Last update: February 22nd, 2023

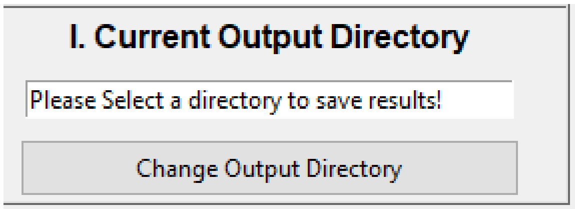
FRAGARIYO .v2

Instructions

Fragariyo is written in Python 3.7 using PyCharm IDE. In this new version it comes with a GUI! Let’s hunt some fragments with Fragariyo!

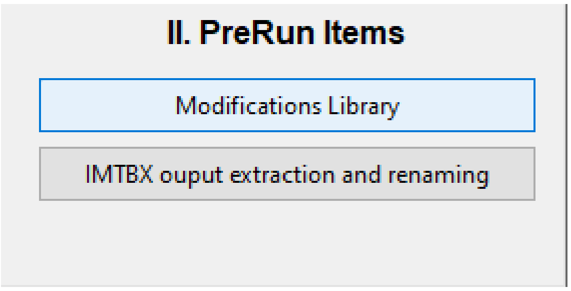
I. Choose where to save results.

Simply press the “Change Output Directory button”. It can be changed at anytime except while Fragariyo is running processes.



Please Select a directory to save results!

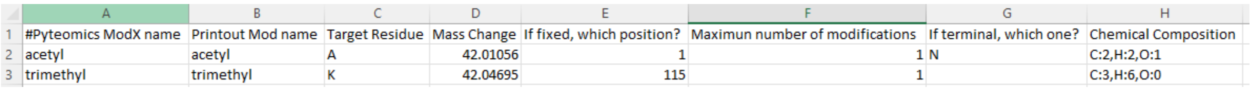
II. PreRun Items



I. Modifications Library. Once clicked, user can upload a .txt file containing modification information (see ModificationsRepository.txt for a template).

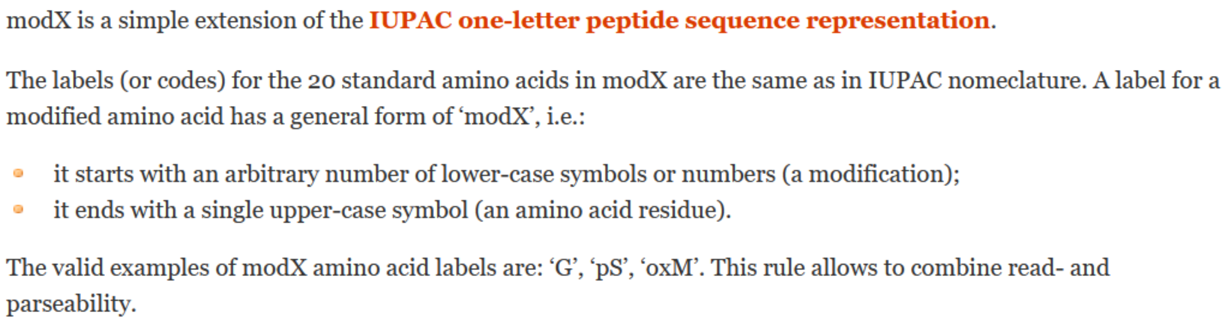
Last update: February 22nd, 2023

ModificationsRepository.txt



	A	B	C	D	E	F	G	H
1	#Pyteomics ModX name	Printout Mod name	Target Residue	Mass Change	If fixed, which position?	Maximun number of modifications	If terminal, which one?	Chemical Composition
2	acetyl	acetyl	A	42.01056	1	1	N	C:2,H:2,O:1
3	trimethyl	trimethyl	K	42.04695	115	1		C:3,H:6,O:0

A. Pyteomics ModX name: String with the modification name compliance with ModX
   a. ModX = https://pyteomics.readthedocs.io/en/latest/api/parser.html



modX is a simple extension of the **IUPAC one-letter peptide sequence representation**.

The labels (or codes) for the 20 standard amino acids in modX are the same as in IUPAC nomeclature. A label for a modified amino acid has a general form of 'modX', i.e.:

- it starts with an arbitrary number of lower-case symbols or numbers (a modification);
- it ends with a single upper-case symbol (an amino acid residue).

The valid examples of modX amino acid labels are: 'G', 'pS', 'oxM'. This rule allows to combine read- and parseability.

