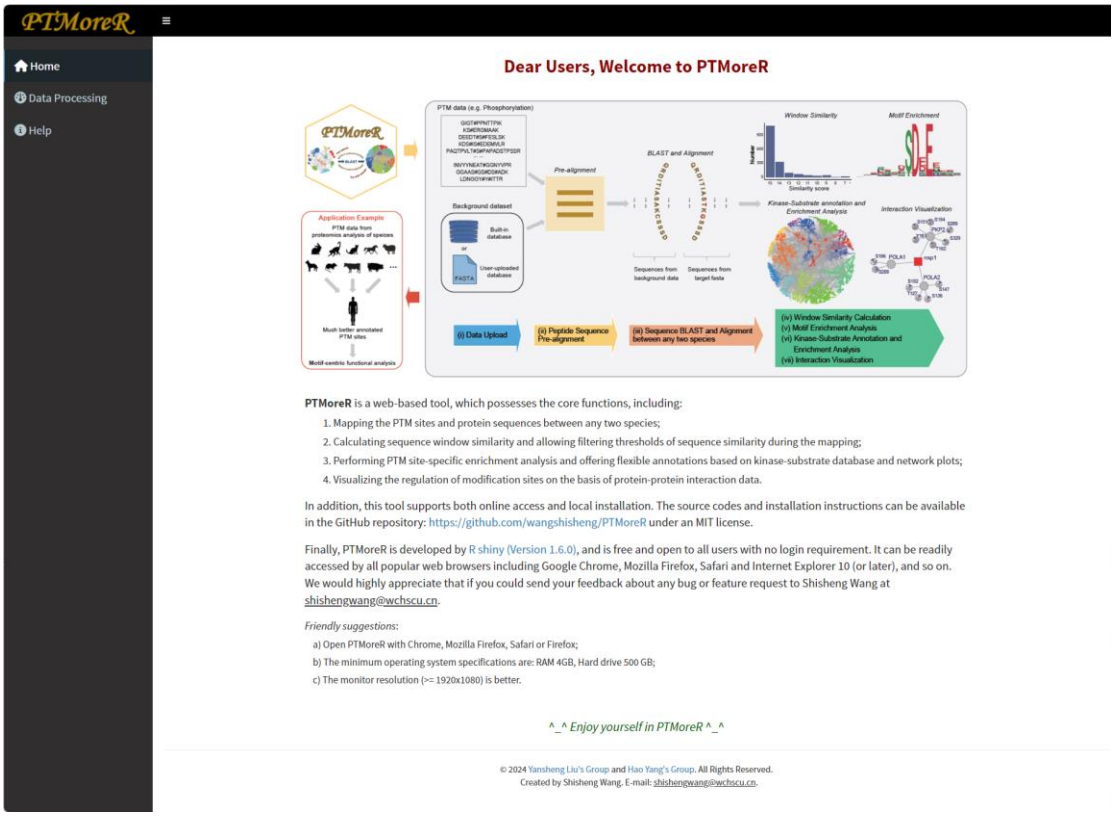
~Supplementary material~

PTMoreR: a motif-centric analysis enabling cross-species PTM mapping and comparative phosphoproteomics across mammals

Supplementary Notes

Overview *PTMoreR* (Post-translational modification ortholog aligner) is not merely a P-site BLAST tool; instead, it considers the surrounding amino acid sequence of PTM sites during BLAST, enabling a motif-centric analysis across species. Particularly, *PTMoreR* supports: 1. Mapping the PTM sites and protein sequences between any two species; 2. Calculating sequence window similarity and allowing filtering thresholds of sequence similarity during the mapping; 3. Processing PTM site-specific enrichment analysis and offering flexible annotation based on kinase-substrate database and network plots; 4. Visualizing the regulation of modification sites on the basis of protein-protein interaction data. Here we present the detailed introduction and operation of *PTMoreR*, by which users can follow to analyze their own data freely and conveniently.

Users can visit this site: https://yslproteomics.shinyapps.io/PTMoreR. The whole source codes: https://github.com/wangshisheng/PTMoreR. Then the website homepage can be shown like this:



1. Data Preparation

PTMoreR supports four basic file formats (.csv, .txt, .xlsx, .xls). Before analysis, users should prepare their peptide sequences with modification and protein background data. The modified peptide sequence data required here could be readily generated based on results of several popular tools such as MaxQuant (6), Spectronaut (7) and so on. The users then can upload the two data into *PTMoreR* with right formats respectively and start subsequent analysis.

1.1. Modified peptide sequences

1.1.1. Modified peptide sequences with normal type

Herein, the first row is the column name (e.g. AnnotatedPeps) and each of the other rows is a modified peptide sequence. Users need to mark those modified residues (e.g. S, T, Y with phosphorylation) with some label they like (such as '#' or '@') in advance. The peptide sequences can be like:

AnnotatedPeps
GIGT#PPNTTPIK
GIGT#PPNT#T#PIK
NGS#PEIK
KS#ERGMAAK
MNGHS#DEESVR
RQIDS#S#EDEDDEDYDNDKR
RYS#GS#DS#DS#ISER
RY#S#GS#DS#DS#ISER
RY#S#GS#DS#DS#IS#ER
KRPY#S#S#FS#NGK

In this situation, PTMoreR will search all proteins that these peptides belong to in the “Step 2. Pre-alignment”.

On the other hand, users could also prepare two columns: one column contains protein ids (i.e. UniProt IDs), the other column contains modified peptide sequences. Users need to mark those modified residues (e.g. S, T, Y with phosphorylation) with some label they like (such as '#' or '@') in advance, as below:

UniProt.ID	Pep.upload
D4A9J4	GIGT#PPNTTPIK
D4A9J4	GIGT#PPNT#T#PIK
D4A9J4	NGS#PEIK
D4A9J4	KS#ERGMAAK
D4AAG9	MNGHS#DEESVR
D4AAG9	RQIDS#S#EDEDDEDYDNDKR
D4AAG9	RYS#GS#DS#DS#ISER
D4AAG9	RY#S#GS#DS#DS#ISER
D4AAG9	RY#S#GS#DS#DS#IS#ER
D4AAG9	KRPY#S#S#FS#NGK
D4AAG9	AASSGPRS#PLDQR
D4AAG9	S#PYGS#RS#PFEHSAEHR
D4AAG9	S#PFEHSAEHR
D4AAG9	S#T#PEHT#WSSR
M0RBE8	VIHS#S#DEGEDQTGEDEEDDEWDD
A0A0G2K130	DEEDT#S#FESLSK
B5DF91	NMALLS#QLY#HS#PAR
B5DF91	NMALLSQLYHS#PAR

In this situation, PTMoreR will only pre-align peptide sequences to the proteins in the first column.

1.1.2. Modified peptide sequences from MaxQuant

If the sequence data are obtained from MaxQuant, then users can find the modified peptide sequences in the modification txt file, for example, the Phospho (STY)Sites.txt file in the output tables from MaxQuant. The peptide sequences are like this:

MaxQuant_Outputs
LFLDGEEEKEWAFEES(1)K
FDEGEDGEGS(0.996)NY(0.004)KKLC
ALVADEPEDLDT(1)EDEGLISFEEER
TYS(0.98)S(0.02)SGSSGGSHPSSR
ELILGS(0.002)ET(0.052)PS(0.779)S(0.167)PR
S(0.008)KS(0.992)PS(0.999)PPRLT(0.001)EDR
AAKLS(1)EGS(1)QPAEEEEDQETPSR
AAKLS(1)EGS(1)QPAEEEEDQETPSR
QEPT(1)QEHKQEEGQKQEEQEEEQEEEGK
NIGFKVNS(1)K

1.1.3. Modified peptide sequences from Spectronaut

If the sequence data are obtained from Spectronaut, then users can found the modified peptide sequences in the Standard Report part of Spectronaut, for example, export the Peptide Quant results from the Pivot Report and extract the modified peptide sequeces from the EG.ModifiedSequence column. The peptide sequences are like this:

Spectronaut_Outputs
INS[Phospho (STY)]APSS[Phospho (STY)]PIK
MLISAVS[Phospho (STY)]PEIR
KINS[Phospho (STY)]APSS[Phospho (STY)]PIK
KINS[Phospho (STY)]APS[Phospho (STY)]SPIK
INSAPSS[Phospho (STY)]PIK
EGSQGEPWT[Phospho (STY)]PTANLK
EGSQGEPWTPT[Phospho (STY)]ANLK
SHMSGS[Phospho (STY)]PGPGGSNTAPSTPVIGGSDKPGMEEK
SHMSGSPGPGGSNT[Phospho (STY)]APSTPVIGGSDKPGMEEK
SS[Phospho (STY)]SS[Phospho (STY)]LLASPGHISVK

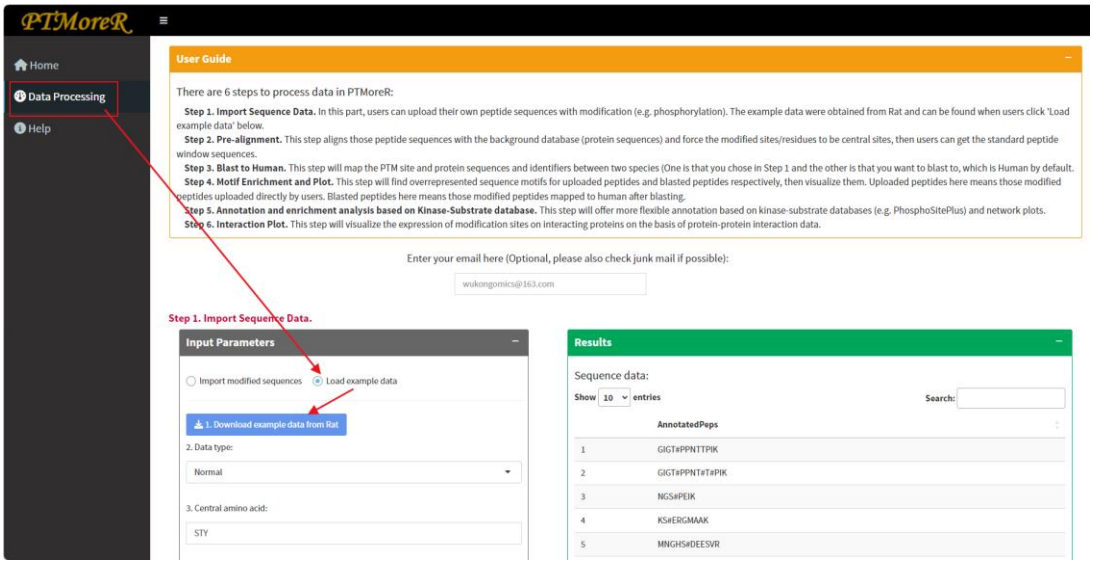
1.2. Background data

Background data here means the protein sequences of a species (.fasta format). Users should use the same protein sequence file as the background database. For example, users can download the protein sequences from UniProt (https://www.uniprot.org/) (3). The protein sequences like this:

```
>sp|Q64578|AT2A1_RAT Sarcoplasmic/endoplasmic reticulum calcium ATPase 1 OS=Rattus norvegicus OX=10116 GN=Atp2a1 PE=1 SV=1
MEAAHSKSTEECLSYFGVSETTGLTPDQVKRHLEKYGPNELPAEEGKSLWELVVEQFEDL
LVRILLLAACISFVLAWFEEGEETVTAFVEPFVILLILIANAIVGVWQERNAENAIEALK
EYEPEMGKVYRADRKSVQRIKARDIVPGDIVEVAVGDKVPADIRILSIKSTTLRVDQSIL
TGESVSVIKHTDPVPDPRAVNQDKKNMLFSGTNIAAGKAVGIVATTGVSTEIGKIRDQMA
ATEQDKTPLQQKLDEFGEQLSKVISLICVAVWLINIGHFNDPVHGGSWFRGAIYYFKIAV
ALAVAAIPEGLPAVITTCLALGTRRMAKKNAIVRSLPSVETLGCTSVICSDKTGTLTTNQ
MSVCKMFIIDKVDGDICSLNEFSITGSTYAPEGEVLKNDKPVRAGQYDGLVELATICALC
NDSSLDFNETKGVYEKVGEATETALTTLVEKMNVFNTEVRSLSKVERANACNSVIRQLMK
KEFTLEFSRDRKSMSVYCSPAKSSRAAVGNKMFVKGAPEGVIDRCNYVRVGTTRVPLTGP
VKEKIMSVIKEWGTGRDTLRCLALATRDTPPKREEMVLDDSAKFMEYEMDLTFVGVVGML
DPPRKEVTGSIQLCRDAGIRVIMITGDNKGTAIAICRRIGIFSENEEVADRAYTGREFDD
LPLAEQREACRRACCFARVEPSHKSKIVEYLQSYDEITAMTGDGVNDAPALKKAEIGIAM
GSGTAVAKTASEMVLADDNFSTIVAAVEEGRAIYNNMKQFIRYLISSNVGEVVCIFLTAA
LGLPEALIPVQLLWVNLVTDGLPATALGFNPPDLDIMDRPPRSPKEPLISGWLFFRYMAI
GGYVGAATVGAAAWWFLYAEDGPHVSYHQLTHFMQCTEHNPEFDGLDCEVFEAPEPMTMA
LSVLVTIEMCNALNSLSENQSLLRMPPWVNIWLLGSICLSMSLHFLILYVDPLPMIFKLR
ALDFTQWLMVLKISLPVIGLDELLKFIARNYLEG
>sp|Q64568|AT2B3_RAT Plasma membrane calcium-transporting ATPase 3 OS=Rattus norvegicus OX=10116 GN=Atp2b3 PE=1 SV=2
MGDMANSSIEFHPKPQQQREVPHVGGFGCTLAELRSLMELRGAEALQKIQEAYGDVSGLC
RRLKTSPTEGLADNTNDLEKRRQIYGQNFIPPKQPKTFLQLVWEALQDVTLIILEVAAIV
SLGLSFYAPPGEESEACGNVSGGAEDEGEAEAGWIEGAAILLSVICVVLVTAFNDWSKEK
QFRGLQSRIEQEQKFTVIRNGQLLQVPVAALVVGDIAQVKYGDLLPADGVLIQGNDLKID
ESSLTGESDHVRKSADKDPMLLSGTHVMEGSGRMVVTAVGVNSQTGIIFTLLGAGGEEEE
KKDKKGKQQDGAMESSQTKAKKQDGAVAMEMQPLKSAEGGEMEEREKKKANVPKKEKSVL
QGKLTKLAVQIGKAGLVMSAITVIILVLYFVIETFVVDGRVWLAECTPVYVQYFVKFFII
GVTVLVVAVPEGLPLAVTISLAYSVKKMMKDNNLVRHLDACETMGNATAICSDKTGTLTT
NRMTVVQSYLGDTHYKEIPAPSALTPKILDLLVHAISINSAYTTKILPPEKEGALPRQVG
NKTECALLGFILDLKRDFQPVREQIPEDQLYKVYTFNSVRKSMSTVIRMPDGGFRLFSKG
ASEILLKKCTNILNSNGELRGFRPRDRDDMVKKIIEPMACDGLRTICIAYRDFSAIQEPD
WDNENEVVGDLTCIAVVGIEDPVRPEVPEAIRKCQRAGITVRMVTGDNINTARAIAAKCG
IIQPGEDFLCLEGKEFNRRIRNEKGEIEQERLDKVWPKLRVLARSSPTDKHTLVKGIIDS
TTGEQRQVVAVTGDGTNDGPALKKADVGFAMGIAGTDVAKEASDIILTDDNFTSIVKAVM
WGRNVYDSISKFLQFQLTVNVVAVIVAFTGACITQDSPLKAVQMLWVNLIMDTFASLALA
TEPPTESLLLRKPYGRDKPLISRTMMKNILGHAVYQLTIIFTLLFVGELFFDIDSGRNAP
LHSPPSEHYTIIFNTFVMMQLFNEINARKIHGERNVFDGIFSNPIFCTIVLGTFGIQIVI
VQFGGKPFSCSPLSTEQWLWCLFVGVGELVWGQVIATIPTSQLKCLKEAGHGPGKDEMTD
EELAEGEEEIDHAERELRRGQILWFRGLNRIQTQMEVVSTFKRSGSFQGAVRRRSSVLSQ
LHDVTNLSTPTHIRVVKAFRSSLYEGLEKPESKSCIHNFMATPEFLINDYTHNIPLIDDT
DVDENEERLRAPPPPPPNQNNNAIDSGIYLTTHATKSATSSAFSSRPGSPLHSMETSL
```

