Pragmatic Computational Proteomics

Taejoon Kwon

UNIST, Republic of Korea
Xenopus Bioinformatics Workshop, May 2016

Images from Dr. Kyowon Jeong (UCSD)
Goal of this session

• How does MS/MS data look like?
  – RAW, mzXML, mzML, MGF, DAT, ...
  – “ProteoWizard saves the day!”

• How to analyze it?
  – Comet: open-source SEQUEST

• How does search outcome look like?
  – pepXML, OUT, SQT, etc....
MS/MS data is a collection of peaks (Martin already told us)

http://www.piercenet.com/method/quantitative-proteomics
MS/MS data is a collection of peaks

<table>
<thead>
<tr>
<th>Time (min)</th>
<th>Purity Absorbance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>11.67</td>
</tr>
<tr>
<td>0.05</td>
<td>22.01</td>
</tr>
<tr>
<td>0.10</td>
<td>39.07</td>
</tr>
<tr>
<td>0.15</td>
<td>58.31</td>
</tr>
<tr>
<td>0.20</td>
<td>88.62</td>
</tr>
<tr>
<td>0.25</td>
<td>114.73</td>
</tr>
</tbody>
</table>

Creation Date: Tue May 13 15:24:09 2014
Extractor: ProteoWizard
Extractor version: Xcalibur software
Source file: Kwon201207_XENLA_st10a_1.raw

<table>
<thead>
<tr>
<th>RT</th>
<th>BPI</th>
<th>BPM</th>
<th>TIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01202167</td>
<td>33873.3</td>
<td>430.8882</td>
<td>3600783</td>
</tr>
<tr>
<td>0.01542088</td>
<td>4965.117</td>
<td>1542.088</td>
<td>1542.088</td>
</tr>
<tr>
<td>0.01322318</td>
<td>18582.8</td>
<td>301.1617</td>
<td>564.8149</td>
</tr>
<tr>
<td>0.01570301</td>
<td>340.2546</td>
<td>340.2546</td>
<td>340.2546</td>
</tr>
<tr>
<td>0.01915907</td>
<td>220.9704</td>
<td>220.9704</td>
<td>220.9704</td>
</tr>
<tr>
<td>0.02727461</td>
<td>159.0758</td>
<td>159.0758</td>
<td>159.0758</td>
</tr>
<tr>
<td>0.03062102</td>
<td>297.1285</td>
<td>297.1285</td>
<td>297.1285</td>
</tr>
<tr>
<td>0.03396608</td>
<td>360.2810</td>
<td>360.2810</td>
<td>360.2810</td>
</tr>
<tr>
<td>0.03731821</td>
<td>449.3398</td>
<td>449.3398</td>
<td>449.3398</td>
</tr>
<tr>
<td>0.04273833</td>
<td>520.3821</td>
<td>520.3821</td>
<td>520.3821</td>
</tr>
<tr>
<td>0.04814845</td>
<td>578.3949</td>
<td>578.3949</td>
<td>578.3949</td>
</tr>
<tr>
<td>0.05355857</td>
<td>651.3688</td>
<td>651.3688</td>
<td>651.3688</td>
</tr>
<tr>
<td>0.05896869</td>
<td>729.4212</td>
<td>729.4212</td>
<td>729.4212</td>
</tr>
<tr>
<td>0.06437881</td>
<td>797.4848</td>
<td>797.4848</td>
<td>797.4848</td>
</tr>
<tr>
<td>0.06978893</td>
<td>865.5485</td>
<td>865.5485</td>
<td>865.5485</td>
</tr>
</tbody>
</table>

http://1.bp.blogspot.com/-QHvLM46g8Hc/UpX8PqiBoFl/AAAAAAAACcg/MRt50VI1n-c/s1600/TMT10_whole_thing.jpg
XML – “Formal way to present data”
(but it is a pain of ass!)

<note date="05-14-2014">
<to> Tove </to>
<from> Jani </from>
<heading> Reminder </heading>
<body>
Don't forget me this weekend!
</body>
</note>
ProteoWizard: A Swiss-knife

ProteoWizard

The ProteoWizard Library and Tools are a set of modular and extensible open-source, cross-platform tools and software libraries that facilitate proteomics data analysis.

The libraries enable rapid tool creation by providing a robust, pluggable development framework that simplifies and unifies data file access, and performs standard chemistry and LCMS dataset computations.

