Supporting Information for:

motifeR: An integrate web server for identification of protein post-translational modification motifs and inference of kinase activity

Shisheng Wang,[†] Yue Cai,[‡] Jingqiu Cheng,[†] Yansheng Liu,[*ξ] and Hao Yang[*†]

† West China-Washington Mitochondria and Metabolism Research Center; Key Lab of Transplant Engineering and Immunology, MOH, West China Hospital, Sichuan University, Chengdu, China.

‡ Ministry of Science and Technology, West China Hospital, Sichuan University, Chengdu, China.

ξ Department of Pharmacology, Yale University School of Medicine, New Haven, CT 06520, USA.

Corresponding Author

*Email address: yansheng.liu@yale.edu; yanghao@scu.edu.cn.

Table of Contents

Supplementary Figures

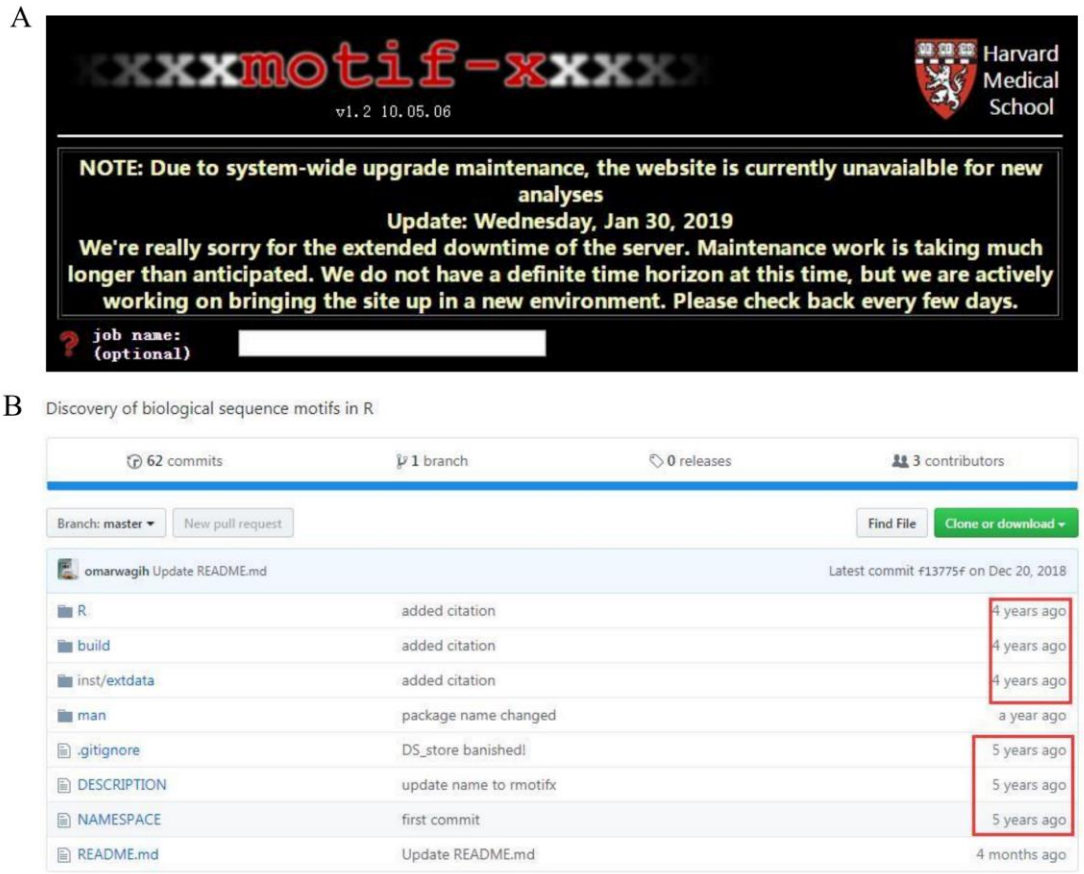


Figure S1. The current status of some published tools. A. motif-x has been in maintenance for a while and users don't know when it can be back. (Link: http://motif-x.med.harvard.edu/motif-x.html). B. rmotifx package has not been updated in a long time. (Link: https://github.com/omarwagih/rmotifx/).

