Bioinformatics, 32(13), 2016, 2035–2037
doi: 10.1093/bioinformatics/btw091
Advance Access Publication Date: 18 February 2016
Applications Note

Sequence analysis

RIPPER: a framework for MS1 only metabolomics and proteomics label-free relative quantification

Susan K. Van Riper[1,2,*], LeeAnn Higgins[3], John V. Carlis[4] and Timothy J. Griffin[3]

[1]Department of Biomedical Informatics and Computational Biology, University of Minnesota, Rochester, [2]University of Minnesota Informatics Institute, University of Minnesota, St Paul, [3]Department of Biochemistry, Molecular Biology, and Biophysics and [4]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on July 22, 2015; revised on February 12, 2016; accepted on February 15, 2016

Abstract

Summary: RIPPER is a framework for mass-spectrometry-based label-free relative quantification for proteomics and metabolomics studies. RIPPER combines a series of previously described algorithms for pre-processing, analyte quantification, retention time alignment, and analyte grouping across runs. It is also the first software framework to implement proximity-based intensity normalization. RIPPER produces lists of analyte signals with their unnormalized and normalized intensities that can serve as input to statistical and directed mass spectrometry (MS) methods for detecting quantitative differences between biological samples using MS.
Availability and implementation: http://www.z.umn.edu/ripper.
Contact: vanr0014@umn.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Biological studies using liquid chromatography (LC) and high-resolution mass spectrometry (MS) have the potential to aid the investigation of molecular mechanisms by discovering biological markers. Two important 'omics research areas utilizing LC-MS are metabolomics and proteomics; both face similar challenges in producing repeatable and reproducible data (Kall and Vitek, 2011). To address these challenges, we recently introduced proximity-based intensity normalization (PIN) (Van Riper *et al.*, 2014). To implement PIN, we created a new quantitative framework, RIPPER, for label-free quantification (LFQ). RIPPER has since become a valuable in-house tool for new MS-based quantitative method development. Because RIPPER does not rely on analyte identification before quantification, users can process MS1-only data for proteomic MS-based studies and generate an inclusion list for directed MS-based identification. By simply adjusting properties through a graphical user interface, users can also use RIPPER to extract analyte information for metabolomic studies, statistically analyze the results, and conduct further fragmentation experiments on metabolites of interest.

Here, we describe RIPPER, a Java-based framework that takes in mzXML files, extracts analyte features from MS1 spectral data and reports validated analyte signal intensities. Because it allows user-specified options and is written in Java, RIPPER accommodates a wide array of users on various platforms (OSX, Window and Linux). Finally, we compare RIPPER to other freely available software, namely MaxQuant (Cox and Mann, 2008), XCMS (Tautenhahn, *et al.*, 2012) and MZMine (Pluskal *et al.*, 2010). MaxQuant couples LFQ to identified peptides and proteins and is therefore not suitable to process MS1-only data. Furthermore, MaxQuant cannot process metabolomic data. XCMS is an online service for metabolomic data and is not well suited for proteomic data. MZMine can process both proteomic and metabolomic data

and does not rely on identification prior to quantification. We therefore chose MZMine as a benchmark for RIPPER. We benchmarked RIPPER versus MZMine 2.18.1 using high-resolution metabolomic data and found that RIPPER identified six times as many analytes six times faster (see Supplementary Materials).

2 Algorithms

Using a simple graphical user interface, a user specifies input mzXML files (Pedrioli *et al.*, 2004), output destination, experimental name, and advanced options. RIPPER then employs a series of algorithms for pre-processing, extracting analyte signals, normalization, retention time alignment and grouping analyte signals (Fig. 1). It outputs a comma-delimited file containing analyte signals, and for each, its un-normalized and, optionally, its normalized intensities. This output serves as input to external statistical methods, e.g. Student's *t*-test, to find statistically significant quantitative analyte differences between biological samples.

2.1 Pre-processing

RIPPER extracts data from mzXML files using JRAP, Institute for Systems Biology (www.proteomecenter.org/software.php). It then uses a series of previously-described algorithms (Horn *et al.*, 2000) for baseline correction using local medians, signal-to-noise thresholding and isotopic peak envelope detection. For each peak envelope, RIPPER constructs a deisotoped peak as a monoisotopic peak *m/z*—summed peak intensity pair. The result is a list of scans objects, each containing a list of deisotoped, real peaks.

2.2 Extract analyte signals

In LC-MS, analytes elute from the LC column over time based on their hydrophobicity or other physiochemical property. When measured via MS, each analyte generates a characteristic set of isotopically related peaks that forms an extracted ion chromatogram (XIC). A typical analyte XIC contains peak intensities that appear, maximize and disappear over time (Bellew *et al.*, 2006). Furthermore, an analyte's abundance correlates well with the area under the curve of its measured XIC (Bondarenko *et al.*, 2002). RIPPER first extracts candidate XIC's by clustering deisotoped peaks along the *m/z* and time dimensions with user supplied tolerances. Next, RIPPER selects valid XIC's from the candidate XIC set that meet user specified criteria. Finally, we construct an analyte signal peak (*m/z*—intensity pair) for each valid XIC. The analyte signal's *m/z* is the XIC's apex *m/z*, and the intensity is the area under the curve using the trapezoidal rule for approximating the definite integral of XIC peak intensities. RIPPER produces a list of valid XIC deisotoped peaks and a vector of constructed analyte signals with their aggregated raw and optionally PIN normalized intensities.

