Bioinformatics, 31(9), 2015, 1493–1495
doi: 10.1093/bioinformatics/btu813
Advance Access Publication Date: 19 December 2014
Applications Note

OXFORD

Gene expression

Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics

Franck Giacomoni[1,†], Gildas Le Corguillé[2,†], Misharl Monsoor[2], Marion Landi[1], Pierre Pericard[2], Mélanie Pétéra[1], Christophe Duperier[1], Marie Tremblay-Franco[3], Jean-François Martin[3], Daniel Jacob[4], Sophie Goulitquer[2], Etienne A. Thévenot[5,*] and Christophe Caron[2,*]

[1]INRA, UMR 1019, PFEM, 63122 Saint Genes Champanelle, [2]CNRS, UPMC, FR2424, ABiMS, Station Biologique, 29680 Roscoff, [3]INRA, UMR 1331, PF MetaToul-AXIOM, Toxalim, F-31027 Toulouse, [4]INRA, Metabolome Facility of Bordeaux Functional Genomics Center, IBVM, 33140 Villenave d'Ornon and [5]CEA, LIST, Laboratory for Data Analysis and Smart Systems (LADIS), MetaboHUB Paris, F-91191 Gif-sur-Yvette, France

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Ziv Bar-Joseph

Received on August 22, 2014; revised on November 2, 2014; accepted on December 4, 2014

Abstract

Summary: The complex, rapidly evolving field of computational metabolomics calls for collaborative infrastructures where the large volume of new algorithms for data pre-processing, statistical analysis and annotation can be readily integrated whatever the language, evaluated on reference datasets and chained to build ad hoc workflows for users. We have developed Workflow4Metabolomics (W4M), the first fully open-source and collaborative online platform for computational metabolomics. W4M is a virtual research environment built upon the Galaxy web-based platform technology. It enables ergonomic integration, exchange and running of individual modules and workflows. Alternatively, the whole W4M framework and computational tools can be downloaded as a virtual machine for local installation.

Availability and implementation: http://workflow4metabolomics.org homepage enables users to open a private account and access the infrastructure.

W4M is developed and maintained by the French Bioinformatics Institute (IFB) and the French Metabolomics and Fluxomics Infrastructure (MetaboHUB).

Contact: contact@workflow4metabolomics.org

1 Introduction

Metabolomics, the high throughput analysis of small molecules in biological samples, heavily depends on data pre-processing, statistical analysis and chemical and biological annotation, which are complex, transdisciplinary processes involving both computation and interpretation (Holmes *et al.*, 2008). Since analytical technologies and protocols evolve rapidly, computational metabolomics is a field of intensive methodological research, resulting in a large volume of proposed algorithms written in various languages, making their evaluation by the bioinformatics community (including reviewers) and their chaining within ad hoc workflows by experimenters difficult (Smith *et al.*, 2013).

A few user-oriented online platforms for metabolomics data pre-processing and analysis have been described recently, including MeltDB (Neuweger *et al.*, 2008), MetaboAnalyst (Xia *et al.*, 2009) and XCMS Online (Tautenhahn *et al.*, 2012). There is, however, an

unmet need for an open source and open development infrastructure which would enable developers to readily integrate new modules, compare their performances on reference datasets or download and modify the existing ones for their own research.

We have thus developed a collaborative online research resource, Workflow4Metabolomics (W4M), for comprehensive metabolomics data pre-processing, statistical analysis and interpretation. W4M is a fully open-source virtual research environment (VRE; Carusi and Reimer, 2010) built upon the Galaxy environment (Goecks *et al.*, 2010) for bioinformatics developers and metabolomics users, with minimal wrapping burden of algorithms into modules, in addition to user-friendly functionalities for workflow management. Moreover, W4M includes unique computational modules for data normalization (signal drift and batch-effect correction), multivariate analysis (orthogonal partial least-squares) and annotation (via multiple databases query).

2 Features

2.1 Framework

The VRE integrates several digital resources over many layers (hardware, software, user interfaces, documentation, tools and workflows), and is based on a High Performance Computing environment (600 cores, 100 TB). The light-weight runner technology has been added to enhance interoperability and integration of components from heterogeneous environments (Linux, Windows, etc.). Multiple workflows can be run in parallel and users can rapidly analyse large datasets: for example, the full pre-processing, statistical analysis and annotation of a cohort dataset (184 raw files, 11 Go) from liquid chromatography coupled to high-resolution mass spectrometry (LC-HRMS) was performed in 5 h, using 1% of the computational resources. Modules include wrappers from existing open-source code, and innovative tools developed with standard languages (e.g. R, Perl, Python or Java). Each tool has a web page including working examples. In addition, a help desk is provided for both users and developers. An 'app-store' based on the Galaxy native toolshed, ToolShed4Metabolomics, fosters the local deployment of modules, facilitates the management of developer contributions, provides wrapper templates and promotes best practice guidelines. Finally, a full W4M portable virtual machine is distributed for local installation (e.g. for development prototyping).