1.3. Download example datasets

If users want to download the example datasets to their own computer and check the data format locally, they can download them from here:

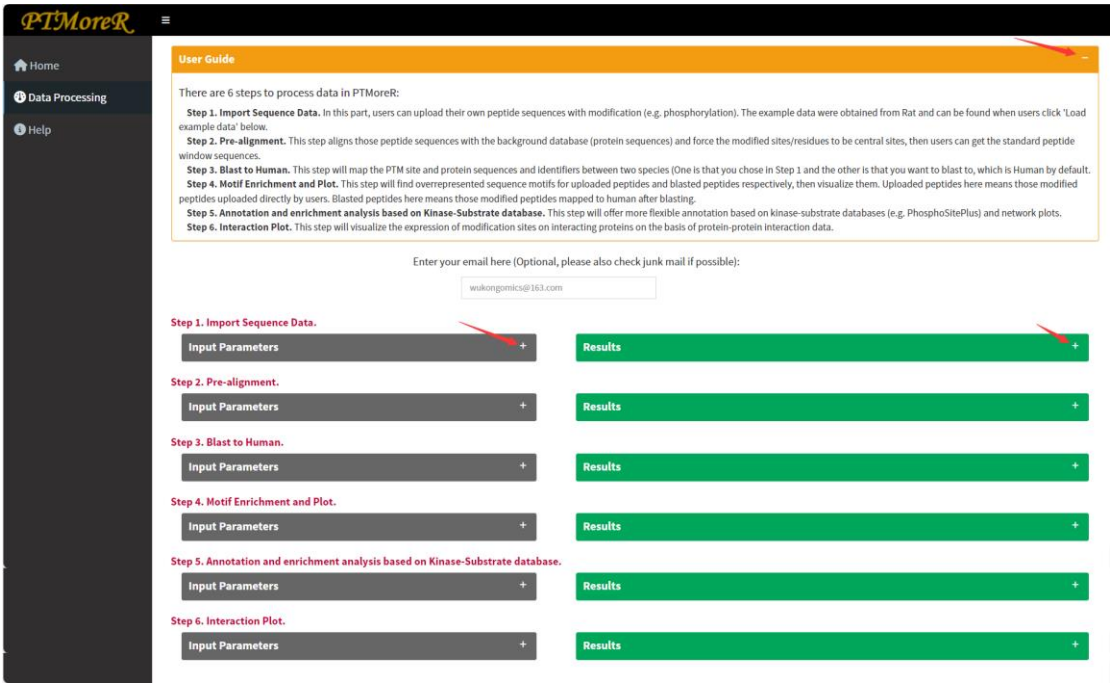


First, select "*Load example data*" and the example data will be shown on the right panel interactively. Second, users can download the example data (the modified peptide sequences) with relative format (*2. Data type*: Normal, MaxQuant, Spectronaut) by clicking the corresponding button. The data sets are saved as .csv format and users can open them in other software, such as Excel.

2. Data processing

Step 1. Import Sequence Data

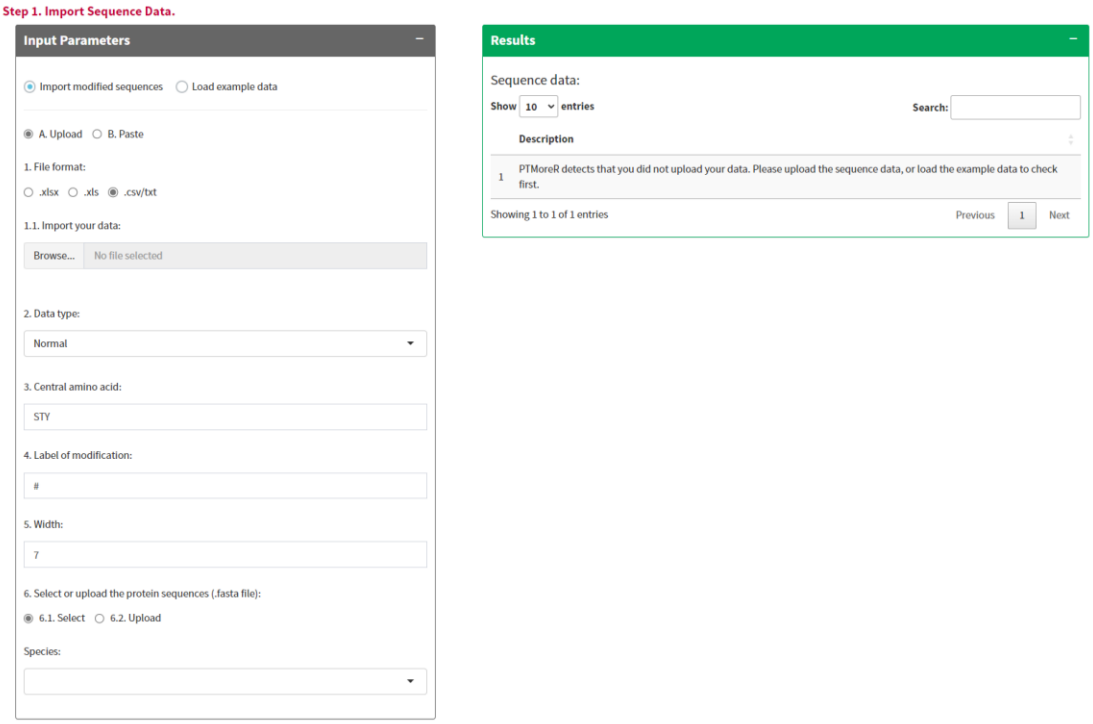
After preparing proper data, users could click “Data processing” part, shown as below. Firstly, users could find a brief user guide (detailed manual can be found in the Help part), which describes six basic steps in this tool. In addition, users can click the top-right corner (-/+) to collapse or expand the contents. Secondly, users can type in their e-mails and PTMoreR will send the blasted results to the e-mail. This is optional. And if users run this tool locally, this function is removed. Thirdly, users can process their data step by step. There are two main panels: right, *parameters panel*, users can adjust parameters here; left, *results panel*, many results after users set the parameters will be shown here and users can also download these results. users can also click the top-right corner (-/+) to collapse or expand the contents.



In the first step, users should upload data here or load the example data with the above data formats. By default, we use the example data to show results of every step.

Step 1.1. Uploading data.

When users prepare their data (the modified peptide sequences and protein background data), they can upload these data from here:

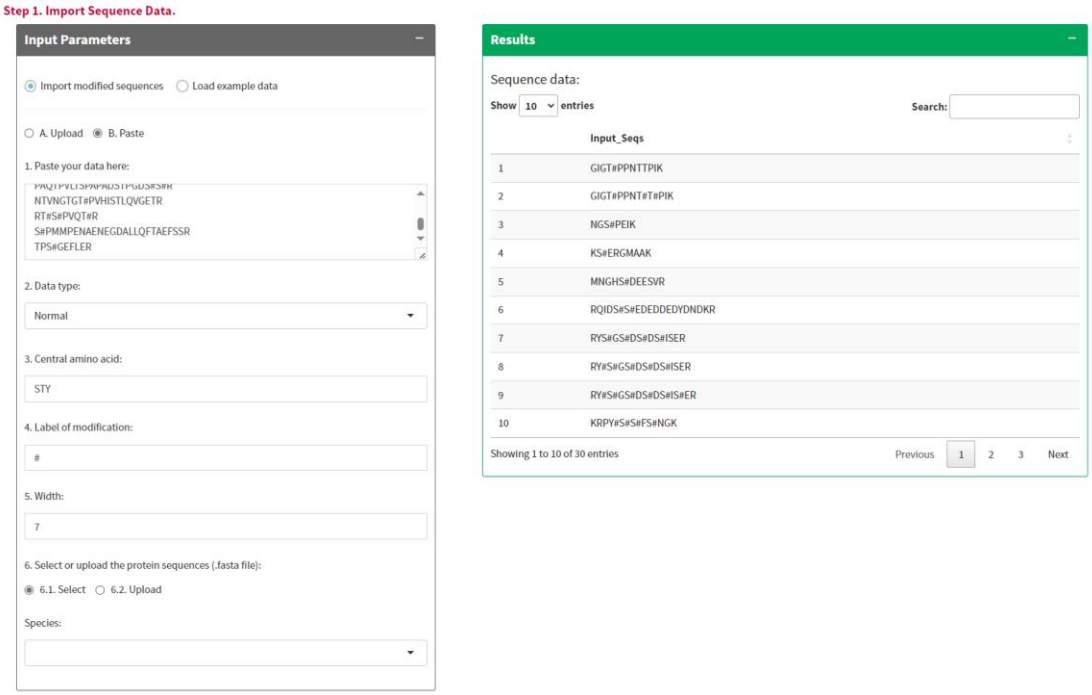


In the *parameters panel* of "*Step 1: Import Sequence Data*", there are two choices:

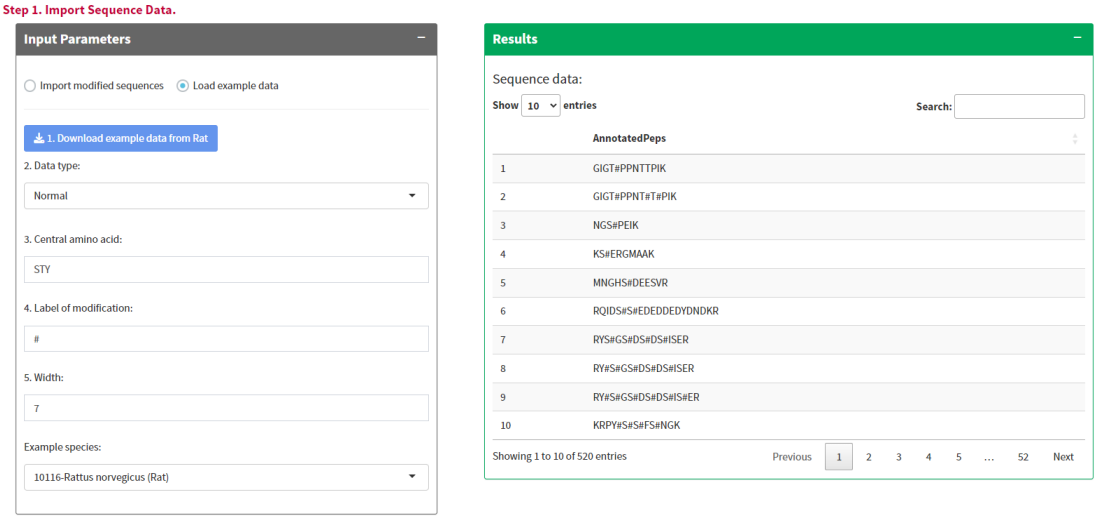
a. Import modified sequences. When users choose this option, they can choose "*A. Upload*" to upload their own data (the modified peptide sequences) here. Users should select the right format (*1. File format*: .csv, .txt, .xlsx, .xls) based on their data and then click "Browse" button (*1.1. Import your data*) to import the data;

In the *results panel* of "*Step 1: Import Sequence Data*", if users don't upload their data, here will show "*PTMoreR detects that you did not upload your data. Please upload the sequence data, or load the example data to check first.*" to warn users.

Users can also directly paste their sequences into *PTMoreR* by choosing "*B. Paste*" like below:

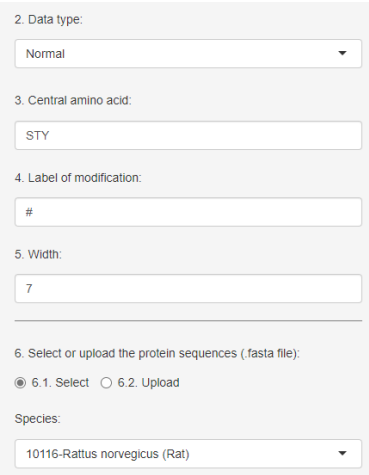


b. Load example data. As described in part 1.3 above, users can choose this option and download the example data to check them locally.



Step 1.2. Parameters

There are some basic parameters that users can change based on their own data, shown as below:



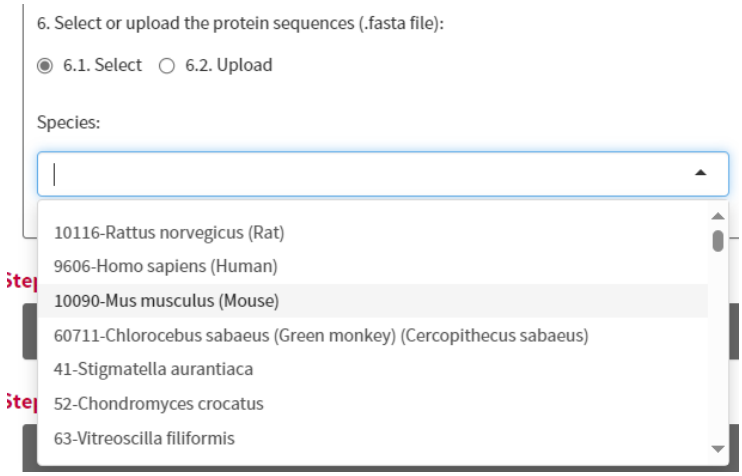
2. Data type: The original post-translational modification data obtained from which kind of search software. If you have processed the PTM data with standard format (e.g. NPT#Y#GSWFTEK), you should choose the "Normal", otherwise, if your PTM data are obtained from MaxQuant or Spectronaut, you should choose the relative type. Shown also as part 1.3 above.

3. Central amino acid: The central residue that users want to analyze, for example, phosphorylation motif analysis, can center on phosphorylated S, T or Y residues. If they want to analyze multi motif sites, here should be “STY”.

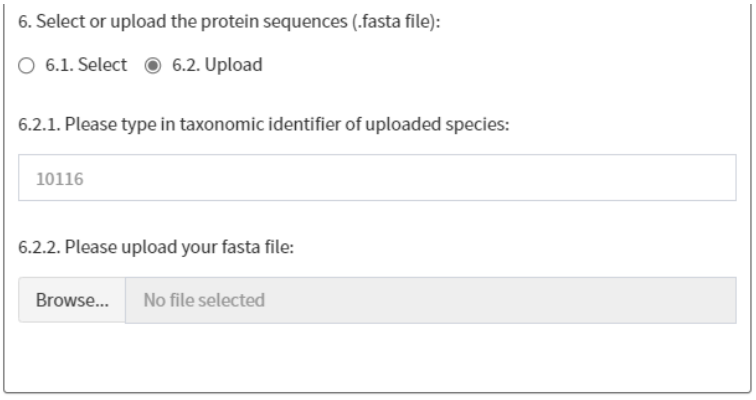
4. Label of modification: The label represents modification, users can use some label they like, such as “#”, “@”, in which “#” is recommended. Here is an example:

5. Width: It is the number of left/right side characters of the central residue.

6. Select or upload the protein sequences (.fasta file): If users want to use the default database, they just select relative species. By default, PTMoreR integrates 27074 species and download automatically the protein sequences (.fasta format) from UniProt database (https://www.uniprot.org).



