Peak List Generation

The output type of a mass spectrometer varies depending on the vendor, and the first step of the workflow consists of converting these raw (binary) files. The converter of choice is MSConvert, part of the Proteowizard[1](#_ENREF_1) package, and the standard format to convert to is mzML[**2**](#_ENREF_2). An mzML file contains all the unprocessed spectra (MS1 and MS2) plus additional spectrum and instrument annotation.

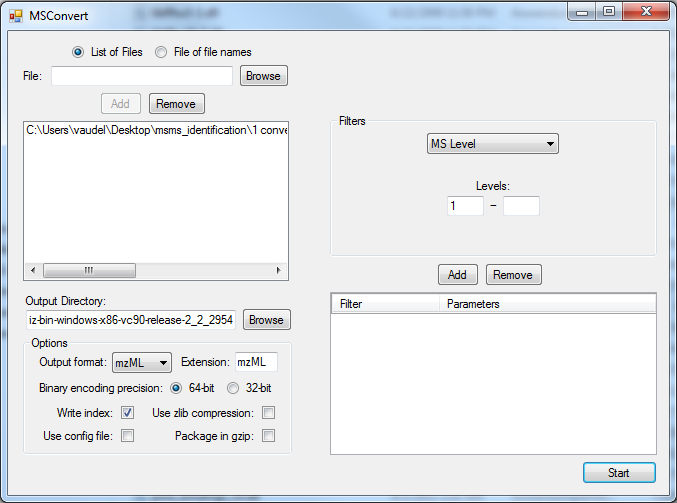
Note that the output of some instruments will require specific signal processing operations which can be performed by tools such as OpenMS[3](#_ENREF_3). For details on how to use OpenMS see <http://open-ms.sourceforge.net>.

Although mzML is the recommended format, the community often uses the mgf format for spectrum identification (mascot generic format, see [www.matrixscience.com/help/data\_file\_help.html#GEN](www.matrixscience.com/help/data_file_help.html%23GEN)). We will therefore convert the raw data into mgf which only contains MS/MS peak lists with some basic information about the precursor.

Raw Files Conversion

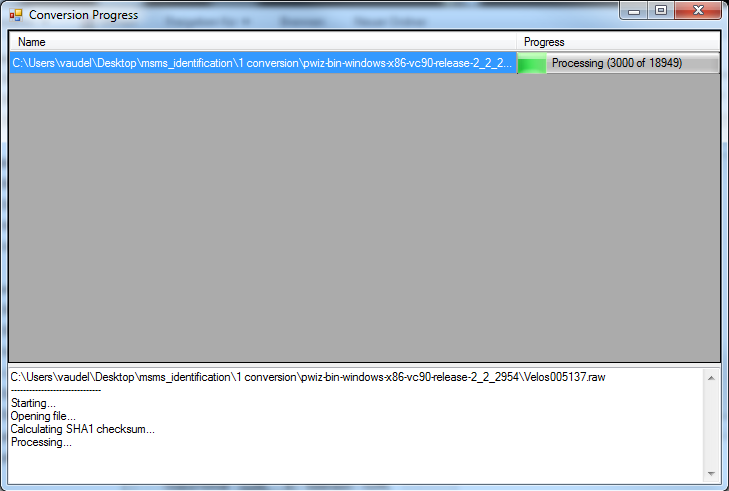
Note that this step is usually platform dependent as it requires vendor libraries. We will describe Windows usage. See <http://proteowizard.sourceforge.net/formats.shtml> for more information about the support for raw data formats.

In the resources folder, you will find a file generated by an Orbitrap Velos (Thermo Scientific, .raw file): Velos005137.raw, and in the software folder, in the proteowizard folder: proteowizard windows-x86-vc90-release-2\_2\_2954. Inside this folder, double click on MSConvertGUI.exe, you should see the MSConvert Graphical User Interface (GUI):



Use the ‘Browse’ button to select our raw file: Velos005137.raw and click ‘Add’. Note that the selected format is mzML, change this to mgf, and then select the output directory.