B. Printout Mod name: String with the modification name the user would like the results to print out as.
C. Target Residue (*Multiple residues support incoming*)
D. Mass Change (decimal number that can be either positive or negative).
E. If modifications is fixed, which position
F. Maximum number of Modifications per peptide
G. If modifications is a terminal one, which terminus to place it at?
H. Chemical Composition for scoring (only implemented in internal fragments searches, *terminal fragment support incoming)*

Element#1[colon]number of atoms of Element#1[comma] Element#2[colon]number of atoms of Element#2[comma]…etc…
DON'T ADD SPACES!

II. IMTBX Output Extraction and Renaming
Script to rename files output from IMTBX/Grppr. It takes .isotopes files from their original .raw file and renames them according to the .raw folder.

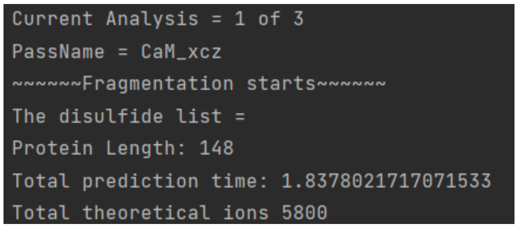
III. Match terminal fragments first

a. Choose Peaklist Files (File types acceptable are):
   i. .isotopes files from IMTBX/Grppr for Waters and from Mobilitron for Agilent files
   ii. mMass files (http://www.mmass.org/)
   iii. csv files (mass-to-charge ratio, charge, intensity with headers m/z,z,int or sample name)

        iv. unmatched files (produced from running terminal fragments. If ions are not matched, the unmatched ones will be placed in that file)

    b. Upload .ions (a previously produced theoretical ion database) files when prompted to, if a theoretical database has been created.
    c. Load the Parameter file (see TerminalFragments_template_example.csv to learn how to fill a parameter file).

