Guideline on running Citrullination_Diagnostic_Ion_Analysis

Requirements

Python requirements

- Python version 3.6 or greater
- Dependencies: pandas, numpy, pyteomics, itertools, collections, statistics, re, io

File requirements

• Input search result file: The input file must be a .csv file(s) containing the following three columns (see next page for the details):

'Title' MS2 spectrum title as written in MGF file

'Peptide' Peptide sequence with modification delta mass rounded up to third decimal places

'Charge' Charge state of the peptide

• Spectrum file: The spectrum file must be a Mascot Generic Format (MGF) file(s) containing MS2 spectra corresponding to those matched to the PSMs in the input search result file. If MS2 spectra in the input search file and spectrum file are not equivalent, only the common MS2 spectra will be retained and subsequently processed.

Input search result file

Example input search result file:

A	В	С
Title	Peptide	Charge
20160312_02_A1.10012.10012.2 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=100	YETSGIGEAR+0.984VK	2
20160312_02_A1.10045.10045.2 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=1004	NIVTPR+0.984TPPPSQGK	2
20160312_02_A1.10116.10116.3 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=101	NIVTPR+0.984TPPPSQGK	3
20160312_02_A1.10222.10222.2 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=102	NIVTPR+0.984TPPPSQGK	2
20160312_02_A1.10334.10334.3 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=103	NIVTPR+0.984TPPPSQGK	3
20160312_02_A1.10362.10362.3 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=1036	TPSTAHLR+0.984VPK	3
20160312_02_A1.10418.10418.2 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=104	NIVTPR+0.984TPPPSQGK	2
20160312_02_A1.10479.10479.2 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=104	AQSR+0.984EQLAALK	2
20160312_02_A1.1054.1054.2 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=1054"	DSR+0.984SGSPM+15.995AR	2
20160312_02_A1.10602.10602.2 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=1060	NIVTPR+0.984TPPPSQGK	2
20160312_02_A1.10646.10646.2 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=1064	Q+0.984KR+0.984LQ+0.984AM+15.995Q+0.984K	2
20160312_02_A1.10671.10671.3 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=106	SGSEAGSPRR+0.984PRRQR	3
20160312_02_A1.1073.1073.2 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=1073"	R+0.984GGGGRR+0.984SK	2
20160312_02_A1.10764.10764.2 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=1076	R+0.984FIN+0.984DMVK	2
20160312_02_A1.10769.10769.2 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=1076	NIVTPR+0.984TPPPSQ+0.984GK	2
20160312_02_A1.10874.10874.2 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=108	MAR+0.984EAEFEAEQER	2
20160312_02_A1.11026.11026.3 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=1102	RGR+0.984PPKDEK	3
20160312_02_A1.11286.11286.2 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=1128	N+0.984R+0.984Q+0.984VIC+57.021VTLK	2
20160312_02_A1.11398.11398.2 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=113	GGTSR+0.984ALAAASSVK	2
20160312_02_A1.11489.11489.3 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=1148	BEFER+0.984Q+0.984N+0.984KQLR	3
20160312_02_A1.11557.11557.2 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=115	TVEMR+0.984DGEVIK	2
20160312_02_A1.11735.11735.2 File:"20160312_02_A1.raw", NativelD:"controllerType=0 controllerNumber=1 scan=117	Q+0.984Q+0.984IADLR+0.984EDLKR	2
> search_result_file ⊕ : 1		

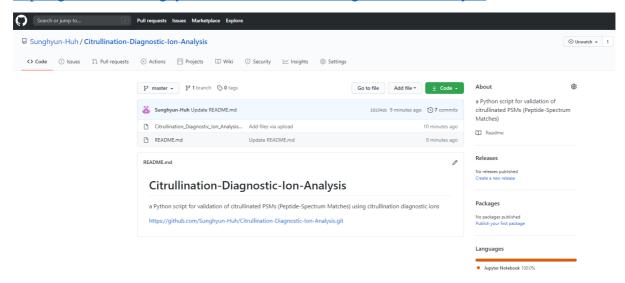
Format for peptide sequences should follow that of MS-GF+ search result. Specifically, modification delta masses should be rounded up to third decimal places. Currently allowed modifications are as follows:

