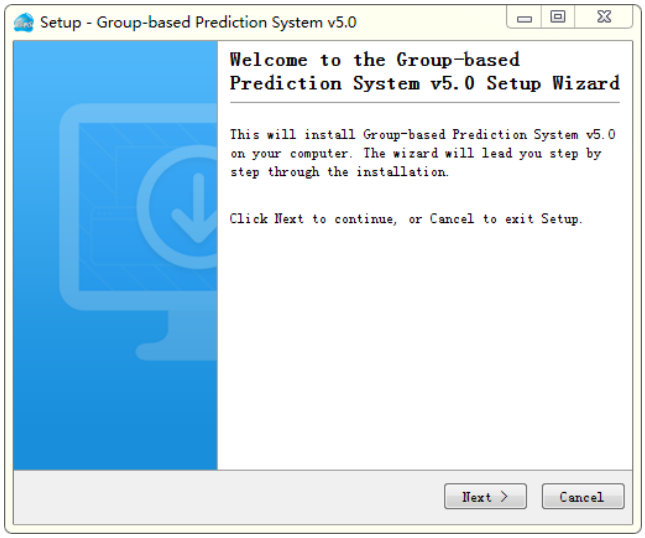


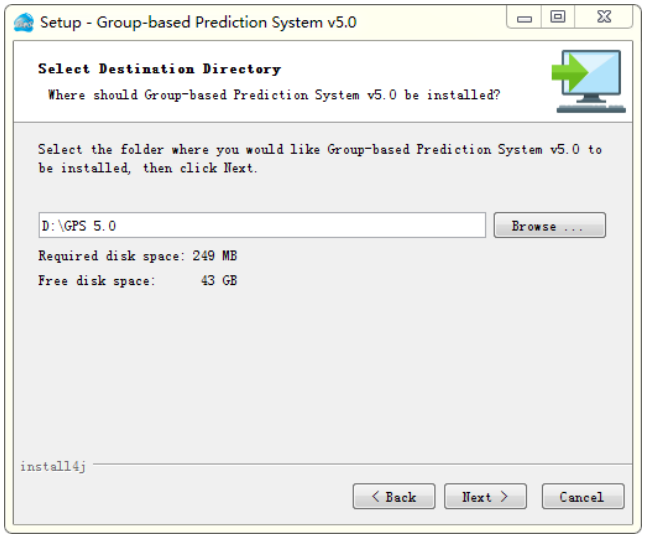
Group-based Prediction System v5.0 User Interface

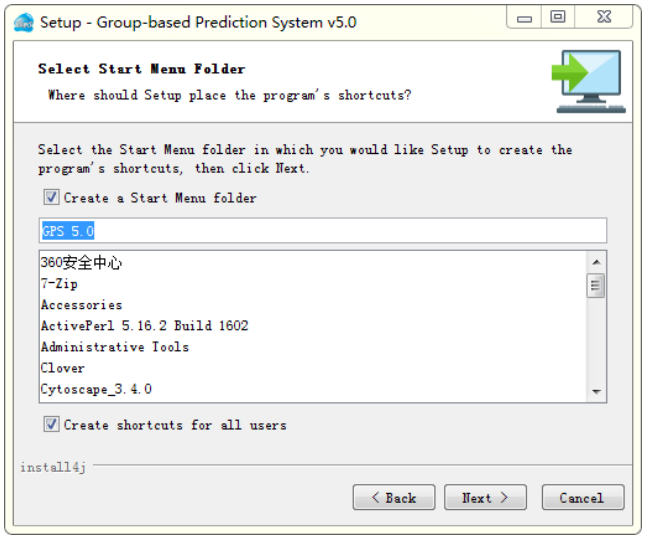
Download & Installation

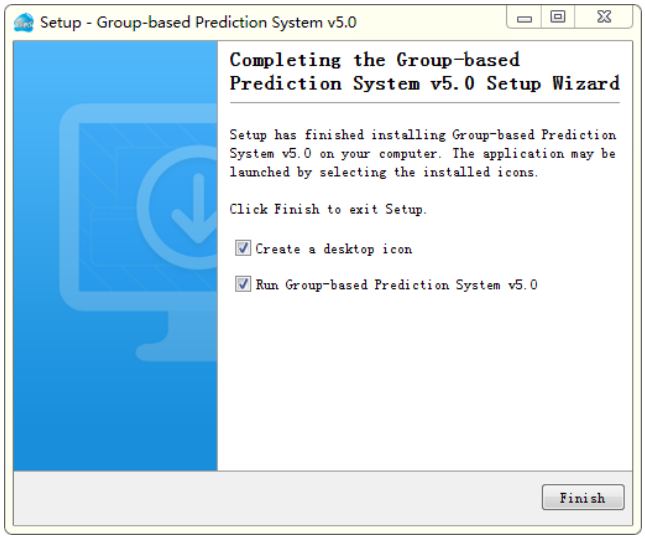
The software of GPS 5.0 was implemented in JAVA, and could be installed on Windows systems. GPS 5.0 distributions for Windows can be found at http://gps.biocuckoo.cn/download.php. We recommend that users could download the latest release.

After downloading, please double-click on the file ***GPS_5.0_windows.exe*** to begin installation. Follow the user prompts through the installation. And snapshots of the setup program are shown below:









Click on the **Finish** button to complete the setup program.

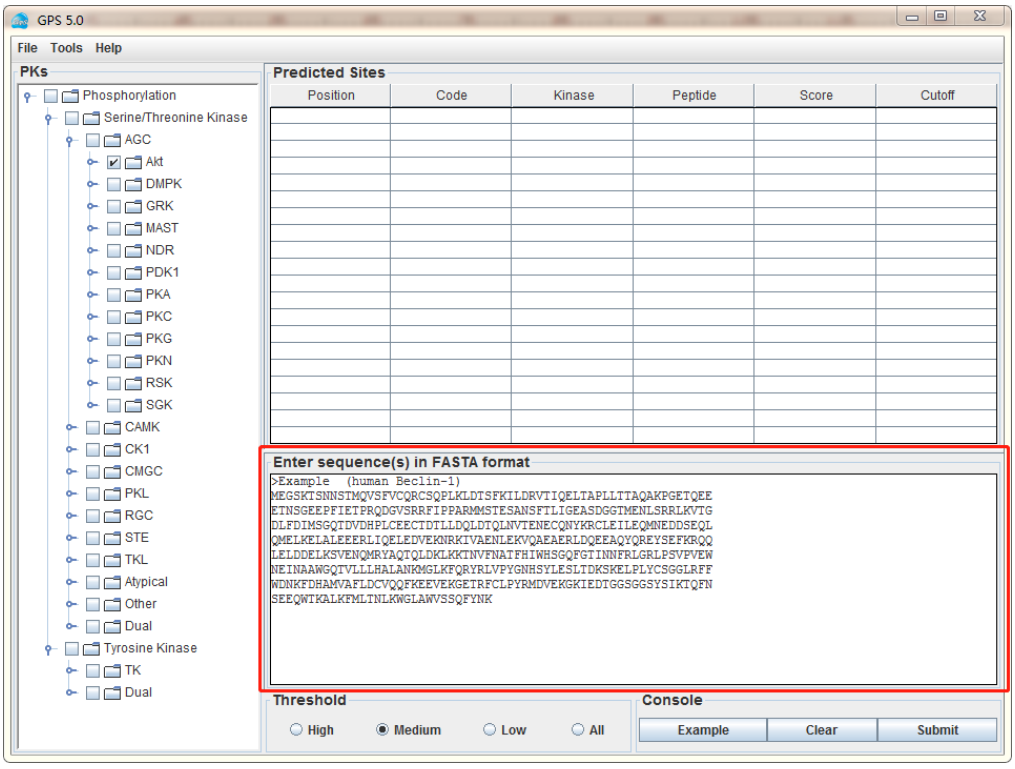
Prediction of Kinase-specific Phosphorylation Sites

