Data Extraction and Analysis for LC-MS Based Proteomics

Instructors
Jake Jaffe\textsuperscript{2}, Deep Jaitly\textsuperscript{1}, and Matt Monroe\textsuperscript{1}

Co-Organizers
Josh Adkins\textsuperscript{1} and Dick Smith\textsuperscript{1}

\textsuperscript{1} Pacific Northwest National Laboratory, Richland, WA 99354
\textsuperscript{2} The Broad Institute, Cambridge, MA 02142
Course Outline

- Introduction
  - Bios
  - Pipelines
  - Data and Tools Availability
- Feature discovery in LC-MS datasets
  - Feature discovery in individual spectra
  - Feature definition over elution time
- Identifying LC-MS Features using an AMT tag DB
- Break
- AMT tag Pipeline Demo
- MAPQUANT, PEPPeR and GenePattern
- Panel Discussion
  - Questions
  - Future Directions
Jake Jaffe – Short Bio

- Ph.D. Biology in 2004
  - Harvard University
  - Dr. George Church and Dr. Howard Berg, advisors
- Research Scientist at The Broad Institute of MIT and Harvard
  - Proteomics Platform
- Co-developer of the Platform for Experimental Proteomic Pattern Recognition (PEPPeR)
  - Landmark Matching Algorithm
  - Proteogenomic Mapping
- Does both Laboratory and Computational Biology
- Likes coffee a lot
  - Would like to thank organizers for inviting him to Seattle!
Deep Jaitly – Short Bio

- Master’s of Mathematics in 2000
  - University of Waterloo
  - Dr. Ming Li and Dr. Paul Kearney, advisors
- Lead algorithm developer for the PNNL proteomics team
- Author or co-author of over 10 peer-reviewed articles
- >10 years experience in programming and algorithm development
  - C++, C#, & VB .NET
  - MatLab
  - Java
- 4 years industrial experience at Caprion with emphasis on algorithm development for LC-MS and LC-MS/MS data
Matt Monroe – Short Bio

- Ph.D. Analytical Chemistry in 2002
  - University of North Carolina, Chapel Hill
  - Dr. James Jorgenson, advisor
- Administers and expands SQL Server databases and associated software that supports the AMT tag pipeline
- Author or Co-author of over 30 peer-reviewed articles
- >15 years experience in programming
  - VB6 / VB .NET
  - SQL
  - LabView
- Key member of PNNL’s proteomics data analysis pipeline
Shotgun or MuDPIT Proteomics

Complex mixture of proteins → Upstream separations

Parent MS spectra → Tandem MS spectra

X!Tandem or SEQUEST w filtering & Archive

LC-MS/MS

CID

Fragmentation from N-Terminus

Fragmentation from C-Terminus
Accurate Mass and Time Tag Approach

High-throughput LC-FTICR-MS Analysis (AMT) tag

SEQUEST and/or X!Tandem Results
• Filtering
• Calculate Exact Mass
• Normalize Observed Elution Time

Complex samples

Accurate Mass and Time (AMT) Tag Data Processing Pipeline

Example Data for the AMT tag Pipeline Demo

- **Salmonella typhimurium, LC-MS/MS**
  - Grown in **LB (Luria-Bertani)** up to log phase
  - Soluble portion of cell lysis
  - “Mini-AMT tag” database, composed of 25 SCX fractions analyzed by LC-MS/MS
  - Mass and time tag database composed from searches using X!Tandem (Log E_Value ≤ -2)
  - Linear alignment of datasets for AMT tag database

- **LC-MS**
  - Different sample, grown and prepared in the same conditions
  - LC-FTICR-MS analysis (11T FTICR)
  - Non-linear alignment and peak matching to the database
PEPPeR Pipeline
Software & Data

AMT tag Pipeline Software

http://ncrr.pnl.gov/

PEPPeR, software within GenePattern

http://www.broad.mit.edu/cancer/software/genepattern/

Salmonella typhimurium data resource

http://www.proteomicsresource.org/
Funding for Tool Development

- NIH
  - National Center for Research Resources
  - National Institute of Allergy and Infectious Diseases
  - National Cancer Institute
  - National Institute of General Medical Sciences
  - National Institute of Diabetes & Digestive & Kidney Diseases
- DOE Office of Biological and Environmental Research
Other Excellent Software Resources

- http://www.ms-utils.org/ (Magnus Palmblad)
- http://open-ms.sourceforge.net/index.php (European consortium)
- http://tools.proteomecenter.org/SpecArray.php (ISB)
- http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics/Peak_Alignment/ (Tobias Kind with Oliver Fiehn)
- http://www.proteomecommons.org/tools.jsp (Phil Andrews and Jayson Falkner)
Course Outline

- Introduction
- Feature discovery in LC-MS datasets
  - Feature discovery in individual spectra
  - Feature definition over elution time
- Identifying LC-MS Features using an AMT tag DB
- Break
- AMT tag Pipeline Demo
- MAPQUANT, PEPPeR and GenePattern
- Panel Discussion
**LC-MS data**

- Complex peptide mixture on a column is separated by liquid chromatography over a period of time
- Changing composition of the mobile phase causes different peptides to elute at different times
- The components eluting from a column is sampled continuously by sequential mass spectra

Mass spectra capture the changing composition of peptides eluting from the column

QC Standards (12 protein digest)
Structure of LC-MS Data

- Each compound is observed as an isotopic pattern in a mass spectrum which is dependent on its chemical composition, charge and resolution of instrument.

Peptide: VKHPSEIVNVGDEINVK
Parent Protein: gi|16759851 30S ribosomal protein S1
Charge: 2+

m/z: 939.0203
Monoisotopic Mass: 1876.0054 Da
Structure of LC-MS Data

- A mass spectrum of a complex mixture contains overlaid distributions of several different compounds
Structure of LC-MS Data

- A mass spectrum of a complex mixture contains overlaid distributions of several different compounds.
Structure of LC-MS Data

- With LC as the first dimension, each compound is observed over multiple spectra, showing a three-dimensional pattern of m/z, elution time and abundance.

Salmonella typhimurium dataset

Peptide: VKHPSEIVNVGDEINVK
Parent Protein: gi|16759851 30S ribosomal protein S1
Charge: 2+
m/z: 939.0203
Monoisotopic Mass: 1876.0054 Da
Elution range: Scans 1539 - 1593
Feature Discovery in LC-MS data

- Goal: Infer \((mass, elution\ time,\ intensity)\) of compounds that are present in data obtained from an LC-MS dataset
  - Since their identities are unknown, the compounds are more appropriately termed features to refer to the idea that these are inferred from a three dimensional pattern

2D view of an LC-MS analysis of Salmonella typhimurium
Feature Discovery in LC-MS data

- Sequential process of finding features in each mass spectrum is followed by grouping of features over multiple spectra together.

2D views of an LC-MS dataset in different stages of processing:

- Raw data
- Deisotoping
- Collapsed monoisotopic features in all spectra
- Elution profile discovery
- LC-MS features
Feature discovery in individual spectra

- Deisotoping

  - Process of converting a mass spectrum ($m/z$, intensity) into a list of species (mass, abundance, charge)

Deisotoping a mass spectrum of 4 overlapping species

<table>
<thead>
<tr>
<th>charge</th>
<th>Monoisotopic MW</th>
<th>abundance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1546.856603</td>
<td>533467</td>
</tr>
<tr>
<td>2</td>
<td>1547.705048</td>
<td>194607</td>
</tr>
<tr>
<td>2</td>
<td>1547.887682</td>
<td>671947</td>
</tr>
<tr>
<td>2</td>
<td>1548.799612</td>
<td>426939</td>
</tr>
</tbody>
</table>
Deisotoping routine for a peak

- Algorithm to detect peptides in a complex spectrum

scan 1545

---

**Charge detection algorithm**

- charge = 2
- avg. mass = 1876.02

**Fitness value**

**Averagine**

- estimated empirical formula: C83 H124 N23 O25 S1

**Mercury**

---

Deisotoping entire spectrum – Modification of THRASH¹

Modified THRASH Routine

Algorithm to detect peptides in a complex spectrum
1. Discover all peaks in a spectrum above a specified S/N and keep in unprocessed list
2. Select most abundant peak still unprocessed
3. Compute charge for peak using charge detection algorithm
4. Compute average mass from observed m/z and predicted charge value
5. Use “Averagine” algorithm to guess empirical formula based on mass and average composition of peptides in database
6. Use Mercury algorithm to generate theoretical spectrum from the predicted empirical formula, charge of peak, and resolution of peak
7. Calculate fitness value for similarity between theoretical and observed spectrum
8. Perform “THRASHING” by overlaying theoretical and observed spectra after applying “isotopic” one dalton shift to the theoretical spectrum. Keep best fit
9. If successful fit was observed, delete isotopic peaks and associated profile of height above specified threshold of most intense ions using the theoretical pattern as template. Otherwise remove current peak only from list of unprocessed peaks.
10. While unprocessed peaks remain, repeat steps starting at step 2.
Charge Detection Algorithm

- Patterson (Autocorrelation) algorithm to detect charge of a peak in a complex spectrum

\[
P(\Delta m/z) = \sum I(m/z_i) \times I(m/z_i + \Delta m/z)
\]

Averagine Algorithm

- Algorithm to guess an average empirical formula for a given mass
- Uses average composition of all peptides in peptide database as the empirical formula for all peptides
- Protein database Averagine formula: $\text{C}_{4.9384} \text{H}_{7.7583} \text{N}_{1.3577} \text{O}_{1.4773} \text{S}_{0.0417}$, Mass = 111.1254
- Average Mass of 1877.025 would give a multiplier of $\frac{1877.025}{111.1254}$