Modification	Mod on peptide						
Carbamidomethyl Cys	C+57.021						
Oxidation Met	M+15.995						
Deamidated Asn	N+0.984						
Deamidated Gln	Q+0.984						
Citrullinated Arg	R+0.984						
Pyro-Glu from Glu	E-17.027						
Pyro-Glu from Gln	Q-18.011						
iTRAQ 4plex Lys	K+144.102						
iTRAQ 8plex Lys	K+304.205						
TMT Lys	K+229.163						
iTRAQ 4plex N-term	+144.102						
iTRAQ 8plex N-term	+304.205						
TMT N-term	+229.163						
Acetyl N-term	+42.011						

Downloading the Python script

The Python script can be downloaded via following GitHub page:

 $\underline{https://github.com/Sunghyun-Huh/Citrullination-Diagnostic-Ion-Analysis}$

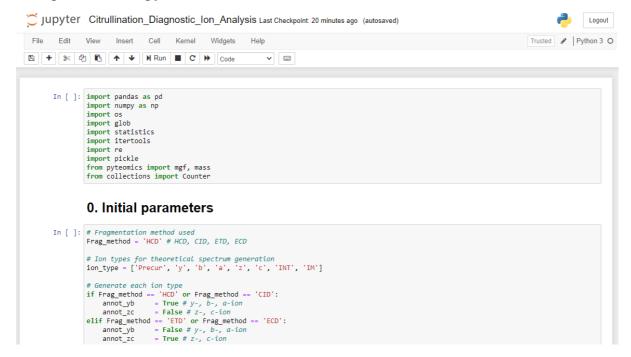


Users can download the Jupyter notebook via the following steps:

- Click on the 'Citrullination_Diagnostic_Ion_Analysis.ipynb'.
- Click on the 'Raw'.
- Press ctrl+s and manually type 'ipynb' after the filename to download as a .ipynb file.

Running the Python script

A snapshot of the Jupyter notebook:



Running the Python script

A snapshot of initial parameters settings:

0. Initial parameters

Explanations of initial parameters are as follows:

Frag_method	Fragmentation method used in the input data (value = 'HCD', 'CID', 'ETD', or 'ECD').
	If set as 'HCD' or 'CID', y-ion, b-ion, and a-ion will be generated for theoretical
	spectrum. If set as 'ETD' or 'ECD', z-ion, c-ion will be generated for theoretical
	spectrum. Commonly, precursor, internal, and immonium ion will be generated for all
	fragmentation method used.
ms2_ppm	MS2 level mass tolerance in ppm (default = 15 ppm)
apply_SNR	Determine whether to apply signal-to-noise filter to remove noise peaks (value = True or
	False; default = True)
SNR	Signal threshold level. The average intensity of noise peaks (as defined in 'low')
	multiplied by this signal threshold level will be the final signal-to-noise filter. If
	'apply_SNR' = True, all peaks below the signal-to-noise filter will be removed (default
	= 2)
low	Proportion of MS2 peaks regarded as noise. If 'apply_SNR' = True, all peaks below this
	noise level will be treated as noise (default = 0.05)
max_charge	Maximum charge state of fragment ions (default $= 2$)
max_NL	Maximum number of neutral losses from a single ion (default = 3)

Running the Python script

A snapshot of codes for loading input files:

1. Input files

```
In []: # Set current working directory
PATH = "F:/Project/"
os.chdir(PATH)

In []: # Input files
    spec_files = glob.glob('spectrum_file.mgf') # MGF file(s)
    search_files = glob.glob('search_result_file.csv') # Search result file(s)
```

Users can upload local input files via the following steps:

- Set the directory in which the input files are located.
- Copy and paste the input filenames. In case of multiple MGF or search result files in the same directory, type in '*.mgf' or '*.csv'.

Output result file

Example output file:

Peptide mod_Peptide P YETSGIGE, YETSGIGEARV NIVTPR+0 NIVTPrTPPPSC NIVTPR+0 NIVTPrTPPPSC		2 655.8335189	it_Count Total_NL_label precNL_la		ei intrit_iabel lotal_INT_II	spei pipeptide_lat										
NIVTPR+0 NIVTP/TPPPSC NIVTPR+0 NIVTP/TPPPSC							A. A.C. MILIT A.C. F 1120 F. L. MILI	3 42	0	22	20	9	4	5	0	
NIVTPR+0 NIVTPrTPPPS0		2 746.9100969					Ar, ArV-NH3,ArV,EAr-H2O,EAr-NH		0	14	17	14	8	6		-
		3 498.2760063					43 TPr-43-H2O,PrT-43-H2O,rTP-4		0	14	17	14	8	6	0	1
NIVTPR+0 NIVTPrTPPPSC		2 746.9100969					43 TPr-43-H2O,PrT-43-H2O,rTP-4		0	19	28	16	10	6	0	- 1
							43 TPr-43-H2O,PrT-43-H2O,rTP-4		0	19	20	13	10	6	0	
NIVTPR+0 NIVTPrTPPPS(3 498.2760063		-43 D7-43-H20		PT-4 TI-43-M2O,PT	43 TPr-43-H2O,PrT-43-H2O,rTP-4	3 21	0		1	13		1	0	
TPSTAHLR TPSTAHL/VPK		3 403.2318493	1 TAHLr-43-H2O		TAHLr-43-H HLr-CO		HLr-CO		0	11	16	12	0	6	0	0.425753061
NIVTPR+0 NIVTP/TPPPS(2 746.9100969					rT- TPr-43-H2O,PrT-43-H2O,rTP-4		0	11		12	6	6	0	1
AQSR+0.9 AQS/EQLAALI		2 608.3387759					43 QSr-43-NH3,SrE-43-H2O,SrE-4	23	0		16		6			1
DSR+0.98-DSrSGSPmAR		2 540.7411109					43 rSG-H2O-NH3,rS IM(r)-NH3	8	0	4	4	8	6	2		0.999999059
NIVTPR+0 NIVTPrTPPPS(2 746.9100969		H2(y10-43-H2			rT- TPr-43-H2O,PrT-43-H2O,rTP-4		0	4	12	9	4	5		0.99999997
Q+0.984Kl qKrLqAmqK		2 575.7928139	1 rLq-43		rLq-43 Kr-H2O-NH			1	0	0	1	2	1	- 1		0.587272168
SGSEAGSF SGSEAGSPR/F		3 566.6354786	1 GSPRr-42-43,GSPRrP-43,b	13-b13-43-59	+ GSPRr-42-4 rPR-42-NHS	3,RrP-42-NH3,PRr-	42- rPR-42-NH3,RrP-42-NH3,PRr-4	2 14	0	2	12	1	0	1		0.99999583
R+0.984GI rGGGGGRrSK		2 495.2659454	2					0	0	0	0	0	0	0		0.127083007
R+0.984RI rEnDMVK		2 512.7607839	1					0	0	0	0	0	0	0	0	0.127083007
NIVTPR+0 NIVTPrTPPPSc		2 747.4021049		T-4y10-43-H2			NITPr-43-H2O,PrT-43-H2O,rTP-4	3 21	0	6	15	10	5	5	0	1
MAR+0.98 MArEAEFEAEC	13	2 798.8521184	1 rE-43-H2O		rE-43-H2O Ar-NH3,rE-4	43-H Ar-NH3,rE-43	H2O	1	0	0	1	2	2	0	0	0.587272168
RGR+0.98 RGrPPKDEK		3 361.8689106	1 rPP-43,rPPKD-43-H2O		rPP-43,rPPK[rP,rPP-43	rP .	rPP-43	2	0	0	2	2	1	1	0	0.790608964
N+0.984R nrqVlcVTLK	10	2 617.3295629	1 Precursor-43-N Precursor-	43-NH3	rqVIcVT-43 rqV		rqV	2	1	0	1	1	0	1	0	0.663001774
GGTSR+0. GGTSrALAAAS	14	2 638.8469649	1 Sr-43-H2O,Sr-43,SrA-43,T	5r-4b5-43-H20	D, Sr-43-H2O, Sr-43-H2O,	Sr-H Sr-43-H2O,Sr	H; TSr-43-H2O,SrA-IM(r)-NH3	21	0	10	11	14	3	11	1	1
EEFER+0.9 EEFErqnKQLR	11	3 493.9078709	1 ErqnKQL-43		ErgnKQL-43			1	0	0	1	0	0	0	0	0.278663657
TVFMR+0 TVFMrDGEVII	11	2 639 3244769	1 rD_43 Mr_43 rD Precursor.	43. y7.43.NH1	t+ r0.43,Mr.41 r0.43,Mr.41	3,m, m.43,Mr.43,r	D. (PDG-43-H2O)/DG-NH3/HG/Mr	52	2	16	84	17	6	11	n	1
Q+0.984Q qqIADLrEDLKI	12	2 744.3891944	1 ADLr-43-H2O,ADLr-43,AD	LrEDLK-43	ADLr-43-H2 Lr-NH3,rE-N	iH3, Lr-NH3,rE-NH	3 DLr-NH3,rED-NH3	3	0	0	3	4	2	2	0	0.973617048
DSR+0.98-DSrSGSPmAR	10	2 540.7411109	1 rS-43-H2O,Sr-43-H2O,rS-	43,563-43-H20	0, rs-43-H2O, rs-43-H2O,	Sr-4 rS-43-H2O,Sr	43 rSG-H2O-NH3,rS IM(r)-NH3	11	0	6	5	8	6	2	1	0.99999995
MFGGPGT MFGGPGTASr	23	3 797.7173133	1 rPSSS-43,SrPSS-43,GTASrF	S-4y15-42-43	+ rPSSS-43,SrPSS-43,GTAS	rPS-43-H2O,GTAS	PSS-43-H2O,SrPSSSR-43,rPSSSRS	- 22	0	3	19	0	0	0	0	0.999999997
