MS-based proteomics using Bioconductor

Laurent Gatto
lg390@cam.ac.uk

Cambridge Centre For Proteomics (CCP)
University of Cambridge

European Bioinformatics Institute (EBI)

1st July 2011
Plan

1. **Introduction**
   - Motivation
   - Mass spectrometry

2. **Data structures**

3. **Application**
   - A typical workflow
   - Use cases

4. **Future work**
Plan

1. **Introduction**
   - Motivation
   - Mass spectrometry

2. Data structures

3. Application
   - A typical workflow
   - Use cases

4. Future work
Motivation

- Many pieces of software are black boxes and just return values.
- Little/no solution to explore raw data and effect of processing/transformation.

Goals of MSnbase

- Apply the Bioconductor software model to MS-based proteomics
- Use robust and annotation rich data structure.
- Re-use algorithms readily available.
- Integration of genetic, genomic, proteomic, metabolomic data.
Plan

1. **Introduction**
   - Motivation
   - Mass spectrometry

2. **Data structures**

3. **Application**
   - A typical workflow
   - Use cases

4. **Future work**
Classes

- MSnExp - MS(MS) experiment.
- Spectrum, Spectrum1 and Spectrum2 – mass spectra.
- ReporterIons defines reporter ions – data(iTRAQ4).
- MSnSet – quantified expression.

- Additional meta-data in MSnProcess and MIAPE.
Plan

1. Introduction
   - Motivation
   - Mass spectrometry

2. Data structures

3. Application
   - A typical workflow
   - Use cases

4. Future work
1. `readMzXMLData()` to create an `MSnExp` instance
2. `plot()` subset of `MSnExp` or `Spectrum`
3. Quality control (see later)
4. Processing: `removePeaks`, `bg.correct`
5. `quantify(MSnExp,ReporterIons)` to create an `MSnSet` instance
6. `purityCorrect(MSnSet,impurities)`
7. `normalise(MSnSet,"vsn")`
8. ...
Number of times a precursor ion has been selected

Optimise MS parameters.

```r
allPrecs <- precursorMz(raw)
number.selection <- c()
ms1scanNums <- ms1scan(raw)
for (mp in unique(allPrecs))
  number.selection <- c(number.selection,
                         length(unique(ms1scanNums[allPrecs==mp])))
names(number.selection) <- unique(allPrecs)
print(table(number.selection))

number.selection
  1   2   3   4
5337  52   2   2
```
QC1 – Experiment-wide visualisation

L. Gatto  
CSAMA 2011 – Brixen
QC1 – Experiment-wide visualisation

- Total ion current
- Peak count
- Precursor M/Z
QC2 – Assessing incomplete dissociation

```r
> foo <- quantify(itraqdata,"trap",iTRAQ5,verbose=FALSE)
> boxplot(exprs(foo),col=iTRAQ5@col,log="y")
```
QC3 – Spectra quality

Histogram of Mass Delta Distributions for PRIDE experiment 12011

L. Gatto  CSAMA 2011 – Brixen
1. Introduction
   - Motivation
   - Mass spectrometry

2. Data structures

3. Application
   - A typical workflow
   - Use cases

4. Future work
for (i in TODO)

- On-disk random access of data (using proteowizard library) – mzR package under development with Bernd Fischer (EMBL) and Steffen Neuman (IPB HALLE, xcms).
- Some processing is embarrassingly easy to parallelise.
- Label-free quantitation.
- Easier integration of identification data.
- ...
More info, other packages

- MSnbase vignettes
- Proteomics sig mailing list – https://stat.ethz.ch/mailman/listinfo/bioc-sig-sequencing
- BiocViews – MassSpectrometry and Proteomics
- CRAN Task View – Chemometrics and Computational Physics

Acknowledgement

- Kathryn Lilley and CCP team.
- BBSRC Tools and Resources Development Fund Award.
- PRIME-XS FP7.

Thank you for your attention.