Supporting Information for:

motifeR: An integrated web software for identification and visualization of protein post-translational modification motifs

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

† 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: yanghao@scu.edu.cn; yansheng.liu@yale.edu.

Table of Contents

I. Supplementary Figures

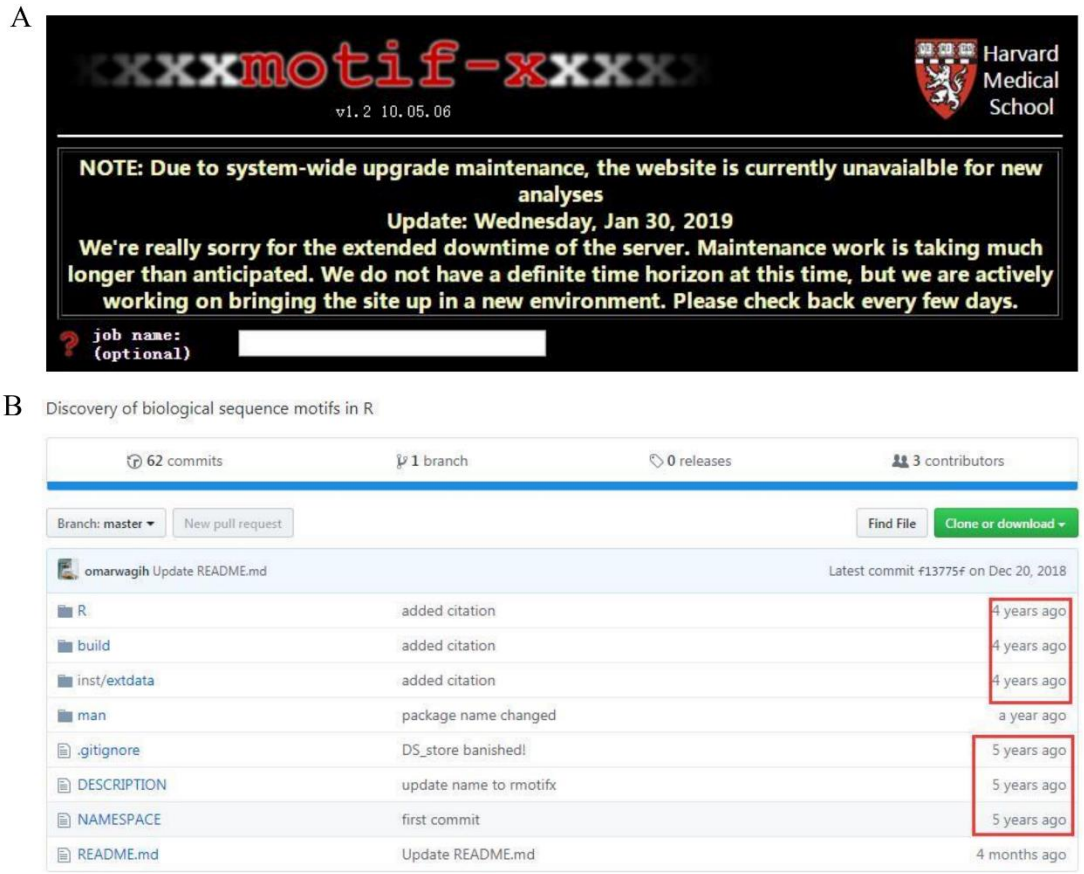


Figure S1. The current status of some published tools. A. motif-x has been in maintenance for a while and users don't know when it can be back. (Link: http://motif-x.med.harvard.edu/motif-x.html). B. rmotifx package has not been updated in a long time. (Link: https://github.com/omarwagih/rmotifx/).

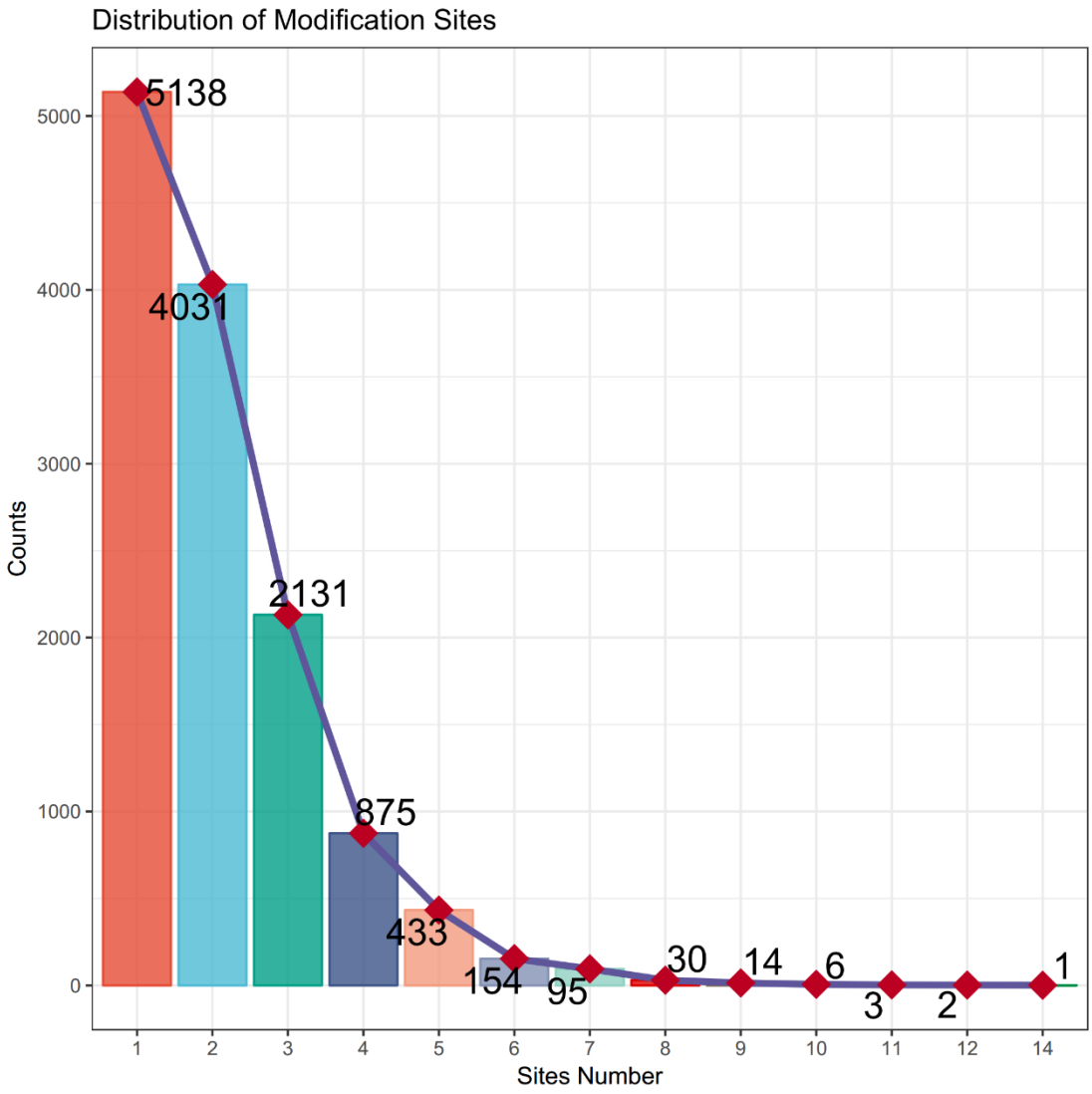


Figure S2. Distribution of original peptide modification site number. This plot can be obtained in "Pre-alignment" module. (This example data from Emdal *et al*.[1])

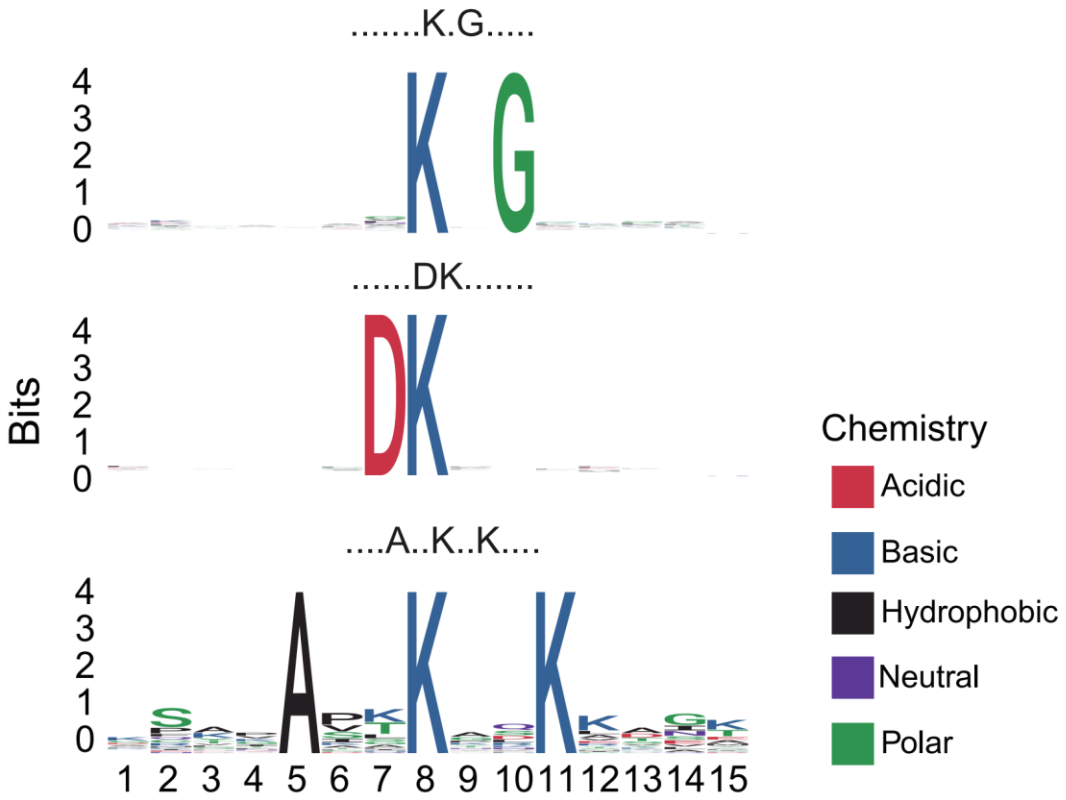


Figure S3. Acetylated peptide motif visualization after enrichment (data from Gu *et al*.[2])

