PPS Manual

PTMs Peptide Scanner

Version 1.0
05/12/2008

Major Authors: Yu Xue & Jian Ren

Contact: Dr. Yu Xue, xueyu@ustc.edu.cn; Dr. Jian Ren, renjian@ustc.edu.cn
The CUCKOO Workgroup
The latest version of PPS is available from http://bioinformatics.lcd-ustc.org/pps

Index

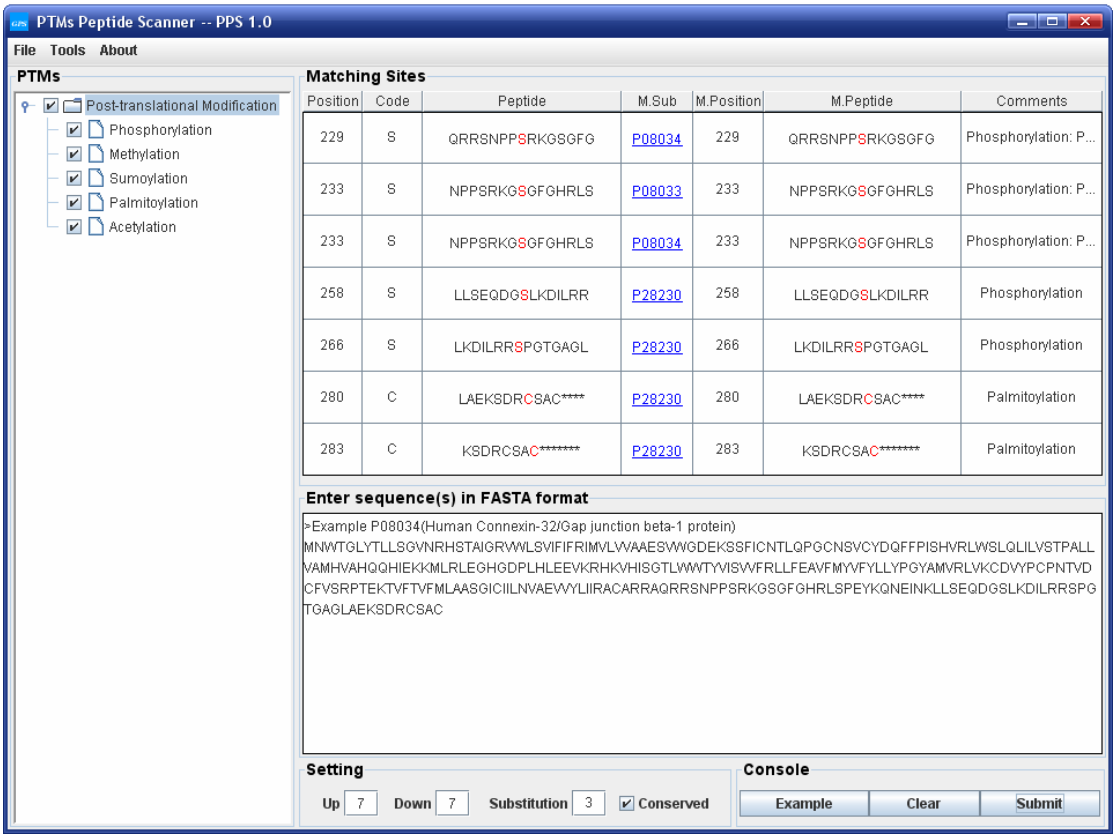
Statement

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**. The work of CUCKOO Workgroup is supported by grants from Chinese 973 project (2006CBOF0503, and 2006CB933300), Chinese Academy of Sciences (KSCX1-YW-R65, KSCX2-YW-21, and KJCX2-YW-M02), and Chinese Natural Science Foundation (39925018, 30270293, 90508002, 30700138).

Introduction

Recently, an interesting and important question has emerged from feedbacks of experimentalists. Frequently, experimental researchers prefer to study novel PTM sites to elucidate new functions. Thus, it's important to show whether there have been any sites experimentally verified as real PTM sites in given proteins. And such information will be greatly useful for researchers to avoid reduplicate work. Although we and other bioinformatists developed numerous tools for PTM sites prediction, these softwares could only regard prediction results as potentially real sites. Currently, several public databases, e.g., Phospho.ELM (Diella, et al., 2004; Diella, et al., 2008) and UniProt (Boutet, et al., 2007), have been developed to contain PTM information of proteins. However, these annotations are usually not integrated. Moreover, due to the diversity and heterogeneity of protein names, it's difficult to fully obtain the known PTM information for a given protein, by database searching.

In this work, we developed a novel software of PTMs Peptide Scanner (PPS), to reveal known or highly potential PTM sites in eukaryotic proteins. Five typical PTMs were considered, including phosphorylation, sumoylation, palmitoylation, methylation and acetylation. And the experimentally verified PTM sites were taken from Phospho.ELM 7.0 (Diella, et al., 2004; Diella, et al., 2008) and our previous studies (Chen, et al., 2006; Li, et al., 2006; Ren, et al., 2008; Xue, et al., 2006; Xue, et al., 2006; Zhou, et al., 2006), containing 18 179 known sites. Based on our previous hypothesis of similar peptides with potentially similar functions, we designed a straightforward approach of conserved peptide matching (CPM) algorithm. Given a protein sequence as input, PPS will compare it to the experimentally verified PTMs peptides to find the identical or highly conserved hits. The identical hits might be bona fide modified peptides in experimentally verified proteins or conserved in their highly similar homologs. Thus, PPS could be useful for annotation of covalent modifications information across eukaryotes. As an application, we computationally revealed 71 663 identical hits in six eukaryotic organisms, including H. sapiens, M. musculus, D. melanogaster, C. elegans, S. pombe, and S. cerevisiae. Obviously, most of these results could be highly potential PTM sites and greatly helpful for further experimental verification. Furthermore, highly conserved hits might also point to potentially conserved modifications. In our results, there were 7 911 highly potential PTM hits (with $\leqslant 3$ conservative substitutions) found in the six eukaryotes. In addition, we carried out a proteome-wide study of creation or disruption of covalent modification sites by alternative splicing (AS) in H. sapiens. Taken together, we proposed PPS could be a multiple useful tool for PTM sites analyses. Finally, the online service and local packages of PPS 1.0 were implemented in JAVA 1.5.

