Baitmet, a computational approach for GC–MS library-driven metabolite profiling

Xavier Domingo-Almenara[1,2,*], Jesus Brezmes[1,2], Gabriela Venturini[3], Gabriel Vivó-Truyols[4], Alexandre Perera[5], Maria Vinaixa[1,2,*]

[1]Metabolomics Platform, Department of Electronic Engineering (DEEEA), Universitat Rovira i Virgili, Tarragona, Catalonia, Spain

[2] Biomedical Research Centre in Diabetes and Associated Metabolic Disorders (CIBERDEM), Madrid, Spain

[3] Lab Genetics and Molecular Cardiology, Heart Institute (InCor), Universidade de São Paulo, Sao Paulo, Brazil

[4] Analytical Chemistry Group, Van't Hoff Institute for Molecular Sciences, Universiteit van Amsterdam, Amsterdam, The Netherlands

[5] B2SLab, Department of ESAII, CIBER of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Universitat Politècnica de Catalunya, Barcelona, Catalonia, Spain

**Corresponding authors:*

Xavier Domingo-Almenara
E-mail: xavier.domingo@urv.cat

Maria Vinaixa
E-mail: maria.vinaixa@urv.cat
Universitat Rovira i Virgili
Department of Electric, Electronic and Automatic Control Engineering (DEEEA).
Pasos Catalans s/n, Sescelades Campus, Tarragona 43007, Catalonia, Spain.
Tel. + 34 977256570

Baitmet, a computational approach for GC–MS library-driven metabolite profiling

Introduction: **Current computational tools for gas chromatography – mass spectrometry (GC–MS) metabolomics profiling do not focus on metabolite identification, that still remains as the entire workflow bottleneck and it relies on manual data reviewing. Metabolomics advent has fostered the development of public metabolite repositories containing mass spectra and retention indices, two orthogonal properties needed for metabolite identification. Such libraries can be used for library-driven compound profiling of large datasets produced in metabolomics, a complementary approach to current GC–MS non-targeted data analysis solutions that can eventually help to assess metabolite identities more efficiently.**

Results: **This paper introduces Baitmet, an integrated open-source computational tool written in R enclosing a complete workflow to perform high-throughput library-driven GC–MS profiling in complex samples. Baitmet capabilities were assayed in a metabolomics study involving 182 human serum samples where a set of 61 metabolites were profiled given a reference library.**

Conclusions: **Baitmet allows high-throughput and wide scope interrogation on the metabolic composition of complex samples analyzed using GC–MS via freely available spectral data. Baitmet is freely available at http://CRAN.R-project.org/package=baitmet.**

Introduction

Reproducibility of electron impact (EI) ionization together with robustness of capillary columns have qualified gas chromatography (GC) coupled to mass spectrometry (MS) as a long-standing analytical platform for metabolomics. Metabolomics has fostered both the expansion of publicly available mass spectral repositories (Hummel et al. 2010; Horai et al. 2010) and the development of metabolic databases containing spectral information (Wishart et al. 2013; Vinaixa et al. 2016). These contain tabulated EI mass spectra together with retention indices (RI), two orthogonal properties needed for metabolite identification and eventual metabolite profiling in GC–MS data (Sumner et al. 2007). In untargeted metabolomics profiling, where there is no previous knowledge of metabolites occurring in samples, pure spectra are usually extracted from GC–MS data using either univariate (Stein 1999) or multivariate (Domingo-Almenara et al. 2016) deconvolutions; and these spectra are posteriorly aligned across samples. Identification is subsequently performed by matching these pure spectra against EI spectral repositories. However, untargeted spectral deconvolution and alignment is a challenging process and therefore, metabolite identification is still relying on user input curation and manual data reviewing. On the other hand, spectral data tabulated in the above mentioned repositories can be used for a wide-scope screening of complex samples. Thus, a more guided approach consisting in profiling anticipated compounds (from which spectral information and RI are a priori known) might overcome some identification challenges.

