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" \o "Kessner, 2008 #14) package, and the standard format to convert to is mzML**[2](#_ENREF_2" \o "Martens, 2011 #13)**. An mzML file contains all the unprocessed spectra (MS1 and MS2) plus additional spectrum and instrument annotation.

Although mzML is the recommended data format, the community often prefers the simpler mgf format for spectrum identification (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, as displayed below:

BEGIN IONS

TITLE=spectrum\_title

RTINSECONDS=173.824

PEPMASS=1467.61962890625 1671.478515625

CHARGE=3

438.2539978 3469.398926

470.861908 1319.134888

472.9784241 1250.332031

479.104187 1144.104004

511.6372375 1121.553833

.

.

.

1938.558105 1074.815063

1952.809082 1343.581421

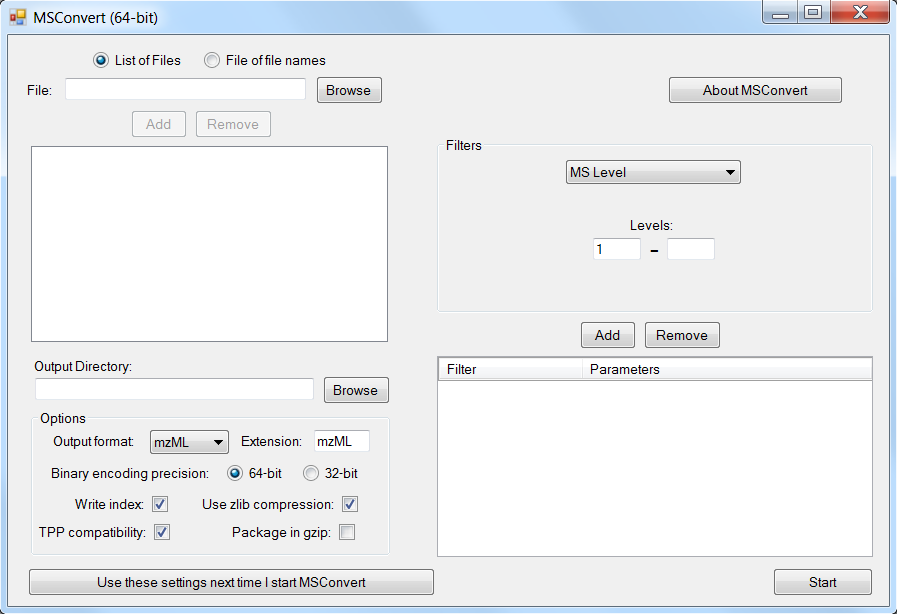
1986.876587 2016.473755

END IONS

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 an example file generated by a Q Exactive (Thermo Scientific, .raw file): qExactive01819.raw. Start MSConvertGUI.exe, delivered with the Proteowizard package, you should see the MSConvert Graphical User Interface (GUI):

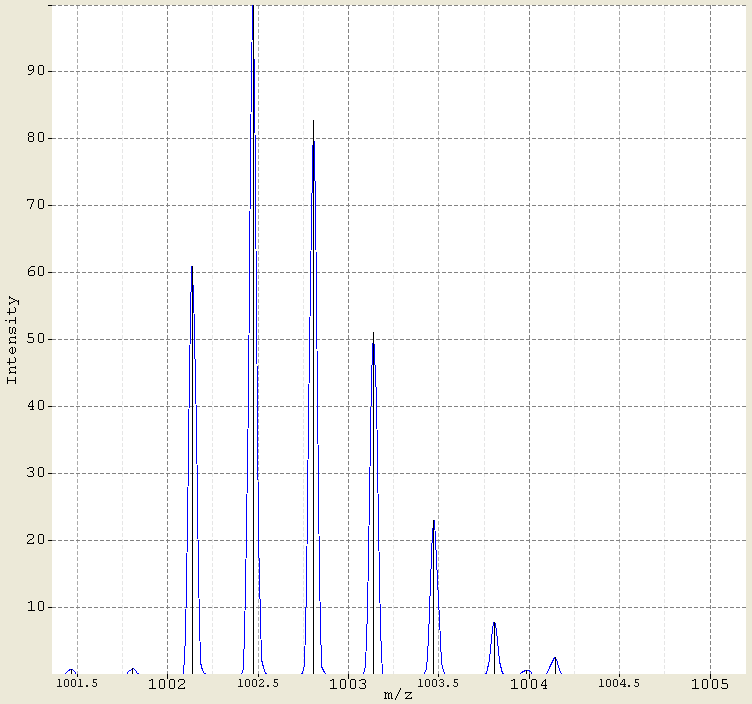


Using the Browse button, select the qExactive01819.raw file and click on Add. Note that MSConvert can process files in parallel very effectively. In the ‘Options’ panel, chose mgf as Output format and leave other settings to default.