<table>
<thead>
<tr>
<th>Element</th>
<th>Copies</th>
<th>Atomicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>4.9834*1877.025/111.1254</td>
<td>84</td>
</tr>
<tr>
<td>H</td>
<td>Remainder = 112</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1.3577*1877.025/111.1254</td>
<td>23</td>
</tr>
<tr>
<td>O</td>
<td>1.4773*1877.025/111.1254</td>
<td>25</td>
</tr>
<tr>
<td>S</td>
<td>0.0417*1877.025/111.1254</td>
<td>1</td>
</tr>
</tbody>
</table>

Empirical formula used for theoretical profile = $\text{C}_{83} \text{H}_{112} \text{N}_{23} \text{O}_{25} \text{S}_{1}$

Theoretical Isotopic Profile

- Mercury algorithm to generate a theoretical profile for a compound
  - Treat each element’s isotopic distribution as a sum of delta ($\delta$) functions

<table>
<thead>
<tr>
<th>Element</th>
<th>Isotope distribution Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbon</td>
<td>$0.98893 , \delta(m-12) + 0.01107 , \delta(m-13.00336)$</td>
</tr>
<tr>
<td>Hydrogen</td>
<td>$0.99985 , \delta(m-1.007825) + 0.00015 , \delta(m-2.014102)$</td>
</tr>
<tr>
<td>Nitrogen</td>
<td>$0.996337 , \delta(m-14.00307) + 0.003663 , \delta(m-15.00011)$</td>
</tr>
<tr>
<td>Oxygen</td>
<td>$0.99759 , \delta(m-15.99491) + 0.000374 , \delta(m-16.99913) + 0.002036 , \delta(m-17.99916)$</td>
</tr>
<tr>
<td>Sulphur</td>
<td>$0.9502 , \delta(m-31.97207) + 0.0075 , \delta(m-32.97145) + 0.0421 , \delta(m-33.96786) + 0.0002 , \delta(m-35.96708)$</td>
</tr>
</tbody>
</table>

Theoretical Isotopic Profile

- Mercury algorithm to generate a theoretical profile for a compound
  - Treat each element’s isotopic distribution as a sum of delta ($\delta$) functions
  - Convert distribution function into frequency domain: delta functions convert to simple exponential functions

<table>
<thead>
<tr>
<th>Element</th>
<th>Frequency Spectrum Function ($f_{\text{Elem}}(\mu)$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbon</td>
<td>$0.98893 , e^{12(i2\pi)\mu} + 0.01107 , e^{13.00336(i2\pi)\mu}$</td>
</tr>
<tr>
<td>Hydrogen</td>
<td>$0.99985 , e^{1.007825(i2\pi)\mu} + 0.00015 , e^{2.014102(i2\pi)\mu}$</td>
</tr>
<tr>
<td>Nitrogen</td>
<td>$0.996337 , e^{14.00307(i2\pi)\mu} + 0.003663 , e^{15.00011(i2\pi)\mu}$</td>
</tr>
<tr>
<td>Oxygen</td>
<td>$0.99759 , e^{15.99491(i2\pi)\mu} + 0.000374 , e^{16.99913(i2\pi)\mu} + 0.002036 , e^{17.99916(i2\pi)\mu}$</td>
</tr>
<tr>
<td>Sulphur</td>
<td>$0.9502 , e^{31.97207(i2\pi)\mu} + 0.0075 , e^{32.97145(i2\pi)\mu} + 0.0421 , e^{33.96786(i2\pi)\mu} + 0.0002 , e^{35.96708(i2\pi)\mu}$</td>
</tr>
</tbody>
</table>
Theoretical Isotopic Profile

- Mercury algorithm to generate a theoretical profile for a compound
  - Treat each element’s isotopic distribution as a sum of delta (δ) functions
  - Convert distribution function into frequency domain: delta functions convert to simple exponential functions.
  - Calculate the isotopic profile for a compound from the convolution of isotopic distributions of individual atoms and the imposition of a peak shape reflecting resolution of instrument
  - Compute convolution using multiplication in the frequency domain and by applying a Fourier transform

For the empirical formula $C_nH_mN_xO_yS_z$

$$F(m) = FT \left[ s(\mu) f_C(\mu)^n f_H(\mu)^m f_N(\mu)^x f_O(\mu)^y f_S(\mu)^z \right]$$

- Inverse transform of shape function
- Frequency spectra of elements
Fit Functions

- Fit functions to quantitate quality of match between theoretical and observed profiles

- Least square area\(^1\): \(\sum (t_i-o_i)^2 / \sum t_i^2\)
- Least square peak: \(\sum (T_j-O_j)^2 / \sum T_j^2\)
- Chi-square area: \(\sum (t_i-o_i)^2 / \sum t_i\)
- Chi-square peaks\(^2\): \(\sum (T_i-O_i)^2 / \sum T_i\)

Chemical Labeling with Tags

• Specify a static tag to be applied to Averagine formula
• Changes the Averagine formula generated

+TOF MS: 0.359 min from bromoadenosine.wiff Agilent

Autocorrelation charge = 1
Average mass = 345.01
Subtract average mass of tag

Autocorrelation charge = 1
Mass = 265.1065
Calculate Averagine formula

\[ 	ext{C}_5 \text{H}_{174} \text{N}_1 \text{O}_1 \text{S}_0 \]

Add tag formula to Averagine formula

\[ 	ext{C}_5 \text{H}_{174} \text{N}_1 \text{O}_1 \text{S}_0 \text{Br}_1 \]

Interesting profile because of the isotopic distribution of bromine (51% 78.91833, 49% 80.91629)
Chemical Labeling with Tags

- Specify a static tag to be applied to Averagine formula
- Changes the Averagine formula generated

+TOF MS: 0.359 min from bromoadenosine.wiff Agilent

autocorrelation charge = 1
Average mass = 345.01

Subtract average mass of tag

autocorrelation charge = 1
mass = 265.1065

Calculate Averagine formula

\[ C_5 H_{174} N_1 O_1 S_0 \]

Add tag formula to Averagine formula

\[ C_5 H_{174} N_1 O_1 S_0 Br_1 \]
16O/18O Mixtures

- Overlapping isotope patterns separated by 4 Da
  - If peaks exist 4 Da before current peak, those are processed first, and only the first four isotopic peaks are removed
Isotopic Composition

• Changing natural abundances

<table>
<thead>
<tr>
<th>Atomicity</th>
<th>Element Name</th>
<th>Element Symbol</th>
<th>Isotope Mass</th>
<th>Isotope Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hydrogen</td>
<td>H</td>
<td>1.00725</td>
<td>0.99985</td>
</tr>
<tr>
<td>1</td>
<td>Hydrogen</td>
<td>H</td>
<td>2.014012</td>
<td>0.00015</td>
</tr>
<tr>
<td>2</td>
<td>Helium</td>
<td>He</td>
<td>3.01603</td>
<td>1E-06</td>
</tr>
<tr>
<td>2</td>
<td>Helium</td>
<td>He</td>
<td>4.0025</td>
<td>0.999999</td>
</tr>
<tr>
<td>3</td>
<td>Lithium</td>
<td>Li</td>
<td>6.015121</td>
<td>0.075</td>
</tr>
<tr>
<td>3</td>
<td>Lithium</td>
<td>Li</td>
<td>7.016003</td>
<td>0.925</td>
</tr>
<tr>
<td>4</td>
<td>Berillium</td>
<td>Be</td>
<td>9.012182</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Boron</td>
<td>B</td>
<td>10.01294</td>
<td>0.199</td>
</tr>
<tr>
<td>5</td>
<td>Boron</td>
<td>B</td>
<td>11.00931</td>
<td>0.801</td>
</tr>
<tr>
<td>6</td>
<td>Carbon</td>
<td>C</td>
<td>12</td>
<td>0.9893</td>
</tr>
<tr>
<td>6</td>
<td>Carbon</td>
<td>C</td>
<td>13.00336</td>
<td>0.01107</td>
</tr>
<tr>
<td>7</td>
<td>Nitrogen</td>
<td>N</td>
<td>14.00307</td>
<td>0.996337</td>
</tr>
<tr>
<td>7</td>
<td>Nitrogen</td>
<td>N</td>
<td>15.00011</td>
<td>0.003663</td>
</tr>
<tr>
<td>8</td>
<td>Oxygen</td>
<td>O</td>
<td>15.99491</td>
<td>0.99759</td>
</tr>
</tbody>
</table>

Helpful Tips
Controls Isotopic Composition

- Loads in the isotopic composition in the form of an XML document
- Values can be edited in the grid and saved in original format or as new XML
Isotopic Composition

- Changing natural abundances
  - $^{13}$C, $^{15}$N depleted media – isotopic composition of atoms is different from those found in nature. Distribution of isotopes of Sulfur predominates the distribution shown below.

