Mass Spectrum Analysis and Data Conversion Tool

(User Guide)

Pavel Cejnar

Department of Computing and Control Engineering, Institute of Chemical Technology,
Technická 3, 166 28 Prague 6, Czech Republic, email: cejnarp@vscht.cz

Mass Spectrum Analysis and Data Conversion Tool was created by Pavel Cejnar (pavel.cejnar@vscht.cz). This software is also based on Martin Strohalm's program mMass (http://www.mmass.org).

Installation on MS Windows

The program is distributed in ZIP archive containing the „ms" application folder. The program executable file is a command line utility *ms.exe* requiring exactly two parameters:

1) the file with configuration options
2) the file with mass spectrometry data to process

If you want to process more than one file, simply use the shell scripting.

Current version of *ms.exe* was successfully tested on Windows 7 64-bit (however the program is a 32-bit application).

Running from Source

The python source code has been compiled with Python 2.7 (http://www.python.org, 32-bit Python 2.7.3 for Windows). You will also need to install NumPy extension (http://www.numpy.org, 32-bit numpy-1.6.2 for python 2.7) and if you want to add some windows gadgets, you can use wxPython (http://www.wxpython.org, 32-bit unicode wxPython 2.8.12.1 for python 2.7 for example).

To compile C-source file calculations.c, use Microsoft Visual Studio Express 2010 for example. When you have the C-compiler installed, open the command line interpreter, move to "mspy" folder and simply run "python setup.py build". Find a "calculations.pyd" file and move it back to "mspy" folder.

If you want to make your own exectutable file for MS Windows you need to have py2exe (http://pypi.python.org/pypi/py2exe, 32-bit py2exe 0.6.9 for python 2.7). After installing the utility you can make the executable file for Windows simply by running the command "python

setup.py py2exe". If you want to distribute the compiled files, make sure you will bundle all the dependencies like msvcrXX.dll and you have the license to distribute them.

Input formats

The program supports several mass spectrometry formats, like XML-based formats *mzData* (http://www.psidev.info), *mzXML* (http://tools.proteomecenter.org) and *mzML* (http://www.psidev.info). The *mMass Spectrum Document (*.msd)* format is also supported. Since it is often impossible to obtain the manufacturer's description of their native file formats, they are not currently supported. However, if you have a *Bruker's CompassXport* tool installed on your computer and the command "CompassXport.exe" is available on the search path, it is automatically used to convert and open raw data from all *Bruker's* instruments. The tool is available for free. However, it is for MS Windows platform only. There are also some configuration options in the configuration file for the cooperation with this tool.

Data Processing

The application is configured by its configuration xml file, which is divided into several sections. The section **<batch>** controls which operation will be executed during the run. The operations are executed in the order croping, smoothing, peakpicking, deisotoping, exporting spectrum, exporting peak list. To execute the operation, set the parameter **value** to 1, to omit the operation, set the parameter **value** to 0.

```
<batch>

    <param name="crop" value="1" type="int" />

    <param name="smoothing" value="1" type="int" />

    <param name="peakpicking" value="1" type="int" />

    <param name="deisotoping" value="1" type="int" />

    <param name="exportSpectrum" value="1" type="int" />

    <param name="exportPeaks" value="1" type="int" />

 </batch>
```

All the parameters for the operations are stored in the <processing> section in an appropriate subsection.

Cropping

This function simply discards all the spectrum data which are out of the m/z range specified by **lowMass** and **highMass** parameters.

Smoothing

You can use this function to smooth the noise which distorts shape of the spectrum. There are three different smoothing methods available - Moving Average, Gaussian and Savitzky-Golay. In general, Moving Average and Gaussian are much faster but causes significant intensity loss for sharp peaks. These methods should be preferentially used to smooth high-mass spectra where peaks are broader. On the other hand, Savitzky-Golay filter

is very slow but intensity loss is much lower. This method should be preferentially used to smooth low-mass spectra where peaks are sharp.

To set the method, set the value “MA”, “GA”, or “SG” to the **method** parameter. Set the appropriate m/z interval size as a smoothing window to the **windowSize** parameter. Set the number of repetitions of the smoothing operation to the **cycles** parameter.

Peak Picking

If you want to automatically find peaks in the spectra, execute the peak picking operation. It is strongly recommended to apply cropping (and/or smoothing) operation before, otherwise you can easily get out of the memory.