II. 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

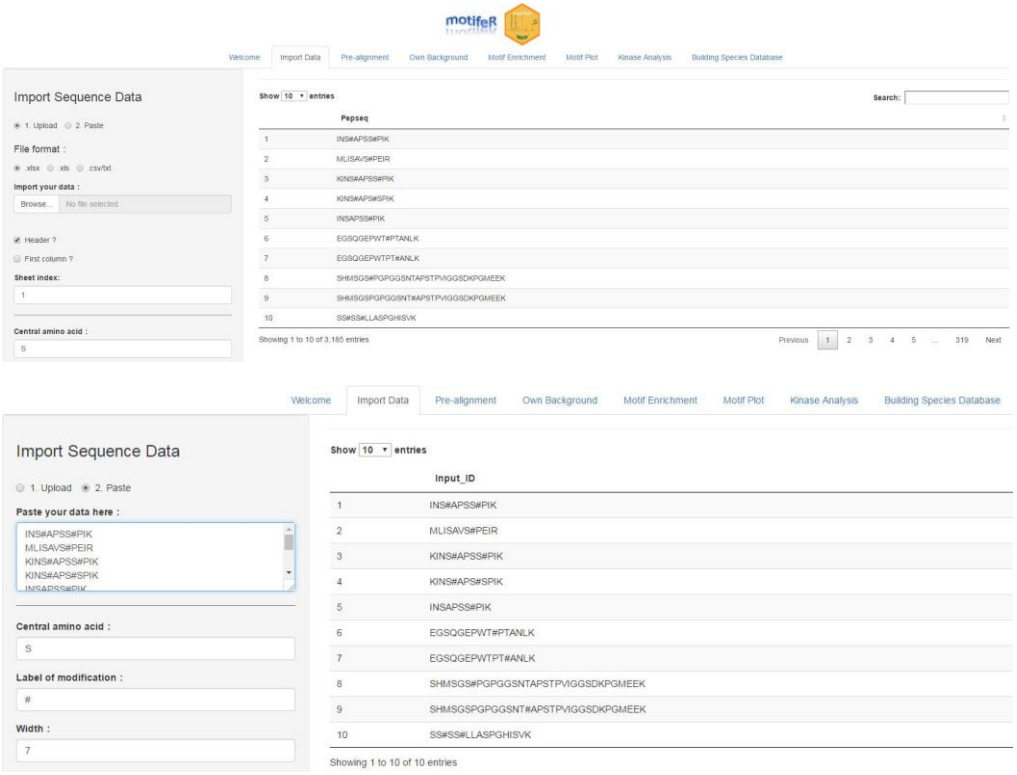
The length of these sequences can be different, and the standard sequences can be obtained in the pre-alignment step. On the other hand, users should mark those modified residues with some label they like (such as “#” or “@”).

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 may 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:

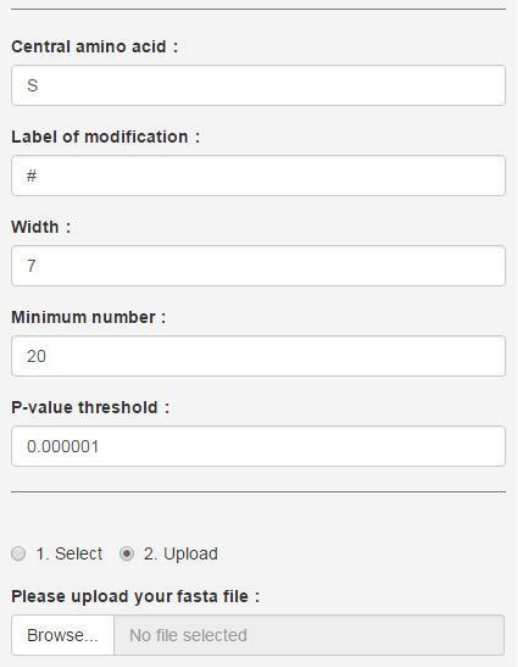
1. Upload, users choose the right format and then click “Browse” button to import there data;

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 “#”, “@”, where “#” is recommended. Here is an example:

Width: it is the number of left/right side characters of the central residue. The default is “7” but can be changed by the user. (The N and C terminal sequences, when not sufficient, will be replaced by “_” as MaxQuant).

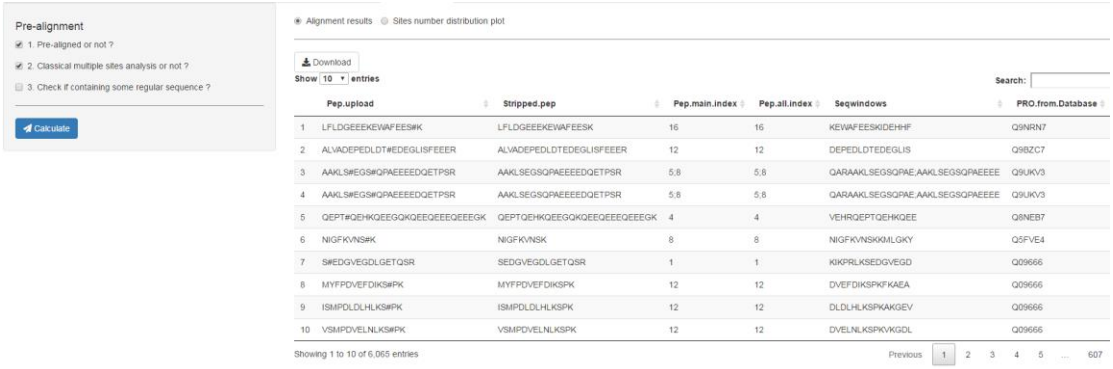
Minimum number: this threshold refers to the minimum number of times you wish each of your extracted motifs to occur in the data set.

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

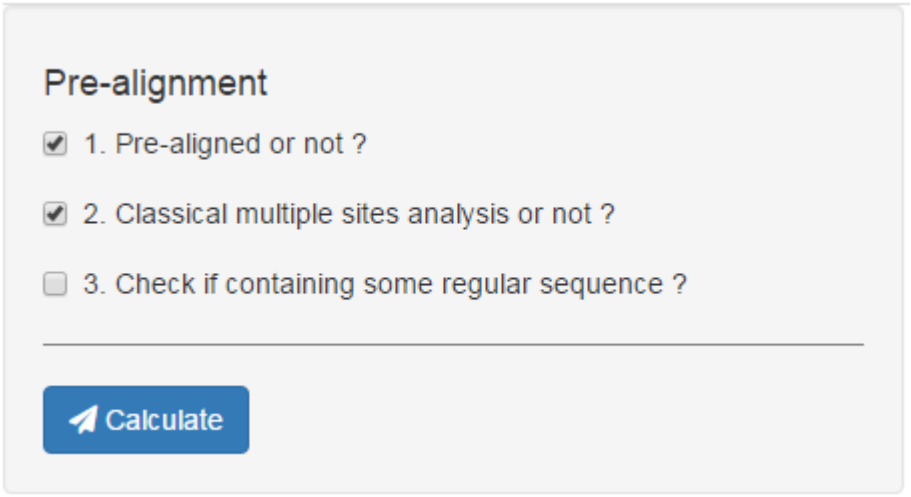
Select or Upload fasta file as background data set: if users want to use the default database, they may just select relative species. Please note, the default database only have Human with 15 length sequences now, more species data will be implemented in the future. Optionally, users can also upload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

3. Pre-alignment

This step aligns 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

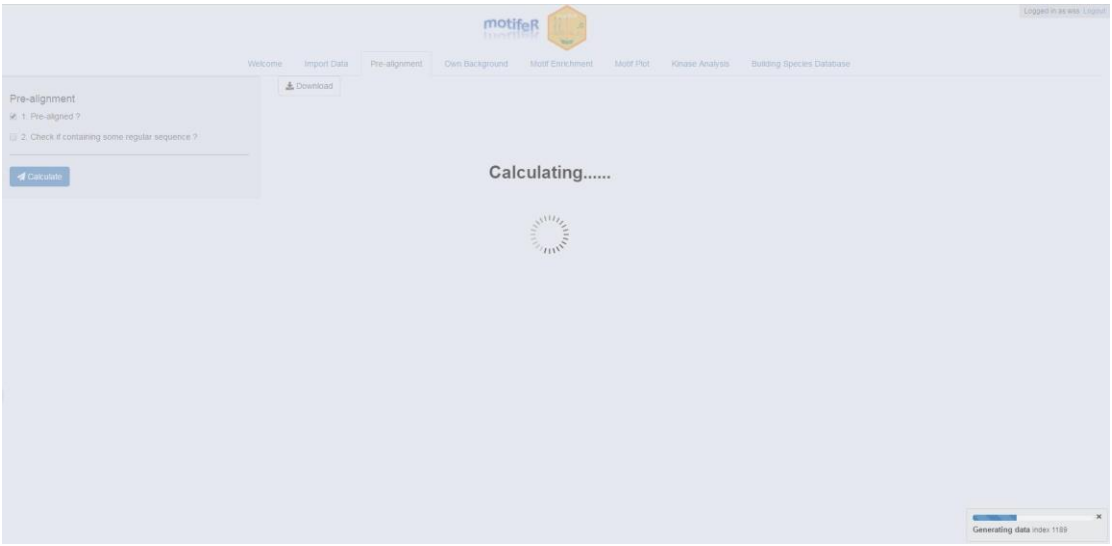


Pre-aligned or not: ask users whether to pre-align their sequences, if your sequences are standard (e.g. 15 length amino acids), you can unselect this parameter. Default is true (i.e., start from tryptic peptide sequences from proteomic database search results).

Classical multiple sites analysis or not: ask users whether to process classical analysis. Classical analysis means not replacing the other modified sites with letter "Z" after pre-alignment, for example 'TSLWNPT#Y#GSWFTEK' to 'TSLWNPTYGSWFTEK', not to 'TSLWNPZYGSWFTEK'. If true, do not process transformation, otherwise, transformation.

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.

Stripped.pep: the peptide skeleton.

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

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

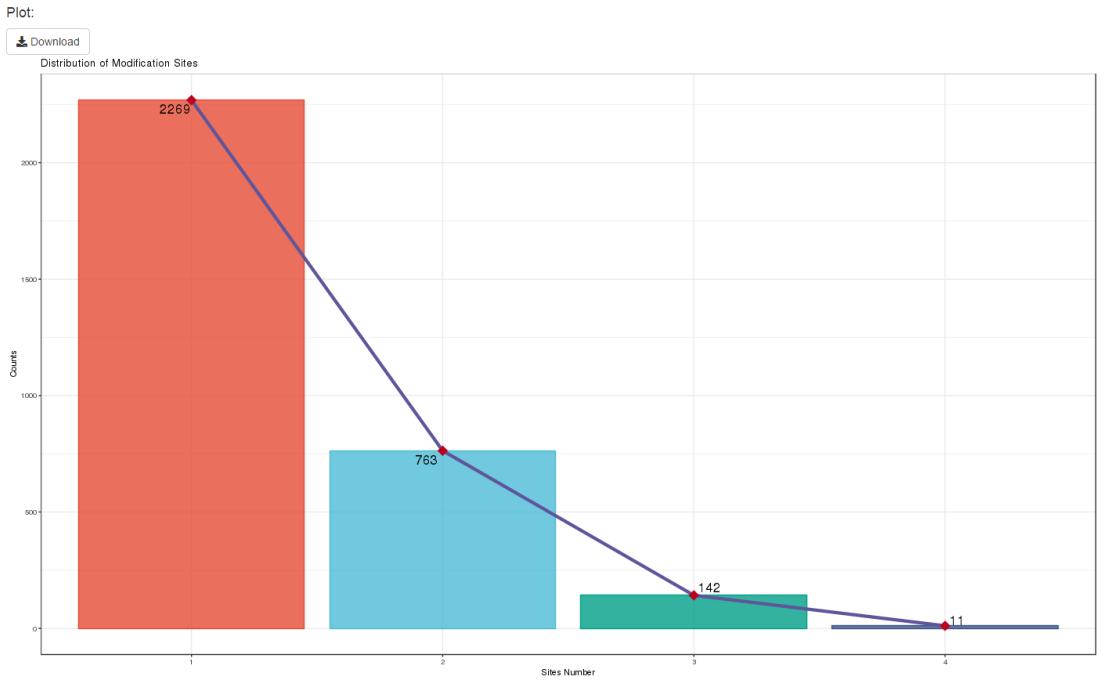
PRO.from.Database: provide the protein name containing this peptide from the fasta file the user uploaded.