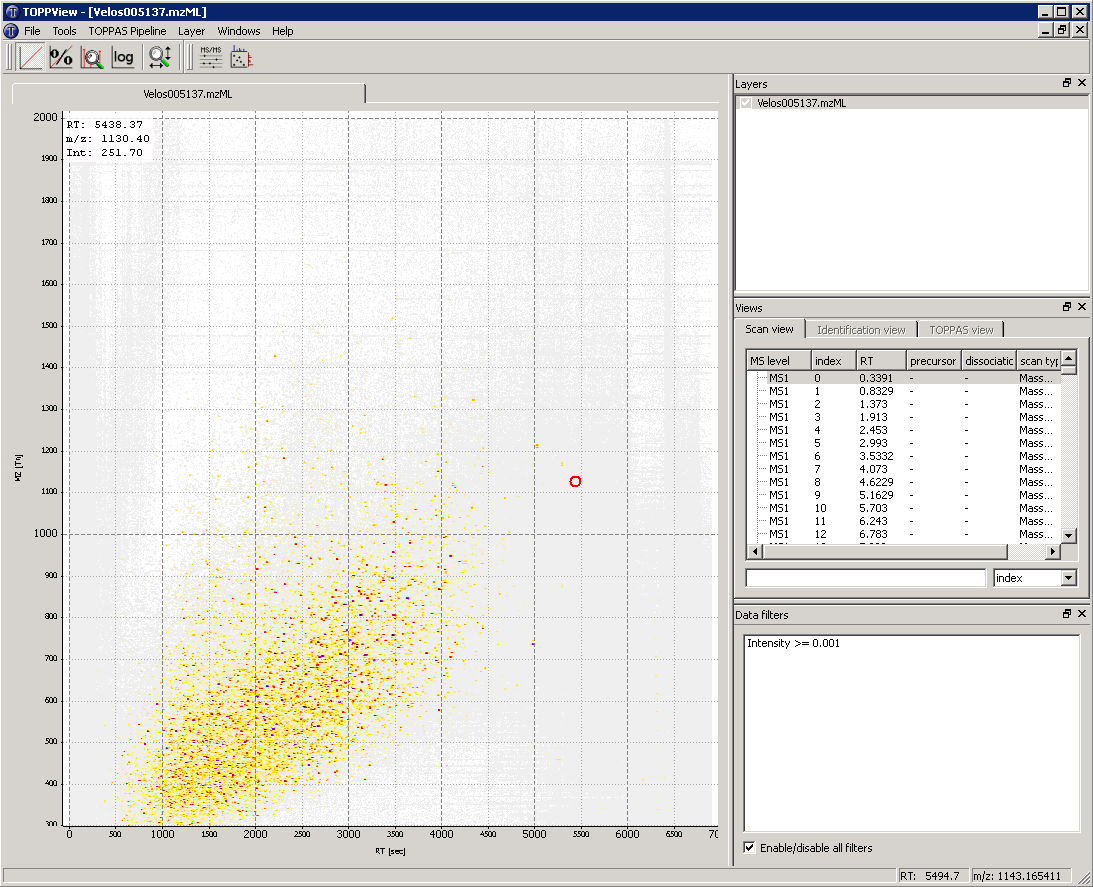
When clicking ‘Start’, the following screen will appear and the file Velos005137.mgf will be generated in the location specified.



MS/MS Processing

Depending on the mass spectrometer, the MS/MS spectra used for identification will require different processing steps. Note that this step is crucial as any imprecision made at this point will affect the rest of the workflow. The tool of choice for processing spectra is OpenMS[**3**](#_ENREF_3). OpenMS is a suite of tools – so called TOPP tools – dedicated to gel free proteomics, it can be downloaded and installed from <http://open-ms.sourceforge.net>.