Direct Prediction

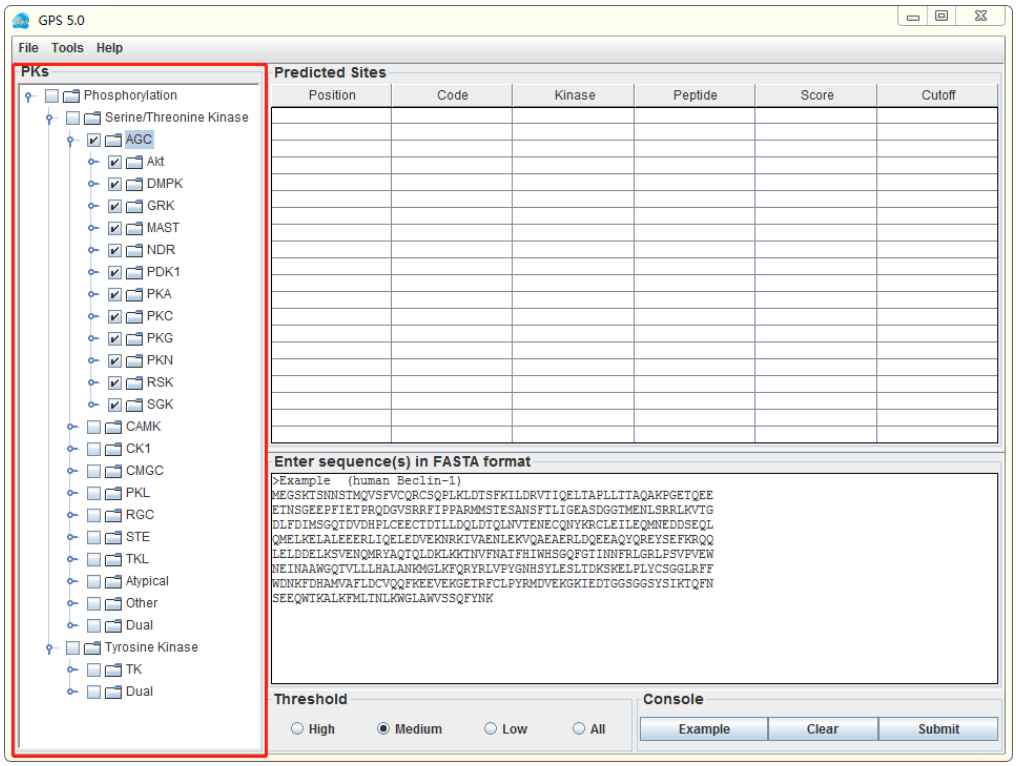
For convenience, the GPS 5.0 allows users to input their protein sequences into the "TEXT form" for prediction. One or more protein sequences should be prepared in **FATSA** format as below:

>protein1
XXXXXXXXXXXXX
XXXXXXXX
>protein2
XXXXXXXXXXXXXXXX…
>protein3
XXXXXXXXXXXX

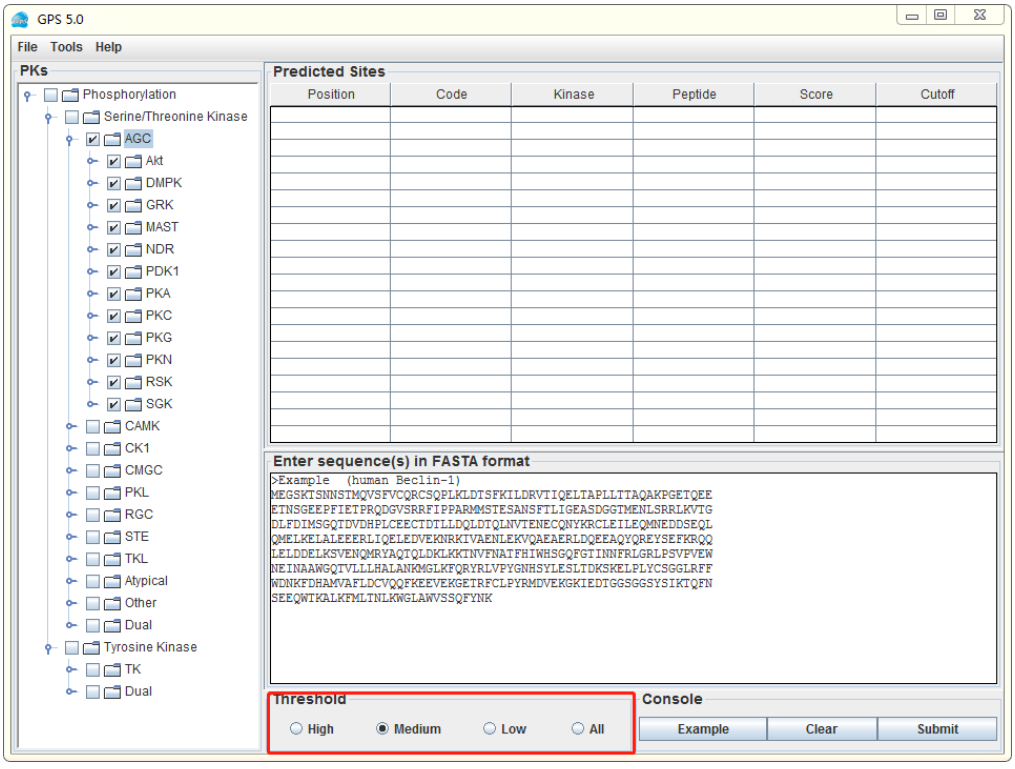
Please note: All irregular words, including non-amino acid word (eg, number) and blank, will be removed automatically. As an instance, we put **human Beclin-1** protein sequence as an example for GPS 5.0. Users could click on the "Example" button to access the example.



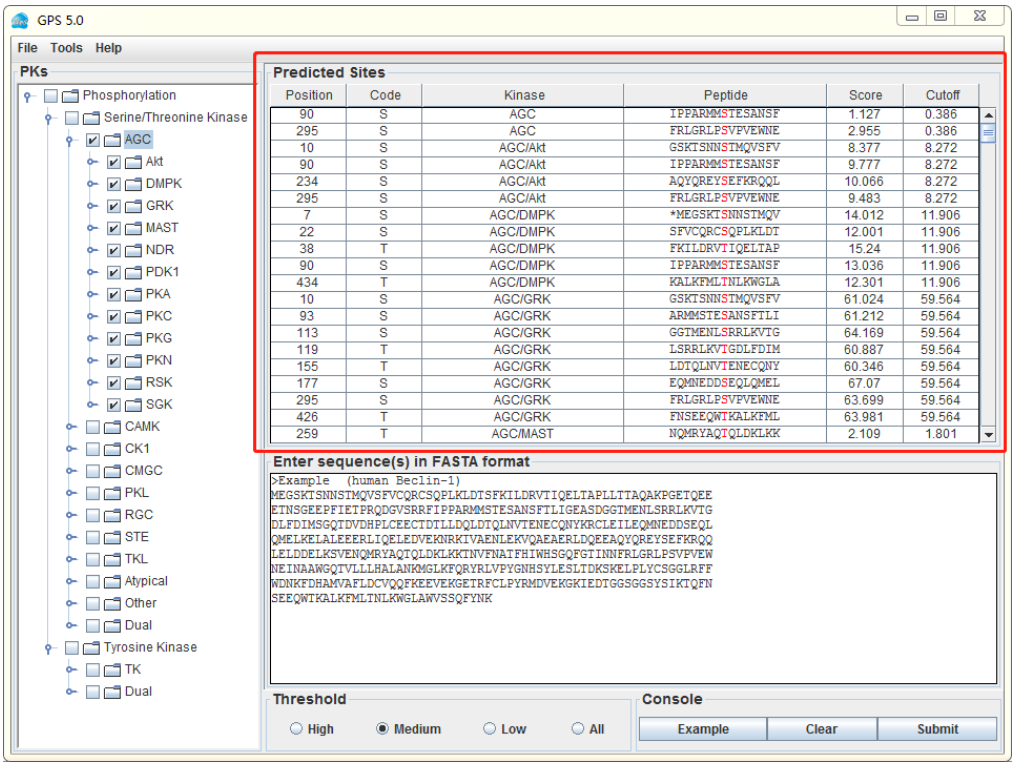
Choose one or more kinases from the **Kinase Hierarchy Tree**



Choose a **Threshold** what you need, default is **Medium**.



Click on the **Submit** button, then the predicted phosphorylation sites will be shown.