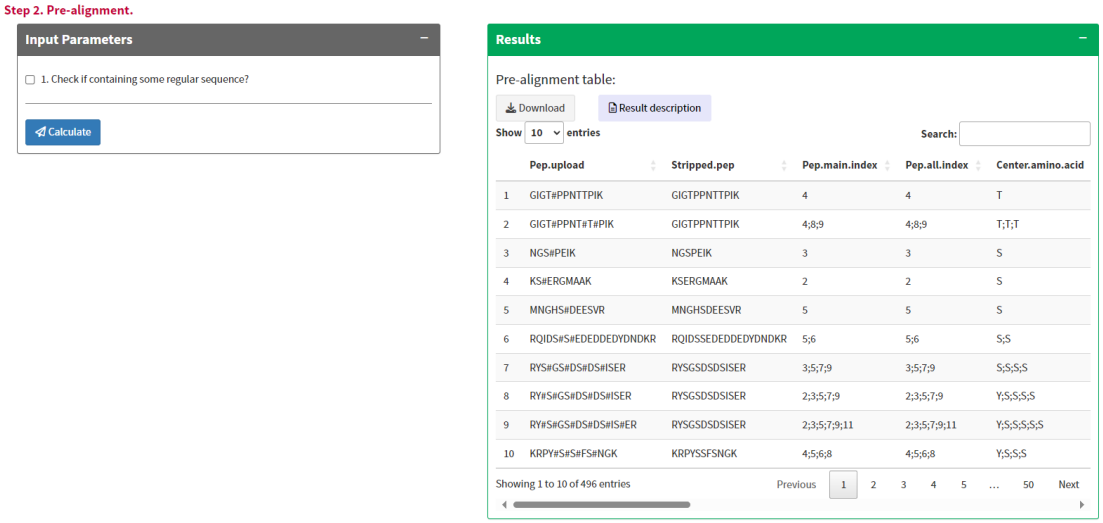
Optionally, if users want to upload their own .fasta file (choose *6.2. Upload*), they should type in the taxonomic identifier in *"6.2.1. Please type in taxonomic identifier of uploaded species"*, for example, rat's taxonomic identifier is 10116.



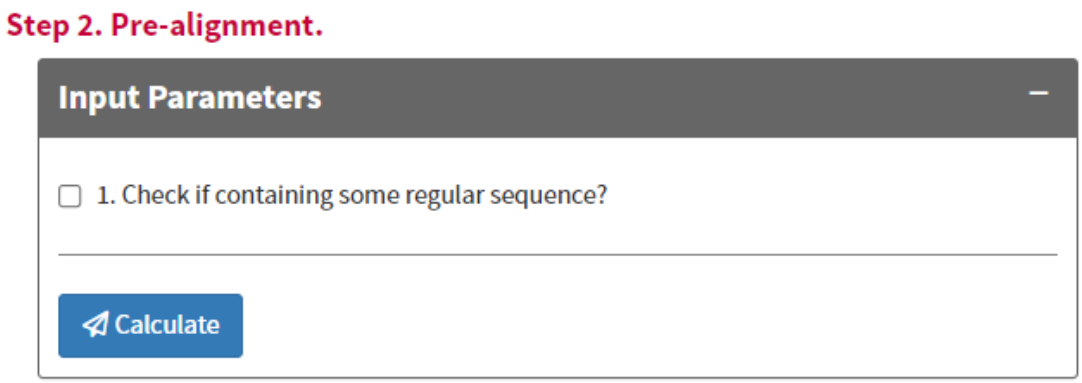
After typing in the taxonomic identifier, users can also upload their own .fasta file (*6.2.2. Please upload your fasta file*), no species limits here, but the calculation time would be longer.

Step 2. Pre-alignment

This step means align those peptide sequences with the background database (protein sequences) and force the modified sites/residues to be central sites, then users can get the standard peptide window sequences.

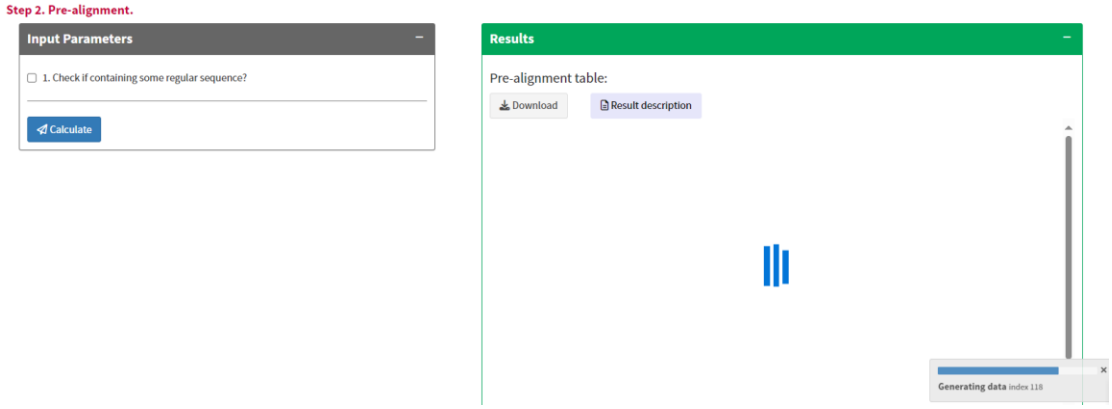


Step 2.1. Parameters



1. Check if containing some regular sequence: if users want to check whether the aligned peptides contain some specific sequences, for example, you want to find those peptides whose 3th and 5th position are R (arginine), then you can select this parameter and type in a simple regular expression, like "^\w{2}R\w{1}R" (more details can be found here: https://en.wikipedia.org/wiki/Regular_expression). Otherwise, you just unselect it.

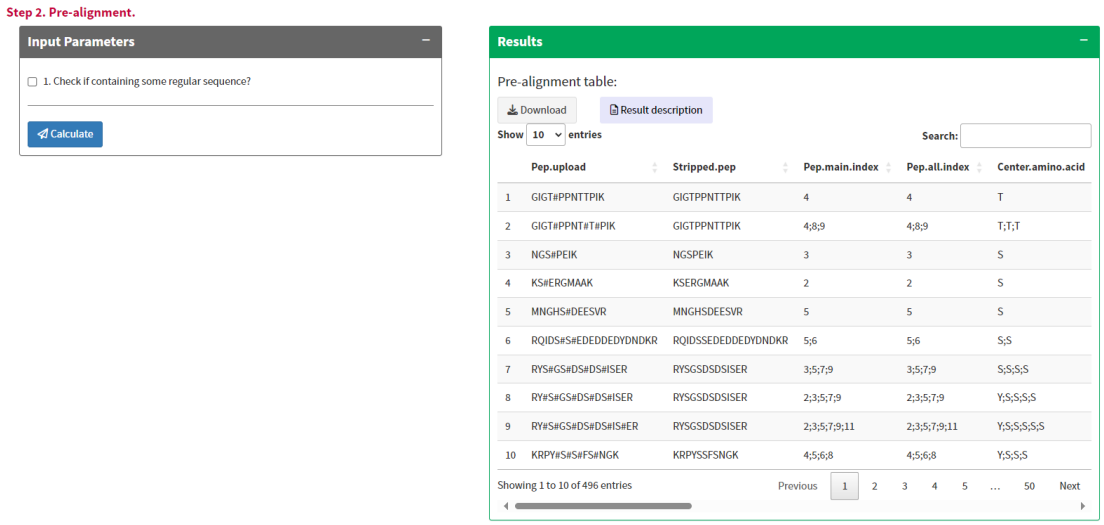
Then, you can click the “Calculate” button, it should be like this:



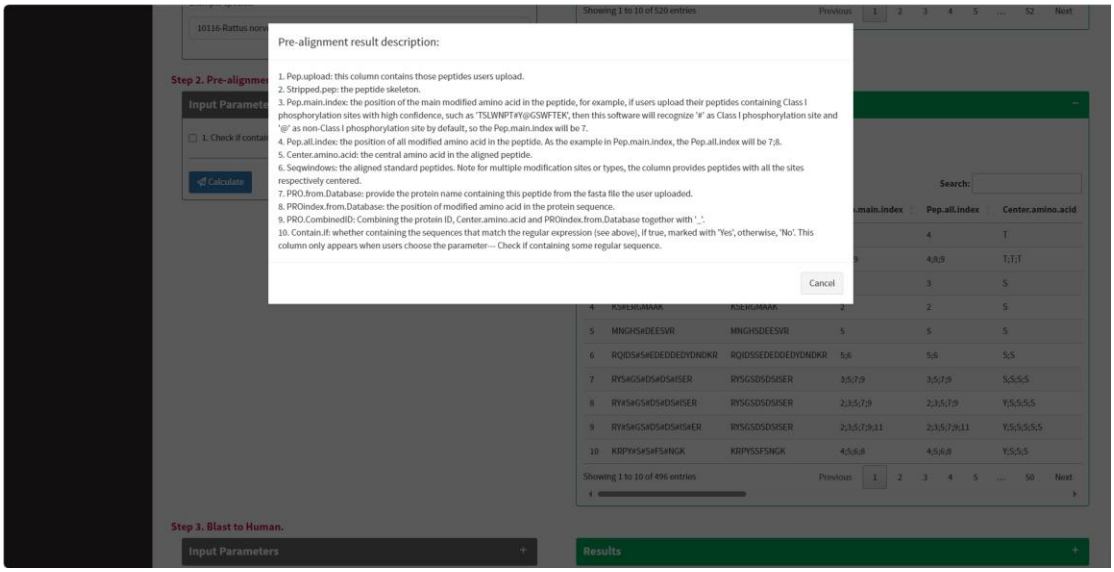
A process bar will appear in the bottom right corner to tell users where it goes.

Step 2.2 Results

Step 2.2.1. Alignment results



Users can click “Result description” button to read the introduction about every column, shown as below:



1. Pep.upload: This column contains those peptides users upload.

2. Stripped.pep: The peptide skeleton.

3. Pep.main.index: The position of the main modified amino acid in the peptide, for example, if users upload their peptides containing Class I phosphorylation sites with high confidence, such as 'TSLWNPT#Y@GSWFTEK', then this software will recognize '#' as Class I phosphorylation site and '@' as non-Class I phosphorylation site by default, so the Pep.main.index will be 7.

4. Pep.all.index: the position of all modified amino acid in the peptide. As the example in Pep.main.index, the Pep.all.index will be 7;8.

5. Center.amino.acid: the central amino acid in the aligned peptide.

6. Seqwindows: the aligned standard peptides. Note for multiple modification sites or types, the column provides peptides with all the sites respectively centered.

7. PRO.from.Database: provide the protein name containing this peptide from the fasta file the user uploaded.

8. PROindex.from.Database: the position of modified amino acid in the protein sequence.

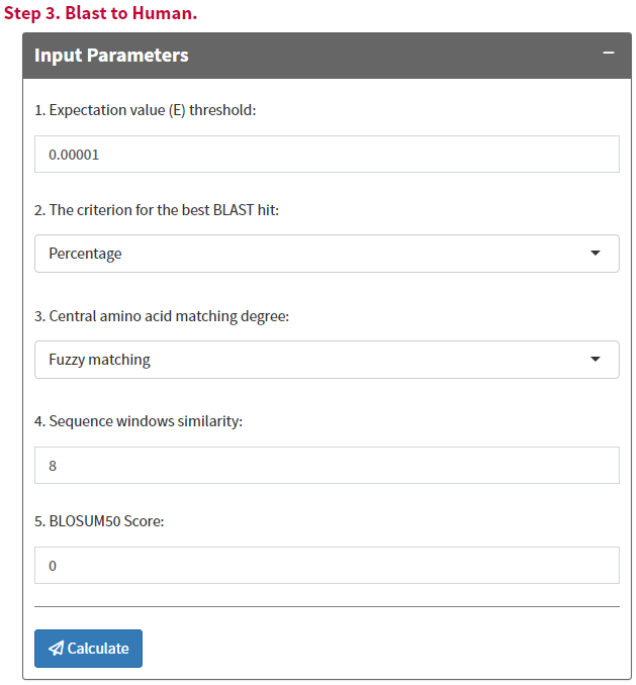
9. PRO.CombinedID: Combining the protein ID, Center.amino.acid and PROindex.from.Database together with '_'.

10. Contain.if: whether containing the sequences that match the regular expression (see above), if true, marked with 'Yes', otherwise, 'No'. This column only appears when users choose the parameter--- Check if containing some regular sequence.

Step 3. Blast to Human

This step will map the PTM site and protein sequences and identifiers between the uploaded species and the other species.

Step 3.1. Parameters



1. Expectation value (E) threshold: Expectation value (E) threshold for saving hits.

2. The criterion for best last hit: This tool performs a BLAST search between query and subject sequences and returns only the best hit based on the selected criterion. “Percentage” means If e-values are identical then the hit with the largest matching percentage is chosen. “Longest alignment length” means If e-values are identical then the hit with the longest alignment length is chosen.

3. Central amino acid matching degree: The matching degree of central amino acids (CAAs) when the uploaded peptides are blasted to Human protein sequences. 1. Exact matching: The CAAs are same, for example, the CAA is "S" in the uploaded peptides and the CAA is also "S" in the blasted sequence. 2. Fuzzy matching: Only for phosphorylation, not for other modification type. For example, the CAA is "S" in the uploaded peptides and the CAA could be "S", "T", or "Y" in the blasted sequence. All: Reporting all results.

4. Sequence windows similarity: The similarity of sequence windows between the uploaded peptides and the blasted peptides. For example, there are 15 amino acids in one sequence window, 8 here means there are 8 amino acids are exactly same (amino acids names and positions in both windows are all the same).

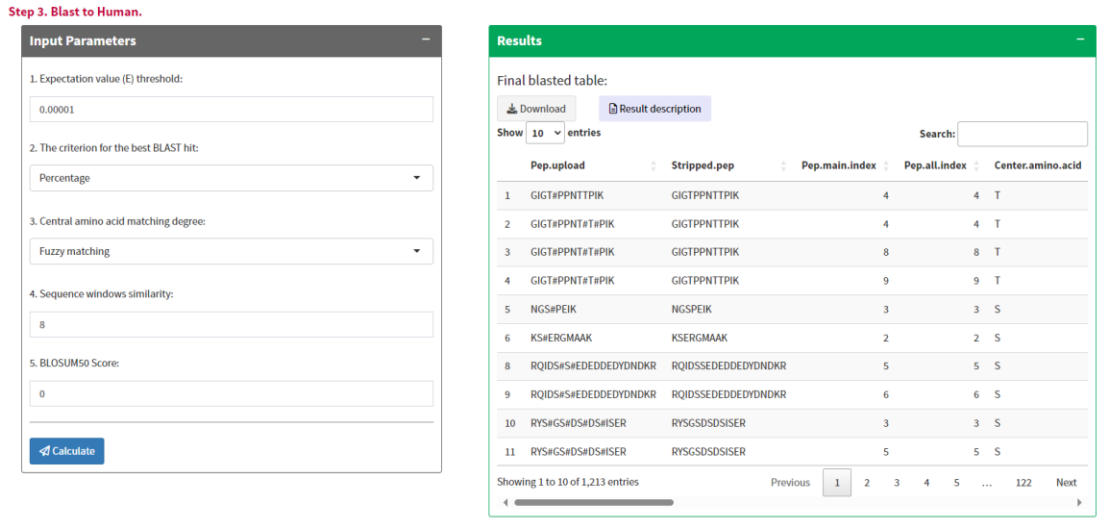
5. BLOSUM50 Score: BLOSUM50 means that the matrix was built using blocks of aligned sequences that had no more than 50% identity, which is used to score alignments between evolutionarily divergent protein sequences. The default BLOSUM50 Score is 0, which means PTMoreR filters the blasted results with score >= 0.