Isotopic distribution of peptide with similar mass and charge (16+), but with natural isotopic distribution of atoms.
Isotopic Composition

- Changing natural abundances
  - Changing $^{12}\text{C}/^{13}\text{C}$, $^{14}\text{N}/^{15}\text{N}$ isotopic abundances from those in nature to appropriate ones results in a better fit
  - As shown, estimated isotopic abundances were still not perfect
LC-MS Feature Discovery

Survey Profile

- Black dots indicate individual m/z values
- Green dots signify successfully deisotoped data
- Shades of red indicate data intensity

Elution Profile

- Black dots indicate individual m/z values
- Green dots signify successfully deisotoped data
- Shades of red indicate data intensity
Course Outline

- Introduction
- Feature discovery in LC-MS datasets
  - Feature discovery in individual spectra
  - Feature definition over elution time
- Identifying LC-MS Features using an AMT tag DB
- Break
- AMT tag Pipeline Demo
- MAPQUANT, PEPPeR and GenePattern
- Panel Discussion
Feature definition over elution time

- Deisotoping collapses original data into data lists

- Goal: Given series of deisotoped mass spectra, group related data across elution time
  - Look for repeated monoisotopic mass values in sequential spectra, allowing for missing data
  - Can also look for expected chromatographic peak shape

<table>
<thead>
<tr>
<th>scan num</th>
<th>charge</th>
<th>abundance</th>
<th>m/z</th>
<th>fit</th>
<th>average mw</th>
<th>monoiso mw</th>
<th>most abu. mw</th>
<th>fwhm</th>
<th>signal noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>1500</td>
<td>1</td>
<td>2772933</td>
<td>759.0649</td>
<td>0.0716</td>
<td>758.5222</td>
<td>758.0576</td>
<td>758.0576</td>
<td>0.0106</td>
<td>718.83</td>
</tr>
<tr>
<td>1500</td>
<td>1</td>
<td>2614913</td>
<td>1103.033</td>
<td>0.1111</td>
<td>1102.698</td>
<td>1102.026</td>
<td>1102.026</td>
<td>0.0222</td>
<td>74.04</td>
</tr>
<tr>
<td>1500</td>
<td>1</td>
<td>2422829</td>
<td>864.4919</td>
<td>0.0156</td>
<td>864.0073</td>
<td>863.4846</td>
<td>863.4846</td>
<td>0.0137</td>
<td>74.75</td>
</tr>
<tr>
<td>1500</td>
<td>2</td>
<td>2297822</td>
<td>563.3253</td>
<td>0.0112</td>
<td>1125.322</td>
<td>1124.636</td>
<td>1124.636</td>
<td>0.006</td>
<td>77.94</td>
</tr>
<tr>
<td>1500</td>
<td>1</td>
<td>1213607</td>
<td>943.9815</td>
<td>0.1025</td>
<td>943.5518</td>
<td>942.9742</td>
<td>942.9742</td>
<td>0.0165</td>
<td>120.36</td>
</tr>
<tr>
<td>1500</td>
<td>3</td>
<td>988761</td>
<td>675.0246</td>
<td>0.02</td>
<td>2023.375</td>
<td>2022.052</td>
<td>2023.0549</td>
<td>0.0086</td>
<td>79.22</td>
</tr>
<tr>
<td>1500</td>
<td>2</td>
<td>734070</td>
<td>688.392</td>
<td>0.0384</td>
<td>1375.694</td>
<td>1374.77</td>
<td>1374.7695</td>
<td>0.009</td>
<td>92.09</td>
</tr>
<tr>
<td>1500</td>
<td>2</td>
<td>663954</td>
<td>642.3243</td>
<td>0.0253</td>
<td>1283.417</td>
<td>1282.634</td>
<td>1282.634</td>
<td>0.0076</td>
<td>109.01</td>
</tr>
<tr>
<td>1500</td>
<td>1</td>
<td>661477</td>
<td>730.1117</td>
<td>0.024</td>
<td>729.5461</td>
<td>729.1045</td>
<td>729.1045</td>
<td>0.0096</td>
<td>39.06</td>
</tr>
<tr>
<td>1500</td>
<td>2</td>
<td>630657</td>
<td>689.3645</td>
<td>0.0446</td>
<td>1377.64</td>
<td>1376.715</td>
<td>1376.7145</td>
<td>0.0088</td>
<td>57.52</td>
</tr>
<tr>
<td>1500</td>
<td>2</td>
<td>569896</td>
<td>591.8343</td>
<td>0.0198</td>
<td>1823.79</td>
<td>1818.654</td>
<td>1818.654</td>
<td>0.0065</td>
<td>111.2</td>
</tr>
<tr>
<td>1500</td>
<td>2</td>
<td>503993</td>
<td>757.8854</td>
<td>0.0706</td>
<td>1513.762</td>
<td>1512.753</td>
<td>1512.753</td>
<td>0.0105</td>
<td>80.4</td>
</tr>
<tr>
<td>1500</td>
<td>2</td>
<td>451007</td>
<td>936.9389</td>
<td>0.0296</td>
<td>1873.091</td>
<td>1871.863</td>
<td>1872.8662</td>
<td>0.0156</td>
<td>46.74</td>
</tr>
</tbody>
</table>
Feature definition over elution time

- Can visualize deisotoped data in two-dimensions

S. typhimurium dataset on 11T FTICR
Feature definition over elution time

- **Charge state view**

  - Plotting monoisotopic mass, but color is based on charge of the original data point seen.
  - Monoisotopic Mass = $(m/z \times \text{charge}) - 1.00728 \times \text{charged}$.
Feature definition over elution time

- Zoom-in view of species
Same species in multiple spectra need to be grouped together. Related peaks found using a weighted Euclidean distance; considers:

- Mass
- Abundance
- Elution time
- Isotopic Fit
Feature definition over elution time

- Grouping uses single linkage clustering
  - Form connections between data points in n-dimensions
  - Compute the Euclidean distance between two points
    - $\text{distance} = \sqrt{\text{weight}_{\text{mass}} \times (\text{mass}_a - \text{mass}_b)^2 + \text{weight}_{\text{abu}} \times (\text{LogAbu}_a - \text{LogAbu}_b)^2 + \text{weight}_{\text{ET}} \times (\text{ET}_a - \text{ET}_b)^2 + \text{weight}_{\text{fit}} \times (\text{fit}_a - \text{Fit}_b)^2}$
  - If distance < threshold, combine points together
Feature definition over elution time

- Determine 6 separate groups
  - Typically require 2 or 3 points per group
Feature definition over elution time

**Feature detail**
- Median Mass: 1904.9399 Da (more tolerant to outliers than average)
- Elution Time: Scan 1757 (0.363 NET)
- Abundance: $1.7 \times 10^7$ counts (area under desired SIC)
  - See both 2+ and 3+ data
  - Stats typically come from the most abundant charge state (2+)

![Graph showing monoisotopic mass and selected ion chromatograms with charge states 1, 2, and 3, and an abundance scale from $0 \times 10^0$ to $2 \times 10^6$ counts. The graph includes a vertical line indicating a 5 ppm tolerance range.](image-url)
Feature definition over elution time

- Second example
  - LC-MS feature eluting over 7.5 minutes

Clustering algorithm allows for missing data, common with chromatographic tailing
Feature definition over elution time

- Second example, feature detail
  - Median Mass: 2068.1781 Da
  - Elution Time: Scan 1809 (0.380 NET)
  - Abundance: $8.7 \times 10^7$ counts (area under 3+ SIC)
    - This example has primarily 3+ data; previous had even mix of 2+ and 3+ data
Feature definition over elution time

- Refining the features
  - Require data spans at least 3 spectra
  - Exclude grouped feature if it is too long (e.g. ≥ 15% of dataset)
  - Sometimes the Euclidean distance results in undesirable clustering
    - Split if elution profile indicates two or more entities with a mass difference ≥ threshold (e.g. 4 ppm)
    - Necessary since hard to define clustering weights and distance constraints that work in all situations

Sometimes the Euclidean distance results in undesirable clustering. Split if elution profile indicates two or more entities with a mass difference ≥ threshold (e.g. 4 ppm). Necessary since hard to define clustering weights and distance constraints that work in all situations.
Feature definition over elution time

- **Example:** S. *typhimurium* dataset on 11T FTICR
  - 100 minute LC-MS analysis (3360 mass spectra)
  - 67 cm, 150 μm I.D. column with 5 μm C\textsubscript{18} particles
  - 78,641 deisotoped peaks
  - Group into 5910 LC-MS Features
Isotopic Pairs Processing

- Paired features typically have identical sequences, with and without an isotopic label
  - e.g. $^{16}\text{O}/^{18}\text{O}$ pairs or $^{14}\text{N}/^{15}\text{N}$ pairs

Data prior to finding features

Control ($^{16}\text{O}$ water)  Perturbed ($^{18}\text{O}$ water)

LC-FTICR-MS
Isotopic Pairs Processing

- Data after finding paired features
  - 4 Da pair spacing due to incorporation of two $^{18}$O atoms

Control ($^{16}$O water)  Perturbed ($^{18}$O water)

LC-FTICR-MS
Paired feature example: $^{16}\text{O}/^{18}\text{O}$ data

- Compute AR using ratio of areas, or
- Compute AR scan-by-scan, then average AR values (members must co-elute)
Isotopic Pairs Processing

- Paired feature example: $^{14}\text{N}/^{15}\text{N}$ data
  - Pair members often do not co-elute
  - Use bulk area ratio, or re-align pair members then compute AR scan-by-scan

30.9 Da, corresponding to 31 N atoms
Matching AMT: GILSGEFDHIEQAFYMVGSIDEAVEK
Empirical formula: $\text{C}_{134}\text{H}_{201}\text{N}_{31}\text{O}_{44}\text{S}$

$\text{AR} = 1.17 \ (\text{Light}_{\text{Area}} \div \text{Heavy}_{\text{area}})$
Feature definition over elution time

- Numerous options for clustering data to form LC-MS features and for finding paired features
Course Outline

- Introduction
- Feature discovery in LC-MS datasets
  - Feature discovery in individual spectra
  - Feature definition over elution time
- Identifying LC-MS Features using an AMT tag DB
- Break
- AMT tag Pipeline Demo
- MAPQUANT, PEPPeR and GenePattern
- Panel Discussion
Assembling an AMT tag DB

- **Accurate Mass and Time (AMT) tag**
  - Unique peptide sequence whose monoisotopic mass and normalized elution time are accurately known
  - AMT tags also track any modified residues in peptide

- **AMT tag DB**
  - Collection of AMT tags

- **AMT tag Approach articles**
Assembling an AMT tag DB

- What can we use an AMT tag DB for?
  - Query LC-MS/MS data to answer questions
    - How many distinct peptides were observed passing filter criteria?
    - Which peptides were observed most often by LC-MS/MS?
    - How many proteins had 2 or more partially or fully tryptic peptides?
  - Correlate LC-MS features to the AMT tags
    - Analyze multiple, related samples by LC-MS using a high mass accuracy mass spectrometer
      - e.g. Time course study, 5 data points with 3 points per sample
    - Characterize the LC-MS features
      - Deisotope to obtain monoisotopic mass and charge
      - Cluster in time dimension to obtain abundance information
    - Match to AMT tags to identify peptides
      - Align in mass and time dimensions
      - Match mass and time of LC-MS features to mass and time of AMT tags
Assembling an AMT tag DB