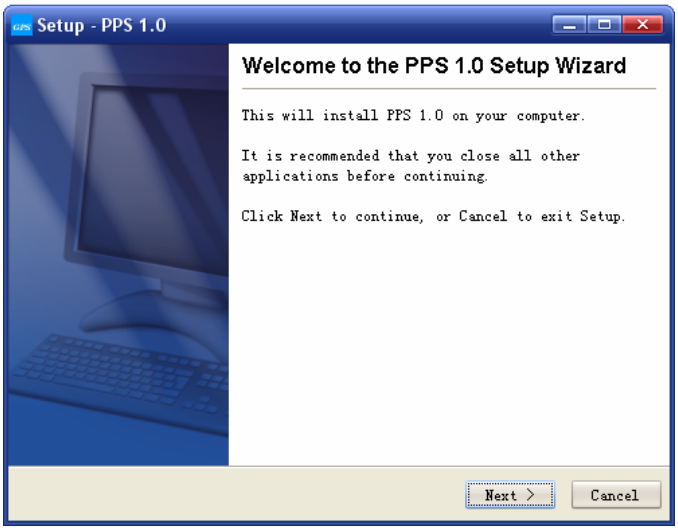


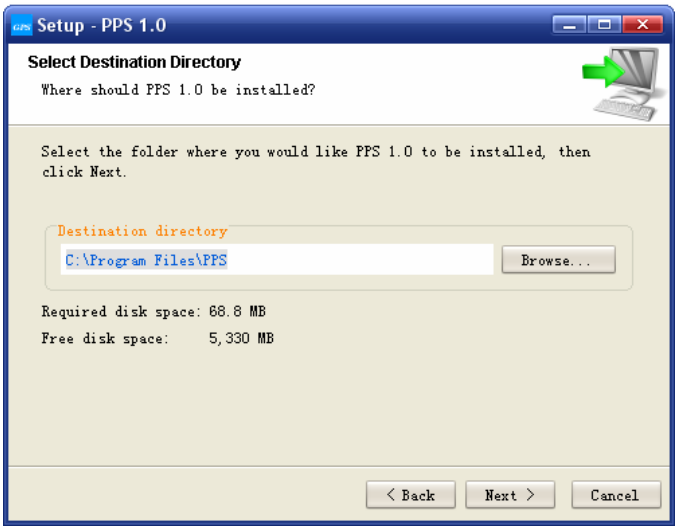
PPS 1.0 User Interface

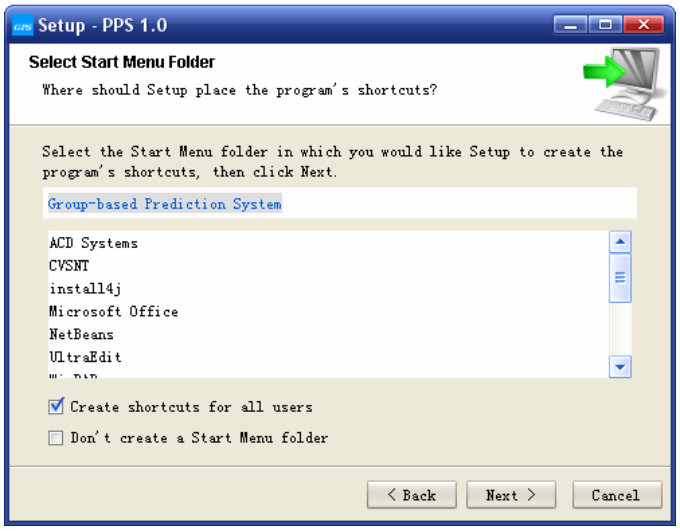
Download & Installation

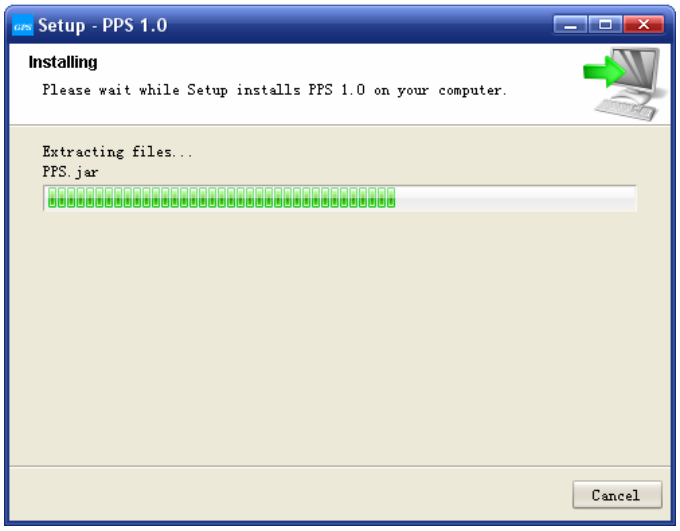
The PPS 1.0 was implemented in JAVA (J2SE), and could support three major Operating Systems (OS), including Windows, Linux/Unix or Mac OS X systems. Both of online web service and local stand-alone packages are available from: http://bioinformatics.lcd-ustc.org/pps/down.php. We recommend that users could download the latest release.

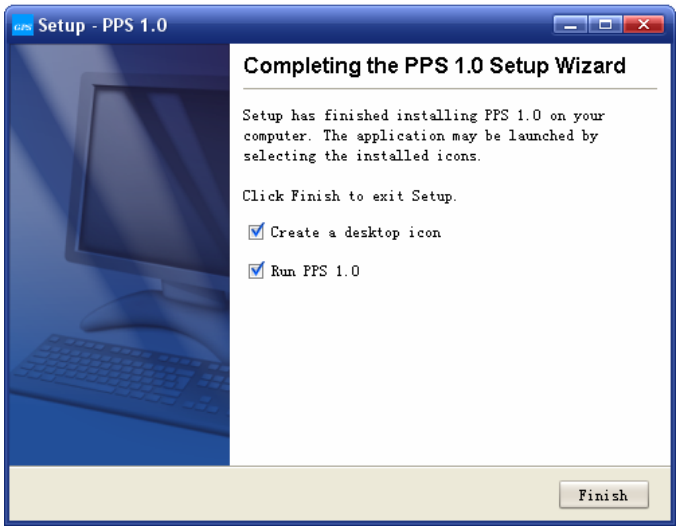
Please choose the proper package to download. After downloading, please double-click on the software package to begin installation, following the user prompts through the installation. And snapshots of the setup program for windows are shown below:











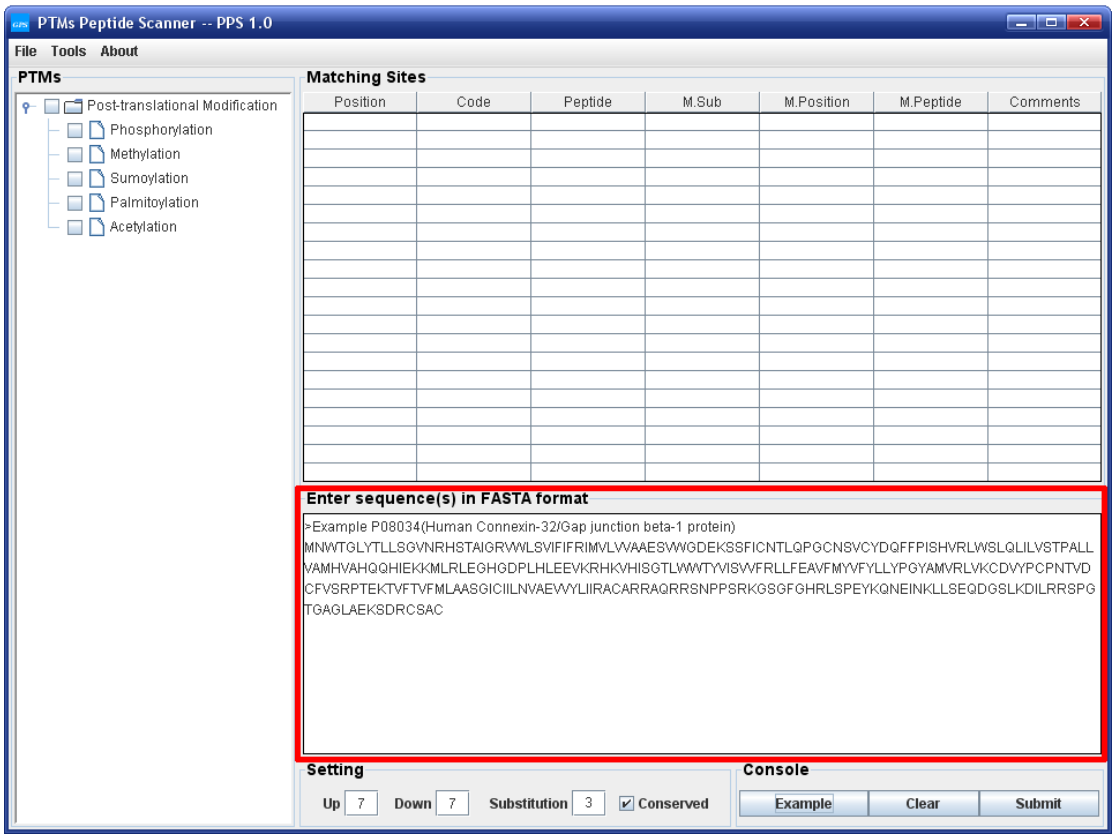
Finally, please click on the **Finish** button to complete the setup program.