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

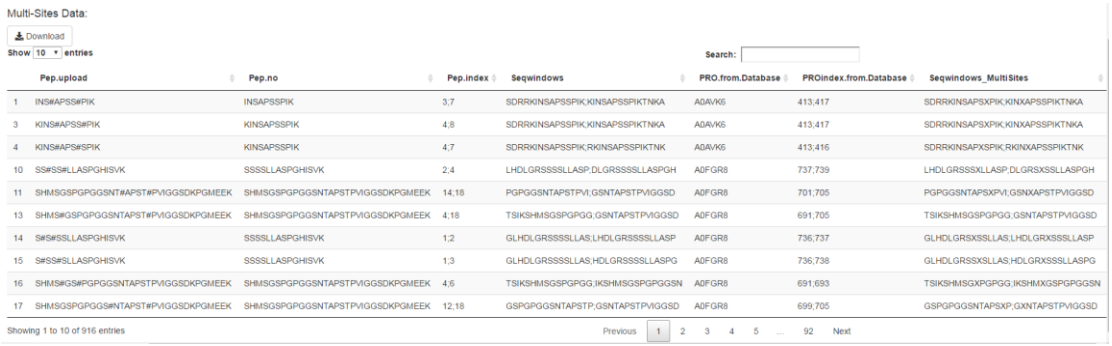
Contain.if: whether containing the sequences that match the regular expression (see above), if true, marked with “Yes”, otherwise, “No”. This column only appears when users choose the parameter--- *Check if containing some regular sequence.*

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 so that the user can choose to perform additional analysis to them or just treat them in the conventional manner as previous tools (the result is similar to that from pre-alignment except the last column).



	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

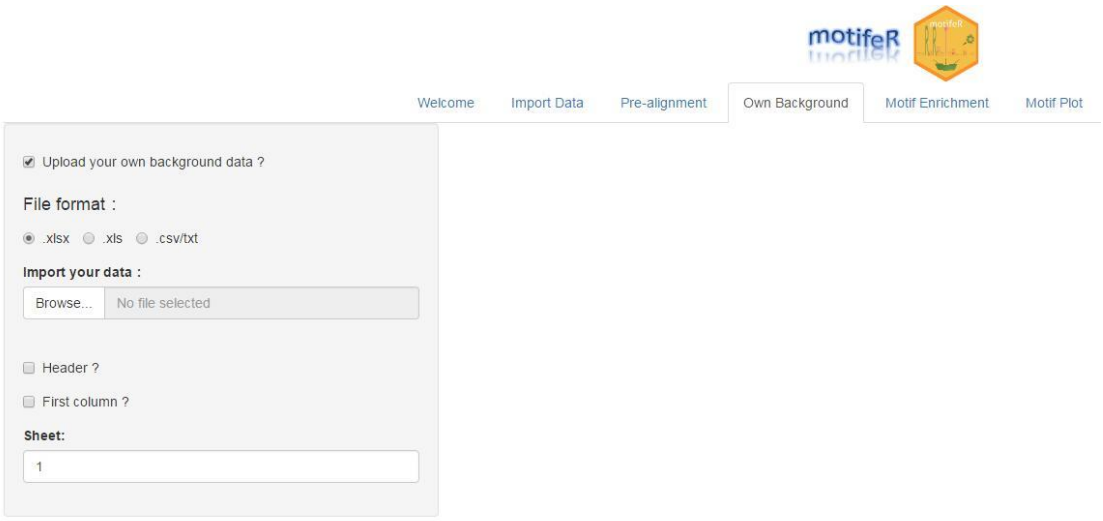
Showing 1 to 10 of 916 entries Previous 1 2 3 4 5 ... 92 Next

Seqwindows_MultiSites: there are two situations here: First, the modified amino acid will be replaced with "X" if it is not the central residue, for example, 'NKPTSLWNPT(0.832)Y(0.168)GSWFTEK' has two phosphosites, one is the 10th amino acid with 0.832 location probability, the other is the 11th amino acid with 0.168 location probability, thus if we transform it like 'NKPTSLWNPT#Y@GSWFTEK' (high probability is replaced with '#', while low probability is replaced with '@'). Then in motifeR, the 10th amino acid will be considered as central residue, the 11th amino acid will be replaced with "X", thus the standard sequence is 'PTSLWNPTYGSWFTE', correspondingly, the Seqwindows_MultiSites should be 'PTSLWNPTXGSWFTE'. Second, if we transform this peptide like 'NKPTSLWNPT#Y#GSWFTEK', the

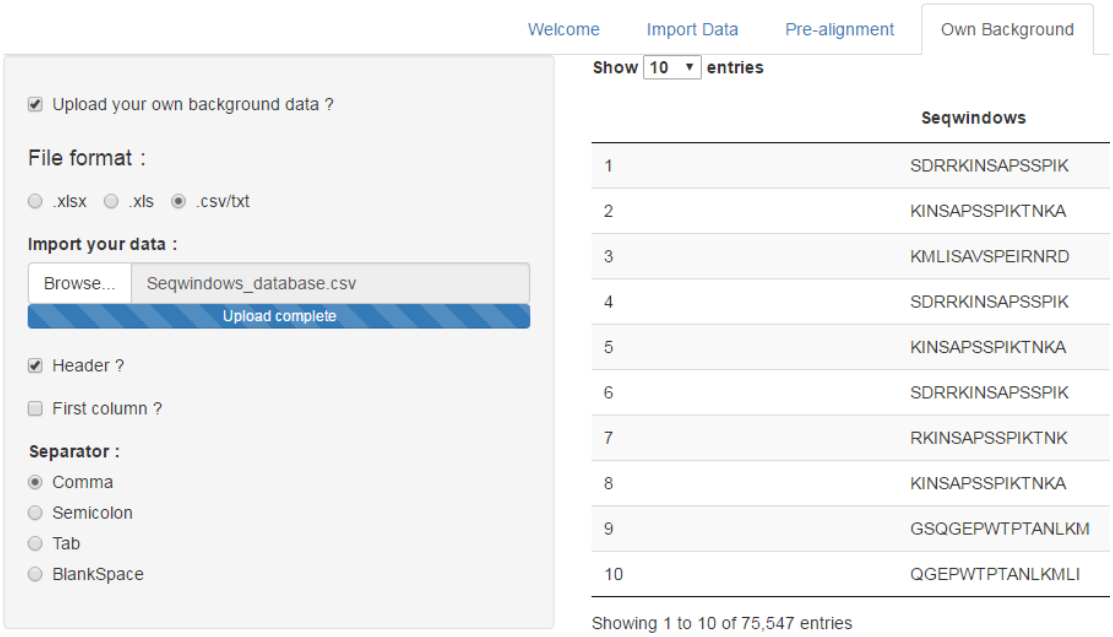
two amino acids will be both considered as central residue, thus the standard sequence is ‘PTSLWNPTYGSWFTE;TSLWNPTYGSWFTEK’, correspondingly, the Seqwindows_MultiSites is still ‘PTSLWNPTYGSWFTE;TSLWNPTYGSWFTEK’.

4. Own Background

Users can upload their own background database, but it is noteworthy that the database must contain peptide sequences with standard length (for example, peptide generated in previous steps by motifeR), not protein sequences.



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

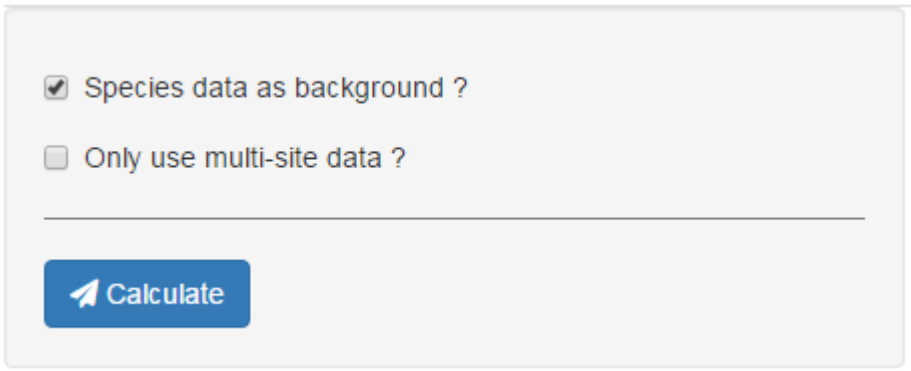


Please note, if you upload you own fasta file as background database, you should unselect this parameter. If you choose this parameter stubbornly, this software will take the data in this step as background database and ignore the fasta file that you upload before.

5. Motif Enrichment

This step will find overrepresented sequence motifs.

5.1 Parameters

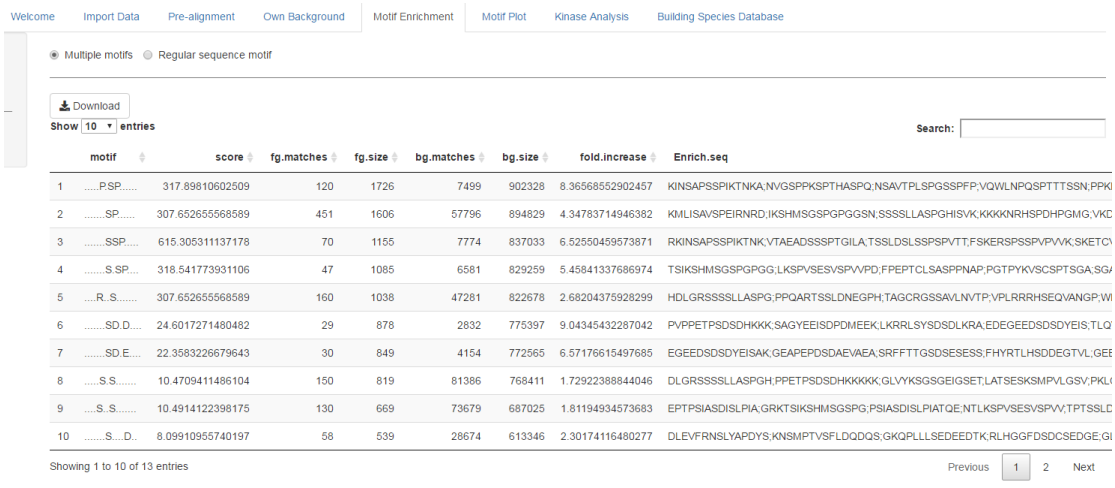


Species data as background: if you upload your own fasta file as background database in the 'Import Data' step, you can ignore this parameter (select or unselect is same). Otherwise, if you choose the database in our system (i.e., human) in the 'Import Data' step, selecting this parameter means this 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.

