

System requirements: Windows system with Java 8 64-bit or newer.

Installation: Unzip the folder. Inside, you need to run the application file “Formaldehyde_XL_Analyzer”

Example input and output files: Two sets of example files are available at:

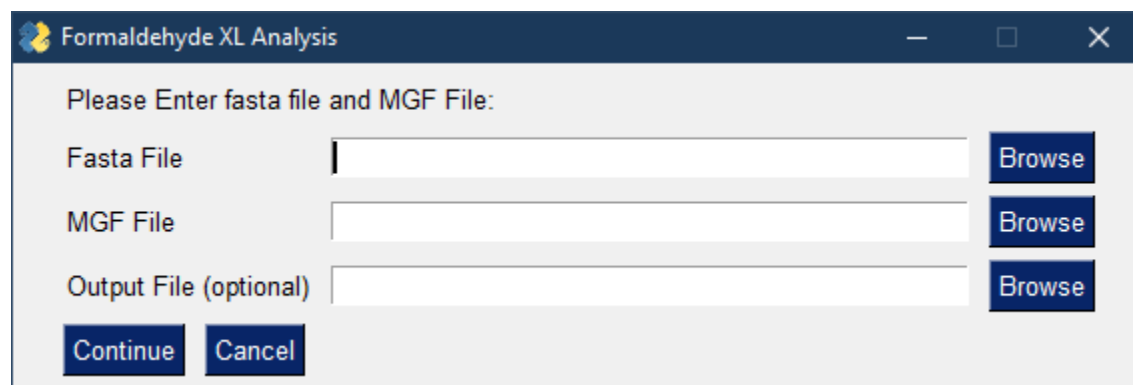
<http://biolchem.huji.ac.il/nirka/software.html>

Example 1 includes mass spectrometry data of a mixture of three proteins that was cross-linked by 4% formaldehyde. The run time of this example is less than a minute on any desktop computer.

Example 2 includes mass spectrometry data of *in situ* cross-linking of PC9 cells in culture by 4% formaldehyde. The search is done against a sequence database of 1692 human proteins that were identified to have medium to high abundance. The run time of this example is about 5 minutes on a strong desktop computer, but could reach up to 30 minutes on slower machines.

Running instruction:

1. When you run “Formaldehyde_XL_Analyzer”, the following window will open:



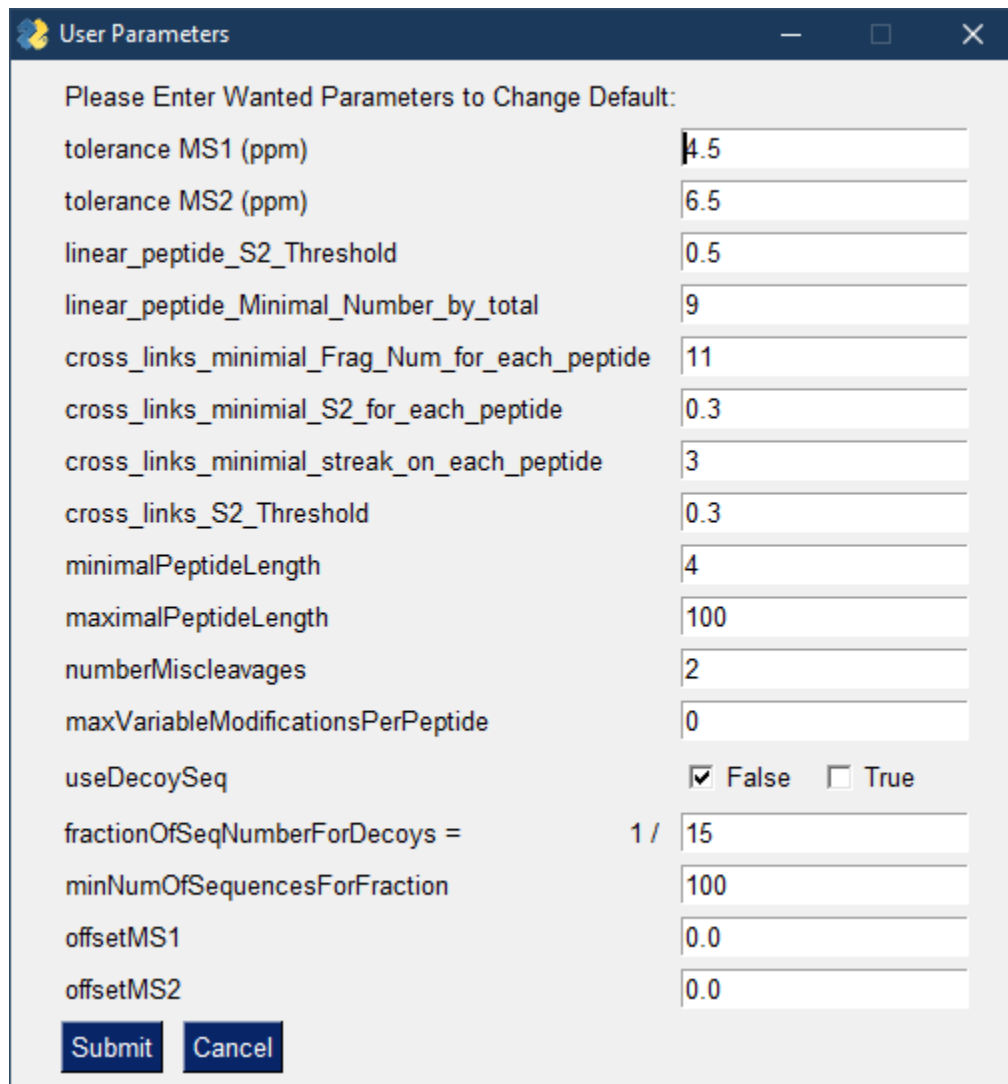
You need to specify the FASTA file with the sequences on which to search, and the MGF file with the mass spectrometry data. You can also specify the output file to which the textual results will be written. Each header in the FASTA file must start according to the UNIPROT format, with three fields separated by two “|”. The first two fields are not used, but the third (marked below in **RED**) is the protein name to be reported in the cross-link:

```
>sp|Q6DD88|ATLA3_HUMAN Atlastin-3 OS=Homo sapiens GN=ATL3 PE=1 SV=1
MLSPQRVAAAASRGADDAMESSKPGPVQVVLVQKDQHSFELDEKALASILLQDHIRLDLV
VVVSVAGAFRKGKSFILDFMLRYLYSQKESGHSNWLGDPEEPLTGFSWRGGSDPETGTGIQ
IWSEVFT...
```

Press “Continue” for the next screen.

2. The next screen allows you to set the parameters of the run. The default parameters are good starting values for MS data from high resolution Orbitrap instruments. Once you press submit the analysis starts. A new window will pop stating that the program is running. This window will disappear after 5 seconds,

and then there will be no other visible indication that the program is running in the background. We are working to add some visual indication about the progress of the analysis, but this is not ready yet. The run may take up to 30 minutes on a slower desktop computers.



The screenshot shows a window titled "User Parameters" with a list of settings. Each setting has a text input field. At the bottom, there are "Submit" and "Cancel" buttons.

Parameter	Value
tolerance MS1 (ppm)	4.5
tolerance MS2 (ppm)	6.5
linear_peptide_S2_Threshold	0.5
linear_peptide_Minimal_Number_by_total	9
cross_links_minimal_Frag_Num_for_each_peptide	11
cross_links_minimal_S2_for_each_peptide	0.3
cross_links_minimal_streak_on_each_peptide	3
cross_links_S2_Threshold	0.3
minimalPeptideLength	4
maximalPeptideLength	100
numberMiscleavages	2
maxVariableModificationsPerPeptide	0
useDecoySeq	<input checked="" type="checkbox"/> False <input type="checkbox"/> True
fractionOfSeqNumberForDecoys =	1 / 15
minNumOfSequencesForFraction	100
offsetMS1	0.0
offsetMS2	0.0

These fields are:

Tolerance MS1 – The maximal relative error (in ppm) of the total mass for reporting a cross-link.

Tolerance MS2 – The maximal relative error (in ppm) for an identification of a fragment in the MS/MS spectra.