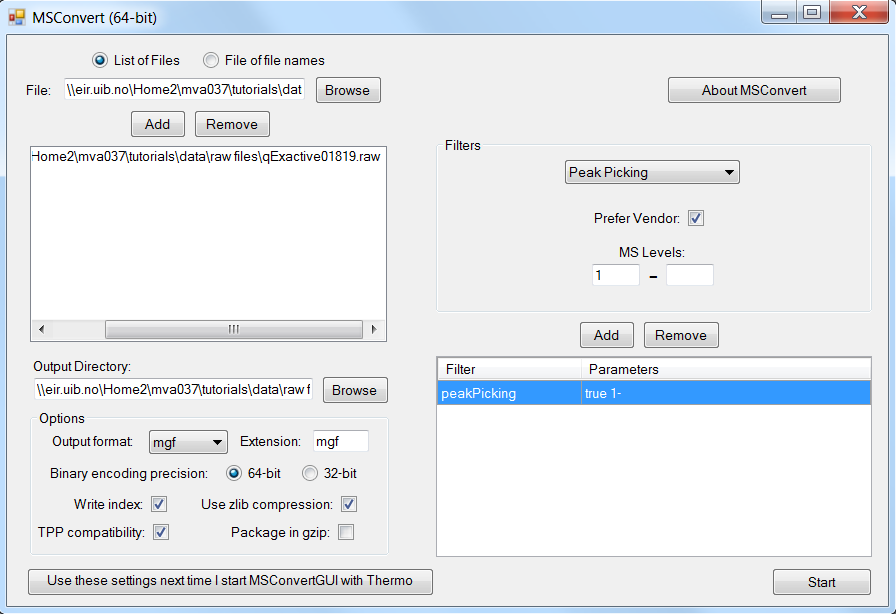
The data recorded by the Q Exactive is in so-called profile mode: spectra are a continuous line of data points. This can be seen when zooming on the spectra as done here in jmzML viewer[3](#_ENREF_3):

\\eir.uib.no\Home2\mva037\tutorials\1- identification\12- Peak list generation\illustrations\spectrum.tif

In order to reduce the amount of data to interpret, we will apply a peak-picker, a program which transforms the bell-shaped peaks into a single data point. As shown below in the TOPPView[4](#_ENREF_4" \o "Sturm, 2009 #1) spectrum viewer:

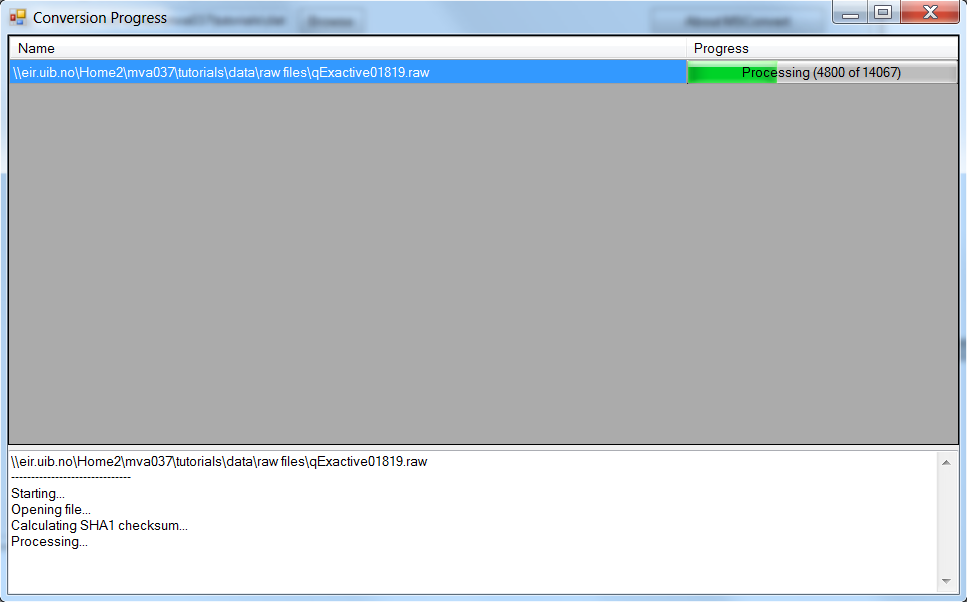


Select “Peak Picking” in the “Filters” panel and click on ‘Add’. You should see the following screen:



*How do you know whether or not you should apply a peak-picker? [1.2a]*

When clicking ‘Start’, the following screen will appear and the file qExactive01819.mgf will be generated in the location specified.



**Tip:**  
*Many files will be generated along the spectrum interpretation workflow.*

*A good organization of the files will save you a lot of time!*

Advanced: Signal Processing

Depending on the mass spectrometer, the MS/MS spectra used for identification can require more advanced 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**[5](#_ENREF_5" \o "Bertsch, 2011 #15)**. 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>.

Our example dataset was acquired on a high resolution mass spectrometer. 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[6](#_ENREF_6" \o "Savitzky, 1964 #311) 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,[7](#_ENREF_7) they are thus advised for quantitative studies.[8](#_ENREF_8) All these tools can be applied by the TOPPAS interface.

Two graphical interfaces allow you to look at your data (TOPPview, see below) and to draw pipelines (TOPPAS).



Note that these come with examples allowing you to familiarize with the software. Moreover, the OpenMS team provides free support.

References

1. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24**, 2534-2536 (2008).

2. Martens, L. et al. mzML--a community standard for mass spectrometry data. *Mol Cell Proteomics* **10**, R110 000133 (2011).

3. Cote, R.G., Reisinger, F. & Martens, L. jmzML, an open-source Java API for mzML, the PSI standard for MS data. *Proteomics* **10**, 1332-1335 (2010).

4. Sturm, M. & Kohlbacher, O. TOPPView: an open-source viewer for mass spectrometry data. *Journal of proteome research* **8**, 3760-3763 (2009).

5. Bertsch, A., Gropl, C., Reinert, K. & Kohlbacher, O. OpenMS and TOPP: open source software for LC-MS data analysis. *Methods Mol Biol* **696**, 353-367 (2011).

6. Savitzky, A. & Golay, M.J.E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry* **36**, 1627-1639 (1964).

7. Lange, E., Gropl, C., Reinert, K., Kohlbacher, O. & Hildebrandt, A. High-accuracy peak picking of proteomics data using wavelet techniques. *Pac Symp Biocomput*, 243-254 (2006).

8. Vaudel, M., Sickmann, A. & Martens, L. Peptide and protein quantification: a map of the minefield. *Proteomics* **10**, 650-670 (2010).