- Characterizing AMT tags
  - Analyze samples by LC-MS/MS
    - 10 minute to 180 minute LC separations
    - Obtain 1000's of MS/MS fragmentation spectra for each sample
  - Analyze spectra using SEQUEST, X!Tandem, etc.
    - SEQUEST: http://www.thermo.com/bioworks/

- Collate results

List of peptide and protein matches
Assembling an AMT tag DB

- AMT tag example
  - R.VKHPSEIVNVGDEINVK.V
  - Observed in scan 11195 of dataset #19 in an SCX fractionation series

3+ species
Match 30 b/y ions
X!Tandem hyperscore = 80
X!Tandem Log(E_Value) = -5.9
Assembling an AMT tag DB

- AMT tag example
  - R.VKHPSEIVNVGDEINVK.V
  - Observed in scan 11195 of dataset #19 in an SCX fractionation series

<table>
<thead>
<tr>
<th>#</th>
<th>Immon.</th>
<th>b</th>
<th>b++</th>
<th>Seq.</th>
<th>y</th>
<th>y++</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>72.08</td>
<td></td>
<td></td>
<td>V</td>
<td>1877.01</td>
<td>939.01</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>101.11</td>
<td>228.17</td>
<td></td>
<td>K</td>
<td>1777.94</td>
<td>889.48</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>110.07</td>
<td>365.23</td>
<td></td>
<td>H</td>
<td>1649.85</td>
<td>825.43</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>70.07</td>
<td>462.28</td>
<td></td>
<td>P</td>
<td>1512.79</td>
<td>756.90</td>
<td>14</td>
</tr>
<tr>
<td>5</td>
<td>60.04</td>
<td>549.31</td>
<td></td>
<td>S</td>
<td>1415.74</td>
<td>708.37</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td>102.06</td>
<td>678.36</td>
<td>339.68</td>
<td>E</td>
<td>1328.71</td>
<td>664.86</td>
<td>12</td>
</tr>
<tr>
<td>7</td>
<td>86.10</td>
<td>791.44</td>
<td>396.22</td>
<td>I</td>
<td>1199.66</td>
<td>600.33</td>
<td>11</td>
</tr>
<tr>
<td>8</td>
<td>72.08</td>
<td>890.51</td>
<td>445.76</td>
<td>V</td>
<td>1086.58</td>
<td>543.79</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>87.06</td>
<td>1004.55</td>
<td>502.78</td>
<td>N</td>
<td>987.51</td>
<td>494.26</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>72.08</td>
<td>1103.62</td>
<td>552.31</td>
<td>V</td>
<td>873.47</td>
<td>437.24</td>
<td>8</td>
</tr>
<tr>
<td>11</td>
<td>30.03</td>
<td>1160.64</td>
<td>580.93</td>
<td>G</td>
<td>774.40</td>
<td>387.70</td>
<td>7</td>
</tr>
<tr>
<td>12</td>
<td>88.04</td>
<td>1275.67</td>
<td>638.34</td>
<td>D</td>
<td>717.38</td>
<td>359.19</td>
<td>6</td>
</tr>
<tr>
<td>13</td>
<td>102.06</td>
<td>1404.71</td>
<td>702.86</td>
<td>E</td>
<td>602.35</td>
<td>301.68</td>
<td>5</td>
</tr>
<tr>
<td>14</td>
<td>86.10</td>
<td>1517.80</td>
<td>759.40</td>
<td>I</td>
<td>473.31</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>87.06</td>
<td>1631.84</td>
<td>816.42</td>
<td>N</td>
<td>360.22</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>72.08</td>
<td>1730.91</td>
<td>865.96</td>
<td>V</td>
<td>246.18</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>101.11</td>
<td></td>
<td></td>
<td>K</td>
<td>147.11</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

3+ species
Match 30 b/y ions
X!Tandem hyperscore = 80
X!Tandem Log(E_Value) = -5.9
Assembling an AMT tag DB

- Align related datasets using elution times of observed peptides
  - One option: utilize NET prediction algorithm to create theoretical dataset to align against
    - NET Prediction uses position and ordering of amino acid residues to predict normalized elution time

<table>
<thead>
<tr>
<th>Peptide</th>
<th>X!Tandem Log (E_Value)</th>
<th>Elution Time</th>
<th>Predicted NET</th>
</tr>
</thead>
<tbody>
<tr>
<td>R.AARPAKSYVDENGETK.T</td>
<td>-6.1</td>
<td>33.958</td>
<td>0.167</td>
</tr>
<tr>
<td>R.LVHGEGLVAAKR.I</td>
<td>-8.8</td>
<td>36.915</td>
<td>0.224</td>
</tr>
<tr>
<td>R.GIIKVGEEEVEIVGIK.E</td>
<td>-8.2</td>
<td>53.003</td>
<td>0.415</td>
</tr>
<tr>
<td>K.RFNDGDPIFIHTGGAPALFAYPHV.-</td>
<td>-7.3</td>
<td>62.583</td>
<td>0.519</td>
</tr>
<tr>
<td>K.KTVGVAQVQELAKGLGDLV.F</td>
<td>-11.6</td>
<td>62.803</td>
<td>0.438</td>
</tr>
<tr>
<td>R.KVAAQIPNGSTLFIDIGTTPEAVALIDGHSNLR.I</td>
<td>-8.9</td>
<td>73.961</td>
<td>0.589</td>
</tr>
<tr>
<td>R.TFAISPGHMNQLRAESIPEAVIAGASALVLTSLVLR.C</td>
<td>-6.5</td>
<td>88.043</td>
<td>0.764</td>
</tr>
</tbody>
</table>

Assembling an AMT tag DB

- Align related datasets using elution times of observed peptides
  - One option: utilize NET prediction algorithm to create theoretical dataset to align against
    - NET Prediction uses position and ordering of amino acid residues to predict normalized elution time
    - Alignment yields NET values based on observed elution times
      - Observed NET = Slope × (Observed Elution Time) + Intercept

Example: 506 unique peptides used for alignment; Log(E_Value) ≤ -6

\[
y = 0.01081x - 0.1829 \\
R^2 = 0.95
\]

VKHPSEIVNVGDEINVK
Elution time: 44.923 minutes
Predicted NET: 0.292
Observed NET: 0.303
Assembling an AMT tag DB

- AMT tag example
  - R.VKHPSEIVNVGDEINVK.V
  - Observed in 7 (of 25) LC-MS/MS datasets in the SCX fractionation series

Analysis 1, scan 11195 → 3+, hyperscore 80, Obs. NET 0.303
Analysis 2, scan 9945 → 3+, hyperscore 69, Obs. NET 0.298
Analysis 3, scan 10905 → 2+, hyperscore 74, Obs. NET 0.301
Analysis 4, scan 9667 → 2+, hyperscore 77, Obs. NET 0.302

Compute monoisotopic mass: 1876.0053 Da
Average Normalized Elution Time: 0.3021 (StDev 0.0021)
Assembling an AMT tag DB

- **Mass and Time Tag Database**
  - Repository for AMT tags
    - Mass, elution time, modified residues, and supporting information for each AMT tag
  - Allows samples of unknown composition to be matched quickly and efficiently, without needing to perform tandem MS
  - Assembled by analyzing a control set of samples, cataloging each peptide identification until subsequent analyses no longer provide new identifications

<table>
<thead>
<tr>
<th>MT Tag ID</th>
<th>Peptide</th>
<th>LC-MS/MS Obs. Count</th>
<th>Calculated Monoisotopic Mass</th>
<th>Average Observed NET</th>
<th>Observed NET StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>1662039</td>
<td>MTGRELKPHDR</td>
<td>1</td>
<td>1338.6826</td>
<td>0.143</td>
<td>0.000</td>
</tr>
<tr>
<td>17683899</td>
<td>SSALNTLTNQK</td>
<td>3</td>
<td>1175.6146</td>
<td>0.235</td>
<td>0.005</td>
</tr>
<tr>
<td>36609588</td>
<td>HRDLLLGATNP…TLR</td>
<td>5</td>
<td>1960.0602</td>
<td>0.379</td>
<td>0.002</td>
</tr>
<tr>
<td>36715875</td>
<td>WVKVDGWDN…FER</td>
<td>11</td>
<td>2590.2815</td>
<td>0.459</td>
<td>0.011</td>
</tr>
<tr>
<td>36843675</td>
<td>MYGHLKGEVA…QER</td>
<td>8</td>
<td>2533.2304</td>
<td>0.557</td>
<td>0.005</td>
</tr>
</tbody>
</table>
Assembling an AMT tag DB

- **Mini AMT tag DB**
  - Database constructed from a relatively small number of datasets
  - e.g. 25 SCX fractionation samples from *S. typhimurium*, each analyzed by LC-MS/MS and then by X!Tandem
  - Protein database: `S_typhimurium_LT2_2004-09-19`
    - 4550 proteins and 1.4 million residues

> STM1834 putative YebN family transport protein (yebN) \{Salmonella typhimurium LT2\}
MFAGGSDFVNGYPGQDVVMHFTATVLLAFGMSMDAFAASIGKGAHLHKPKFSEALRTGLI
FGAVETLTPLIGWGLGILASKFVLEWNHWIAFVLLIFLGGRRMIIGEGRGSSEDDEETPLRR
HSFWLLVTATTATSDAMAVGVGLAFLOQVNNIATALAIIGCATLIMSTLGMMGIRFIGPML
GKRAEILGGVVLIGGVQILWTHFHG

> STM1835 23S rRNA m1G745 methyltransferase (rrmA) \{Salmonella typhimurium LT2\}
MSFTCPLCHQPLTQINNSVICPQRHQFDVAKEGYINLLPVQHKRSRDPGDSAEIMMQARRA
FLDAGHYQPLRDAVINLLRERLDQSATAILDGCGEGGYTHAFAEALPGVTTFGLDVAKT
AIKAAAKRYSQVFVASSHRPLPFADASMDAVERIYAPCKAQELARVVKPGWVVTATPG
PHHLMELKGLVDEVRLHAPYTEQLDGFTLQQOSTRLAYHMQLTAEAAVALLQMPFAWRA
RPDVWEQLAASAGLSCQTDFNLHLWQRNR
Assembling an AMT tag DB