Conserved peptide matching

1. A single protein sequence in FASTA format

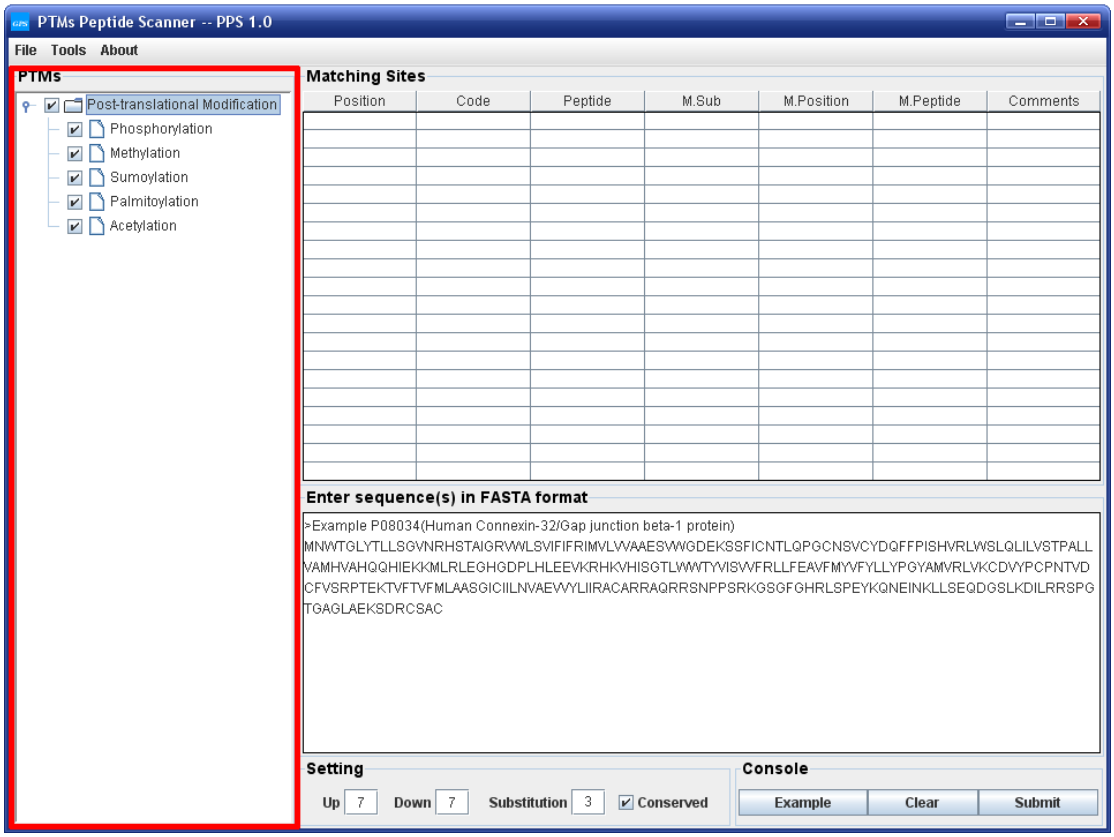
The following steps show you how to use the PPS 1.0 to match the conserved peptides for a single protein sequence in FASTA format.

(1) Firstly, please use “Ctrl+C & Ctrl+V” (Windows & Linux/Unix) or “Command+C & Command+V” (Mac) to copy and paste your sequence into the text form of PPS 1.0

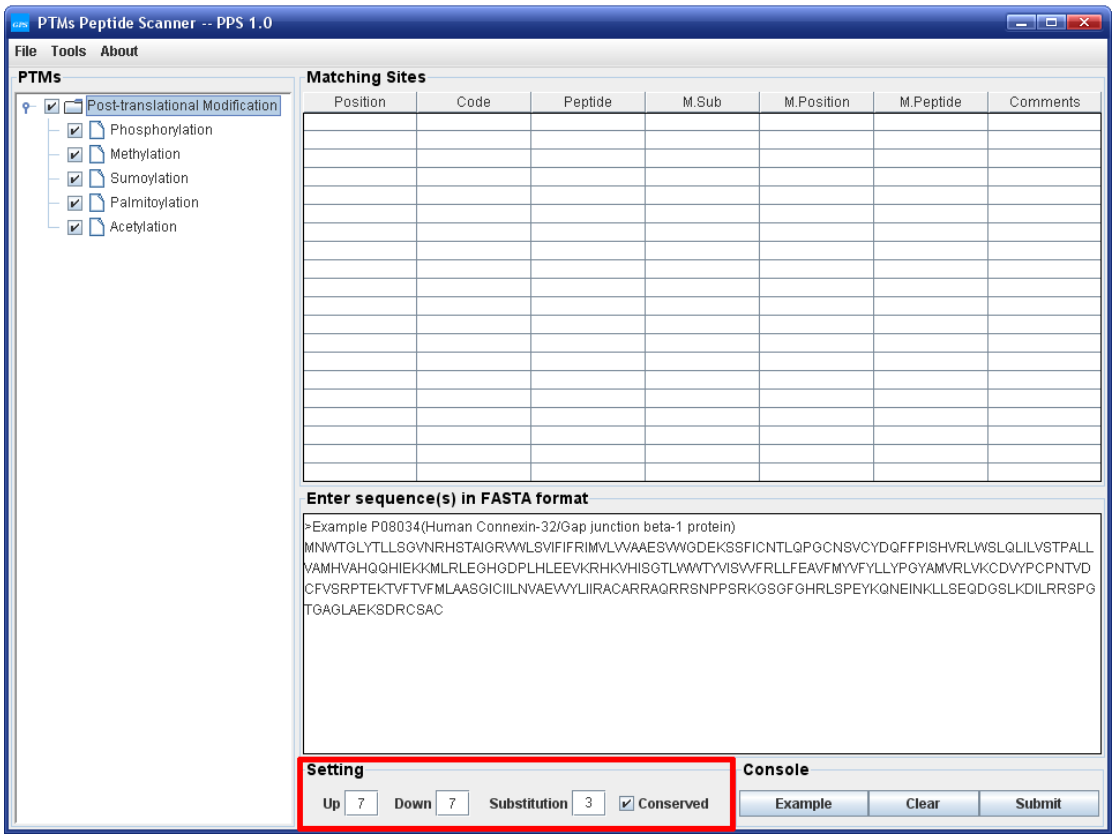


Note: for a single protein, the sequence without a name in raw format is also OK. However, for multiple sequences, the name of each protein should be presented.

(2) Choose one or more kinds of Post-translational Modification from **PTMs Tree**.