Batch Prediction

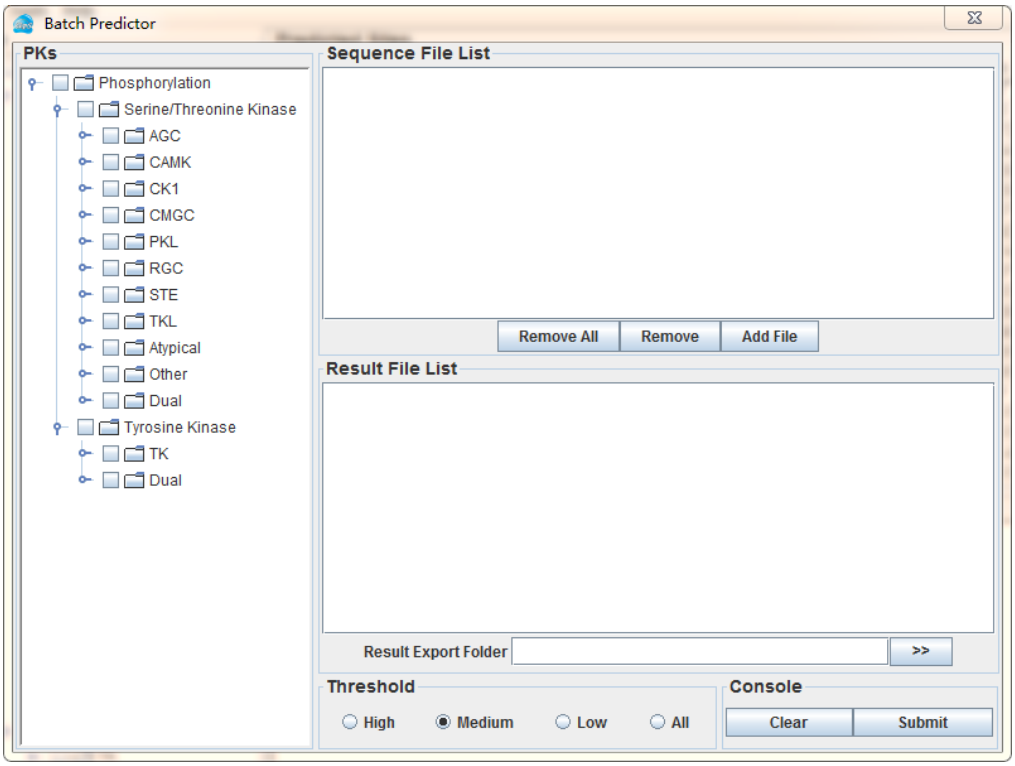
We also provide an alternative approach for processing multiple protein sequences. If the file is large, the **Batch Predictor** will be convenient for users.

The following steps show you how to use it:
Put protein sequences into a file with **FATSA** format as below:

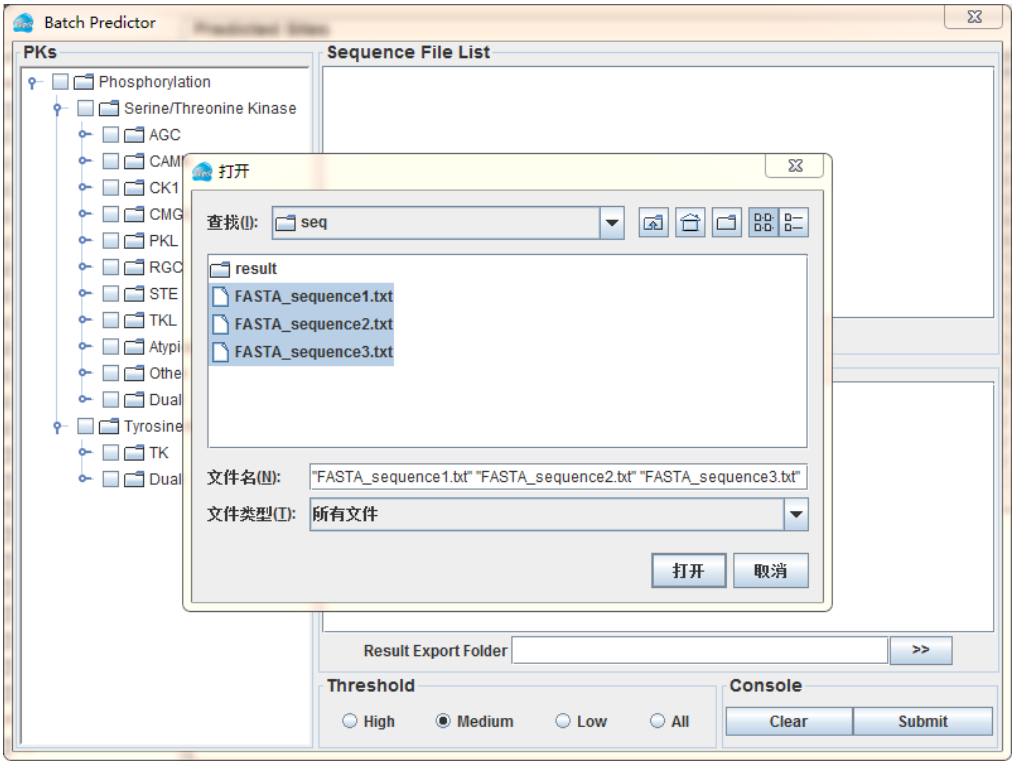
>protein1
XXXXXXXXXXXX
XXXXXXXX
>protein2
XXXXXXXXXXXXXXX…
>protein3
XXXXXXXXXXX

The names of proteins are necessary (the line with “>” and a protein name/accession number).

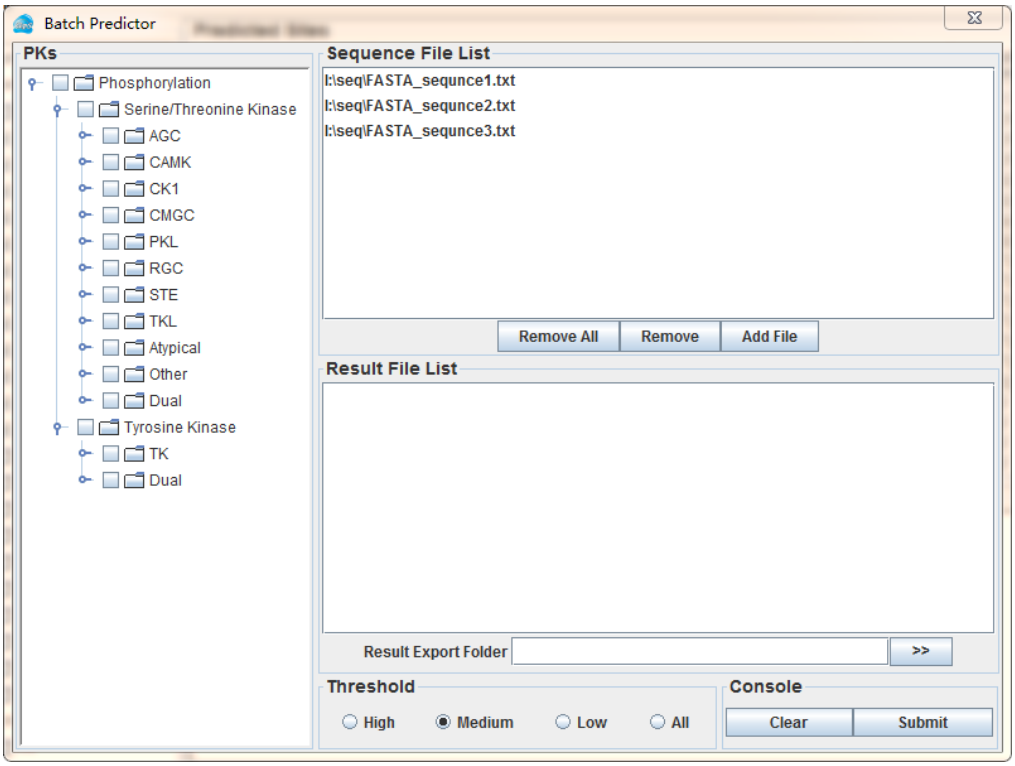
To run the Batch Predictor just select the **Batch Predictor** option in the **Tools** menu.



