A bioinformatic package for automated targeted and global profiling analysis of large direct infusion mass spectrometry datasets

Gonçalo Correia[1], Olivier Cloarec[3], Elena Chekmeneva[1], Queenie Chan[2], Elaine Holmes[1]

[1]Division of Computational and Systems Medicine, Department of Surgery and Cancer, Imperial College London; Sir Alexander Fleming Building, South Kensington Campus, London, SW7 2AZ, UK

[2]Department of Epidemiology and Biostatistics, MRC-HPA Centre for Environment and Health, School of Public Health, Faculty of Medicine, Imperial College London, St. Mary's Campus, London, UK

[3]Korrigan Sciences Ltd, 38 Wakemans, Upper Basildon Reading, RG8 8JE, UK

Mass spectrometry techniques are very attractive for molecular epidemiology studies, as their sensitivity and ability to measure a wide range of compounds aids the unravelling of complex relationships between gene-environment interactions or exposure patterns, and the metabolic phenotypes and outcomes they induce. Hyphenated techniques, where liquid or gas chromatography separation is interfaced with mass spectrometry detection are commonly used for this purpose. However, for analysis of large cohorts, typical in epidemiological studies, their cost and time of analysis per sample quickly becomes a hindrance.

A practical alternative is to use direct infusion mass spectrometry (DIMS), where the samples are injected directly into the mass spectrometer without any pre-separation. This reduces enormously time and cost of analysis. However, these methods have the drawback of worsening the ion suppression phenomenon and removing one dimension for peak deconvolution and species identification. The former problem can be overcome by using nanoelectrospray (nESI) sources. A DI-nanoESI-High resolution MS method for high throughput analysis of human biofluid samples is being developed in our group[1]. It uses the TriVersa Advion NanoMate system in infusion mode for ionization, which greatly reduces ion suppression effects using minimal sample amount. Each specimen can thus be analysed in both positive and negative mode in only 3 minutes. Using this method, around 10,000 24h urine collection samples from the INTERMAP (INTERnational collaborative study of MAcronutrients, micronutrients and blood Pressure)[2] study where measured in only 3 months.

However when it comes to data analysis, to our knowledge no automated software tools for analysis of large DIMS datasets with all the necessary quality control measures and statistics

exists. Most computational work-flows for either targeted/quantification or untargeted analysis of mass spectrometry datasets were made specifically for GC/LC-MS data. To address this shortage we are developing an open-source Python package for automated pre-processing, quality control and analysis of DIMS data, particularly in the context of large scale human biofluid metabolic phenotyping projects. The package contains modules for pre-processing, targeted quantification and untargeted/profiling analysis. Thanks to its object-oriented development, the modules can be used on their own or included in third party application easily.

The initial pre-processing module contains algorithms for reading and parsing data from the standard .mzML format, perform scan averaging, m/z scale correction and interpolation of the averaged scan to obtain identical a standard m/z scale for the whole dataset. Baseline modelling, signal to noise ratio calculation and quality assessment measures for flagging technical problems are also available. All the statistical assumptions and pre-processing algorithms vary according to the MS detector used. To cope with the large size of these datasets, the obtained pre-processed spectra and their respective sample type (*e.g.*, calibration curve or quality control sample) and batch metadata annotations are saved on disk in a specific HDF5[3] binary file format layout. This layout can be read by the targeted and profiling analysis modules of the package.

The targeted analysis module was developed to quantify specific metabolites using the method of standard additions. It requires the input of a library of targeted compounds containing a correspondence between the chemical species to quantify and their labelled internal standard. For each sample, the peaks corresponding to the compounds of interest and their internal standard are identified based on their accurate mass, and integrated after fitting mixtures of gaussian peaks. The quantification results are obtained from an intra-batch calibration curve, after ensuring this calibration series passed the required statistical quality control measures. For untargeted analysis, feature detection, normalization and batch effect analysis and correction procedures appropriate for this type of data are under investigation and refinement and will be added to the package.

These tools are currently being applied in the analysis of DIMS data from the INTERMAP study, for analysis of DI-nanoESI-HR MS data acquired on HRMS QTOF Synapt G2-*Si* (Waters, Manchester, UK).

References:

1 - Elena Chekmeneva et al, Study of pregnancy outcomes: quantification of selected metabolites by highthroughput nano-electrospray HRMS-TOF method, presented at the MSACL 2014 EU

2 - Stamler, et al, INTERMAP: background, aims, design, methods, and descriptive statistics (nondietary), Journal of Human Hypertension (2003)

3 - The HDF Group. Hierarchical Data Format, version 5, 1997-2014. http://www.hdfgroup.org/HDF5/.