Direct Infusion Mass Spectrometry Processing (DIMaSP)

Instructions for use

The following instructions assume the user has a batch of paired sample and blank spectra that have been exported to a comma-separated values (.csv) format matching that of the proprietary software Xcalibur™ (Thermo Scientific™, Bremen, Germany). The same blank can be paired with multiple samples and samples are expected to have replicates. Each processed spectrum is expected to only allow formula assignment using up to 999 atoms of ^{12}C, ^{1}H, ^{16}O, ^{14}N, ^{32}S and up to 1 atom of Na, ^{13}C, and ^{34}S. If there is a mix of ionisation modes and/or polarities, it is best to separate them for analysis of mass shifts since many contaminants are polarity specific.

The instructions provide examples based on a set of two theoretical samples (Samples A and B) with three replicates (1, 2, and 3) with one blank per sample (Blanks A and B).

The scheme currently uses a series of scripts executable through Mathematica (Wolfram Research Inc., UK). The names of the required scripts (also known as notebooks) are:

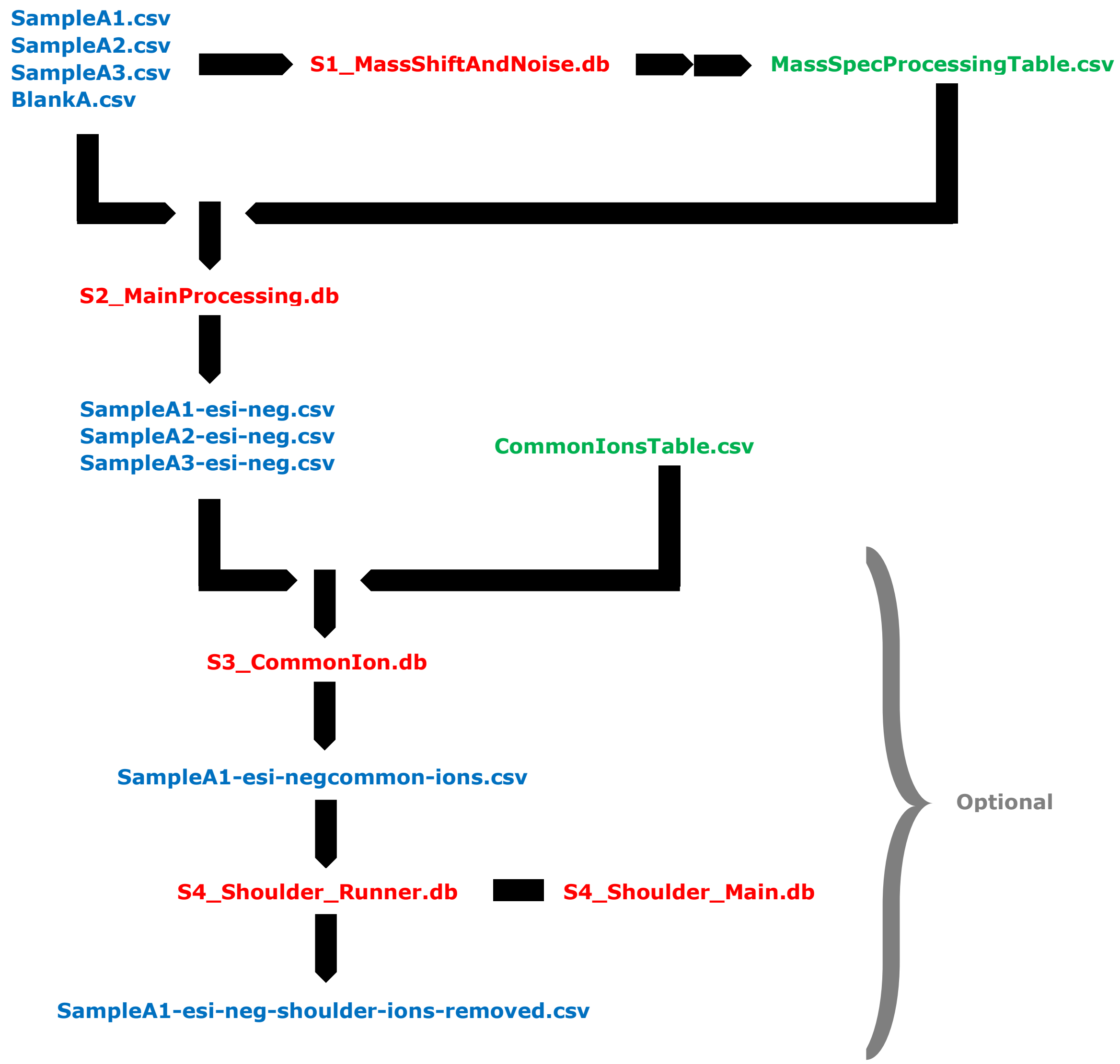
S1_MassShiftAndNoise.nb
S2_MainProcessing.nb
S3_CommonIon.nb
S4_Shoulder_Main.nb
S4_Shoulder_Runner.nb

