BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 5 2012, pages 736–738
doi:10.1093/bioinformatics/bts012

Systems biology

Advance Access publication January 11, 2012

MetExtract: a new software tool for the automated comprehensive extraction of metabolite-derived LC/MS signals in metabolomics research

Christoph Bueschl[1,2], Bernhard Kluger[1], Franz Berthiller[1], Gerald Lirk[2], Stephan Winkler[2], Rudolf Krska[1] and Rainer Schuhmacher[1,*]

[1]Center for Analytical Chemistry, Department for Agrobiotechnology (IFA-Tulln), University of Natural Resources and Life Sciences Vienna, Konrad Lorenz Strasse 20, 3430 Tulln and [2]School of Informatics, Communications and Media, University of Applied Sciences, Upper Austria, Softwarepark 11, 4232 Hagenberg, Austria

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Liquid chromatography–mass spectrometry (LC/MS) is a key technique in metabolomics. Since the efficient assignment of MS signals to true biological metabolites becomes feasible in combination with *in vivo* stable isotopic labelling, our aim was to provide a new software tool for this purpose.

Results: An algorithm and a program (MetExtract) have been developed to search for metabolites in *in vivo* labelled biological samples. The algorithm makes use of the chromatographic characteristics of the LC/MS data and detects MS peaks fulfilling the criteria of stable isotopic labelling. As a result of all calculations, the algorithm specifies a list of *m/z* values, the corresponding number of atoms of the labelling element (e.g. carbon) together with retention time and extracted adduct-, fragment- and polymer ions. Its function was evaluated using native ^{12}C- and uniformly ^{13}C-labelled standard substances.

Availability: MetExtract is available free of charge and warranty at http://code.google.com/p/metextract/. Precompiled executables are available for Windows operating systems.

Contact: rainer.schuhmacher@boku.ac.at

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

Received on September 4, 2011; revised on December 15, 2011; accepted on January 4, 2012

1 INTRODUCTION

The extraction of high-resolution LC/MS signals of true metabolites contained in complex biological samples is still a major challenge in non-targeted metabolomics. This limitation can largely be attributed to the non-specific nature of the electrospray ionisation (ESI) process in LC/MS, leading to full scan spectra, which tend to contain up to $>90\%$ background signals compared with signals measured for metabolites (Keller *et al.*, 2008). *In vivo* stable isotopic labelling offers a powerful tool to circumvent these limitations. Uniformly ^{13}C-labelled metabolites show identical chromatographic behaviour as their non-labelled isotopologues but can be easily differentiated by MS. Recently, (Giavalisco *et al.*, 2009) for example, successfully used ^{13}C-labelled *Arabidopsis thaliana* plants in combination with database searching for the elimination of background signals and annotation of metabolites. Despite the high potential of this approach in metabolomics research, no algorithms have been published for the automated evaluation of LC/MS data originating from *in vivo* labelled biological samples. To the best of our knowledge, a single commercial software programme exists to date, which has been designed to assign differentially expressed metabolites in differently treated biological samples (e.g. control versus treatment) by calculating and graphically illustrating the intensity ratios of the principal ions from isotopologue signal pairs (Jong and Beecher, 2011; http://nextgenmetabolomics.com/). Complementary to this, the presented MetExtract algorithm makes use of biological cultures obtained under identical growth conditions with the only difference being the isotopic composition of the nutrition source. MetExtract mainly aims at the global detection and combination of metabolite-derived corresponding LC/MS signals originating from natural and stable isotopically labelled analogues and their assignment to true biological metabolites.