Linear_peptide_S2_Threshold – The minimal score to assign a linear peptide to an MS/MS event. This score is defined as [total number of b- and y-ions] / [length of peptide]. If a linear peptide fits an MS/MS event above this threshold, then this event will not be searched further for a cross-linked peptide pair.

Linear_peptide_Minimal_number_by_total – In order for an MS/MS event to be assigned to a linear peptide (see previous field), it must also have at least this number of b- and y-fragments (in total).

cross_links_minimal_Frag_Number_for_each_peptide – An annotation of an MS/MS event as a putative cross-link is only reported if there is at least this number of fragments (total of b-, B-, y, and Y-ions) on each peptide.

cross_links_minimal_S2_for_each_peptide – An annotation of an MS/MS event as a putative cross-link is only reported if the ratio [total number of b- , B- , Y- , and y-ions] / [length of peptide] is above this number for each peptide.

cross_links_minimal_streak_on_each_peptide – An annotation of an MS/MS event as a putative cross-link is only reported if there is a streak of consecutive identified fragments [either b- , B- , Y- , and y-ions] longer than this number for each peptide.

cross_links_S2_Threshold – An annotation of an MS/MS event as a putative cross-link is only reported if the ratio [total number of b- , B- , Y- , and y-ions] / [total length of both peptides] is above this threshold.

minimalPeptideLength, maximalPeptideLength, numberMiscleavages – Parameters of the *in-silico* trypsin digestion.

maxVariableModificationsPerPeptide – Currently only the oxidized methionine modification is coded.

useDecoySeq – If you choose True, than the sequence database is spiked with reversed decoy sequences. Every **fractionOfSeqNumberForDecoys** sequence in the database is added in reverse for the decoy database.

minNumberOfSequencesForFraction – You can force to have at least this minimal number of decoys.

offsetMS1 – If the relative MS1 offset of the mass spectrometer is known, you can use this field to negate it. The value is in ppm.

offsetMS2 – If the relative MS2 offset of the mass spectrometer is known, you can use this field to negate it. The value is in ppm.

3. In the output file each row is an annotation of a specific MS/MS event as a possible cross-link. We recommend sorting these rows in Excel according to Column Q in decreasing order. This will sort the identifications in order of decreasing confidence. The columns are:

Column A – Name of first protein.

Column B – Name of second protein.

Column C – The ratio: [total number of fragments]/[total length of both peptides]. Usually, should be >1.5 for high confidence cross-link.

Column D – Number of fragments (total number of b-, B-, Y-, and y-ions) on first peptide. Usually, should be >18 for high confidence cross-link, and >15 for medium confidence cross-link.

Column E – Number of fragments (total number of b-, B-, Y-, and y-ions) on second peptide. Usually, should be >18 for high confidence cross-link, and >15 for medium confidence cross-link.

Column F – Number of the first residue in the first peptide.

Column G – Number of the first residue in the second peptide.

Column H – The ΔM (in ppm) between the measured and theoretical total mass. You can use the average of this column to negate the MS1 offset in the previous screen.

Column I – The first peptide.

Column J – The second peptide.

Column K – The cross-link mass (24Da or 12 Da) according to the following legend: “0” – 24 Da and no error in the identification of the mono-isotopic mass ; “1” – 24 Da and an error in the identification of the mono-isotopic mass by one peak shift ; “2” – 24 Da and an error in the identification of the mono-isotopic mass by two peak shift (this is the situation depicted in Figure S1) ; “3” – 12 Da and no error in the identification of the mono-isotopic mass ; “4” – 12 Da and an error in the identification of the mono-isotopic mass by one peak shift ; “5” – 12 Da and an error in the identification of the mono-isotopic mass by two peak shift (this is the situation depicted in Figure S1) ;

Column L – The mass of the precursor ion (Da).

Column M – The m/z of the precursor ion.

Column N – The charge of the precursor ion.

Column O – The retention time (seconds).

Column P – The number of the MS/MS event in the MGF file.

Column Q – The minimum between columns C and D, i.e. the number of fragments on the weaker peptide.

Column R – The average ΔM (in ppm) of the fragments in the MS/MS spectrum. You can use the average of this column to negate the MS2 offset in the previous screen.

Column S – The standard deviation (in ppm) of ΔM 's of the fragments in the MS/MS spectrum.