BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 7 2012, pages 1048–1049
doi:10.1093/bioinformatics/bts069

Systems biology

Advance Access publication February 4, 2012

IDEOM: an Excel interface for analysis of LC–MS-based metabolomics data

Darren J. Creek[1,2], Andris Jankevics[3,4], Karl E. V. Burgess[1], Rainer Breitling[3,4] and Michael P. Barrett[1,*]

[1]Institute of Infection, Immunity and Inflammation, Wellcome Trust Centre for Molecular Parasitology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8TA, UK, [2]Department of Biochemistry and Molecular Biology, University of Melbourne, Parkville, Victoria 3010, Australia, [3]Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, G12 8QQ, UK and [4]Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, 9700 CC Groningen, The Netherlands

Associate Editor: Jonathan Wren

ABSTRACT

Summary: The application of emerging metabolomics technologies to the comprehensive investigation of cellular biochemistry has been limited by bottlenecks in data processing, particularly noise filtering and metabolite identification. IDEOM provides a user-friendly data processing application that automates filtering and identification of metabolite peaks, paying particular attention to common sources of noise and false identifications generated by liquid chromatography–mass spectrometry (LC–MS) platforms. Building on advanced processing tools such as mzMatch and XCMS, it allows users to run a comprehensive pipeline for data analysis and visualization from a graphical user interface within Microsoft Excel, a familiar program for most biological scientists.

Availability and implementation: IDEOM is provided free of charge at http://mzmatch.sourceforge.net/ideom.html, as a macro-enabled spreadsheet (.xlsb). Implementation requires Microsoft Excel (2007 or later). R is also required for full functionality.

Contact: michael.barrett@glasgow.ac.uk

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on December 1, 2011; revised on January 9, 2012; accepted on January 31, 2012

1 INTRODUCTION

Metabolomics aims to measure all small molecules (metabolites) in a biological system. However, challenges associated with data analysis, in particular the accurate identification of metabolites (Moco *et al*., 2007; Neumann and Böcker, 2010) have restricted progress.

Recent advances in high resolution mass spectrometry (MS) provide accurate mass detection that significantly improves metabolite identification from liquid chromatography–MS (LC–MS) (Breitling *et al*., 2006; Moco *et al*., 2007) although limitations still abound (Kind and Fiehn, 2010; Neumann and Böcker, 2010), and noise or artifact peaks are often incorrectly identified as metabolites (Scheltema *et al*., 2009).

Applications are freely available for detection, quantification and alignment of signals in LC–MS data (Blekherman *et al*., 2011), and numerous multivariate statistical methods have been proposed to extract significant features from these high-dimensional datasets (Madsen *et al*., 2010). However, many of the most powerful tools are implemented in statistical software, and while the graphical user interface (GUI) in MZmine (Pluskal *et al*., 2010) simplifies data pre-processing, recent advances in noise filtering and identification algorithms (Brown *et al*., 2011; Scheltema *et al*., 2011) are difficult to use without special training.

IDEOM is a Microsoft Excel template with a collection of VBA macros that enable automated data processing of high resolution LC–MS data from untargeted metabolomics studies, with a particular focus on removal of noise [which accounts for ~80% of peaks in typical LC–MS metabolomics datasets (Jankevics *et al*., 2012)], metabolite identification and data visualization. Its GUI allows users to exploit the power of data-processing methods such as mzMatch (Scheltema *et al*., 2011) and XCMS (Smith *et al*., 2006) from within Excel, and enables rapid and simple conversion of raw or pre-processed LC–MS data into a filtered, interpretable list of putative metabolites, with associated confidence levels and sample intensities.

2 METHODS

Opening the IDEOM template in Excel provides a spreadsheet on which to specify the parameters for data processing (default values are provided). All IDEOM macros for data processing are then activated by buttons or in-cell hyperlinks, and pop-up dialog boxes are used to specify important parameters for individual processing steps.

Raw LC–MS data files (.mzXML) are processed by IDEOM using the freely available XCMS (Smith *et al*., 2006) and mzmatch.R tools (http://mzmatch.sourceforge.net/index.php) (Scheltema *et al*., 2011), by automatic generation and execution of scripts in the R environment (www.r-project.org) (R, 2008) using readily adjustable parameters. Raw peaks are extracted by XCMS, and mzMatch applies peak matching, noise filtering, gap-filling and annotation of related peaks. Pre-processed peak lists from mzMatch or MZmine can be directly imported into IDEOM as text or csv files. During data import, any number of samples can be assigned to (up to 30) study groups (e.g. blanks, controls and QCs) to improve data filtering, analysis and visualization. Sample consistency, according

*To whom correspondence should be addressed.

to average peak height and internal standards (optional), is automatically checked and normalization can be applied.

Noise filtering within IDEOM removes common sources of noise in high resolution LC–MS data: chromatographic peak shoulders, irreproducible peaks, background or contaminant signals, and fourier transform (FT) and electrospray ionisation (ESI) artifacts including isotopes, adducts and fragments (based on Brown *et al*., 2011 and Scheltema *et al*., 2011), as described in the documentation (Supplementary Material).

