iGPS Manual

GPS algorithm with the interaction filter

Version 1.0.1
27/08/2012

Authors: Yu Xue & Jian Ren

Contact: Dr. Yu Xue, xueyu@mail.hust.edu.cn;
Dr. Jian Ren, renjian.sysu@gmail.com
The software is only free for academic research.
The latest version of iGPS software is available from http://igps.biocuckoo.org

Index

Index....3
Statement....1
Introduction....2
Download & Installation....4
The Usage of iGPS....6
Direct Prediction....6
Batch Prediction....11
Visualization of Protein Phosphorylation Network with PNC 1.0....14
Search pre-calculated ssKSRs in EPNdb....17
References....21
Release Note....23

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**. The work of CUCKOO Workgroup is supported by grants from Chinese 973 project (2010CB945400, 2012CB911200, 2012CB910101), and Chinese Natural Science Foundation (90919001, 30830036, 30900835, 31071154, 91019020, 31171263).

Introduction

Protein kinase (PK)-catalyzed phosphorylation is one of the most important and ubiquitous post-translational modifications (PTMs) of proteins. This process temporally and spatially modifies approximately 30% of all cellular proteins and plays a crucial role in regulating a variety of biological processes such as signal transduction and the cell cycle [1-6]. The human genome encodes 518 PK genes (approximately 2% of the genome), with different PKs exhibiting distinct recognition specificities; each PK modifies only a limited subset of substrates, thereby guaranteeing the fidelity of cell signaling [1-6]. It is accepted that short linear motifs (SLMs) around the phosphorylation sites (p-sites) provide primary specificity [2,6-9], and a variety of additional contextual factors, including co-localization, co-expression, co-complex, and physical interaction of the PKs with their targets, contribute additional specificity *in vivo* [10-15]. Aberrances of PKs or key substrates disrupt normal function, rewire signaling pathways, and are implicated in various diseases and cancers [4,16-19]. In this regard, the identification of kinase-specific p-sites and the systematic elucidation of site-specific kinase-substrate relations (ssKSRs) would provide a fundamental basis for understanding cell plasticity and dynamics and for dissecting the molecular mechanisms of various diseases, while the ultimate progress could suggest potential drug targets for future biomedical design [11,12,15].

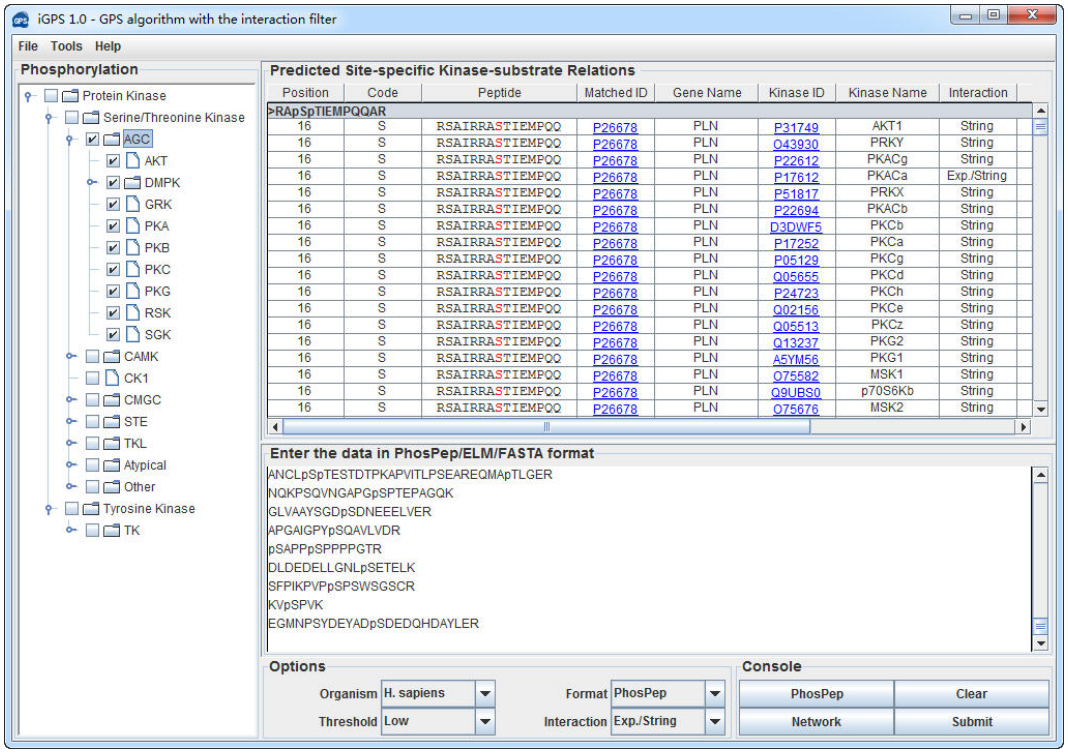
Conventional experimental identification of ssKSRs, performed in a one-by-one manner, is labor-intensive, time-consuming and expensive. There are only 3,508 known kinase-specific p-sites in the 1,390 proteins collected in the Phospho.ELM 8.2 database (released in April 2009) [20]. In 2005, Ptacek *et al*. detected more than 4,000 *in vitro* kinase-substrate relations (KSRs) in *Saccharomyces cerevisiae* using protein chip technology, although the exact phosphorylation sites were not determined [3]. Recently, rapid advances in phosphoproteomics have provided an opportunity to systematically assess phosphorylation [1,21,22]. State-of-the-art high-throughput mass spectrometry (HTP-MS) techniques have the ability to detect thousands of p-sites in cells or tissues in a single experiment [1,21,23,24]. We have collected 145,646 eukaryotic p-sites, primarily from these large-scale assays (Supplemental Table S1); the regulatory PKs for 97.6% of these sites remain to be characterized.

Alternatively, the *in silico* prediction of ssKSRs can generate useful information for subsequent experimental manipulation. In 2001, Yaffe *et al*. developed the SLM-based software Scansite for the prediction of ssKSRs directly from protein primary sequences [10]. Later, the strategy was employed in a variety of kinase-specific predictors [25], including our group-based prediction system (GPS) program [26]. These tools may guarantee partially correct predictions for *in vitro* phosphorylation, but they are far from being

adequate for *in vivo* hits because the contributions of various contextual factors cannot be neglected. To address this problem, Linding *et al*. developed a predictor of NetworKIN by combining an SLM-based approach with network contextual information to predict *in vivo* ssKSRs, and a potential *in vivo* human phosphorylation network (HPN) was modeled by annotating the phosphoproteomic data [11,12].

In this work, we developed a software package of iGPS (GPS algorithm with the interaction filter, or *in vivo* GPS) mainly for the prediction of *in vivo* ssKSRs. Eukaryotic PKs were classified into a hierarchy with four levels: group, family, subfamily, and single PK [4]. Based on the hypothesis that similar PKs recognize similar SLMs, we selected a predictor in GPS 2.0 [26] for each PK and directly predicted the potential PKs for the non-annotated p-sites from the phosphoproteomic studies. Consequently, protein-protein interaction (PPI) information was used as the major contextual factor to filtrate potentially false-positive hits. The performance of iGPS was shown by critical evaluations and comparisons to be promising for the accurate prediction of *in vivo* ssKSRs. Based on the prediction results of iGPS, we modeled eukaryotic protein phosphorylation networks (PPNs) at different levels, including whole proteome, pathway and tissues/organs. By additionally computational analyses, we obtained a substantial number of potentially new observations, which can be subjected to further experimental manipulation. This study provides useful information for the understanding of the functional organization and diversity of eukaryotic phosphoproteomes at a systemic level and can be a model for analyzing other PTM-regulating proteomes.

The iGPS 1.0 is freely available at: http://igps.biocuckoo.org.

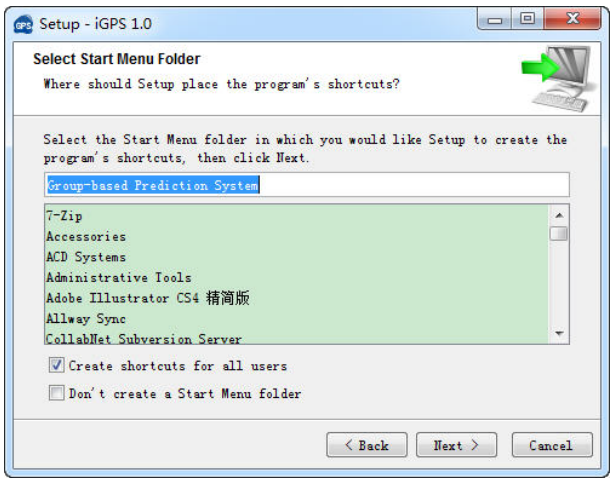


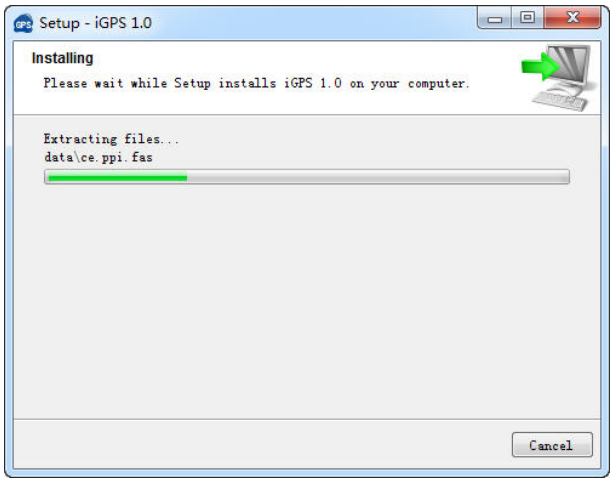
iGPS v1.0 User Interface

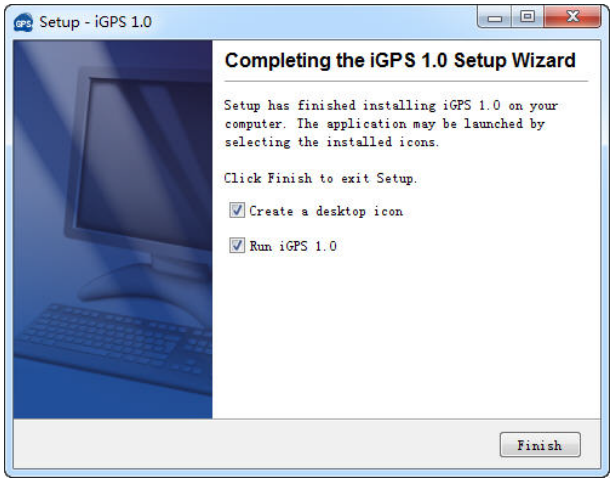
Download & Installation

The software of iGPS 1.0 has been implemented in JAVA, and can be installed on Windows, Mac OS X or Linux systems. The iGPS 1.0 is freely available for academic research at http://igps.biocuckoo.org/down.php. We recommend that users can download the latest release.