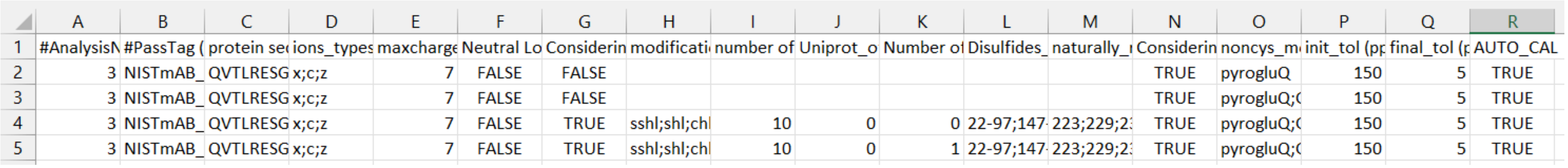
Terminal Matching starts. Fragariyo GUI will freeze, but progress can be followed in the terminal window running with Fragariyo. Fragmentation will start. The analysis number will appear as well as the current pass (PassName). An analysis is a group of passes.



```
Current Analysis = 1 of 3
PassName = CaM_xcz
~~~~~~Fragmentation starts~~~~~~
The disulfide list =
Protein Length: 148
Total prediction time: 1.8378021717071533
Total theoretical ions 5800
```

Last update: February 22nd, 2023

Terminal Fragment Parameter File

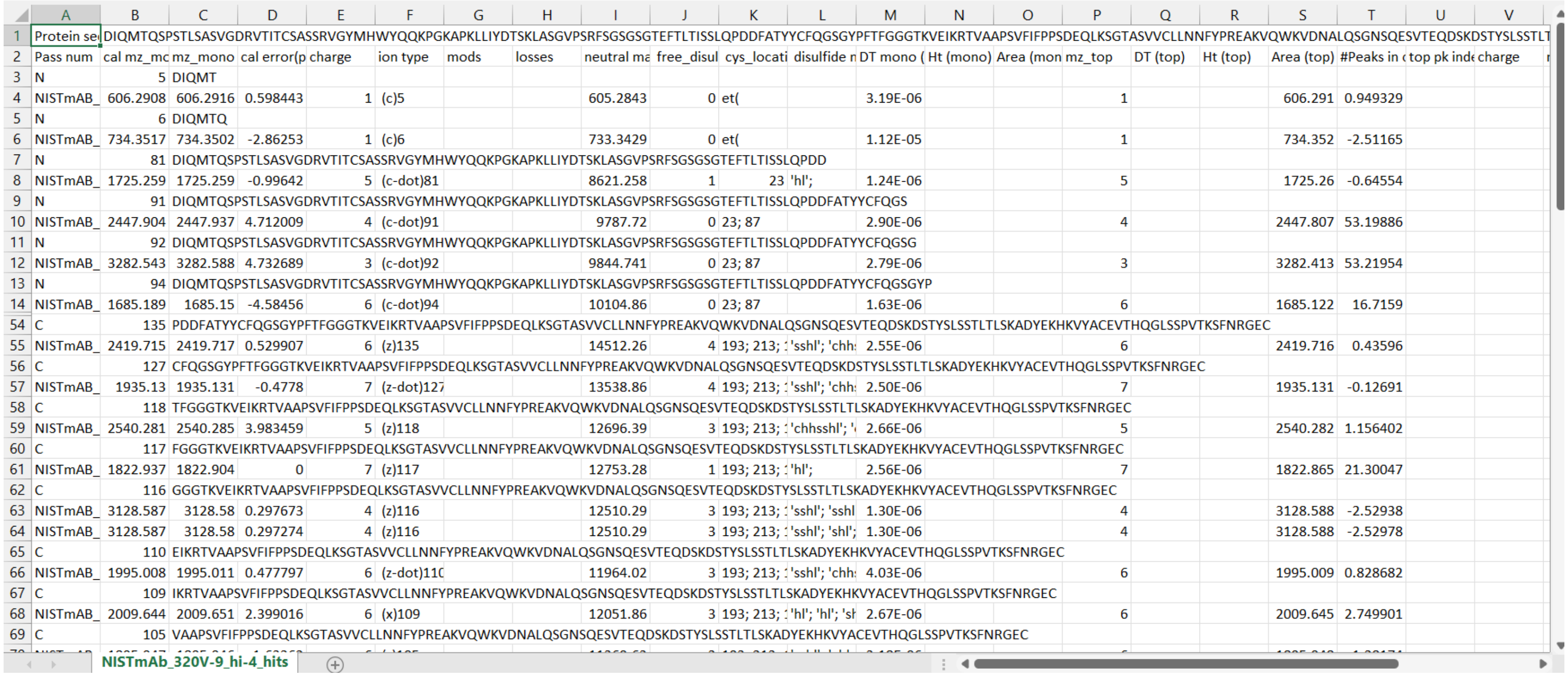


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	#AnalysisN	#PassTag (	protein se	ions_types	maxcharge	Neutral Lo	Considerin	modificati	number of	Uniprot_o	Number of	Disulfides_	naturally_	Considerin	noncys_m	init_tol (pp	final_tol (p	AUTO_CAL
2	3	NISTmAB_	QVTLRESG	x;c;z	7	FALSE	FALSE							TRUE	pyrogluQ	150	5	TRUE
3	3	NISTmAB_	QVTLRESG	x;c;z	7	FALSE	FALSE							TRUE	pyrogluQ;(	150	5	TRUE
4	3	NISTmAB_	QVTLRESG	x;c;z	7	FALSE	TRUE	sshl;shl;chl	10	0	0	22-97;147-	223;229;2	TRUE	pyrogluQ;(	150	5	TRUE
5	3	NISTmAB_	QVTLRESG	x;c;z	7	FALSE	TRUE	sshl;shl;chl	10	0	1	22-97;147-	223;229;2	TRUE	pyrogluQ;(	150	5	TRUE

A. Analysis number
B. Name for the pass (it can be something informative about the pass e.g. ion type, mutant, etc)
C. One-letter-code protein sequence
D. Ion types (by inputting z-dot and c-dot ions are automatically included)
E. Maximum charge state to be used
F. Neutral losses bool. TRUE will considered neutral losses.
G. Disulfide bond bool. TRUE makes the script to determine the amount of disulfide bonds and where they are and to separate them if the two cysteine partners end up in different fragments
H. What modifications to use for the reduced cysteines
   a. sshl;shl;chhsshl;hl;sh;h
I. What is the maximum number of modifications to be used for the cysteine modifications (maximum number of cysteines that can be reduced)
J. Uniprot_offset, int. Number that the sequence index must be offset to match the disulfide bond positions , if these positions were obtain from the uniport database. Put “0” is no offset needed
K. Number of disulfide bonds to be broken even if their cysteines are in the same fragment
L. Disulfide bond list
   a. Cys1_index-Cys1partner_index; Cys2_index-Cys2partner_index (e.g NISTmAb disulfide bonds: 22-97;147-203;223-1000;229-1000;232-1000;264-324;370-428).