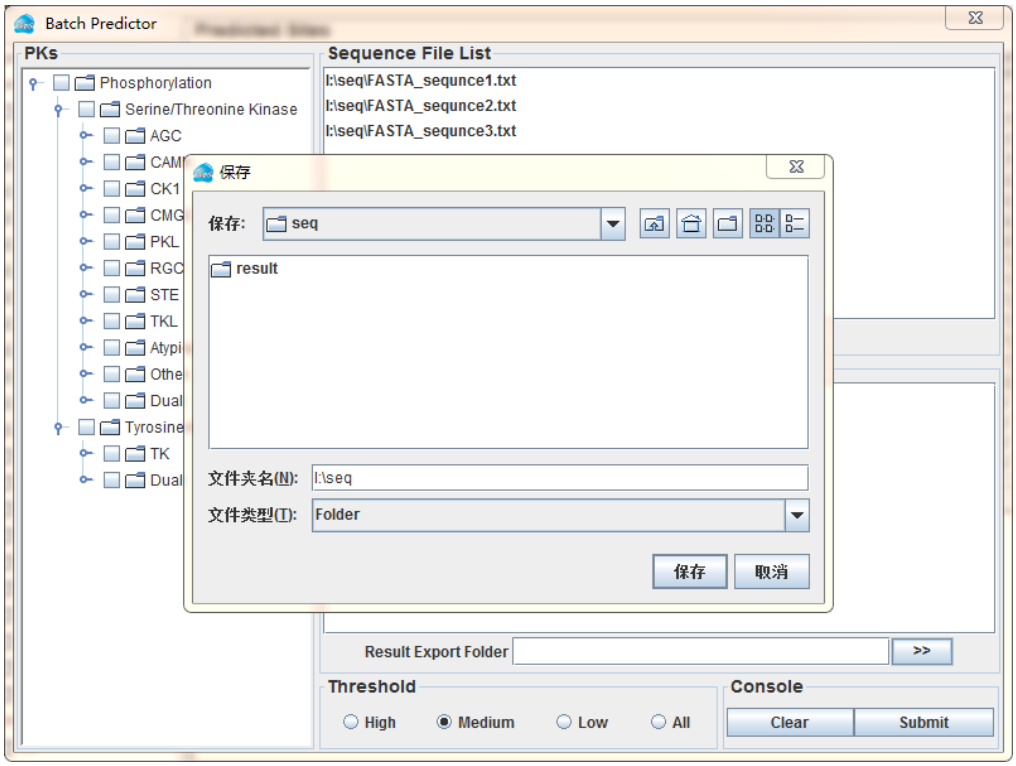
Click on the **Add File** button and add one or more protein sequence files in your hard disk.



The name of added files will be shown in the **Sequence File List**

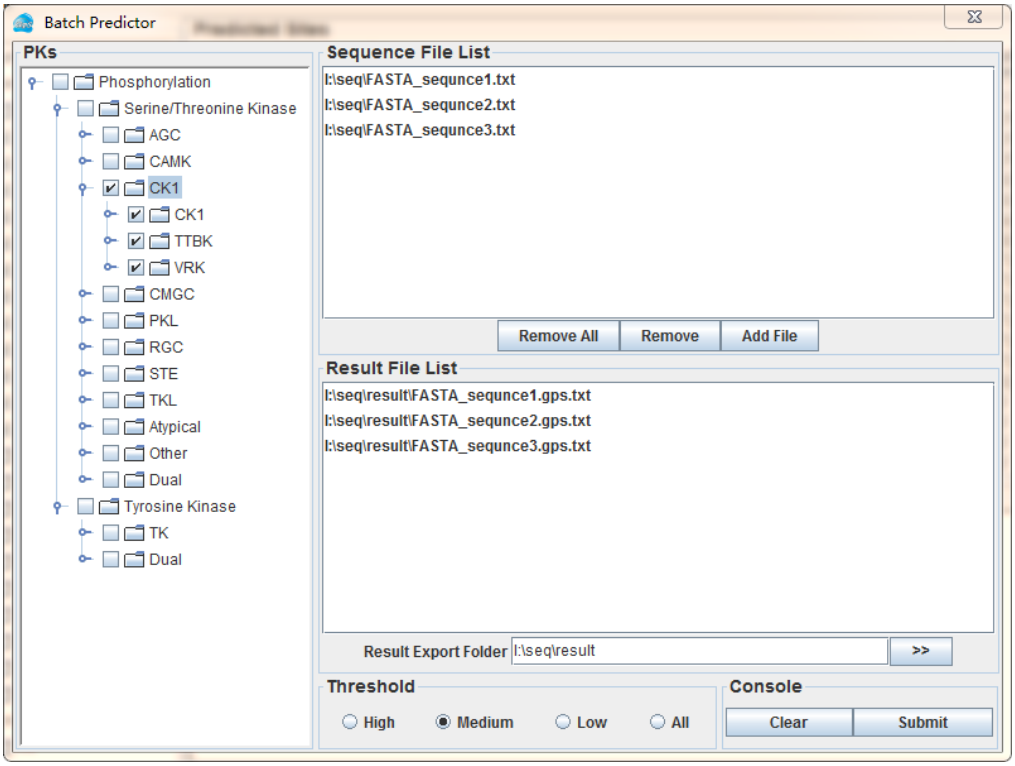


The output directory of prediction results should also be defined. Please click on the "**>>**" button to specify the export file fold.



5. Choose one or more kinases from **Kinase Hierarchy Tree,** and then pick a **Threshold**, click on the **Submit** button, then the **Batch Predictor** will begin to process all of the sequence files that have been added to the list. The results of predictions will be exported to the **Prediction Export Fold,** and the name of result files will be shown in the **Prediction**

File List.

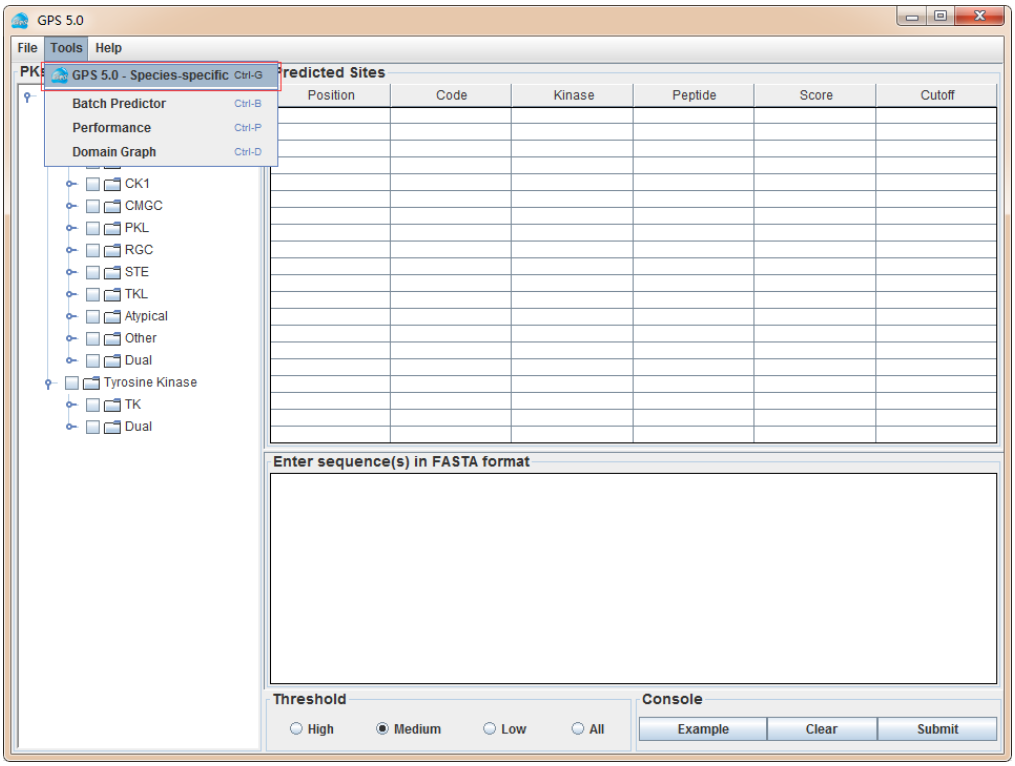

Batch Predictor
PKs
Phosphorylation
Serine/Threonine Kinase
AGC
CAMK
CK1
CK1
TTBK
VRK
CMGC
PKL
RGC
STE
TKL
Atypical
Other
Dual
Tyrosine Kinase
TK
Dual
Sequence File List
I:\seq\FASTA_sequnce1.txt
I:\seq\FASTA_sequnce2.txt
I:\seq\FASTA_sequnce3.txt
Remove All
Remove
Add File
Result File List
I:\seq\result\FASTA_sequnce1.gps.txt
I:\seq\result\FASTA_sequnce2.gps.txt
I:\seq\result\FASTA_sequnce3.gps.txt
Result Export Folder
I:\seq\result
>>
Threshold
High
Medium
Low
All
Console
Clear
Submit

Prediction of Species-specific Kinase-specific Phosphorylation Sites

Direct Prediction

An advance feature of GPS 5.0 is Prediction of Species-specific Kinase-specific Phosphorylation Sites. Users can predict phosphorylation sites in different species. Since there isn't a standard principle classification for all animals and plants, we applied the kinases classification for eukaryotes from a new published database - iEKPD, which is a eukaryotic protein kinases and protein phosphatases database. Kinases and phosphatases in 164 species are classified in a hierarchy structure, including group, family and single protein.