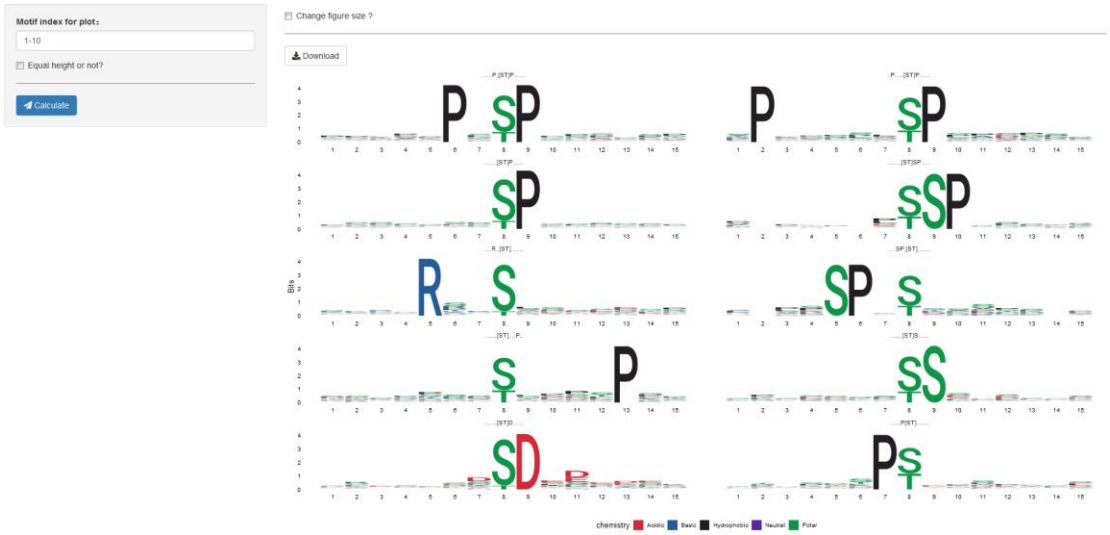
Enrich.pro: those proteins in which the peptides exist from Enrich.seq.

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:



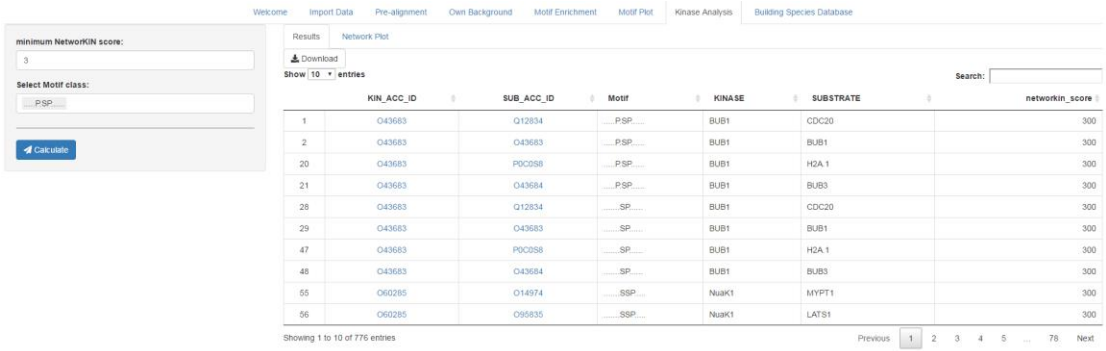
In addition, if users select '*Equal height or not?*' parameter, which means whether all residues in the figure have equal height. This plot will be like:

7. Kinase-Subtrate Analysis

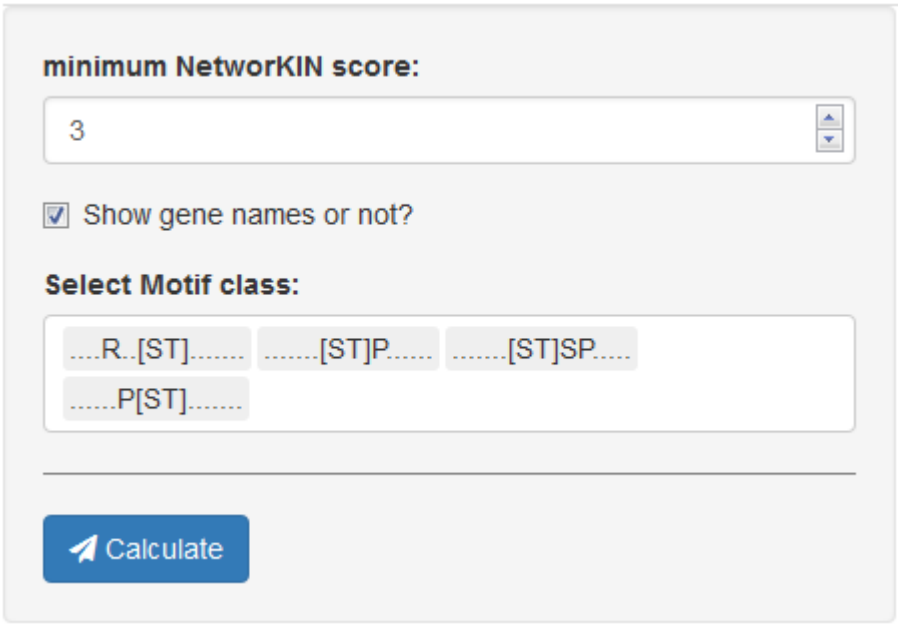
This step will take a formatted phoshoproteomics data input and perform kinase-substrate analysis calculations to infer relative kinase activities. Users should note here:

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

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



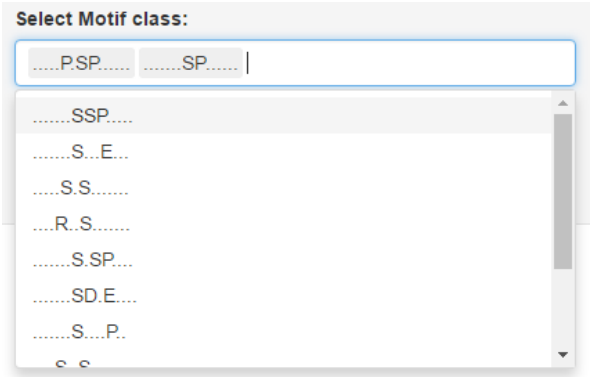
7.1 Parameters



minimum NetworKIN score: a numeric value between 1 and infinity setting the minimum NetworKIN score.[3]

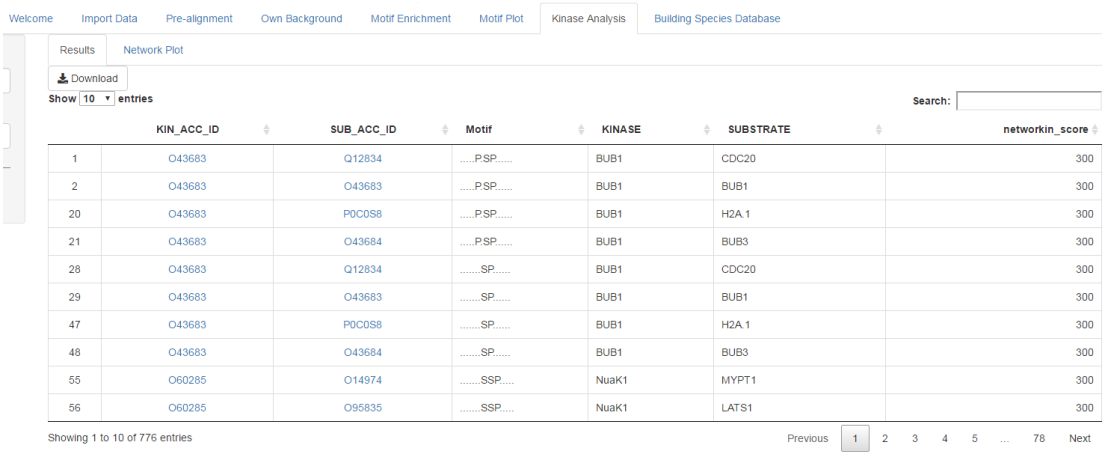
Show gene names or not: if true, the gene names will be appeared in the network plot, otherwise, the uniprot ids will be shown.

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.

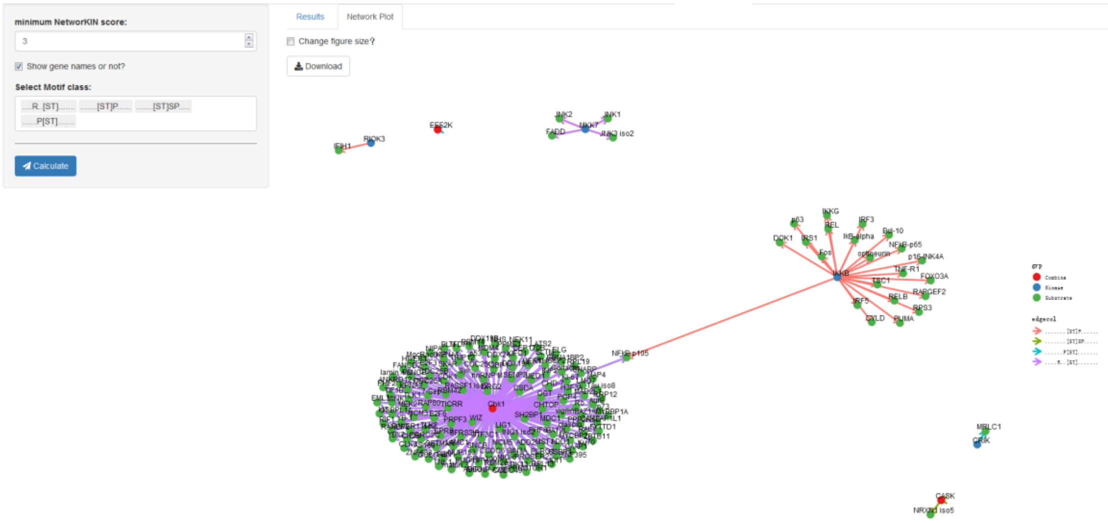
Motif: the overrepresented motif.

KINASE: kinase gene name.

SUBSTRATE: substrate name.