1.1 Scope of the MetExtract algorithm

The presented algorithm was designed to extract as many metabolite-derived MS signals as possible from LC high-resolution full scan data with the aim to obtain a global picture of the metabolome of the respective organism. It requires that the studied organism is cultivated in parallel on two nutrition media, which only differ in the isotopic composition of a specified element *X* (e.g. carbon), either almost solely in form of stable 'isotope *A*' (e.g. ^{12}C in natural glucose) or 'isotope *B*' (e.g. ^{13}C in labelled $^{13}C_6$ glucose). The developed algorithm makes use of the following characteristics, typically observed in mass spectra of a 1+1 mixture of non-labelled as well as fully labelled biological samples: (i) the monoisotopic non-labelled and the completely labelled isotopologues should form the principal ions of their corresponding isotopic patterns and (ii) these two patterns have to be separated by at least one atomic mass unit. For example, if carbon is used as the labelling element and the first nutrition medium contains carbon at its natural isotopic composition (98.93% ^{12}C and 1.07% ^{13}C), the second medium should contain the same carbon source(s) as medium 1 but each of the C-sources with a high ^{13}C/^{12}C ratio. Assuming for example, a final ^{13}C/^{12}C ratio of 0.985 for the metabolites of the '^{13}C-culture sample', all metabolite ions containing between

[*]To whom correspondence should be addressed.

2 and 65 C-atoms fulfil the above-mentioned criteria and can be detected by MetExtract. Moreover, since the resulting ESI-LC/MS metabolite mass spectra contain two distinct isotopic patterns, the mass difference of the corresponding principal ions will be proportional to the number of atoms of the labelling element. MetExtract can also be used to investigate the metabolization of uniformly labelled tracer substrates as long as the MS signals of the resulting tracer derivatives fulfil the above-mentioned criteria. It shall be noted, however, that the algorithm has not been specifically designed to work with nutrition sources of partially labelled or mixtures of labelled and non-labelled precursors. Partly labelled metabolite pools, originating from the metabolization / incorporation of non- or partly labelled substrates can only be detected by the algorithm if the formed (and partially labelled) derivatives lead to mass spectra with an isotopic pattern containing a clearly assignable principal ion and corresponding isotopic peaks, which can be linked to the corresponding non-labelled isotopologue.

2 DESCRIPTION OF THE ALGORITHM FOR METABOLITE EXTRACTION

MetExtract uses a brute force approach for MS signal extraction and was implemented in C++ using the Qt SDK (v. 4.7.2, http://qt.nokia.com/products/). The programme is capable of processing high-resolution MS full scan data, independent of the MS instrument type or manufacturer. It comprises several distinct operation steps, which will be described in the following.

In a first step, full profile, full scan raw LC/MS data files have to be centroided and converted to the common data format ‘mzxml’ (Pedrioli *et al.*, 2004).

M/z value picking: detection of MS signals originating from isotopically labelled metabolites is done by iterating over all MS signal data in each mass spectrum. Since the different native (non-labelled) ^{12}C- and uniformly ^{13}C-labelled isotopologues are not separated by LC, each mass spectrum of a metabolite contains mass peaks of both isotopic forms. Each MS signal above a selected intensity threshold in the investigated mass spectra is initially considered to be the monoisotopic peak of a non-labelled metabolite ion containing only the lighter isotope of the labelling element *X*. Since the number of atoms of element *X* per metabolite ion cannot directly be determined from the selected *m/z* value, the algorithm uses a defined range of atom numbers to find the corresponding peak of the fully labelled isotopologue by testing for postulated theoretical masses of the fully labelled isotopologues. If a mass peak with a postulated *m/z* value has been found in the same mass spectrum, the presence of isotopic peaks is verified by the algorithm. Only if a pre-defined number of isotopic peaks are observed for both non- and fully labelled isotopologues and if their intensity ratios compared with the monoisotopic or fully labelled mass peaks are within a pre-defined intensity range window, the signal for which the search was started is considered a MS peak of a putative metabolite. The *m/z* value picking is finalized by calculating the Pearson’s Correlation Coefficient (corr) for the extracted ion current chromatograms of both the non-labelled and fully labelled isotopic molecules (Dalgaard, 2008). Only, if the signals of the isotopic ions are found and corr is at least the pre-set threshold value, the investigated mass peak is added to the list of picked *m/z* values. Otherwise, the assumption of the processed mass peak being caused by a true metabolite is discarded. Subsequently, the next mass peak in the mass spectrum is tested by the same procedure until all signals of all spectra have been evaluated. The result of this step is a list of *m/z* values representing the native isotopic form together with the corresponding numbers of atoms of element *X* contained in the respective metabolite ion.