To predict phosphorylation sites in 161 species, users could open the species-specific prediction interface by clicking the item "**GPS 5.0 – Species-specific**" item in "**Tools**" menu.

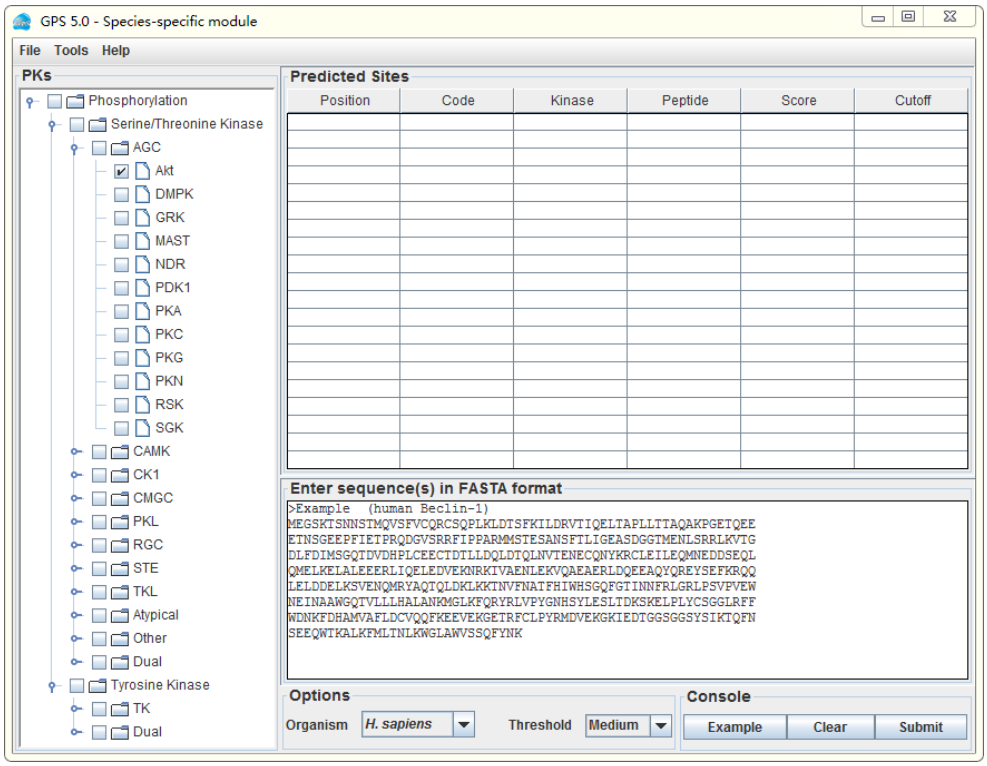


Similar with the classical version, this prediction interface also allows users to input their protein sequences into the "TEXT form" for prediction. One or more protein sequences should be prepared in **FATSA** format as below:

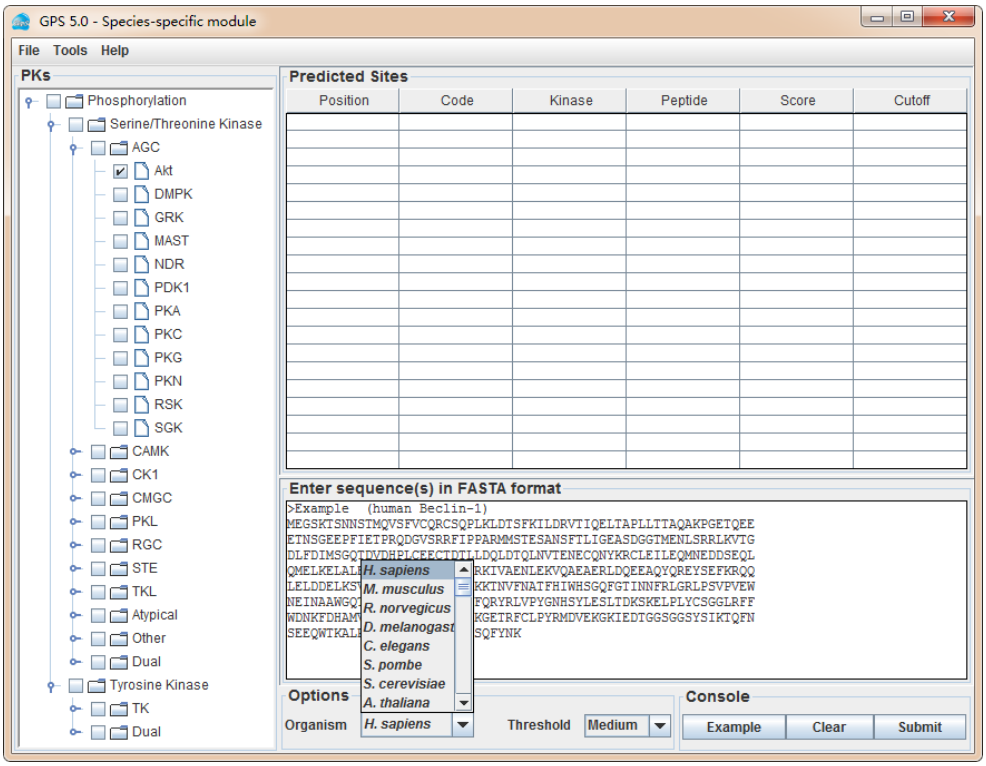
>protein1

XXXXXXXXXXXXX
XXXXXXXX
>protein2
XXXXXXXXXXXXXXXX…
>protein3
XXXXXXXXXXXX

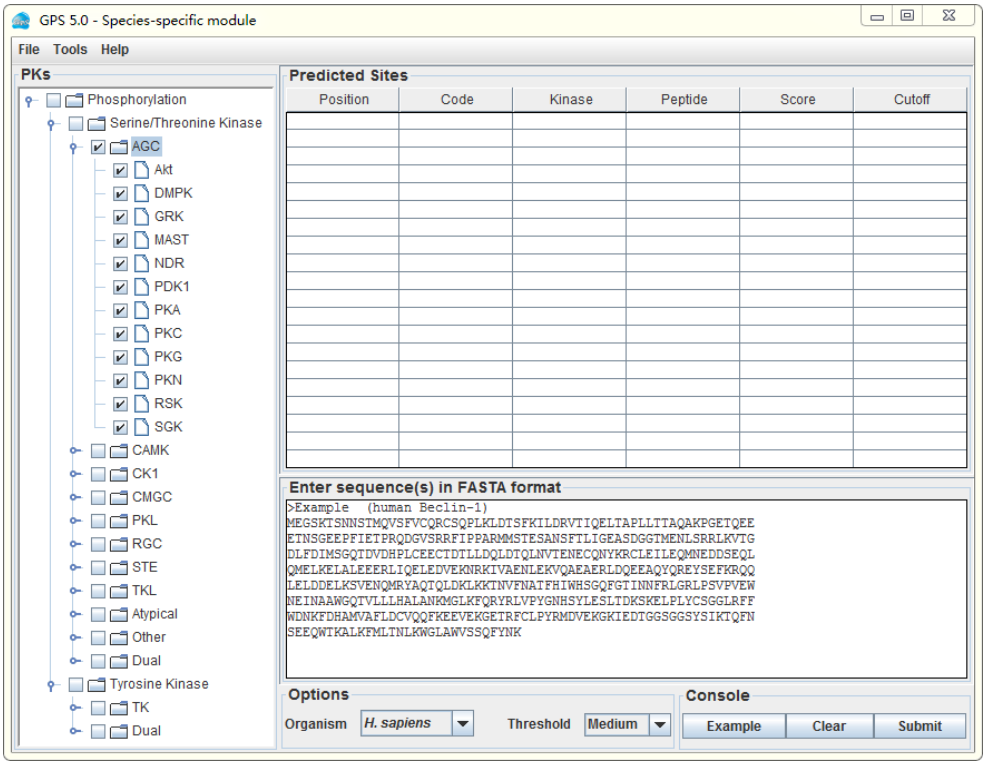
Please note: All irregular words, including non-amino acid word (eg, number) and blank, will be removed automatically. As an instance, we put **human Beclin-1** protein sequence as an example for GPS 5.0. Users could click on the "Example" button to access the example.