Step 3.2. Results

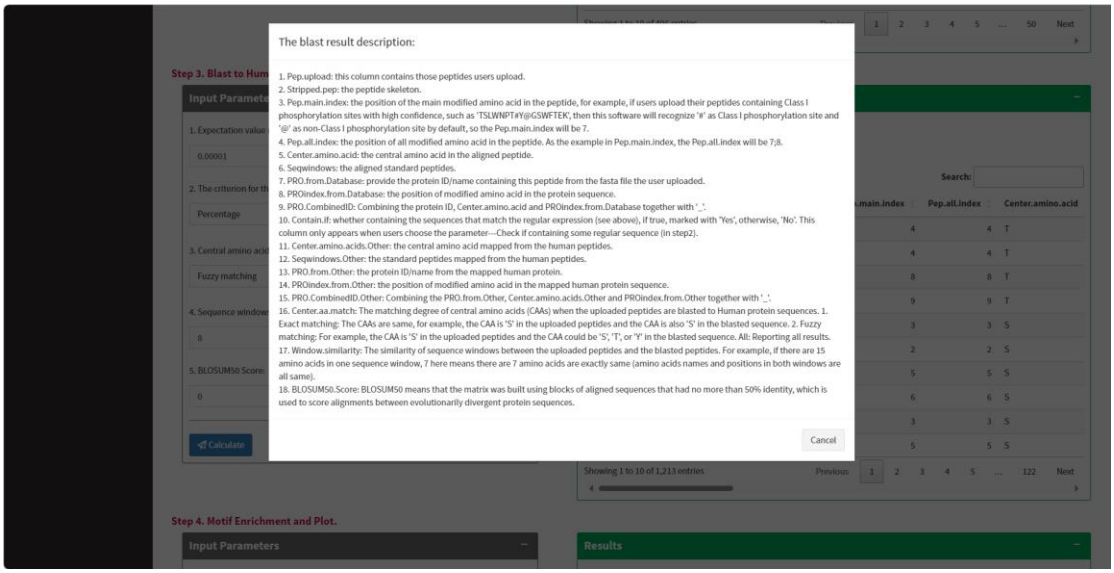
After setting proper parameters, users can click “Calculate” button and the results will be shown in the right panel.

Step 3.2.3. Final blast results

When users click “3. Final blast results” here, this tool will calculate the final result, like below:



Users can click “Result description” button to read the introduction about every column, shown as below:



1. Pep.upload: this column contains those peptides users upload.

2. Stripped.pep: the peptide skeleton.

3. Pep.main.index: the position of the main modified amino acid in the peptide, for example, if users upload their peptides containing Class I phosphorylation sites with high confidence, such as 'TSLWNPT#Y@GSWFTEK', then this software will recognize '#' as Class I phosphorylation site and '@' as non-Class I phosphorylation site by default, so the Pep.main.index will be 7.

4. Pep.all.index: the position of all modified amino acid in the peptide. As the example in

Pep.main.index, the Pep.all.index will be 7;8.

5. Center.amino.acid: the central amino acid in the aligned peptide.

6. Seqwindows: the aligned standard peptides.

7. PRO.from.Database: provide the protein ID/name containing this peptide from the fasta file the user uploaded.

8. PROindex.from.Database: the position of modified amino acid in the protein sequence.

9. PRO.CombinedID: Combining the protein ID, Center.amino.acid and PROindex.from.Database together with '_'.

10. Contain.if: whether containing the sequences that match the regular expression (see above), if true, marked with 'Yes', otherwise, 'No'. This column only appears when users choose the parameter---Check if containing some regular sequence (in step2).

11. Center.amino.acids.Other: the central amino acid mapped from the human peptides.

12. Seqwindows.Other: the standard peptides mapped from the human peptides.

13. PRO.from.Other: the protein ID/name from the mapped human protein.

14. PROindex.from.Other: the position of modified amino acid in the mapped human protein sequence.

15. PRO.CombinedID.Other: Combining the PRO.from.Other, Center.amino.acids.Other and PROindex.from.Other together with '_'.

16. Center.aa.match: The matching degree of central amino acids (CAAs) when the uploaded peptides are blasted to Human protein sequences. 1. Exact matching: The CAAs are same, for example, the CAA is 'S' in the uploaded peptides and the CAA is also 'S' in the blasted sequence. 2. Fuzzy matching: For example, the CAA is 'S' in the uploaded peptides and the CAA could be 'S', 'T', or 'Y' in the blasted sequence. All: Reporting all results.

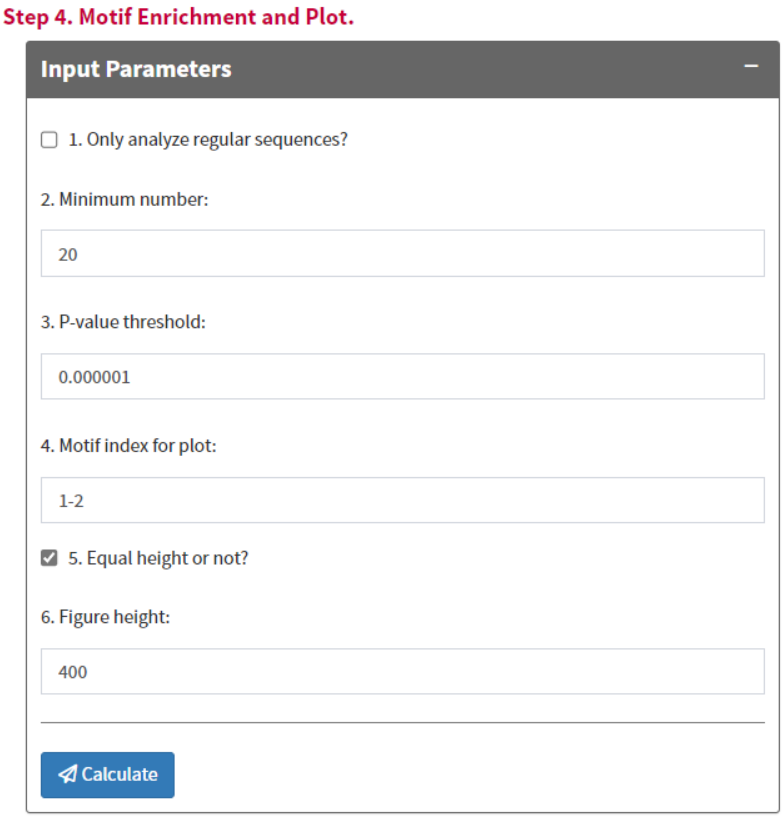
17. Window.similarity: The similarity of sequence windows between the uploaded peptides and the blasted peptides. For example, if there are 15 amino acids in one sequence window, 7 here means there are 7 amino acids are exactly same (amino acids names and positions in both windows are all same).

18. BLOSUM50.Score: BLOSUM50 means that the matrix was built using blocks of aligned sequences that had no more than 50% identity, which is used to score alignments between evolutionarily divergent protein sequences.

Step 4. Motif Enrichment and Plot

This step will find overrepresented sequence motifs as we previous described for single species (8).

Step 4.1. Parameters



1. Only analyze regular sequences? If true, this tool will only analyze the peptides contain some specific sequences (regular sequences) obtain from “3. Pre-alignment” part above.

2. Minimum number: This threshold refers to the minimum number of times you wish each of your extracted motifs to occur in the data set.

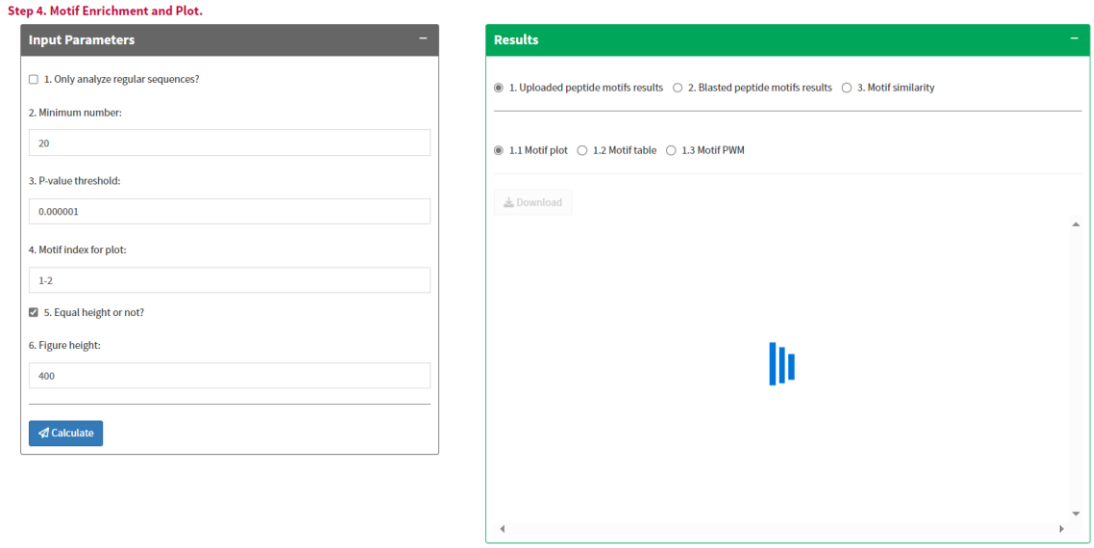
3. P-value threshold: The p-value threshold for the binomial probability. This is used for the selection of significant residue/position in the motif.

4. Motif index for plot: Which motif would be plotted. If users only type in one number, it will plot the relative motif. If users type in "1-10", it will plot the 1th to 10th motifs.

5. Equal height or not? Whether all residues in the figure have equal height. Default is false.

6. Figure height: The height of the figure.

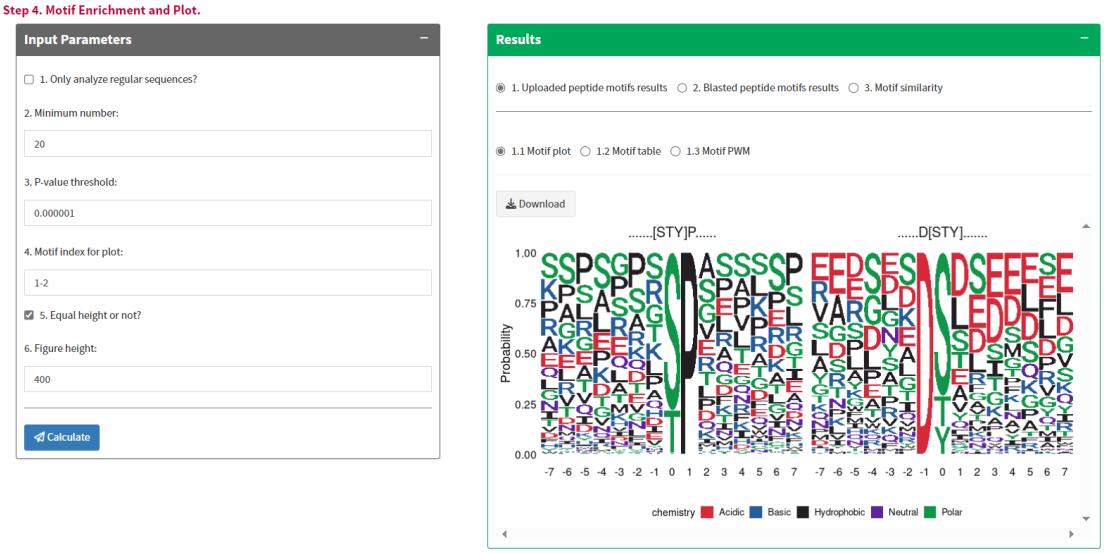
Then, users can click the “Calculate” button, this tool will process motif enrichment and plot those motifs:



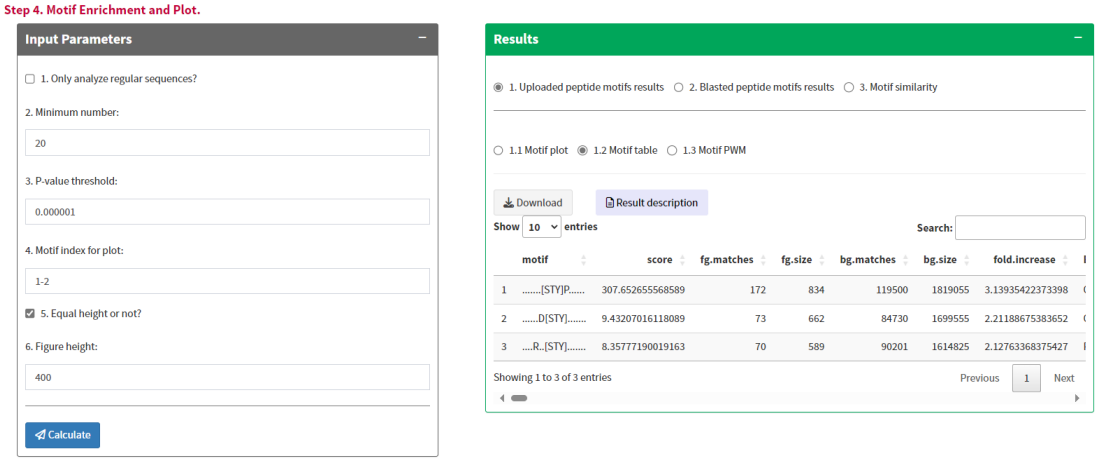
Step 4.2. Results

Step 4.2.1. Uploaded peptide motifs results

All results here are obtained from the uploaded peptides. When users click “1.1. Motif plot”, the motif plot is shown as below:



When users click “1.2. Motif table”, the motif table is shown as below:



Every column in this table means (Users can also check this by clicking “Result description” button):

1. motif: the overrepresented motif.

2. score: the motif score, which is calculated by taking the sum of the negative log probabilities used to fix each position of the motif. Higher motif scores typically correspond to motifs that are more statistically significant as well as more specific.

3. fg.matches: frequency of sequences matching this motif in the foreground set.

4. fg.size: total number of foreground sequences.

5. bg.matches: frequency of sequences matching this motif in the background set.