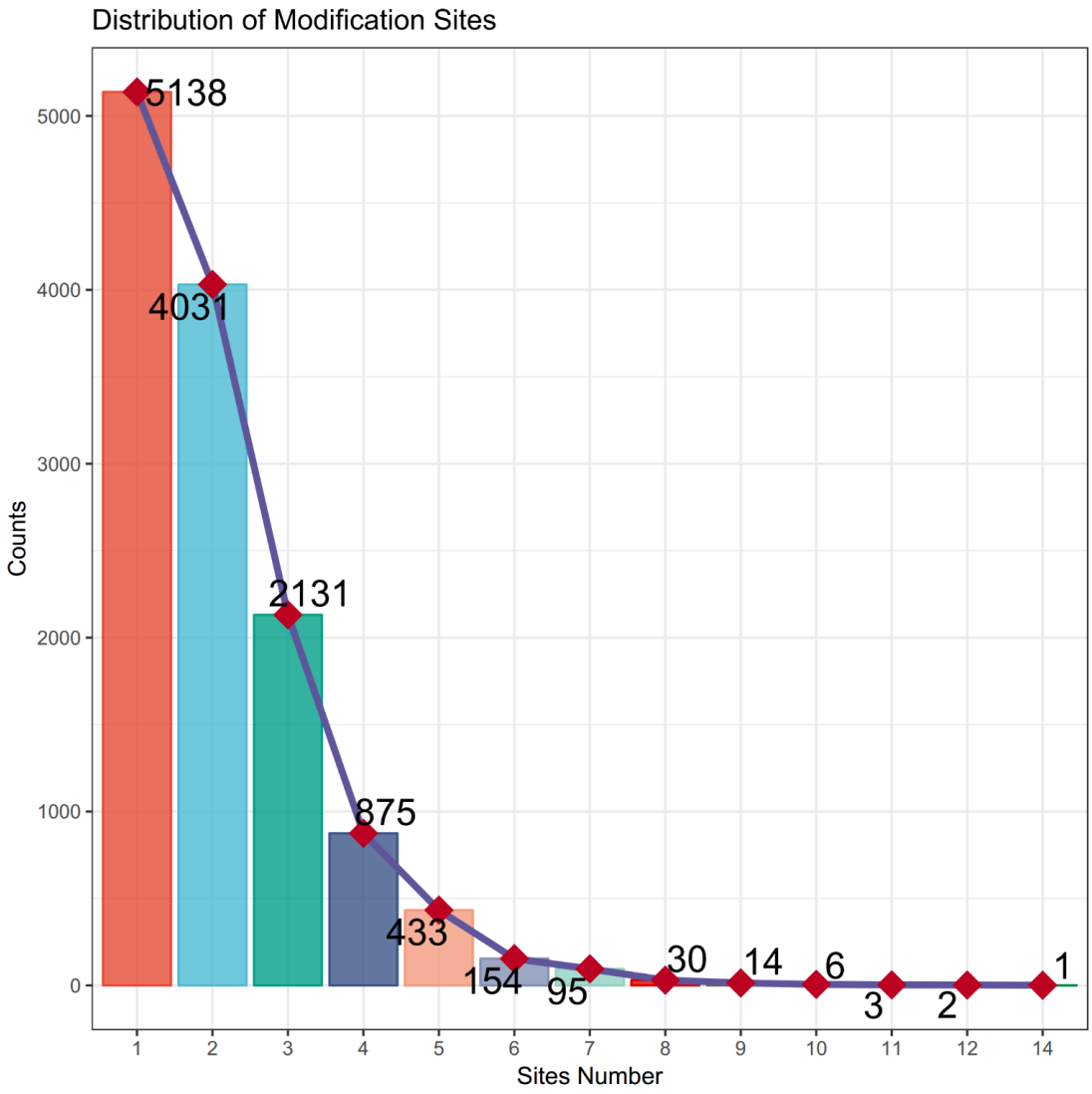


Figure S2. Distribution of peptide modification site number. This plot can be obtained in "Pre-alignment" module.

Supplementary notes:

motifeR a powerful and comprehensive web server, can provide three main functions (sequence pre-alignment, motif discovery, kinase-substrate analysis) to help users find the patterns of residues along the short span of a protein or polypeptide and process kinase-substrate analysis subsequently. Here we present the detailed introduction and operation of motifeR, users can follow this manuscript to analyze their own data freely and conveniently.

Users can visit this site: https://www.omicsolution.org/wukong/motifeR. Then the website homepage can be shown like this:

1. Data Preparation

1.1 Foreground data set, which can be obtained from many softwares, such as MaxQuant, Proteome Discovery, Spectronaut and so on. Then users just prepare the peptide sequences like this:

	A
1	Pep
2	TQVPASVESQK#PR
3	LIDGK#DK#AAR
4	VSSPPPTITQQGK#K
5	VAHPQDPHHSSEK#PVIHCHK
6	DLTDYLMK#ILTER
7	DIK#EKLCYVALDFENEMATAASSSSLEK
8	EK#LCYVALDFENEMATAASSSSLEK
9	EITALAPSTMK#IK
10	EITALAPSTMKIK#
11	DSYVGDEAQSK#R
12	RGILTLK#YPIEHGIITNWDDMEK
13	LNKDDPIGNINLAMEIAEK#HLDIPK
14	TPEK#TMQAMQK
15	QSILAIQNEVEK#VIQSYSIR
16	ELPPDQAQYCIK#R
17	SEFYK#HIVLSGGSTMYPGLPSRLER
18	SALSGHLETVILGLLK#TPAQYDASELK
19	SVCHLQK#VFER
20	QTFK#SHFGR
21	LILGLMMPPAHYDAKQLK#
22	GAGTDEK#TLTR
23	GSYNIK#SR
24	LHTEGDK#AFVEFLTDEIKEEK

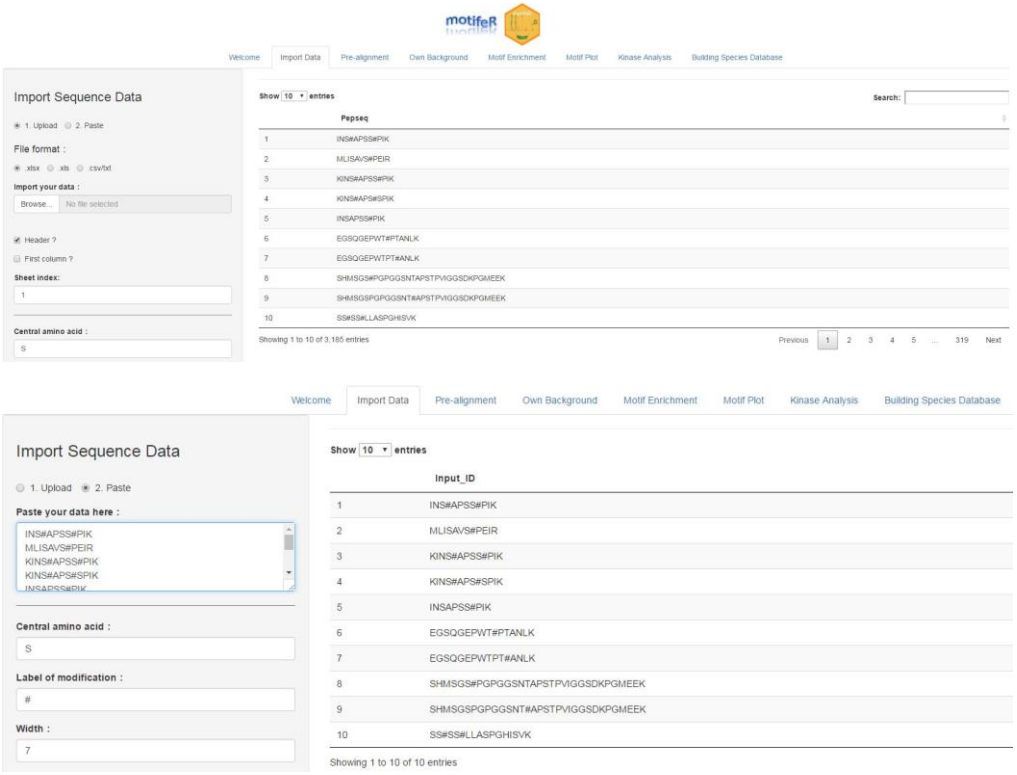
The length of these sequences can be different, and the standard sequences can be obtained in the pre-alignment step. On the other hand, users should mark those modified residues with some label they like (such as “#” or “@”).