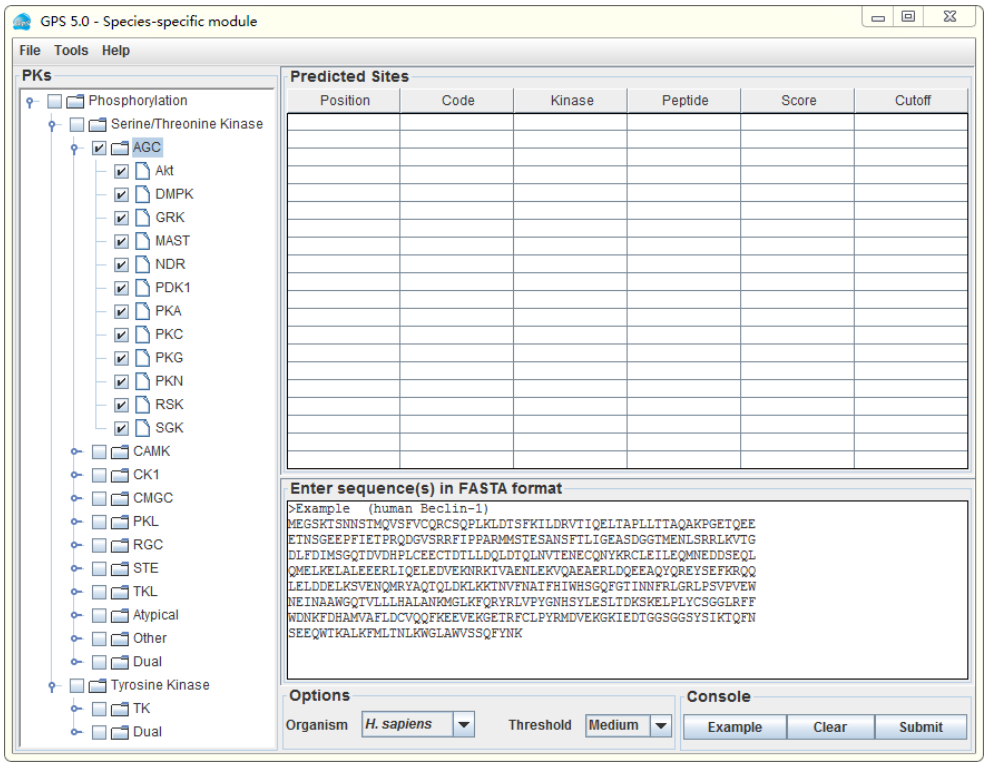
Choose a species from the **Organism List.**



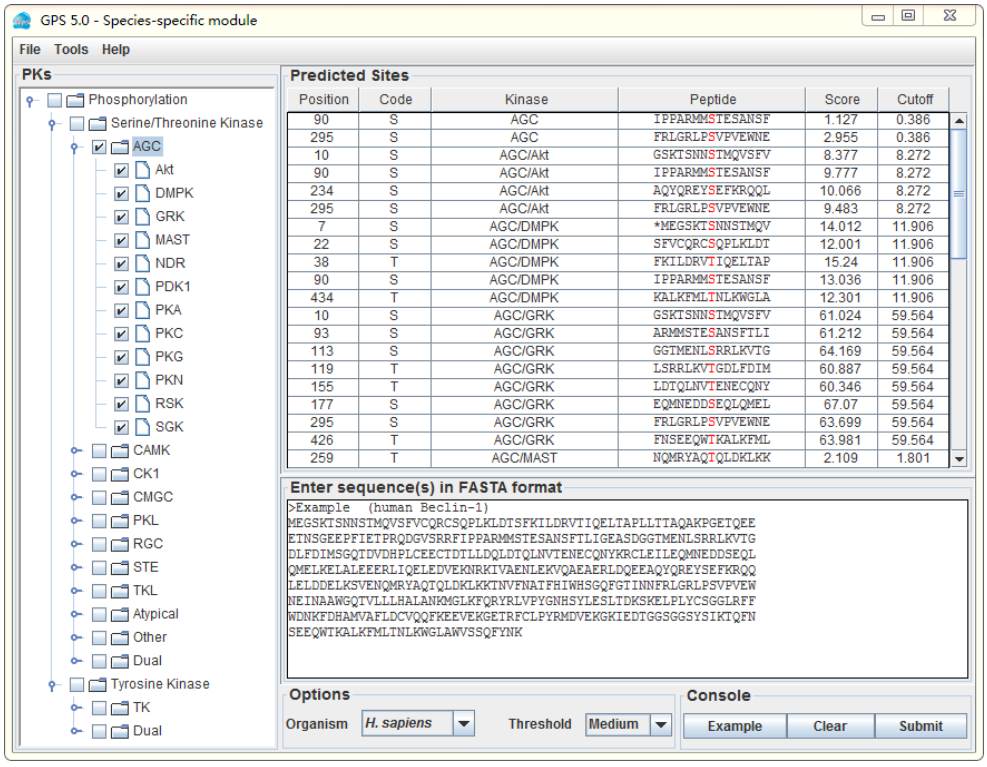
Choose one or more kinases from the **Kinase Hierarchy Tree**



Choose a **Threshold** what you need, default is **Medium**



Click on the **Submit** button, then the predicted phosphorylation sites will be shown.



Batch Prediction

In GPS 5.0 – Species-specific prediction interface, we also provide an alternative approach for processing multiple protein sequences. If the file is large, the **Batch Predictor** will be convenient for users. The usage is similar to the **Batch Prediction** of **Direct Prediction**.

Algorithms and Prediction Performance

Algorithm Design

To predict kinase-specific phosphorylation sites, a series of GPS (Group-based Phosphorylation Scoring) algorithm were maintained and improved in the past few years [2,3]. The basic hypothesis of GPS algorithm is that highly similar short peptides bear similar biochemical properties for the modification. Therefore, 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 the amino acid substitution matrix BLOSUM62 to calculate the similarity between two PSP (m, n) peptides.

In GPS 5.0, a new hypothesis is that long flanking regions around p-sites might be important for the recognition of PKs, which are bulky molecules to interact with phosphorylatable residues. Thus the GPS 5.0 is upgraded in basis of GPS 2.1 and contained two parts: the scoring strategy and performance improvement.

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

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

Where *L* is the length of the PSP(30, 30) peptide and equal to 61 to represent 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 ($i = 1, 2, 3, \ldots, N$). W_j is the weight value of position *j*, and M_{train} represents the optimized amino acid substitution matrix.

The performance improvement comprises two parts, the MLS and MaM in GPS 2.1 were upgraded into PWD and SMO, respectively.

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

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

Initially, the weight value of each position W_j in the PSP(30, 30) peptide was set to 1. Then the one-vs-rest (OVR) classifier with the ridge (L2) penalty of the LR algorithm was used to optimize W_j values, by applying the "newton-cg" solver in the class

LogisticRegressionCV of scikit-learn v0.21.0 (https://scikit-learn.org/). 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 of 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:

$$W_j = W_{-30}, \ldots, W_{-1}, W_0, W_1, \ldots, W_{30}$$

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.

(ii) 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 was defined as S_{ab}:

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

Where C_j is the number of *ab* amino acid pairs at position *j*. In BLOSUM62, there were 24 types of characters including 20 types of amino acids and 4 non-canonical characters (B, Z X and *). Thus, a number of [24*(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}:

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

Evaluation of Prediction Performances

Performance measurements

To evaluate the prediction performances, four standard measurements were used, including accuracy (*Ac*), sensitivity (*Sn*), specificity (*Sp*) and Mathew correlation coefficient (*MCC*). Accuracy (*Ac*) represents the correct ratio between both positive (+) and negative (-) data sets, while sensitivity (*Sn*) and specificity (*Sp*) illustrate the correct prediction ratios of positive (+) and negative data (-) sets respectively. Since the number of positive data and negative data differed too much from each other, the Mathew correlation coefficient (*MCC*) was also included. The value of *MCC* ranges from -1 to 1, and a larger *MCC* value stands for better prediction performance.

