DIA-Umpire signal extraction module (Step A)

DIA_Umpire_SE.jar provides the signal extraction module for DIA data (regular SWATH with fixed isolation window size, variable window SWATH, MSX, MS^E) which generates pseudo MS/MS spectra to be searched against a protein database using conventional proteomics search engines such as X!Tandem, SEQUEST, MSGF+, OMSSA, etc.

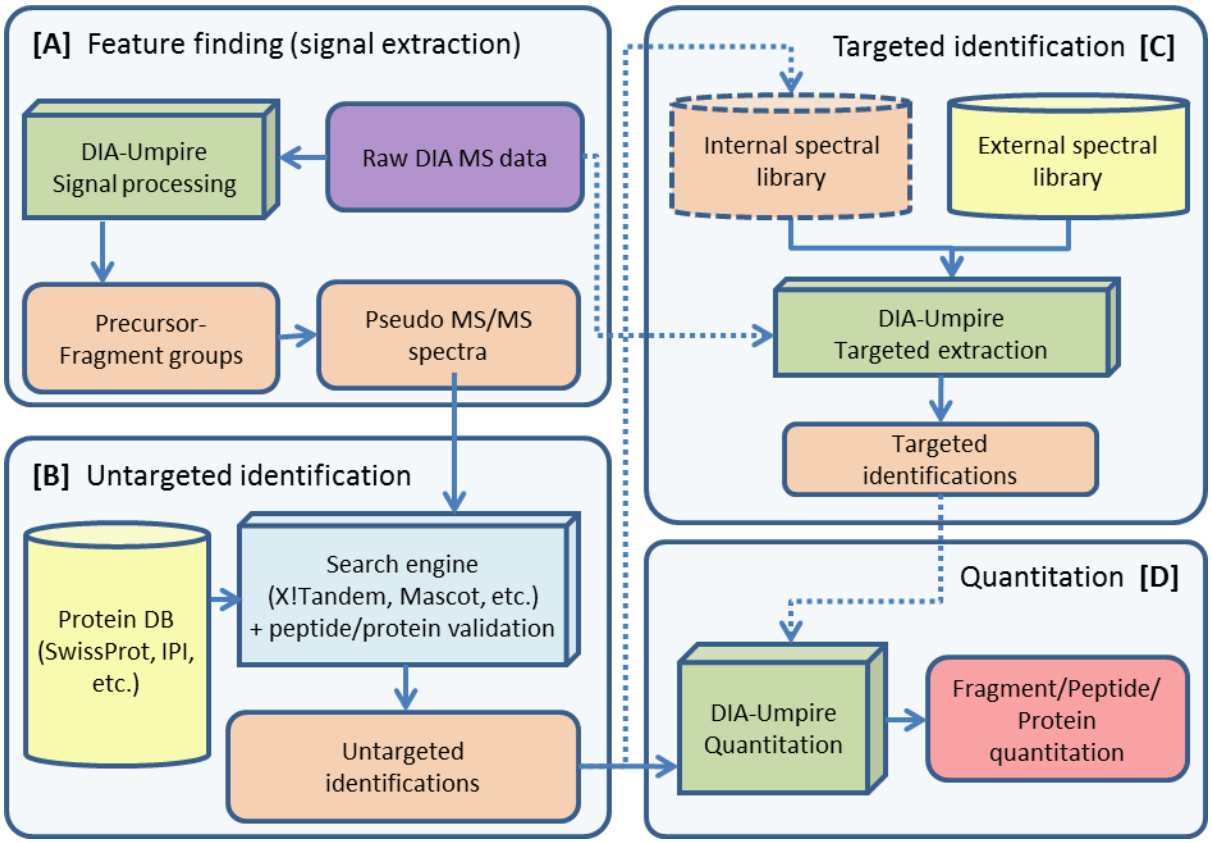

Figure 1. DIA-Umpire hybrid workflow. Green boxes are DIA-Umpire modules, yellow – external data, orange – intermediate generated data, purple – input experimental data, red – output quantitation results from the whole pipeline.

Requirements

Java 7 or higher (download link: Java SE Runtime Environment 7).