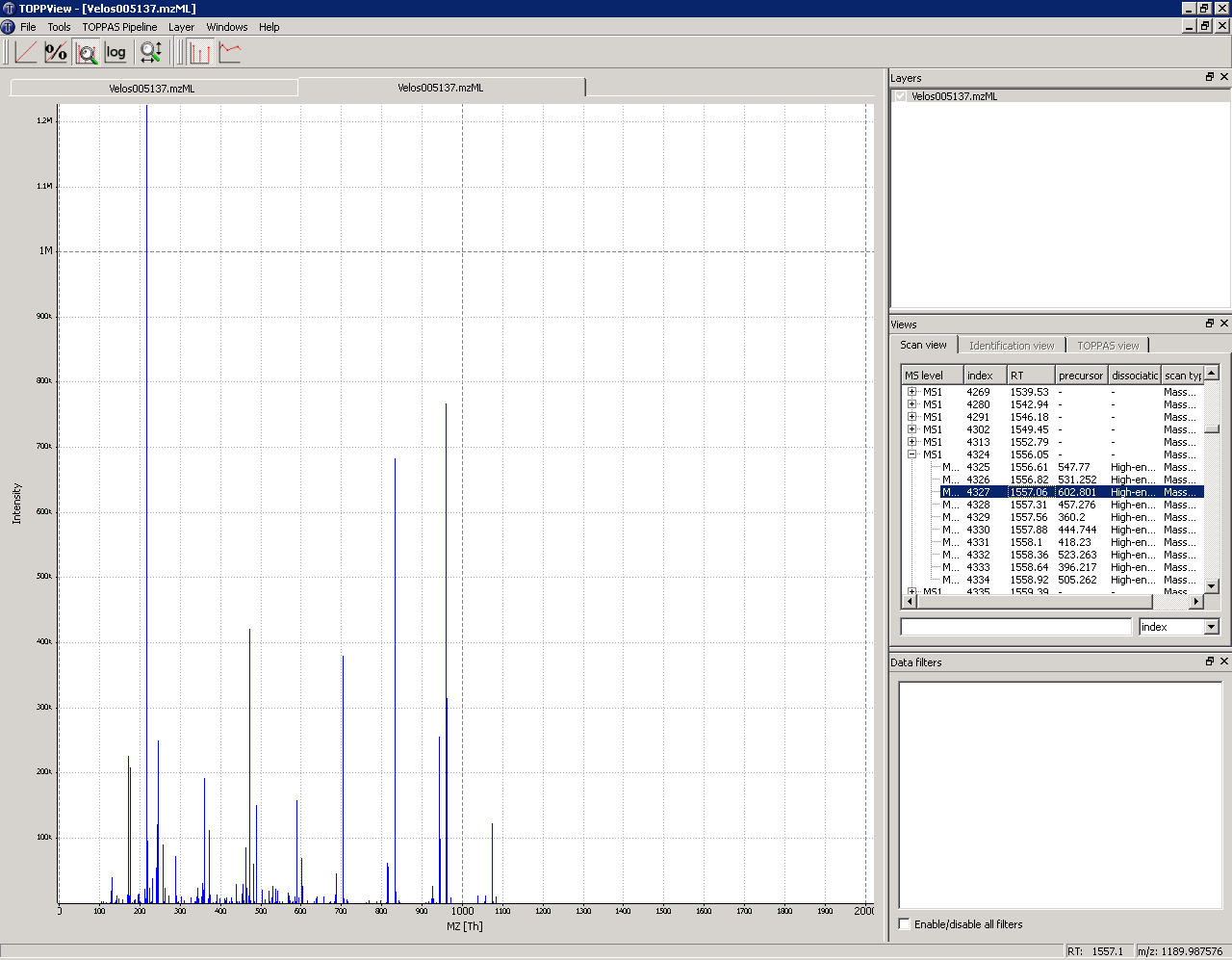
Two graphical interfaces allow you to look at your data (TOPPview) and to draw pipelines (TOPPAS). Open the previously generated mzML file (note, the file is also in the resources folder) with TOPPview, you should see this screen:



You recognize here the MS1 map of intensities as acquired by the mass spectrometer in the two dimensions of retention time and mass over charge ratio (m/z). After zooming, you can toggle the 3D view with a right-click on the map:



You can select an MS/MS spectrum using the scan view on the right side of the interface:

