**Data and text mining**

**MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data**

Mikko Katajamaa¹, Jarkko Miettinen² and Matej Orešič²,*

¹Turku Centre for Biotechnology, Turku, Finland and ²VTT Technical Research Centre of Finland, Espoo, Finland

Received on November 25, 2005; revised on December 21, 2005; accepted on January 3, 2006

Advance Access publication January 10, 2006
Associate Editor: Jonathan Wren

**ABSTRACT**

Summary: New additional methods are presented for processing and visualizing mass spectrometry based molecular profile data, implemented as part of the recently introduced MZmine software. They include new features and extensions such as support for mzXML data format, capability to perform batch processing for large number of files, support for parallel processing, new methods for calculating peak areas using post-alignment peak picking algorithm and implementation of Sammon's mapping and curvilinear distance analysis for data visualization and exploratory data analysis.

Availability: MZmine is available under GNU Public license from http://mzmine.sourceforge.net/

Contact: matej.oresic@vtt.fi

**INTRODUCTION**

Mass spectrometry coupled to liquid or gas chromatography, or capillary electrophoresis (LC/MS, GC/MS or CE/MS, respectively) is increasingly utilized for differential profiling of biological samples. The applications of such an approach can be found in domains of systems biology, functional genomics and biomarker discovery. One of the ongoing challenges of such molecular profiling approaches is the development of better data processing methods.

We have recently introduced a suite of tools for the processing of mass spectrometry based profile data (Katajamaa and Orešić, 2005). MZmine implements solutions for several stages of data processing, including input file manipulation, spectral filtering, peak detection, chromatographic alignment, normalization, visualization and data export. MZmine (version 0.55) is a stand-alone Java application requiring Java Runtime Environment 5.0 or higher. It is therefore platform-independent, and successful installations have been reported on systems running Linux, Windows and Mac OS X, utilizing the software to process data from a variety of LC/MS and GC/MS instruments.

In this paper we report new developments of the software that include solutions for automated processing of large numbers of spectra, enhanced secondary peak picking method, as well as extension of software to post-processing by implementation of two methods for non-linear mapping of high-dimensional profile data into two-dimensional space.

**IMPORT AND PROCESSING OF FILES**

MZmine supports import of the NetCDF as well as mzXML (Pedrioli et al., 2004) raw data formats. New tools for manipulating the raw data files are available, including methods for noise reduction by filtering in chromatographic direction, cropping raw data range and removing scans by their width.

Stages of spectral data processing are sequential, and once parameter values for a specific type of platform are known, the process can be automated. MZmine enables the set up of data processing as a batch process, as well as an option to store the data processing parameters into the template files that can be loaded for future runs using the data from the same platform. In addition, the data processing can be set up to run on multiple processors, which is particularly useful for stages that are trivially parallelizable such as peak picking.

**ESTIMATION OF AREAS FOR MISSED PEAKS**

Following peak detection and subsequent alignment, many of the peaks have none or only few matches in other samples. There are various possible reasons for the misses: peak may not be present in the sample; peak detection may have failed because of noisy raw data or inaccurate parameter settings may have been used for peak detection and chromatographic alignment methods. The empty gaps caused by missing peaks are often troublesome to handle during subsequent steps in the data analysis and it is therefore worthwhile to return to raw data and check for the presence of corresponding peaks based on detected peaks in select samples.

We implemented a gap-filler method which estimates heights and areas for missed peaks. This method first searches for a local intensity maximum within a selected chromatographic region corresponding to expected location of a missed peak, which is used as an estimate for peak height. The peak area estimate is then calculated by moving from the maximum to both directions along the extracted ion chromatogram as long as the peak curve is monotonously decreasing within the pre-specified tolerance limits.

The gap-filler method increases the number of low-intensity peaks included in data analysis (Fig. 1), and advances our ability to utilize the differential profiling for quantitative measurements of metabolites. As a limitation, the current alignment and gap-filler methods cannot distinguish different molecular species if present at the same retention time and m/z value.

**Data visualization**

While a variety of excellent data analysis tools exist for software packages such as R (http://cran.r-project.org/) or Matlab (MathWorks, Inc), visualization capabilities enabling exploration
of high-dimensional profile data embedded into MZmine facilitate quality control and first-pass data exploration.

We incorporated two methods, curvilinear distance analysis (CDA) (Lee et al., 2000) and Sammon’s non-linear mapping (NLM) (Sammon Jr., 1969). They both try to preserve distances between points in original $N$-dimensional space and in lower dimensional projection space $P$ ($P$ being 2 in our case). Both use iterative process to find minimum of their respective error function. In the brief summary of the two methods, $d_{ij}^N$ will denote distance between points $i$ and $j$ in $N$-dimensional original space and $d_{ij}^P$ will denote distance between same points in $P$-dimensional projection space.

Fig. 1. Comparison of peak heights and areas for two different aligned samples from the analysis on UPLC-MS (QToF Premier from Waters, Inc.). Each dot is a peak with a specific $m/z$ value and retention time. Peaks found in primary peak picking are shown as black dots and those found after gap filling are white circles.

Fig. 2. Screenshot of MZmine, based on lipidomic profiling of two cell lines (five samples each). Chromatograms of two samples are shown, along with the CDA plot of all 10 samples.
Sammon’s non-linear mapping

Sammon’s NLM tries to minimize its error function $E$

$$E = \frac{1}{\sum_{i,j}^{N} d_{ij}} \sum_{i,j}^{N} \frac{(d_{ij}^{0} - d_{ij})^2}{d_{ij}^2},$$

by iterative steepest gradient descent. Its strengths include ease of implementation and use. On the other hand, generally it converges slowly and its error function is biased towards the small distances.

Curvilinear distance analysis

Unlike Sammon’s NLM, CDA uses stochastic gradient descent to minimize its error function $E$

$$E = \frac{1}{2} \sum_{i} \sum_{j \neq i} (d_{ij} - d_{ij}^{0})^2 F(d_{ij}, \lambda(k)),$$

where $F(d_{ij}^{0}, \lambda(k))$ denotes weight function and $\lambda(k)$ is the neighborhood radius. The initial parameters are the starting learning rate $\alpha_0$ and the starting neighborhood radius $\lambda_0$. CDA reduces its workload by quantizing points in N-space to centroids, followed by creating a graph in which every centroid connects to a select number of centroids. Distances from every centroid to every other centroid, called curvilinear distances and denoted with $d_{ij}$, are then calculated using Dijkstra’s shortest path algorithm. The distances are therefore calculated along the structures in N-dimensional space, not through them, therefore CDA provides a powerful distance metric for dimensionality reduction approaches.

Screenshot of MZmine with application of CDA included is shown in Figure 2.

CONCLUSIONS

The development of MZmine has been motivated by the need to create a software platform that enables easy incorporation of new algorithms and applications for data processing of mass spectrometry based molecular profile data.

Our current development areas are implementation of new normalization algorithms, extending the software to handle multiple spectra from the same sample (e.g. MS or MS²), and enabling database connectivity.

ACKNOWLEDGEMENTS

The authors thank Tuulikki Seppänen-Laakso and Tapani Suortti for performing most of the LC/MS analyses utilized during the MZmine development process. M.K. was funded by Academy of Finland SYSBIO Programme. M.O. was partially funded by EU Marie Curie International Reintegration Grant.

Conflict of Interest: none declared.

REFERENCES