After downloading, please double-click on the install package 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.

The Usage of iGPS

Direct Prediction

The main propose of iGPS 1.0 was designed for the prediction of ssKSRs from the phosphoproteomic data. For convenience, the iGPS 1.0 allows users to input their data into the “TEXT form” for prediction. Three formats such as **PhosPep**, **ELM** or **FATSA** are adopted as below:

(1) PhosPep format:
LVEDKPGpSR
GEpSENAGTNQETR
pSRpSNpSKSKPNLPSESR
SKPNLPpSESR
pTSEETISTVQEK
pSLQPLAPR
KDpSLLKPGLR
pSGGQLPSLQEETTR
…
Please note: The p must be annotated in front of the phosphorylation site. Only the annotated phosphorylation sites will be predicted. We **strongly recommend** users to input the data in this format, because it can be easily prepared from the phosphoproteomic experiments.

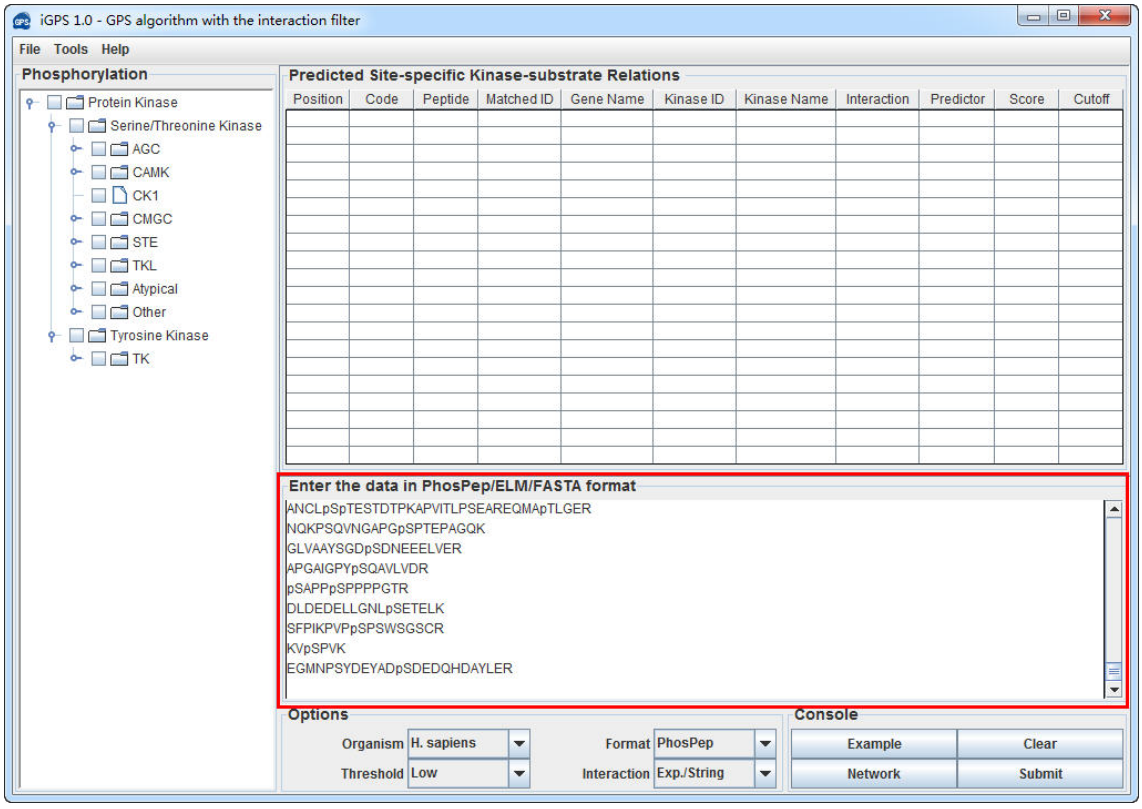
(2) ELM format:
acc sequence position code pmids kinases source entry_date
Q99640
MLERPPALAMPMPTEGTPPPLSGTPIPVPAYFRHAEPGFSLKRPRGLSRSLPPPPPAKGSIPISRLFPPR
TPGWHQLQPRRVSFRGEASETLQSPGYDPSRPESFFQQSFQRLSRLGHGSYGEVFKVRSKEDGRLYAVKR
SMSPFRGPKDRARKLAEVGSHEKVGQHPCCVRLEQAWEEGGILYLQTELCGPSLQQHCEAWGASLPEAQ
VWGYLRDTLLALAHLHSQGLVHLDVKPANIFLGPRGRCKLGDFGLLVELGTAGAGEVQEGDPRYMAPELLQ
GSYGTAADVFSLGLTILEVACNMELPHGGEGWQQLRQGYLPPEFTAGLSSELRSVLVMMLEPDPKLRATAEA
LLALPVLRQPRAWGVLWCMAAEALSRGWALWQALLALLCWLWHGLAHPASWLQPLGPPATPPGSPPCSLL
LDSSLSSNWDDDSLGPSLSPEAVLARTVGSTSTPRSRCTPRDALDLSDINSEPPRGSFPSFEPRNLLSLFED
TLDPT 426 S 12738781 PLK1 LTP 2004-12-31 00:00:00+01
…
Please note: This format was defined in the Phospho.ELM database [20]. All data should be delimited with **Tab**. Only phosphorylation sites with defined positions will be predicted.

(3) FASTA format:
>protein1
XXXXXXXXXXXXX
XXXXXXXX
>protein2
XXXXXXXXXXXXXXX…
>protein3
XXXXXXXXXXXX
...

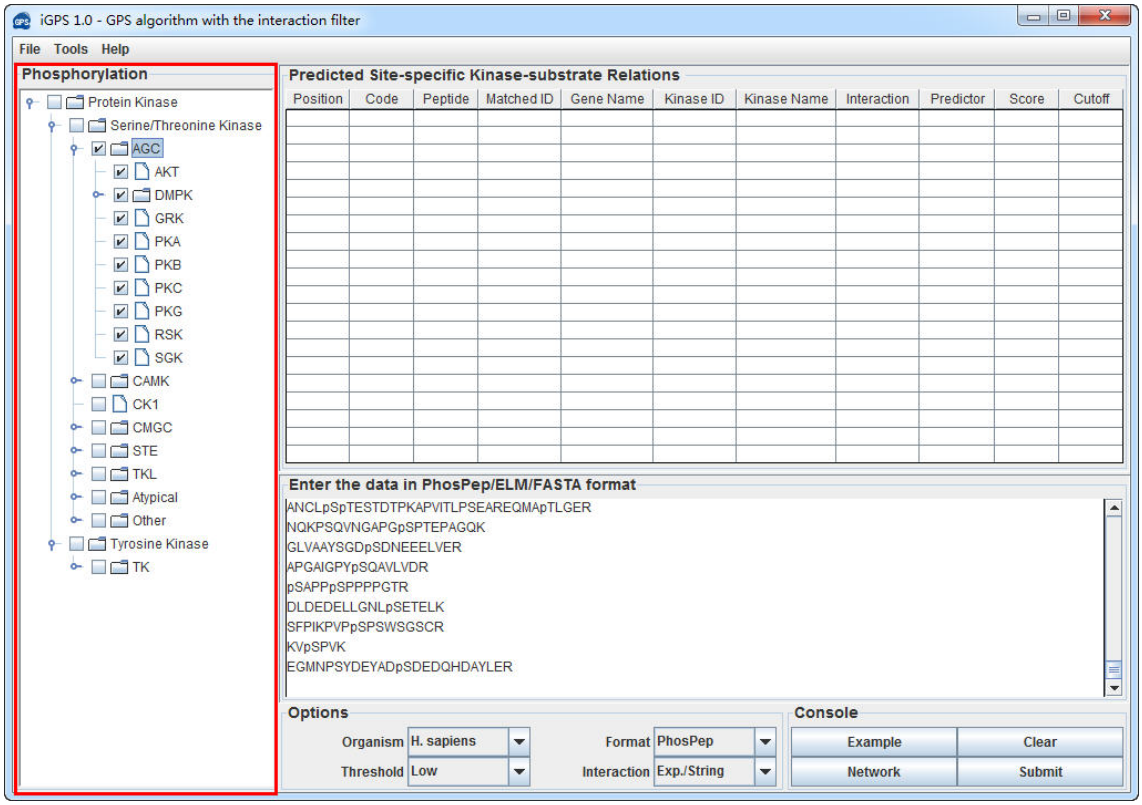
Please note: All irregular words, including non-amino acid word (e.g., number) and blank, will be removed automatically. All of the potential phosphorylation sites in the sequences will be predicted. Here, we strongly recommend users **NOT** to input the data in this format, unless you can make sure the positions of real phosphorylation sites in the proteins. *Ab initio* prediction of ssKSRs merely from protein sequences with iGPS 1.0 will generate too much false positive hits.

Please choose a proper organism based on your input data. The iGPS can predict ssKSRs for five eukaryotic organisms, including *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus*, and *Homo sapiens*.