Metabolite identification is achieved by matching the accurate mass and retention time of observed peaks to metabolites in the included database, which incorporates all likely metabolites from a wide range of biological databases, and can be updated by users for specific applications. Retention times for authentic standards, and a retention time prediction model, are included for ZIC-HILIC chromatography data (Creek *et al*., 2011); however, as retention times are instrument specific, users are encouraged to upload standard retention times from their own platform with the macro provided. Subsequent to initial metabolite identification, a data-dependent polynomial mass recalibration step can be applied to correct for mass-dependent calibration errors. The final lists of identified, and rejected, peaks are annotated with confidence scores regarding the identification of each metabolite, based on retention times, organism-specific (or user-defined) databases and annotation as possible 'related peaks' (Scheltema *et al*., 2009). Univariate statistics (mean, relative intensity, SD, *t*-test and Fisher ratio) are calculated in Excel. Multivariate statistics are obtained by calls to the R environment.

3 RESULTS

The automated pre-processing steps in IDEOM drastically reduce the need for manual curation of LC–MS data by applying filters to remove hundreds of false-identifications (Creek *et al*., 2011). The example tutorial dataset (Supplementary Material) demonstrates putative identification of 1314 metabolites and rejection of 1942 false-identifications from serum samples. The putatively identified metabolites are displayed in an interactive table that includes metabolite information, sample intensities, comparative statistics, LC–MS data, and links to websites and graphs (Fig. 1). Putative identifications can be rapidly scrutinized by double-click access to LC–MS metadata, or adjusted by selecting alternative isomers from dropdown lists. Data visualization is enhanced by Excel's conditional formatting, filtering and sorting, allowing users to view results according to sample intensities, significance, pathways or other properties.

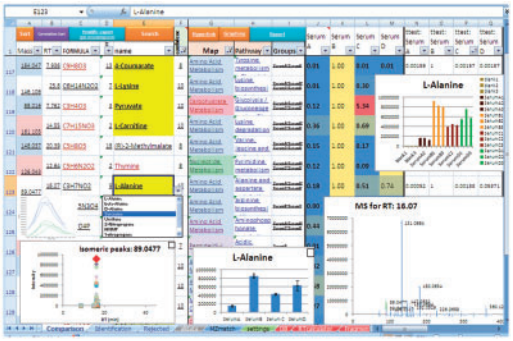

Fig. 1. Screenshot of results visualization showing the dropdown list selection of isomers and some of the automatically generated hyperlinked charts.

Additional IDEOM tools include: merging dual-polarity data, chemical formula determination for unidentified masses, stable isotope tracking and targeted analysis. Full documentation, tutorials, source code and the IDEOM template are freely available at http://mzmatch.sourceforge.net/ideom.html.

IDEOM provides a user-friendly interface for analysis of complex metabolomics datasets without the need for specialist bioinformatics skills, allowing for the rapid production of meaningful, interactive results for biological interpretation of untargeted metabolomics data sets.

ACKNOWLEDGEMENTS

The authors thank Gavin Blackburn, Isabel Vincent and Anubhav Srivastava for testing and feedback. Example datasets were provided by the Scottish Metabolomics Facility, funded by the Scottish Universities Life Sciences Alliance (SULSA).

Funding: Australian National Health and Medical Research Council postdoctoral training fellowship. Funding for AJ was provided by an Netherlands Organisation for Scientific Research NWO-VIDI grant to RB.

Conflict of Interest: none declared.

REFERENCES

Blekherman,G. *et al.* (2011) Bioinformatics tools for cancer metabolomics. *Metabolomics*, **7**, 329–343.

Breitling,R. *et al*. (2006) Precision mapping of the metabolome. *Trends Biotech.*, **24**, 543–548.

Brown,M. *et al.* (2011) Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets. *Bioinformatics*, **27**, 1108–1112.

Creek,D.J. *et al.* (2011) Towards global metabolomics analysis with Liquid Chromatography-Mass Spectrometry: improved metabolite identification by retention time prediction. *Anal. Chem.*, **83**, 8703–8710.

Jankevics,A. *et al*. (2012) Separating the wheat from the chaff: a prioritisation pipeline for the analysis of metabolomics datasets. *Metabolomics* [Epub ahead of print, doi:10.1007/s11306-011-0341-0, July 30, 2011].

Kind,T. and Fiehn,O. (2010) Advances in structure elucidation of small molecules using mass spectrometry. *Bioanal. Rev.*, **2**, 23–60.

Madsen,R. *et al*. (2010) Chemometrics in metabolomics—a review in human disease diagnosis. *Anal. Chim. Acta*, **659**, 23–33.

Moco,S. *et al*. (2007) Metabolomics technologies and metabolite identification. *Trends Analyt. Chem.*, **26**, 855–866.

Neumann,S. and Böcker,S. (2010) Computational mass spectrometry for metabolomics: identification of metabolites and small molecules. *Anal. Bioanal. Chem.*, **398**, 2779–2788.

Pluskal,T. *et al*. (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinf.*, **11**, 395.

R Development Core Team. (2008) R: a language and environment for statistical computing. In *R Foundation for Statistical Computing.* R Development Core Team, Vienna, Austria. http://www.R-project.org.

Scheltema,R. *et al*. (2009) Simple data-reduction method for high-resolution LCMS data in metabolomics. *Bioanalysis*, **1**, 1551–1557.

Scheltema,R.A. *et al.* (2011) PeakML/mzMatch: a File Format, Java Library, R Library, and Tool-Chain for mass spectrometry data analysis. *Anal. Chem.*, **83**, 2786–2793.

Smith,C. *et al*. (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.*, **78**, 779–787.