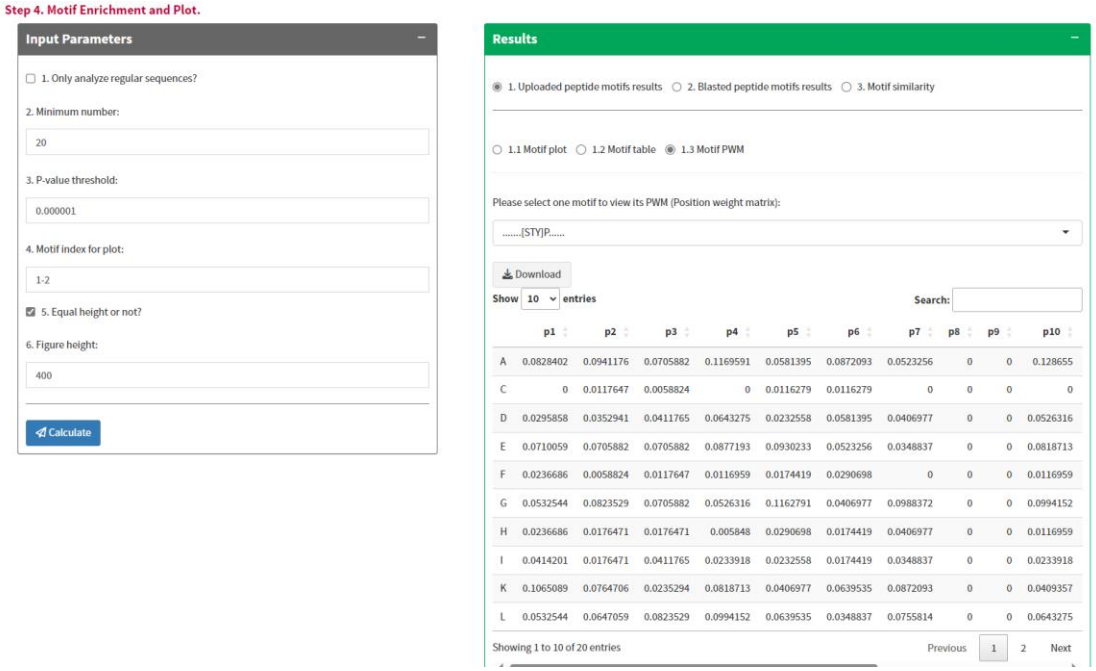
6. bg.size: total number of background sequences.

7. fold.increase: An indicator of the enrichment level of the extracted motifs. Specifically, it is calculated as (foreground matches/foreground size)/(background matches/background size).

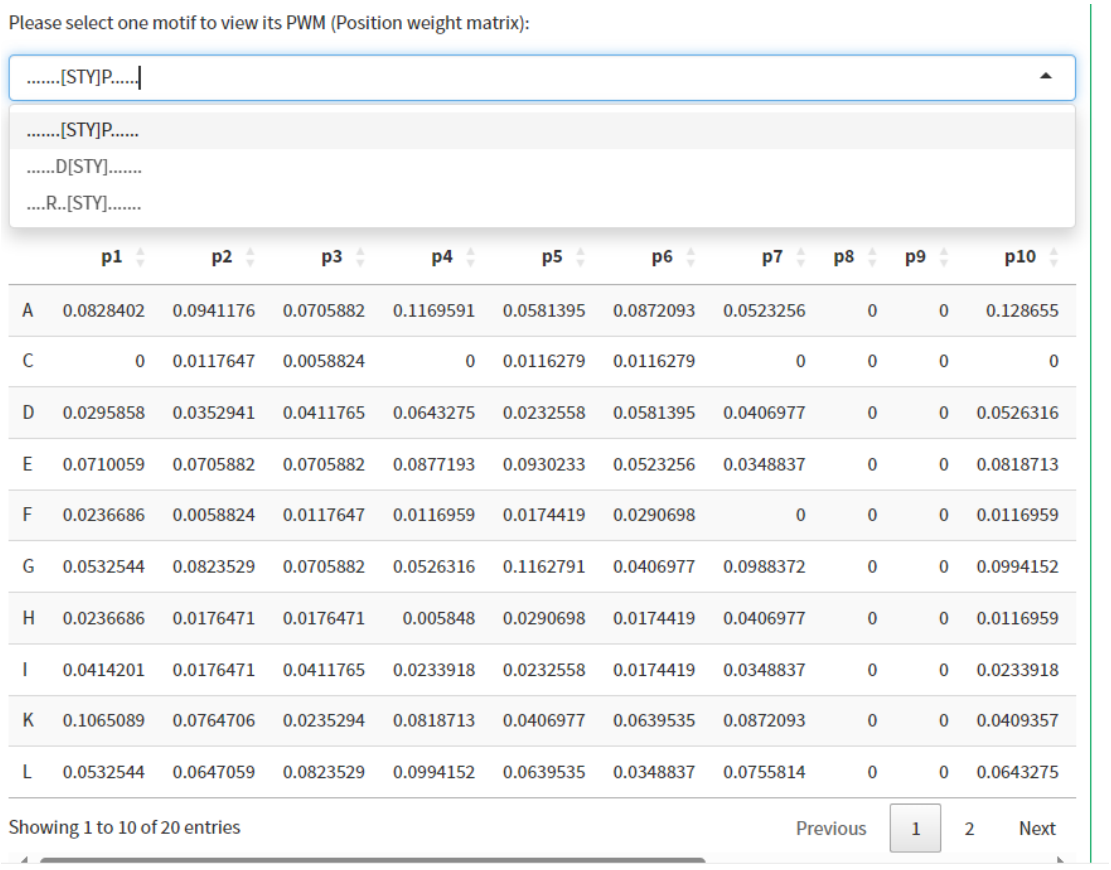
8. Enrich.seq: those peptides are overrepresented in this motif.

9. Enrich.pro: those proteins in which the peptides exist from Enrich.seq.

When users click “1.3. Motif PWM”, PWM here means position weight matrix, the motif table is shown as below:

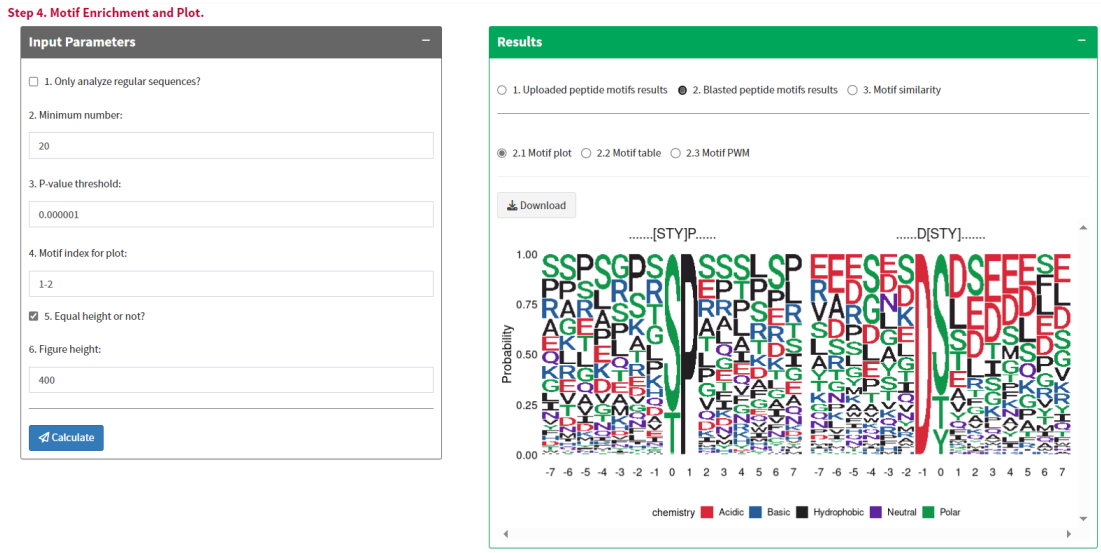


Users can select one motif to view its PWM. The motif here are obtained from “1.2. Motif table”, like below:

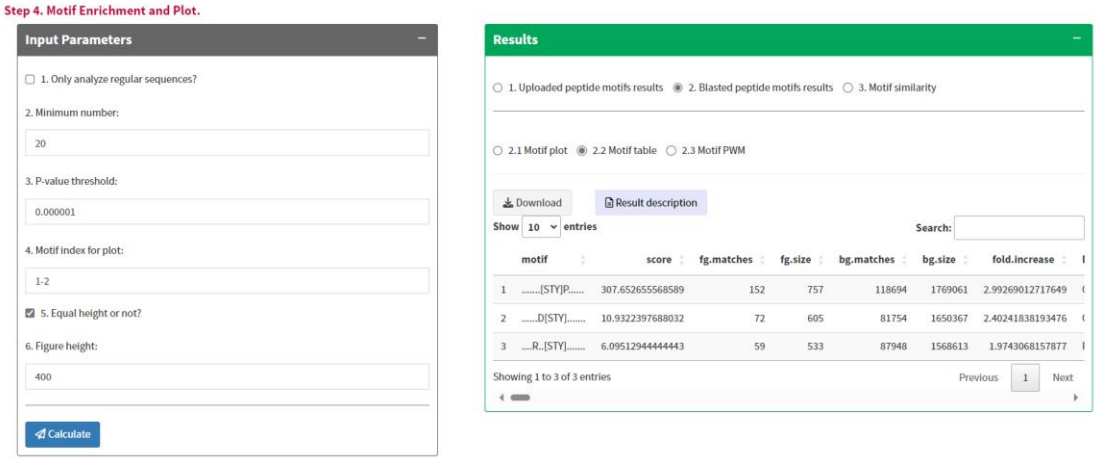


Step 4.2.2. Blasted peptide motifs results

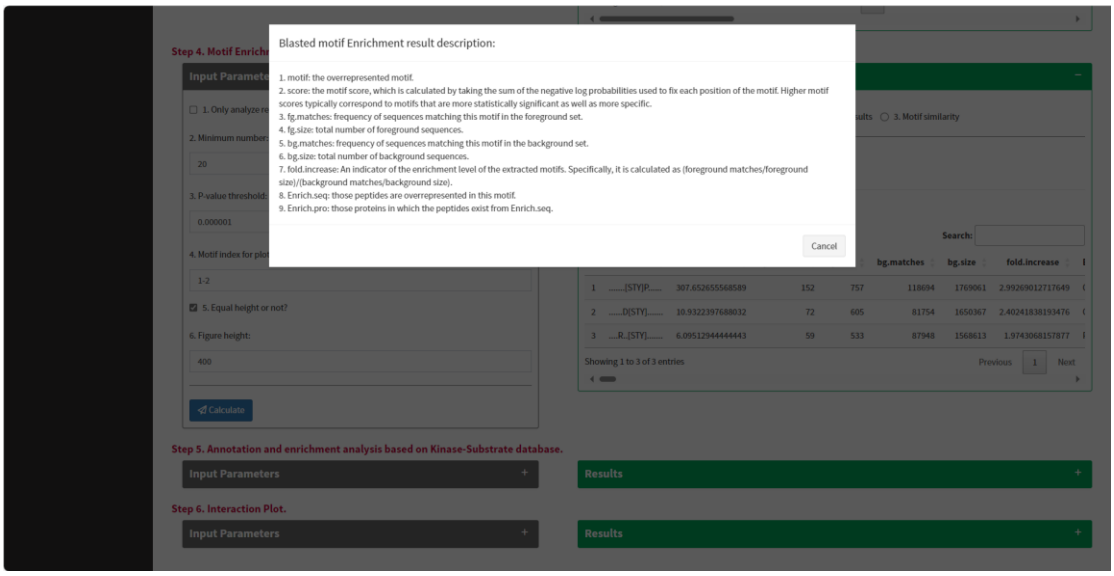
All results here are obtained from the blasted peptides. Similar to the above, when users click "2.1. Motif plot", the motif plot is shown as below:



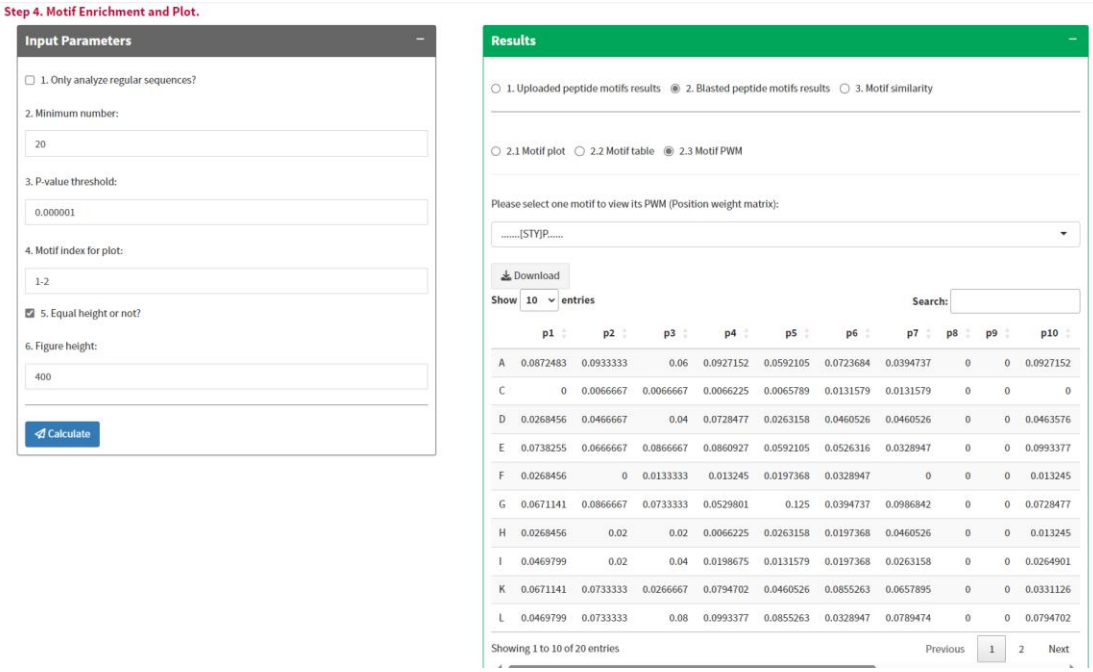
When users click "2.2. Motif table", the motif table is shown as below:



Every column means the same thing as above (Users can also check this by clicking “Result description” button).



Just as above, when users click “2.3. Motif PWM”, PWM here means position weight matrix, the motif table is shown as below:



Users can select one motif to view its PWM. The motif here are obtained from “2.2. Motif table”, like below:

Please select one motif to view its PWM (Position weight matrix):

.....[STY]P.....

.....[STY]P.....

.....D[STY].....

....R..[STY].....