YLATASTM YLATASTMDH	18	4 516.005272	1 Ar-43,ArH-43,HAr-43,DHA	r-4 y8-43++,yf	9- Ar-43,ArH-4 Ar-43,ArH-4	3,H ₂ Ar-43	ArH-43,HAr-43	18	0	5	13	2	1	- 1	0	0.999999956
YLATASTN YLATASTmDH	18	3 687.6710876	1 Ar-43,ArH-43,HAr-43,DHA	r-4 y9-43++y	7- Ar-43,ArH-4 Ar-43,ArH-4	3,H ₂ Ar-43	ArH-43,HAr-43	18	0	5	13	2	1	- 1	0	0.999999956
EKEEKEEAI EKEEKEEArLrA	13	2 809.4263074	2 rA-43,Ar-43,EAr-43-H2O,E	ArL-43,EEArL-4	13 rA-43,Ar-43, rA-43,Ar-43,	Ar-1 rA-43,Ar-43,A	-N EAr-43-H2O,EAr-NH3	11	0	0	11	4	2	2	0	0.999988977
DSR+0.98-DSrSGSPmAR	10	2 540.7411109	1 rS-43-H2O,Sr-43-H2O,rS-	43,5 b3-43-H20	0,1rs-43-H2O,1rs-43-H2O,1	Sr-4 rS-43-H2O,Sr	43 rSG-H2O-NH3,r\$ IM(r)-NH3	6	0	2	4	8	6	2	- 1	0.999993371
ITPGAR+0 ITPGArGAPSEI	14	2 763.8784579	1 rG-43,Ar-43,GA Precursor-	43-y9-43-NH3	8+ rG-43,Ar-43 rG-43,Ar-43	rG- rG-43,Ar-43,rt	3-N GAr-43.ArG-43.rGA-43.ArG-NH	5 42	1	16	25	8	5	3	0	1
LEMVVVN LEMVVVNGrG	14	3 500.5932293	1 rGD-43,VNGrGD-43-NH3,	7 Y7-43-H2C	rGD-43,VNC Gr-NH3,rG-	NHEGr-NH3,rG-NI	H3 GrG-CO,rGD-43	5	0	1	4	3	1	2	0	0.992668379
STLAR+0.5 STLArVIVDK	10	2 551.8275124					Ar, ArV-43,LAr-43,ArV-NH3,LAr-NH	30	2	14	14	15	6	9	0	1
Q+0.984A gAPRAAPAAP	16	4 436.2305062	1	1				0	0	0	0	0	0	0	0	0.127083007
STLAR+0.5 STLAYVIVDK	10	2 551.8275124	1 Ar-43 rV-43 LAr Precursor-	43 v6-43++.v	7- Ar-43.rV-43. Ar-43.Ar-NF	13.r\ Ar-43.Ar-NH3	rV LAr-43.ArV-NH3.LAr-NH3.ArV.L	A 18	1	7	10	13	6	7	0	- 1
DSR+0.98-DSrSGSPmAR		2 540.7411109					43 rSG-H2O-NH3 IM(r)-NH3	8	0	4	4	7	6	1	- 1	0.999998193
GAAGR+0 GAAGrPLELSE	13	3 463.9094346	1 AGr-43 rPLE-43 GrPLELSD-		AGr-43 rPLE AGr-43 rP.A		AGr-43 AGr-CO AGr	3	0	0	3	4	1	3	0	0.973617048
ILSDDVPIR ILSDDVPIrDYF		3 488.2566083	1 Pir-43.VPirDY-43-H2O				D Pir-43.irD-H2O-NH3.Pir-NH3.ir	0 2	0	0	2	8	2	6		0.994727688
DSR+0.98-DSrSGSPmAR		2 540.7411109		43 (b3.43.H20			43 rSG-H2O-NH3 IM(r)-NH3	6	0	2	4	7	6	1		0.999987277
FADLSEAA FADLSEAANIN		2 889.4273694					-4 ANr-43-NH3.ANr-NH3-NH3.AN	r 19	0	10	9	4	1	3		0.999999996
RPFKR+0.5 RPFK/mNYSDv								7	0	- 1	6	1	0	1		0.996151439
			1					0	0	0	0	2	2	0		0.349054012
SELO+09ESE oS JECUR				V16-10 43 M2			NA IEC MINO		0	1			4	1		0.999892692
SFLQ+0.9E SFLqSLEcLrR NLDIER+0 NLDIErPTYTNI																0.555052052
RPFKR+0.5 R	FLqSLEcLrR	FLqSLEcLrR 11	FLqSLEcLrR 11 3 470.9069253	FLqSLEcLrR 11 3 470,9069253 1	FLqSLEcLrR 11 3 470,9069253 1	FLqSLEcLrR 11 3 470.9069253 1 Lr-NH3,Lr	PLqSLEcLrR 11 3 470.9069253 1 Lr-NH3,Lr Lr-NH3,Lr	FLqSLEcLrR 11 3 470.9069253 1 Lr-NH3,Lr Lr-NH3,Lr	FLqSLEcLrR 11 3 470,9069253 1 Lr-NH3,Lr Lr-NH3,Lr 0	RQSLECLER 11 3 4709069253 1 LF-43-H2O,Er-43,DIEP-43,YI(Y)0-43-H2O,Er-43-H2O,Er-43-H2O,F-N;Er-43-H2O,P-N;IEr-H2O 8 0	RqSLEctur 11 3 4709089253 1 U-1NH3_Lr U-1NH3_Lr U-1NH3_Lr 0 0 0 0 UDIEPTYTNI 14 3 5739607429 1 Er-43-H2O_Er-43,DIEP-43y1(y10-43-H2O_Er-43-H2O_P-N-1Er-43-H2O_P-N-1Er-H2O 8 0 3	RQSEREUR 11 3 4709999253 1	RQSEEUR 11 3 4709069253 1	RQSEEUR 11 3 470998933 1 LOUDEPYTN 14 3 5739807429 1 E-43-H2QE-43,DIEP-43,VI(y)-43-H2Q E-43-H2Q)-F-43-H2Q)	RADIEGE 1 3 17-0500235 1 0-43-400 (-4-3-400 (-	RqStRctrR 11 3 4709069255 1 Ur-NH3_Lr Ur-NH3_Lr 0 0 0 0 2 2 0 0