Core code and libraries are under the Apache open source license; the vendor libraries fall under various vendor-specific licenses.

Features

- reference implementation of the new HUPO-PSI mzML standard mass spectrometry data format
- implementation of the new HUPO-PSI mzIdentML standard mass spectrometry data format
- modern C++ techniques and design principles
- cross-platform support (MSVC on Windows, gcc on Linux, XCode on OSX)
- extensive extensibility
- of data analysis tools
- both academic and commercial projects (Apache v2)
- many vendor raw data formats (on Windows)

# convert data.RAW to data.mzML
msconvert data.RAW

# convert data.RAW to data.mzXML
msconvert data.RAW --mzXML

# put output file in my_output_dir
msconvert data.RAW -o my_output_dir

# extract scan indices 5...10 and 20...25
msconvert data.RAW --filter "index [5,10] [20,25]"

# extract MS1 scans only
msconvert data.RAW --filter "msLevel 1"

# extract MS2 and MS3 scans only
msconvert data.RAW --filter "msLevel 2-3"

http://proteowizard.sourceforge.net/
Basic concept of “search”
(skip it if Martin is already covered)

Peptide
QVSVVVDLTLTNR

Experimental spectrum

MS/MS

DB search engine

Compare

Protein sequence database

Theoretical spectrum

m/z

m/z

Ranked list of peptide matches

Rank Peptide Score
1 QVSVVVDLTLTNR 3.3
2 VVEELCTPEGK 2.1
3 DLLLQWCWENGK 1.9
4 ECDVVSNTIIAEK 1.7
5 GDAVFVIDALNR 1.6
...

n SYLFCEMAEAEK 0.2

Ranked list of Peptide matches

Assign peptide To best match

SEQUEST: a pioneer (1994)

An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database

Jimmy K. Eng, Ashley L. McCormack, and John R. Yates, III
Department of Molecular Biotechnology, University of Washington, Seattle, Washington, USA

Multiple versions of SEQUEST

Crux (Bill Noble group, UW)

Score comparison for xcorr

Tide (Bill Noble group, UW)

Diament & Noble, J Prot Res (2011)
Comet: An open-source MS/MS sequence database search tool

Jimmy K. Eng¹, Tahmina A. Jahan¹ and Michael R. Hoopmann²

¹ Department of Genome Sciences, University of Washington, Seattle, WA, USA
² Institute for Systems Biology, Seattle, WA, USA

Welcome to the Comet project!

com·et n. ˈkō-mēt\n> a celestial body with a fuzzy head of ice and dust and characteristic long tail
> a common variety of fancy goldfish with a deep, forked tail
> an open source tandem mass spectrometry (MS/MS) sequence database search tool

Comet MS/MS

Searching uninterpreted tandem mass spectra of peptides against sequence databases is the most common method used to identify peptides and proteins. Since this method was first developed in 1993, many commercial, free, and open source tools have been created over the years that accomplish this task.

Although its history goes back two decades, the Comet search engine was first made publicly available in August 2012 on SourceForge under the Apache License, version 2.0. Comet is multithreaded, supports multiple input and output formats, and binaries are available for both Windows and Linux operating systems.

http://comet-ms.sourceforge.net
MaxQuant & Persus: all-in-one (but Windows only)

http://www.maxquant.org
Check out YouTube Channel for “MaxQuant Summer School” (on the course wiki)
Dataset for today

- Two biological replicates of unfertilized eggs
  - No ‘yolk protein depletion’; two injections/sample.
- Two biological replicates of stage 10 embryos
  - No ‘yolk protein depletion’; two injections/sample.
- RAW and mzXML are available at /opt/xenopus/Proteomics/
- Database (also at /opt/xenopus/annot/)
  - (forward) XENLA_2014may.prot_annot_longest.fa
  - (reversed) XENLA_2014may.prot_annot_longest.fa
  - (combined) XENLA_2014may.prot_combined.fa
Xenopus provides unique Opportunities for Protoemic experiments

• Amount of protein needed for deep proteomics ~ 100ug protein per condition is: 3 X.l embryo, 100 staged Drosophila embryos, 250 mouse embryos, $10^6$ somatic cells per sample

• Amount of protein for phospho- and ubiquitin proteomics is ~ 100 times greater

• Superbly timed cell cycle or developmental stages easily obtainable in large quantities

• Unique extract system for biochemical isolation and reconstitution

• Physical manipulation: Animal caps/Germinal vesicle isolation

From Martin Wuehr’s slide, Xenopus Bioinformatics Workshop 2014
Mass-spectrometer measures mass/charge.