	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10
A	0.0872483	0.0933333	0.06	0.0927152	0.0592105	0.0723684	0.0394737	0	0	0.0927152
C	0	0.0066667	0.0066667	0.0066225	0.0065789	0.0131579	0.0131579	0	0	0
D	0.0268456	0.0466667	0.04	0.0728477	0.0263158	0.0460526	0.0460526	0	0	0.0463576
E	0.0738255	0.0666667	0.0866667	0.0860927	0.0592105	0.0526316	0.0328947	0	0	0.0993377
F	0.0268456	0	0.0133333	0.013245	0.0197368	0.0328947	0	0	0	0.013245
G	0.0671141	0.0866667	0.0733333	0.0529801	0.125	0.0394737	0.0986842	0	0	0.0728477
H	0.0268456	0.02	0.02	0.0066225	0.0263158	0.0197368	0.0460526	0	0	0.013245
I	0.0469799	0.02	0.04	0.0198675	0.0131579	0.0197368	0.0263158	0	0	0.0264901
K	0.0671141	0.0733333	0.0266667	0.0794702	0.0460526	0.0855263	0.0657895	0	0	0.0331126
L	0.0469799	0.0733333	0.08	0.0993377	0.0855263	0.0328947	0.0789474	0	0	0.0794702

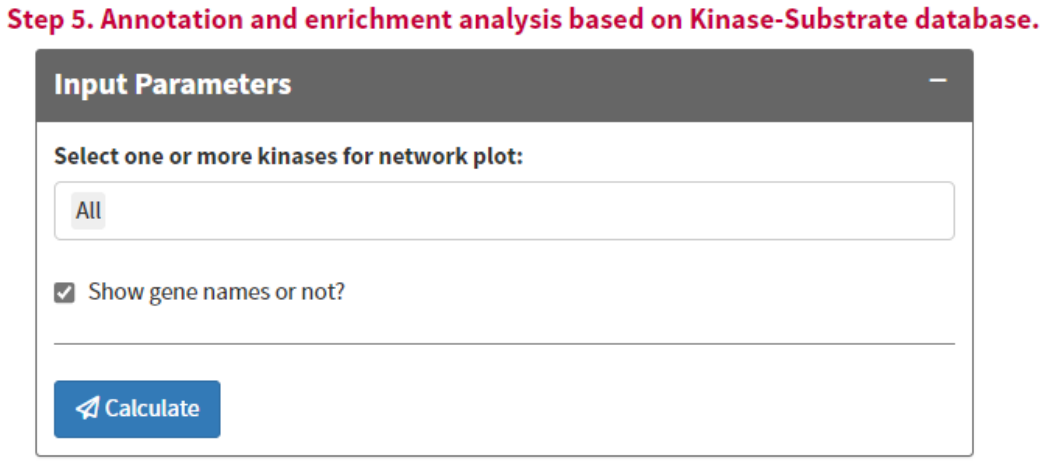
Showing 1 to 10 of 20 entries

Previous 1 2 Next

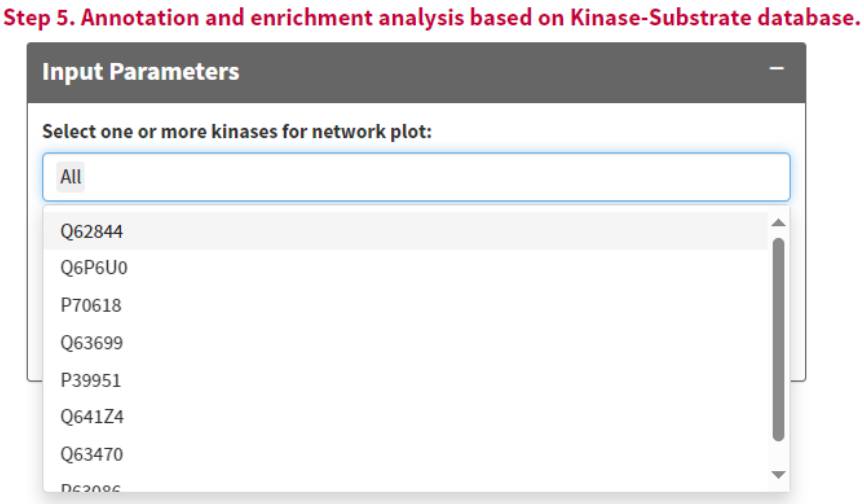
Step 5. Annotation and enrichment analysis based on Kinase-Substrate database

This step will perform kinase-substrate annotation and enrichment for every kinase using Fisher test based on PhosphoSitePlus (9), to facilitate a site-specific analysis on phosphorylation regulation. Users should note that here is only for phoshoproteomics data, other modification data are not inappropriate here.

Step 5.1. Parameters



Select one or more kinases for network plot: This means which kinase users want to select to show a network plot. If selecting “All”, all kinases identified in users’ data will be used to plot network. User can also just select one or more kinases, like below:



Show gene names or not? If true, the node table, the edge table, and the network plot will show the gene names, otherwise, it shows the IDs (e.g. UniProt IDs by default).

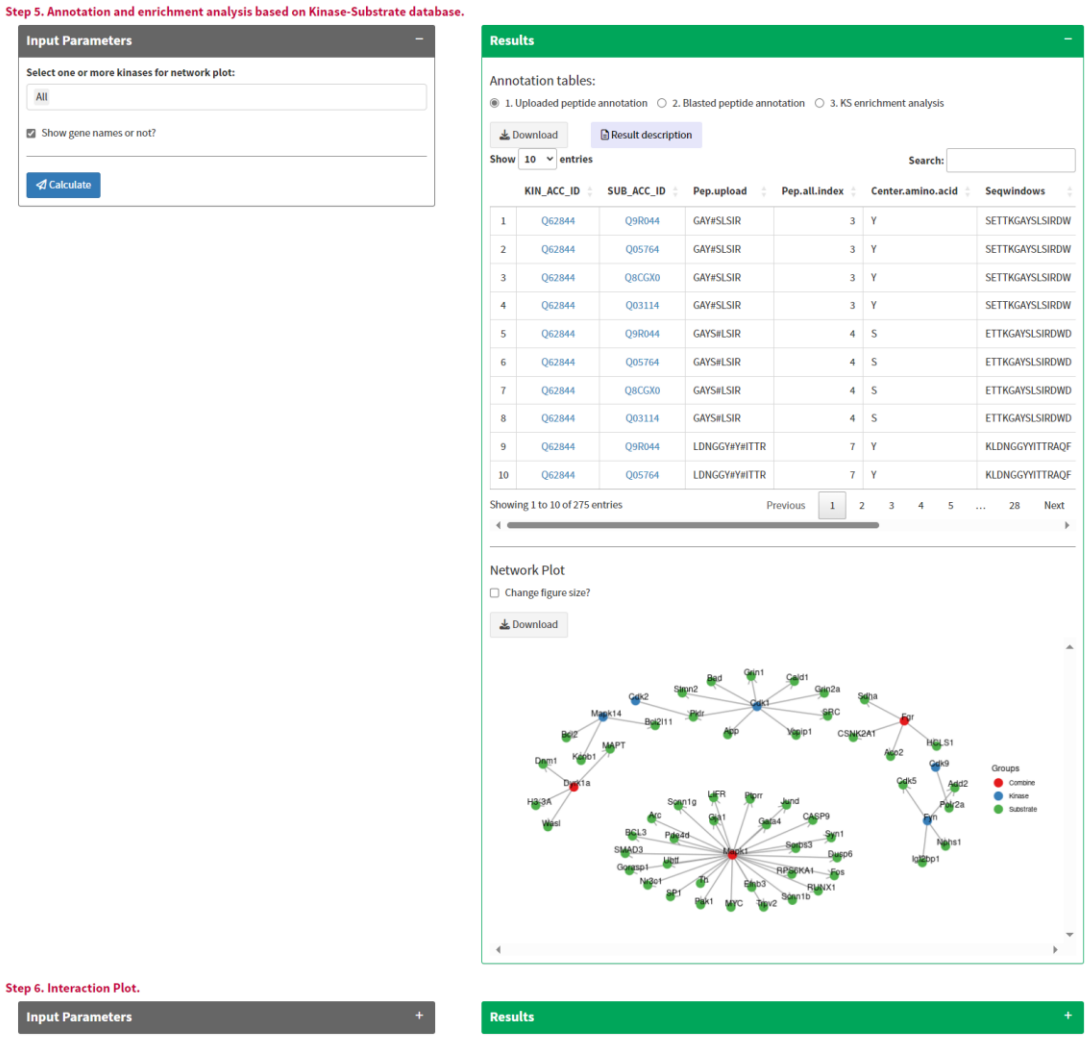
Then, users can click the “Calculate” button.

Step 5.2. Results

Here shows the annotation tables for the uploaded peptides (*1. Uploaded peptide annotation*) and the blasted peptides (*2. Blasted peptide annotation*) based on the kinase-substrate database from PhosphoSitePlus (9).

Step 5.2.1. Uploaded peptide annotation

When users click "*1. Uploaded peptide annotation*", they will obtain the annotation table for the uploaded peptides:



Every column in this table means (Users can also check this by clicking "Result description" button):

1. KIN_ACC_ID: kinase uniprot id.

2. SUB_ACC_ID: substrate uniprot id.

3. Pep.upload: the original peptide.

4. Pep.all.index: the position of all modified amino acid in the peptide.

5. Center.amino.acid: the central amino acid in the aligned peptide. Or, Center.amino.acids.Other: the central amino acid mapped from the human peptides.

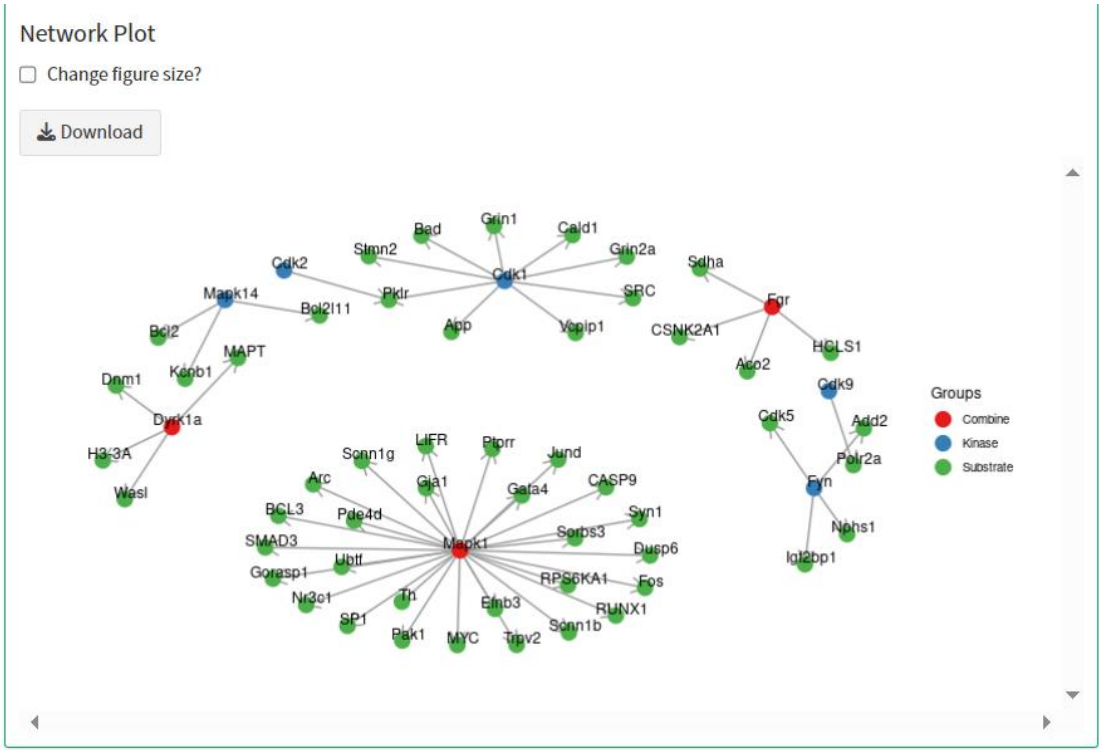
6. Seqwindows: the aligned standard peptides. Or, Seqwindows.Other: the standard peptides mapped from the human peptides.

7. PROindex.from.Database: the position of modified amino acid in the protein sequence. Or, PROindex.from.Other: the position of modified amino acid in the mapped human protein sequence.

8. GENE: kinase gene name.

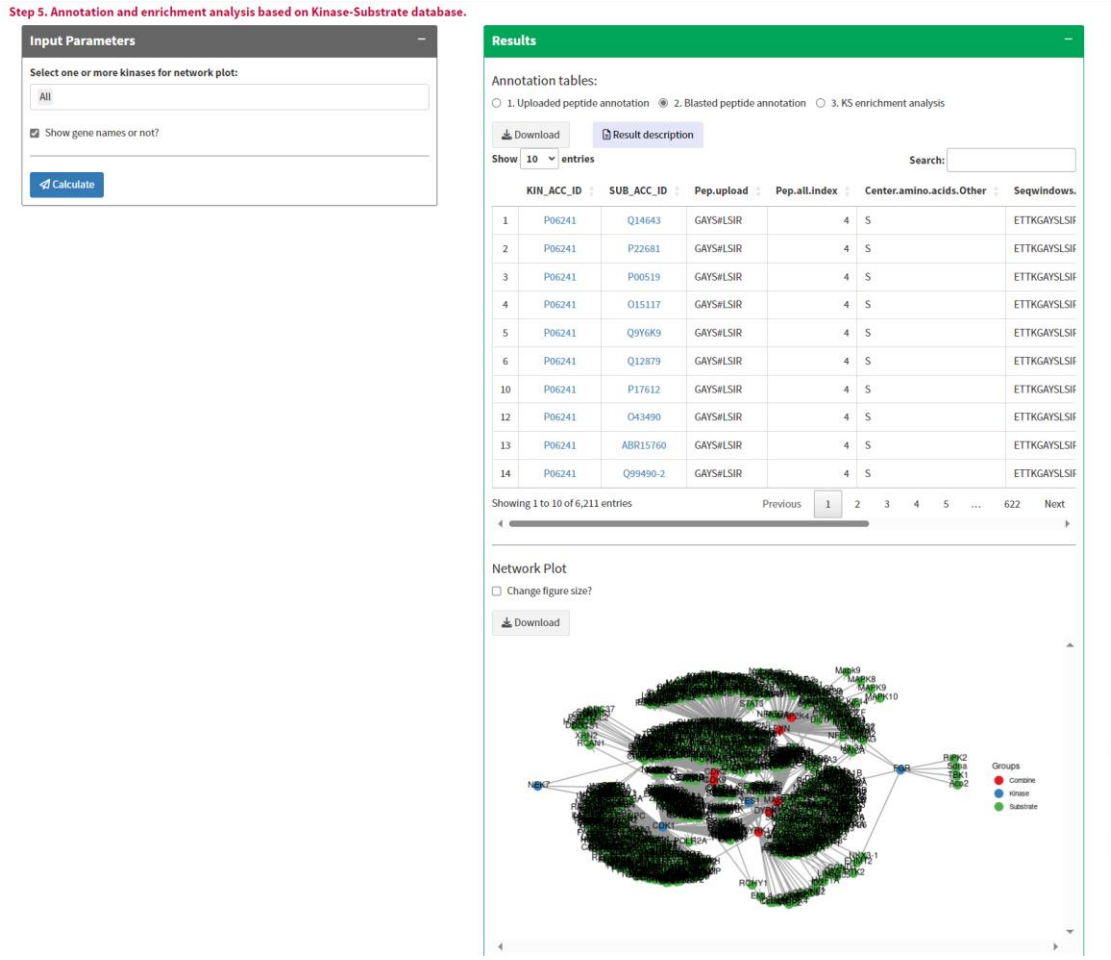
9. SUB_GENE: substrate gene name.