1.2 Database/Background data set. motifeR supports users to upload their own database, if they don’t want to use the default data. Herein the database formats can be .fasta, .xlsx, .xls, .csv, or .txt. Particularly, if users want to upload the .fasta file, it should contain all protein sequences of the species they study, otherwise, the other formats (.xlsx, .xls, .csv, .txt.) should contain standard sequences like below:

Windows
_____MSKVFKKTS
SKVFKKTSSNGKLSI
KVFKKTSSNGKLSIY
TSSNGKLSIYLGKRD
LQVVPAESSSPQGPL
QVVPAESSSPQGPLT
VVPAESSSPQGPLTV
MVTNLPCSVTLQPGP
GIDFEVKSFCAENPE
ENPEETVSKRDYVRL
PEAGPGPSAQTIRRF

2. Import data.

2.1 Uploading data. When users prepare their data (Foreground and Background data set), they can upload these data from here:



There are two choices for users:

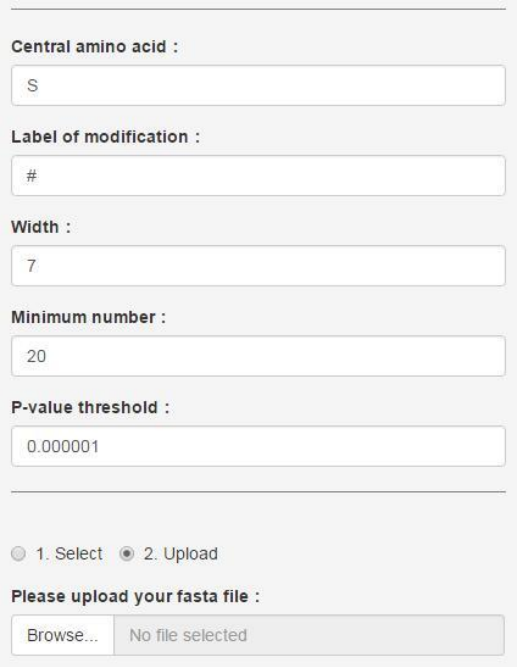
1. Upload, users choose the right format and then click “Browse” button to import there data;

Header: this means whether the first row is column names. If true, you should choose this parameter.

First column: this means whether the first column is row names. If true, you should choose this parameter.

2. Paste, users can also paste their sequences in the box.

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



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”.

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

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

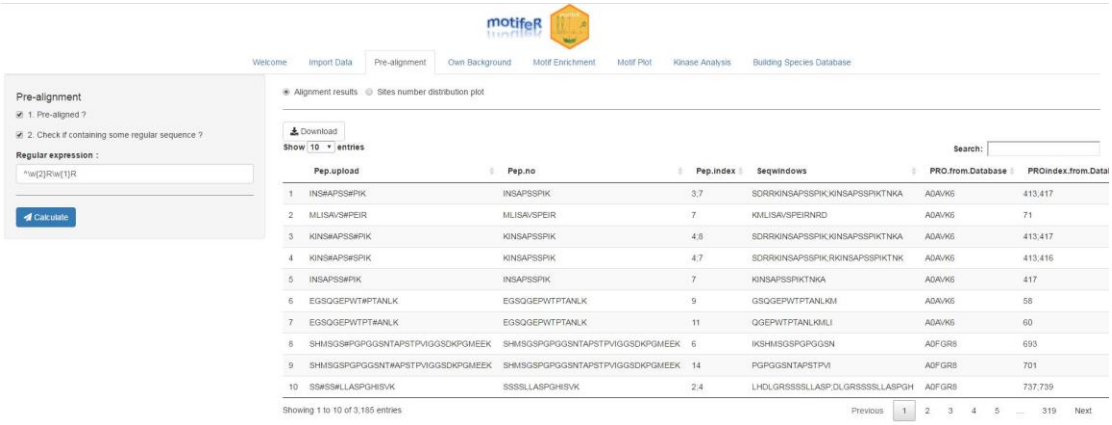
Minimum number: this threshold refers to the minimum number of times you wish each of your extracted motifs to occur in the data set.

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

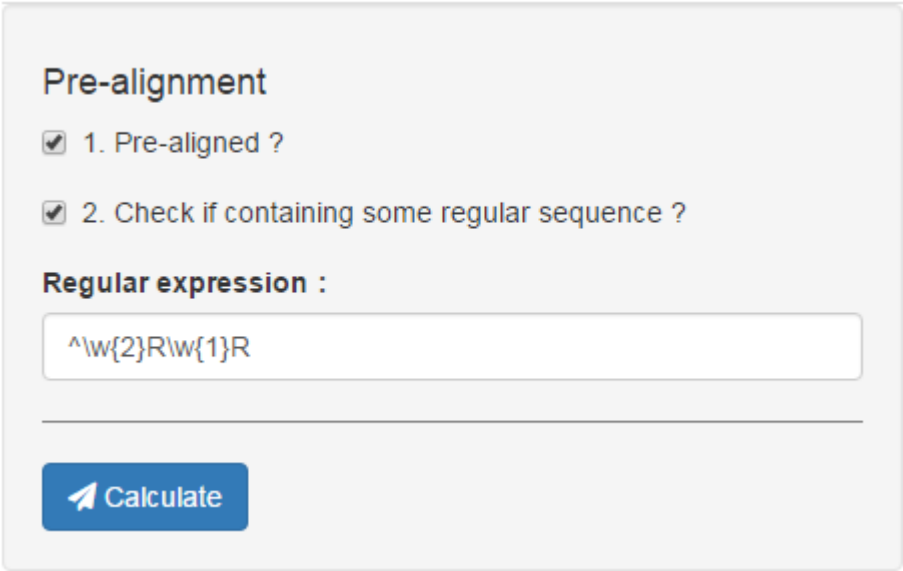
Select or Upload fasta file as background data set: if users want to use the default database, they just select relative species, please note, the default database only have Human with 15 length sequences now, more species data will be implemented in the future. Optionally, users can also upload their own fasta file, no species limits here, but the calculation time would be longer.

3. 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.



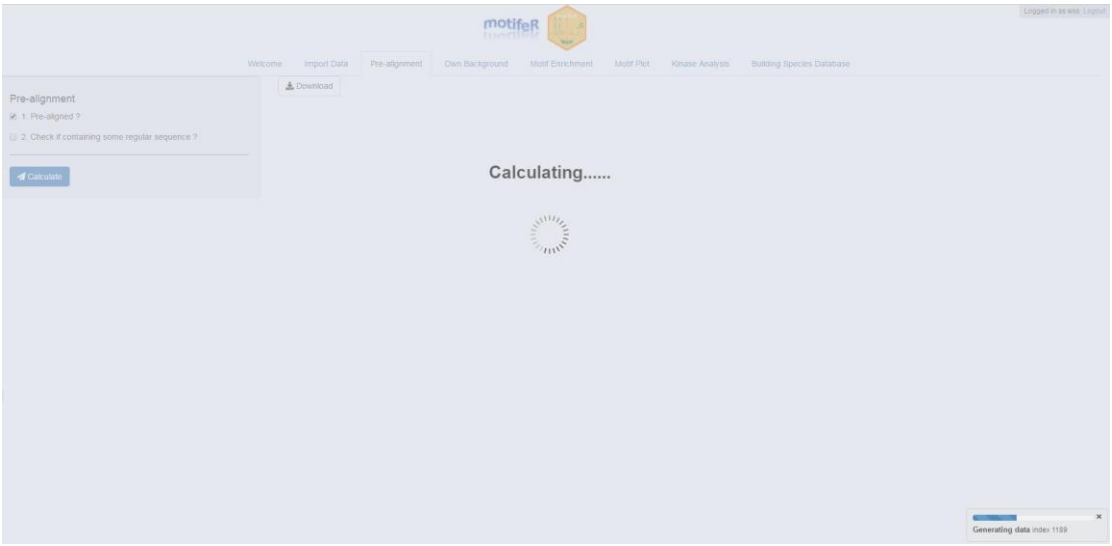
3.1 Parameters



Pre-aligned: ask users whether pre-align their sequences, if your sequences are standard (e.g. 15 length amino acids), you can unselect this parameter. Default is true.

