RockerBox

Filtering massive Mascot search results at the .dat level
Challenges

• “Big” experiments
• High amount of data
• Large raw and .dat files (> 2GB)

• How to handle our results??
  – The ‘2.2’ peptide summary could not be made by Mascot
  – MSQuant couldn’t load the result files
Proteomics workflows

Sample

Cultured cells, yeast, tissue samples, stem cells, ...

Protein digestion

Trypsin, LysN, LysC, Chymo, V8, Pepsin

Sample fractionation

SCX, ZIC-HILIC, ZIC-cHILIC, Ti-IMAC

Database search, statistical analysis

Peptide fragmentation

CID, ETD, HCD

NanoLC-MS
Boxes

• ‘proteomics’ mass spectrometers
  – 3 Orbitrap (Thermo)
  – 2 Orbitrap Velos(Thermo)
  – Quadropole TOFs (Agilent, Waters and AB Sciex)
  – 2 Triple Quad (AB Sciex, Thermo)

• 70 Terabytes of stored data

• Software:
  – Preprocessing scripts: in-house, MaxQuant, Proteome Discoverer, Scaffold, MSQuant…
  – Mascot 2.3 (Linux)
Large MS experiment

Biological Replica 1

Biological Replica 2

“Big SCX”

“Small SCX”

40 LC-MS/MS (CID)

14 LC-MS/MS (ETD)

PTM ids

Protein ids

40 LC-MS/MS (CID)

14 LC-MS/MS (ETD)

PTM ids

Protein ids
Experimental design and MS results

- Biological replica (label swap)
- 196 LC-MS/MS (3 h gradient)
- LTQ-Orbitrap: CID/ETD
- 2,440,583 MS/MS spectra collected
- 568,054 PSMs (FDR=1.02%)
- 68,172 unique peptides
- **10,683 unique proteins**

130 GB raw files
12 GB .dat files
ROCKERBOX

Meeting the challenges
What is RockerBox?

• Filtering .dat file peptide spectrum matches (PSMs)
• Charting of search results
• Combining .dat files (new)
• Exporting text files with PSMs

• Cross-platform usability (Java)

Example data set

Mascot ions score over Mass delta (ppm) plot of F274658.dat.db
Wide search window

Mascot ions score over Mass delta (ppm) plot of F274658.dat.db

Mascot ions score over Mass delta (ppm) plot of F275563.dat.db
Mascot ions score over Mass delta (ppm) plot of
TT1_110516_MLH_PG_HeLa_Alba3_1ug_180min_excl_window_25s_top20_50ms_rolling_CE_thres
hold500.dat.db
Mass delta (ppm) over Retention time plot of TT1_110516_MLH_PG_HeLa_Alba3_1ug_180min_excl_window_25s_top20_50ms_rolling_CE_thres_hold500.dat.db
Mass-based calibration applied

Mass delta (ppm) over Retention time plot of
TT1_110516_MLH_PG_HeLa_Alba3_1ug_180min_excl_window_25s_top20_50ms_rolling_CE_thres
hold500_recal_371_445.dat.db

Retention time

Mass Delta (ppm)
Mass-based calibration applied

Mascot ions score over Mass delta (ppm) plot of
TT1_110516_MLH_PG_HeLa_Alba3_1ug_180min_excl_window_25s_top20_50ms_rolling_CE_thres
hold500_recal_371_445.dat.db
Workflow
Removing PSMs?

- Many spectra are not matched to a correct peptide sequence
  - Low quality real spectra (signal/noise ratio)
  - Spectra from non-peptide origins
  - Mixed peptide spectra
  - Spectra from peptides not in the database
- These low quality matches are abundant
  - Typically around 50%
- Which PSMs really matter?
ROCKERBOX FILTERING

METHODS

An overview
Manual filter: full control

- Mascot score
- Modifications
- Mass delta
Manual filtering

Mascot ions score over Mass delta (ppm) plot of F274658.dat.db
Manual filter results

Mascot ions score over Mass delta (ppm) plot of F274658.dat.db

18853 → 5335

Is this optimal?

What’s the FDR?
What’s an FDR

• False Discovery Rate

• The FDR is the proportion of matches in the result set, expected to be false
  – Usually a percentage
FDR estimation methods

\[
T_s = \text{Accepted target (known) sequences} \\
D_s = \text{Accepted decoy (nonsense) sequences} \\
\text{FDR}_s \approx \frac{D_s}{(T_s + D_s)}
\]

Competitive

- Decoy and target sequences combined in one database
- A spectrum matches either a decoy or a target sequence

Non-competitive

- Search separate Decoy and Target databases
- A spectrum can match both decoy and target sequences
Automatic FDR based filtering

- FDR guaranteed
- 50% Automatic
- Mass window
- Different decoy strategies
- Possibility to use on separate mass spec runs
FDR based filtered file

Mascot ions score over Mass delta (ppm) plot of F274658.dat.db

cutoff: 15.6
18853 → 5880
Fractions are not the same...