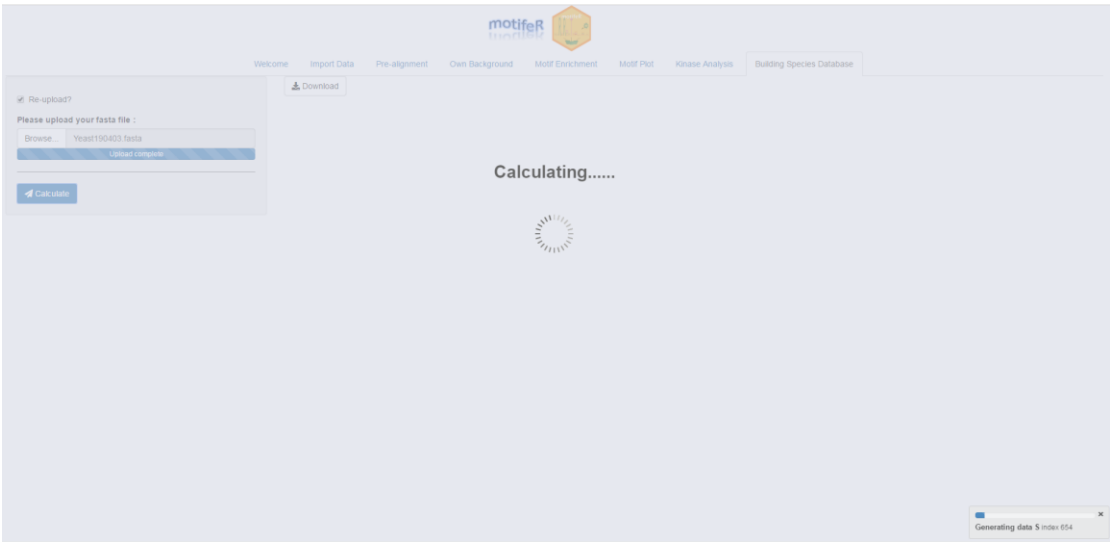
networkin_score: the prediction score from networKIN (https://networkin.info/).

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

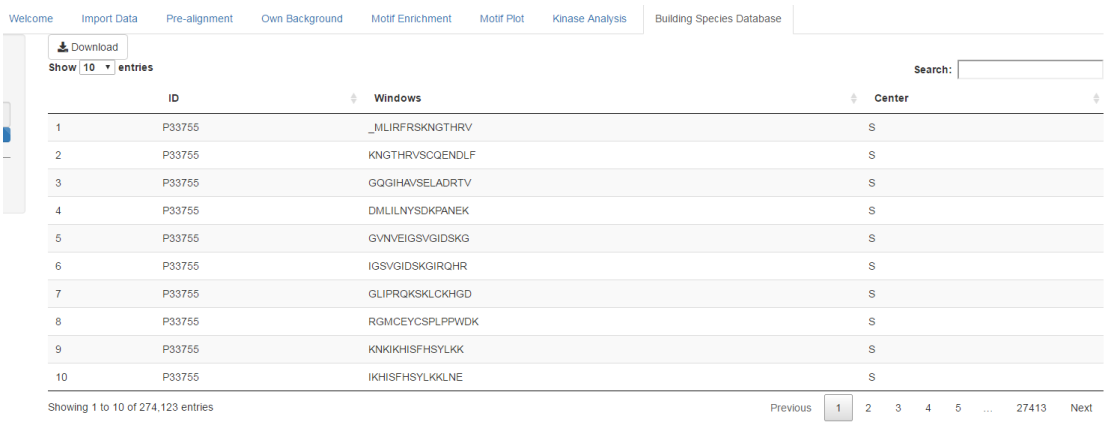

minimum NetworKIN score:
3
Show gene names or not?
Select Motif class:
Calculate
Results
Network Plot
Change figure size?
Download
RIOK3
EEF2K
MKK7
FADD
JNK1
JNK2 iso2
IKKB
IKKG
IRF3
DOK1
Bcl-10
NFkB-p65
FOXO3A
RAPGEF2
RPS3
RELB
CYLD
PUMA
NFkB-p105
Chk1
MYLC1
CRIK
CASK
NRXN1 iso5

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.

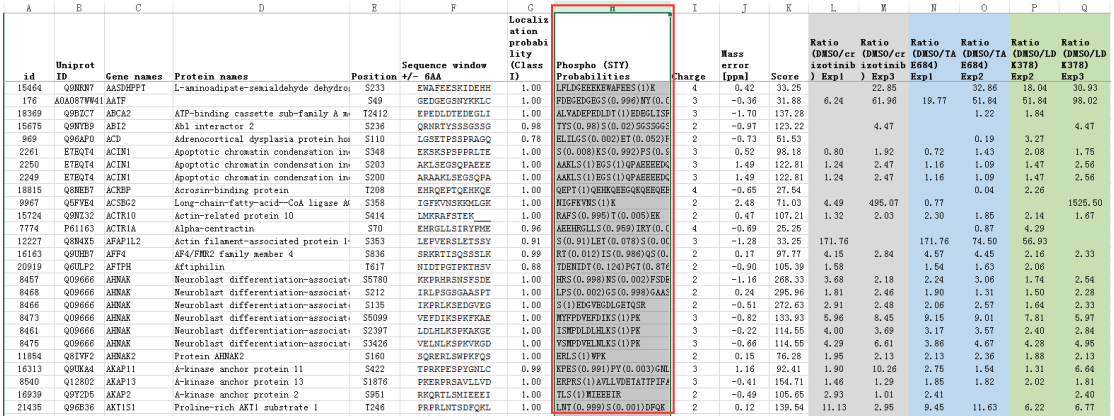
III. Case Study

The detailed parameters introduction can be found in supplementary notes above. Herein we mainly analyze a published data as a real example for users to operate this software better.

Step 1, data preparation

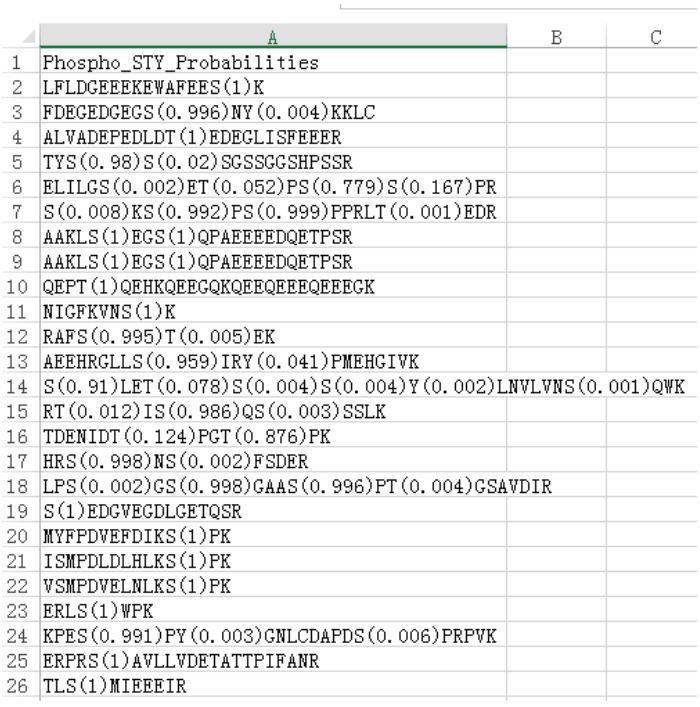
Here we just take a recently published data as an example, users can prepare their own data in this way. The example data can be downloaded from here (doi: 10.1126/scisignal.aap9752)[1]: https://stke.sciencemag.org/highwire/filestream/214464/field_highwire_adjunct_files/2/aap9752_Data_File_S3.xlsx

Download this data and then open it in Excel, the data is like this:



id	Uniprot ID	Gene names	Protein names	Position	Sequence window +/- 6AA	Localization probability (Class I)	Phospho (STY) Probabilities	Charge	Mass error [ppm]	Score	Ratio (DMSO/crizotinib) Exp1	Ratio (DMSO/crizotinib) Exp3	Ratio (DMSO/TAE684) Exp1	Ratio (DMSO/TAE684) Exp2	Ratio (DMSO/LDK378) Exp2	Ratio (DMSO/LDK378) Exp3
15464	Q9NRN7	AASDHPPT	L-aminoadipate-semialdehyde dehydro	S233	EWAFEESKIDEHH	1.00	LFLDGEEEKEWAFEES(1)K	4	0.42	33.25		22.85		32.86	18.04	30.93
176	A0A087WW41	AATF		S49	GEDGEGSNYKKLC	1.00	FDEGEDGEGS(0.996)NY(0.0	3	-0.36	31.88	6.24	61.96	19.77	51.84	51.84	98.02
18369	Q9BZC7	ABCA2	ATP-binding cassette sub-family A m	T2412	EPEDLDTEDEGLI	1.00	ALVADEPEDLDT(1)EDEGLISF	3	-1.70	137.28				1.22	1.84	
15675	Q9NYB9	ABI2	Abl interactor 2	S236	QRNRTYSSSGSSG	0.98	TYS(0.98)S(0.02)SGSSGGS	2	-0.97	123.22		4.47				4.47
969	Q96AP0	ACD	Adrenocortical dysplasia protein ho	S110	LGSETPSSPRAGQ	0.78	ELILGS(0.002)ET(0.052)F	2	-0.73	51.53				0.19	3.27	
2261	E7EQT4	ACIN1	Apoptotic chromatin condensation in	S348	EKSKSPSPPRLTE	1.00	S(0.008)KS(0.992)PS(0.9	3	0.52	98.18	0.80	1.92	0.72	1.43	2.08	1.75
2250	E7EQT4	ACIN1	Apoptotic chromatin condensation in	S203	AKLSEGSQPAEEE	1.00	AAKLS(1)EGS(1)QPAEEEEDQ	3	1.49	122.81	1.24	2.47	1.16	1.09	1.47	2.56
2249	E7EQT4	ACIN1	Apoptotic chromatin condensation in	S200	ARAAKLSEGSQPA	1.00	AAKLS(1)EGS(1)QPAEEEEDQ	3	1.49	122.81	1.24	2.47	1.16	1.09	1.47	2.56
18815	Q8NEB7	ACRBP	Acrosin-binding protein	T208	EHRQEPTQEHKQE	1.00	QEPT(1)QEHKQEEGQKQEEQEE	4	-0.65	27.54				0.04	2.26	
9967	Q5FVE4	ACSBG2	Long-chain-fatty-acid--CoA ligase A	S358	IGFKVNSKKMLGK	1.00	NIGFKVNS(1)K	2	2.48	71.03	4.49	495.07	0.77			1525.50
15724	Q9NZ32	ACTR10	Actin-related protein 10	S414	LMKRAFSTEK___	1.00	RAFS(0.995)T(0.005)EK	2	0.47	107.21	1.32	2.03	2.30	1.85	2.14	1.67
7774	P61163	ACTR1A	Alpha-centractin	S70	EHRGLLSIRYPME	0.96	AEEHRGLLS(0.959)IRY(0.0	4	-0.69	25.25				0.87	4.29	
12227	Q8NAX5	AFAP1L2	Actin filament-associated protein 1	S353	LEPVERSLETSSY	0.91	S(0.91)LET(0.078)S(0.00	3	-1.28	33.25	171.76		171.76	74.50	56.93	
16163	Q9UHB7	AFF4	AF4/FMR2 family member 4	S836	SRKRTISQSSSLK	0.99	RT(0.012)IS(0.986)QS(0.	2	0.17	97.77	4.15	2.84	4.57	4.45	2.16	2.33
20919	Q6ULP2	AFTPH	Aftiphilin	T617	NIDTPGTPKTHSV	0.88	TDENIDT(0.124)PGT(0.876	2	-0.90	105.39	1.58		1.54	1.63	2.06	
8457	Q09666	AHNAK	Neuroblast differentiation-associat	S5780	KKPRHRSNSFSDE	1.00	HRS(0.998)NS(0.002)FSDE	2	-1.16	268.33	3.68	2.18	2.24	3.06	1.74	2.54
8468	Q09666	AHNAK	Neuroblast differentiation-associat	S212	IRLPSGSGAASPT	1.00	LPS(0.002)GS(0.998)GAAS	2	0.24	295.96	1.81	2.46	1.90	1.31	1.50	2.28
8466	Q09666	AHNAK	Neuroblast differentiation-associat	S135	IKPRLKSEDGVEG	1.00	S(1)EDGVEGDLGETQSR	2	-0.51	272.63	2.91	2.48	2.06	2.57	1.64	2.33
8473	Q09666	AHNAK	Neuroblast differentiation-associat	S5099	VEFDIKSPKFKAE	1.00	MYFPDVEFDIKS(1)PK	3	-0.82	133.93	5.96	8.45	9.15	9.01	7.81	5.97
8461	Q09666	AHNAK	Neuroblast differentiation-associat	S2397	LDLHLKSPKAKGE	1.00	ISMPDLDLHLKS(1)PK	3	-0.22	114.55	4.00	3.69	3.17	3.57	2.40	2.84
8475	Q09666	AHNAK	Neuroblast differentiation-associat	S3426	VELNLKSPKVKGD	1.00	VSMPDVELNLKS(1)PK	3	-0.66	114.55	4.29	6.61	3.86	4.67	4.28	4.95
11854	Q8IVF2	AHNAK2	Protein AHNAK2	S160	SQRERLSWPKFQS	1.00	ERLS(1)WPK	2	0.15	76.28	1.95	2.13	2.13	2.36	1.88	2.13
16313	Q9UKA4	AKAP11	A-kinase anchor protein 11	S422	TPRKPESPYGNLC	0.99	KPES(0.991)PY(0.003)GNL	3	1.16	92.41	1.90	10.26	2.75	1.54	1.31	6.64
8540	Q12802	AKAP13	A-kinase anchor protein 13	S1876	PKERPRSAVLLVD	1.00	ERPRS(1)AVLLVDETATTPIFA	3	-0.41	154.71	1.46	1.29	1.85	1.82	2.02	1.81
16939	Q9Y2D5	AKAP2	A-kinase anchor protein 2	S951	RKQRTLSMIEEEI	1.00	TLS(1)MIEEEIR	2	-0.49	105.65	2.93	1.01	2.41			2.40
21435	Q96B36	AKT1S1	Proline-rich AKT1 substrate 1	T246	PRPRLNTSDFQKL	1.00	LNT(0.999)S(0.001)DFQK	2	0.12	139.54	11.13	2.95	9.45	11.63	6.22	6.77