M. Index for cysteines that do not participate in any disulfide bonds.
N. Considering modifications (not involved in disulfide bonds)
O. Modifications list (separate modifications with semi colons (e.g. NISTmAb modification for the n-term and glycan: pyrogluQ;GiiF)
   a. See Modifications.py section for more information
P. Initial error tolerance
Q. Final error tolerance
R. Auto Calibration bool.If TRUE, Auto calibration will be performed using

Last update: February 22nd, 2023

Terminal Fragment Results



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Protein se	DIQMTQSPSTLSASVGDRVTITCSASSRVGYMHWYQQKPGKAPKLLIYDTSKLASGVPSRFSGSGSGTEFTLTISSLQPDDFATYYCFQGSGYPFTFGGGTKVEIKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLT																				
2	Pass num	cal mz_mc	mz_mono	cal error(p	charge	ion type	mods	losses	neutral ma	free_disul	cys_locati	disulfide n	DT mono (	Ht (mono)	Area (mon	mz_top	DT (top)	Ht (top)	Area (top)	#Peaks in	top pk inde	charge
3	N	5	DIQMT																			
4	NISTmAB_	606.2908	606.2916	0.598443	1	(c)5			605.2843	0	et(		3.19E-06			1			606.291	0.949329		
5	N	6	DIQMTQ																			
6	NISTmAB_	734.3517	734.3502	-2.86253	1	(c)6			733.3429	0	et(		1.12E-05			1			734.352	-2.51165		
7	N	81	DIQMTQSPSTLSASVGDRVTITCSASSRVGYMHWYQQKPGKAPKLLIYDTSKLASGVPSRFSGSGSGTEFTLTISSLQPDD																			
8	NISTmAB_	1725.259	1725.259	-0.99642	5	(c-dot)81			8621.258	1	23	'hl';	1.24E-06			5			1725.26	-0.64554		
9	N	91	DIQMTQSPSTLSASVGDRVTITCSASSRVGYMHWYQQKPGKAPKLLIYDTSKLASGVPSRFSGSGSGTEFTLTISSLQPDDFATYYCFQGS																			
10	NISTmAB_	2447.904	2447.937	4.712009	4	(c-dot)91			9787.72	0	23; 87		2.90E-06			4			2447.807	53.19886		
11	N	92	DIQMTQSPSTLSASVGDRVTITCSASSRVGYMHWYQQKPGKAPKLLIYDTSKLASGVPSRFSGSGSGTEFTLTISSLQPDDFATYYCFQGSG																			
12	NISTmAB_	3282.543	3282.588	4.732689	3	(c-dot)92			9844.741	0	23; 87		2.79E-06			3			3282.413	53.21954		
13	N	94	DIQMTQSPSTLSASVGDRVTITCSASSRVGYMHWYQQKPGKAPKLLIYDTSKLASGVPSRFSGSGSGTEFTLTISSLQPDDFATYYCFQGSGYP																			
14	NISTmAB_	1685.189	1685.15	-4.58456	6	(c-dot)94			10104.86	0	23; 87		1.63E-06			6			1685.122	16.7159		
54	C	135	PDDFATYYCFQGSGYPFTFGGGTKVEIKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC																			
55	NISTmAB_	2419.715	2419.717	0.529907	6	(z)135			14512.26	4	193; 213; :	'sshl'; 'chh:	2.55E-06			6			2419.716	0.43596		
56	C	127	CFQGSGYPFTFGGGTKVEIKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC																			