A flow diagram of the processing scheme is shown on the next page.

Only one method of blank subtraction (by mass) is included here. If interested in blank subtraction by formula, please contact authors or wait for the release of the Python version of the code.

Note that that the notebooks must be run one at a time. When evaluating notebooks, pressing "Yes" or "No" are both acceptable if prompted "Do you want to automatically evaluate all the initialization cells in the notebook…".

Instructional flow diagram



Step 1: Extracting mass shift and noise parameters

Place all raw spectra (*e.g.* SampleA1.csv, SampleA2.csv, …, SampleB3.csv, BlankB.csv) in the same directory (*e.g.* C:\Users\Name\Raw) on your computer. Do not include any other .csv files (other than output from the notebook). Open the **S1_MassShiftAndNoise.nb** notebook and change the **folder** variable to match the given directory taking care to add an extra black slash (\) at each location to match the required syntax (*e.g.* **folder** = "C:\\Users\\Name\\Raw";).

Change the variable **formlist** to include a list of chemical formulae expected in the samples and/or blanks. These will be used to estimate the mass shift. The chemical formulae should be written in the order and format given by Xcalibur:

C# [13]C# H# O# N# Na# S# [34]S#

Where # denotes the number of assigned atoms for the given element and if #=0 the entire label (element and number) are removed. An example is
formlist = {"C9 H17 O2 ", " C10 H19 O2 ", " C12 H23 O2 ", " C14 H27 O2 ",
" C16 H31 O2 ", " C17 H31 O3 ", " C20 H39 O2 ", " C22 H43 O2 "};

The other options are described in Table 1 below and in the notebook itself. Once the options are set correctly, run the notebook be going to *Evaluate* > *Evaluate Notebook*. The main output will be a new file in the given directory named "Mass Shift and Noise Output" with the date and time appended to the end. The file will contain the individual mass shifts for each chemical assignment in each spectrum as well as the associated statistics for the mass shift and noise approximation.

While running the code will state the percentage completion and state if there are any outliers (for mass shift). Do not worry about the errors that appear if they state "Part 5 of {} does not exist". This is expected (see the Important Notes in the notebook itself).

Once complete, save and exit the notebook.

Step 2: Main processing sequence

This step performs the main processing of the data including blank subtraction and several filters such as the nitrogen rule check. An input table must be created before running next notebook (**S2_MainProcessing.nb**).

The input table consists of the extracted mass shift and noise parameters from the previous step and pairs the sample and blank filenames. The table can have as many rows as desired and can mix different ionisation modes. It must be saved in the .csv format. An example input table is shown in Table 8 with a description of each column value in Table 2. Note that the column order between the output of Step 1 and the input table match to make copying data easier.

Once the input table is completed, open the **S2_MainProcessing.nb** notebook and change the **folder** and **tablename** variables to match. The **folder** should contain the sample and blank .csv files as well as the input table. The **tablename** should be the name of with input table including the .csv extension (*e.g.* **tablename** = "MassSpecProcessingTable.csv").

As before, the other options are described in Table 3 below and in the notebook itself.

The code will show which file it is processing and, once that file is complete, it will state the number of assignments at various stages. The program will output a new file for each row in **tablename** with the filename being the sample name with the ionisation mode appended (*e.g.* SampleA1-esi-neg.csv).