Number of unique peptides

Peptide cut-off score

2 P  N-Ac  1 P increasing # basic residues
FDR based filtered file

Mascot ions score over Mass delta (ppm) plot of F274658.dat.db

18853 → 5992
# PSM properties

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>Identifier. RockerBox uses the form <em>db</em>_querynumber_rank, in which <em>db</em> may be ‘target’ for Mascot automatic decoy real database, ‘decoy’ for Mascot automatic decoy scrambled database or ‘combined’ for a concatenated decoy strategy</td>
</tr>
<tr>
<td>label</td>
<td>-1 if decoy, 1 if target PSM</td>
</tr>
<tr>
<td>charge</td>
<td>Precursor charge</td>
</tr>
<tr>
<td>score</td>
<td>Mascot score</td>
</tr>
<tr>
<td>deltaScore</td>
<td>Difference between current rank score and ‘next’ rank score</td>
</tr>
<tr>
<td>mr</td>
<td>Measured precursor mass</td>
</tr>
<tr>
<td>deltaM</td>
<td>Delta mass between precursor mass and matched peptide mass</td>
</tr>
<tr>
<td>deltaMPpm</td>
<td>deltaM relative to matched peptide mass</td>
</tr>
<tr>
<td>absDeltaM</td>
<td>Absolute value of deltaM</td>
</tr>
<tr>
<td>absDeltaMPpm</td>
<td>Absolute value of deltaMPpm</td>
</tr>
<tr>
<td>isoDeltaM</td>
<td>Delta mass allowing for 1, 2, 3 or 4 Dalton difference</td>
</tr>
<tr>
<td>isoDeltaMPpm</td>
<td>isoDeltaM relative to matched peptide mass</td>
</tr>
<tr>
<td>missedCleavages</td>
<td>Number of missed cleavages</td>
</tr>
<tr>
<td>fragMassError*</td>
<td>RMS error of the MS2 spectrum to the theoretical spectrum</td>
</tr>
<tr>
<td>totalIntensity*</td>
<td>Total intensity of the MS2 spectrum</td>
</tr>
<tr>
<td>intMatchedTot*</td>
<td>Total intensity of matched MS2 peaks</td>
</tr>
<tr>
<td>relIntMatchedTot*</td>
<td>intMatchedTot divided by totalIntensity</td>
</tr>
<tr>
<td>fraclonsMatched*</td>
<td>Fraction of all MS2 peaks matched</td>
</tr>
<tr>
<td>peptide</td>
<td>Peptide sequence</td>
</tr>
<tr>
<td>proteins</td>
<td>The list of proteins from the search database that contain the peptide sequence</td>
</tr>
</tbody>
</table>
Using the Percolator algorithm

- Fully automatic
- Both decoy strategies
- Different score outputs
- Apply to separate spectrometry runs
Percolator filtered file
# Overview of filtering methods

<table>
<thead>
<tr>
<th>File</th>
<th>Size</th>
<th>Number of PSMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original .dat</td>
<td>95 MB</td>
<td>18853</td>
</tr>
<tr>
<td>Manually filtered</td>
<td>12 MB</td>
<td>5335</td>
</tr>
<tr>
<td>FDR 1%</td>
<td>11 MB</td>
<td>5880</td>
</tr>
<tr>
<td>FDR 1% per fraction</td>
<td>15 MB</td>
<td>5992</td>
</tr>
<tr>
<td>Percolator FDR 1%</td>
<td>15 MB</td>
<td>6671</td>
</tr>
</tbody>
</table>
PHOSPHORYLATION SITE COUNTS

Use case
Experiment

- Separate distinct peptide populations using SCX
- Enrich Phosphopeptide using Ti-IMAC
- Test in multiple fragmentation methods
Use case: phosphorylation counting

• Houjiang Zhou:

  Which unique sites (protein level) do I have?

  Maximize the number of phosphopeptides identified

  Reliable data with known FDR

  I need it tomorrow, the reviewer is waiting

  Which fragmentation method is best for phosphopeptides? CID, ETD, HCD
Data analysis (WIP)

- Use RockerBox to filter the .dat file
  - using Percolator, FDR 1%, with a minimum Mascot score of 20
- Extract *PTM delta scores* using RockerBox csv export
- Count the phosphorylated peptides and phosphorylation sites on the proteins
Results

19,692 uni.phosphopeptides
16,624 uni.phosphosites
3862 phosphoproteins
Conclusions

• RockerBox helps to alleviate size problems
  – Complex research problems can be addressed more easily
Acknowledgements

Lukas Käll
Javier Muñoz, Reinout Raijmakers, Shabaz Mohammed

Bas van Breukelen, Albert J.R. Heck
Thank you

http://www.hecklab.nl
Availability

http://trac.nbic.nl/rockerbox