Explanations of output columns are as follows:

Column	Explanation							
mod_peptide	Simplified peptide with a predefined set of symbols for modifications							
Pep_length	Peptide length							
mz_Precursor	Theoretical precursor <i>m/z</i>							
Cit_Count	Number of citrullinated sites							
Total_NL_label	Annotations of all diagnostic neutral loss ions							
precNL_label	Annotations of precursor neutral losses							
seqNL_label	Annotations of sequence ion neutral losses							
intNL_label	Annotations of internal ion neutral losses							
Total_INT_label	Annotations of all diagnostic internal ions							
Dipeptide_label	Annotations of diagnostic dipeptides							
Tripeptide_label	Annotations of diagnostic tripeptides							
IM_NH3_label	Annotation of IM(Cit)-NH ₃							
Total_NL_count	Number of all diagnostic neutral loss ions							
precNL_count	Number of precursor neutral losses							
seqNL_count	Number of sequence ion neutral losses							
intNL_count	Number of internal ion neutral losses							
Total_INT_count	Number of all diagnostic internal ions							
Dipeptide_count	Number of diagnostic dipeptides							
Tripeptide_count	Number of diagnostic tripeptides							
IM_NH3_count	Number of IM(Cit)-NH ₃							
Cit_probability	Probability (P) of citrullination status calculated by the EN model (HCD data only)							
Cit_prediction	Classification of citrullination status using a P cutoff >0.5 (HCD data only)							