2.2 Computational tools, workflows and services

W4M currently contains 19 modules covering all steps of LC-HRMS data analysis:

- Format conversion: raw data can be converted from commercial formats (e.g. Thermo Fisher.RAW) to open formats (including mzXML, Pedrioli *et al.*, 2004, and mzML, Deutsch, 2008), via a recently developed toolshed wrapper implementation of the ProteoWizard software (Kessner *et al.*, 2008).
- Pre-processing: all wrappers of the reference XCMS (Smith *et al.*, 2006) and CAMERA (Kuhl *et al.*, 2012) functions are available to perform peak extraction, retention time alignment and annotation of isotopes and adducts.
- Normalization: signal drift and batch-effects, which are two major source of bias in MS data (van der Kloet *et al.*, 2009), can be corrected by fitting linear or local polynomial regression models to quality control samples.
- Statistical analysis: in addition to parametric and non-parametric univariate tests, W4M offers unique functionalities for multivariate modelling, including orthogonal partial least-squares (Trygg and Wold, 2002), with all numerical and graphical results and diagnostics (optimal number of components estimated by cross-validation, variable importance in projection, model significance by permutation testing, outlier detection).
- Annotation: a formula generator based on the HiRes (High Resolution) algorithm (Kind and Fiehn, 2007) is provided, in addition to several modules for public database query which allow the user to define specific annotation strategies (e.g. by searching from general to more specialized resources).

Metabolomics scientists can access W4M with a simple web browser, upload their data, select analysis parameters or choose the default settings, and run their workflows in batch mode. In addition, W4M provides functionalities for creating interactive web-based documents showing the results of the analyses, and sharing them with collaborators directly on W4M. To get started easily, pre-configured workflows and corresponding histories are publicly shared for pre-processing, statistical analysis and annotation, respectively. A real LC-HRMS dataset (Roux *et al.*, 2012) is provided as a reference for new module and workflow evaluation.

3 Conclusion

The W4M infrastructure enables both experimental users with no specific programming skills and advanced developers to perform cutting-edge and reproducible computational analyses from raw data to metabolite annotation. W4M can be further extended to integrate external workflows running on desktop platforms (e.g. Taverna, KNIME), or acquisition instruments. The statistical modules from W4M can be used to analyse other 'omics' data, or can be combined with existing Galaxy workflows (e.g. in transcriptomics), thus enabling multi-omics analyses in a global systems-biology approach. In the coming months, modules for NMR data pre-processing will be integrated into W4M and the infrastructure will be connected to MetExplore (Cottret *et al.*, 2010) for genome-scale network analysis. W4M is therefore an innovative open-source computational VRE bridging the data-intensive bioinformatics and metabolomics communities.

Funding

This work was supported by Biogenouest®, Lifegrid (Auvergne), and by the IDEALG project [ANR-10-BTBR-04], IFB [ANR-11-INBS-0013] and MetaboHUB [ANR-11-INBS-0010] grants.

Conflict of Interest: none declared.

References

Carusi,A. and Reimer,T. (2010) *Virtual Research Environment Collaborative Landscape Study*. JISC, Bristol, *106 pp*.

Cottret,L. *et al.* (2010) MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. *Nucleic Acids Res.*, **38**, W132–W137.

Deutsch,E. (2008) mzML: a single, unifying data format for mass spectrometer output. *Proteomics*, **8**, 2776–2777.

Goecks,J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

Holmes,E. *et al.* (2008) Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature*, **453**, 396–400.

Kessner,D. *et al.* (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics,* **24**, 2534–2536.

Kind,T. and Fiehn,O. (2007) Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, **8**, 105.

Kuhl,C. *et al.* (2012) CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.*, **84**, 283–289.

Neuweger,H. *et al.* (2008) MeltDB: a software platform for the analysis and integration of metabolomics experiment data. *Bioinformatics,* **24**, 2726–2732.

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.

Roux,A. *et al.* (2012) Annotation of the human adult urinary metabolome and metabolite identification using ultra high performance liquid chromatography coupled to a linear quadrupole ion trap-orbitrap mass spectrometer. *Anal. Chem.,* **84**, 6429–6437.

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.

Smith,R. *et al.* (2013) Novel algorithms and the benefits of comparative validation. *Bioinformatics*, **29**, 1583–1585.

Tautenhahn,R. *et al.* (2012) XCMS online: a web-based platform to process untargeted metabolomic data. *Anal. Chem.,* **84**, 5035–5039.

Trygg,J. and Wold,S. (2002) Orthogonal projection to latent structures (O-PLS). *J. Chemometr.,* **16**, 119–128.

van der Kloet,F.M. *et al.* (2009) Analytical error reduction using single point calibration for accurate and precise metabolomic phenotyping. *J. Proteome Res.,* **8**, 5132–5141.

Xia,J. *et al.* (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res.,* **37**, W652–W660.