A software complement to AMDIS for processing GC-MS metabolomics data

Volker Behrends,[1] Gregory D. Tredwell,[1] Jacob G. Bundy*

Imperial College London, Department of Surgery and Cancer, Faculty of Medicine, Sir Alexander Fleming Building, London SW7 2AZ, UK

[1]These authors contributed equally.

*Corresponding author. Email: j.bundy@imperial.ac.uk

The software package AMDIS performs GC-MS peak deconvolution, but tends to produce false positives and leaves missing values where peaks are found in only a proportion of a set of chromatograms. We have developed a software complement to AMDIS that (a) allows rapid manual inspection of chromatographic peaks across all samples, to confirm data quality; and (b) for a given sample set, integrates peak areas across all samples, even where AMDIS deconvolution would leave missing values. The software is a freely available package that runs within the commercial Matlab environment, and is useful for metabolomics and other situations where GC-MS is used to profile many peaks from complex mixtures.

Keywords: GC-MS, AMDIS, deconvolution, metabolomics, software

GC-MS is widely used for metabolomics, owing to its excellent chromatographic behaviour for volatile compounds (i.e. most metabolites require derivatization) and the production of reproducible mass spectral fingerprints with electron impact ionization [1]. It is a mature and robust technology, and relatively low cost, especially for quadrupole instruments. It is particularly suited to a 'targetted metabolomics' approach of profiling a mixture and comparing the results to a compound library [2, 3]. Peak deconvolution is necessary to discriminate co-eluting compounds, by using *a priori* information from multiple ions. The free software package AMDIS is widely used for deconvolution of GC-MS data [4]. Lu et al. [5] concluded that its performance compared favourably to commercial software packages for GC-TOF-MS data, although it did tend to give a high false positive rate. However, it does have an additional disadvantage when used in a typical metabolomics study, when a small set of chromatograms (e.g. tens to hundreds) are compared: if a compound is detected in some but not all of the chromatograms, this leads to missing values when the peak table (matrix of samples against analytes) is generated. This is problematic because it affects the performance of the multivariate or biostatistical methods that are required to extract information from the complex metabolomic datasets. It is of course possible to replace the missing values with zeros, but this introduces additional problems, as the actual distribution of the variables (i.e. signal from analyte + signal from instrument noise) is no longer accurately described. Data transformations have been developed to deal specifically with the issue where analyte concentrations are on a similar level to instrument noise [6, 7], but are not applicable where zeros replace missing values.

We have developed a software package which takes the AMDIS output plus the raw data and gives a ‘cleaned’ peak table (Fig. 1). Specifically, it (a) includes a step for visual inspection of all peaks across all samples by the analyst, as this is still more powerful and flexible than fully-automated algorithms; and (b) reintegrates peaks from extracted ion chromatograms across an entire sample set, thus giving a back-filled sample table with no missing values, ready for further data analysis. Unlike standalone packages such as MET-IDEA [8], which also back-fills across all samples in a set, it is designed to be used only in combination with AMDIS.

We use AMDIS to process Agilent raw GC-MS data files (acquired using Chemstation), using the Fiehn spectral library [2] or NIST libraries, and typically find that AMDIS does not consistently identify targets (metabolites) across all related samples in a set. The output from AMDIS batch processing of multiple spectra consists of two text files for each sample (spectrumname.ELU and spectrumname.FIN). The .FIN file contains information about identified targets, such as the integrated peak area, retention time, name, and net score. When processing a single spectral file a simple text report file can be produced that summarizes the results in a metabolite list, but it is harder to combine the results from multiple spectra into a single data matrix. To address this problem, we developed a GC-MS Assignment Validator and Integrator (GAVIN) script for Matlab, which is available to download from http://www.box.net/shared/dj7f2zfmey. We use a Perl script (called from within GAVIN) to combine the results from multiple AMDIS files to produce a peak table (sample against metabolite matrix) of metabolites identified in at least one spectrum as a delimited text file. This initial peak table is then imported into the Matlab workspace, together with a spreadsheet containing sample ID and user-

defined metadata (e.g. experimental treatment, replicate number) to aid graphic representation of the results. This forms the basic data structure that is used as the starting point for visual inspection and back filling of the data matrix.