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.

3.2 results

There are two results here.

3.2.1 Alignment results

	Pep.upload	Pep.no	Pep.main.index	Pep.all.index	Seqwindows	PRO.from.Database	PROindex.from.Database	Contain.if
1	INS#APSS#PIK	INSAPSSPIK	3;7	3;7	SDRRKINSAPSSPIK;KINSAPSSPIKTNKA	A0AVK6	413;417	No
2	MLISAVS#PEIR	MLISAVSPEIR	7	7	KMLISAVSPEIRNRD	A0AVK6	71	No
3	KINS#APSS#PIK	KINSAPSSPIK	4;8	4;8	SDRRKINSAPSSPIK;KINSAPSSPIKTNKA	A0AVK6	413;417	No
4	KINS#APS#SPIK	KINSAPSSPIK	4;7	4;7	SDRRKINSAPSSPIK;RKINSAPSSPIKTNK	A0AVK6	413;416	No
5	INSAPSS#PIK	INSAPSSPIK	7	7	KINSAPSSPIKTNKA	A0AVK6	417	No
6	EGSQGEPWT#PTANLK	EGSQGEPWTPTANLK	9	9	GSQGEPWTPTANLKM	A0AVK6	58	No
7	EGSQGEPWTPT#ANLK	EGSQGEPWTPTANLK	11	11	QGEPWTPTANLKMLI	A0AVK6	60	No
8	SHMSGS#PGPGGSNTAPSTPVIGGSDKPGMEEK	SHMSGSPGPGGSNTAPSTPVIGGSDKPGMEEK	6	6	IKSHMSGSPGPGGSN	A0FGR8	693	No
9	SHMSGSPGPGGSNT#APSTPVIGGSDKPGMEEK	SHMSGSPGPGGSNTAPSTPVIGGSDKPGMEEK	14	14	PGPGGSNTAPSTPVI	A0FGR8	701	No
10	SS#SS#LLASPGHISVK	SSSSLLASPGHISVK	2;4	2;4	LHDLGRSSSSLLASP;DLGRSSSSLLASPGH	A0FGR8	737;739	No

Showing 1 to 10 of 3,185 entries

Previous 1 2 3 4 5 ... 319 Next

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

Pep.no: the peptide skeleton.

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

Seqwindows: the standard peptides.

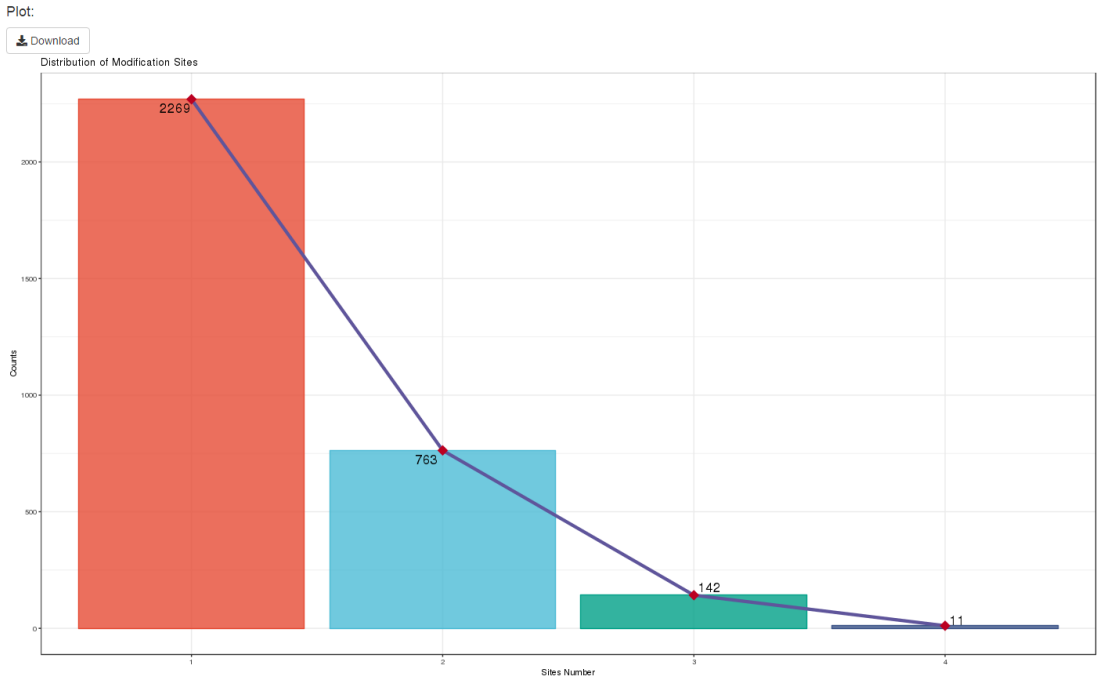
PRO.from.Database: which protein contains this peptide.

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

Contain.if: whether containing the sequences that match the regular expression, if true, marked with “Yes”, otherwise, “No”. This column only appears when users choose the parameter--- *Check if containing some regular sequence.*

3.2.2 Sites number distribution plot

First, this software counts the number of modified sites and plot it:



Second, this tool also extracts those peptides with multi modified sites, the result is similar to that from pre-alignment except the last column.