After ionization, acceleration, and selection of single velocity particles, the ions move into a mass spectrometer region where the radius of the path and thus the position on the detector is a function of the mass.

\[ r = \frac{mv}{qB} = \frac{mE_s}{qBB_s} \]

From Martin Wuehr’s slide, Xenopus Bioinformatics Workshop 2014
Fragmentation of Peptide Ions

Note: Only break one bond per molecule on average

From Martin Wuehr’s slide, Xenopus Bioinformatics Workshop 2014
YMR134W, yeast protein involved in iron metabolism

From Martin Wuehr’s slide, *Xenopus* Bioinformatics Workshop 2014
“Shotgun Sequencing” of Complex Peptide Mixtures

From Martin Wuehr’s slide, Xenopus Bioinformatics Workshop 2014
Hybrid Instrument: Linear Ion Trap - Orbitrap

From Martin Wuehr’s slide, *Xenopus* Bioinformatics Workshop 2014

1. Ions are stored in the Linear Trap
2. … are axially ejected
3. … and trapped in the C-trap
4. … they are squeezed into a small cloud and injected into the Orbitrap
5. … where they are electrostatically trapped, while rotating around the central electrode and performing axial oscillation

The oscillating ions induce an image current into the two outer halves of the orbitrap, which can be detected using a differential amplifier

Ions of only one mass generate a sine wave signal

From Martin Wuehr’s slide, *Xenopus* Bioinformatics Workshop 2014
The axial oscillation frequency follows the formula

$$\omega = \sqrt{\frac{k}{m/z}}$$

Where

- $\omega$ = oscillation frequency
- $k$ = instrumental constant
- $m/z$ = …. well, we have seen this before

Many ions in the Orbitrap generate a complex signal whose frequencies are determined using a Fourier Transformation

From Martin Wuehr’s slide, *Xenopus* Bioinformatics Workshop 2014
Spectral Matching: SEQUEST/Mascot

From Martin Wuehr’s slide, *Xenopus* Bioinformatics Workshop 2014
SEQUEST – Under the Hood

\[ C_{ij} = \int_{-\infty}^{\infty} x(t)y(t)dt \]

\[ R_x = \sum_{i=0}^{n} x(i)y(i) \]

Aquired Spectrum

Cross-Correlation

In silico Spectrum

delta-Cn

From Martin Wuehr’s slide, *Xenopus* Bioinformatics Workshop 2014
ANDROMEDA/Mascot

Cox, Mann, 2010

From Martin Wuehr’s slide, Xenopus Bioinformatics Workshop 2014
Basic assumptions

1. Essentially no peptides are in common between target and decoy sequence databases

2. The likelihood of incorrectly selecting a peptide is the same for target and decoy sequence databases

Decoy peptide hits tell what type of and how many target hits are likely to be incorrect

From Martin Wuehr’s slide, Xenopus Bioinformatics Workshop 2014
Estimation of error with target-decoy searches

Target sequence database

Decoy sequence database

From Martin Wuehr’s slide, Xenopus Bioinformatics Workshop 2014
You need to make a decoy DB first

• Several different ways to generate decoys.
  – Reverse proteins
  – Shuffle proteins
  – Reverse tryptic peptides
  – Shuffle tryptic peptides

• If you aren’t really interested in their differences, just use ‘reverse proteins’.

> protA
PROTEIN

> rev_protA
NIETORP

J. Proteome Res., 2009, 8 (4):1782–1791
From Martin Wuehr’s slide, *Xenopus* Bioinformatics Workshop 2014
TMT/iTRAQ: labeling after harvesting

The tagged peptides behave exactly the same, except during fragmentation.

This spectra indicates that this protein is upregulated in patient 1 approximately 2 fold.

http://2.bp.blogspot.com/-R4HWJ2N_vZk/UIzwMIXyKoI/AAAAAAAAAnk/_lOlWaLuw84/s1600/itraq.jpg
SILAC: labeling before harvesting

- Light: \(^{12}\text{C6-Arg}\)
- Heavy: \(^{13}\text{C6-Arg}\)

1. Mix lysates 1:1
2. Digest with enzymes
3. nanoLC-MS/MS analysis

http://www.intechopen.com/source/html/43637/media/image6.png