We can extract the column “Phospho (STY) Probabilities” data in a new file, as this:



	A	B	C
1	Phospho_STY_Probabilities		
2	LFLDGEEEKEWAFEES(1)K		
3	FDEGEDGEGS(0.996)NY(0.004)KKLC		
4	ALVADEPEDLDT(1)EDEGLISFEEER		
5	TYS(0.98)S(0.02)SGSSGGSHPSSR		
6	ELILGS(0.002)ET(0.052)PS(0.779)S(0.167)PR		
7	S(0.008)KS(0.992)PS(0.999)PPRLT(0.001)EDR		
8	AAKLS(1)EGS(1)QPAEEEEDQETPSR		
9	AAKLS(1)EGS(1)QPAEEEEDQETPSR		
10	QEPT(1)QEHKQEEGQKQEEQEEEQEEEGK		
11	NIGFKVNS(1)K		
12	RAFS(0.995)T(0.005)EK		
13	AEEHRGLLS(0.959)IRY(0.041)PMEHGIVK		
14	S(0.91)LET(0.078)S(0.004)S(0.004)Y(0.002)LNVLVNS(0.001)QWK		
15	RT(0.012)IS(0.986)QS(0.003)SSLK		
16	TDENIDT(0.124)PGT(0.876)PK		
17	HRS(0.998)NS(0.002)FSDER		
18	LPS(0.002)GS(0.998)GAAS(0.996)PT(0.004)GSAVDIR		
19	S(1)EDGVEGDLGETQSR		
20	MYFPDVEFDIKS(1)PK		
21	ISMPDLDLHLKS(1)PK		
22	VSMPDVELNLKS(1)PK		
23	ERLS(1)WPK		
24	KPES(0.991)PY(0.003)GNLCDAPDS(0.006)PRPVK		
25	ERPRS(1)AVLLVDETATTPIFANR		
26	TLS(1)MIEEEIR		

Then we need convert these this location probability into a specific label, so that motifeR can recognize. For example, here we use ‘#’ as the specific label:

‘FDEGEDGEGS(0.996)NY(0.004)KKLC’ to ‘FDEGEDGEGS#NY#KKLC’

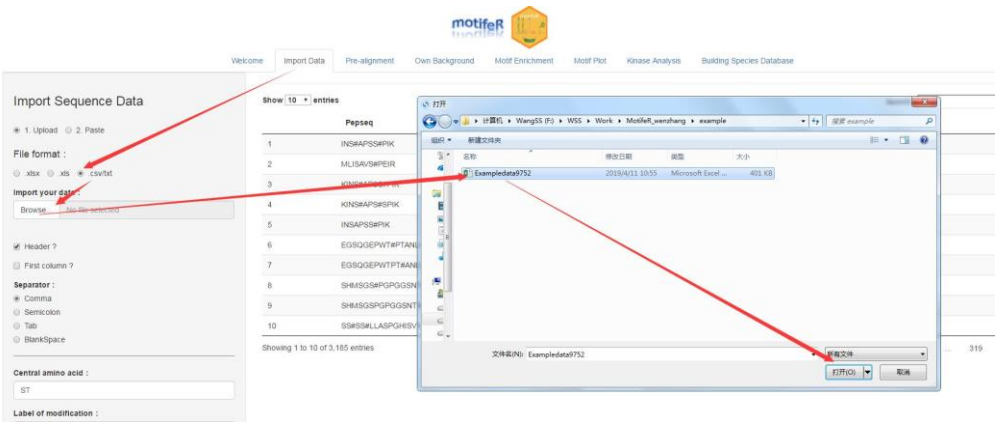
No matter what method you use (e.g. writing some R/Python codes to convert in bulk, or process it manually in Excel, and so on), your final data should be like this:

	A
1	Phospho_STY_Probabilities
2	LFLDGEEEKEWAFEES#K
3	FDEGEDGEGS#NY#KKLC
4	ALVADEPEDLDT#EDEGLISFEEER
5	TYS#S#SGSSGGSHPSSR
6	ELILGS#ET#PS#S#PR
7	S#KS#PS#PPRLT#EDR
8	AAKLS#EGS#QPAEEEEDQETPSR
9	AAKLS#EGS#QPAEEEEDQETPSR
10	QEPT#QEHKQEEGQKQEEQEEEQEEEGK
11	NIGFKVNS#K
12	RAFS#T#EK
13	AEEHRGLLS#IRY#PMEHGIVK
14	S#LET#S#S#Y#LNVLVNS#QWK
15	RT#IS#QS#SSLK
16	TDENIDT#PGT#PK
17	HRS#NS#FSDER
18	LPS#GS#GAAS#PT#GSAVDIR
19	S#EDGVEGDLGETQSR
20	MYFPDVEFDIKS#PK
21	ISMPDLDLHLKS#PK
22	VSMPDVELNLKS#PK
23	ERLS#WPK
24	KPES#PY#GNLCDAPDS#PRPVK
25	ERPRS#AVLLVDETATTPIFANR

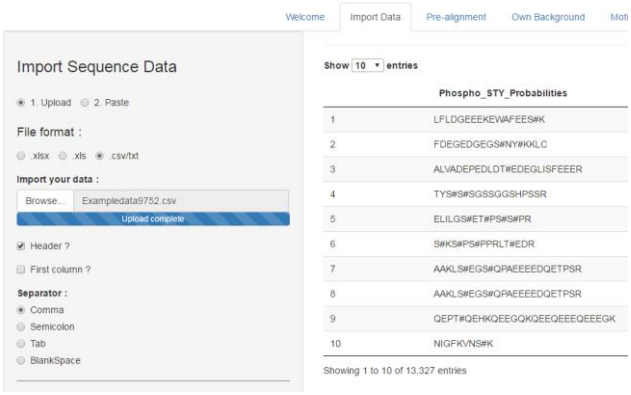
Save this data in a .xlsx, .xls, .csv or .txt file, here we save it in the “Exampledata9752.csv” file.

Step 2, Analysis in motifeR

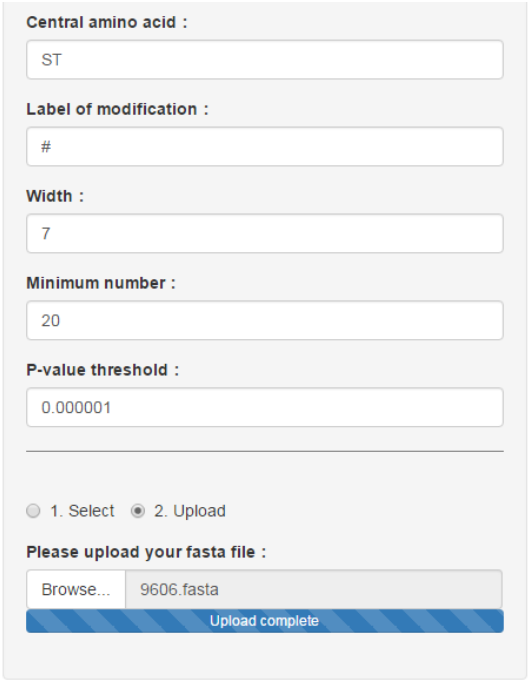
First, open the software: https://www.omicsolution.org/wukong/motifeR. Click “Import Data”, find our data and then import:



Users must set right parameters based on their own data. Now we import the example data:

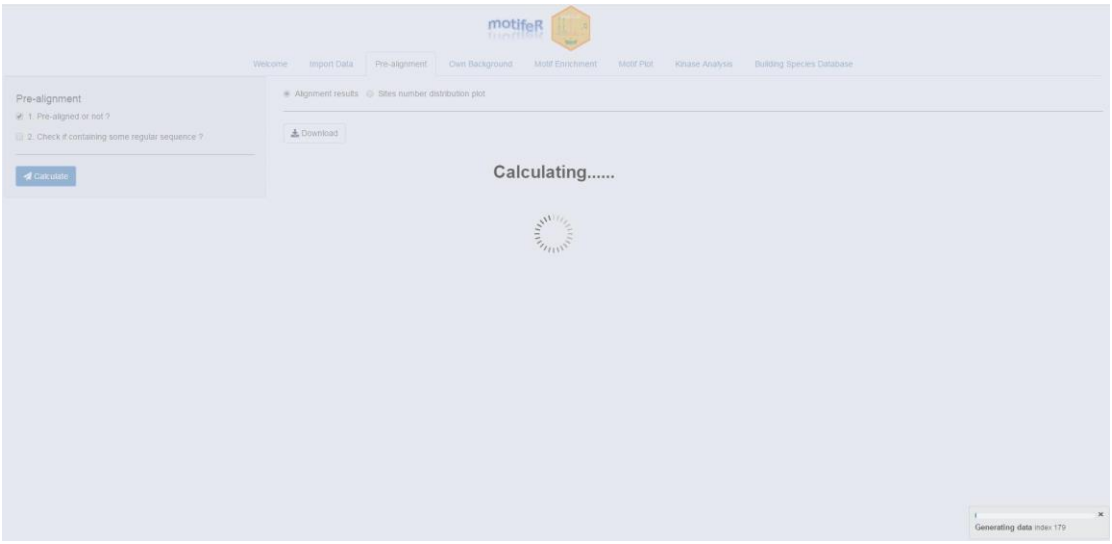


Second, set the basic parameters:

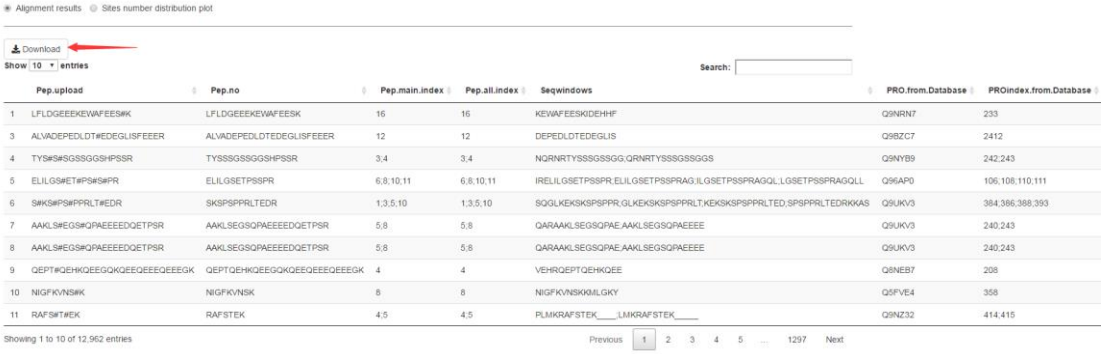


Base on the example data, we just leave basic parameter as default, whereas, we upload a new fasta file as the background database.

Third, Click “Pre-alignment”. Herein we don’t want to check if the example data contain some regular sequences, but the peptides in example data are not standard, we need pre-align them with the background database. Then click “Calculate” button:



After a while (normally in a few seconds), this step can be done. The total calculation time is directly proportional to the data size. We can get the alignment results like this:



	Pep.upload	Pep.no	Pep.main.index	Pep.all.index	Seqwindows	PRO.from.Database	PROindex.from.Database
1	LFLDGEEEKEWAFEES#K	LFLDGEEEKEWAFEESK	16	16	KEWAFEESKIDEHHF	Q9NRN7	233
3	ALVADEPEDLDT#EDEGLISFEEER	ALVADEPEDLDTEDEGLISFEEER	12	12	DEPEDLDTEDEGLIS	Q9BZC7	2412
4	TYS#S#SGSSSGGSHPSSR	TYSSSGSSSGGSHPSSR	3;4	3;4	NQRNRTYSSSGSSGG;QRNRTYSSSGSSGGS	Q9NYB9	242;243
5	ELILGS#ET#PS#S#PR	ELILGSETPSSPR	6;8;10;11	6;8;10;11	IRELILGSETPSSPR;ELILGSETPSSPRAG;ILGSETPSSPRAGQL;LGSETPSSPRAGQLL	Q96AP0	106;108;110;111
6	S#KS#PS#PPRLT#EDR	SKSPSPPRLTEDR	1;3;5;10	1;3;5;10	SQGLKEKSKSPSPPR;GLKEKSKSPSPPRLT;KEKSKSPSPPRLTED;SPSPPRLTEDRKKAS	Q9UKV3	384;386;388;393
7	AAKLS#EGS#QPAEEEEDQETPSR	AAKLSEGSQPAEEEEDQETPSR	5;8	5;8	QARAAKLSEGSQPAE;AAKLSEGSQPAEEEE	Q9UKV3	240;243
8	AAKLS#EGS#QPAEEEEDQETPSR	AAKLSEGSQPAEEEEDQETPSR	5;8	5;8	QARAAKLSEGSQPAE;AAKLSEGSQPAEEEE	Q9UKV3	240;243
9	QEPT#QEHKQEEGQKQEEQEEEQEEEGK	QEPTQEHKQEEGQKQEEQEEEQEEEGK	4	4	VEHRQEPTQEHKQEE	Q8NEB7	208
10	NIGFKVNS#K	NIGFKVNSK	8	8	NIGFKVNSKKMLGKY	Q5FVE4	358
11	RAFS#T#EK	RAFSTEK	4;5	4;5	PLMKRAFSTEK_____LMKRAFSTEK_____	Q9NZ32	414;415

Users can also click the “Download” button to download the results into their own computer. Then we can click “Sites number distribution plot” to obtain multiple sites information:

Multi-Sites Data:

Download

Show 10 entries

Search:

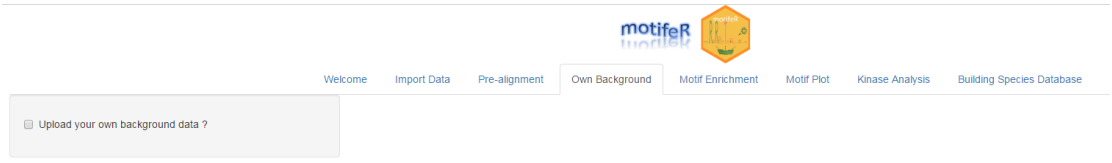
	Pep.upload	Pep.no	Pep.main.index	Pep.all.index	Seqwindows
4	TYS#S#SGSSSGGSHPSSR	TYSSSGSSSGGSHPSSR	3;4	3;4	NQRNRTYSSSGSSSGG;QRNRTYSSSGSSSGGS
5	ELILGS#ET#PS#S#PR	ELILGSETPSSPR	6;8;10;11	6;8;10;11	IRELILGSETPSSPR;ELILGSETPSSPRAG;ILGSETPSSPRAGQL;LGSETPSSPRAGQLL
6	S#KS#PS#PPRLT#EDR	SKSPSPPRLTEDR	1;3;5;10	1;3;5;10	SQGLKEKSKSPSPPR;GLKEKSKSPSPPRLT;KEKSKSPSPPRLTED;SPSPPRLTEDRKKAS
7	AAKLS#EGS#QPAEEEEDQETPSR	AAKLSEGSQPAEEEEDQETPSR	5;8	5;8	QARAAKLSEGSQPAE;AAKLSEGSQPAEEEE
8	AAKLS#EGS#QPAEEEEDQETPSR	AAKLSEGSQPAEEEEDQETPSR	5;8	5;8	QARAAKLSEGSQPAE;AAKLSEGSQPAEEEE
11	RAFS#T#EK	RAFSTEK	4;5	4;5	PLMKRAFSTEK____;LMKRAFSTEK_____
12	AEEHRGLLS#IRY#PMEHGIVK	AEEHRGLLSIRYPMEHGIVK	9;12	9;12	EEHRGLLSIRYPMEH;RGLLSIRYPMEHGIV
13	S#LET#S#S#Y#LNVLVNS#QWK	SLETSSYLNVLVNSQWK	1;4;5;6;7;14	1;4;5;6;7;14	SLEPVERSLETSSYL;PVERSLETSSYLNVL;VERSLETSSYLNVLV;ERSLETSSYLNVLVN;RSLETSSYLNVLVNS;
14	RT#IS#QS#SSLK	RTISQSSSLK	2;4;6	2;4;6	EHGSRKRTISQSSSL;GSRKRTISQSSSLKS;RKRTISQSSSLKSSS
15	TDENIDT#PGT#PK	TDENIDTPGTPK	7;10	7;10	RTDENIDTPGTPKTH;ENIDTPGTPKTHSVP

Showing 1 to 10 of 7,824 entries

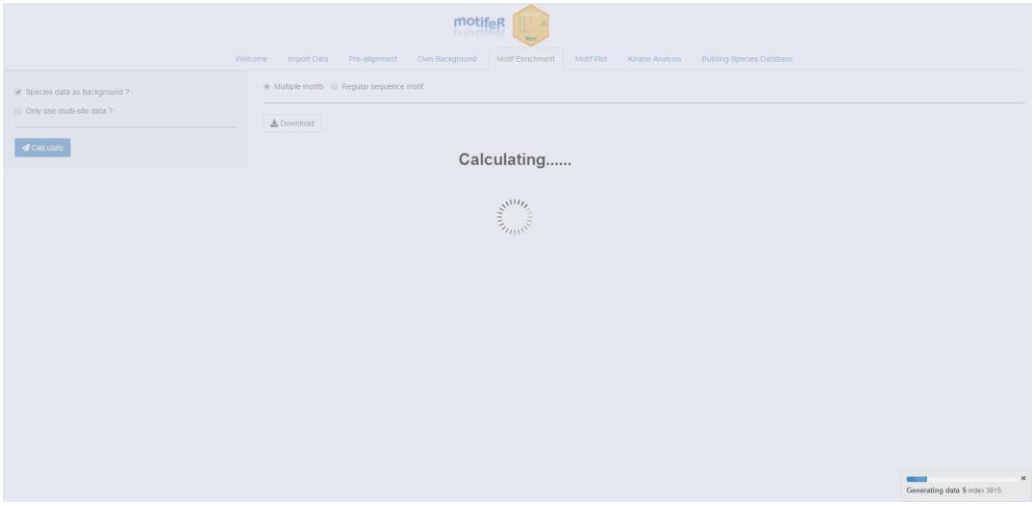
Previous 1 2 3 4 5 … 783 Next

If there are multiple modifications in peptides, this part can give users necessary information. From here we can know there are total 7824 peptides with >2 phosphorylation sites.