Multi-Sites Data:

Download

Show 10 entries

Search:

	Pep.upload	Pep.no	Pep.index	Seqwindows	PRO.from.Database	PROindex.from.Database	Seqwindows_MultiSites
1	INS#APSS#PIK	INSAPSSPIK	3;7	SDRRKINSAPSSPIK,KINSAPSSPIKTNKA	A0AVK6	413;417	SDRRKINSAPSXPIK,KINXAPSSPIKTNKA
3	KINS#APSS#PIK	KINSAPSSPIK	4;8	SDRRKINSAPSSPIK,KINSAPSSPIKTNKA	A0AVK6	413;417	SDRRKINSAPSXPIK,KINXAPSSPIKTNKA
4	KINS#APS#SPIK	KINSAPSSPIK	4;7	SDRRKINSAPSSPIK,RKINSAPSSPIKTNK	A0AVK6	413;416	SDRRKINSAPXSPIK,RKINXAPSSPIKTNK
10	SS#SS#LLASPGHISVK	SSSSLLASPGHISVK	2;4	LHDLGRSSSSLLASP;DLGRSSSSLLASPGH	A0FGR8	737;739	LHDLGRSSSXLLASP;DLGRSXSSLLASPGH
11	SHMSGSPGPGGSNT#APST#PVIGGSDKPGMEEK	SHMSGSPGPGGSNTAPSTPVIGGSDKPGMEEK	14;18	PGPGGSNTAPSTPVI;GSNTAPSTPVIGGSD	A0FGR8	701;705	PGPGGSNTAPSXPVI;GSNXAPSTPVIGGSD
13	SHMS#GSPGPGGSNTAPST#PVIGGSDKPGMEEK	SHMSGSPGPGGSNTAPSTPVIGGSDKPGMEEK	4;18	TSIKSHMSGSPGPGG;GSNTAPSTPVIGGSD	A0FGR8	691;705	TSIKSHMSGSPGPGG;GSNTAPSTPVIGGSD
14	S#S#SSLLASPGHISVK	SSSSLLASPGHISVK	1;2	GLHDLGRSSSSLLAS;LHDLGRSSSSLLASP	A0FGR8	736;737	GLHDLGRSXSSLLAS;LHDLGRXSSSLLASP
15	S#SS#SLLASPGHISVK	SSSSLLASPGHISVK	1;3	GLHDLGRSSSSLLAS;HDLGRSSSSLLASPG	A0FGR8	736;738	GLHDLGRSSXSLLAS;HDLGRXSSSLLASPG
16	SHMS#GS#PGPGGSNTAPSTPVIGGSDKPGMEEK	SHMSGSPGPGGSNTAPSTPVIGGSDKPGMEEK	4;6	TSIKSHMSGSPGPGG;IKSHMSGSPGPGGSN	A0FGR8	691;693	TSIKSHMSGXPGPGG;IKSHMXGSPGPGGSN
17	SHMSGSPGPGGS#NTAPST#PVIGGSDKPGMEEK	SHMSGSPGPGGSNTAPSTPVIGGSDKPGMEEK	12;18	GSPGPGGSNTAPSTP;GSNTAPSTPVIGGSD	A0FGR8	699;705	GSPGPGGSNTAPSXP;GXNTAPSTPVIGGSD

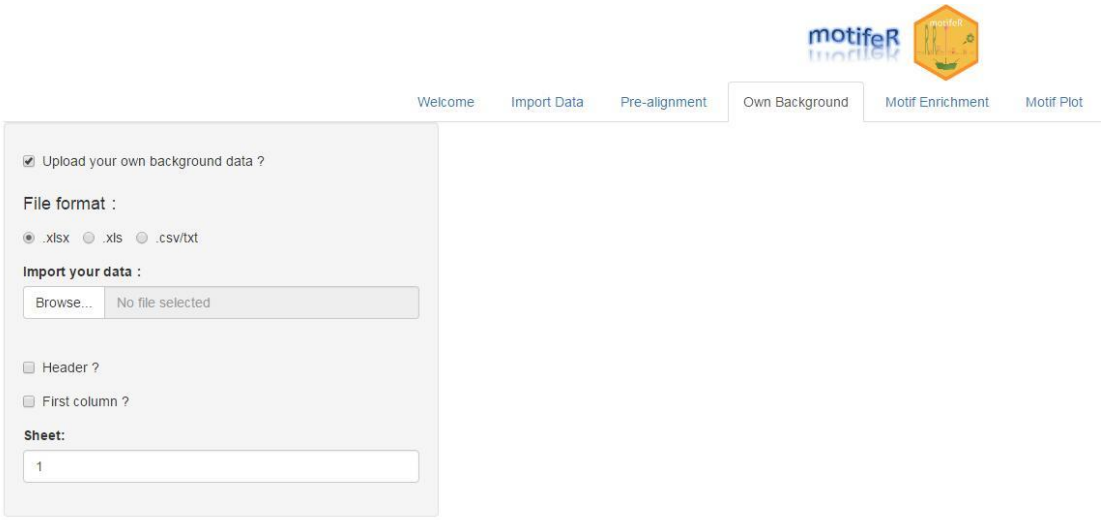
Showing 1 to 10 of 916 entries

Previous 1 2 3 4 5 ... 92 Next

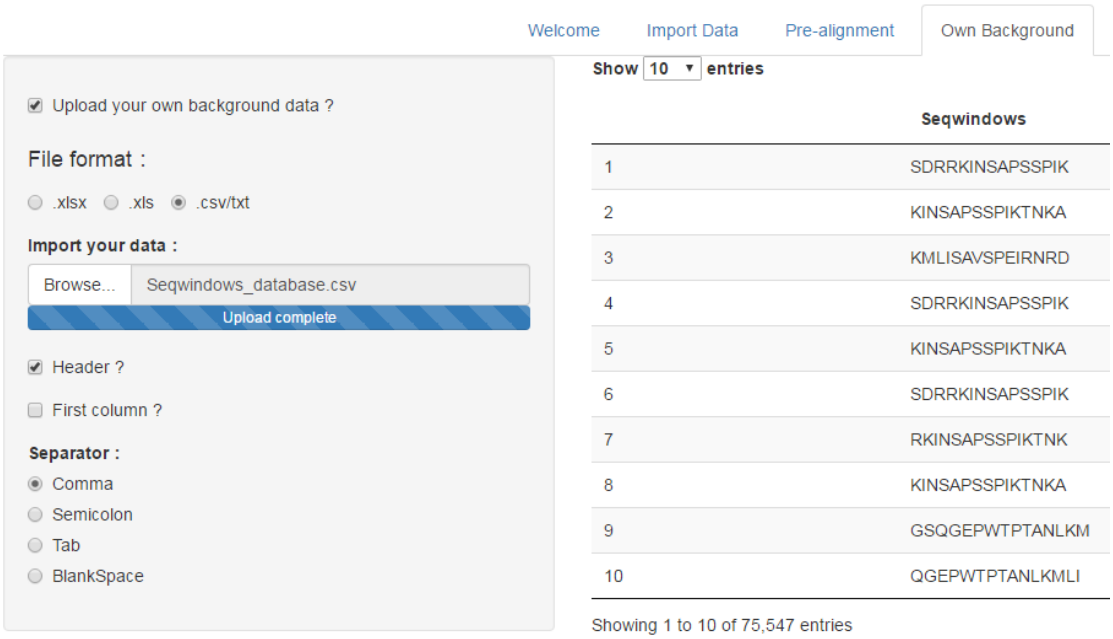
Seqwindows_MultiSites: the modified amino acid will be replaced with “X” if it is not the central residue, for example, ‘YES#DEDS#LGSSGR’, when the 3th amino acid is considered as central residue, the 7th amino acid will be replaced with “X”, thus the standard sequence is ‘YSNRKYESDEDXLGS’, conversely, the standard sequence should be ‘KYEXDEDSLGSSGRV’.

