Bioinformatics, 31(14), 2015, 2415–2417
doi: 10.1093/bioinformatics/btv152
Advance Access Publication Date: 18 March 2015
Applications Note

OXFORD

Data and text mining

Efficient visualization of high-throughput targeted proteomics experiments: TAPIR

Hannes L. Röst[1,2], George Rosenberger[1,2], Ruedi Aebersold[1] and Lars Malmström[1,*]

[1]ETH Zurich, Institute of Molecular Systems Biology, CH-8093 Zurich, Switzerland and [2]Ph.D. Program in Systems Biology, University of Zurich and ETH Zurich, CH-8057 Zurich, Switzerland

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on December 10, 2014; revised on February 24, 2015; accepted on March 11, 2015

Abstract

Motivation: Targeted mass spectrometry comprises a set of powerful methods to obtain accurate and consistent protein quantification in complex samples. To fully exploit these techniques, a cross-platform and open-source software stack based on standardized data exchange formats is required.

Results: We present TAPIR, a fast and efficient Python visualization software for chromatograms and peaks identified in targeted proteomics experiments. The input formats are open, community-driven standardized data formats (mzML for raw data storage and TraML encoding the hierarchical relationships between transitions, peptides and proteins). TAPIR is scalable to proteome-wide targeted proteomics studies (as enabled by SWATH-MS), allowing researchers to visualize high-throughput datasets. The framework integrates well with existing automated analysis pipelines and can be extended beyond targeted proteomics to other types of analyses.

Availability and implementation: TAPIR is available for all computing platforms under the 3-clause BSD license at https://github.com/msproteomicstools/msproteomicstools.

Contact: lars@imsb.biol.ethz.ch

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

1 Introduction

In mass spectrometry-based proteomics, most currently available analysis and visualization software is focused on discovery proteomics workflows and, consequentially, on the representation of fragment ion spectra (Aebersold and Mann, 2003). Recent advances in targeted proteomics such as the development of high-throughput selected reaction monitoring (SRM) and SWATH-MS, have sparked interest in the computational analysis of chromatographic traces (Gillet *et al.*, 2012; Röst *et al.*, 2014a). These methods have substantially increased the number of transitions that can be concurrently analyzed in a single LC-MS/MS run from a few hundred to several hundred thousand.

Although multiple algorithms have been proposed in the literature to analyze such data (Reiter *et al.*, 2011; Röst *et al.*, 2014a; Teleman *et al.*, 2014), insufficient visualization capability and inability to manually verify the results have hampered their adoption in the research community. The lack of common data formats and missing cross-platform support are some of the main challenges for targeted proteomics analysis, which make combining and integrating multiple analysis workflows difficult. Often, a user is thus restricted to a single software environment, causing lock-in effects and preventing optimal data analysis. In addition, most currently available software tools were developed with low-throughput targeted proteomics data (SRM/MRM data) in mind and may not scale well to the large data volumes generated by next-generation high-throughput targeted proteomics pipelines such as SWATH-MS.

Here, we present TAPIR, a fast and efficient cross-platform software for visualizing chromatograms and corresponding peaks identified in targeted proteomics experiments. TAPIR uses standardized and open file formats for data access [mzML and TraML; Martens *et al.* (2011), Deutsch *et al.* (2012)] and is able to visualize

experiments whose size substantially exceeds system memory through efficient access to the raw data.

2 Implementation

The TAPIR Software (TArgeted Proteomics Information Representation Software) uses the Python scripting language together with Qt and Qwt (PyQt and PyQwt in Python) to implement robust plotting capabilities. It uses pymzML (Bald *et al.*, 2012) and pyOpenMS (Röst *et al.*, 2014b) to access mass spectrometric raw data natively through the standardized indexed mzML data format (Martens *et al.*, 2011), which allows for memory-efficient access to individual data vectors. Whenever the visualization of a specific compound is requested, the PyQt layer triggers a data load event. Next, the pymzML software library uses the pre-computed binary index to load the data into system memory and sends the data to the plotting device which displays the data using the guiqwt library. Modern software architecture principles following the model-view paradigm separate the representation of the data and its presentation. The meta-data (e.g. the grouping of transitions to a peptide precursor, the grouping of precursors to peptides and peptides to proteins) is clearly separated from the raw data and encoded in the TraML format, a standard format by the Proteomics Standards Initiative (Deutsch *et al.*, 2012). Thus, TAPIR can be readily integrated in any targeted proteomics pipeline that supports standardized data formats, such as the OpenSWATH pipeline (Röst *et al.*, 2014a).

Due to its implementation in the Python programming language, the TAPIR software is available on all three major platforms (Mac OS X, Linux and Windows). This will allow a large number of researchers to use it directly on their preferred operating system without having to install a different, potentially proprietary, software environment first. In addition, it allows rapid modification of the source code which can be achieved even by novice programmers.

3 Results

TAPIR is a novel tool designed to display and analyze targeted proteomics data to validate experimental results. It presents a highly interactive user interface allowing to zoom, pan, investigate individual data traces and export data as images or tables in a highly customizable fashion (Fig. 1). The tool is scalable to a large amount of input data (dozens of input files with several hundred thousands of transitions each) which makes it suitable to display and analyze high-throughput data. This is achieved by using the indexed mzML data format which allows fast indexed seeks and loading of only the requested chromatograms into memory. In addition, TAPIR uses highly optimized data structures based on the numerical numpy library to represent data and, to generate the plots, employs the efficient guiqwt plotting library. Even though no data is held in system memory, loading the data and generating the plots is generally done in less than a second, providing a fast and responsive user experience on the desktop.