Among the data with positive hits by GPS 5.0, the real positives were defined as *true positives* (*TP*), while the others were defined as *false positives* (*FP*). Among the data with negative predictions, the real positives were defined as *false negatives* (*FN*), while the others were defined as *true negatives* (*TN*). The four measurements of sensitivity (*Sn*), specificity (*Sp*), accuracy (*Ac*), and Mathew correlation coefficient (*MCC*) were defined as below:

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

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

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

The self-consistency, leave-one-out validation and n-fold cross-validation

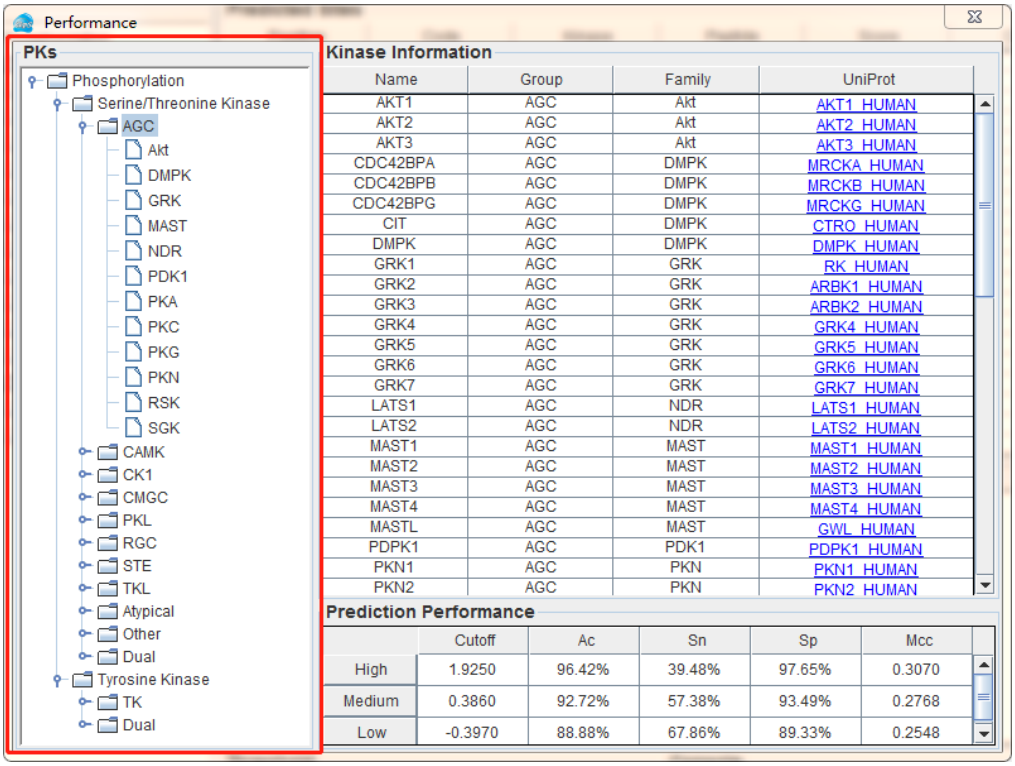
The self-consistency used the training positive data and negative data directly to evaluate the prediction performance, and represented the computational power of the prediction system. However, the robustness and stability of the software should also be evaluated by leave-one-out validation and *n*-fold cross-validation. In the leave-one-out validation, which is also called as Jack-Knife validation, each sites in the dataset 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 *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 remanent part was taken as a testing data set. This process was repeated twenty times and the average performance of *n*-fold cross- validation was used to estimate the performance. In this work, the performances of validation were calculated for all PK groups. And the 10-fold cross-validations were performed for PK groups with the number of positive data ≥ 30 and leave-one-out validation were calculated

for PK groups with the number of positive data ＜ 30. All validation performances of GPS 5.0 have been included in **Performance** option in the **Tools** menu.

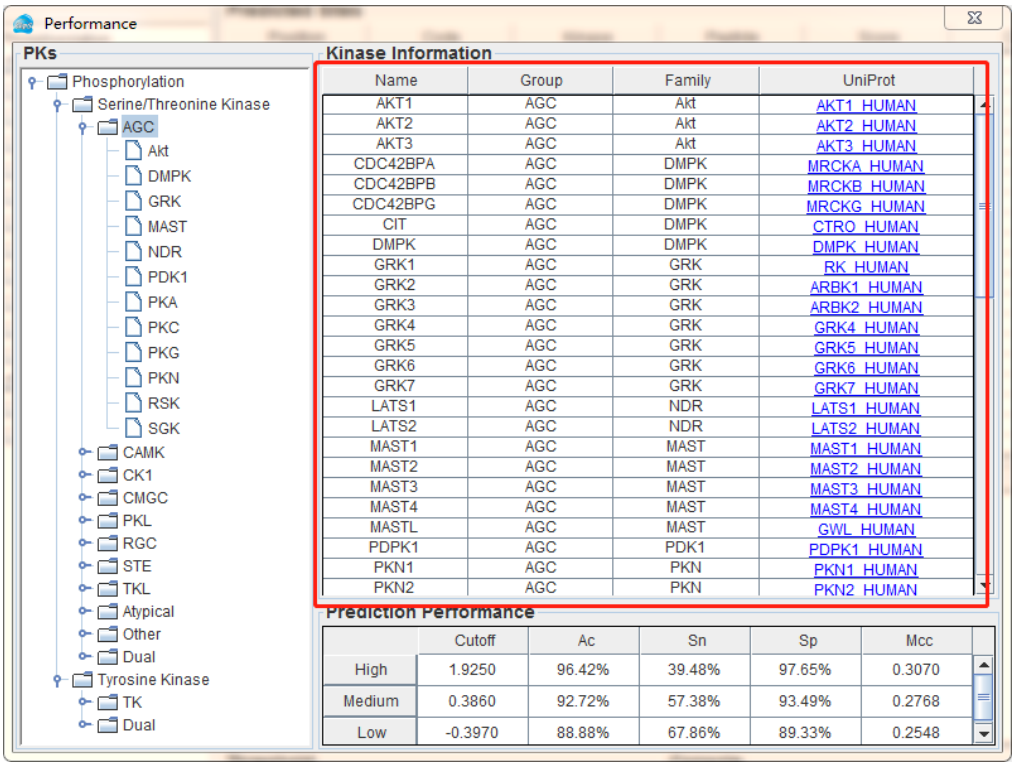
Accessing of prediction performances of GPS 5.0 with related information

To check the prediction performance of the **479** human PKs just select the **Performance** option in the **Tools** menu.

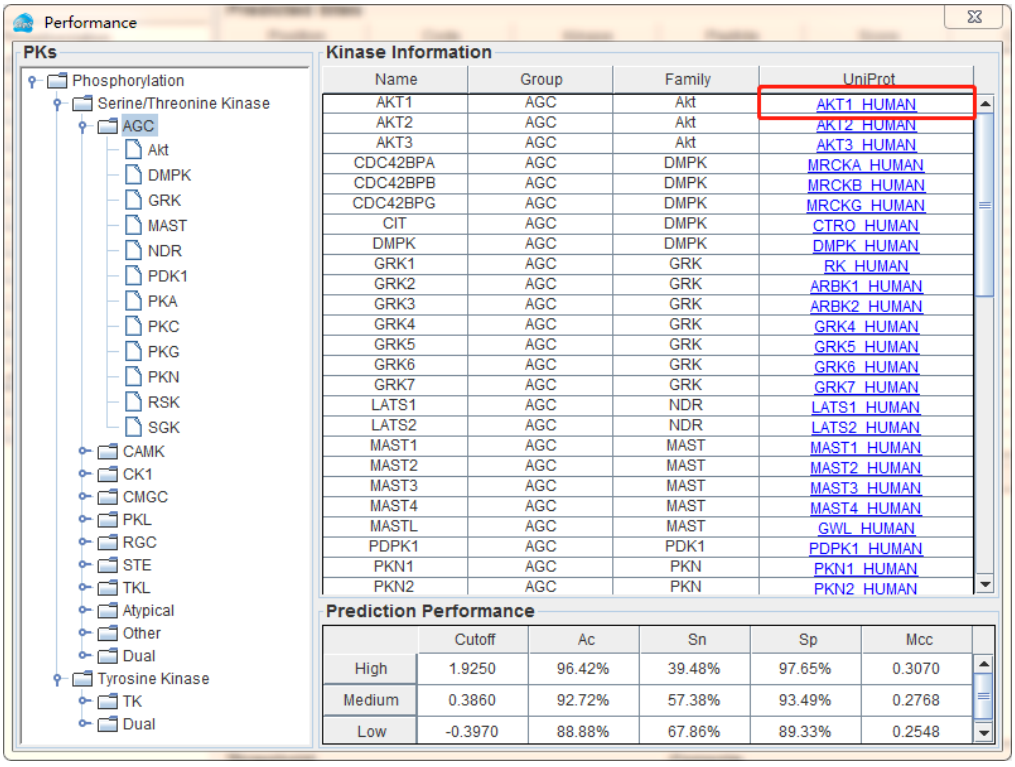
Choose a kinase what you want check from the **Kinase Hierarchy Tree**.



