BIOINFORMATICS APPLICATIONS NOTE

Vol. 30 no. 9 2014, pages 1336–1337
doi:10.1093/bioinformatics/btu019

Systems biology

Advance Access publication January 17, 2014

ProbMetab: an *R* package for Bayesian probabilistic annotation of LC–MS-based metabolomics

Ricardo R. Silva[1], Fabien Jourdan[2,3], Diego M. Salvanha[1,4], Fabien Letisse[2,3,5], Emilien L. Jamin[2,3], Simone Guidetti-Gonzalez[6], Carlos A. Labate[6,7] and Ricardo Z. N. Vêncio[1,*]

[1]LabPIB, Department of Computing and Mathematics FFCLRP-USP, University of Sao Paulo, Ribeirao Preto, Brazil, [2]INRA UMR1331, Toxalim, Research Centre in Food Toxicology, [3]Universit de Toulouse, INSA, UPS, INP; LISBP, Toulouse, France, [4]Institute for Systems Biology, Seattle, Washington, USA, [5]CNRS, UMR5504, Toulouse, France, [6]Department of Genetics ESALQ-USP, University of Sao Paulo, Piracicaba, Brazil and [7]Laboratorio Nacional de Ciencia e Tecnologia do Bioetanol CTBE, Campinas, Brazil

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: We present ProbMetab, an R package that promotes substantial improvement in automatic probabilistic liquid chromatography–mass spectrometry-based metabolome annotation. The inference engine core is based on a Bayesian model implemented to (i) allow diverse source of experimental data and metadata to be systematically incorporated into the model with alternative ways to calculate the likelihood function and (ii) allow sensitive selection of biologically meaningful biochemical reaction databases as Dirichlet-categorical prior distribution. Additionally, to ensure result interpretation by system biologists, we display the annotation in a network where observed mass peaks are connected if their candidate metabolites are substrate/product of known biochemical reactions. This graph can be overlaid with other graph-based analysis, such as partial correlation networks, in a visualization scheme exported to Cytoscape, with web and stand-alone versions.

Availability and implementation: ProbMetab was implemented in a modular manner to fit together with established upstream (xcms, CAMERA, AStream, mzMatch.R, etc) and downstream R package tools (GeneNet, RCytoscape, DiffCorr, etc). ProbMetab, along with extensive documentation and case studies, is freely available under GNU license at: http://labpib.fmrp.usp.br/methods/probmetab/.

Contact: rvencio@usp.br

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

Received on November 16, 2013; revised on January 6, 2014; accepted on January 8, 2014

1 INTRODUCTION

Metabolomics is an emerging field of study in post-genomics, which aims at comprehensive analysis of small organic molecules in biological systems. Techniques of mass spectrometry coupled to liquid chromatography [liquid chromatography–mass spectrometry (LC–MS)] stand out as dominant methods in metabolomic experiments.

Although computational strategies have been used to filter and annotate mass peaks in LC–MS experiments (Dunn *et al*., 2012), these methods do not include the addition of external information into a mathematical model in a principled way. Recently, Rogers *et al*. (2009) put forward a proof-of-concept in which information incorporated to a probabilistic model provides better annotation (Breitling *et al*., 2013). Their Bayesian model, by means of appropriate prior distribution selection, introduces the elegant idea of using a set of known chemical reactions among candidate compounds to improve annotation, as certain combinations, detected together, would make more biochemical sense than others.

The state-of-the-art in probabilistic annotation established by Rogers *et al*. (2009) did not include an integrative computational implementation, a practical connection to public biological databases such as KEGG or MetaCyc (Altman *et al*., 2013) or a network-based output visualization schema. Therefore, our contribution is to fulfill these specific needs allowing easy access to this powerful statistical model for all metabolomic bioinformatics community.

2 RESULTS AND CONCLUSION

The platform chosen for implementation of these ideas was the well-known and established *R* programming environment, which incorporates a wide range of analyses including successful tools that perform preprocessing of spectral data required for metabolite annotation (Supplementary Fig. S1) (Kuhl *et al*., 2011; Smith *et al*., 2006).

Following Rogers *et al*. (2011) brief suggestion on how their previous method could be extended to incorporate additional experimental information and metadata, we implemented modifications to the likelihood term. Expanding the likelihood function L in multiplicative independent terms allows one to account for additional orthogonal (independent) information sources: $L = L_N \cdot L_{rt} \cdot L_{iso}$, where subindexes N, rt and iso stand for measurement noise model, retention time error model and isotope profile error model, respectively. For a complete model's description, we refer the interested reader to the Supplementary Material.

[*]To whom correspondence should be addressed.