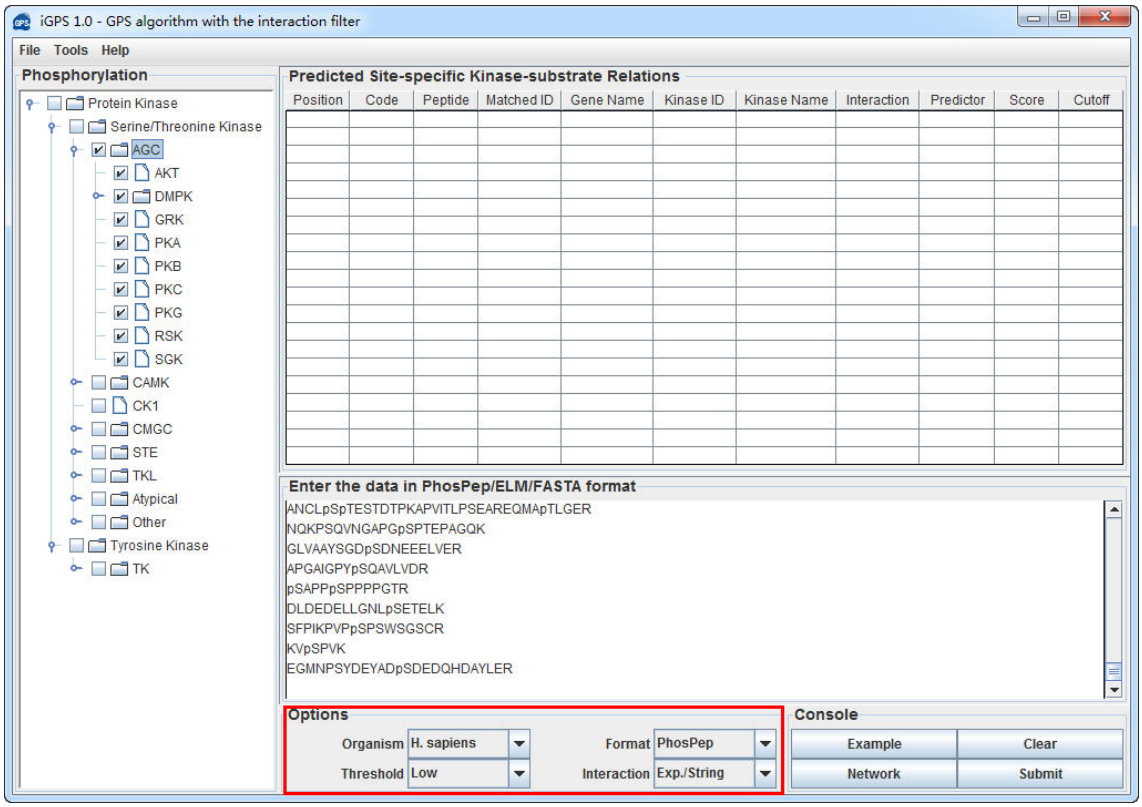
One example was selected for each format. Users can click on the “Example” button to access the examples.



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



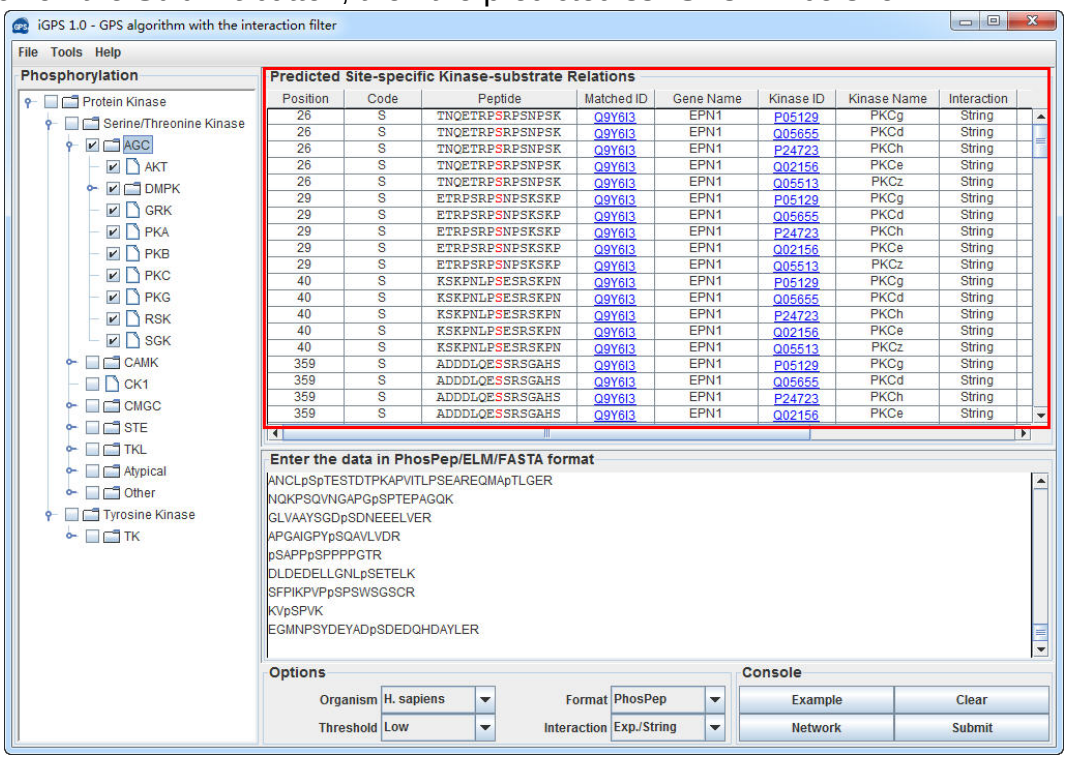
Choose a **Threshold** that you need, while the default threshold is **Low**.



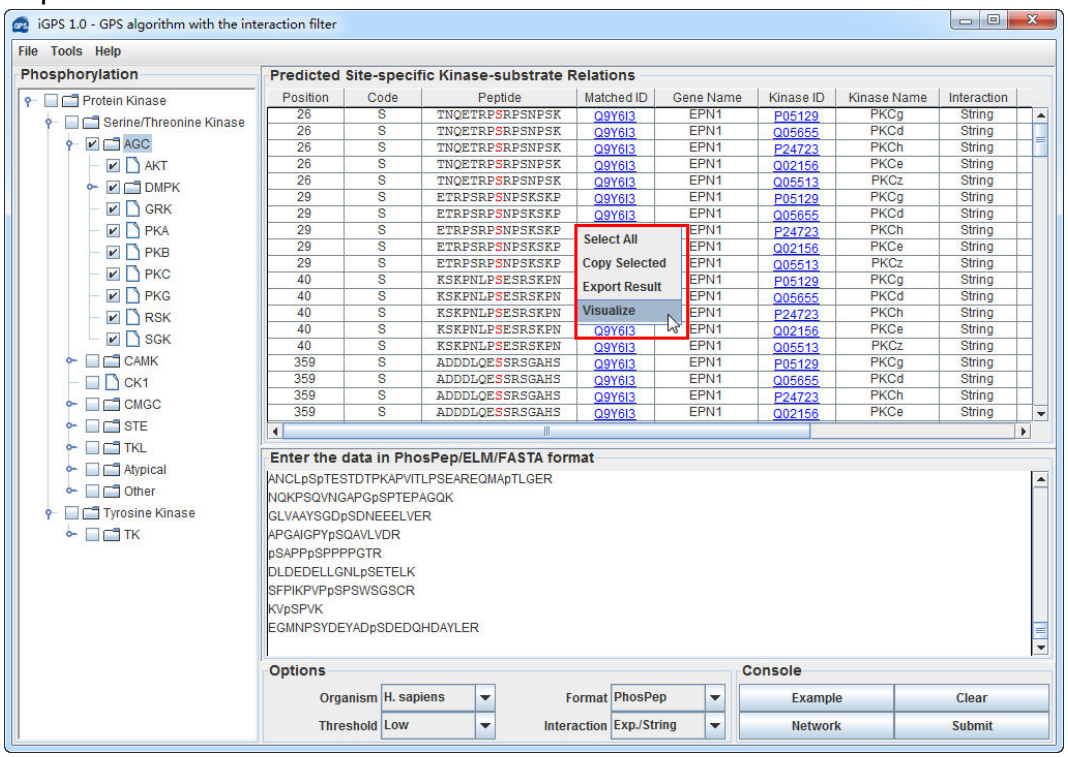
Also, the PPI information was adopted as a filter to reduce false positive hits. Both

experimentally identified and predicted PPI information was used, while the default parameter is all PPIs.

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



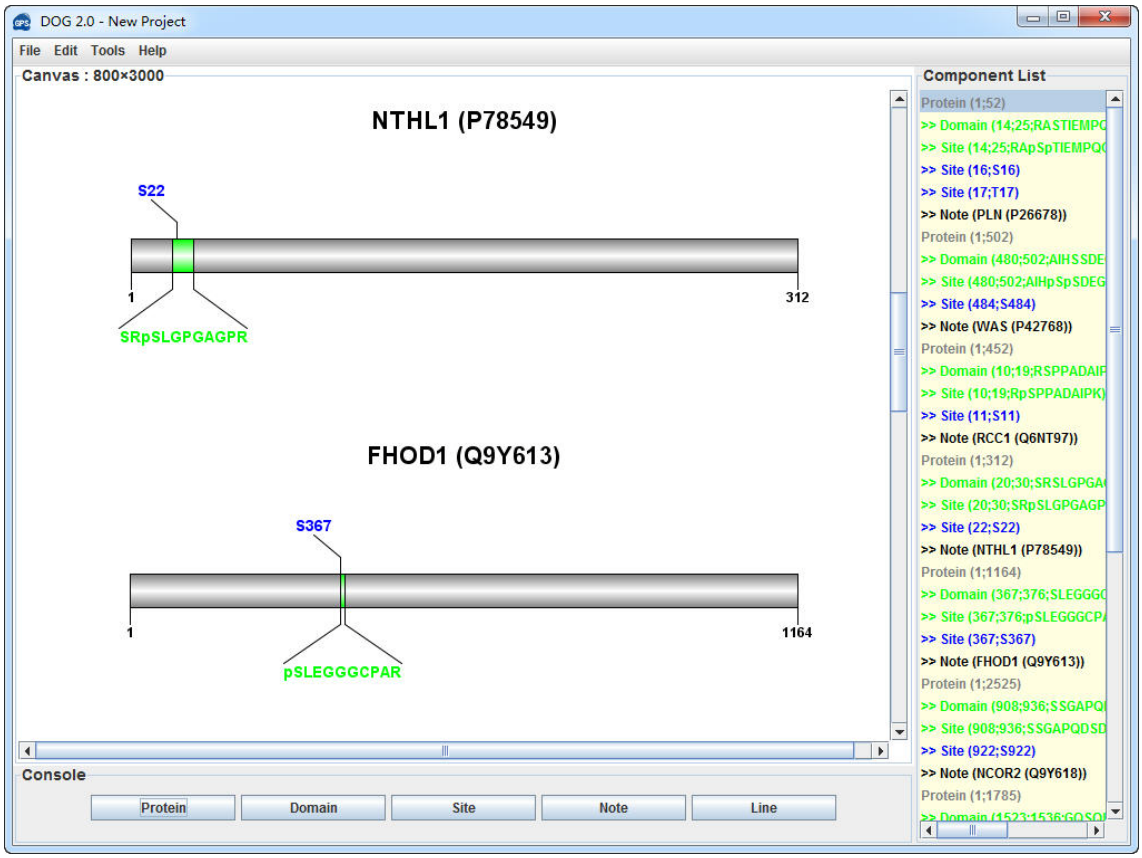
Then please click on the **RIGHT** mouse button in the results form.