- Database Relationships
  - Minimum information required:
    - Single table with Mass and NET

- Expanded schema:

<table>
<thead>
<tr>
<th>T_Mass_Tags</th>
<th>PK</th>
<th>Mass_Tag_ID</th>
<th>Peptide</th>
<th>Monoisotopic_Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>PK,FK1</td>
<td>Mass_Tag_ID</td>
<td>PK,FK2 Ref_ID</td>
<td>Avg_GANET</td>
<td>Cnt_GANET</td>
</tr>
</tbody>
</table>

PK := Primary Key
FK := Foreign Key
Assembling an AMT tag DB

- Microsoft Access DB Relationships
- Full schema to track individual peptide observations

<table>
<thead>
<tr>
<th>T_Analysis_Description</th>
<th>T_Peptides</th>
<th>T_Mass_Tags</th>
<th>T_Mass_Tags_to_Protein_Map</th>
</tr>
</thead>
<tbody>
<tr>
<td>PK, Job</td>
<td>PK, Peptide_ID</td>
<td>PK, Mass_Tag_ID</td>
<td>PK, FK1, FK2, Ref_ID</td>
</tr>
<tr>
<td></td>
<td>PK, Analysis_ID</td>
<td>PK, Peptide_ID</td>
<td>Mass_Tag_ID</td>
</tr>
<tr>
<td></td>
<td>PK, Scan_Number</td>
<td>PK, Mass_Tag_ID</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PK, Number_Of_Scans</td>
<td>PK, Mass_Tag_ID</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PK, Charge_State</td>
<td>PK, Mass_Tag_ID</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PK, MH</td>
<td>PK, Mass_Tag_ID</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PK, Multiple_Proteins</td>
<td>PK, Mass_Tag_ID</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PK, Mass_Tag_ID</td>
<td>PK, Mass_Tag_ID</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PK, GANET_Obs</td>
<td>PK, Mass_Tag_ID</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PK, Scan_Time_Peak_Apex</td>
<td>PK, Mass_Tag_ID</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PK, Peak_Area</td>
<td>PK, Mass_Tag_ID</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PK, Peak_SN_Ratio</td>
<td>PK, Mass_Tag_ID</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PK, PK1</td>
<td>PK, Peptide_ID</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PK, FK1</td>
<td>PK, Peptide_ID</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PK, FK2</td>
<td>PK, Peptide_ID</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>T_Peptides</th>
<th>T_Mass_Tags</th>
<th>T_Mass_Tags_NET</th>
<th>T_Proteins</th>
<th>V_Filter_Set_Overview_Ext</th>
</tr>
</thead>
<tbody>
<tr>
<td>PK, Peptide_ID</td>
<td>PK, Mass_Tag_ID</td>
<td>PK, Mass_Tag_ID</td>
<td>PK, Ref_ID</td>
<td>Filter_Type</td>
</tr>
<tr>
<td>PK, Mass_Tag_ID</td>
<td></td>
<td>PK, Mass_Tag_ID</td>
<td>Reference</td>
<td>Filter_Set_Name</td>
</tr>
<tr>
<td>PK, Mass_Tag_ID</td>
<td></td>
<td>PK, Mass_Tag_ID</td>
<td>Description</td>
<td>Filter_Set_Description</td>
</tr>
<tr>
<td>PK, Mass_Tag_ID</td>
<td></td>
<td>PK, Mass_Tag_ID</td>
<td>Protein_Sequence</td>
<td>Extra_Info</td>
</tr>
<tr>
<td>PK, Mass_Tag_ID</td>
<td></td>
<td>PK, Mass_Tag_ID</td>
<td>Protein_Residue_Count</td>
<td>Filter_Set_ID</td>
</tr>
<tr>
<td>PK, Mass_Tag_ID</td>
<td></td>
<td>PK, Mass_Tag_ID</td>
<td>Monoisotopic_Mass</td>
<td>Filter_Name</td>
</tr>
<tr>
<td>PK, Mass_Tag_ID</td>
<td></td>
<td>PK, Mass_Tag_ID</td>
<td>Protein_Collection_ID</td>
<td>Filter_Type</td>
</tr>
<tr>
<td>PK, Mass_Tag_ID</td>
<td></td>
<td>PK, Mass_Tag_ID</td>
<td>Last_Affected</td>
<td>Filter_Provider</td>
</tr>
</tbody>
</table>

- PK = Primary Key
- FK = Foreign Key

[Schema Diagram with Tables and Relationships]
Assembling an AMT tag DB

- Example data

<table>
<thead>
<tr>
<th>Mass_Tag_ID</th>
<th>Peptide</th>
<th>Monoisotopic_Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>24847</td>
<td>VKHPSEIVNVGDEINVK</td>
<td>1876.00533</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mass_Tag_ID</th>
<th>Avg_GANET</th>
<th>Cnt_GANET</th>
<th>Std_GANET</th>
</tr>
</thead>
<tbody>
<tr>
<td>24847</td>
<td>0.3021</td>
<td>7</td>
<td>2.11E-03</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Peptide_ID</th>
<th>Peptide</th>
<th>Mass_Tag_ID</th>
<th>Job</th>
<th>Scan Number</th>
<th>Charge State</th>
</tr>
</thead>
<tbody>
<tr>
<td>53428</td>
<td>R.VKHPSEIVNVGDEINVK.V</td>
<td>24847</td>
<td>206386</td>
<td>11195</td>
<td>3</td>
</tr>
<tr>
<td>57461</td>
<td>R.VKHPSEIVNVGDEINVK.V</td>
<td>24847</td>
<td>206387</td>
<td>9945</td>
<td>3</td>
</tr>
<tr>
<td>61511</td>
<td>R.VKHPSEIVNVGDEINVK.V</td>
<td>24847</td>
<td>206388</td>
<td>10905</td>
<td>2</td>
</tr>
<tr>
<td>65386</td>
<td>R.VKHPSEIVNVGDEINVK.V</td>
<td>24847</td>
<td>206389</td>
<td>9667</td>
<td>2</td>
</tr>
<tr>
<td>69081</td>
<td>R.VKHPSEIVNVGDEINVK.V</td>
<td>24847</td>
<td>206390</td>
<td>9118</td>
<td>2</td>
</tr>
<tr>
<td>72556</td>
<td>R.VKHPSEIVNVGDEINVK.V</td>
<td>24847</td>
<td>206391</td>
<td>9159</td>
<td>2</td>
</tr>
<tr>
<td>76263</td>
<td>R.VKHPSEIVNVGDEINVK.V</td>
<td>24847</td>
<td>206392</td>
<td>9421</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Peptide_ID</th>
<th>Hyperscore</th>
<th>Log(E_Value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>53428</td>
<td>80.2</td>
<td>-5.89</td>
</tr>
<tr>
<td>57461</td>
<td>69.2</td>
<td>-4.92</td>
</tr>
<tr>
<td>61511</td>
<td>74</td>
<td>-12.85</td>
</tr>
<tr>
<td>65386</td>
<td>77.2</td>
<td>-12.80</td>
</tr>
<tr>
<td>69081</td>
<td>69</td>
<td>-12.82</td>
</tr>
<tr>
<td>72556</td>
<td>78</td>
<td>-13.77</td>
</tr>
<tr>
<td>76263</td>
<td>60.3</td>
<td>-11.27</td>
</tr>
</tbody>
</table>
# Assembling an AMT tag DB

## Processing steps

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
</table>
| 1.   | Thermo-Finnigan LTQ .Raw files  
| 2.   | MS/MS spectra files  
       | Process _Dta.txt files with X!Tandem (round 1 partially tryptic; round 2 dynamic oxidized methionine) |
| 3.   | X!Tandem Results  
       | Convert X!Tandem .XML files to tab-delimited files using the Peptide Hit Results Processor application |
| 4.   | Tab delimited text files  
       | Align datasets using MTDB Creator application |
| 5.   | Summarized result files  
       | Load into database |
| 6.   | Microsoft Access DB |

---

**Diagram:**

- **Thermo-Finnigan LTQ .Raw files**
  - Convert to .Dta using Extract_MSn.exe
  - Concatenate .Dta files into _Dta.txt file using Perl script
  - Improved application (under development): Decon_MSn.exe

- **MS/MS spectra files**
  - Process _Dta.txt files with X!Tandem (round 1 partially tryptic; round 2 dynamic oxidized methionine)

- **X!Tandem Results**
  - Convert X!Tandem .XML files to tab-delimited files using the Peptide Hit Results Processor application

- **Tab delimited text files**
  - Align datasets using MTDB Creator application

- **Summarized result files**
  - Load into database

---

**Notes:**

- Thermo-Finnigan LTQ .Raw files are converted to .Dta using Extract_MSn.exe. The .Dta files are then concatenated into _Dta.txt using Perl script. An improved application (Decon_MSn.exe) is under development.

- MS/MS spectra files are processed with X!Tandem (round 1 partially tryptic; round 2 dynamic oxidized methionine).

- The X!Tandem .XML files are converted to tab-delimited files using the Peptide Hit Results Processor application.

- Tab delimited text files are aligned using MTDB Creator application.