As a complementary approach to non-targeted GC–MS data analysis solutions (Wehrens et al. 2014; Domingo-Almenara et al. 2016) here we introduce Baitmet, an integrated open-source R package allowing high-throughput metabolite relative quantification and identification through the projection of an entire mass spectral library into full-scan acquired GC–MS data. Baitmet uses MS and RI libraries as a bait, to profile metabolites (met). Baitmet can quantify compounds using either (i) selective mass ions for each compound or (ii) multivariate methods which implies that no prior information about the selective masses is required. The latter confers advantages over current library-driven compound profiling solutions revolving around the concept of spectral mass tags (MSTs) (Luedemann et al. 2008) or where selective ions for extraction peak apex intensities should be specified (Cuadros-Inostroza et al. 2009).

Methods

Baitmet operational modules and computational workflow are summarized in Figure 1. Baitmet requires as input data i) GC–MS files in commonly accepted chromatography interchange open standard formats (either netCDF or mzXML); ii) an EI spectral library containing retention indices (RI) and iii) RT/RI reference curve for the chromatographic method obtained from series of either internal or external RI reference standards such as n-alkanes (ALK) or fatty acid methyl esters (FAME) (Fig. 1A). First, a preprocessing step is applied with baseline drift removal and a de-noising using moving-minimum and Savitzky-Golay filters respectively. After this preprocessing Baitmet workflow iterates for each entry in the library as follows: an Expected Elution Window (EEW) is determined by extrapolating the corresponding tabulated compound RI into the initial RT/RI reference curve (Fig. 1B, left panel). All empirical spectra recorded in scans within this EEW are correlated against corresponding compound tabulated EI spectrum. RT maximizing this correlation is retained as the center of a region of interest (ROI) with boundaries four times the minimum compound full width at half maximum ($FWHM_{MIN}$), a user-defined value (in seconds) (Fig~~.~~ 1B, bottom panel). Next, the resolved compound chromatographic profile is reconstructed from this ROI using least absolute deviation (LAD) regression, a special case of least squares approach that balances all ions weight in the regression model. Likewise, orthogonal signal deconvolution (OSD) (Domingo-Almenara et al. 2015) is used to extract the corresponding compound pure empirical spectrum which is compared to the tabulated one by computing matching factor using either the Stein and Scott's composite similarity (Stein and Scott 1994) or the dot product (Fig. 1B, right panel). Finally, Baitmet computes empirical RI values. Of mention, Baitmet can compute empirical RI values by adapting the RT/RI initial curve using either co-injected standards or naturally occurring – and user-defined – compounds in samples as internal RI reference. Additionally, in absence of co-injected standards, Baitmet includes an automatic RT/RI curve correction to handle possible variations of the input RT/RI curve in each particular sample caused by small instrumental fluctuations (Supplementary Fig. S1). Refer to Supplementary Methods for an extended Baitmet computational workflow and built-in functionalities descriptions and a more detailed explanation of Baitmet output format.

Results and Discussion

Baitmet capabilities for library-driven compound profiling were evaluated using a GC–MS dataset consisting of analytical triplicates of serum samples from 56 age and body weight-matched subjects diagnosed with chronic kidney disease and the respective quality controls (182 GC–MS injections in total, see Supplementary Material for extended details on sample preparation and GC–MS analysis). The complete processing of the entire dataset was performed in 2 hours and 15 minutes (45 seconds per sample in a 2.9 GHz Intel Core i5 computer). The library used was a subset of the the Golm Metabolome Database (GMD, Version at 2011-11-21) (Hummel et al. 2007) including only those entries containing KEGG reference (a total of 1152) and excluding FAME. Initial RT/RI curve was determined using mean RT across samples for a series of internal FAME (C10 – 25) spiked to each sample. The RI error was set to 0.5% according to thresholds proposed by Strehmel et al. (Strehmel et al. 2008) and the $\mathrm{FWHM}_{\mathrm{MIN}}$ was set to 1 second. Baitmet detected 127 metabolites (RI error<1%) with 61 of them accounting for spectral matching factors above 85%, appearing in at least 80% of samples and known to be present in serum according to HMDB. The identity of these metabolites was reported with a level 2 according to The Metabolomics Standards Initiative (MSI) guidelines (Sumner et al. 2007). The Baitmet typical output for these 61 metabolites is summarized in Table S1. Spectral matching factors (MF) are indicated together with coefficient of determination obtained from regressing areas computed from Baitmet reconstructed chromatographic profiles against areas from the extracted ion chromatograms of particularly selected quantitative ions. Additionally, for each compound in Table S1 a comparison–between relative RI error deviations (RIe) is shown for empirical RI computed either using RT/RI internal calibration curve (internal standards co-injection) or automatic Baitmet RI/RT curve correction. In the first case, Baitmet input library included a set of MS and RI for each FAME spiked in the samples and these FAME were automatically detected by Baitmet. In the second case, the FAME information was removed from the library, and Baitmet computed RI making use of a set of pre-defined naturally occurring metabolites instead. The absolute mean difference between internal and automated Baitmet calibration was 0.02%, which is significantly less than the commonly accepted identification RI error (0.5 – 1%) (Strehmel et al. 2008) (Supplementary Fig. S3). Thus, internal RI standards addition to each individual sample can be avoided by adapting external RI calibration to each sample instead. This prevents sample chromatograms being cluttered with unnecessary peaks that can otherwise mask potential compound peaks. Moreover, the majority of metabolites listed in Table S1 showed coefficients of determination close to 0.90 and matching factors above 85%. This demonstrates Baitmet capability to reconstruct chromatographic profiles and extract pure mass spectra in real samples enabling library-driven metabolite profiling in GC– MS metabolomics.