57	NISTmAB_	1935.13	1935.131	-0.4778	7	(z-dot)127			13538.86	4	193; 213; :	'sshl'; 'chh:	2.50E-06			7			1935.131	-0.12691		
58	C	118	TFGGGTKVEIKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC																			
59	NISTmAB_	2540.281	2540.285	3.983459	5	(z)118			12696.39	3	193; 213; :	'chhsshl'; '	2.66E-06			5			2540.282	1.156402		
60	C	117	FGGGTKVEIKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC																			
61	NISTmAB_	1822.937	1822.904	0	7	(z)117			12753.28	1	193; 213; :	'hl';	2.56E-06			7			1822.865	21.30047		
62	C	116	GGGTKVEIKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC																			
63	NISTmAB_	3128.587	3128.58	0.297673	4	(z)116			12510.29	3	193; 213; :	'sshl'; 'sshl	1.30E-06			4			3128.588	-2.52938		
64	NISTmAB_	3128.587	3128.58	0.297274	4	(z)116			12510.29	3	193; 213; :	'sshl'; 'shl';	1.30E-06			4			3128.588	-2.52978		
65	C	110	EIKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC																			
66	NISTmAB_	1995.008	1995.011	0.477797	6	(z-dot)11C			11964.02	3	193; 213; :	'sshl'; 'chh:	4.03E-06			6			1995.009	0.828682		
67	C	109	IKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC																			
68	NISTmAB_	2009.644	2009.651	2.399016	6	(x)109			12051.86	3	193; 213; :	'hl'; 'hl'; 'sh	2.67E-06			6			2009.645	2.749901		
69	C	105	VAAPSVFIFPPSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC																			

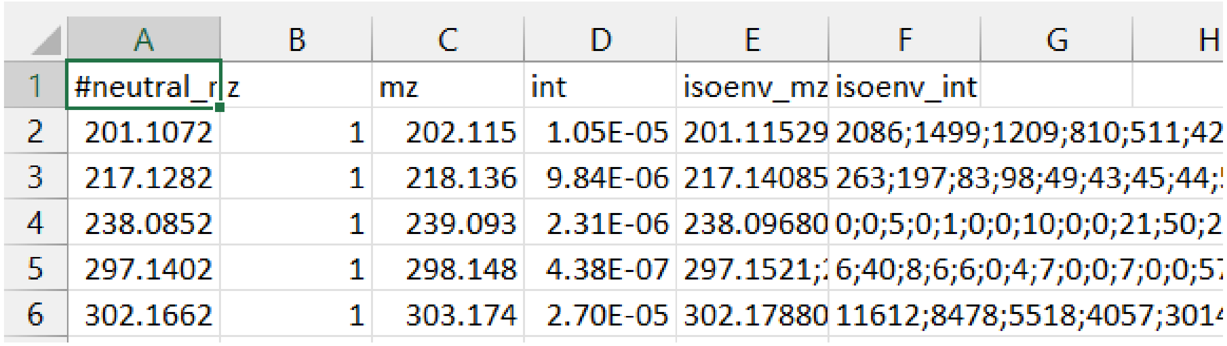
The results are organized by termini and fragment site. The results are saved as.csv files and in binary .hits files (created with Pickle).

IV. Internal Fragmentation

I. Input Generator

Internal fragment matches are scored by comparing experimental with theoretical isotope envelopes. Therefore experimental isotope information is extracted from files containing the xy coordinates of the spectrum of interest. First a .csv file with experimental ions (mass-to-charge ratio, charge, intensity with headers m/z,z,int) is requested. Next Fragariyo will ask for the xy coordinates (currently only accepting .csv files from Agilent and xy files from Breuker. For the Breuker ones add the headers: 'X(MassToCharge)'[space]'Y(Counts)'.

Input ready for Internal Fragment matching:



	A	B	C	D	E	F	G	H
1	#neutral_r	z	mz	int	isoenv_mz	isoenv_int		
2	201.1072	1	202.115	1.05E-05	201.11529	2086;1499;1209;810;511;42		
3	217.1282	1	218.136	9.84E-06	217.14085	263;197;83;98;49;43;45;44;		
4	238.0852	1	239.093	2.31E-06	238.09680	0;0;5;0;1;0;0;10;0;0;21;50;2		