As a rule of thumb, it is recommended have at least double the amount of RAM as the average size of your *mzXML* files (*mzXML* written in 32-bit format without zlib compression). If *mzXML* is in 64-bit format, then RAM requirements should be approximately the size of the file.

Input (DIA-Umpire signal extraction module)

1. Spectral data in *mzXML* format

   ***Important*:** for data from AB SCIEX instruments, use AB SCIEX MS Data Converter:

   - http://www.AB SCIEX.com/downloads/software-downloads

   Use it for *.wiff* -> *.mzML* conversion, then use MSConvert for *.mzML* -> *.mzXML*. Read "Raw spectral data files conversion to mzXML" section on page 16 for more details.

2. Parameter file (An example "diaumpire.se_params" is provided with the package)

Runtime parameters (parameter file)

Precursor-fragment grouping parameters

PrecursorRank: (integer) Determines how many precursors a single fragment is allowed to be grouped to. Precursors are first sorted by Pearson correlation of elution profiles; this option specifies the rank of a precursor in this sorted list. Lowering the value for this parameter increases the stringency of precursor-fragments grouping.

FragmentRank: (integer) Determines how many fragments a single precursor is allowed to have. Fragments are first sorted by Pearson correlation of elution profiles; this option specifies the rank of a fragment in this sorted list. The lower - the more stringent.

CorrThreshold: (0.0~1.0) Minimum Pearson correlation between a precursor and a fragment to be considered, the higher, the more stringent.

DeltaApex: (Unit: minute) Maximum retention time difference of LC profile apexes between precursor and fragment (the lower, the more stringent).

BoostComplementaryIon: (true or false) set to ***true*** if you want to boost complementary ions' intensity

AdjustFragIntensity: (true or false) set to ***true*** if you want to adjust fragment intensity by the Pearson correlation between a precursor and a fragment.

Signal extraction parameters

SE.MS1PPM: (Unit: ppm) Maximum mass error for two MS1 peaks in consecutive spectra to be considered signal of the same ion. Used in MS1 signal detection and precursor alignment between samples/runs.

Recommended value: Depends on the instrument. Typical values are 5-10 ppm for Thermo Orbitrap, 20-40pm for AB SCIEX Triple TOF 5600.

SE.MS2PPM: (Unit: ppm) Maximum mass error for two MS2 peaks in consecutive spectra to be considered signal of the same ion.

Recommended value: Depends on the instrument. If fragmentation spectra are measured with the same detector as MS1 spectra, set the same as Para.MS1PPM or a little higher, e.g. if you've set Para.MS1PPM=30 ppm for AB SCIEX Triple TOF 5600, consider setting to 40ppm.

SE.MinMSIntensity: Minimum signal intensity for a peak in an MS1 spectrum to be considered as a valid signal. Any MS1 peak having intensity lower than this threshold will be ignored. It is the main parameter controlling how many peaks and isotopic envelopes will be detected.

Recommended value: Depends on the data. Check raw data for average noise-levels. E.g. TOF data often have thousands of random small intensity peaks.
Warning: Setting this parameter too low (or zero) in such a case will significantly increase processing time and memory requirements.

SE.MinMSMSIntensity: Same as *para.MinMSIntensity*, but for MS2 signals.

SE.MaxCurveRTRange: (Unit: minute) The maximum allowed retention time (RT) range for elution profile of a single ion. If a detected elution profile exceeds that time span, it will be trimmed around the apex to fit into this range. Used to avoid having lots of ions which elute during the whole LC/MS run or over a very long period of time, as this greatly complicates grouping of precursors to fragments. Such long-eluting ions are likely to be contaminants, lock-mass ions, calibrants, etc.

Recommended value: The expected maximum peak chromatographic time. E.g. set to several percent of the whole run time, if the run was 100 min long, set to 5 min.

SE.SN: Minimum signal-to-noise threshold for MS1 precursor signal detection. It is not the real S/N value, but rather a multiplier for *para.MinMSIntensity*, if a detected elution profile is less intense in the apex than (*para.SN* x *para.MinMSIntensity*) it will be discarded.