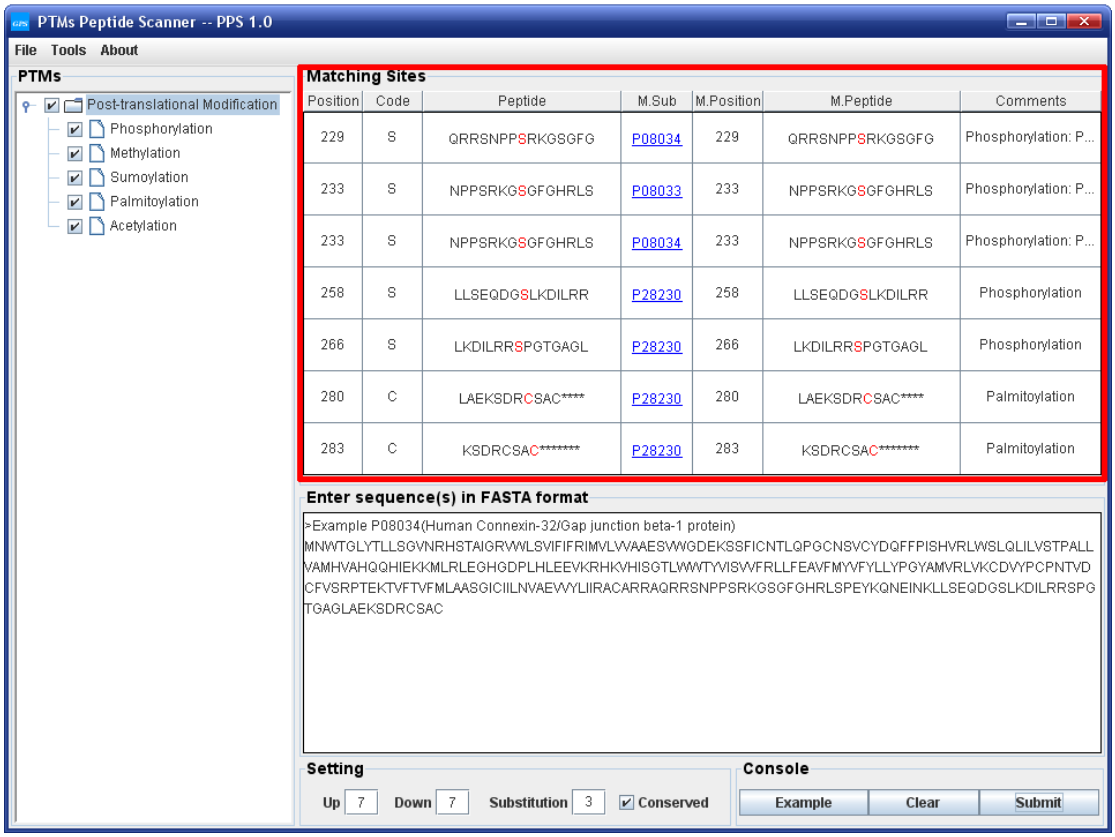


(3) Set the basic parameters of matching.



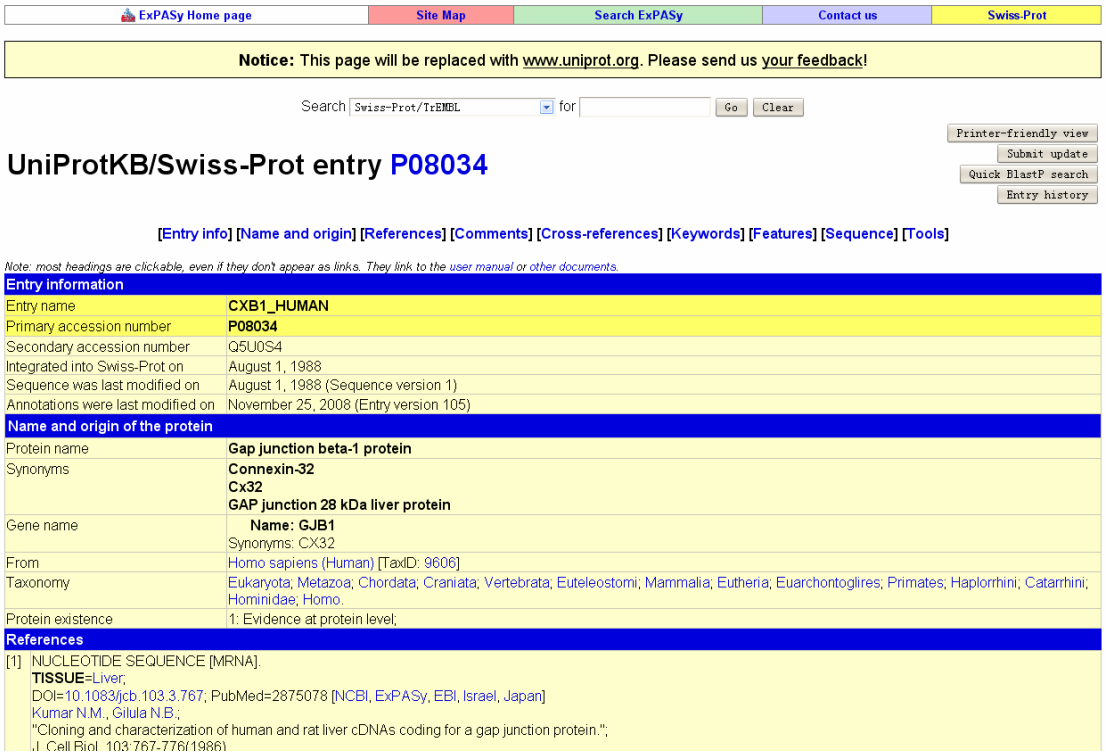
Here we defined a *potentially modified peptide* PMP(m, n) as a modified site flanked by m residues upstream and n residues downstream. Based on a previous article of residues clustering[26], we classified twenty common amino acids into five conservative groups, including the residues with large hydrophobic side chains (L, V, I, M), the hydrophobic aromatic residues (F, Y, W), the long-chain positively charged residues (K, R), the alcohol residues (S, T) and the charged/polar residues (E, D, N, Q). At the same position, two different amino acids of the same group will be regarded as a conserved substitution. The parameters of PMP(7, 7) with $\leqslant 3$ conserved substitutions were chosen as default.

(4) Click on the **Submit** button, then the matched conserved sites will be shown.