5	297.1402	1	298.148	4.38E-07	297.1521;	6;40;8;6;6;0;4;7;0;0;7;0;0;5		
6	302.1662	1	303.174	2.70E-05	302.17880	11612;8478;5518;4057;301		

II. Internal Run

i. Choose Peaklist Files (in-house .csv file). Files can be obtained from unmatched files obtained from matching terminals or they can be created from ions obtained from other sources.

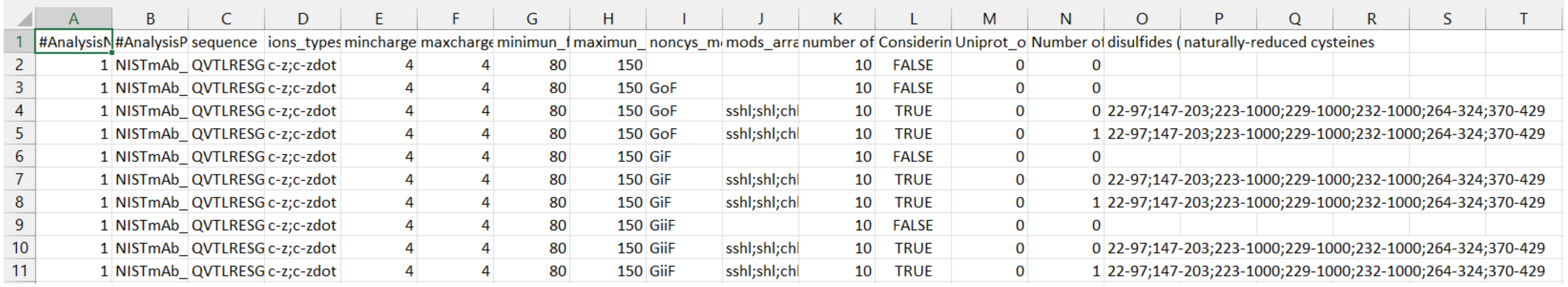
A. Neutral mass of experimental ion obtained from charge (z) and m/z values
B. Charge
C. C
D. Mass-to-charge ratio
E. Values *m/z* corresponding to the peak isotope envelope

      F. Values *intensity* corresponding to the peak isotope envelope

   ii. Load .ions file, if a previous created database will be used

   iii. Choose parameter file, if a new database needs to be created

      A. Analysis number



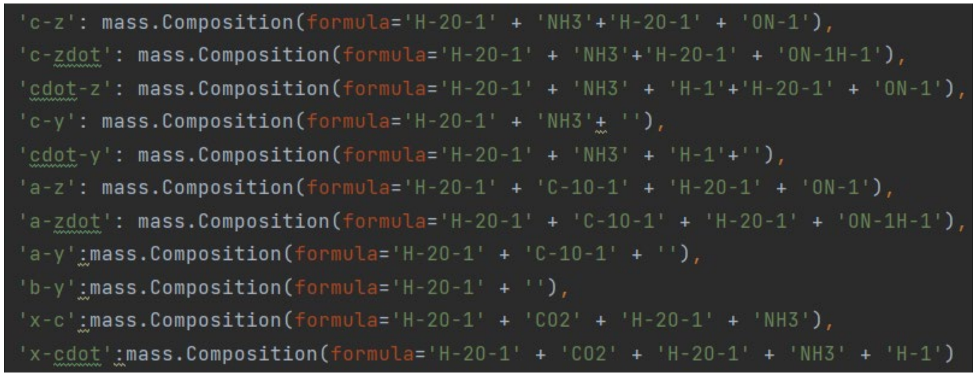
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	#AnalysisN	#AnalysisP	sequence	ions_types	mincharge	maxcharge	minimun_f	maximun_	noncys_m	mods_arra	number of	Considerin	Uniprot_o	Number of	disulfides (	naturally-reduced cysteines				
2	1	NISTmAb_	QVTLRESG	c-z;c-zdot	4	4	80	150			10	FALSE	0	0						
3	1	NISTmAb_	QVTLRESG	c-z;c-zdot	4	4	80	150	GoF		10	FALSE	0	0						
4	1	NISTmAb_	QVTLRESG	c-z;c-zdot	4	4	80	150	GoF	sshl;shl;chl	10	TRUE	0	0	22-97;147-203;223-1000;229-1000;232-1000;264-324;370-429					
5	1	NISTmAb_	QVTLRESG	c-z;c-zdot	4	4	80	150	GoF	sshl;shl;chl	10	TRUE	0	1	22-97;147-203;223-1000;229-1000;232-1000;264-324;370-429					
6	1	NISTmAb_	QVTLRESG	c-z;c-zdot	4	4	80	150	GiF		10	FALSE	0	0						
7	1	NISTmAb_	QVTLRESG	c-z;c-zdot	4	4	80	150	GiF	sshl;shl;chl	10	TRUE	0	0	22-97;147-203;223-1000;229-1000;232-1000;264-324;370-429					
8	1	NISTmAb_	QVTLRESG	c-z;c-zdot	4	4	80	150	GiF	sshl;shl;chl	10	TRUE	0	1	22-97;147-203;223-1000;229-1000;232-1000;264-324;370-429					
9	1	NISTmAb_	QVTLRESG	c-z;c-zdot	4	4	80	150	GiiF		10	FALSE	0	0						
10	1	NISTmAb_	QVTLRESG	c-z;c-zdot	4	4	80	150	GiiF	sshl;shl;chl	10	TRUE	0	0	22-97;147-203;223-1000;229-1000;232-1000;264-324;370-429					
11	1	NISTmAb_	QVTLRESG	c-z;c-zdot	4	4	80	150	GiiF	sshl;shl;chl	10	TRUE	0	1	22-97;147-203;223-1000;229-1000;232-1000;264-324;370-429					

      B. Pass name (something descriptive about the pass (e.g ions searched for, mods searched for, etc)

      C. Sequence = Protein sequence one-letter code

      D. Iontypes allowed for internal fragments:

      E. Minimum charge to be considered