To filter according to s/n threshold, first the baseline (zero-noise level) must be computed. The baseline is computed as a median intensity of the signal in selected signal window. Set the parameter **baselinePrecision** to specify the selected window. The higher the value of the **baselinePrecision** parameter, the shorter the window, from which the baseline will be computed, i.e. the baseline will shape according to the spectrum. Set the value 1 to compute the baseline from a widest possible window. Set the value 0 to compute the constant baseline from the whole spectrum. Set the parameter **baselineOffset** for relative correction of the computed baseline, i.e. to lower the baseline by the specified relative amount of the computed noise deviation. Then set the parameter **snThreshold**. All the peaks bellow the S/N threshold will not be reported. Set the parameter **pickingHeight** to find at which relative height of the peak should be computed the m/z center of the peak and thus its intensity at that point.

Deisotoping

The main purpose of this tool is to cluster the peaks to appropriate groups and if required to remove peak isotopes or peaks that don’t belong to any peak cluster. Starting from specified **maxCharge**, for every peak its isotopes are searched using corresponding isotopic mass difference (1.00287/abs(z)) ± **massTolerance**. If at least one isotope is found, the peak is set as parent peak (monoisotopic peak) with current charge state. If no isotope is found, current charge state is decreased (abs(z) - 1) and search cycle starts again for the same peak. Because of possible peak overlaps, theoretical isotopic pattern needs to be taken into account. While searching for isotopes, intensity of every found peak is also compared with its isotopic theoretical value. If the intensity is matching theoretical value ± (**intTolerance** * theoretical value), corresponding peak is set as the isotope peak for given parent peak and discarded from any subsequent search cycle. If the difference is over the tolerance, the corresponding peak will be used as a possible parent (monoisotopic) peak in a subsequent search cycle. For isotopes, the default isotope distance is used (1.00287). You can change this value by setting **isotopeShift** parameter to the value to add to default isotope distance. If you do not want to report the isotopes, but only the parental peaks, set the parameter **removeIsotopes** to value 1. If you do not want to report peaks that were not assigned to any peak cluster, set the parameter **removeUnknown** to 1.

Exporting Spectrum

This operation exports the processed spectrum (the m/z and the absolute processed intensity) to a text file. Any previous operations are applied according to configuration file settings and then the spectrum is exported. The output text file is stored to the same directory as the read spectrum.

Set the parameter **spectrumHeader** to value 1, if you want to add the first line to the exported file, containing the names of the columns. Set the **spectrumSeparator** parameter to the column delimiter character. Use HTML escape sequences for special characters. For the tabulator character use the value “tab”.

Exporting Peaklist

This operation exports the processed peak list to the text file. Any previous operations are applied according to configuration file settings and then the peaks are exported. The output text file is stored to the same directory as the read spectrum.

Set the parameter **peaklistHeader** to value 1, if you want to add the first line to the exported file, containing the names of the columns. Set the **peaklistSeparator** parameter to the column delimiter character. Use HTML escape sequences for special characters. For the tabulator character use the value “tab”. Set the **peaklistColumns** parameter to choose which columns to export. The possible columns are (in order of appearance):

1) **mz** – m/z of the peak
2) ai – absolute processed intensity of the peak
3) base – computed baseline intensity for given m/z
4) int – intensity of the peak (i.e. = ai – base)
5) rel – relative intensity of the peak in %. The base (100%) is the highest peak in the spectrum
6) sn – signal-to-noise ratio
7) z – peak charge
8) mass – peak mass parameter computed from its m/z and z value.
9) fwhm – full width at half maximum of the peak
10) pickheight_b – m/z start of the peak at peak picking height
11) pickheight_e – m/z end of the peak at peak picking height
12) resol – peak resolution (i.e. = (m/z) / fwhm)
13) deisotoped – whether peak is a part of any peak cluster. Possible values are: ‘None’ – not a part of any cluster, ‘False’ – peak is a parent peak of a peak cluster, ‘True’ – peak is some subsequent peak in a peak cluster
14) deisotoping_grp – number of cluster which given peak belongs to or ‘None’ if it is not a part of any peak cluster

Use the semicolon character (;) as a separator in the parameter string. The order of the columns doesn’t depend on the order in the parameter string. The order is always as listed above.

License