Recommended value: Typical values depend on the *para.MinMSIntensity* setting. If you've set *para.MinMSIntensity* to a very low value, consider setting this one to some small number in range 1.0 – 5.0.

SE.MS2SN: Same as *para.SN*, but for possible unfragmented precursors in MS2 data (i.e. for selecting precursors to generate Q3 tier pseudo spectra).

SE.Resolution: Used only if the input spectra are stored in profile mode (i.e. not centroided, e.g. by using "Peak Picking" option in MSConvert when converting raw spectral data to mzXML format).

Profile spectra will be centroided using a sliding window. The window is moved across the entire mass range of a spectrum. Only the most intense peak in the window centered at the peak m/z is kept, others are discarded. The window width is calculated based on this parameter as: *width = mz / para.Resolution*.

Recommended value: Depends on the instrument and acquisition settings. Either check raw data to see the real average resolution of peaks in spectra or consult vendor specifications for the instrument. For AB SCIEX TripleTOF 5600 we use 15000-20000.

SE.StartCharge: The minimum charge state for MS1 precursor ion to be detected during isotopic peak grouping.

SE.EndCharge: The maximum charge state for MS1 precursor ion to be detected during isotopic peak grouping.

Recommended value: it is not recommended to set this parameter higher than 5 for typical proteomic experiments, as it is unlikely to observe peptides of higher charge states.

SE.MS2StartCharge: The minimum charge state for MS2 ***unfragmented precursor*** ion to be detected during isotopic peak grouping.

SE.MS2EndCharge: The maximum charge state for MS2 ***unfragmented precursor*** ion to be detected during isotopic peak grouping.

Recommended value: it is not recommended to set this parameter higher than 5 for typical proteomic experiments, as it is unlikely to observe peptides of higher charge states.

SE.NoMissedScan: Maximum number of consecutive "gaps" allowed during extraction of elution profile (scans, in which the precursor mass being traced was not detected). E.g. if set to 1 and a particular mass can be found at every second scan, the algorithm will trace such a peak unless it can't find the peak in 2 scans in a row.

SE. MinFrag: Minimum number of fragments for a precursor. Precursors which have less than the set number of fragments will be removed from pseudo MS/MS spectra.

SE. EstimateBG: (true or false) set to ***true*** if you want to perform the background detection algorithm to determine minimum intensities for MS1 and MS2 spectra. Note that the settings of SE.MinMSMSIntensity and SE.MinMSIntensity will be ignored.

DIA isolation window settings

WindowType: DIA experiment type. DIA is implemented differently by different vendors and current support for data-formats is lacking, so the program needs additional info to properly interpret input spectral data.

Supported values in this version:

- SWATH - fixed window size SWATH, as described in the original SWATH paper. If you're using this option, it's mandatory to specify ***WindowSize*** option as well.
- V_SWATH - variable window size SWATH. If you're using this option, it's mandatory to specify ***Variable SWATH window setting*** (see section below).
- MSX – 2Da isolation window, its position is shuffled randomly until the whole MS1 range is covered, the process is then repeated but coverage of MS1 range by isolation windows will be different because of randomization.
- MSE – as originally implemented in Waters instruments. The full MS1 range is being fragmented at once.

WindowSize: Isolation window size setting for fixed window SWATH.

***Note*:** The window size is to be specified including overlapping regions. I.e. if your windows are: 399.5-425.5, 424.5 – 450.5, etc., then the window size should be set to 26.

Note: Was tested only on AB SCIEX TripleTOF 5600 and Thermo Q-Exactive data.

Variable SWATH window setting: Isolation settings for variable window size SWATH. Should be a tab-delimited list of m/z low and high values, one window per row. List begins with "==window setting begin" on a separate line and ends at "==window setting end". Example (2 windows: 400-451m/z and 449-600m/z):

```
==window setting begin
400	451
449	600
==window setting end
```

Other parameters

Thread: the maximum number of processing threads to be used.

ExportPrecursorPeak: set to ***true*** if you want detailed information about detected MS1 signals to be written to plain text file.