Once complete, save and exit the notebook.

Step 3: Common ion selection (optional)

This step is an optional post-processing step. It is used when multiple repeats of the same sample have been analysed and keeps assignments present in a (user-defined) fraction of the repeats.

To begin, place all processed spectra (*e.g.* SampleA1-esi-neg.csv, SampleA2-esi-neg.csv, etc.) in the same directory (*e.g.* C:\Users\Name\Processed). Then create a new input table in the same directory. This step requires an input table to group replicates and state how many replicates must have the same assignment to keep it in the final output. See Table 9 for an example with a description of each column value in Table 4. Once again, this table must be saved in the .csv format.

Open the **S3_CommonIon.nb** notebook. Set the **directory** variable to be the path with the processed files (*e.g.* **directory** = "C:\\Users\\Name\\Processed";). Then set **tablename** to the new input table name (*e.g.* **tablename** = "CommonIonsTable.csv";). The other options are described in Table 5 below and in the notebook itself. Note that the number of replicates for each sample must be the same. If some samples have 3 replicates, while others have 4 you must process them separately.

The code will output a new file with "common-ions" appended to the end of the first sample in each set (*e.g.* "SampleA1-esi-negcommon-ions.csv"). It will be in a new folder within the **directory** (*e.g.* C:\Users\Name\Processed\Common_Ion_Folder). In a similar fashion, a folder full of plots will be saved within a new folder (*e.g.* C:\Users\Name\Processed\Plots_Common_Ion_Folder). Plots are named the same as the output file.

Once complete, save and exit the notebook.

Step 4: Shoulder ion removal (optional)

This step is an optional post-processing step. It is used to remove any shoulder ions in the final spectra and takes the output of either Step 3 (if applicable) or Step 2 (if step 3 was skipped).

This step requires two notebooks: **S4_Shoulder_Main.nb** and **S4_Shoulder_Runner.nb**. The "main" notebook contains the code involved in removing shoulder ions while the "runner" code allows for processing multiple files.

Place the files to be processed in a new directory (*e.g.* C:\Users\Name\Shoulder) as well as both notebooks. Then set the **directory** variable the "runner" notebook to match (*e.g.* **directory** = "C:\\Users\\Name\\Shoulder";). The other options are described in Tables 6 and 7. Note that options exist in the "main" code

Once ready, evaluate the **S4_Shoulder_Runner.nb** notebook to process all the .csv files in the folder. Therefore, only have the notebooks and files you wish to process in the **directory**. Note that only is required **S4_Shoulder_Runner.nb** to be open when evaluating (**S4_Shoulder_Main.nb** can be closed).

The code will output two new folders with the final files in the **directory**\Shoulder_Ion_Folder and plots of the shoulder ions in **directory**\Plots_Shoulder_Ion_Folder.

Once complete, save and exit the notebook.

Table 1: Options for running S1_MassShiftAndNoise.nb

Parameter Name	Description	Sample Values
folder	Folder directory containing files to analyse. Sets data to be analysed.	"C:\\Users\\Name\\Raw"[a]
formlist	List of chemical formulae used to check for mass error. Can be adjusted for different known contaminants but should have at least 10 compounds.	{"C8 H13 O2 ", "C9 H17 O2 ", "C10 H19 O2 ", "C14 H27 O2 "}[b]
α	Confidence level for Grubbs' Test. Used to determine whether a given mass error is an outlier.	0.01
nmaxlimit	Intensity ceiling for fitting purposes of noise to speed up processing. The lower the limit, the quicker the processing becomes.	5000
exportfits	Boolean variable for if noise-fitted plots should be exported.	0 (no) or 1 (yes)
upperppm	Sets the additive mass shift term used to provide an upper mass shift limit (in ppm).	0.5
lowerppm	Sets the subtractive mass shift term used to provide a lower mass shift limit (in ppm).	0.5
minimumbinsize	Sets minimum bin width (in intensity units) for the noise histogram used to estimate noise level. Bin size is otherwise calculated using Freedman-Diaconis rule.	1