```
'c-z': mass.Composition(formula='H-2O-1' + 'NH3'+'H-2O-1' + 'ON-1'),
'c-zdot': mass.Composition(formula='H-2O-1' + 'NH3'+'H-2O-1' + 'ON-1H-1'),
'cdot-z': mass.Composition(formula='H-2O-1' + 'NH3' + 'H-1'+'H-2O-1' + 'ON-1'),
'c-y': mass.Composition(formula='H-2O-1' + 'NH3'+ ''),
'cdot-y': mass.Composition(formula='H-2O-1' + 'NH3' + 'H-1'+''),
'a-z': mass.Composition(formula='H-2O-1' + 'C-1O-1' + 'H-2O-1' + 'ON-1'),
'a-zdot': mass.Composition(formula='H-2O-1' + 'C-1O-1' + 'H-2O-1' + 'ON-1H-1'),
'a-y':mass.Composition(formula='H-2O-1' + 'C-1O-1' + ''),
'b-y':mass.Composition(formula='H-2O-1' + ''),
'x-c':mass.Composition(formula='H-2O-1' + 'CO2' + 'H-2O-1' + 'NH3'),
'x-cdot':mass.Composition(formula='H-2O-1' + 'CO2' + 'H-2O-1' + 'NH3' + 'H-1')
```

      F. Maximum charge to be considered (right now max and min are the same to keep the passes somehow smaller. In a future update hopefully this will be simplified).

      G. Minimum fragment length in amount of residues

      H. Maximun fragment length in amount of residues

I. Modifications for non-cysteiens residues separate by semicolons (examples shoes an analysis with several passes where a different NISTmAb glycan is searched for).

J. What modifications to use for the reduced cysteines

K. What is the maximum number of modifications to be used for the cysteine modifications (maximum number of cysteines that can be reduced)

L. Disulfide bond bool. TRUE makes the script to determine the amount of disulfide bonds and where they are and to separate them if the two cysteine partners end up in different fragments

M. Uniprot_offset, int. Number that the sequence index must be offset to match the disulfide bond positions , if these positions were obtain from the uniport database. Put “0” is no offset needed

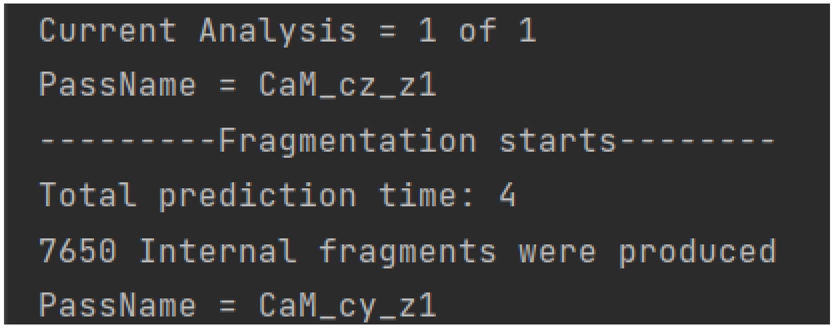
N. Number of disulfide bonds to be broken even if their cysteines are in the same fragment

O. Disulfide bond list

   a. Cys1_index-Cys1partner_index; Cys2_index-Cys2partner_index (e.g NISTmAb disulfide bonds: 22-97;147-203;223-1000;229-1000;232-1000;264-324;370-428).

1. Index for cysteines that do not participate in any disulfide bonds.

iv. Script will run (a protein a large as BSA took several days and breaking the analysis down to several pieces)



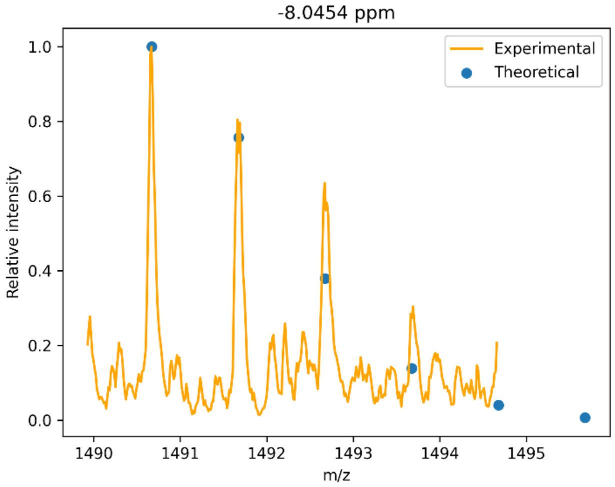
```
Current Analysis = 1 of 1
PassName = CaM_cz_z1
---------Fragmentation starts--------
Total prediction time: 4
7650 Internal fragments were produced
PassName = CaM_cy_z1
```

Last update: February 22nd, 2023

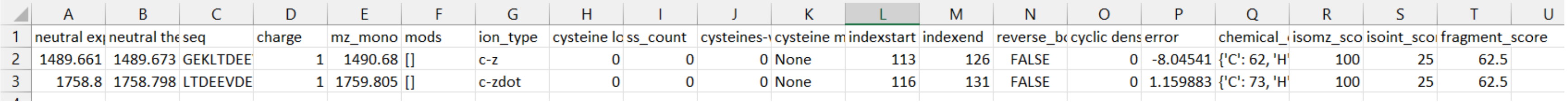
Internal Fragment Results

The results will be saved in a folder created with the name of the experimental ion .csv file: The folder contains:

1. A .theofrags file (a.ka. .ions file) a binary version of the theoretical database
2. A.txt.frags. A human readable version of the theoretical databased (they can be quite large). To be removed.
3. PNG files showing the matching of the theoretical isotopic envelope and experimental isotopic envelope



4. TSV files (tab-delimited files) containing the details of the matched peaks



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	neutral ex	neutral the	seq	charge	mz_mono	mods	ion_type	cysteine lo	ss_count	cysteines-	cysteine m	indexstart	indexend	reverse_b	cyclic dens	error	chemical_	isomz_sco	isoint_sco	fragment_score	
2	1489.661	1489.673	GEKLTDEE	1	1490.68	[]	c-z	0	0	0	None	113	126	FALSE	0	-8.04541	{'C': 62, 'H'	100	25	62.5	
3	1758.8	1758.798	LTDEEVDE	1	1759.805	[]	c-zdot	0	0	0	None	116	131	FALSE	0	1.159883	{'C': 73, 'H'	100	25	62.5	

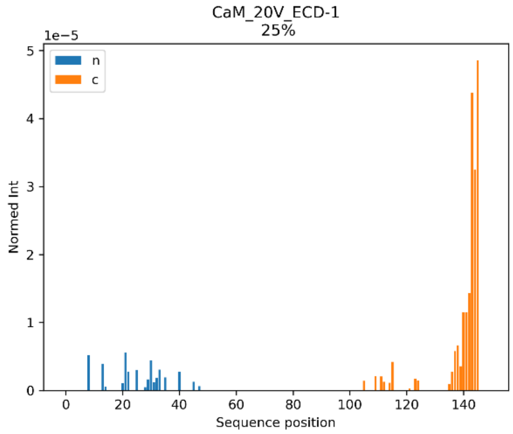
A. Neutral mass (experimental ion)
B. Neutral mass (theoretical ion)
C. Fragment sequence
D. Charge
E. Monoisotopic *m/z* value
F. Modifications
G. Ion type
H. Indices where the cysteines are located
I. Number of disulfide bonds in internal fragment
J. How many cysteines are being modified
K. Cystine modification if disulfide bonds were reduced
L. Index where the fragment starts
M. Index where the fragment ends
N. Bool. If TRUE the sequence is a reverse sequence. To be used as FDR (False Discovery Rate) calculations in the future.
O. Cyclic density (number of cyclic regions form by disulfide bonds/ total residue number)
P. Matching error
Q. Chemical compositions (in python dictionary form = {'C': 44, 'H': 72, 'N': 14, 'O': 18, 'S': 0, 'Fe': 0})
R. Scoring in *m/z*
S. Scoring in *intensity*
T. Compound score

Last update: February 22nd, 2023

V. Data Analysis

They are divided into terminal and internal sections

Sequence Coverage Calculations

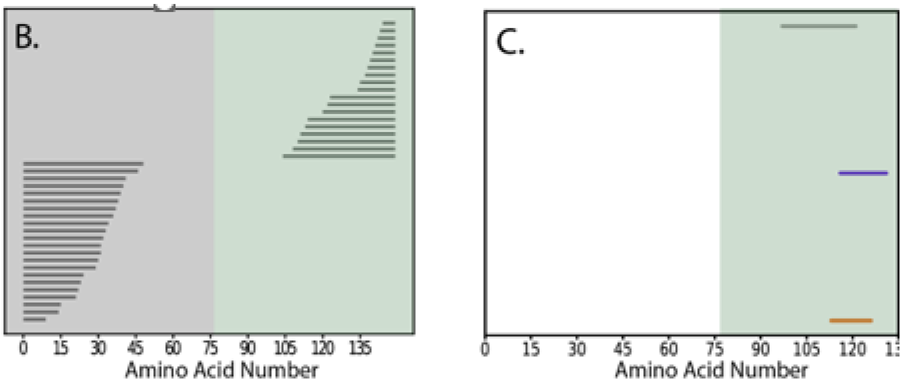


- Outputs example: A graph showing sequence coverage (coordinate information to replicate this graph is saved as a file: samplename_analysis.csv) It can be done per file or for a combination of two files.

ClipsMS Plotter

Based on ClipsMS (Software from the Loo lab) graphing code. It will transform terminal fragment input to be compatible with ClipsMS plotter. There is a terminal and an internal fragments version.

❖ Output:B. Terminal Fragments C. Internal Fragments (color has been added using Illustrator.



Internal Fragments Outputs

In the Internal fragments analysis tools there is the option of merging the output into one file. The other option is to “average” them: Only retain fragments that appear across all files.

Last update: February 22nd, 2023