To perform the back-filling, the GAVIN package imports the raw GC-MS data in netCDF format (netCDF support is native in Matlab release 2008b and newer versions), and a manual validation procedure is carried out on a per compound basis. As a first step, the user is asked to define a threshold that only selects metabolites found in a certain percentage of samples. Based on this subset of compounds, GAVIN starts with the compound that has the lowest mean retention time and extracts a quantification ion (QI) and two validation ions (VI1 and VI2) from the raw data for each sample. The QI is taken from the user's AMDIS library; however if one is not defined than the model ion that AMDIS uses for its peak matching is used. The validation ions are defined as the next two most abundant ions in the compound mass spectra; however some ions that are commonly occurring, such as 73 and 74, as well as ions one or two mass units away from the QI, are excluded from being validation ions. These extracted ions are displayed as tiled subplots for each sample, with the three ion chromatograms coloured differently. There is no alignment of chromatograms. A screenshot of the user interface is given in Figure 2. If the package is called with the 'Overlay' option set to 'Yes', the data of the QIs and respective VIs are overlaid for all samples and displayed in three windows, one for each ion. This option should be selected if a large number of samples are to be analysed simultaneously. In each plot, the dashed lines represent the integration region for the QI, the solid black line represents the centre of that region. Yellow backgrounds (default mode) or ion chromatograms coloured yellow (overlay mode)

signal that the compound was not identified in that sample by AMDIS. As seen in Figure 2, the user has various options for each compound. If ‘Integrate’ is selected, the QI is integrated over the integration region and passed to the result matrix. The ‘Delete’ option deletes the peak for all samples, effectively erasing the compound from the results matrix. ‘Manipulate’ allows the user to change the parameters of the integration region. ‘Narrow Interval’ narrows the integration region in case of peak overlap, and ‘Widen Interval’ widens the integration region in case of broad peaks. ‘Force RT’ enables the user to manually set the centre the integration region. This is useful for QIs that are common to several compounds, for which multiple peaks might be found within the RT window, and so for which the automatic peak picking (the centre of the integration region is set to the maximum value in the displayed RT window) might pick the wrong peak. Additional options for improving consistent peak picking in normal mode (i.e. non-overlay) are the RT shift penalty options explained in the supplementary information.

Once the manual validation procedure is completed for each compound in the data matrix the package generates a samples-by-compounds array without missing values, and also three data structures containing a) the metadata and the back-filled integrals (‘curatedData’), b) the information retrieved from the AMDIS output files (‘keptCompounds’) and c) the raw data, including the GC dimension (‘rawData’).

Optional inputs: GAVIn can be called with several command line options, which are described in the supplementary documentation. Briefly, ‘Overlay’ toggles between two options to display and integrate the data. In non-overlay mode, the extracted ions

are displayed in a different plot for each sample and the RT is determined by finding the maximum value of the QI in the integration window. In contrast, the RT is defined by the maximum value of the QI across all samples in overlay mode. ‘SameScale’ toggles if the intensity scale (y-axis) of the extracted ion chromatograms is defined on a per-sample or per-batch basis. ‘rtPenalty’ can be used to facilitate peak picking by imposing penalties based on RT differences from the expected RT. Finally, after each compound, GAVIN deposits the data structure PrevCompVal into the Matlab workspace. This acts as a failsafe should the integration / validation procedure be terminated prematurely. To continue at the last validated compound, GAVIN can be called with PrevCompVal as an input argument.

In summary, we have developed a freely-available software tool for backfilling missing values obtained from AMDIS processed GCMS spectra, producing a data matrix more suitable for subsequent chemometric analysis. Users perform visual inspection of mass spectral information for all metabolites to give a greater confidence in assignments and can adjust the position and size of the integration window.

Acknowledgements

VB was funded by the BBSRC. GDT was funded by the Bioprocessing Research Industry Club (BRIC), a partnership between BBSRC, EPSRC, and a consortium of leading companies. We thank Manuel Liebeke for testing the software and his helpful suggestions.

References

[1] Fiehn, O., Extending the breadth of metabolite profiling by gas chromatography coupled to mass spectrometry, Trends Analyt Chem 27 (2008) 261-269.

[2] Kind, T., G. Wohlgemuth, Y. Lee do, Y. Lu, M. Palazoglu, S. Shahbaz, O. Fiehn, FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry, Anal Chem 81 (2009) 10038-10048.