- The Summarized result files are loaded into a Microsoft Access DB.
Assembling an AMT tag DB

- PHRP Relationships

<table>
<thead>
<tr>
<th>Results_Info</th>
<th>Result_To_Seq_Map</th>
<th>Seq_Info</th>
<th>Mod_Details</th>
<th>Seq_to_Protein_Map</th>
</tr>
</thead>
<tbody>
<tr>
<td>PK</td>
<td>Result_ID</td>
<td>PK,FK1</td>
<td>Unique_Seq_ID</td>
<td>PK,FK1</td>
</tr>
<tr>
<td>FK1</td>
<td>Unique_Seq_ID</td>
<td>PK,FK1</td>
<td>Result_ID</td>
<td>PK,FK2</td>
</tr>
<tr>
<td></td>
<td>Group_ID</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Scan</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Charge</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Peptide_MH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Peptide_Hyperscore</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Peptide_Expectation_Value_Log(e)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Multiple_Protein_Count</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Peptide_Sequence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DeltaCn2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>y_score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>y_ions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>b_score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>b_ions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Delta_Mass</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Peptide_Intensity_Log(I)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Assembling an AMT tag DB

- Database histograms – filtered on $\log(E\_Value) \leq -2$
AMT Tag DB Growth Trend

- Trend for Mini AMT Tag DB
  - 25 SCX fractionation datasets of a single growth condition

- Trend for Mature AMT Tag DB
  - 521 different samples from ~25 different growth conditions
  - Slope of curve decreases as more datasets added and fewer new peptides are seen

Filtered on Log(E_Value) \leq -2

Filtered on Peptide Prophet Probability $\geq 0.99$
Identifying LC-MS Features

- VIPER software
  - Visualize and find features in LC-MS data
  - Match features to peptides (AMT tags)
  - Graphical User Interface and automated analysis mode
Identifying LC-MS Features

- Peak Matching Steps
  - Load LC-MS peak lists from Decon2LS
  - Filter data
  - Feature definition over elution time
    - Select AMT tags to match against
    - Optionally, find paired features (e.g. $^{16}\text{O}/^{18}\text{O}$ pairs)
    - Align LC-MS features to AMT tags using LCMSWarp
    - Broad AMT tag DB search
    - Search tolerance refinement
    - Final AMT tag DB search
  - Report results
Identifying LC-MS Features

- AMT Tag database selection

Connect to mass tag system (MTS) if inside PNNL or use Standalone Microsoft Access DB
Alignment using LCMSWarp

- Align scan number (i.e. elution time) of features to NETs of peptides in given AMT tag database
  - Match mass and NET of AMT tags to mass and scan number of MS features
  - Use LCMSWarp algorithm to find optimal alignment to give the most matches

**Calculated monoisotopic mass**

**Deisotoped monoisotopic mass**

**AMTs**

**LC-MS Features**

**Average observed NET**

**Observed scan number**
Alignment using LCMSWarp

- LCMSWarp computes a similarity score from conserved local mass and retention time patterns

N. Jaitly, M.E. Monroe et. al., *Analytical Chemistry* 2006, 78, 7397-7409.
Alignment using LCMSWarp

- Similarity scores between LC-MS features and AMT tags are used to generate a score graph of similarity.
- Best alignment is found using a dynamic programming algorithm that determines the transformation function with maximum likelihood.

Heatmap of similarity score between LC-MS features and AMT tags (z-score representation)

Alignment using LCMSWarp

- Transformation function is used to convert from scan number to NET
  - Features centered at same scan number get the same obs. NET value
  - When matching LC-MS Features to AMTs, we will search +/- a NET tolerance, which effectively allows for LC-MS features to shift around a little in elution time

<table>
<thead>
<tr>
<th>LC-MS Feature Scan</th>
<th>Matching AMT tag NET</th>
<th>LC-MS Feature NET</th>
</tr>
</thead>
<tbody>
<tr>
<td>1011</td>
<td>0.1519</td>
<td>0.1569</td>
</tr>
<tr>
<td>1019</td>
<td>0.1626</td>
<td>0.1589</td>
</tr>
<tr>
<td>1019</td>
<td>0.1507</td>
<td>0.1589</td>
</tr>
<tr>
<td>1021</td>
<td>0.1653</td>
<td>0.1594</td>
</tr>
<tr>
<td>1027</td>
<td>0.1509</td>
<td>0.1609</td>
</tr>
<tr>
<td>1037</td>
<td>0.1519</td>
<td>0.1633</td>
</tr>
<tr>
<td>1042</td>
<td>0.183</td>
<td>0.1645</td>
</tr>
<tr>
<td>1055</td>
<td>0.1652</td>
<td>0.1677</td>
</tr>
<tr>
<td>1056</td>
<td>0.1862</td>
<td>0.1679</td>
</tr>
<tr>
<td>1056</td>
<td>0.1697</td>
<td>0.1679</td>
</tr>
<tr>
<td>1056</td>
<td>0.1682</td>
<td>0.1679</td>
</tr>
</tbody>
</table>
Alignment using LCMSWarp

- **NET Residual Plots**
  - Difference between NET of LC-MS feature and NET of matching AMT tag
  - Indicates quality of alignment between features and AMT tags
  - This data shows nearly linear alignment between features and AMTs, but the algorithm can easily account for non-linear trends

NET Residuals if a linear mapping is used

NET Residuals after LCMSWarp
Alignment using LCMSWarp

- Non-linear alignment example #1
  - Identical LC separation system, but having column flow irregularities

NET Residuals if a linear mapping is used

NET Residuals after LCMSWarp

S. typhimurium on 9T
Alignment using LCMSWarp

- Non-linear alignment example #2
  - PMT Tag DB from C18 LC-MS/MS analyses using ISCO-based LC (exponential dilution gradient)
  - LC-MS analysis used C18 LC-MS via Agilent linear gradient pump

S. oneidensis on LTQ-Orbitrap

NET Residuals after LCMSWarp

NET Residuals if a linear mapping is used
Alignment using LCMSWarp

- Non-linear alignment example #3
  - PMT Tag DB from C₁₈ LC-MS/MS analyses using ISCO-based LC
  - LC-MS analysis used C₁₈ LC-MS via Agilent linear gradient pump

QC Standards (12 protein digest) on LTQ-Orbitrap

NET Residuals after LCMSWarp

NET Residuals if a linear mapping is used
Alignment using LCMSWarp

- LCMSWarp Features
  - Fast and robust
    - Previous method used least-squares regression, iterating through a large range of guesses (slow and often gave poor alignment)
  - Requires that a reasonable number of LC-MS features match the AMT Tag DB

S. typhimurium on 11T match against 18,617 S. typhimurium PMTs

S. typhimurium on 11T match against 65,193 S. oneidensis PMTs
In addition to aligning data in time, we can also recalibrate the masses of the LC-MS features because mass and time values are available for both LC-MS features and AMT tags.

Two options for mass re-calibration:
- Bulk linear correction
- Piece-wise correction via LCMSWarp

Visualize mass differences using mass error histogram or mass residual plot.
**Mass Error Histogram**

- List of binned mass error values
  - Difference between feature's mass and matching AMT tag's mass
  - Bin values to generate a histogram
  - Typically observe background false positive level

<table>
<thead>
<tr>
<th>LC-MS Feature Mass (Da)</th>
<th>AMT Tag Mass (Da)</th>
<th>Delta Mass (Da)</th>
<th>Mass Error (ppm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1570.9005</td>
<td>1570.883</td>
<td>0.01745</td>
<td>11.1</td>
</tr>
<tr>
<td>1571.74325</td>
<td>1571.726</td>
<td>0.01770</td>
<td>11.3</td>
</tr>
<tr>
<td>1571.8498</td>
<td>1571.831</td>
<td>0.01912</td>
<td>12.2</td>
</tr>
<tr>
<td>1571.9107</td>
<td>1571.892</td>
<td>0.01848</td>
<td>11.8</td>
</tr>
<tr>
<td>1573.8381</td>
<td>1573.832</td>
<td>0.00569</td>
<td>3.6</td>
</tr>
</tbody>
</table>

Match Tolerances:
- Mass: ±25 ppm
- NET: ±0.05 NET

**Likely true positive identifications**

**Likely false positive identifications**
**Mass Calibration**

- **Option 1: Bulk linear correction**
  - Use location of peak in mass error histogram to adjust masses of all features
  - Shift by ppm mass; absolute shift amount increases as monoisotopic mass increases

\[
\Delta_{\text{mass}} = -11.6 \text{ ppm} \times \frac{\text{mass}_{\text{old}}}{1 \times 10^6 \text{ ppm/Da}}
\]

For 1+ feature at 1570.9005 Da,
\[
\Delta_{\text{mass}} = -0.0182 \text{ Da}
\]

For 3+ feature at 2919.4658 Da,
\[
\Delta_{\text{mass}} = -0.0339 \text{ Da}
\]

---

**Peak Center of mass:** 11.6 ppm
**Peak Width:** 2 ppm at 60% of max
**Peak Height:** 404 counts/bin
**Noise level:** 19 counts/bin
Mass Calibration

- Option 2: Piece-wise correction via LCMSWarp
  - Examine sections of the data to determine a custom mass shift for each section
  - One option is to divide into time sections

S. typhimurium on 11T
Mass Calibration

- **Option 2: Piece-wise correction via LCMSWarp**
  - Second option is to divide into m/z sections
  - LCMSWarp utilizes a hybrid correction based on both mass error vs. time and mass error vs. m/z

![Mass Calibration Diagram]

- Mass Error (ppm) vs. m/z
- Mass Error (ppm) vs. m/z after correction

*S. typhimurium* on 11T
Mass Calibration

- Comparison of the three methods
  - Mass error histogram gets taller, narrower, and more symmetric
  - Linear $\rightarrow$ Mass error vs. m/z $\rightarrow$ Mass error vs. time $\rightarrow$ Hybrid
  - Not all datasets show the same trends, but Hybrid mass recalibration is generally superior

![Mass Error Histograms for S. typhimurium and S. oneidensis]
Identifying LC-MS Features

- Peak Matching Steps
  - Load LC-MS peak lists from Decon2LS
  - Filter data
  - Feature definition over elution time
  - Select AMT tags to match against
  - Optionally, find paired features (e.g. $^{16}\text{O}/^{18}\text{O}$ pairs)
  - Align LC-MS features to AMT tags using LCMSWarp

