BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 10 2012, pages 1404–1405
doi:10.1093/bioinformatics/bts150

Data and text mining

Advance Access publication March 29, 2012

OCAP: an open comprehensive analysis pipeline for iTRAQ

Penghao Wang[1,*], Pengyi Yang[2] and Jean Yee Hwa Yang[1]

[1]School of Mathematics and Statistics and [2]School of Information Technologies, University of Sydney, Camperdown, NSW2006, Australia

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Mass spectrometry-based iTRAQ protein quantification is a high-throughput assay for determining relative protein expressions and identifying disease biomarkers. Processing and analysis of these large and complex data involves a number of distinct components and it is desirable to have a pipeline to efficiently integrate these together. To date, there are limited public available comprehensive analysis pipelines for iTRAQ data and many of these existing pipelines have limited visualization tools and no convenient interfaces with downstream analyses. We have developed a new open source comprehensive iTRAQ analysis pipeline, OCAP, integrating a wavelet-based preprocessing algorithm which provides better peak picking, a new quantification algorithm and a suite of visualizsation tools. OCAP is mainly developed in C++ and is provided as a standalone version (OCAP_standalone) as well as an R package. The R package (OCAP) provides the necessary interfaces with downstream statistical analysis.

Availability: OCAP is freely available and can be downloaded at http://www.maths.usyd.edu.au/u/penghao

Contact: penghao.wang@sydney.edu.au

Received on October 20, 2011; revised on March 12, 2012; accepted on March 23, 2012

1 INTRODUCTION

Accurate identification and quantification of protein expressions are crucial in developing new diagnostic, prognostic and therapeutic products for the treatment of various diseases. With the introduction of isobaric quantification technologies, such as iTRAQ and TMT, researchers are able to determine relative expressions of thousands of proteins simultaneously. However, analysis of iTRAQ data remains a very challenging task.

Typical workflow of iTRAQ data analysis can be viewed as two major components, the preprocessing of the data and higher statistical analysis. The first component can be further divided into three main stages: (i) spectrum peak picking; (ii) peptide and protein identification; and (iii) protein quantification. The second component consists of quality control, and higher level statistical analyses such as identification of differentially expressed proteins and prediction. Some and/or all of the components are usually combined in an analysis pipeline.

Currently, there are several pipelines designed for other purposes, e.g. TOPP (Kohlbacher *et al.,* 2007) for label-free quantification, maxQuant (Cox and Mann, 2008) for SILAC analysis. However, there are a limited number of public available pipelines specifically designed for iTRAQ from the initial preprocessing phase right through to higher level statistical analysis, and existing pipelines include the Trans-Proteomic Pipeline (TPP; Keller *et al.,* 2005), Multi-Q (Lin *et al*., 2006) and MSnbase (Gatto and Lilley, 2011). Many existing open source pipelines that support iTRAQ focus primarily on preprocesssing of iTRAQ data and offer limited visualization tools for efficiently exploring the data, e.g. TPP pipeline. As the eventual aim of iTRAQ analysis includes identifying differentially expressed proteins and finding biomarkers for good prediction outcome, it is important that the output from the three main stages from the preprocessing component is effectively integrated with major statistical softwares such as R, SPSS and SAS to facilitate downstream analyses.

To this end, we have developed a new comprehensive iTRAQ data analysis pipeline with the three main stages of the first component forming a standalone software. In addition, we have provided an R-interface (OCAP) of this software that includes additional visualization tools for exploring the data. This interface also provides easy access to the suite of downstream analytical packages including limma (Smyth, 2005), pamR (Tibshirani *et al.,* 1999) and isoBar (Breitwieser *et al*., 2011).

2 ANALYSIS COMPONENTS

OCAP expects input as mzXML raw spectra and generates analytical results in an automatic or a separate manner. Under the automatic mode, the pipeline directly produces peptide and protein level identification and quantification results through a single function pipeline_analyse. Figure 1a provides a diagrammatic view of OCAP and its main R functions. The analysis can also be completed separately for each component. We will illustrate the utility of our pipeline through a few figures using a published 4-plex iTRAQ dataset (Karp *et al*., 2010).

(1) *Peak picking*: OCAP utilizes DyWave algorithm (Wang *et al*., 2010) for spectrum peak picking. It dynamically adjusts the peak model based on the spectrum and takes account of additional information such as shapes of the signal during peak detection. OCAP provides users several visualization tools for evaluating preprocessing results and exploring the data at spectrum level. For example, showspectrum shows the ion series of a specific preprocessed spectrum and compare_spectrum enables comparison of spectra. These functions provide users a way to evaluate the reliability of peptide-to-spectrum matches.

(2) *Protein identification*: OCAP adopts the widely used open source X!Tandem (Craig and Beavis, 2004) for protein identification. A FASTA format protein sequence database is required for identification. OCAP uses the expect value (*E*-score) of X!Tandem as the final protein scoring model. Only unique

*To whom correspondence should be addressed.