Binning of m/z values: the next step is to group the picked MS signals to bins according to the similarity of *m/z* values and retention times. This is done by performing several hierarchical clustering algorithms (Hastie *et al.*, 2009, pp. 520–528) using Euclidean distance and average linking. Each hierarchical clustering is performed only with mass peaks with the same number of *X* atoms. This step successfully groups potential metabolite ions with the same *m/z* values into bins, but it does not consider structural isomers. As a consequence, putative metabolites having the same sum formula but different molecular structures are located in the same bin even if they were separated during the LC step of the analysis.

Separating of m/z value bins according to expected chromatographic peak width: since the extracted metabolites can be expected to elute into the MS detector in form of chromatographic peaks with typical peak widths in the order of seconds, the minimum and maximum number of scans contained in an *m/z* value bin can be selected with the program. If, for the selected *m/z* bin no signal was recorded for a pre-defined time interval, the algorithm splits the inspected bin in to two parts: one representing the detected mass peaks before the chromatographic gap and one with MS signals of this particular *m/z* value after the gap.

Combining of m/z value bins and database searching: subsequently, the algorithm searches for bins with *m/z* values of pre-defined adduct-, fragment- and polymer ions using an *m/z* value offset list. If two bins have a specified *m/z* offset value as well as the same number of *X* atoms and are eluted within a specified time window, the algorithm can combine the related adduct ions to a single group and calculate the accurate mass of the putative metabolite which can be used to perform an automated database search on the extracted metabolites using in-house or commercial databases such as e.g. Antibase (Laatsch, 2007).

Results of the algorithm: for a given LC/MS dataset observed for stable isotopic labelled metabolites, MetExtract offers a list of *m/z* values which were identified to represent true biological metabolites. The final output specifies binned *m/z* values, the corresponding number of atoms of the specified labelling element (e.g. carbon) together with retention time and found adduct-, fragment-and polymer ions. If database search function is used, database entries matching the selected assignment criteria will also be listed.

3 VERIFICATION OF THE ALGORITHM

The 15 native ^{12}C- and uniformly ^{13}C-labelled fungal standard substances were spiked to a non-labelled *Fusarium graminearum* culture sample. The mixture was measured by high-resolution LC/MS using positive ionization for verification. Under the described measurement conditions and parameter settings, 55 LC/MS signals from all 15 metabolites were extracted by the developed MetExtract algorithm (Table 1, Supplementary

Materials). For each standard substance, the correct number of C-atoms was assigned. No false positives were detected by the algorithm. Further, the algorithm detected two ions within our standards that showed characteristic pairs of isotopic patterns but could not be assigned to one of the spiked standard substances. We concluded that these ions originate from impurities in the used standard solutions, which partly originated directly from a production stock solution. For experimental details and MetExtract parameter settings, please refer to the Supplementary Materials.

ACKNOWLEDGEMENTS

Roman Labuda and Georg Haeubl, Romer Labs Diagnostics GmbH are greatly acknowledged for providing the labelled standards.

Funding: Austrian Science Fund (project SFB Fusarium 3706-B11); WWTF (project Toxi-Genome 9793008037); Federal Country Lower Austria in co-operation with the European Regional Development Fund (ERDF) of the European Union.

Conflict of Interest: none declared.

REFERENCES

Dalgaard,P. (2008) *Introductory Statistics with R*. Springer New York, Berlin, Heidelberg.

Giavalisco,P. *et al.* (2009) ^{13}C isotope-labeled metabolomes allowing for improved compound annotation and relative quantification in liquid chromatography-mass spectrometry-based metabolomic research. *Anal. Chem.*, **81**, 6546–6551.

Hastie,T. *et al.* (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, Berlin, Heidelberg.

Jong,F.D. and Beecher,C. (2011) Iroa makes Metabolic Profiling Easy. http://nextgenmetabolomics.com/ (29 November 2011, date last accessed).

Keller,B.O. *et al.* (2008) Interferences and contaminants encountered in modern mass spectrometry. *Anal. Chim. Acta.*, **627**, 71–81.

Laatsch,H. (2007) *AntiBase 2007: The Natural Product Identifier*. Wiley-VCH GmbH. Weinheim, Germany.

Pedrioli,P.G.A., *et al.* (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.*, **22**, 1459–1466.