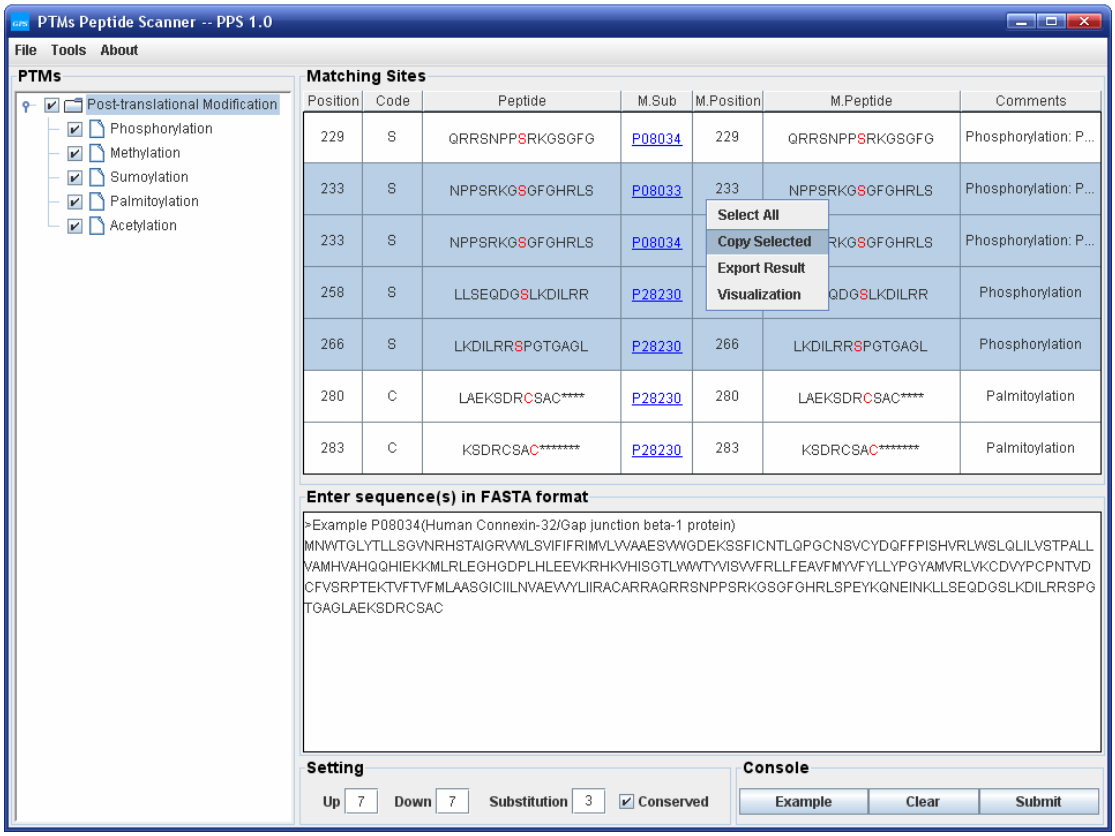


The first three columns contain the site information of peptides in the input protein sequence. The last four columns show the information of matched sites which have been experimentally verified as real PTM sites.

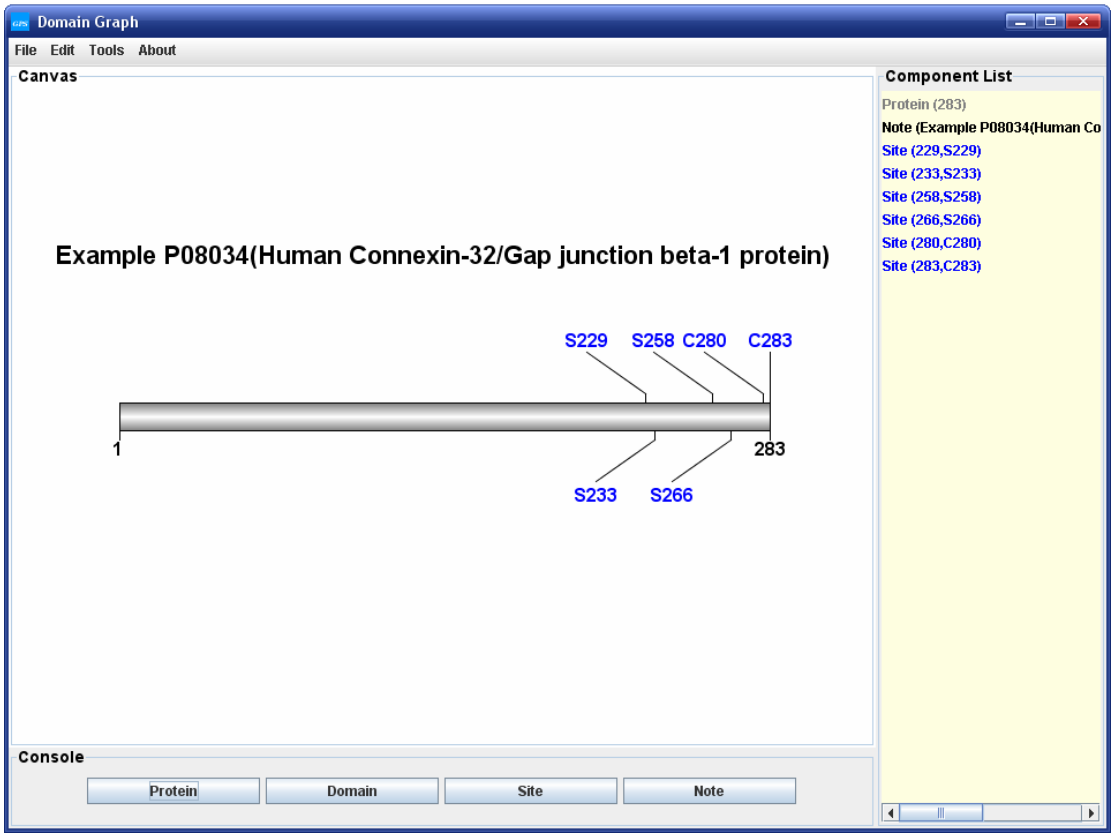
If you want get more information of the matched proteins, you can click on the hyperlinks in the table. The hyperlinks will access the **UniProt** database and show you the detailed information.



(5) Then please click on the **RIGHT** button in the matching 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 **Visualization** function, the given protein and its matching sites will be visualized with DOG (Domain Graph, Ver 1.0), an illustrator of protein domain structures.

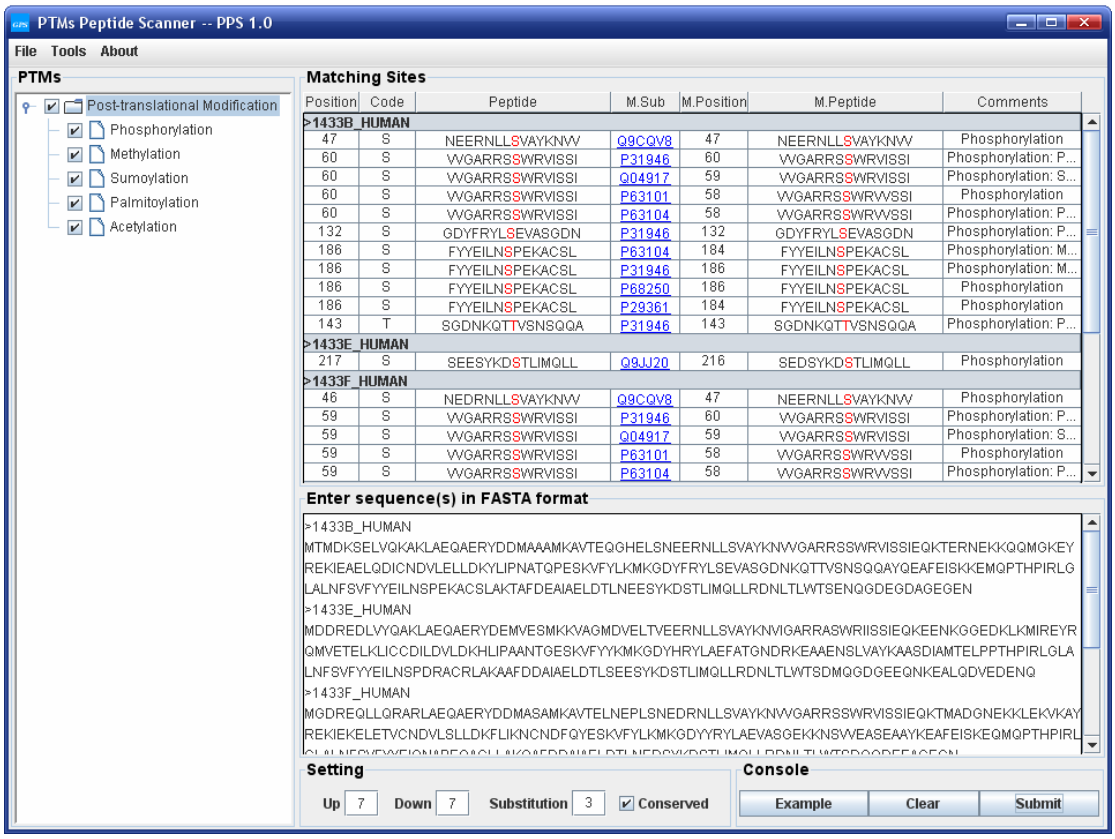


2. Multiple protein sequences in FASTA format

For multiple protein sequences, there are two ways to use the PPS 1.0.