- Broad AMT tag DB search
- Search tolerance refinement
- Final AMT tag DB search
- Report results
Identifying LC-MS Features

- Match Features to LC-MS/MS IDs
- *S. typhimurium* DB, from 25 LC-MS/MS analyses
  - 18,617 AMT tags, all fully or partially tryptic
  - Look for AMT tags within a broad mass range, e.g., ±25 ppm and ±0.05 NET of each feature

**S. typhimurium** AMT Tag Database

- 18,617 AMT tags

**S. typhimurium** on 11T FTICR

- 5,934 features
- 4,678 features have match, matching 6,242 AMT tags

Average observed NET

Observed NET
Search tolerance refinement

- Can use mass error and NET error histograms to determine optimal search tolerances

Examine distribution of errors to determine optimal tolerance using expectation maximization algorithm

±1.76 ppm
Identifying LC-MS Features

- Repeat search with final search tolerances
  - 5,934 features
  - 3,866 features with matches
  - 3,958 out of 18,617 AMT tags matched using ±1.76 ppm

Match Tolerances
- Mass: ±1.76 ppm
- NET: ±0.0203 NET

Observed NET vs. Monoisotopic Mass
Identifying LC-MS Features

- Given feature can match more than one AMT tag
- Need measure of ambiguity

### Match Tolerances
- Mass: ±4 ppm
- NET: ±0.02 NET

<table>
<thead>
<tr>
<th>AMT Tag ID</th>
<th>Peptide</th>
<th>Mass (Da)</th>
<th>NET</th>
</tr>
</thead>
<tbody>
<tr>
<td>35896216</td>
<td>T.RALMQLDEALRPSLR.S</td>
<td>1767.9777</td>
<td>0.373</td>
</tr>
<tr>
<td>105490</td>
<td>K.DLETIVGLQTDAPLKR.A</td>
<td>1767.9730</td>
<td>0.380</td>
</tr>
<tr>
<td>36259992</td>
<td>R.SIGIAPDVILICRGDRAI.P</td>
<td>1767.9664</td>
<td>0.392</td>
</tr>
</tbody>
</table>

**Δ mass**
- Δ mass = 2.8 ppm
- Δ NET = -0.010
- Δ mass = 0.17 ppm
- Δ NET = -0.003
- Δ mass = -3.5 ppm
- Δ NET = 0.009

**Monoisotopic Mass**
- 1767.9727 Da
- NET: 0.383
Identifying LC-MS Features

\[ d_{ij}^2 = \left( \frac{m_i - \mu_{mj}}{\sigma_{mj}^2} \right)^2 + \left( \frac{t_i - \mu_{tj}}{\sigma_{tj}^2} \right)^2 \]

\[ \sigma_{mj} = 4 \text{ ppm}, \sigma_{tj} = 0.025 \]

\[ p_{ij} = \frac{(\sigma_{mj} \sigma_{tj})^{-1} \exp(-d_{ij}^2 / 2)}{\left( \sum_{k=1}^{N} (\sigma_{mk} \sigma_{tk})^{-1} \exp(-d_{ik}^2 / 2) \right)} \]

**Match Tolerances**

<table>
<thead>
<tr>
<th>AMT Tag ID</th>
<th>Mass (Da)</th>
<th>NET</th>
<th>( d_{ij}^2 )</th>
<th>Numerator</th>
<th>( p_{ij} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>35896216</td>
<td>1767.9777</td>
<td>0.373</td>
<td>3.012</td>
<td>6273.3</td>
<td>0.16</td>
</tr>
<tr>
<td>105490</td>
<td>1767.9730</td>
<td>0.380</td>
<td>0.090</td>
<td>27042.5</td>
<td>0.70</td>
</tr>
<tr>
<td>36259992</td>
<td>1767.9664</td>
<td>0.392</td>
<td>3.267</td>
<td>5521.4</td>
<td>0.14</td>
</tr>
</tbody>
</table>

**Sum:** 38837.2

Identifying LC-MS Features

- **SLiC**: Spatially Localized Confidence Score
  - Measures uniqueness of match

<table>
<thead>
<tr>
<th>AMT Tag ID</th>
<th>Peptide</th>
<th>Mass (Da)</th>
<th>NET</th>
<th>SLiC Score</th>
<th>Average XCorr</th>
<th>Avg Disc Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>35896216</td>
<td>T.RALMQLDEALRPSLR.S</td>
<td>1767.9777</td>
<td>0.373</td>
<td>0.16</td>
<td>3.13</td>
<td>0.61</td>
</tr>
<tr>
<td>105490</td>
<td>K.DLETIVGLQTDAPLKR.A</td>
<td>1767.9730</td>
<td>0.380</td>
<td>0.70</td>
<td>3.68</td>
<td>0.97</td>
</tr>
<tr>
<td>36259992</td>
<td>R.SIGIAPDVLCRGRDRA.I.P</td>
<td>1767.9664</td>
<td>0.392</td>
<td>0.14</td>
<td>2.15</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Monoisotopic Mass

Effect of search tolerances on Mass Error histogram

- If mass error plot not centered at 0, then narrow mass windows exclude valid data
- Decreasing mass and/or NET tolerance reduces background false positive level

Mass error histograms with linear mass correction

Mass error histograms with LCMSWarp mass correction
Identifying LC-MS Features

- Peak Matching Steps
  - Load LC-MS peak lists from Decon2LS
  - Filter data
  - Feature definition over elution time
  - Optionally, find paired features (e.g. $^{16}\text{O}/^{18}\text{O}$ pairs)
  - Align LC-MS features to AMT tags using LCMSWarp
  - Broad AMT tag DB search
    - ±25 ppm and ±0.05 NET
  - Search tolerance refinement
  - Final AMT tag DB search
    - e.g. ±1.8 ppm and ±0.02 NET
  - Report results
Automated Peak Matching

- Automated processing using VIPER
  - Processing steps and parameters defined in .Ini file
  - Separate .Ini file for $^{14}\text{N}/^{15}\text{N}$ pairs and $^{16}\text{O}/^{18}\text{O}$ pairs
**Peak Matching Results**

- Browsable result folders for visual QC of each dataset
  - *S. typhimurium* on 11T FTICR

  ![Image of 2D Plot Metrics](image1.png)

  **2D Plot Metrics**
  - Reasonable number of matches
  - NET range ≈ 0 to 1

  ![Image of Mass Error Histogram Metrics](image2.png)

  **Mass Error Histogram Metrics**
  - Well defined, symmetric mass error peak centered at 0 ppm

### Data Searched

- [Image of Mass Errors Before Refinement](image3.png)

### Data With Matches

- [Image of Mass Errors After Refinement](image4.png)
**Peak Matching Results**

- Browsable result folders for visual QC of each dataset
  - *S. typhimurium* on 11T FTICR

**NET Error Histogram Metrics**
- Well defined, symmetric NET error peak centered at 0

**Chromatogram Metrics**
- Narrow peaks evenly distributed throughout separation window

**Images**
- **NET Errors Before Refinement**
- **NET Errors After Refinement**
- **Total Ion Chromatogram (TIC)**
- **Base Peak Intensity (BPI) Chromatogram**
Peak Matching Results

- Browsable result folders for visual QC of each dataset
  - *S. typhimurium* on 11T FTICR

- **NET Alignment Surface Metrics**
  - Should show a smooth, bright yellow, diagonal line

- **NET Alignment Residual Metrics**
  - Data after recalibration should be narrowly distributed around zero
Peak Matching Results

- What about the unmatched LC-MS features?
  - Could align LC-MS features across datasets
  - Find the unmatched ones that show interesting trends
  - Generate list of the mass and elution times for the interesting features
  - Re-analyze the sample to perform targeted LC-MS/MS
  - Alignment example for 36 datasets using prototype software tool

After alignment
**LC-MS Feature Discovery**

- Similar approaches and software tools: High Res LC-MS
  - SpecArray (Pep3D, mzXML2dat, PepList, PepMatch, PepArray)
  - msInspect
  - PEPPeR
  - XCMS (for Metabolite profiling)
  - Surromed label-free quantitation software (MassView)
LC-MS Feature Discovery

- Similar approaches and software tools: Low Res LC-MS
  - Signal maps software
  - Informatics platform for global proteomic profiling using LC-MS
  - Computational Proteomics Analysis System (CPAS)
Course Outline

- Introduction
- Feature discovery in LC-MS datasets
  - Feature discovery in individual spectra
  - Feature definition over elution time
- Identifying LC-MS Features using an AMT tag DB
- Break
- AMT tag Pipeline Demo
- MAPQUANT, PEPPeR and GenePattern
- Panel Discussion
Alternative Approaches to Mass Spec Features: Picking, Assignment, and Statistical Discovery Tools

Jacob D. Jaffe
The Broad Institute of Harvard and MIT Proteomics Platform
Section Outline

- MAPQUANT: an image-based feature picking engine
- PEPPeR: a self-contained web-based Biomarker Discovery pipeline
- GenePattern: a suite of analysis and visualization tools that works with just about anything
A picture is worth 1000 parameters…
MAPQUANT

- MAPQUANT treats LCMS runs as images
- Can leverage tried and true image processing techniques
- Scripting language allows fine tuning and optimization for your data
  - Supporting both hi-res (FTMS) and lo-res (ITMS) platforms
- High computational overhead
  - Supports parallelization on a Linux cluster

Common Image Operations

- Structuring Elements

- Morphological Operations

Source: http://rkb.home.cern.ch/rkb/AN16pp/node178.html
Opening with a Cross Element
Opening with an Ellipse Element
Closing with a Cross Element
Filtering and Smoothing

- Eases ‘jagged’ peaks
- Gaussian, Boxcar, and Savitsky-Golay Filters
Rough Peak
Gaussian Smoothed Peak
Segmentation

- Find the rough peak borders (think paint bucket!)