4. Own Background

Users can upload their own background database, but it is noteworthy that the database must contain peptide sequences with standard length, not protein sequences.



Users can click “Browse” button and import their data, like this:

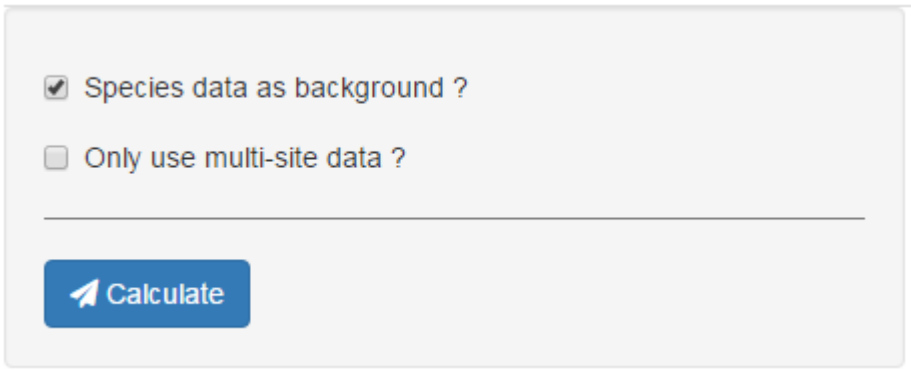


Please note, if you upload you own fasta file as background database, you should unselect this parameter. If you choose this parameter stubbornly, this software will take the data in this step as background database and ignore that you upload before.

5. Motif Enrichment

This step will find overrepresented sequence motifs.

5.1 Parameters

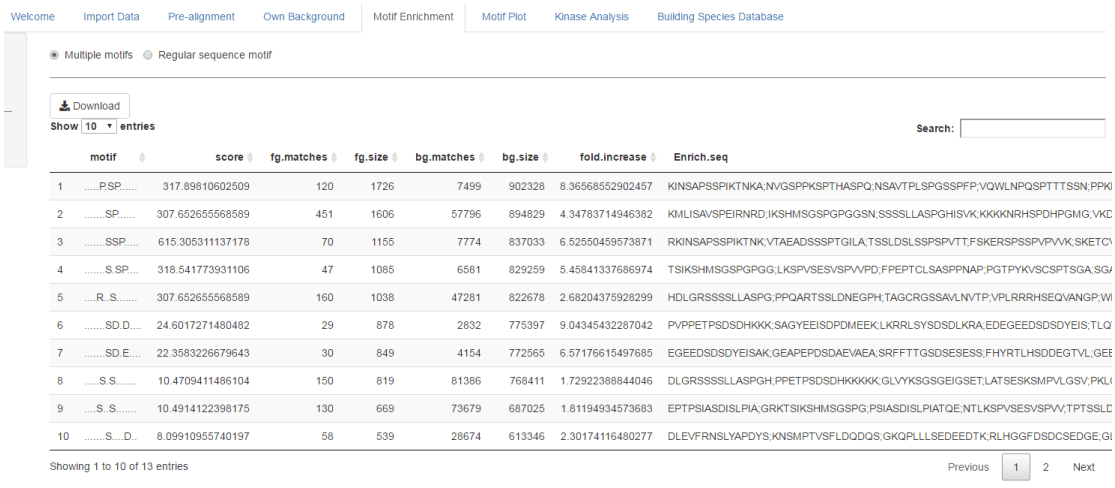


Species data as background: if you upload your own fasta file as background database in the 'Import Data' step, you can ignore this parameter (select or unselect is same). Otherwise, if you choose the database in our system in the 'Import Data' step, selecting this parameter means this software will take the database in our system as background database. If you don't choose, the software will take the foreground data as background database.

Only use multi-site data: if selected, this tool will only take the peptides with multi modification sites as foreground data, that is, it will use the sequences in the *Seqwindows_MultiSites* column obtain from 'Pre-alignment' step as foreground data.

5.2 Results

The enrichment results like this:



motif: the overrepresented motif.

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.

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

fg.size: total number of foreground sequences

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

bg.size: total number of background sequences

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).

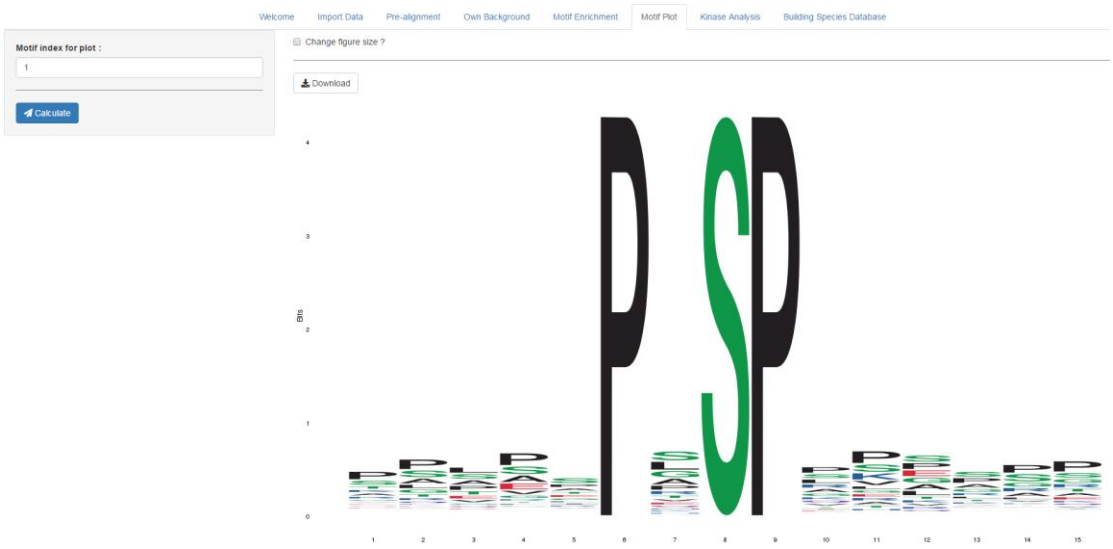
Enrich.seq: those peptides are overrepresented in this motif.

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

6. Motif Plot

This step will plot the motifs from the enrichment results.