You can use the "**Select All**" and "**Copy Selected**" to copy the selected results into Clipboard. Then please copy the results into a file, e.g., an EXCEL file for

further consideration. Also, you can choose “**Export Result**” to export the results into a tab-delimited text file.

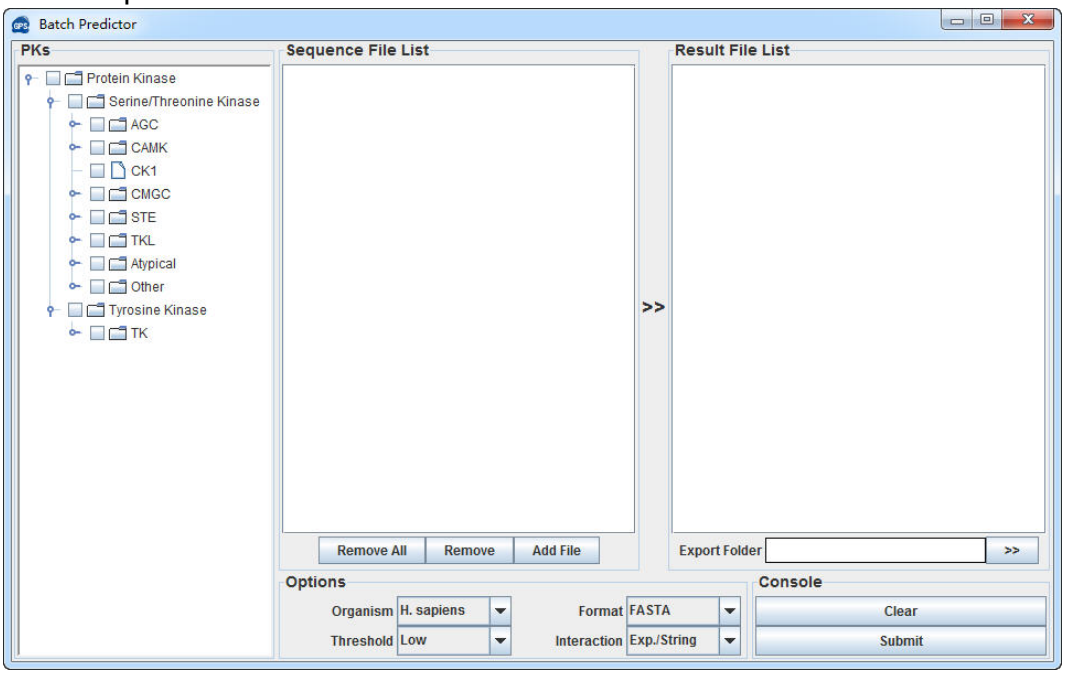
If you choose the **Visualize** function, the given protein, its phosphorylation sites and phospho-peptides will be visualized with DOG (Domain Graph, Version 2.0), an illustrator of protein domain structures.



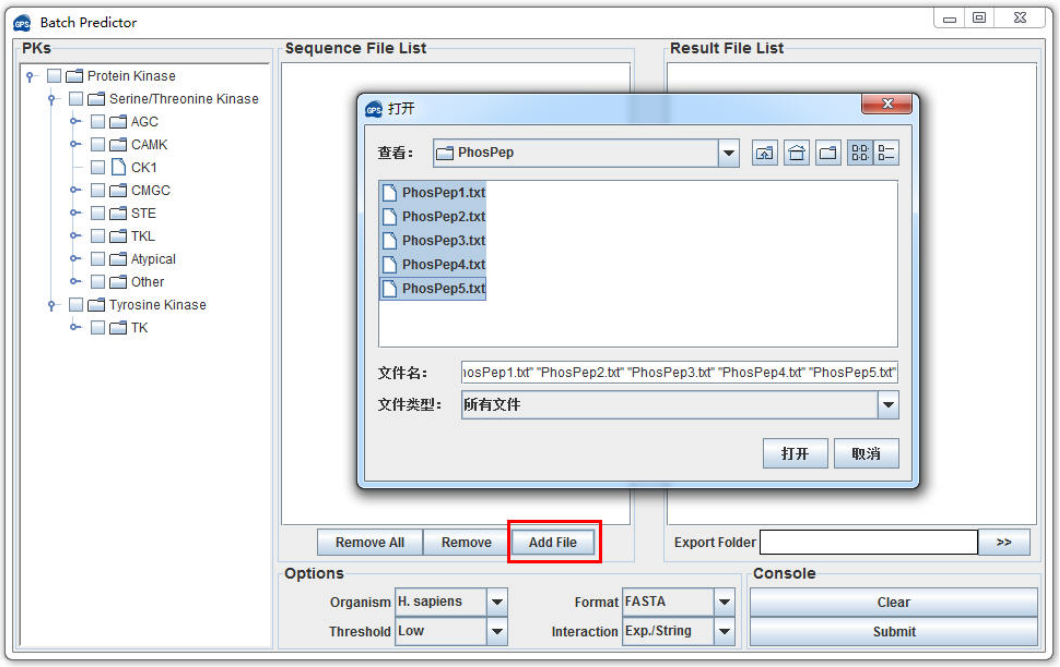
Batch Prediction

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

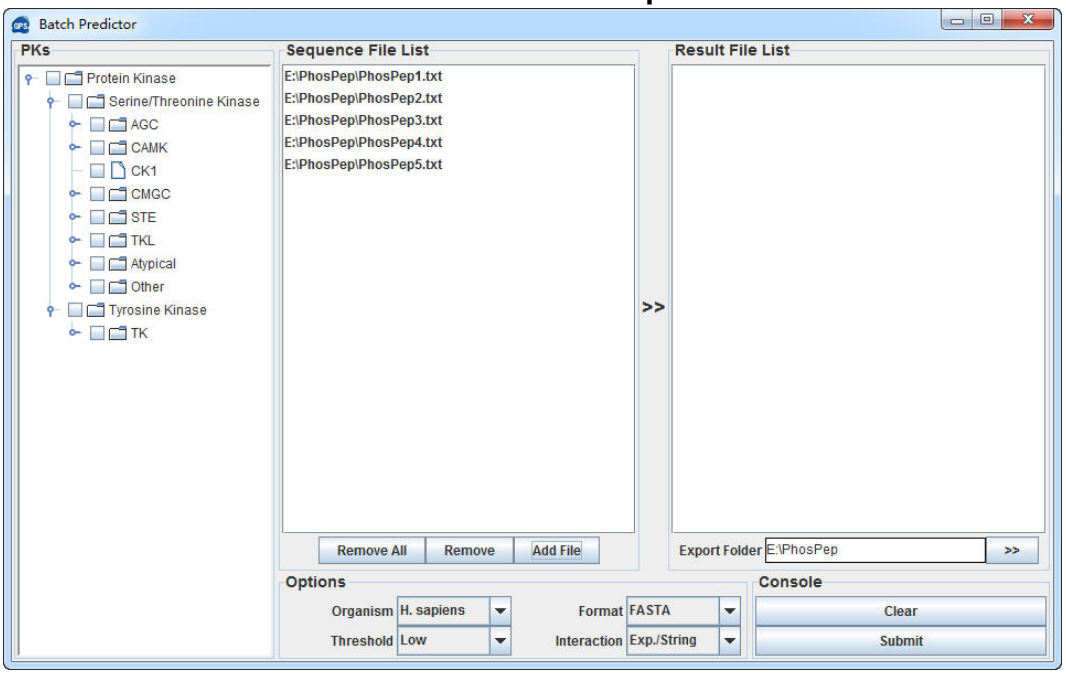
The inputted data should be prepared in **PhosPep**, **ELM** or **FATSA** format. The mixed formats are not permitted. To run the Batch Predictor just select the **Batch Predictor** option in the **Tools** menu.



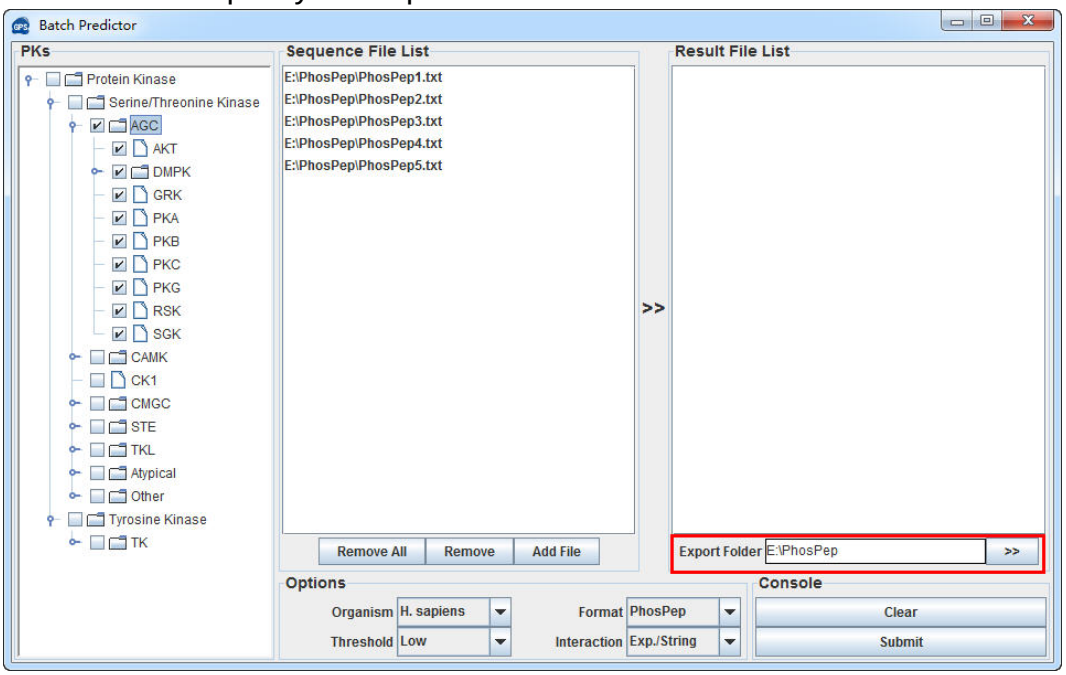
Click on the **Add File** button and add one or more files in your hard disk.