Then they can also obtain the network plot:



Step 5.2.2. Blasted peptide annotation

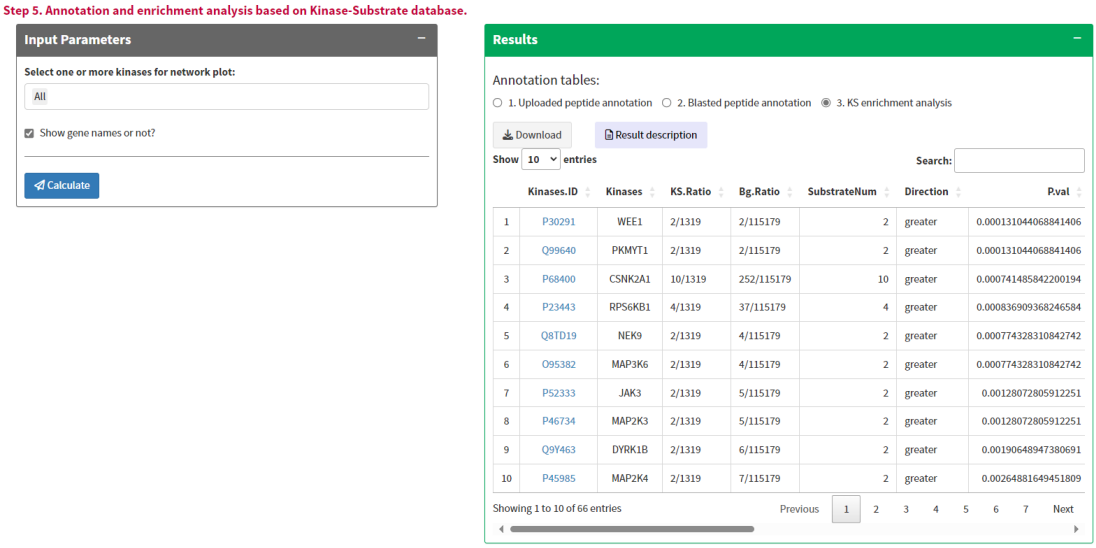
When users click "*2. Blasted peptide annotation*", they will obtain the annotation table for the blasted peptides:



The annotation table and network plot are similar to the above.

Step 5.2.3. KS enrichment analysis

When users click "*3. KS enrichment analysis*", this function will process the enrichment analysis at phosphosite levels for every kinases. Then, the results are shown as below:



Every column in this table means (Users can also check this by clicking "Result description" button):

1. Kinases.ID: Kinase uniprot ids.

2. Kinases: Kinase gene names.

3. KS.Ratio: k/n, k means the overlap between phosphosites-of-interest and the uploaded phosphosite set, n means the number of all unique phosphosites-of-interest.

4. Bg.Ratio: M/N, M means the number of substrate phosphosites of each kinase in the whole phosphosite set, N means the number of phosphosites in the whole phosphosite set.

5. SubstrateNum: Same as k.

6. Direction: If KS.Ratio >= Bg.Ratio, the value is 'greater', otherwise, 'less'.

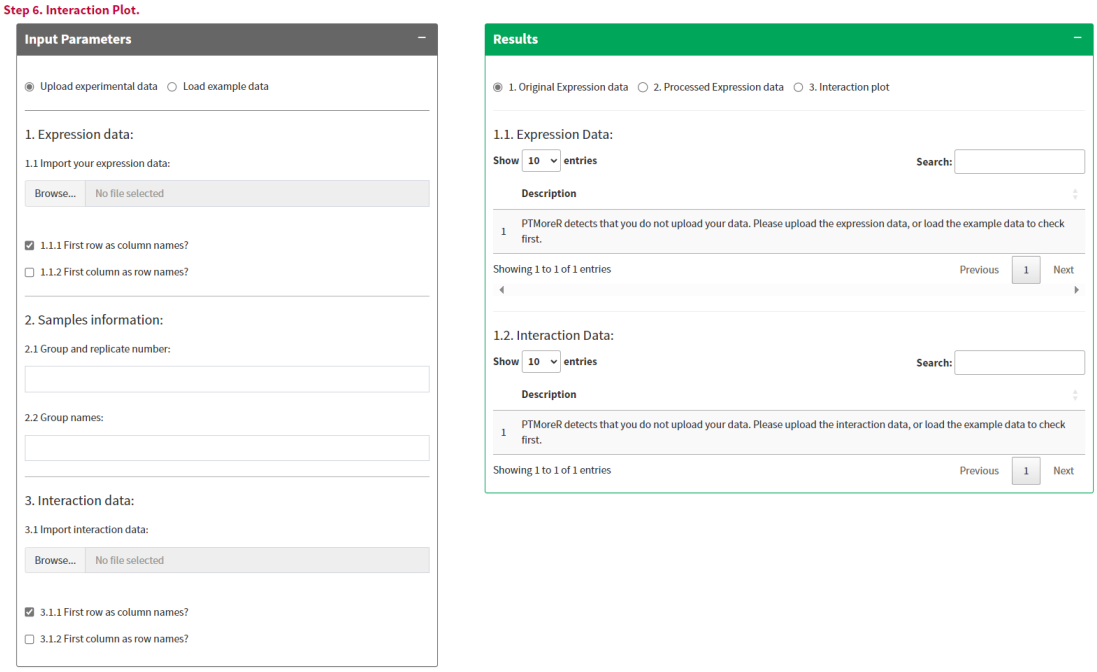
7. P.val: Original P value obtained from Fisher test.

8. P.adj: Adjusted P value based on the BH method.

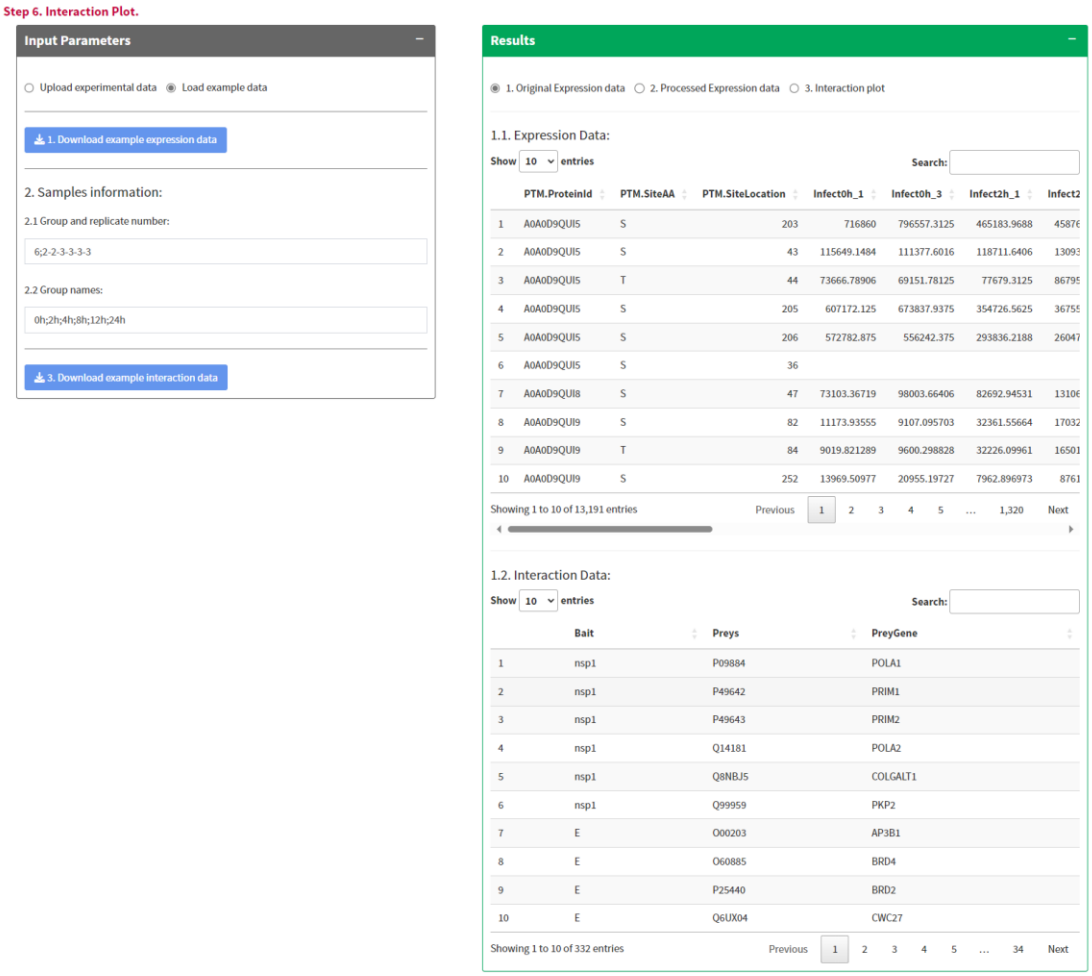
9. Substrates: Substrate phosphosites-of-interest.

Step 6. Interaction Plot

In some cases, it may not be convenient for the users to display a network diagram of protein-protein-modification sites relationships. This step mainly shows the plot of the interaction between the uploaded and the blasted peptides/proteins. In this step, users need to upload two kinds of data (Expression data and Interaction data) and type in the samples information, shown as below:



Users can click “Load example data”. And shown in the “*1. Original Expression data*” part below, the example data are from the African green monkey (*Chlorocebus sabaeus*) cell phosphoproteome quantified at 6 time points after SARS-CoV-2 infection (0, 2, 4, 8, 12, or 24 h) using Spectronaut (18.0.230605.50606) (11).



Step 6.1. Data preparation

Here users should prepare two kinds of data: The modification site expression data and the interaction data.

Step 6.1.1. The modification site expression data

Users should note that the modification site expression data should be obtained from the same species as processed in the previous steps. As shown in the example data below, the first three columns are protein ids (e.g. PTM.ProteinId), modification site amino acid (e.g. PTM.SiteAA), modification site position (e.g. PTM.SiteLocation). And the other columns are samples (e.g. The African green monkey cell samples with 6 time points after SARS-CoV-2 infection (0, 2, 4, 8, 12, or 24 h)). These data can be extracted from many popular software, such as MaxQuant, Spectronaut etc., and saved in a .csv file. The missing values are NA and shown as blank after uploaded into *PTMoreR*.

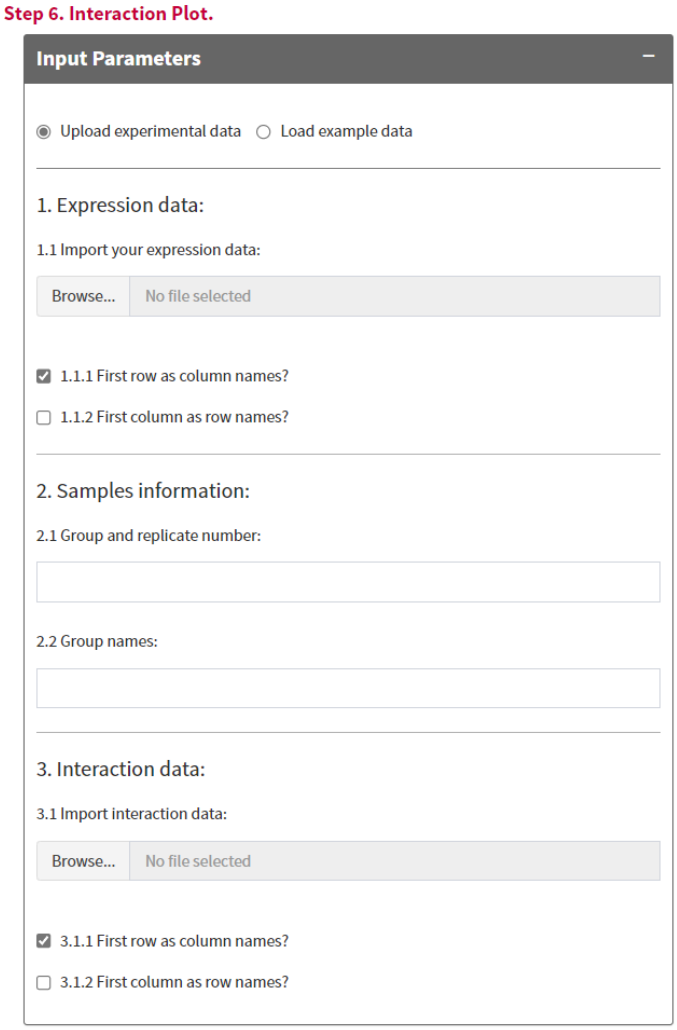
PTM.ProteinId	PTM.SiteAA	PTM.SiteLocation	Infect0h_1	Infect0h_3	Infect2h_1	Infect2h_3	Infect4h_1	Infect4h_2	Infect4h_3
A0A0D9QUI5	S	203	716860	796557.3125	465183.9688	458769.1875	467674.1875	517208.6875	563238.9375
A0A0D9QUI5	S	43	115649.1484	111377.6016	118711.6406	130935.4453	91683.72656	106065.1563	109223.6016
A0A0D9QUI5	T	44	73666.78906	69151.78125	77679.3125	86795.76563	60218.52344	69550.14063	71446.70313
A0A0D9QUI5	S	205	607172.125	673837.9375	354726.5625	367551.6875	375132.9375	433767.8125	444802.3125
A0A0D9QUI5	S	206	572782.875	556242.375	293836.2188	260473.8125	358492.0313	304453.4063	395302.4063
A0A0D9QUI5	S	36						418.7579041	
A0A0D9QUI8	S	47	73103.36719	98003.66406	82692.94531	131065.3516	88582.30469	156939.7969	121457.4609
A0A0D9QUI9	S	82	11173.93555	9107.095703	32361.55664	17032.79297	37902.76563	34069.21094	28763.04102
A0A0D9QUI9	T	84	9019.821289	9600.298828	32226.09961	16501.47852	38630.11328	35007.21875	29710.75586
A0A0D9QUI9	S	252	13969.50977	20955.19727	7962.896973	8761.28125	11051.95215	10059.21973	9189.84668

Step 6.1.2. The interaction data

Users should know the interaction data in advance, which could be protein-protein interaction data, or drug-protein interaction data and so on. Three are three columns in this table: The first one is protein ids/names from one species (e.g. SARS-CoV-2 protein names), or drug names; the second one is protein ids from the species that users blast to in the previous steps (e.g. human protein ids, as users blast the phosphopeptides from green monkey to human); the third one is gene names relative to the protein ids in the second column.

Bait	Preys	PreyGene
nsp1	P09884	POLA1
nsp1	P49642	PRIM1
nsp1	P49643	PRIM2
nsp1	Q14181	POLA2
nsp1	Q8NBJ5	COLGALT1
nsp1	Q99959	PKP2
E	O00203	AP3B1
E	O60885	BRD4
E	P25440	BRD2
E	Q6UX04	CWC27

Step 6.2. Parameters



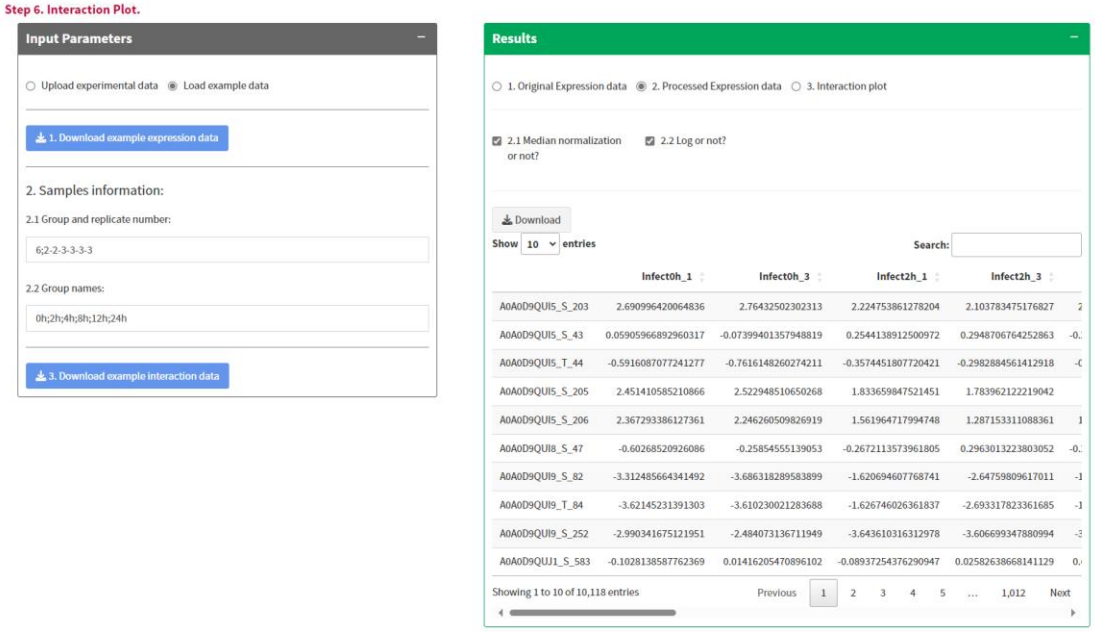
The parameters in the "*1. Expression data*" and the "*3. Interaction data*" parts are similar and easy. Users just click the "*Browse*" and find their data in a .csv file, then notice that whether the first row/column is used as row/column names, if true, they should select relative parameters.