Conclusions

Here we introduced Baitmet, an integrated modular and open-source R package for high throughput GC–MS library-driven metabolite profiling. It allows dumping publicly available EI-based spectral repositories such as MassBank or GMD into GC–MS experimental data. Baitmet is implemented as an easy-to-use workflow with a high runtime performance allowing high throughput processing of large datasets typically measured in metabolomics. Additionally, Baitmet provides an easy-to-interpret output that simplifies user-guided metabolite identification and assignment review, a common bottleneck in other GC–MS analysis pipelines. A Baitmet distinctive feature from other library-driven approaches is that it offers the possibility to profile metabolites without prior selective masses input. Additionally, it includes a novel strategy to automatically correct an external RT/RI calibration curve for each particular sample, allowing accurate computation of RI values without internal calibration (standards co-injection). Altogether, we present an integrated open-source tool that allows a high-throughput and wide scope interrogation on the metabolic composition of complex samples analyzed using GC–MS via freely available spectral data. Baitmet is freely available at http://CRAN.R- project.org/package=baitmet.

Funding: This research was partially funded by MINECO grant TEC2014–60337-R and TEC2015–69076-P; and FAPESP grant 2012/12042-7. CIBER–BBN and CIBER–DEM are initiatives of the Spanish ISCIII.

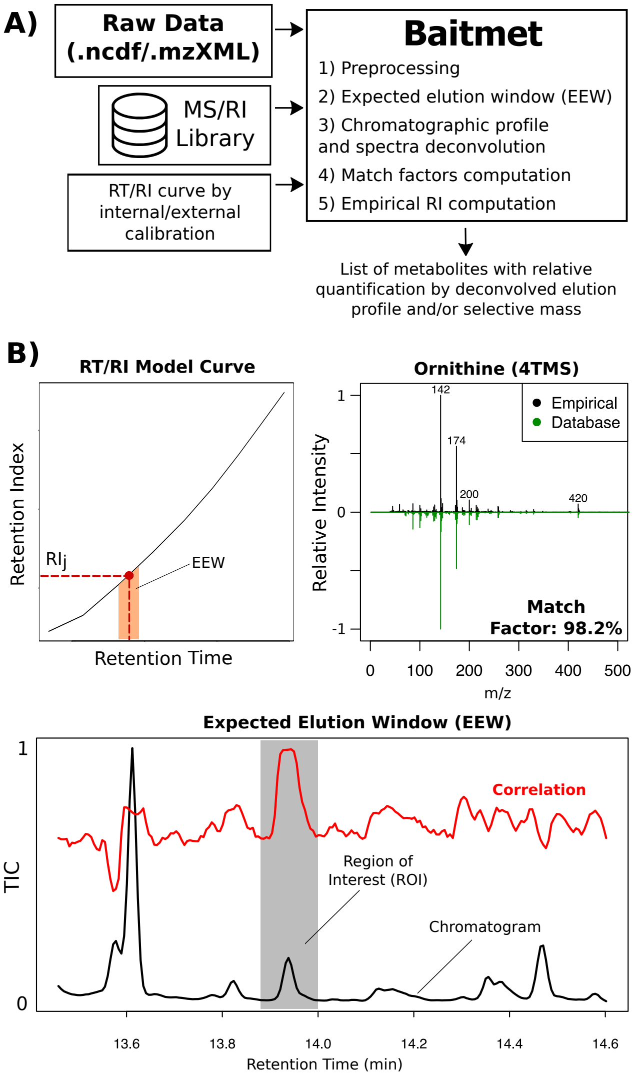

Figure 1 (A) Baitmet workflow. (B) For each compound in the library, an expected elution profile window (EEW) is determined by projecting corresponding tabulated RIj into the initial RT/RI model (left panel). MS spectrum for each tabulated compound is correlated against all spectra acquired within the scan range falling in this EEW and a ROI (region of interest) is defined with center position at RT maximizing this correlation (bottom panel). For each compound, pure empirical spectrum (black) is extracted which is further compared to the tabulated reference spectrum in the MS library (green) (right panel).

Compliance with ethical standards

Ethical approval: The ethics committee of the Hospital das Clinicas, University of São Paulo (Brazil) approved the study (protocol number 3759/12/015).

Informed consent: Informed written consent was obtained from all participants in the study.

Conflict of Interest: The authors declare no conflict of interest.

References

Cuadros-Inostroza, A., Caldana, C., Redestig, H., Kusano, M., Lisec, J., Peña-Cortés, H., *et al* (2009). TargetSearch - a Bioconductor package for the efficient preprocessing of GC–MS metabolite profiling data. *BMC Bioinformatics*, 10, 428.