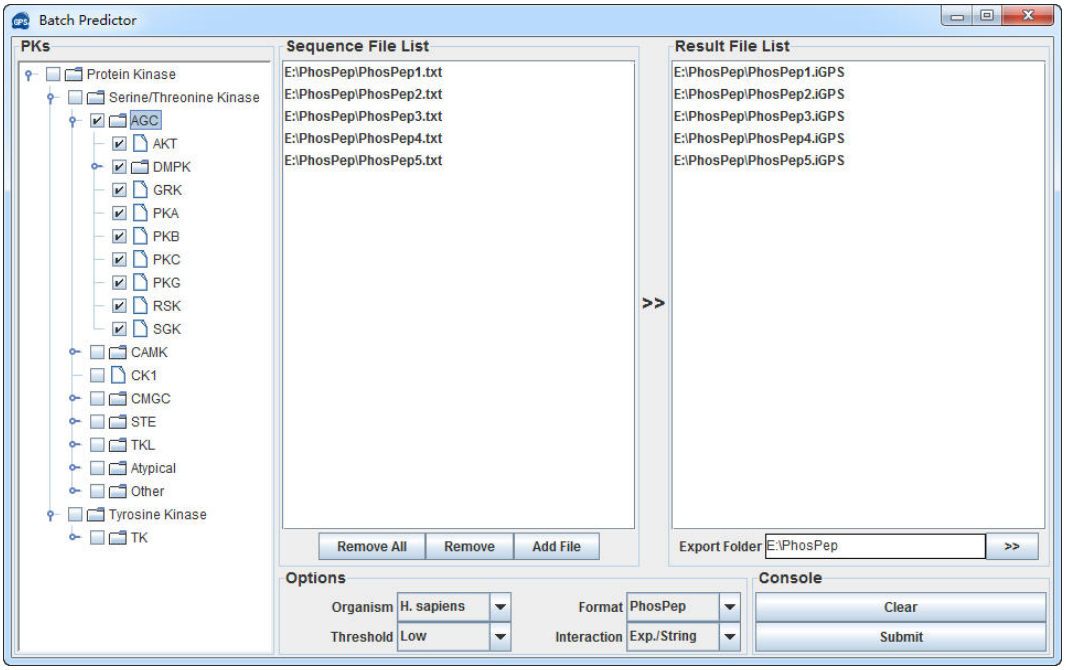
The name of added files will be shown in the **Sequence File List**



The output directory of prediction results should also be defined. Please click on the “>>” button to specify the export file folder.



Choose one or more kinases from **Kinase Hierarchical 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 **Result Export Folder,** and the name of result files will be shown in the **Result File List**.


Batch Predictor
PKs
Protein Kinase
Serine/Threonine Kinase
AGC
AKT
DMPK
GRK
PKA
PKB
PKC
PKG
RSK
SGK
CAMK
CK1
CMGC
STE
TKL
Atypical
Other
Tyrosine Kinase
TK
Sequence File List
E:\PhosPep\PhosPep1.txt
E:\PhosPep\PhosPep2.txt
E:\PhosPep\PhosPep3.txt
E:\PhosPep\PhosPep4.txt
E:\PhosPep\PhosPep5.txt
Remove All
Remove
Add File
>>
Result File List
E:\PhosPep\PhosPep1.iGPS
E:\PhosPep\PhosPep2.iGPS
E:\PhosPep\PhosPep3.iGPS
E:\PhosPep\PhosPep4.iGPS
E:\PhosPep\PhosPep5.iGPS
Export Folder
E:\PhosPep
>>
Options
Organism
H. sapiens
Threshold
Low
Format
PhosPep
Interaction
Exp./String
Console
Clear
Submit

Visualization of Protein Phosphorylation Network with PNC 1.0

After prediction, you can click on the **Network** button or use the right button menu option to build a protein phosphorylation network (PPN).

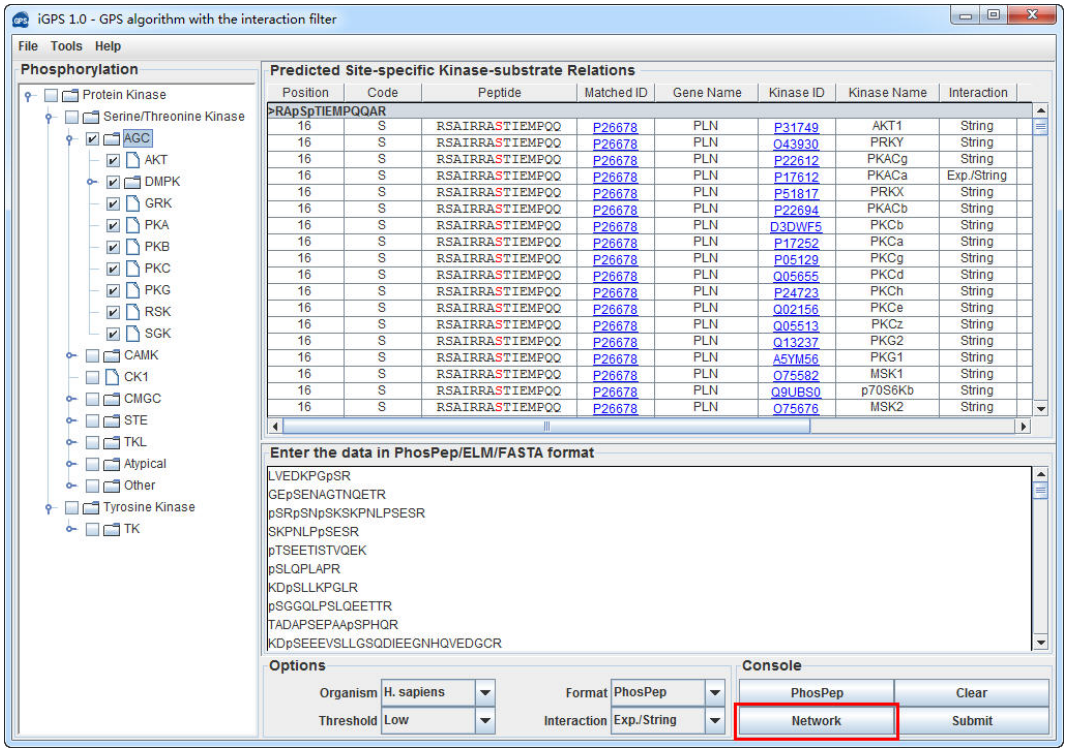
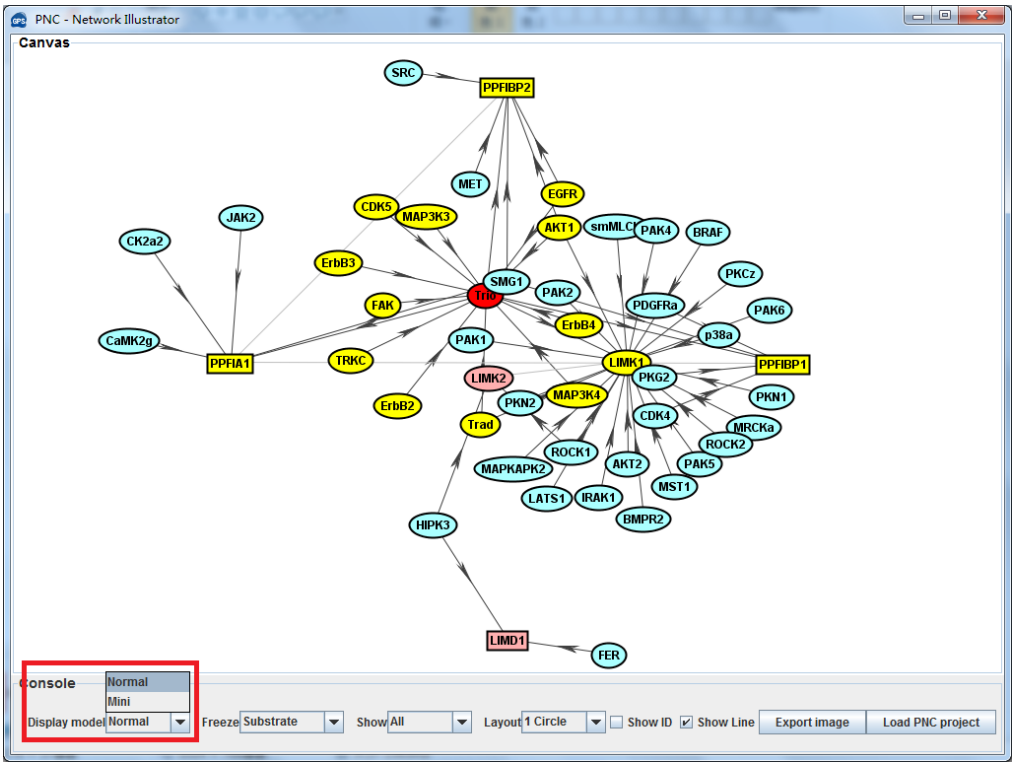


Diagram:

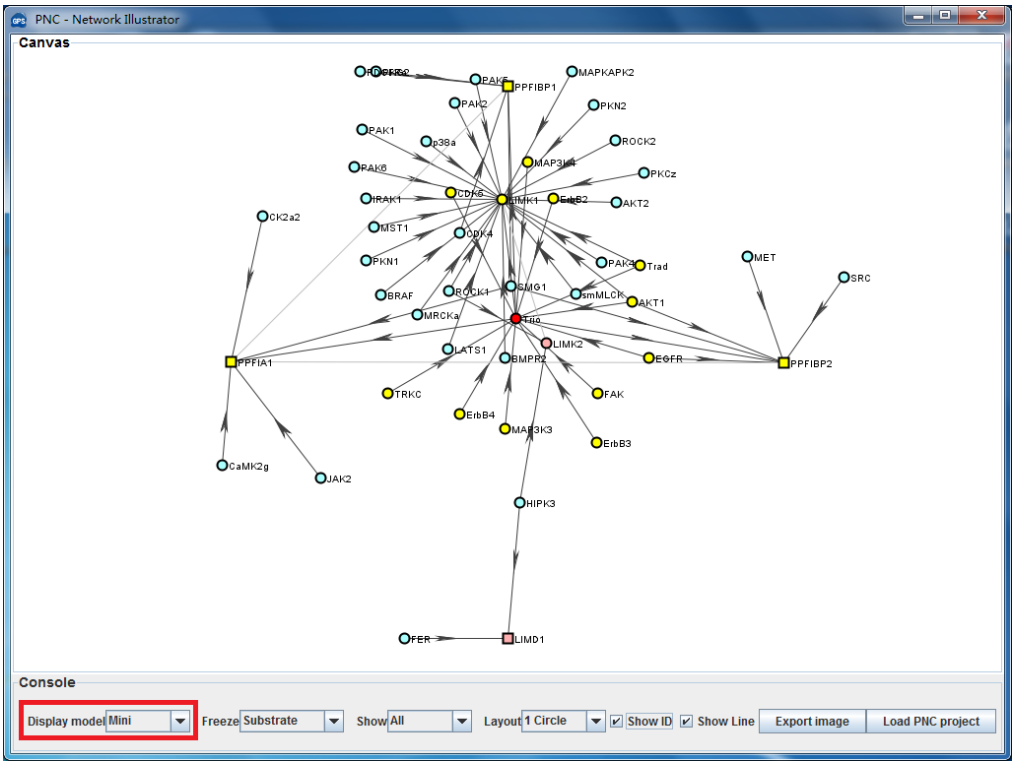
	Normal display model	Mini display model
Substrate in Input		
Kinase in Input		
Regulatory Kinase		
Selected Unit		
Neighbor Unit		

The orientation was defined as Kinase -> Substrate

Normal display model:



Mini display model:

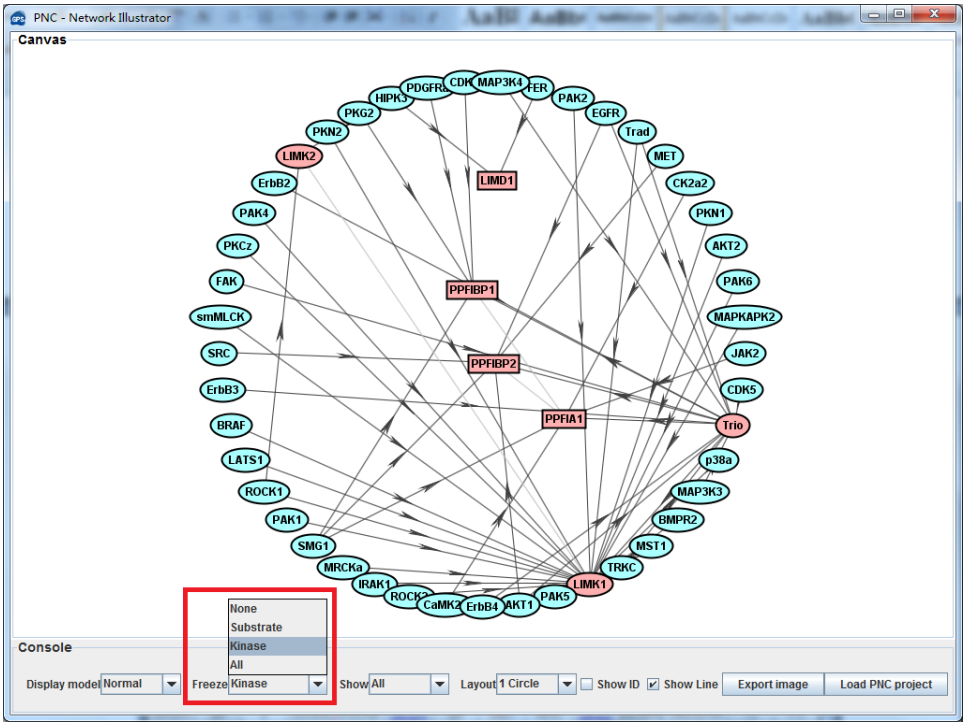


Freeze menu options

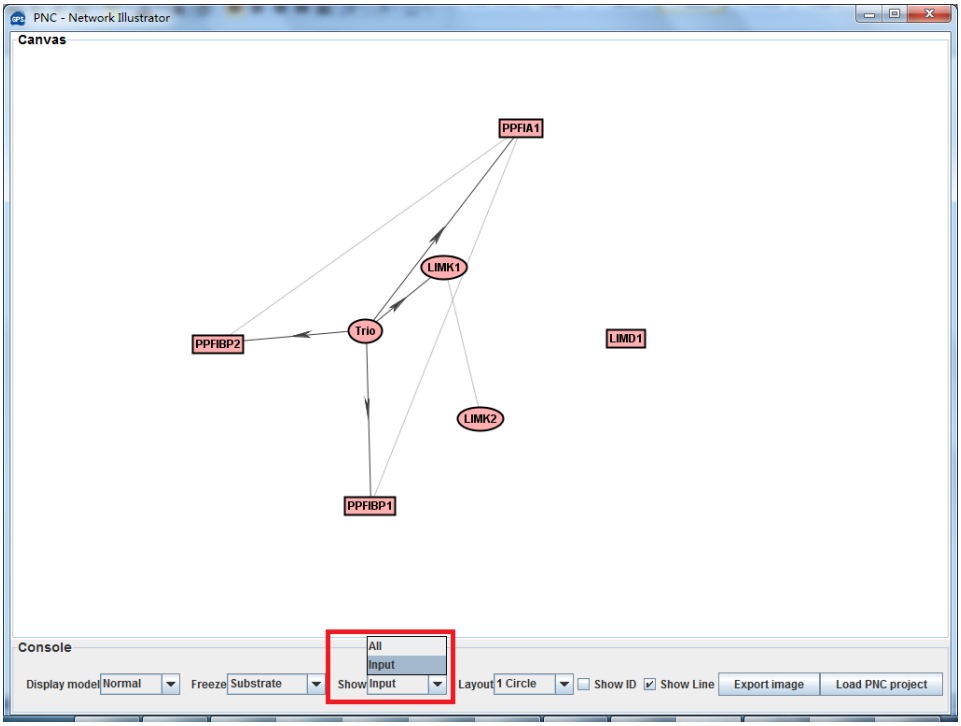
Kinase: freeze input kinase and related kinase. If you choose a **Circle Layout** model, the kinases will be located as hollow circles.

Substrate: freeze the input substrate. If you choose a **Circle Layout** model, the substrates will be located as hollow circles.

All: freeze all display units. **None**: release all display units.



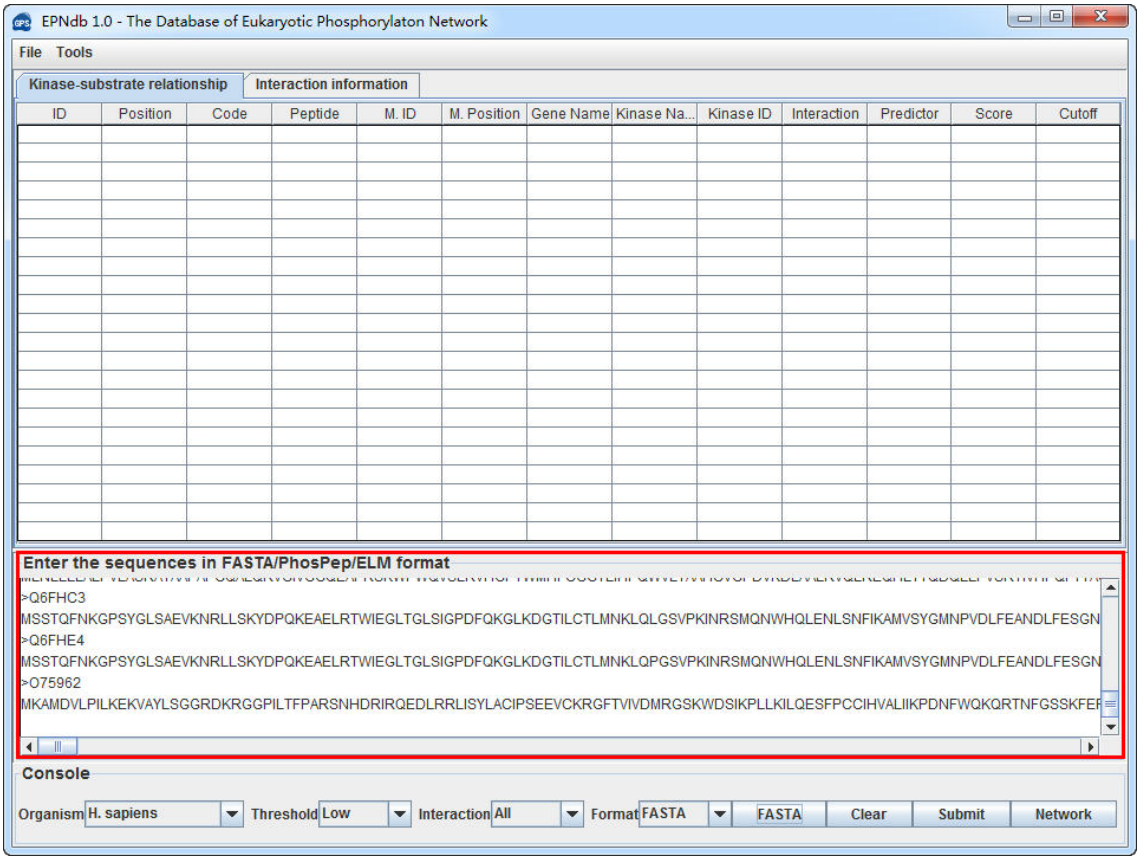
If you choose the **Input** option in the **Show** menu, only the kinases and substrates in the input data will be shown.