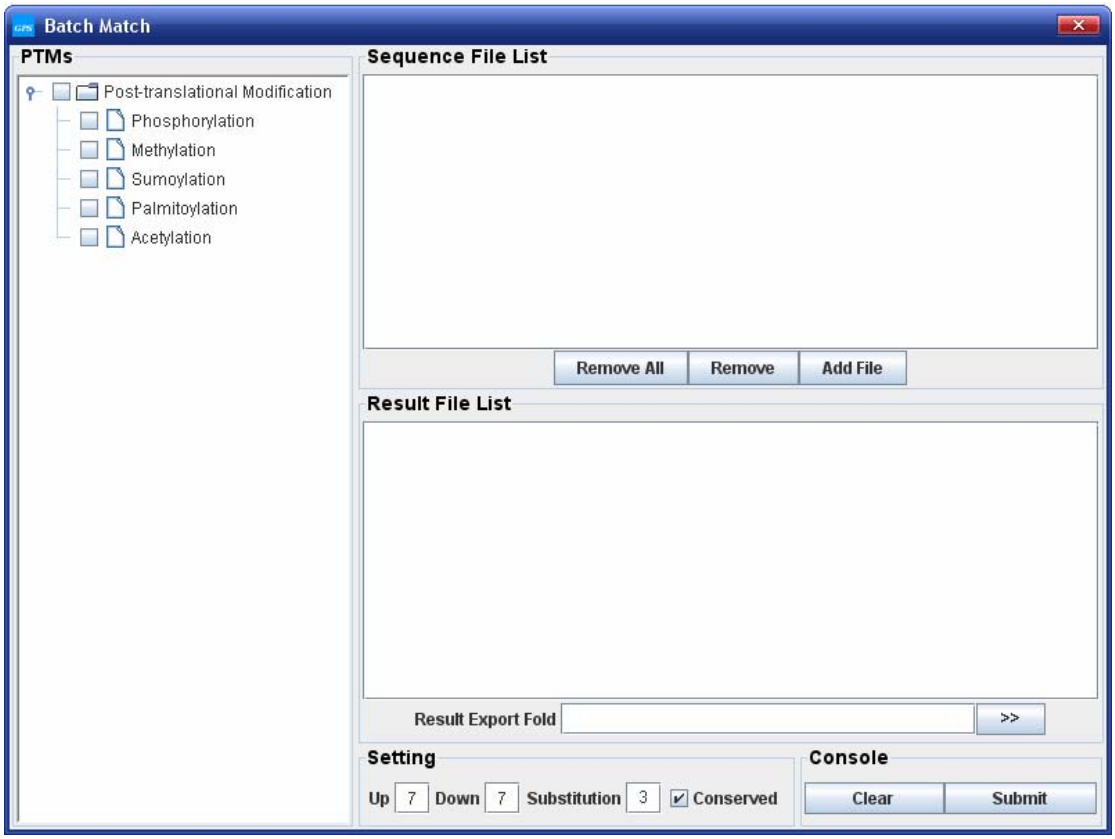
A. Input the sequences into text form directly. (Num. of Seq ≤ 2,000)

If the number of total protein sequences is not greater than 2,000, you can just use “Ctrl+C & Ctrl+V” (Windows & Linux/Unix) or “Command+C & Command+V” (Mac) to copy and paste your sequences into the text form of PPS 1.0 for matching. The name of each protein should be presented.



B. Use Batch Match tool.

If the number of protein sequences is very large, e.g., yeast or human proteome, please use the **Batch Match Tool**. Please click on the "**Batch Match**" button in the **Tools** menu.

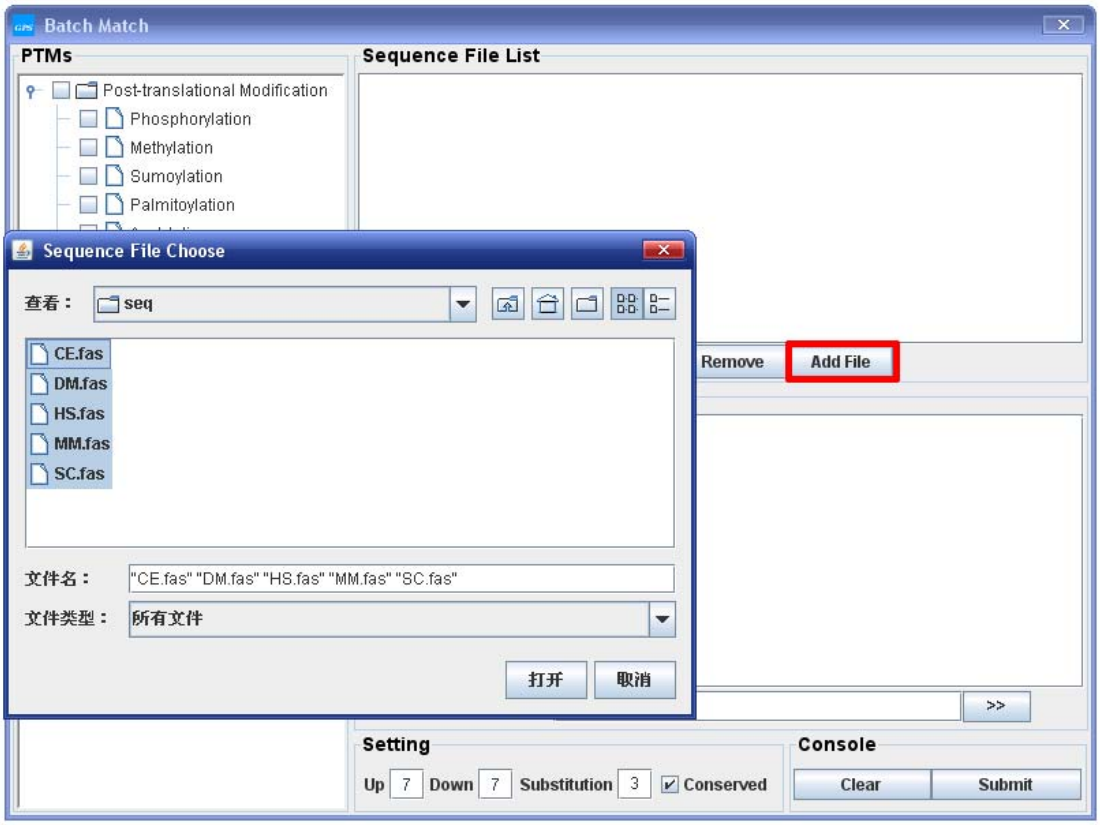


The following steps show you how to use it:

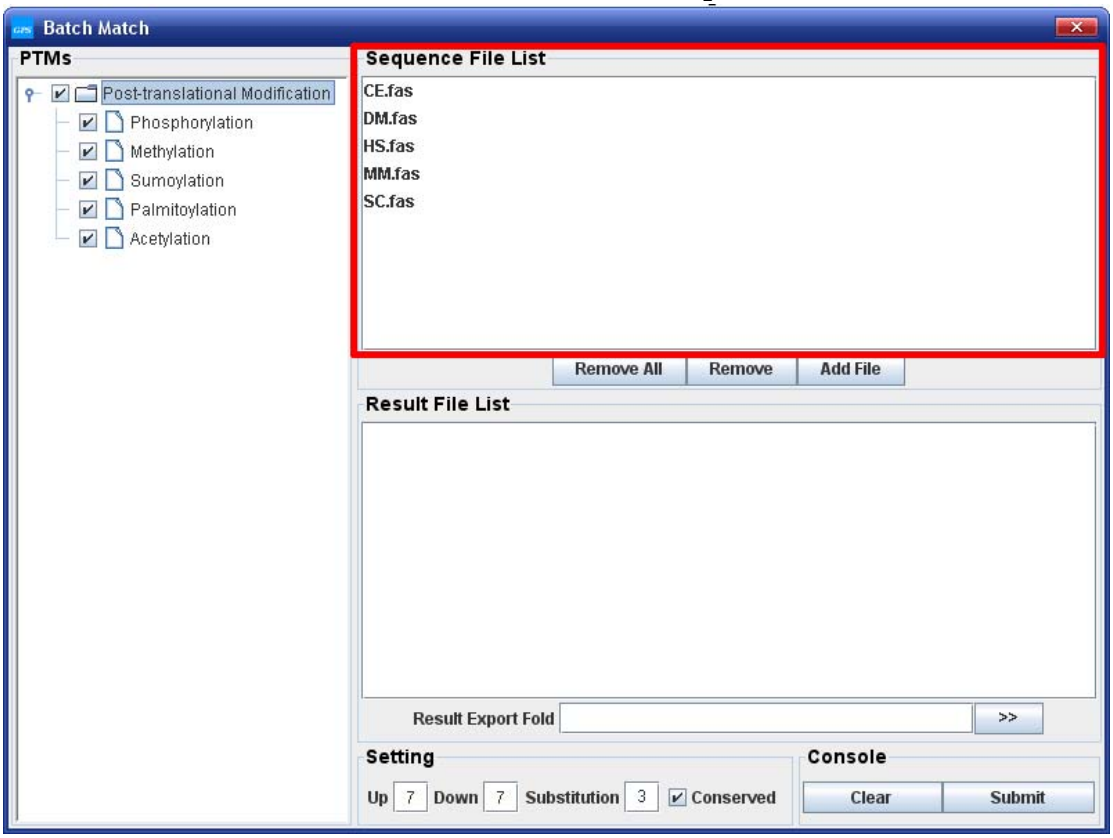
(1) Put protein sequences into one or several files (e.g., SC.fas, CE.fas, and etc) with FATSA format as below:

>protein1
XXXXXXXXXXXXX
XXXXXXXX
>protein2
XXXXXXXXXXXXXXXX…
>protein3
XXXXXXXXXXXX
...
Most importantly, the name of each protein should be presented.

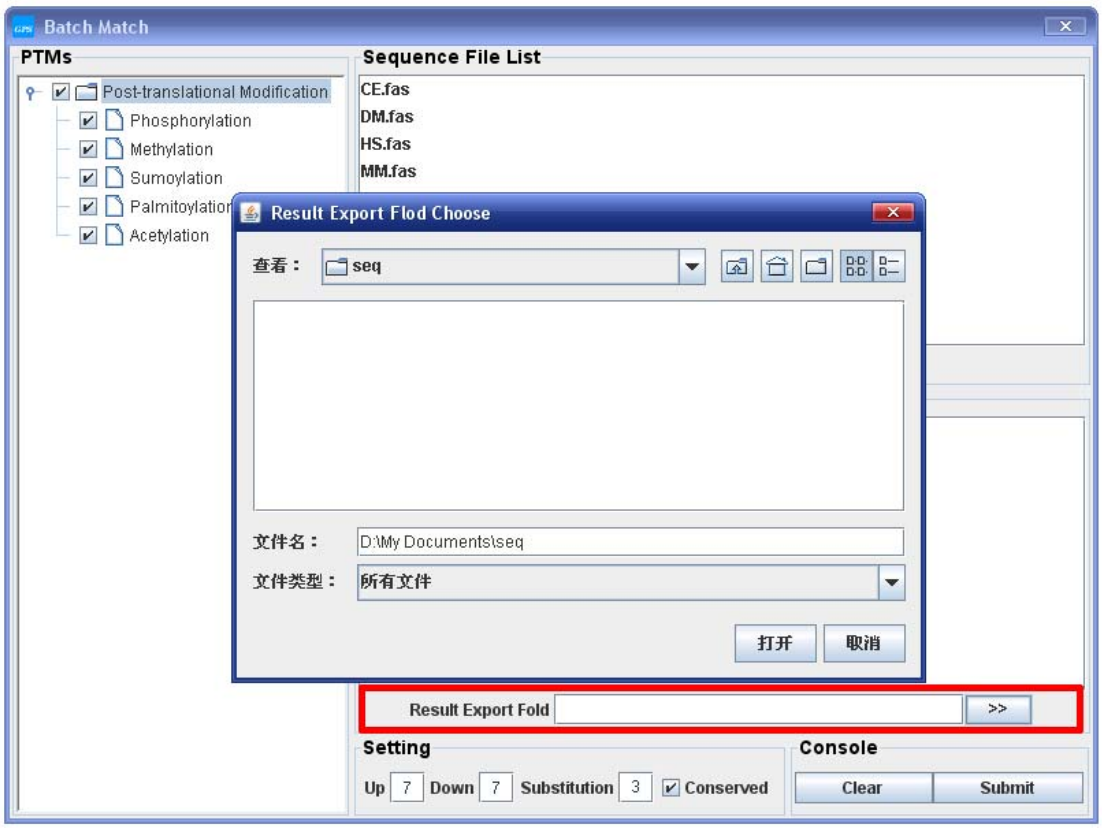
(2) Click on the **Add File** button and add one or more protein sequence files in your hard disk.



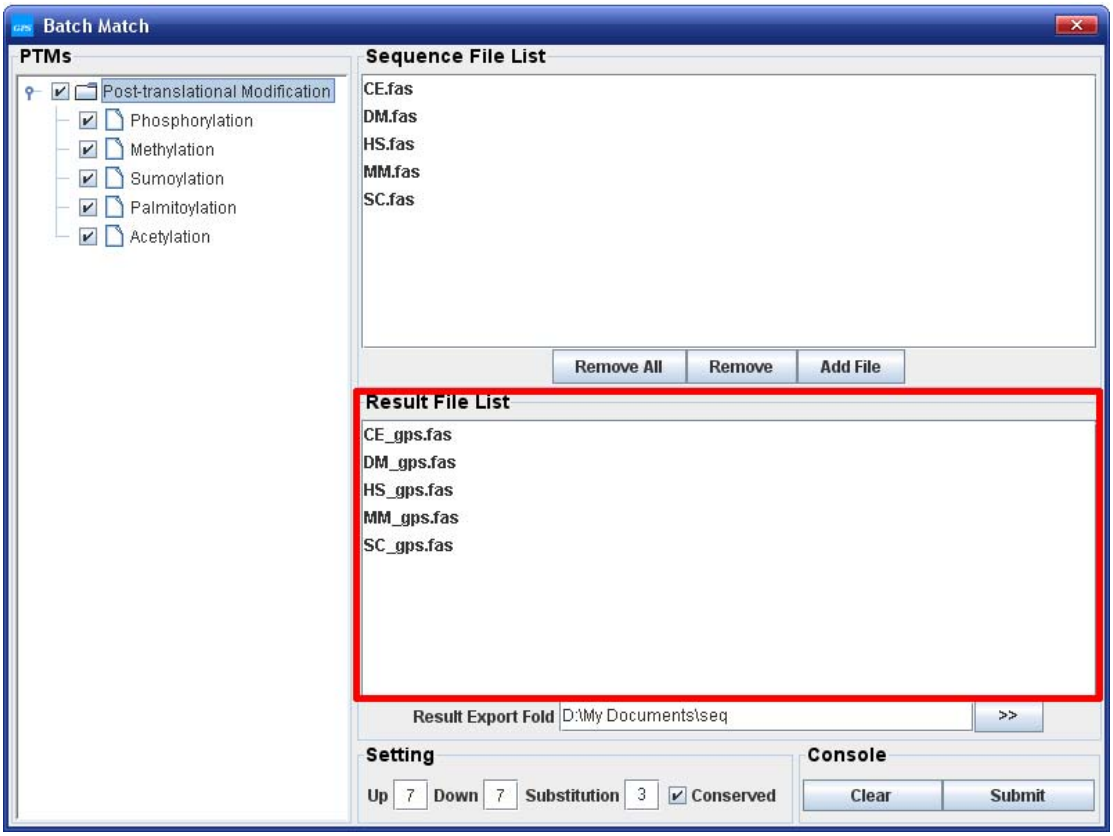
Then the names of added files will be shown in the **Sequence File List**.