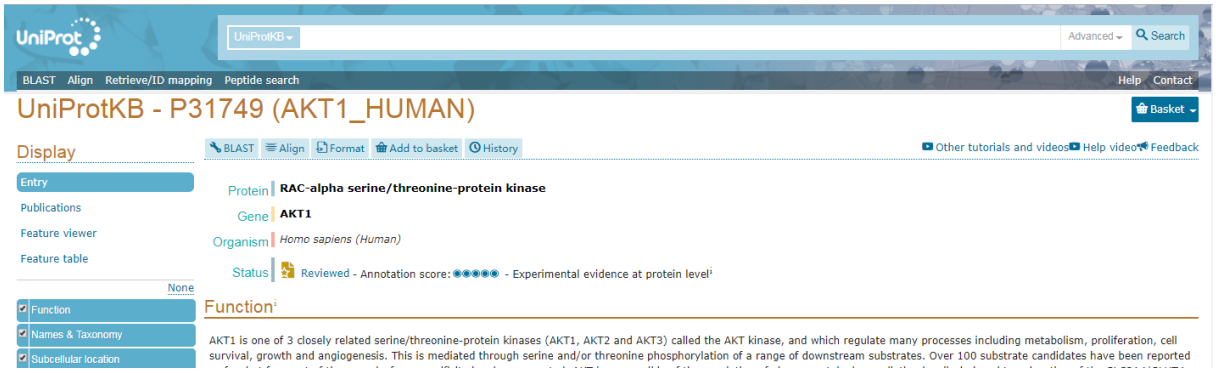
Then **Kinase Information** and **Prediction Performance** of the kinase are shown in the tables.



If you want get more kinase information, you can click on the hyperlinks in the table.



The hyperlinks will access the **UniProt** database and show you the detailed information of the kinases.


UniProt
UniProtKB
Advanced
Search
BLAST Align Retrieve/ID mapping Peptide search
Help Contact
UniProtKB - P31749 (AKT1_HUMAN)
Basket
Display
BLAST Align Format Add to basket History
Other tutorials and videos Help video Feedback
Entry
Publications
Feature viewer
Feature table
None
Function
Names & Taxonomy
Subcellular location
Protein RAC-alpha serine/threonine-protein kinase
Gene AKT1
Organism Homo sapiens (Human)
Status Reviewed - Annotation score: - Experimental evidence at protein level
Function
AKT1 is one of 3 closely related serine/threonine-protein kinases (AKT1, AKT2 and AKT3) called the AKT kinase, and which regulate many processes including metabolism, proliferation, cell survival, growth and angiogenesis. This is mediated through serine and/or threonine phosphorylation of a range of downstream substrates. Over 100 substrate candidates have been reported

References

1. G. Manning, D. B. Whyte, R. Martinez et al., *Science* **298** (5600), 1912 (2002).
2. Y. Xue, F. Zhou, M. Zhu et al., *Nucleic Acids Res* **33** (Web Server issue), W184 (2005).
3. F. F. Zhou, Y. Xue, G. L. Chen et al., *Biochem Biophys Res Commun* **325** (4), 1443 (2004).
4. Y. Xue, A. Li, L. Wang et al., *BMC Bioinformatics* **7**, 163 (2006).
5. N. Blom, T. Sicheritz-Ponten, R. Gupta et al., *Proteomics* **4** (6), 1633 (2004).
6. J. C. Obenauer, L. C. Cantley, and M. B. Yaffe, *Nucleic Acids Res* **31** (13), 3635 (2003).
7. H. D. Huang, T. Y. Lee, S. W. Tzeng et al., *Nucleic Acids Res* **33** (Web Server issue), W226 (2005).
8. J. H. Kim, J. Lee, B. Oh et al., *Bioinformatics* **20** (17), 3179 (2004).
9. R. I. Brinkworth, R. A. Breinl, and B. Kobe, *Proc Natl Acad Sci U S A* **100** (1), 74 (2003).
10. T. Li, F. Li, and X. Zhang, *Proteins* (2007).
11. G. Neuberger, G. Schneider, and F. Eisenhaber, *Biology direct* **2**, 1 (2007).
12. Y. H. Wong, T. Y. Lee, H. K. Liang et al., *Nucleic Acids Res* **35** (Web Server issue), W588 (2007).
13. F. Diella, S. Cameron, C. Gemund et al., *BMC Bioinformatics* **5** (1), 79 (2004).
14. P. V. Hornbeck, I. Chabra, J. M. Kornhauser et al., *Proteomics* **4** (6), 1551 (2004).
15. L. M. Iakoucheva, P. Radivojac, C. J. Brown et al., *Nucleic Acids Res* **32** (3), 1037 (2004).
16. N. Blom, S. Gammeltoft, and S. Brunak, *J Mol Biol* **294** (5), 1351 (1999).
17. C. R. Ingrell, M. L. Miller, O. N. Jensen et al., *Bioinformatics* **23** (7), 895 (2007).
18. Y. R. Tang, Y. Z. Chen, C. A. Canchaya et al., *Protein Eng Des Sel* **20** (8), 405 (2007).
19. R. Linding, L. J. Jensen, G. J. Ostheimer et al., *Cell* **129** (7), 1415 (2007).

Release Note

1. Jan. 1st, 2008, the online service and the local stand-alone packages of GPS 2.0 were released. The stand-alone software of GPS 2.0 could support Windows Operating Systems.
2. Jan. 29th, 2008, a bug was found that the version 2.0 couldn't be used under non-English Operating Systems. We fixed the bug and released the version 2.0.1 beta version. We thank Dr. Miguel Angel Sanchez (Malaga, Spain) and Dr. Gilles Vachon (Universite J. Fourier, France) to send us feedbacks.
3. Apr. 13th, 2008, The GPS 2.0.1 was released, with online service and local packages, to support three major Operating Systems, including Windows, Linux/Unix and Mac. Also, the GPS 2.0.1 manual was updated and included in the packages.
4. Mar. 1st, 2009, The GPS 2.1 was released, with online service and local packages and support three major Operating Systems including Windows, Linux/Unix and Mac. In this version, the newly developed motif length selection (MLS) method was introduced and the robustness of the prediction system was greatly improved.
5. Jul. 21st, 2009, The GPS 2.1.1 was released, with online service and local packages and support three major Operating Systems including Windows, Linux/Unix and Mac.
6. Sep. 13th, 2012, The GPS 2.1.2 was released, with online service and local packages and support three major Operating Systems including Windows, Linux/Unix and Mac.
7. Dec. 14th, 2014, The GPS 3.0 was released, with online service and local packages and support three major Operating Systems including Windows, Linux/Unix and Mac. In this version, *k-means* clustering, peptide selection (PS), and weight training (WT) procedures were added to enhance the prediction performance.
8. Jul. 20th, 2019, The GPS 5.0 was released, with online service and local packages and support three major Operating Systems including Windows, Linux/Unix and Mac. In this version, two novel methods of position weight determination (PWD) and scoring matrix optimization (SMO) were developed to improve the performance for kinase-specific phosphorylation sites prediction. Meanwhile, the GPS 5.0 manual was updated and included in the packages.