Search pre-calculated ssKSRs in EPNdb

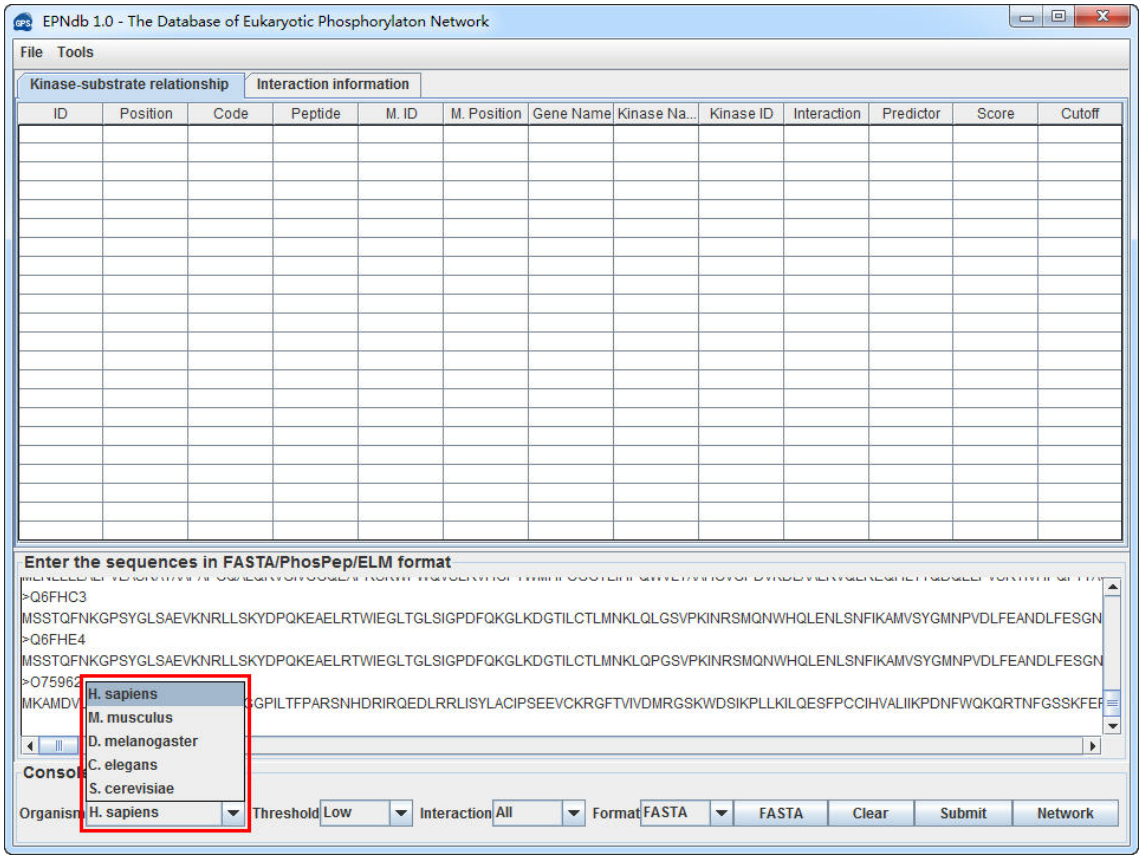
From public databases and the scientific literature, we collected 145,646 experimentally identified phosphorylation sites in 28,457 substrates, with 14,534, 5,555, 15,622, 49,119 and 60,816 phosphorylation sites in *S. cerevisiae*, *C. elegans*, *D. melanogaster*, *M. musculus* and *H. sapiens*, respectively. With the PPI information, we predicted 186,922 (total PPIs) and 34,873 (experimental PPIs) ssKSRs. With these prediction results, we constructed the EPNdb 1.0. The users can search this database for pre-calculated ssKSRs.

Please input data into the “TEXT form”.

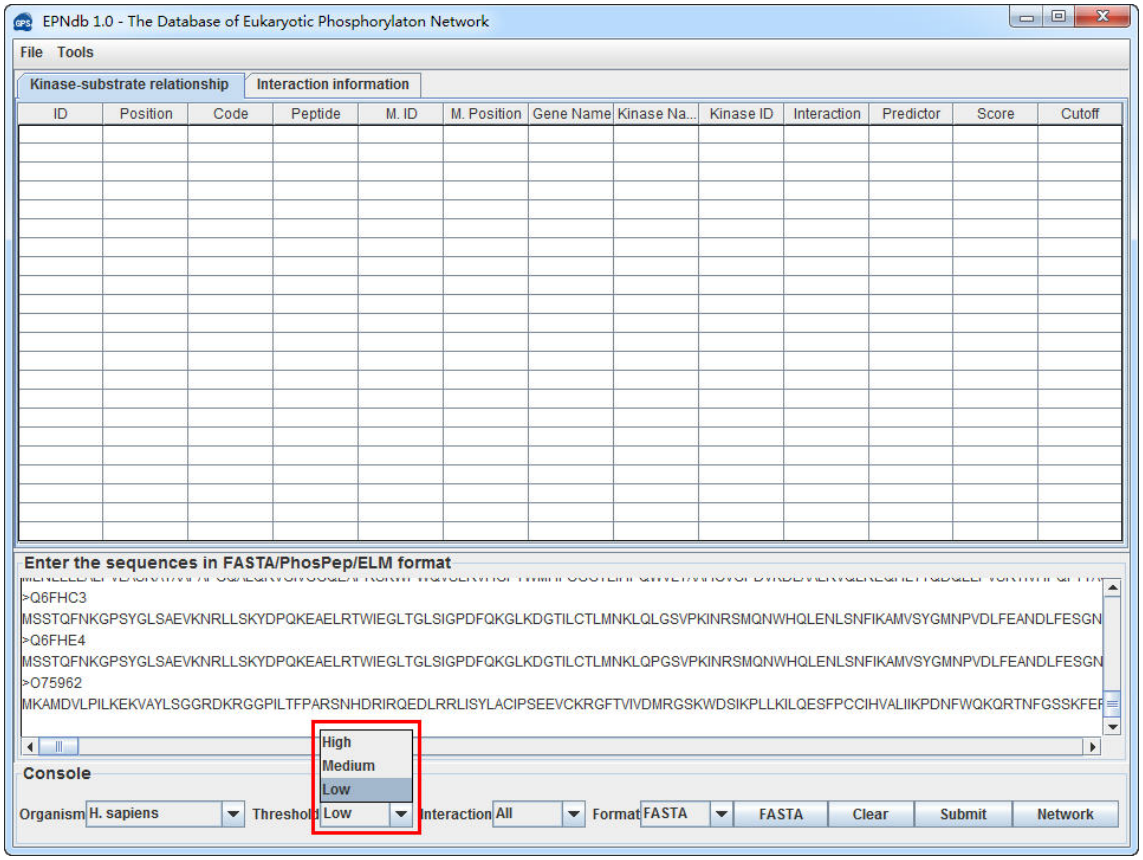


The data should be prepared in **FATSA**, **PhosPep** or **ELM** format.

Please choose a proper organism based on your input data.



Choose a **Threshold** what you need, while the default is **Low**.

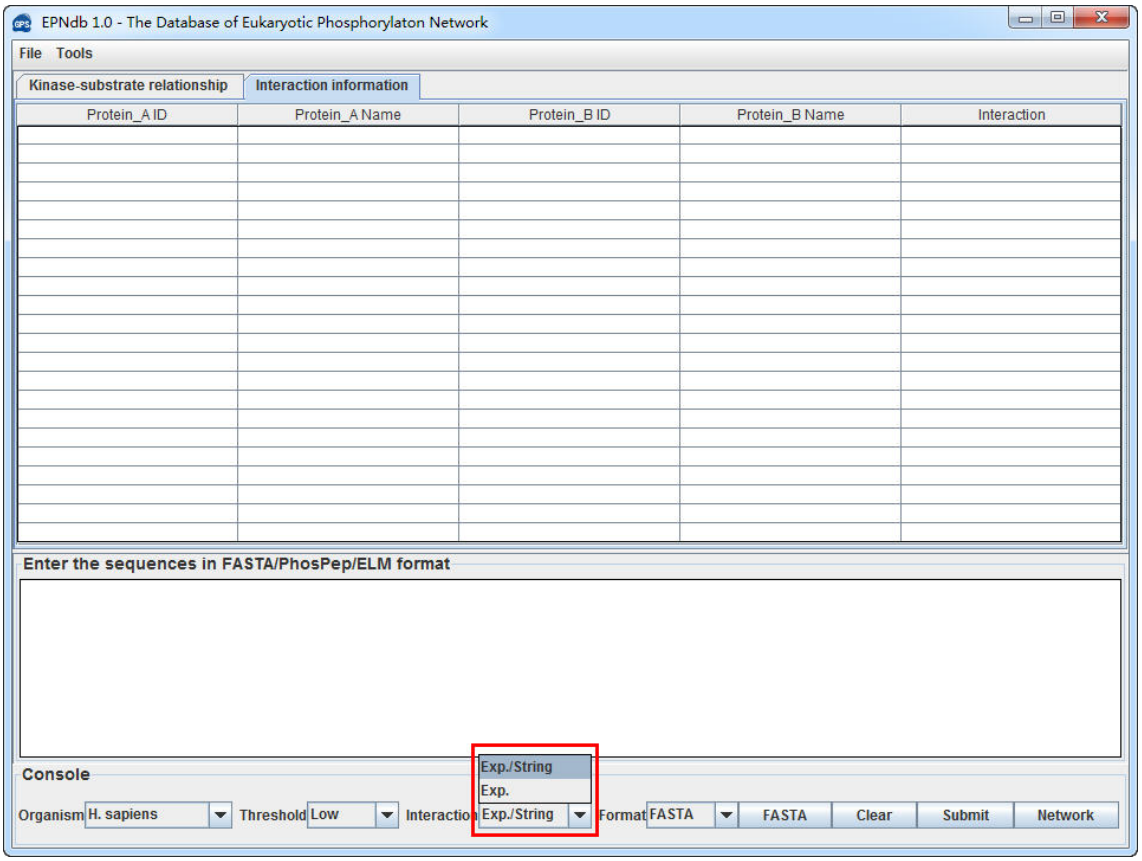


The protein-protein interaction (PPI) information was used as the major filter to reduce false positive predictions.