[3] Smart, K.F., R.B. Aggio, J.R. Van Houtte, S.G. Villas-Boas, Analytical platform for metabolome analysis of microbial cells using methyl chloroformate derivatization followed by gas chromatography-mass spectrometry, Nat Protoc 5 (2010) 1709-1729.

[4] Stein, S.E., An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data, J Am Soc Mass Spectr 10 (1999) 770-781.

[5] Lu, H.M., W.B. Dunn, H.L. Shen, D.B. Kell, Y.Z. Liang, Comparative evaluation of software for deconvolution of metabolomics data based on GC-TOF-MS, Trac-trend Anal Chem 27(3) (2008) 215-227.

[6] Purohit, P.V., D.M. Rocke, M.R. Viant, D.L. Woodruff, Discrimination models using variance-stabilizing transformation of metabolomic NMR data, OMICS 8 (2004) 118-130.

[7] Rocke, D.M., B. Durbin, Approximate variance-stabilizing transformations for gene-expression microarray data, Bioinformatics 19 (2003) 966-972.

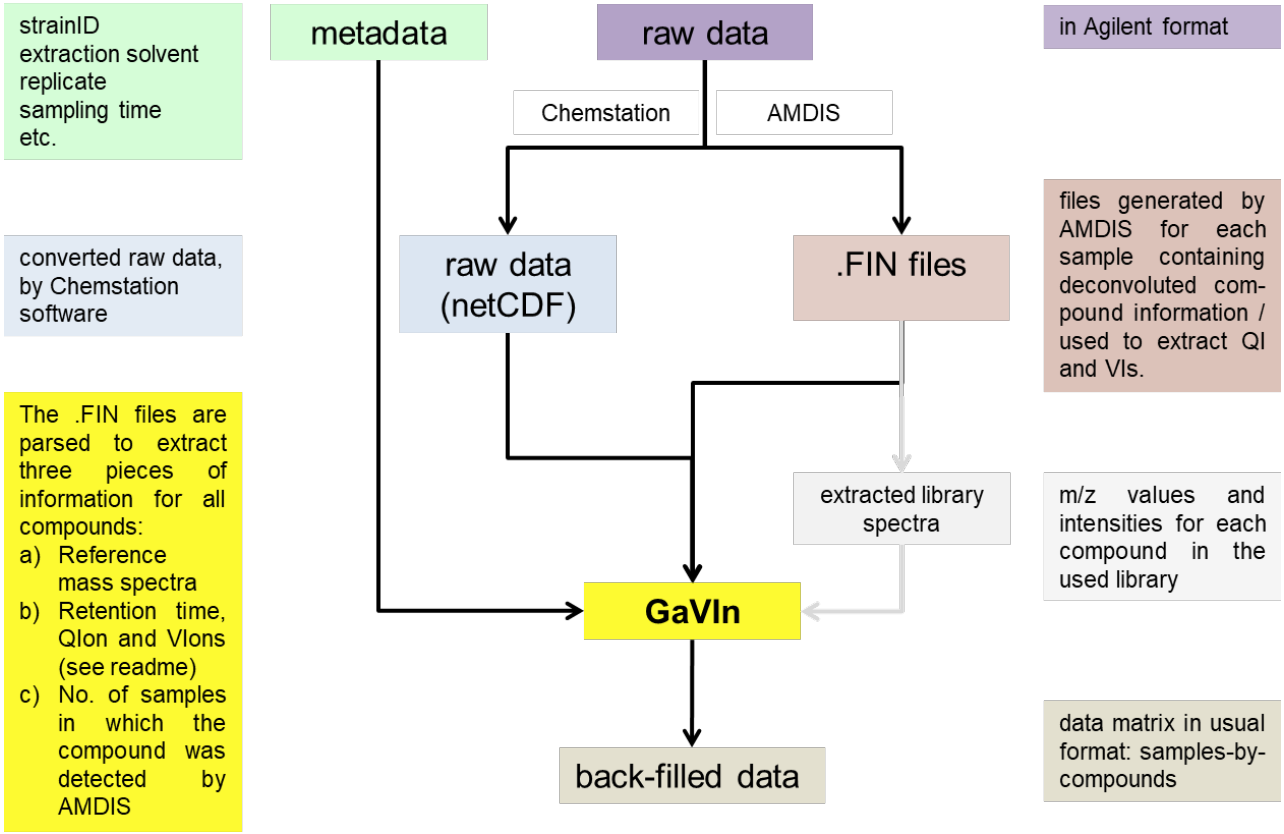
[8] Broeckling, C.D., I.R. Reddy, A.L. Duran, X. Zhao, L.W. Sumner, MET-IDEA: data extraction tool for mass spectrometry-based metabolomics, Anal Chem 78 (2006) 4334-4341.