For the "*2. Samples information*" part, the "*2.1. Group and replicate number*" means users can type in the group number and the biological replicate number here. Please note, the group number and replicate number are linked with ";", and the replicate number of each group is linked with "-". For example, if you have two groups, each group has three replicates, then you should type in "2;3-3" here. Similarly, if you have 3 groups with 5 replicates in every groups, you should type in "3;5-5-5".

2.2 Group names: Type in the group names of your samples. Please note, the group names are linked with ";". For example, there are 6 time points after SARS-CoV-2 infection (0, 2, 4, 8, 12, or 24 h), you can type in "0h;2h;4h;8h;12h;24h".

Step 6.3. Processed Expression data

After uploading the proper data and setting up the right parameters, users can click “2. Processed Expression data” (see below), and this tool will process the data automatically for users. By default, those sites with over 50% missing ratios across all samples were removed and missing values were imputed with the k-Nearest Neighbor algorithm provided in NAguideR (12).

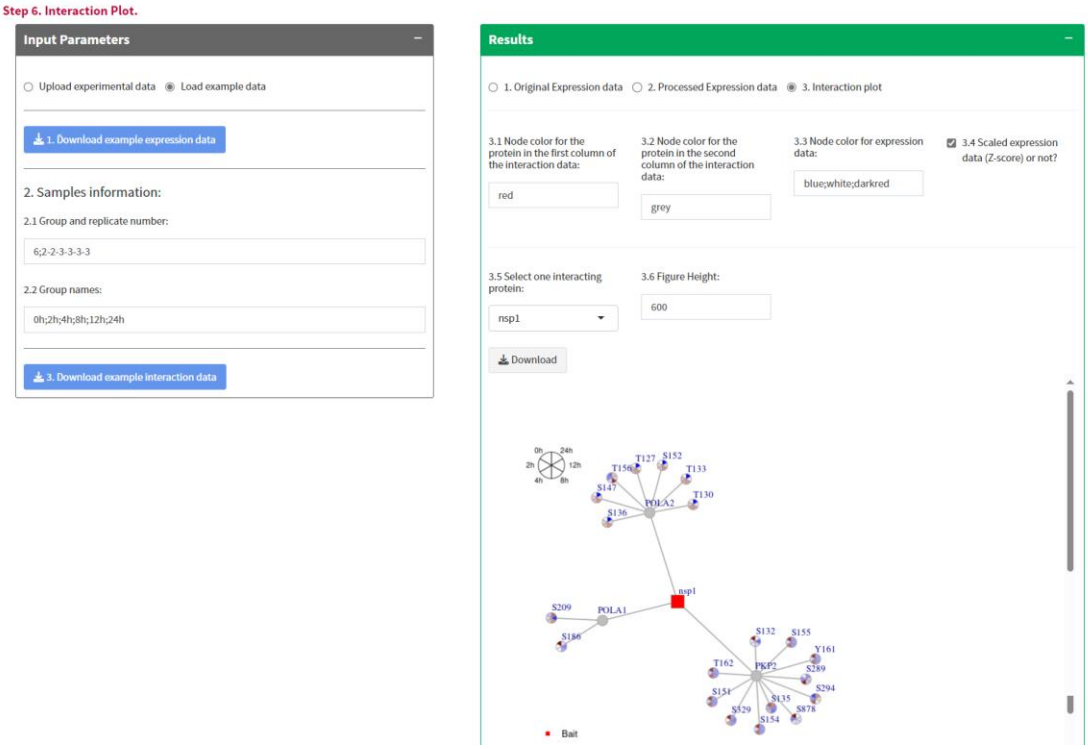


2.1. Median normalization or not: if true, *PTMoreR* will process median normalization for original data. (Note, *PTMoreR* was not designed to perform sophisticated normalization analysis. Any normalized datasets with NA can be accepted for analysis).

2.2. Log or not: if true, the data will be transformed to the logarithmic scale with base 2.

Step 6.4. Interaction plot

In this step, PTMoreR will show the final interaction plot based on the uploaded expression data and the interaction data.



3.1 Node color for the protein in the first column of the interaction data: Three are three columns in the interaction data (see "*7.1.2. The interaction data*" part), the node color for one protein in the first column (e.g. One SARS-CoV-2 protein in the *Bait* column).

3.2 Node color for the protein in the second column of the interaction data: Similar as above, the node color for one protein in the second column (e.g. One human protein in the *Preys* column).

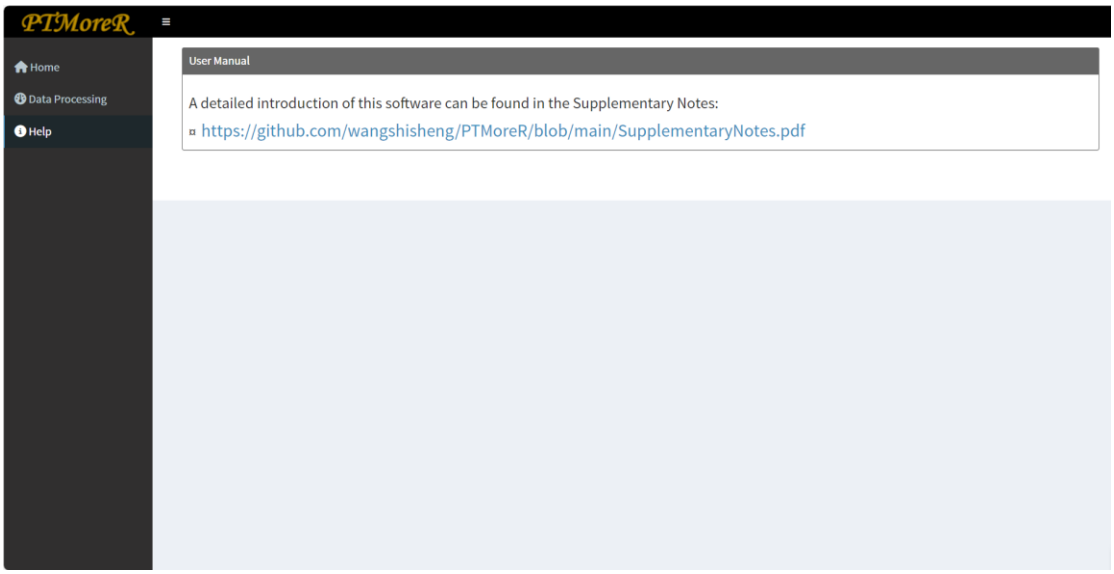
3.3 Node color for expression data: Three color names which are linked with ";" for the expression data (e.g. blue;white;darkred). The first color (e.g. blue) indicates the lowest expression value, the second one (e.g. white) indicates the middle expression value, and the third one (e.g. darkred) indicates the largest expression value.

3.4 Scaled expression data (Z-score) or not? If true, the expression data will be scaled by rows, which means the mean of all of the values in each row is 0 and the standard deviation is 1.

3.5 Select one interacting protein: Users can change this parameter to select any protein in the first column (e.g. One SARS-CoV-2 protein in the *Bait* column) and the plot will show the selected protein.

3. Help

This part provides a detailed operation manual about *PTMoreR*, which is saved as a pdf file as below:



4. How to run this tool locally?

PTMoreR is an open source software for non-commercial use and all codes can be obtained on our GitHub: https://github.com/wangshisheng/PTMoreR. If users want to run *PTMoreR* on their own computer independent of the internet speed, they should operate as below:

As this tool was developed with R, you may:

a) Install R. You can download R from here: https://www.r-project.org/.

b) Install RStudio. (Recommendatory but not necessary). You can download RStudio from here: https://www.rstudio.com/.

c) Check packages. After installing R and RStudio, you should check whether you have installed these packages (shiny, shinyjs, shinyWidgets, shinyBS, DT, data.table, openxlsx, Biostrings, GenomicFeatures, rtracklayer, stringi, stringr, ggsci, ggplot2, ggrepel, msa, tidyr, ggraph, graphlayouts, scales, impute, igraph, scatterpie, plotfunctions, mapplots, devtools, KinSwingR). You may run the codes below to check them:

```
if(!require(pacman)) install.packages("pacman")
pacman::p_load(shiny, shinyjs, shinyWidgets, shinyBS, DT, data.table, openxlsx, Biostrings, GenomicFeatures, rtracklayer, stringi, stringr, ggsci, ggplot2, ggrepel, msa, tidyr, ggraph, graphlayouts, scales, impute, igraph, scatterpie, plotfunctions, mapplots, devtools, KinSwingR)
```

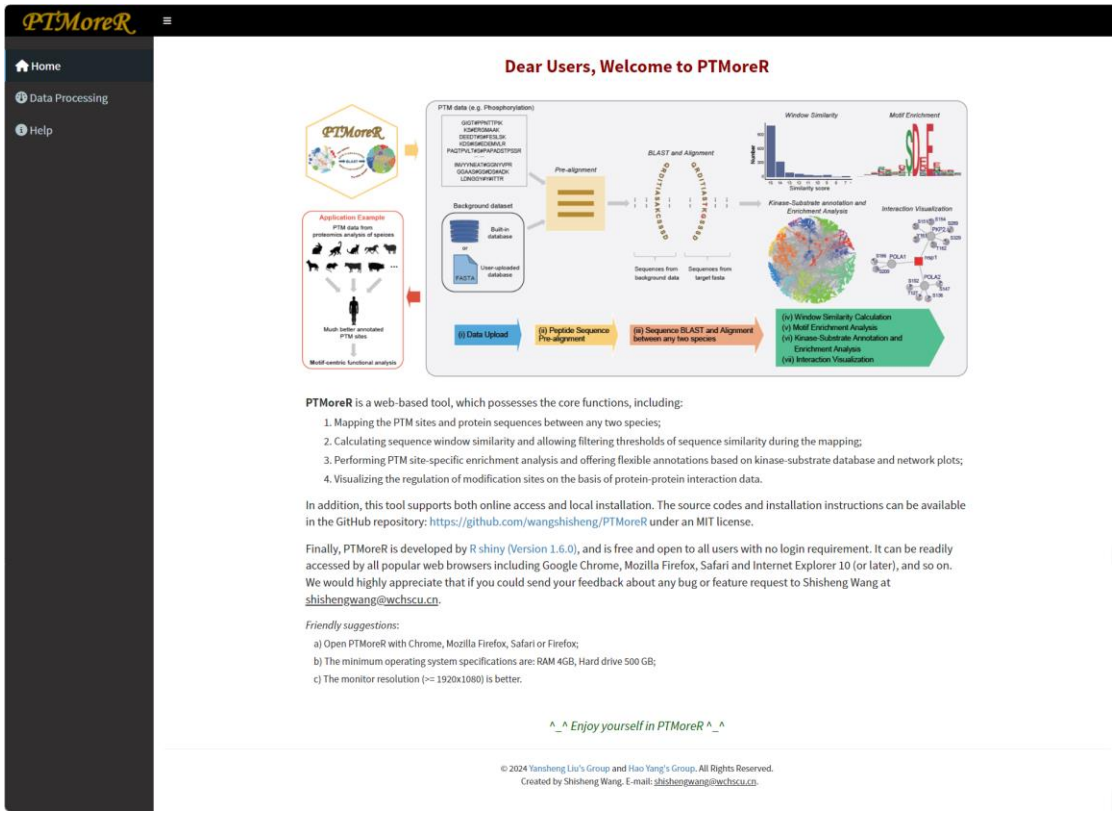
Then install some packages from GitHub, as below:

```
devtools::install_github("drostlab/metablastr", build_vignettes = TRUE, dependencies = TRUE)
devtools::install_github('omarwagih/rmotifx')
devtools::install_github("omarwagih/ggseqlogo")
```

d) Run this tool locally

```
if(!require(PTMoreR)) devtools::install_github("wangshisheng/PTMoreR")
library(PTMoreR)
PTMoreR_app()
```

Then PTMoreR will be started as below, and the detailed operation about PTMoreR can be found in the Supplementary Notes part 1-4 above.



III. References

1. Giansanti, P., Samaras, P., Bian, Y., Meng, C., Coluccio, A., Frejno, M., Jakubowsky, H., Dobiasch, S., Hazarika, R.R., Rechenberger, J. et al. (2022) Mass spectrometry-based draft of the mouse proteome. Nat Methods, 19, 803-811.
2. Ochoa, D., Jarnuczak, A.F., Vieitez, C., Gehre, M., Soucheray, M., Mateus, A., Kleefeldt, A.A., Hill, A., Garcia-Alonso, L., Stein, F. et al. (2020) The functional landscape of the human phosphoproteome. Nat Biotechnol, 38, 365-373.
3. UniProt, C. (2021) UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res, 49, D480-D489.
4. Ba, Q., Hei, Y., Dighe, A., Li, W., Maziarz, J., Pak, I., Wang, S., Wagner, G.P. and Liu, Y. (2022) Proteotype coevolution and quantitative diversity across 11 mammalian species. Sci Adv, 8, eabn0756.
5. Grubbs, F.E. (1950) Sample Criteria for Testing Outlying Observations. The Annals of Mathematical Statistics, 21, 27-58, 32.
6. Tyanova, S., Temu, T. and Cox, J. (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat Protoc, 11, 2301-2319.
7. Bruderer, R., Bernhardt, O.M., Gandhi, T., Miladinović, S.M., Cheng, L.-Y., Messner, S., Ehrenberger, T., Zanotelli, V., Butscheid, Y. and Escher, C. (2015) Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. Mol Cell Proteomics, 14, 1400-1410.
8. Wang, S., Cai, Y., Cheng, J., Li, W., Liu, Y. and Yang, H. (2019) motifeR: An Integrated Web Software for Identification and Visualization of Protein Posttranslational Modification Motifs. Proteomics, 19, e1900245.
9. Hornbeck, P.V., Kornhauser, J.M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V. and Sullivan, M. (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res, 40, D261-270.
10. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res, 13, 2498-2504.
11. Bruderer, R., Bernhardt, O.M., Gandhi, T., Miladinovic, S.M., Cheng, L.Y., Messner, S., Ehrenberger, T., Zanotelli, V., Butscheid, Y., Escher, C. et al. (2015) Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. Mol Cell Proteomics, 14, 1400-1410.
12. Wang, S., Li, W., Hu, L., Cheng, J., Yang, H. and Liu, Y. (2020) NAguideR: performing and prioritizing missing value imputations for consistent bottom-up proteomic analyses. Nucleic Acids

Res, 48, e83.