Towards reproducible MSMS data preprocessing, quality control and quantification

Laurent Gatto and Kathryn S. Lilley

Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge

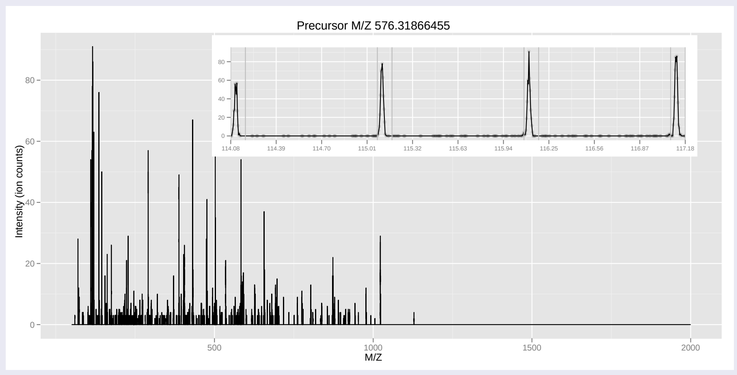
BSPR/EBI Proteomics Meeting – 13 - 15th July 2010

Introduction

- The development of `MSnbase` aims at providing researchers dealing with labelled quantitative proteomics data with a transparent, portable, extensible and open-source collaborative framework to easily manipulate and analyse MS^2-level tandem mass spectrometry data.
- `MSnbase` has been developed following an object-oriented programming paradigm: all information that is manipulated by the user is encapsulated in *ad hoc* data containers to hide it's underlying complexity.
- The implementation in R gives users and developers a great variety of powerful tools to be used in a controlled and reproducible way.

Data inspection

- Data is seamlessly loaded from `mzXML` format and full or specific regions of MS^2 spectra can be plotted.

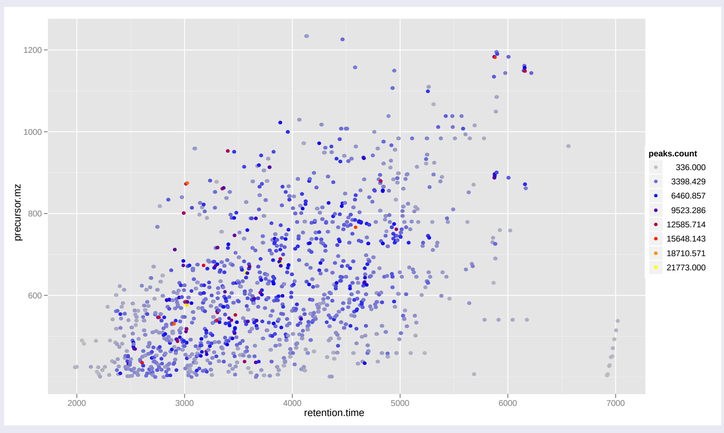


- Peptide identification data can easily be incorporated.
- **Traceability**: data, meta data and processing logging.

```
- - - Meta data  - - -
 Data description: iTRAQ4 Spiked-In Experiment
 Loaded from:
    /Data/QTofPremier/ksliTRAQ4mix_1to100_REG_1.mzXML
- - - Processing information - - -
 Data Loaded: Mon Jul  5 10:34:59 2010
 Curves <= 2 set to '0': Wed Jul  7 09:40:42 2010
 Quantification by trapezoidation iTRAQ4: Wed Jul 7 ...
 Normalised (sum): Wed Jul  7 13:50:54 2010
```

Precursor data inspection

- Several information (retention time, peaks count, ...) about the precursor ions is readily available and can be used for quality control and instrument setup optimisation.



- We can, for example, compare the MS1 acquisition numbers of precursor ions and identify which ones have been selected multiple times.

	number.selection
1	1003
2	52
3	7
4	2

Data processing

- Relevant spectra can be extracted (based on the precursor MZ values or peptide identification) and *identical* spectra (based on MZ values of peptide identification) can be *merged*.
- Data can be *cleaned* up (removal of low intensity peaks, background subtraction).

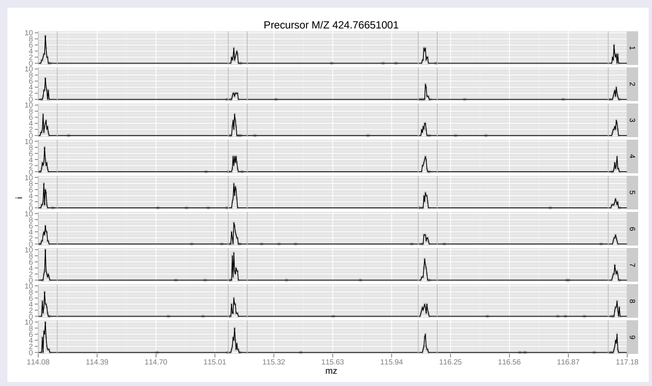
This work was funded by a BBSRC Tools and Resources Development Fund Award

Assessing incomplete dissociation

- We created a new set of reporter ions, including the 4 iTRAQ tags and a virtual tag at MZ 145, which corresponds to partially fragmented reporter tags and balance groups and quantified these 5 peaks.
- In 99.4% of the cases, no peaks were quantified at MZ 145. When some signal was detected, it was not significant, indicating that incomplete collision induced dissociation is not an issue for this data set.