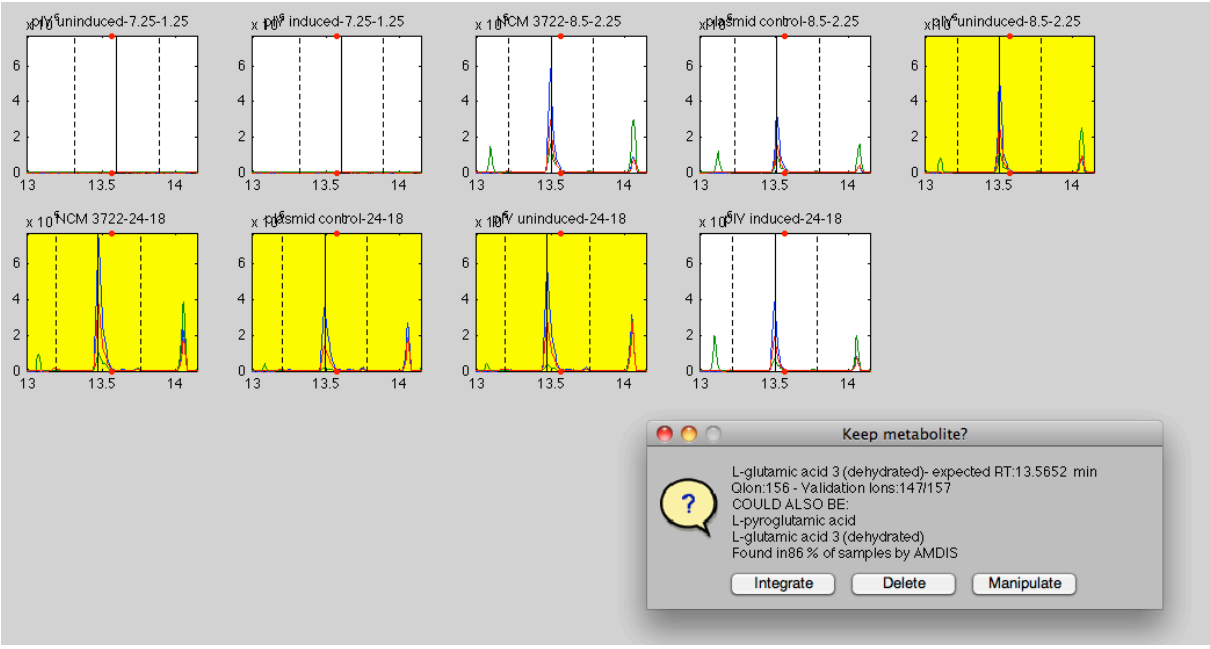
Legends to figures and online supplementary information

Figure 1. Schematic of operations performed by the software package.

Figure 2. Screenshot showing example of software in use in tiled mode. Yellow panels are samples where AMDIS has not annotated a compound as present.

Supplementary information: A zipped directory containing all Matlab files required to run GAVIN, plus a README file with detailed instructions for use.




pIV uninduced-7.25-1.25
pIV induced-7.25-1.25
NCM 3722-8.5-2.25
plasmid control-8.5-2.25
pIV uninduced-8.5-2.25
NCM 3722-24-18
plasmid control-24-18
pIV uninduced-24-18
pIV induced-24-18
Keep metabolite?
L-glutamic acid 3 (dehydrated)- expected RT:13.5652 min
Qion:156 - Validation Ions:147/157
COULD ALSO BE:
L-pyroglutamic acid
L-glutamic acid 3 (dehydrated)
Found in86 % of samples by AMDIS
Integrate
Delete
Manipulate