OCAP: An Open Comprehensive Analysis Pipeline for iTRAQ

Penghao Wang¹,*, Pengyi Yang² and Jean Yee Hwa Yang¹
¹School of Mathematics and Statistics, University of Sydney, Australia.
²School of Information Technologies, University of Sydney, Australia.

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 visualisation 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 visualisation 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

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 (1) spectrum peak picking, (2) peptide and protein identification, and (3) 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 MSnibase (Galto and Lilley 2011). Many existing open source pipelines that support iTRAQ focus primarily on preprocessing of iTRAQ data and offer limited visualisation 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 3 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 3 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 visualisation 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. Fig. 1(a) 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 utilises 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 visualisation tools for evaluating preprocessing results and exploring the data at spectrum level. For example, showspectrum shows the ion series of a specific pre-processed 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 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 dataframe. 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 explorative visualisation 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. 1(b)) shows the peptide identification score distribution and identification confidence 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. Fig. 1(c) 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 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 utilise 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 visualisation 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. Funding: ARC Discovery Grant (DP0984267); NICTA scholarship (to P.Y.). We thank Dr. Vivek Jayaswal and Kaushala Jayawardana from our group for testing the package.

REFERENCES