Quantitation

- `MSnbase` is flexible and extensible (works also for TMT6 and iTRAQ8-plex).
- Allows facile inspection of individual MS^2 spectra.



- Allows quantitation at MS^2 level.

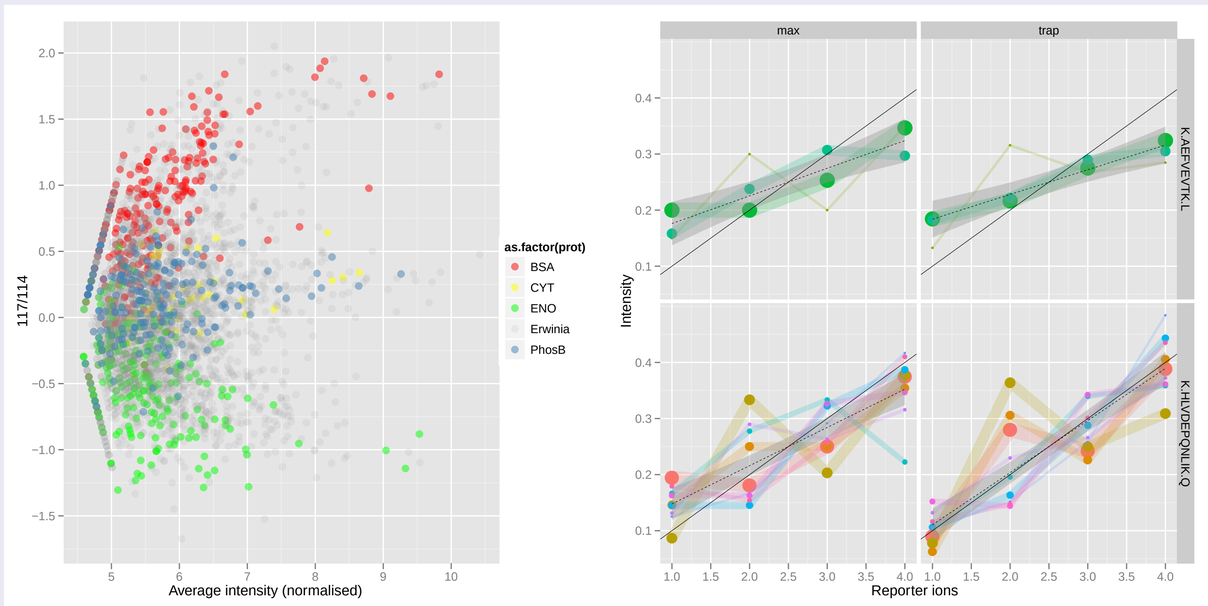


Figure: Four exogenous proteins have been spiked in an *Erwinia* background and labelled with 4 iTRAQ labels (BSA 1:2:3:4, cytochrome C and phosphorylase B 1:1:1:1, Enolase 4:3:2:1). On the left, we illustrate the 117/114 ratios for all the individual MS^2 spectra. Details for two peptides from the BSA protein, are shown on the right. The plots show the normalised intensities (y axes) for the four reporter ions (along the x axes) and are organised by quantitation method along the columns (maximum or area of the reporter peaks), and by peptide sequence (along the rows). Different colours represent different MS^2 spectra and the size of the points is proportional to the precursor peak counts. Solid lines represent the theoretical values and dashed lines are linear models fitted to the data.

Conclusions and perspectives

- Softwares are more than mere programs that are used to transform an input into an output and should be recognised as the effective implementation of the analytical methodology that is applied. Research aims to be reproducible across time and teams and reliably documented open software is of paramount importance. `MSnbase` is a framework that allows researchers to investigate their data in depth in a traceable and reproducible way.
- The functionality currently implemented in `MSnbase` will be expanded and will be used to investigate MSMS data at the MS^2 spectrum level from different experimental designs and instruments. Reciprocally, the insights provided by `MSnbase` are vital input for instrument setup optimisation.
- `MSnbase` will be publicly released under an open licence. If you wish to be kept informed of its release and future development, please do not hesitate to share your contact details at the conference or contact us later.

Nature Precedings : doi:10.1038/npre.2010.5010.1 : Posted 14 Oct 2010