Table 2: Input table parameters for S2_MainProcessing.nb

Parameter Name	Description	Sample Values
Filename	Sample filename with matching data to follow (pairs with blank filename below). Filename should point towards a .csv file.	SampleA1
Sample Mass Shift (ppm)	Sample mass shift mean in ppm.[c]	-0.172941176
Sample Mass Shift Standard Deviation (ppm)	Sample mass shift standard deviation in ppm.[c]	0.659874654
Number of peaks used to estimate SAMPLE mass shift	Number of peaks used to estimate sample mass shift parameters.[c]	17
Sample Noise Limit	Noise limit of the sample file.[c]	239.8823901
Sample upper ppm limit	Upper limit of the sample mass shift.[c]	2
Sample lower ppm limit	Lower limit of the sample mass shift.[c]	-1.59
Blank Filename	Blank filename with matching data to follow (pairs with blank filename above). Filename should point towards a .csv file.	BlankA
Blank Mass Shift (ppm)	Blank mass shift mean in ppm.[c]	-0.044705882
Blank Mass Shift Standard Deviation (ppm)	Blank mass shift standard deviation in ppm.[c]	0.271503169
Number of peaks used to estimate BLANK mass shift	Number of peaks used to estimate blank mass shift parameters.[c]	17
Blank Noise Limit	Noise limit of the blank file.[c]	128.0483771
Minimum Sample to Blank Ratio	Sample to blank ratio used for blank subtraction (for intensities).	10
Mode and Polarity (ESI NEG, ESI POS, APPI NEG, or APPI POS)	Mode and polarity used for mass spectrometry.	ESI NEG or ESI POS or APPI NEG or APPI POS

Table 3: Options for running S2_MainProcessing.nb

Parameter Name	Description	Sample Values
`folder`	Folder directory containing files (samples and blanks) to analyse and input table. Sets data to be analysed.	`"C:\\Users\\Name\\Raw"`[a]
`tablename`	Table filename including extension (i.e. ".csv") including the information as described above.	`"MassSpecProcessingTable.csv"`
`deleteduplicates`	Boolean variable for controlling if duplicate formulae (for a given peak) should be removed (minimum absolute value of corrected delta is kept).	`0` (no) or `1` (yes)
`hcmin`	Minimum allowable H/C ratio.	`0.3`
`hcmax`	Maximum allowable H/C ratio.	`2.5`
`ocmax`	Maximum allowable O/C ratio.	`2`
`ocmin`	Minimum allowable O/C ratio.	`0`
`ncmax`	Maximum allowable N/C ratio.	`1.3`
`scmax`	Maximum allowable S/C ratio.	`0.8`
`upper13c`	Boolean toggle for checking isotopic ratio (13C) - should not be adjusted.	`1`
`lower13c`	Boolean toggle for checking isotopic ratio (13C) - should not be adjusted.	`0`
`upper34s`	Boolean toggle for checking isotopic ratio (34S) - should not be adjusted.	1
`lower34s`	Boolean toggle for checking isotopic ratio (34S) - should not be adjusted.	`0`
`c1213`	Isotopic ratio for 12C/13C check. Typically based on natural abundance of isotope.	`0.011`
`s3234`	Isotopic ratio for 32S/34S check. Typically based on natural abundance of isotope.	`0.045`

Table 4: Input table parameters for S3_CommonIon.nb		
Parameter Name	**Description**	**Sample Values**
Accessed name	New name repeated samples in format rep%%-## where the %% value is the set number and the ## value is the number within a set.	rep1-1
File name	Actual filename corresponding to the newly chosen accessed name. Does not include file extension.	Sample-1
Common ions	Number of files within a given set (i.e. same %% designation) that a peak must appear to be kept. If equal to maximum ## then must be present in all repeats.	3