If users only input one number in the '*Motif index for plot*' parameter, it will plot the relative motif, shown as below:



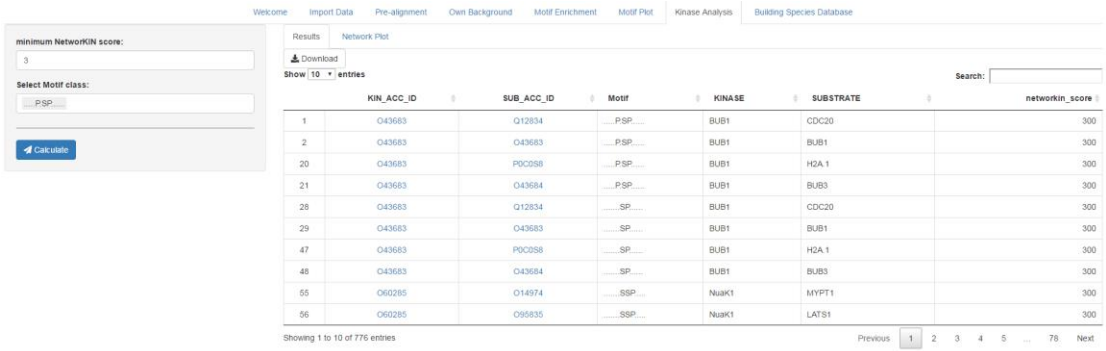
If users type in '1-10', it will plot the 1th to 10th motifs, like this:

7. Kinase-Subtrate Analysis

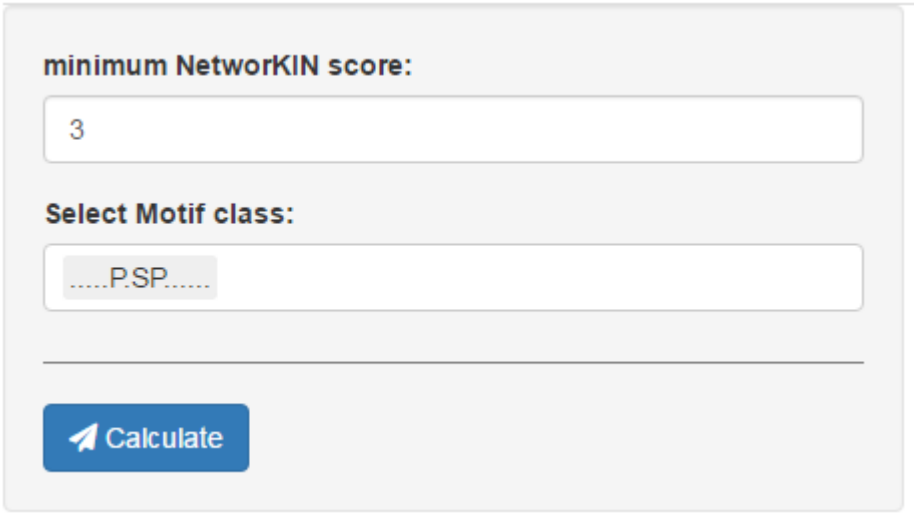
This step will take a formatted phoshoproteomics data input and perform kinase-substrate analysis calculations to infer relative kinase activities. Users should note here:

a. There are only Human database for this analysis in this system at present.

b. This is only for phoshoproteomics data, other modification data are not inappropriate here.

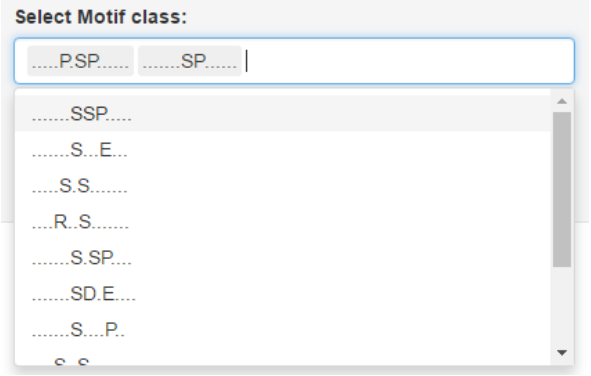


7.1 Parameters



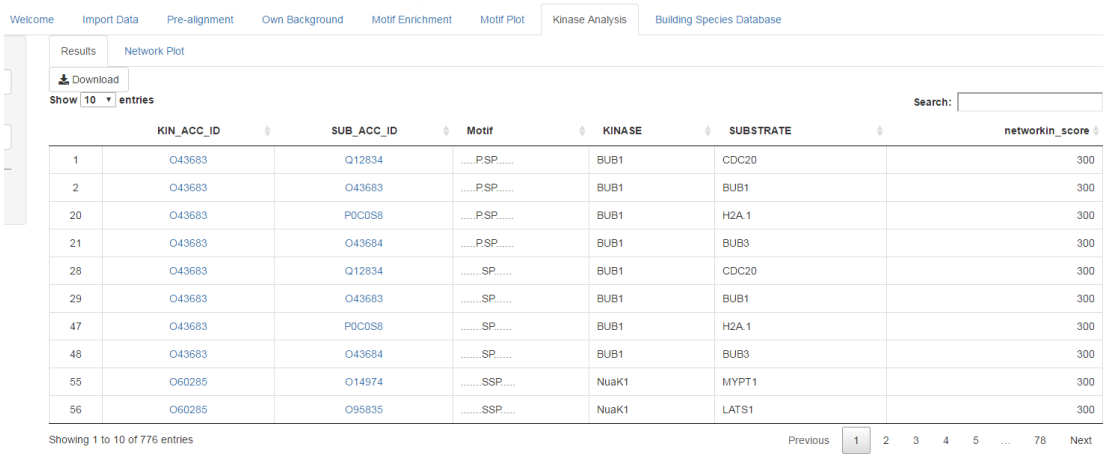
minimum NetworKIN score: a numeric value between 1 and infinity setting the minimum NetworKIN score.

Select Motif class: select those motifs you want to analyze, these motifs are obtained from 'Motif Enrichment' step:



7.2 Results

First, this software give a result table containing kinases and substrates information.



	KIN_ACC_ID	SUB_ACC_ID	Motif	KINASE	SUBSTRATE	networkin_score
1	O43683	Q12834	.....P.SP.....	BUB1	CDC20	300
2	O43683	O43683	.....P.SP.....	BUB1	BUB1	300
20	O43683	P0C0S8	.....P.SP.....	BUB1	H2A.1	300
21	O43683	O43684	.....P.SP.....	BUB1	BUB3	300
28	O43683	Q12834	.....SP.....	BUB1	CDC20	300
29	O43683	O43683	.....SP.....	BUB1	BUB1	300
47	O43683	P0C0S8	.....SP.....	BUB1	H2A.1	300
48	O43683	O43684	.....SP.....	BUB1	BUB3	300
55	O60285	O14974	.....SSP.....	NuaK1	MYPT1	300
56	O60285	O95835	.....SSP.....	NuaK1	LATS1	300

KIN_ACC_ID: kinase uniprot id.