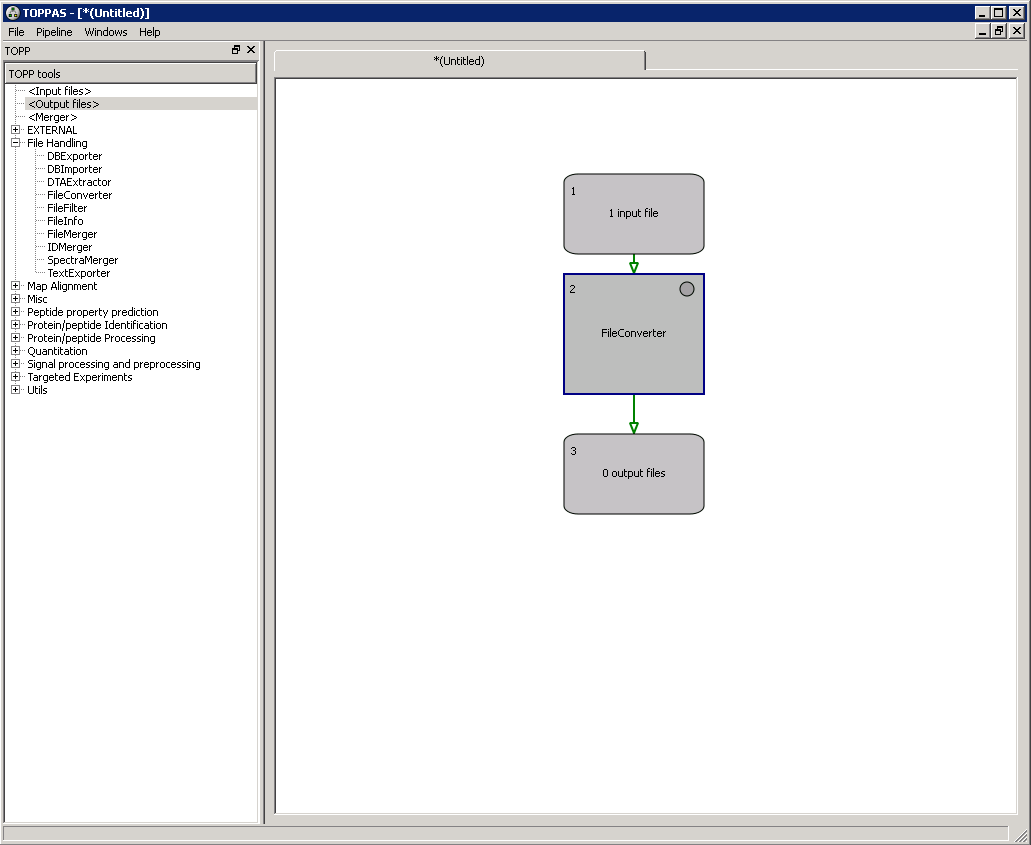


*Looking at the spectrum, what processing steps are necessary here?*

This example was acquired on a high resolution mass spectrometer in centroid mode. There is thus no specific processing required. The processing steps which are usually needed are:

* Baseline reduction: typically for TOF mass spectrometers, use this tool if the zero-intensity line of the spectrum is not stable or presents an offset.
* Noise filtering: for low resolution mass spectrometers, OpenMS provides a Savitzky-Golay[4](#_ENREF_4) filter in order to reduce the noise.
* Peak-picking:when the data is acquired in profile mode, every peak consists of several points which need to be summarized into one single peak before further processing. This step reduces the amount of data to be handled in the following. OpenMS provides two peak-pickers: a wavelet based peak-picker dedicated to low resolution mass spectrometers and a high resolution peak-picker for high resolution mass spectrometers. OpenMS peak-pickers are usually more efficient than the vendor’s peak-pickers[5](#_ENREF_5), they are thus advised for quantitative studies[6](#_ENREF_6). All these tools can be applied by the TOPPAS interface.

In this example we simply need to convert the mzML file into an mgf file for later processing. Start TOPPAS and draw the following pipeline: input file -> file converter -> output file. If you double click on the corresponding box, you can select the input file and set the file converter output type to 'mgf':



Select 'Pipeline' -> 'Run' in the menu at the top, an mgf file containing all our processed MS/MS spectra will be produced. *Note that OpenMS comes with other examples if you need more advanced workflows.*

Here is an example of a single MS/MS spectrum exported from an mgf file:

BEGIN IONS

TITLE=824.836730957031\_212.9232

PEPMASS=824.836730957031

RTINSECONDS=212.9232

CHARGE=2

118.936477661133 429.616

122.26781463623 354.588

138.923324584961 369.316

188.516448974609 367.936

268.502807617188 408.742

269.640869140625 414.032

291.39013671875 405.47

301.587707519531 425.719

326.118194580078 429.315

340.996948242188 657.018

355.069671630859 2271.57

357.8935546875 477.306

708.145690917969 479.876

731.38818359375 348.131

1201.05639648438 385.268

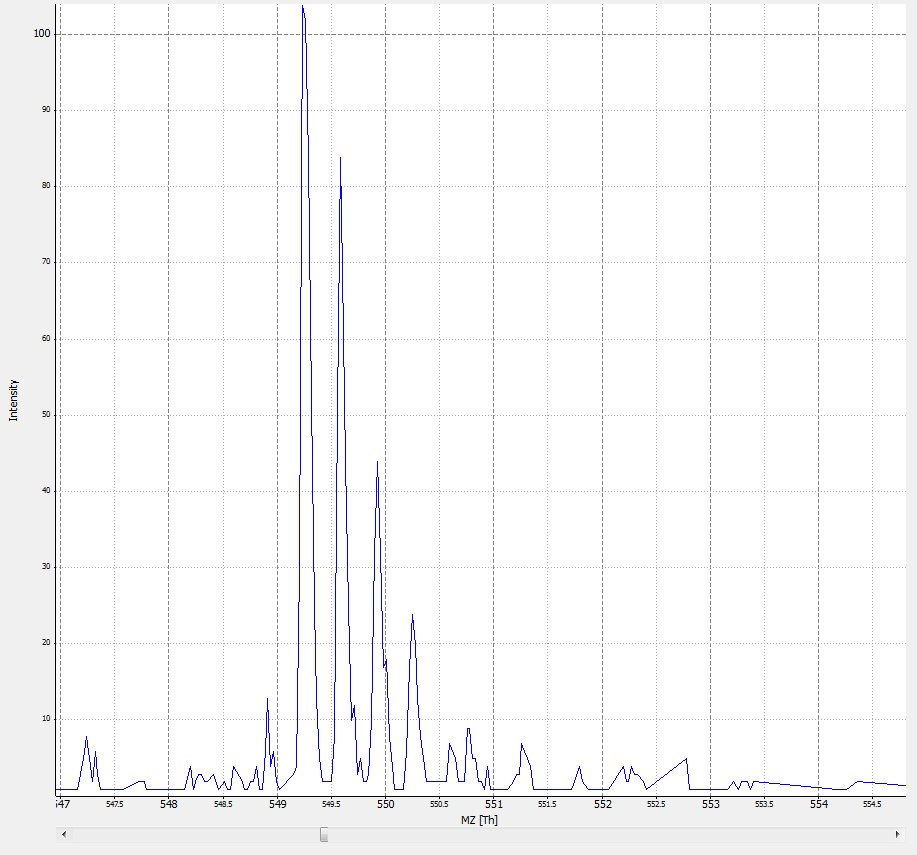
1364.15832519531 385.311

END IONS

*What information do you have about the spectrum? What is missing in comparison with the mzML file viewed in OpenMS?*

Advanced

In the resources folder, you will find the file QstarE04588.mzML obtained on another mass spectrometer. Load it in TOPPview, if you zoom, you should see these kinds of features in MS/MS spectra:



Design a workflow in TOPPAS and generate peak lists for these spectra.

References

(1) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **2008**, *24*, 2534.

(2) Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Rompp, A.; Neumann, S.; Pizarro, A. D.; Montecchi-Palazzi, L.; Tasman, N.; Coleman, M.; Reisinger, F.; Souda, P.; Hermjakob, H.; Binz, P. A.; Deutsch, E. W. mzML--a community standard for mass spectrometry data. *Mol Cell Proteomics* **2011**, *10*, R110 000133.

(3) Bertsch, A.; Gropl, C.; Reinert, K.; Kohlbacher, O. OpenMS and TOPP: open source software for LC-MS data analysis. *Methods Mol Biol* **2011**, *696*, 353.

(4) Savitzky, A.; Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry* **1964**, *36*, 1627.

(5) Lange, E.; Gropl, C.; Reinert, K.; Kohlbacher, O.; Hildebrandt, A. High-accuracy peak picking of proteomics data using wavelet techniques. *Pac Symp Biocomput* **2006**, 243.

(6) Vaudel, M.; Sickmann, A.; Martens, L. Peptide and protein quantification: a map of the minefield. *Proteomics* **2010**, *10*, 650.