Table 5: Options for running S3_CommonIon.nb		
Parameter Name	**Description**	**Sample Values**
directory	Folder directory containing files (sample repeats) to analyse and input table. Sets data to be analysed.	"C:\\Users\\Name\\Processed"[a]
mzmin	Minimum m/z range to be plotted. All peaks analysed.	100
mzmax	Maximum m/z range to be plotted. All peaks analysed.	900
absintensitymin	Minimum intensity to be plotted. All peaks analysed.	0
absintensitymax	Maximum intensity to be plotted. All peaks analysed.	5000000
tablename	Table filename including extension (i.e. ".csv") including the information as described above.	"CommonIonsTable.csv"
numberofreplicates	Number of replicates in each set (%%).[d]	3

Table 6: Options for running S4_Shoulder_Main.nb

Parameter Name	Description	Sample Values
intensitycutoff	Intensity cut off for considering peaks which may have shoulder ions	1000000
shouldermassrange	Mass range (plus minus) in m/z units to consider peaks potential shoulder ions.	0.01
percentparentiontoshoulderion	Percentage cut off for relative intensity based on local peaks. Peaks below cut off (and within mass range) are considered shoulder ions.	1
mzmin	Minimum m/z range to be plotted. All peaks analysed.	100
mzmax	Maximum m/z range to be plotted. All peaks analysed.	900
absintensitymin	Minimum intensity to be plotted. All peaks analysed.	0
absintensitymax	Maximum intensity to be plotted. All peaks analysed.	5000000

Table 7: Options for running S4_Shoulder_Runner.nb

Parameter Name	Description	Sample Values
directory	Folder directory containing main code (i.e. Shoulder Ions code above) and .csv files together.	"C:\\Users\\Name\\Shoulder"[a]
main	Name of main shoulder ions code with extension.	"S4_Shoulder_Main.nb"

Footnotes:

Table parameters (as described in blue sections) should be set as column headers in a .csv file with values put into rows below it for as many files as needed

[a] "\\" denotes the typical "\" for describing computer directories in the Mathematica syntax

[b] Additional space after each formula is required to match the format of the Xcalibur output. Note that the formulae shown are for illustrative purposes only.

[c] Output from **S1_MassShiftAndNoise.nb**. All data is sorted in output for ease of copying for the input table.

[d] Must be the same for all files listed in the table for analysis. Different input tables are required for different replicate numbers.

Table 8: Sample input table for S2_MainProcessing.nb (simulated to match Excel .csv view):

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Sample Filename	Sample Mass Shift (ppm)	Sample Mass Shift Standard Deviation (ppm)	Number of peaks used to estimate SAMPLE mass shift	Sample Noise Limit	Sample upper ppm limit	Sample lower ppm limit	Blank Filename	Blank Mass Shift (ppm)	Blank Mass Shift Standard Deviation (ppm)	Number of peaks used to estimate BLANK mass shift	Blank Noise Limit	Minimum Sample to Blank Ratio	Mode and Polarity (ESI NEG, ESI POS, APPI NEG, or APPI POS)
2	SampleA1	-0.172	0.659	17	239.8824	2	-1.59	BlankA	-0.0447	0.271	17	128.0	10	ESI NEG
3	SampleA2	-0.150	0.565	17	378.8827	1.91	-1.28	BlankA	-0.0447	0.271	17	128.0	10	ESI NEG
4	SampleA3	-0.047	0.464	18	288.4917	1.73	-1.27	BlankA	-0.0447	0.271	17	128.0	10	ESI NEG
5	SampleB1	0.275	0.205	10	4640.827	0.97	-0.44	BlankB	0.3275	0.377	9	95.5	10	APPI POS
6	SampleB2	0.325	0.231	8	4761.559	1.06	-0.42	BlankB	0.3275	0.377	9	95.5	10	APPI POS
7	SampleB3	0.300	0.218	9	4746.917	1.01	-0.43	BlankB	0.3275	0.377	8	95.5	10	APPI POS

Table 9: Sample input table for S3_CommonIon.nb (simulated to match Excel .csv view):

	A	B	C
1	Accessed name	File name	common ions
2	rep1-1	SampleA1-esi-neg	3
3	rep1-2	SampleA2-esi-neg	3
4	rep1-3	SampleA3-esi-neg	3
5	rep2-1	SampleB1-appi-pos	3
6	rep2-2	SampleB1-appi-pos	3
7	rep2-3	SampleB1-appi-pos	3