Domingo-Almenara, X., Perera, A., Ramirez, N., Canellas, N., Correig, X., Brezmes, J. (2015). Compound identification in gas chromatography/mass spectrometry-based metabolomics by blind source separation. *Journal of Chromatography A*, 1409, 226–233.

Domingo-Almenara, X., Brezmes, J., Vinaixa, M., Samino, S., Ramirez, N., Ramon-Krauel, M. *et al* (2016). eRah: a computational tool integrating spectral deconvolution and alignment with quantification and identification of metabolites in GC–MS-based metabolomics. *Analytical Chemistry*, 88(19), 9821–9829.

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K. *et al* (2010). MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45(7), 703–714.

Hummel, J., Selbig, J., Walther, D., Kopka, J. (2007). The Golm Metabolome Database: a database for GC–MS based metabolite profiling. *Topics in Current Genetics*, 18, 75–95.

Hummel, J., Strehmel, N., Selbig, J., Walther, D., Kopka, J. (2010). Decision tree supported substructure prediction of metabolites from GC–MS profiles. *Metabolomics*, 6(2), 322–333.

Luedemann, A., Strassburg, K., Erban, A., Kopka, J. (2008). TagFinder for the quantitative analysis of gas chromatography–mass spectrometry (GC–MS)-based metabolite profiling experiments. *Bioinformatics*, 24(5), 732–737.

Stein, S. E. and Scott, D. R. (1994). Optimization and testing of mass spectral library search algorithms for compound identification. Journal of the American Society for Mass Spectrometry, 5(9), 859–866.

Stein, S. E. (1999). An integrated method for spectrum extraction and com- pound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry*, 10(8), 770–781.

Strehmel, N., Hummel, J., Erban, A., Strassburg, K., Kopka, J. (2008). Retention index thresholds for compound matching in GC–MS metabolite profiling. *Journal of Chromatography B*, 871(2), 182–190.

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C.

A. *et al* (2007). Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics*, 3(3), 211–221.

Vinaixa, M., Schymanski, E. L., Neumann, S., Navarro, M., Salek, R. M., Yanes, O. (2016). Mass spectral databases for LC/MS and GC/MS-based metabolomics: state of the field and future prospects. *Trends in Analytical Chemistry*, 78, 23–35.

Wehrens, R. Weingart, G., Mattivi F. (2014). metaMS: an open-source pipeline for GC–MS-based untargeted metabolomics. *Journal of Chromatography B*, 966, 109–116.

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y. *et al* (2013). HMDB 3.0-The Human Metabolome Database in 2013 *Nucleic Acids Research*, 41, 801–807.