The main product of a probabilistic annotation is a list of compound candidates ranked by their probabilities (Supplementary Fig. S2). To easily navigate over ProbMetabs results, we display tabular and dynamic network outputs along with supporting information, which assists practitioners to ultimately decide on most parsimonious annotations instead of forcing them to simplistically rely on the top probability assignment. All mass peaks are viewed as graph's nodes. Edges between two nodes are drawn if any candidate compound assigned to the outgoing node can be metabolized to any candidate compound assigned to the incoming node by means of a known biochemical reaction (Supplementary Fig. S3). ProbMetab is capable of producing reaction graphs and export them as standard Cytoscape input files or broadcasting the necessary graph data and attributes (colour, shapes, etc) directly to Cytoscape Desktop using RCytoscape (Shannon *et al*., 2013). This information can be easily overlaid with other widely used systems biology strategies such as correlation or partial correlation networks. If a mass spectra time-series or biological replicates are available, ProbMetab uses third-party packages integrated downstream to export correlation or partial correlation graphs, along with their intersection/difference with the reaction graph.

Alternatively, a biologist can visualize ProbMetab's results in a simplified searchable web interface. Our package has a function that is responsible to consume an online web-service, which checks and renders the broadcast results as a web page. The visualization approach was developed taking advantage of the cytoscape.js library (Lopes *et al*., 2010) and its dependencies and can be easily integrated or embedded into any html5 web application.

ProbMetab's documentation brings two detailed case studies in which all its features are explored. Moreover, to highlight integration with downstream and upstream third-party *R* packages, data analysis examples mentioned are carried out from raw data, following through preprocessing until it reaches ProbMetab's specific point of action. We used publicly available data from *Trypanosoma brucei*, causative agent of sleeping sickness, and an original dataset from *Saccharum officinarum* (sugarcane), an important biofuel source, to illustrate several points in typical metabolomics analysis sections.

The *T. brucei* dataset, obtained from the mzMatch.R project website, was chosen because it presents a set of metabolites identified with the aid of internal control standard compounds, being specially suited for performance evaluation. With this validation dataset, we compare the MetSamp (http://www.dcs.gla.ac.uk/inference/metsamp/) implementation from Rogers *et al*. (2009) with ProbMetab's implementation and show that, the efficient *R*/c++ integrated function (Eddelbuettel and François, 2011) had a 3-fold running time improvement over the MATLAB implementation. For both implementations, the higher probability candidate was the true identity in up to 60% of the metabolites. However, instead of reporting only the higher probability candidate identity as proposed by Rogers *et al*. (2009), we show that exporting the complete ranking in summarized visualizations, up to 90% of metabolite identities are among the top three higher probabilities. The full or filtered ranking allows the experimenter to associate the candidates with additional information present in this output and attribute the correct identity.

The sugarcane dataset was chosen to exemplify differential expression of annotated metabolites in contrasting environmental perturbation. We successfully recovered changes in a known stress response pathway (flavone and flavonol biosynthesis), showing the importance of a network-centric visualization for metabolite annotation to track metabolism changes. The benchmark dataset confirms, as preconceived by Rogers *et al*. (2009), that a probabilistic model using orthogonal data and metadata yields better automatic mass peak annotation. The perturbation dataset shows that probabilistic annotation can produce otherwise impossible interpretation for differential network connectivity.

We implemented a method to annotate compounds in a computational framework that allows the introduction of prior knowledge and additional spectral information. With the *R* package ProbMetab, we provide ways to summarize the results of series of analysis needed to extract information from complex high-dimensional MS data, and help the experimenter to track metabolism changes in the process of interest.

ACKNOWLEDGEMENTS

The authors thank Bruno M. Della Vecchia (USP), Dr Oscar Yanes's team (Rovira i Virgili University) and Paul Shannon (Fred Hutchinson Cancer Research Center) for critical discussions. They thank EBI/EMBO's courses for catalyzing cooperation among coauthors.

Funding: Microsoft Research and FAPESP (09/53161-6); FAPESP Temático grant (08/56100-5) and fellowships (10/14926-4, 12/05392-1 and 10/15417-6); CAPES through Pós-graduação em Genética FMRP-USP.

Conflict of Interest: none declared.

REFERENCES

Altman,T. *et al*. (2013) A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics*, **14**, 112.

Breitling,R. *et al*. (2013) Modeling challenges in the synthetic biology of secondary metabolism. *ACS Synth. Biol*., **2**, 373–378.

Dunn,W.B. *et al*. (2012) Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*, **9**, S44–S66.

Eddelbuettel,D. and François,R. (2011) Rcpp: seamless R and c++ integration. *J. Stat. Softw*., **40**, 1–18.

Kuhl,C. *et al*. (2011) CAMERA: an integrated strategy for compound spectra extraction and annotation of LC/MS data sets. *Anal. Chem*., **84**, 283–289.

Lopes,C.T. *et al*. (2010) Cytoscape web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.

Rogers,S. *et al*. (2009) Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics*, **25**, 512–518.

Rogers,S. *et al*. (2011) Bayesian approaches for mass spectrometry based metabolomics. In: Stumpf,M.P.H., Balding,D.J. and Girolami,M. (eds) *Handbook of Statistical Systems Biology*, John Wiley & Sons Ltd, Chichester, UK, pp. 467–476.

Shannon,P.T. *et al*. (2013) RCytoscape: tools for exploratory network analysis. *BMC Bioinformatics*, **14**, 217.

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