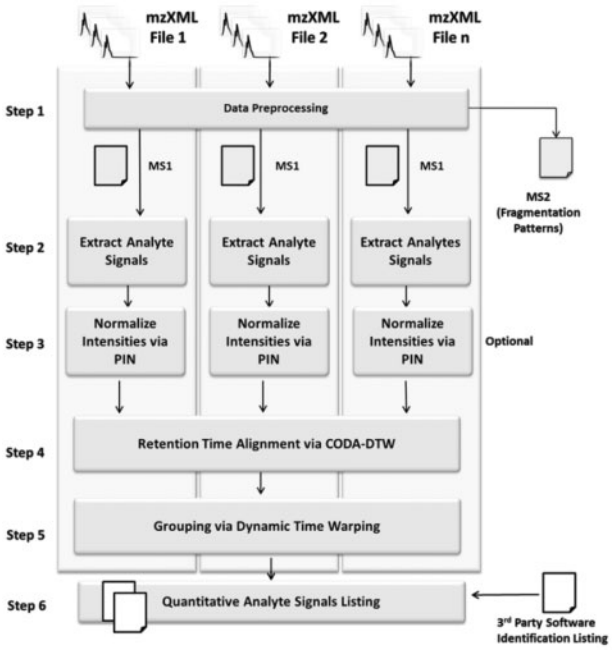

Fig. 1. RIPPER's processing steps

2.3 Retention time alignment

Retention time drift due to changes in experimental or physical conditions, (column performance, matrix effects, humidity and so forth), hinders accurate matching of analyte signals across LC-MS runs. To overcome this impediment, RIPPER uses a variant of the dynamic time warping (DTW)-component detection algorithm (Christin *et al.*, 2010). This variant saves memory consumption by using either integer and binary distance matrices rather than double or floating point matrices to align extracted analyte signal's retention times.

2.4 Group analyte signals across runs

RIPPER uses an analyte signal matching algorithm adapted from the FeatureCluster module available in msInspect (Bellew *et al.*, 2006). Our analyte signal grouper takes in multiple analyte signal vectors and selects the vector with the largest number of elements to be the reference model. It then performs a recursive pseudo-hierarchical clustering and DTW to generate a consensus analyte signal map. The map contains analyte signal objects, each with an *m/z* value, intensity, and a retention time.

2.5 Identify analytes

For peptidomic and proteomic applications, RIPPER does not identify peptide sequences associated with *m/z* values directly. However, researchers can develop a simple R script to match RIPPER extracted analyte signals to peptide and protein or metabolite identifications from external software applications. For metabolomic studies, we recommend further fragmentation experiments and identification via spectral library search.

3 Conclusion

RIPPER is the first framework implementing PIN and enabling the proportionality paradigm for LC-MS compositional LFQ workflows. Also, RIPPER is extensible and, therefore, easily maintained and updated. Finally, we have used RIPPER for numerous LC-MS 'omics experiments, most notably, differential proteomics (directed MS) and metabolomics (tracer analysis) with success. In sum, RIPPER's introduction is an important step in advancing LFQ workflows for investigating molecular machinery and biomarker discovery.

Acknowledgements

We thank the Center for Mass Spectrometry and Proteomics at the University of Minnesota and the Minnesota Supercomputing Institute for instrumental access and infrastructure support.

Funding

This research was funded in part by National Institutes of Health [1R01DE017734], National Science Foundation [1147079] and the Doctoral Dissertation Fellowship from the University of Minnesota Graduate School.

Conflict of Interest: none declared.

References

Bellew,M. *et al*. (2006) A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics*, **22**, 1902–1909.

Bondarenko,P.V. *et al*. (2002) Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Anal. Chem*., **74**, 4741–4749.

Christin,C. *et al*. (2010) Time alignment algorithms based on selected mass traces for complex LC-MS data. *J. Proteome Res*., **9**, 1483–1495.

Cox,J. and Mann,M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol*., **26**, 1367–1372.

Horn,D.M. *et al*. (2000) Automated reduction and interpretation of high-resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom*., **11**, 320–332.

Kall,L. and Vitek,O. (2011) Computational mass spectrometry-based proteomics. *Plos Comput. Biol*., **7**, e1002277.

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.

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

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

Van Riper,S.K. *et al*. (2014) Improved intensity-based label-free quantification via proximity-based intensity normalization (PIN). *J. Proteome Res*., **13**, 1281–1292.