To demonstrate the application of TAPIR, we have used the software to visualize multiple types of chromatographic data. First, we have applied it to an experiment investigating the virulence mechanisms of *Streptococcus pyogenes* using SWATH-MS and a targeted data analysis strategy (Supplemental Material). Specifically, several thousand peptides (more than 60 000 transitions) were monitored over multiple conditions and TAPIR allowed the visual inspection of peptides reported to have differential expression between conditions. The software displays the extracted ion chromatograms for each peptide and highlights the correct peak (as determined by OpenSWATH, for example). Further information relating to each peak (q-value, intensity, peak boundaries) is displayed alongside the chromatogram, facilitating visual validation of data obtained in high-throughput experiments. Next, we have extended the TAPIR software framework and adopted it for a metabolomics use-case. In the Supplementary Material, we show multiple extracted ion chromatograms in a metabolomics GC-MS (gas chromatography) experiment where three metabolites were quantified. Because the analysis also relied on mzML and TraML for data storage, no changes were needed to adopt TAPIR to GC-MS data of metabolites.

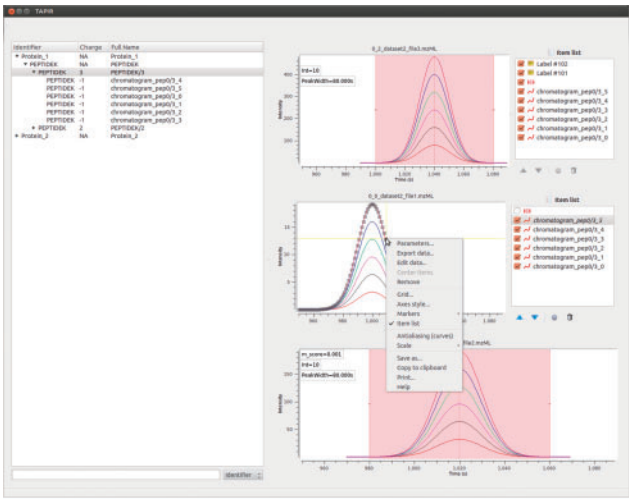

Fig. 1. Screenshot of user interaction with TAPIR displaying chromatographic data. The TAPIR software is highly flexible and interactive, allowing for investigation of single data traces and data points. Each graph item can be selected and inspected individually, allowing for customization of the visualization and production of publication-quality figures. Data can be exported as an image or in table format and used for further analysis; individual traces can be removed or re-added and all graph settings (such as color, line width, line style etc.) are fully customizable. The implementation relies on guiqwt for these features. Here, simulated data is shown

By design, TAPIR is built for high-throughput applications, is completely open-source and supports multiple computing platforms, scaling well with the number of transitions analyzed and the number of runs in a single experiment. This sets it apart from other visualization software for targeted proteomics, such as PeakView, Skyline (MacLean *et al.*, 2010) or TOPPView (Sturm and Kohlbacher, 2009); see also Supplemental Discussion. The design of TAPIR allows for rapid visual validation of results obtained from automated software tools and deep exploration of signals that were highlighted by downstream statistical analysis to confirm the correct identification and quantification of the analyte. Manual inspection by life science researchers increases confidence in the results and allows a tight integration of automated analysis pipelines with efficient visualization software. We hope that the availability of such a tool improves the quality of reported results in targeted proteomics and can contribute to increased confidence and transparency in the biological findings of future mass spectrometric studies.

Acknowledgements

The authors would like to thank Henning Kuich for providing us with a GC-MS run for visualization. H.R. was funded by ETH Zurich (ETH-30 11-2).

References

Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.

Bald,T. *et al.* (2012) pymzML–python module for high-throughput bioinformatics on mass spectrometry data. *Bioinformatics*, **28**, 1052–1053.

Deutsch,E.W. *et al.* (2012) TraML–a standard format for exchange of selected reaction monitoring transition lists. *Mol. Cell. Proteomics*, **11**, R111.015040.

Gillet,L.C. *et al.* (2012) Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics*, **11**, O111.016717.

MacLean,B. *et al.* (2010) Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*, **26**, 966–968.

Martens,L. *et al.* (2011) mzML–a community standard for mass spectrometry data. *Mol. Cell. Proteomics*, **10**, R110.000133.

Reiter,L. *et al.* (2011) mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat. Methods*, **8**, 430–435.

Röst,H.L. *et al.* (2014a) OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.*, **32**, 219–223.

Röst,H.L. *et al.* (2014b) pyOpenMS: a Python-based interface to the OpenMS mass-spectrometry algorithm library. *Proteomics*, **14**, 74–77.

Sturm,M. and Kohlbacher,O. (2009) TOPPView: an open-source viewer for mass spectrometry data. *J. Proteome Res.*, **8**, 3760–3763.

Teleman,J. *et al.* (2014) DIANA—algorithmic improvements for analysis of data-independent acquisition MS data. *Bioinformatics*, **31**, 555–562.