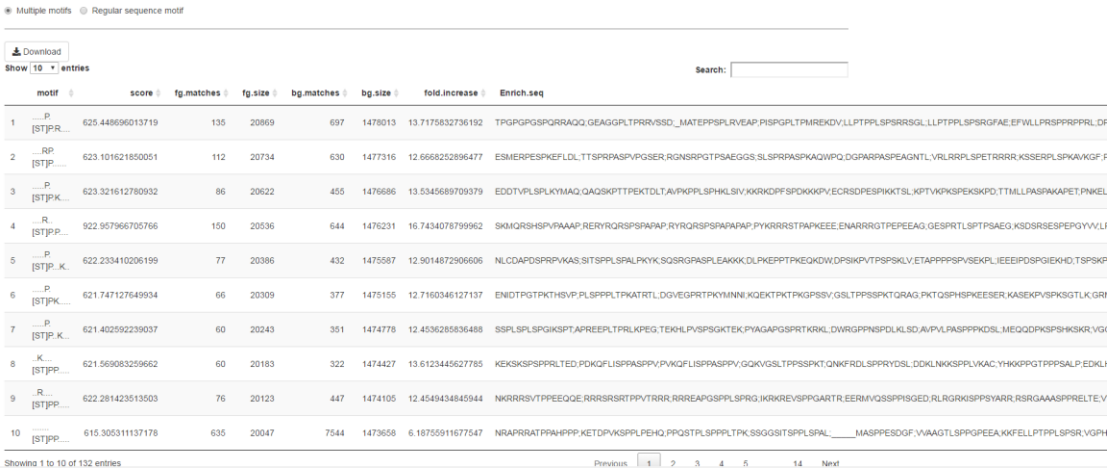
Fourth, click “Own Background”. As we have uploaded fasta file as background database, here we needn’t upload background data again, just keep it as default.



Fifth, click “Motif Enrichment”. As we upload fasta file as background database and don’t check these peptides if containing some regular sequence, here the default parameter set is ok. Then click “Calculate” button.



Then motifeR would generate the standard background database based on the fasta file we upload and then calculate the significant motifs. This step will take relatively a long time, especially when we analyze >10000 peptides. Now we can take a cup of coffee and wait some minutes.

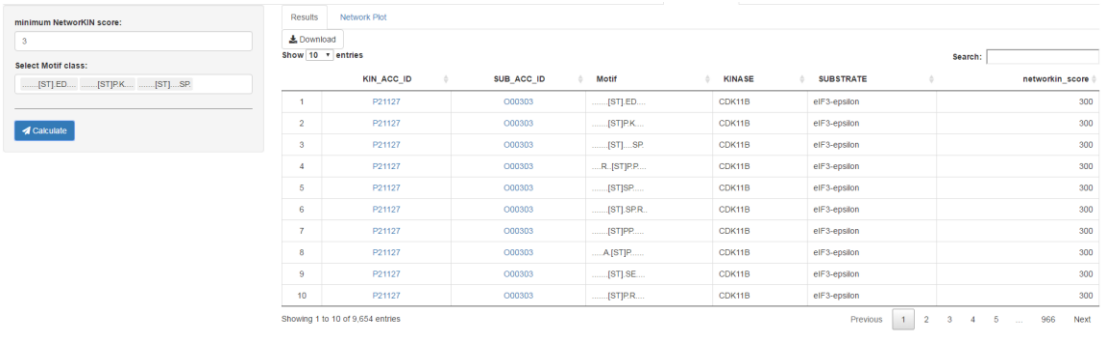


From this results, we can find there are 132 significant motifs and users can click “Download” button to save this results.

Sixth, click “Motif Plot” to get the visualization of these motifs. By default, the first motif will be plotted:

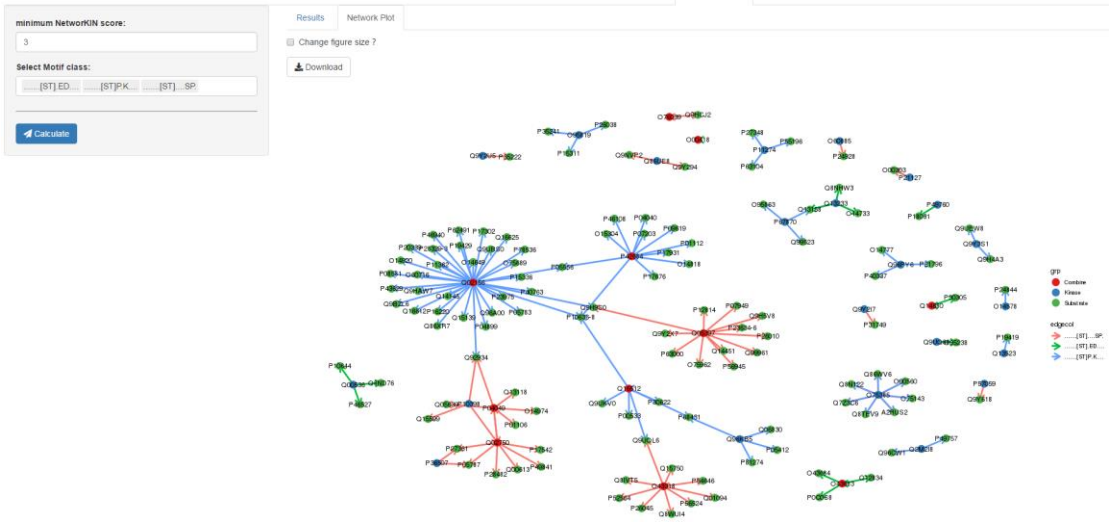
We can change the “Motif index for plot” parameter to plot more motifs:

Seventh, as this example data were obtained from Human samples, here we can continue. Click “Kinase Analysis”. (Warning: if your data are not obtained from Human samples, you need not continue.)



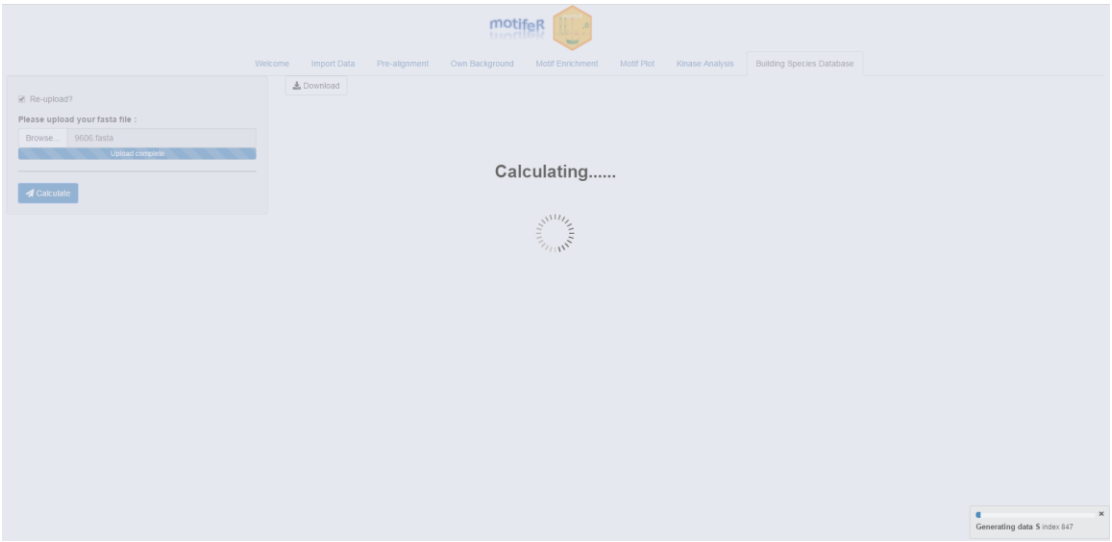
	KIN_ACC_ID	SUB_ACC_ID	Motif	KINASE	SUBSTRATE	networkin_score
1	P21127	O00303	.....[ST].ED....	CDK11B	eIF3-epsilon	300
2	P21127	O00303	.....[ST]P.K....	CDK11B	eIF3-epsilon	300
3	P21127	O00303	.....[ST]....SP.	CDK11B	eIF3-epsilon	300
4	P21127	O00303	....R..[ST]P.P....	CDK11B	eIF3-epsilon	300
5	P21127	O00303	.....[ST]SP.....	CDK11B	eIF3-epsilon	300
6	P21127	O00303	.....[ST].SP.R..	CDK11B	eIF3-epsilon	300
7	P21127	O00303	.....[ST]PP.....	CDK11B	eIF3-epsilon	300
8	P21127	O00303	....A.[ST]P.....	CDK11B	eIF3-epsilon	300
9	P21127	O00303	.....[ST].SE....	CDK11B	eIF3-epsilon	300
10	P21127	O00303	.....[ST]P.R....	CDK11B	eIF3-epsilon	300

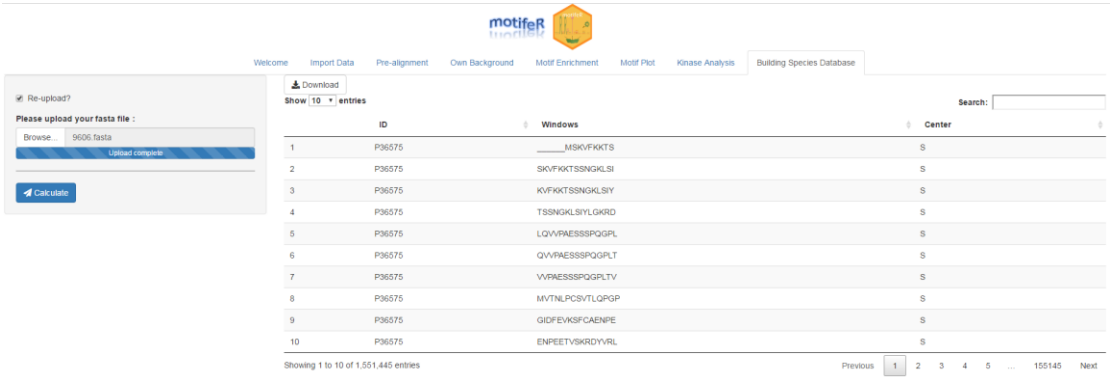
This part will give users some kinase-substrate information and then plot the network:



Herein we choose three motifs and this software plots their network based on the matched results. Perhaps, these information are useful and helpful for some users, thus we provide them here as an optional function.

Eighth, here we don't want to build the background database and download it, so we can ignore the "Building Species Database". However, if users are interested in the background database, they can re-upload the fasta file and obtain the standard background database as below:





This result may be huge based on the species fasta file. If users think this information is also useful for them, just click "Download" button to save it in the local computer. Note: this result

can be regulated based on the parameter in “Import Data” part.

Reference

(1) Emdal, K. B.; Pedersen, A.-K.; Bekker-Jensen, D. B.; Lundby, A.; Claeys, S.; De Preter, K.; Speleman, F.; Francavilla, C.; Olsen, J. V. *Sci. Signal.* **2018**, *11*, eaap9752.

(2) Gu, H.; Stokes, M. P.; Silva, J. C. In *Analysis of Post-Translational Modifications and Proteolysis in Neuroscience*; Springer, 2015, pp 1-29.

(3) Linding, R.; Jensen, L. J.; Pasculescu, A.; Olhovsky, M.; Colwill, K.; Bork, P.; Yaffe, M. B.; Pawson, T. *Nucleic acids research* **2007**, *36*, D695-D699.