(3) The output directory of prediction results should also be defined. Please click on the >> button to specify the export fold.



(4) Please choose one or more kinds of PTMs from **PTMs Tree** before matching. Then click on the **Submit** button and the **Batch Match Tool** begin to process all of the sequence files that have been added to the list. The result of prediction will be export to the **Result Export Fold,** and the name of result files will be shown in the **Result File List**.



References

1. Huber, S. C.; Hardin, S. C., Numerous posttranslational modifications provide opportunities for the intricate regulation of metabolic enzymes at multiple levels. Curr Opin Plant Biol 2004, 7, (3), 318-22.
2. Jensen, O. N., Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. Curr Opin Chem Biol 2004, 8, (1), 33-41.
3. Mann, M.; Jensen, O. N., Proteomic analysis of post-translational modifications. Nat Biotechnol 2003, 21, (3), 255-61.
4. Walsh, C. T.; Garneau-Tsodikova, S.; Gatto, G. J., Jr., Protein posttranslational modifications: the chemistry of proteome diversifications. Angew Chem Int Ed Engl 2005, 44, (45), 7342-72.
5. Walsh, G.; Jefferis, R., Post-translational modifications in the context of therapeutic proteins. Nat Biotechnol 2006, 24, (10), 1241-52.
6. Blom, N.; Gammeltoft, S.; Brunak, S., Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 1999, 294, (5), 1351-62.
7. 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, (6), 1633-49.
8. Rodriguez, M. S.; Dargemont, C.; Hay, R. T., SUMO-1 conjugation in vivo requires both a consensus modification motif and nuclear targeting. J Biol Chem 2001, 276, (16), 12654-9.
9. Zhou, F.; Xue, Y.; Yao, X.; Xu, Y., A general user interface for prediction servers of proteins' post-translational modification sites. Nat Protoc 2006, 1, (3), 1318-21.
10. Zhou, F. F.; Xue, Y.; Chen, G. L.; Yao, X., GPS: a novel group-based phosphorylation predicting and scoring method. Biochem Biophys Res Commun 2004, 325, (4), 1443-8.
11. Diella, F.; Cameron, S.; Gemund, C.; Linding, R.; Via, A.; Kuster, B.; Sicheritz-Ponten, T.; Blom, N.; Gibson, T. J., Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. BMC Bioinformatics 2004, 5, (1), 79.

12. 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, (Web Server issue), W184-7.
13. Xue, Y.; Ren, J.; Gao, X.; Jin, C.; Wen, L.; Yao, X., GPS 2.0: Prediction of kinase-specific phosphorylation sites in hierarchy. Mol Cell Proteomics 2008.
14. Xue, Y.; Zhou, F.; Fu, C.; Xu, Y.; Yao, X., SUMOsp: a web server for sumoylation site prediction. Nucleic Acids Res 2006, 34, (Web Server issue), W254-7.
15. Zhou, F.; Xue, Y.; Yao, X.; Xu, Y., CSS-Palm: palmitoylation site prediction with

a clustering and scoring strategy (CSS). Bioinformatics 2006, 22, (7), 894-6.
16. Ren, J.; Wen, L.; Gao, X.; Jin, C.; Xue, Y.; Yao, X., CSS-Palm 2.0: an updated software for palmitoylation sites prediction. Protein Eng Des Sel 2008.
17. Xue, Y.; Chen, H.; Jin, C.; Sun, Z.; Yao, X., NBA-Palm: prediction of palmitoylation site implemented in Naive Bayes algorithm. BMC Bioinformatics 2006, 7, 458.
18. Chen, H.; Xue, Y.; Huang, N.; Yao, X.; Sun, Z., MeMo: a web tool for prediction of protein methylation modifications. Nucleic Acids Res 2006, 34, (Web Server issue), W249-53.
19. Li, A.; Xue, Y.; Jin, C.; Wang, M.; Yao, X., Prediction of Nepsilon-acetylation on internal lysines implemented in Bayesian Discriminant Method. Biochem Biophys Res Commun 2006, 350, (4), 818-24.
20. Diella, F.; Gould, C. M.; Chica, C.; Via, A.; Gibson, T. J., Phospho.ELM: a database of phosphorylation sites--update 2008. Nucleic Acids Res 2008, 36, (Database issue), D240-4.
21. Boutet, E.; Lieberherr, D.; Tognolli, M.; Schneider, M.; Bairoch, A., UniProtKB/Swiss-Prot: The Manually Annotated Section of the UniProt KnowledgeBase. Methods Mol Biol 2007, 406, 89-112.
22. Li, W.; Godzik, A., Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006, 22, (13), 1658-9.
23. Kim, N.; Alekseyenko, A. V.; Roy, M.; Lee, C., The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. Nucleic Acids Res 2007, 35, (Database issue), D93-8.
24. Castrignano, T.; D'Antonio, M.; Anselmo, A.; Carrabino, D.; D'Onorio De Meo, A.; D'Erchia, A. M.; Licciulli, F.; Mangiulli, M.; Mignone, F.; Pavesi, G.; Picardi, E.; Riva, A.; Rizzi, R.; Bonizzoni, P.; Pesole, G., ASPicDB: a database resource for alternative splicing analysis. Bioinformatics 2008, 24, (10), 1300-4.
25. Kent, W. J., BLAT--the BLAST-like alignment tool. Genome Res 2002, 12, (4), 656-64.
26. Li, T.; Fan, K.; Wang, J.; Wang, W., Reduction of protein sequence complexity by residue grouping. Protein Eng 2003, 16, (5), 323-30.
27. Wei, C. J.; Xu, X.; Lo, C. W., Connexins and cell signaling in development and disease. Annu Rev Cell Dev Biol 2004, 20, 811-38.
28. Villen, J.; Beausoleil, S. A.; Gerber, S. A.; Gygi, S. P., Large-scale phosphorylation analysis of mouse liver. Proc Natl Acad Sci U S A 2007, 104, (5), 1488-93.
29. Locke, D.; Koreen, I. V.; Harris, A. L., Isoelectric points and post-translational modifications of connexin26 and connexin32. Faseb J 2006, 20, (8), 1221-3.
30. Xu, Q.; Lee, C., Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. Nucleic Acids Res 2003, 31, (19), 5635-43.
31. Hiller, M.; Huse, K.; Platzer, M.; Backofen, R., Creation and disruption of protein features by alternative splicing -- a novel mechanism to modulate function. Genome Biol 2005, 6, (7), R58.

32. Resch, A.; Xing, Y.; Modrek, B.; Gorlick, M.; Riley, R.; Lee, C., Assessing the impact of alternative splicing on domain interactions in the human proteome. J Proteome Res 2004, 3, (1), 76-83.

Release Note

1. Dec. 5th, 2008, the online service and the local stand-alone packages of PPS 1.0 were released.