SUB_ACC_ID: substrate uniprot id.

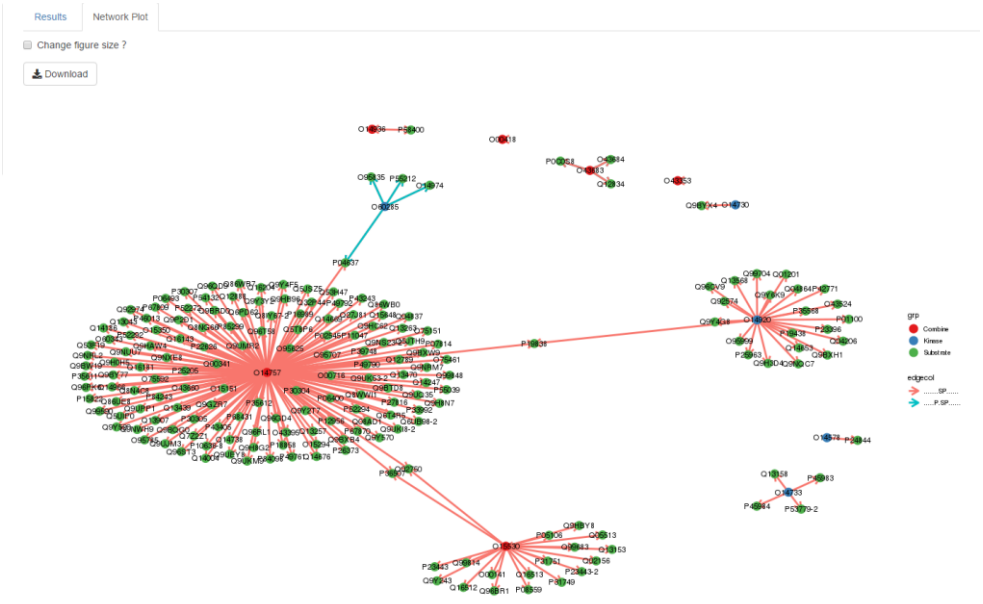
Motif: the overrepresented motif.

KINASE: kinase gene name.

SUBSTRATE: substrate name.

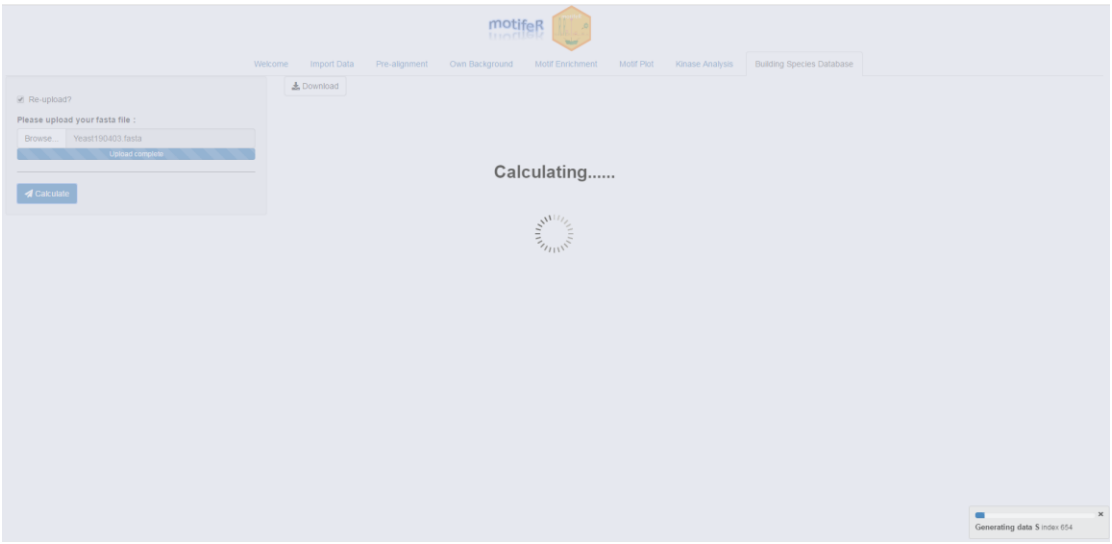
networkin_score: the prediction score from networKIN (https://networkin.info/).

Second, this tool will plot the kinase-substrate network:

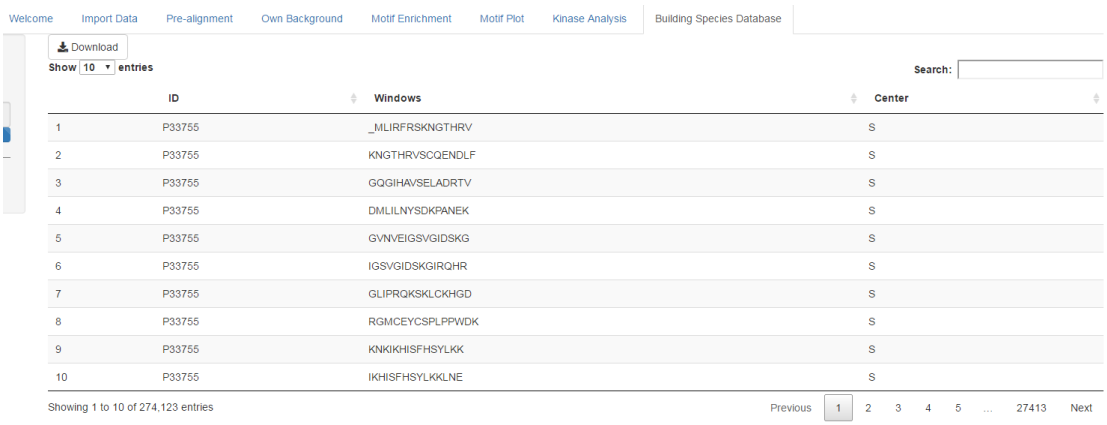


8. Building Species Database

This step can build the standard database based on the fata file that users upload, herein there is no species limit. And this results can also be used in ‘Own background’ step.



If users want to build their own database, they can select the ‘Re-upload’ parameter and then upload a fasta file, the results will be shown as below:



ID: uniprot ids.

Windows: the standard peptides.

Center: Center residue.

Those basic parameters in ‘Import Data’ step are also usable here. Therefore, if you want to get different results, just change those parameters.

All results can be saved to .pdf or .csv files by clicking corresponding “Download” button.