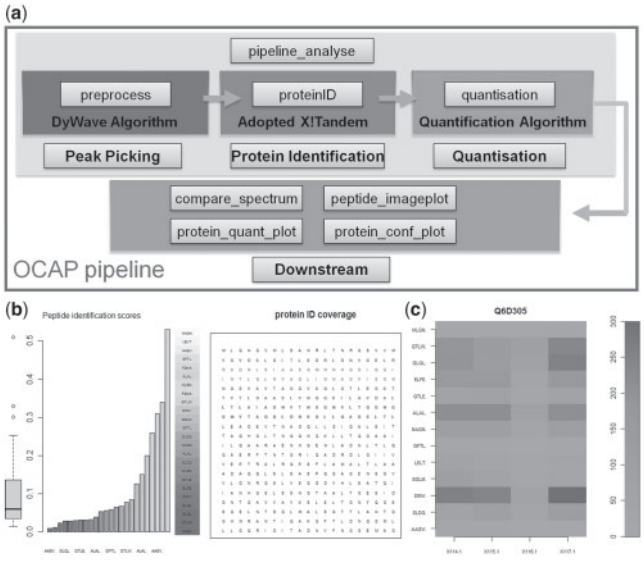

Fig. 1. (**a**) Overview of OCAP pipeline, its analysis components and its major R functions; (**b**) the identification coverage and confidence graph for a protein; (**c**) the protein quantification image-plot.

peptides are considered for protein assignment and quantification. The identification results will be automatically parsed and loaded for quantification. At this stage, users have the flexibility to output the peptide identification and protein assignment results in a tab-delimited file for separate analysis.

(3) *Protein quantification*: OCAP uses a wavelet-based algorithm for quantification. The algorithm firstly applies a continuous wavelet approach similar to DyWave to dynamically identify iTRAQ reporter ions and the peak centroids. Secondly, the algorithm applies the spatially selective signal filtration technique (Xu *et al.,* 1994) to detect the edges of the identified reporter ions, extracts the iTRAQ reporter ion signals from noise, and finally automatically corrects isotope impurity. Thus, spectrum artefacts such as mass shift, baseline effect and noise interface can be significantly alleviated comparing to the traditional approach of summing all peak intensity within a predefined mass window as quantification. OCAP also provides the option to use an intensity approach if users so desire. Impurity of iTRAQ reagents is automatically corrected as specified by the manufacturer. OCAP provides peptide level and protein level quantification within R as a data.frame. Two text-readable files for peptide and protein level results may be exported, which can be loaded to Excel or other preferred statistical analysis software for further analysis.

(4) OCAP provides a number of exploratory visualization tools for iTRAQ data. These tools can significantly facilitate our understanding of the data and quality control. For example, users may display all peptides for a protein and determine if some peptides have inconsistent quantifications. These provide a visual quality check of the matched spectra and an option to remove spurious peptides. A protein identification graph (Fig. 1b) shows the peptide identification score distribution and identification coverage which provides an indication of the protein identification confidence. In this example, the coverage is moderate, and the peptide identifications scatter at the 1st half of the protein, thus indicating that the identification and quantification from the 2nd half may not be so reliable. Figure 1c presents an image-plot for evaluating protein quantification, which can help users to get an overview of the concordance of peptide expressions for a protein across samples. It demonstrates that for this protein the expression is down-regulated in 116.1 sample while 114.1 and 117.1 samples seem to have the highest expression. Most of the peptides show consistent expression for this trend, but there are two peptides that have low expression (as shown in light green), and users may want to remove these peptides for quality control.

At this stage, depending on the analytical aim, users can utilize other R packages such as: pamR for classification and identifying protein biomarkers; limma for differential protein analysis; KEGG for functional grouping and pathway analysis and many others.

3 CONCLUSION

Being open source, OCAP can be easily extended and modified to fit specific analyses. It provides an alternative workflow to the TPP pipeline. OCAP also incorporates a range of visualization tools for exploring the iTRAQ data and a convenient interface to many downstream analyses, greatly facilitating the understanding of the underlying biological problem.

ACKNOWLEDGEMENTS

We adapted source codes from TPP for processing mzXML, and X!Tandem source codes for identification. We thank Dr Vivek Jayaswal and Kaushala Jayawardana from our group for testing the package.

Funding: The work is supported by ARC Discovery Grant (DP0984267). P.Y is supported by NICTA scholarship.

Conflict of Interest: none declared.

REFERENCES

Breitwieser,F.P. *et al.* (2011) General statistical modelling of data from protein relative expression isobaric tags. *J. Proteome Res.*, **10**, 2758–2766.

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

Craig,R. and Beavis,R. (2004) TANDEM: matching proteins with mass spectra. *Bioinformatics,* **20**, 1466–1467.

Gatto,L. and Lilley,K.S. (2011) MSnbase – an R/Bioconductor package for isobaric tagged mass spectrometry data visualisation, processing and quantitation. *Bioinformatics*, [Epub ahead of print, doi: 10.1093/bioinformatics/btr64, November 2011].

Karp,N.A. *et al.* (2010) Addressing accuracy and precision issues in iTRAQ. *Mol. Cell. Proteomics*, **9**, 1885–1897.

Keller,A. *et al.* (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.*, **1**, E1–E8.

Kohlbacher,O. *et al.* (2007) TOPP-the OpenMS proteomics pipeline. *Bioinformatics,* **23**, e191–e197.

Lin,W.T. *et al.* (2006) Multi-Q: a fully automated tool for multiplexed protein quantitation. *J. Proteome Res.,* **5**, 2328–2338.

Smyth,G.K. (2005) Limma: linear models for microarray data. In: Gentleman,R. *et al.* (eds). *Bioinformatics and Computational Biology Solutions using R and Bioconductor.* New York: Springer; pp. 397–420.

Tibshirani,R. *et al.* (1999) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.

Wang,P. *et al.* (2010) A dynamic wavelet-based algorithm for pre-processing tandem mass spectrometry data. *Bioinformatics,* **26**, 2242–2249.

Xu,Y. *et al.* (1994) Wavelet transform domain filters: a spatially selective noise filtration technique. *IEEE Trans. Image Process.,* **3**, 747–758.