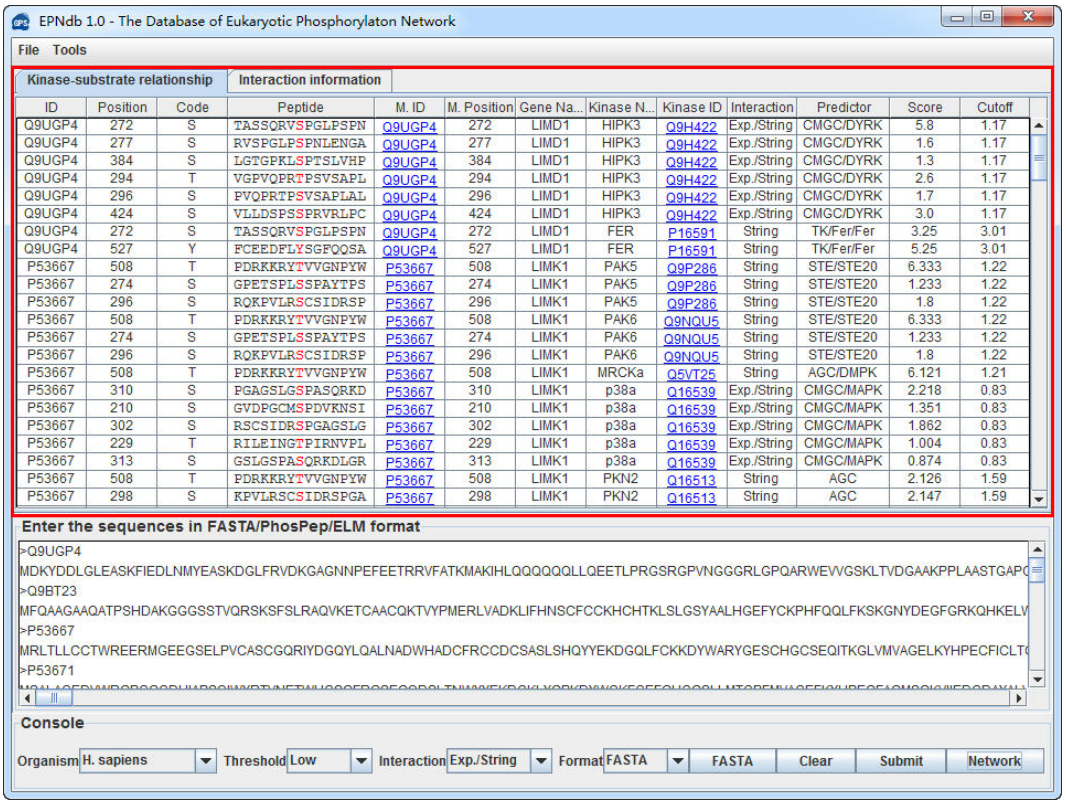
Interaction menu options

Exp.: filter the predictions with the experimentally identified physical interactions between PKs and substrates.

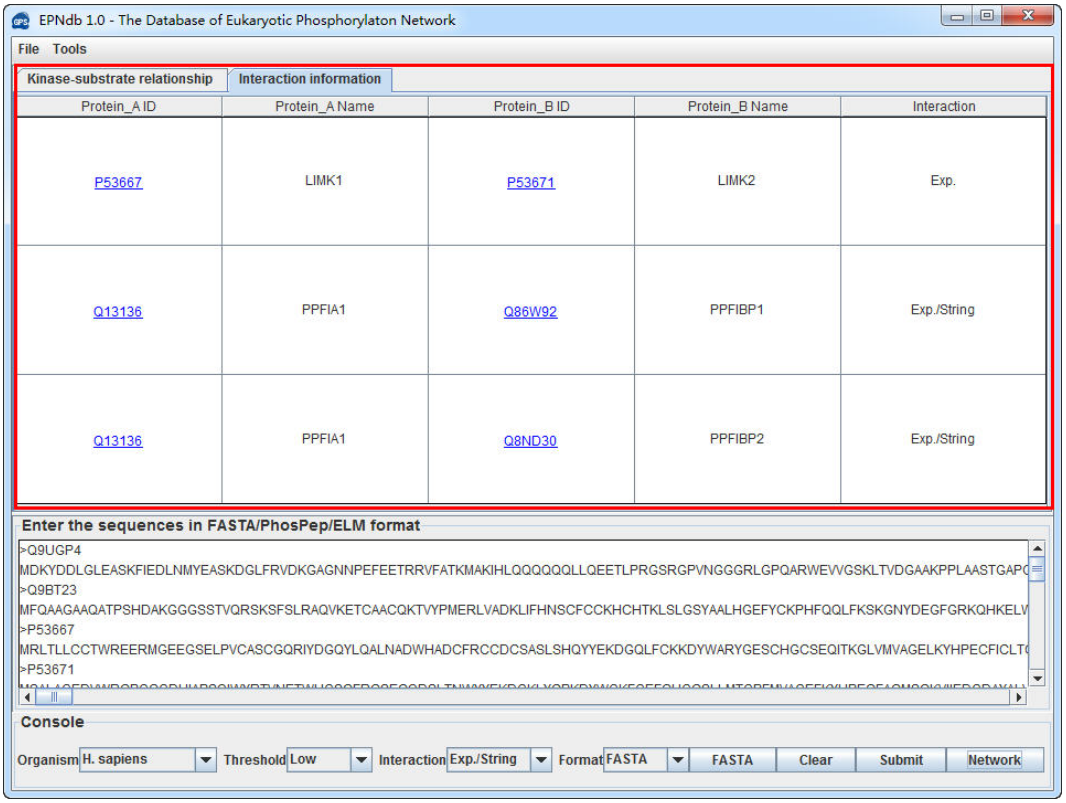
All: filter the predictions with both STRING and experimental PPI.



Click on the **Submit** button, then the predicted site-specific kinase-substrate relations will be shown.



The interactions among input substrates also displayed.



References

1. Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, et al. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell 127: 635-648.
2. Ubersax JA, Ferrell JE, Jr. (2007) Mechanisms of specificity in protein phosphorylation. Nat Rev Mol Cell Biol 8: 530-541.
3. Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, et al. (2005) Global analysis of protein phosphorylation in yeast. Nature 438: 679-684.
4. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. Science 298: 1912-1934.
5. Ptacek J, Snyder M (2006) Charging it up: global analysis of protein phosphorylation. Trends Genet 22: 545-554.
6. Kobe B, Kampmann T, Forwood JK, Listwan P, Brinkworth RI (2005) Substrate specificity of protein kinases and computational prediction of substrates. Biochim Biophys Acta 1754: 200-209.
7. Hjerrild M, Gammeltoft S (2006) Phosphoproteomics toolbox: computational biology, protein chemistry and mass spectrometry. FEBS Lett 580: 4764-4770.
8. Kreegipuu A, Blom N, Brunak S, Jarv J (1998) Statistical analysis of protein kinase specificity determinants. FEBS Lett 430: 45-50.
9. Songyang Z, Lu KP, Kwon YT, Tsai LH, Filhol O, et al. (1996) A structural basis for substrate specificities of protein Ser/Thr kinases: primary sequence preference of casein kinases I and II, NIMA, phosphorylase kinase, calmodulin-dependent kinase II, CDK5, and Erk1. Mol Cell Biol 16: 6486-6493.
10. Yaffe MB, Leparc GG, Lai J, Obata T, Volinia S, et al. (2001) A motif-based profile scanning approach for genome-wide prediction of signaling pathways. Nat Biotechnol 19: 348-353.
11. Linding R, Jensen LJ, Ostheimer GJ, van Vugt MA, Jorgensen C, et al. (2007) Systematic discovery of in vivo phosphorylation networks. Cell 129: 1415-1426.
12. Linding R, Jensen LJ, Pasculescu A, Olhovsky M, Colwill K, et al. (2008) NetworKIN: a resource for exploring cellular phosphorylation networks. Nucleic Acids Res 36: D695-699.
13. Biondi RM, Nebreda AR (2003) Signalling specificity of Ser/Thr protein kinases through docking-site-mediated interactions. Biochem J 372: 1-13.
14. Holland PM, Cooper JA (1999) Protein modification: docking sites for kinases. Curr Biol 9: R329-331.
15. Tan CS, Linding R (2009) Experimental and computational tools useful for (re)construction of dynamic kinase-substrate networks. Proteomics 9: 5233-5242.
16. Erxleben C, Liao Y, Gentile S, Chin D, Gomez-Alegria C, et al. (2006)

Cyclosporin and Timothy syndrome increase mode 2 gating of CaV1.2 calcium channels through aberrant phosphorylation of S6 helices. Proc Natl Acad Sci U S A 103: 3932-3937.

17. Gentile S, Martin N, Scappini E, Williams J, Erxleben C, et al. (2008) The human ERG1 channel polymorphism, K897T, creates a phosphorylation site that inhibits channel activity. Proc Natl Acad Sci U S A 105: 14704-14708.
18. Ren J, Jiang C, Gao X, Liu Z, Yuan Z, et al. (2010) PhosSNP for systematic analysis of genetic polymorphisms that influence protein phosphorylation. Mol Cell Proteomics 9: 623-634.
19. Radivojac P, Baenziger PH, Kann MG, Mort ME, Hahn MW, et al. (2008) Gain and loss of phosphorylation sites in human cancer. Bioinformatics 24: i241-247.
20. Diella F, Gould CM, Chica C, Via A, Gibson TJ (2008) Phospho.ELM: a database of phosphorylation sites--update 2008. Nucleic Acids Res 36: D240-244.
21. Villen J, Beausoleil SA, Gerber SA, Gygi SP (2007) Large-scale phosphorylation analysis of mouse liver. Proc Natl Acad Sci U S A 104: 1488-1493.
22. Grimsrud PA, Swaney DL, Wenger CD, Beauchene NA, Coon JJ (2010) Phosphoproteomics for the masses. ACS Chem Biol 5: 105-119.
23. Matsuoka S, Ballif BA, Smogorzewska A, McDonald ER, 3rd, Hurov KE, et al. (2007) ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. Science 316: 1160-1166.
24. Song C, Ye M, Han G, Jiang X, Wang F, et al. (2010) Reversed-phase-reversed-phase liquid chromatography approach with high orthogonality for multidimensional separation of phosphopeptides. Anal Chem 82: 53-56.
25. Hu ZZ, Narayanaswamy M, Ravikumar KE, Vijay-Shanker K, Wu CH (2005) Literature mining and database annotation of protein phosphorylation using a rule-based system. Bioinformatics 21: 2759-2765.
26. Xue Y, Ren J, Gao X, Jin C, Wen L, et al. (2008) GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. Mol Cell Proteomics 7: 1598-1608.

Release Note

1. May. 26th, 2010, the alpha version of GPS-PNC was constructed for testing.
2. Oct. 28th, 2010, the beta version of GPS-PNC software was released for testing and debugging.
3. Dec. 16th, 2010, the stand-alone packages of GPS-PNC 1.0 were released.
4. Jul. 23rd, 2011, some bugs were fixed in GPS-PNC 1.0.
5. Jan. 3rd, 2012, the GPS-PNC 1.0 was renamed in EPNdb 1.0, while the main program was re-designed as iGPS 1.0. The EPNdb 1.0 was a plugin of iGPS 1.0. The stand-alone packages of iGPS 1.0 were released.
6. Aug. 27 th, 2012, iGPS 1.0.1 was released. The output format was changed from FASTA to TAB.