- Break into many subtasks
Peak Centering/Modeling

- Find the Center: Structuring Elements Again
  - Raster structure element over segment
  - Find local maxima of products

- Model as a 3D Gaussian/oid
  - Model peaks to contain 95% of abundance
  - Altered model available to better fit chromatographic tailing effects
Isotope Clusters

- Find features that are close in time and not more than 1 m/z unit apart
  - Single-linkage clustering

- Deconvolve the clustered peaks
  - Iterative – once a successful deconvolution is made, these peaks are subtracted and the residual is deconvolved again until no more deconvolution can be done
Clustered, Deisotoped, Deconvolved
MAPQUANT Last Words

- Chain these operations into scripts for automated processing

- Break LCMS run into m/z stripes for parallelization on a cluster

- Output:
  - Parameterized peaks with m/z, r.t., abundance, z, carbon content

- Also visualization tools based on scripting language for method development / trouble shooting

- Cross-platform: Win32 and Linux
Multiple LCMS Experiments: Good with the Bad

- There is a lot of information in there
  - Peptide/protein IDs
  - Quantitative data
  - Statistical assessment

- The information may be noisy
  - Retention time drift
  - Instrument response noise

- Are there methods to leverage this information?
  - Without ‘perfect’ chromatography?
  - Without strict alignment?
PEPPeR Concepts – Samples and Data Acquisition
PEPPeR Concepts – Processing Continued...
PEPPeR Concepts – Analysis and Follow up

Landmark Matching

Peak Matching

Marker Discovery

Targeted MS/MS Identifications
Landmark Matching: Identity Propagation

- Use accurate mass, relative retention order comparison to identify peaks

Current Experiment

- m/z = 999.4991
- m/z = 999.4996
Landmark Matching: Identity Propagation

- Use accurate mass, relative retention order comparison to identify peaks

Current Experiment

Comparison Experiment

APEPTIDEK
m/z=999.4993

APDITEPEK
m/z=999.4993
Landmark Matching: Identity Propagation

Is \( X = M \) ?

Current Experiment

Comparison Experiment

APEPTIDEK
\[ m/z=999.4993 \]

APDITEPEK
\[ m/z=999.4993 \]
Landmark Matching: Identity Propagation

Current Experiment

Comparison Experiment

Is X = N ?

m/z=999.4991

APEPTIDEK
m/z=999.4993

APDITEPEK
m/z=999.4993

m/z=999.4996
Landmark Scoring and Confidence

\[ S = \sum_{i=1}^{w} \left[ \xi(\Lambda_{-i}, \Lambda_0) + \xi(\Lambda_0, \Lambda_i) \right] \]

\[ \xi(m,n) = \begin{cases} 
1 \text{ if } \tau(m) < \tau(n) \\
0.5 \text{ if } \tau(n) - \tau(m) < \delta \text{ and } \mu(m) + \sigma(m) > \mu(n) - \sigma(n) \\
-1 \text{ if else} \\
0 \text{ if else}
\end{cases} \]

\[ P_{\text{overall}} = \frac{P_{m/z} P_{\text{landmark}}}{P_{m/z \mid \text{landmark}}} \]

Let:

\( \Lambda \) be a list of peptides observed in the comparison experiment ordered by elution time. Here, elution time is defined by the centroid of all MS/MS scans leading to the identification of the peptide.

\( \Lambda_0 \) is defined as the position of the putative assignment in \( \Lambda \)

\( \mu(x) \) be the centroid of elution time of peptide \( x \) in the comparison experiment (in scans)

\( \sigma(x) \) be the standard deviation of elution time of peptide \( x \) in the comparison experiment (in scans)

\( \tau(x) \) be the centroid of elution time of peptide \( x \) in the current experiment (in seconds)

\( \delta \) be the average retention time peak width, such that peptides eluting within \( \delta \) sec are considered to be co-eluting (typically \( \delta = 30 \) s)

\( w \) the number of peptides to consider before and after the putative assignment on the landmark list (typically \( w = 3 \))
Nuts and bolts: How it works

- Match features to sequenced peptides in a single LCMS run
- Refine/recalibrate m/z tolerance
- Re-match features to sequenced peptides in a single LCMS run
- Now compare list of all features to Basis Set for mass, relative elution order matches given landmarks as reference points – propagation of identified features across multiple experiments
Peak Matching: Recognizing Identical Features

- Use landmarks to derive corrections and tolerances for clustering of features across LCMS experiments
- Break down the problem to make it parallelizable
Peak Matching: Recognizing Identical Features

- Use landmarks to derive corrections and tolerances for clustering of features across LCMS experiments

![Diagram showing m/z and Retention Time axes with examples of peaks across experiments.](Image)
Peak Matching: Recognizing Identical Features

- Use landmarks to derive corrections and tolerances for clustering of features across LCMS experiments
  - Gaussian mixture model (GMM) with parameters determined by maximizing Likelihood ratio using Expectation Maximization (EM)
  - Number of clusters determined using Bayesian Information Criterion (BIC)
  - Coalesce clusters if M/Z and RT variation is within tolerance
## Parameterized Peaks

<table>
<thead>
<tr>
<th>Peak ID</th>
<th>m/z</th>
<th>R.T.</th>
<th>z</th>
<th>Run 1</th>
<th>Run 2</th>
<th>Run 3</th>
<th>Run ...</th>
<th>Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>490.3144</td>
<td>62.0</td>
<td>3</td>
<td>607.6</td>
<td>544.2</td>
<td>581.0</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>743.3549</td>
<td>56.2</td>
<td>3</td>
<td>694.4</td>
<td>682.6</td>
<td>691.4</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>999.4991</td>
<td>22.5</td>
<td>2</td>
<td>209.6</td>
<td>247.6</td>
<td>232.6</td>
<td>...</td>
<td>APEPTIDEK</td>
</tr>
<tr>
<td>4</td>
<td>396.7187</td>
<td>20.5</td>
<td>3</td>
<td>321.7</td>
<td>344.9</td>
<td>318.5</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>934.6045</td>
<td>31.7</td>
<td>2</td>
<td>722.7</td>
<td>753.0</td>
<td>701.3</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>678.1993</td>
<td>32.4</td>
<td>3</td>
<td>371.2</td>
<td>387.2</td>
<td>441.4</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>999.4994</td>
<td>56.8</td>
<td>2</td>
<td>857.1</td>
<td>811.0</td>
<td>750.5</td>
<td>...</td>
<td>APDITEPEK</td>
</tr>
<tr>
<td>8</td>
<td>526.6502</td>
<td>46.0</td>
<td>3</td>
<td>183.6</td>
<td>169.0</td>
<td>155.2</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1105.3597</td>
<td>69.4</td>
<td>3</td>
<td>1130.1</td>
<td>1075.7</td>
<td>1075.1</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1292.0880</td>
<td>34.5</td>
<td>2</td>
<td>709.7</td>
<td>614.0</td>
<td>656.0</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

**Marker Discovery**

**Targeted MS/MS Identifications**
Experimental Parameters

- Thermo LTQ-FT Mass Spectrometer
  - Resolution 100,000
  - 1 MS, 3 MS/MS for Quantification Experiments
  - 1 MS, 10 MS/MS for Identification Experiments

- Agilent 1100 Nano-flow Chromatograph
  - 12.5 cm x 75 μm Reprosil 3 μm C18AQ material
  - 200 nl/min, 50’ RP gradient, 15 μm tip opening

- SpectrumMill for MS/MS searches
  - 0.05 Da tolerance, Autovalidation

- MapQuant for feature detection
Calibration and Landmark Performance

Scale Mixture

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aprotinin</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>100</td>
<td>200</td>
<td>300</td>
</tr>
<tr>
<td>Ribonuclease A</td>
<td>300</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>Myoglobin</td>
<td>200</td>
<td>300</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>100</td>
</tr>
<tr>
<td>beta-Lactoglobulin</td>
<td>100</td>
<td>200</td>
<td>300</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>10</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>alpha Casein</td>
<td>30</td>
<td>100</td>
<td>200</td>
<td>300</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Carbonic anhydrase</td>
<td>20</td>
<td>30</td>
<td>100</td>
<td>200</td>
<td>300</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>Ovalbumin</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>100</td>
<td>200</td>
<td>300</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Fibrinogen</td>
<td>3</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>100</td>
<td>200</td>
<td>300</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>BSA</td>
<td>2</td>
<td>3</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>100</td>
<td>200</td>
<td>300</td>
<td>1</td>
</tr>
<tr>
<td>Transferrin</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Plasminogen</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>beta-Galactosidase</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

All concentrations in fmol/ul (nM)
Inject 1 ul x 5 replicates each

Peaks with IDs (avg. per run):
165 ⇒ 281         +70%
False positive rate:
93%  p < 0.005
100%  p < 0.05
False negative rate:
~2%
**Measurement of Ratios with Variability**

### Variability Mixture

<table>
<thead>
<tr>
<th></th>
<th>Person A</th>
<th>Person B</th>
<th>Person C</th>
<th>Person D</th>
<th>Person E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aprotinin</td>
<td>100 5</td>
<td>100 5</td>
<td>100 5</td>
<td>100 5</td>
<td>100 5</td>
</tr>
<tr>
<td>Ribonuclease A</td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
</tr>
<tr>
<td>Myoglobin</td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
</tr>
<tr>
<td>beta-Lactoglobulin</td>
<td>50 1</td>
<td>50 1</td>
<td>50 1</td>
<td>50 1</td>
<td>50 1</td>
</tr>
<tr>
<td>alpha Casein</td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
</tr>
<tr>
<td>Carbonic anhydrase</td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
</tr>
<tr>
<td>Ovalbumin</td>
<td>5 10</td>
<td>5 10</td>
<td>5 10</td>
<td>5 10</td>
<td>5 10</td>
</tr>
<tr>
<td>BSA</td>
<td>200 200</td>
<td>200 200</td>
<td>200 200</td>
<td>200 200</td>
<td>200 200</td>
</tr>
<tr>
<td>Transferrin</td>
<td>10 5</td>
<td>10 5</td>
<td>10 5</td>
<td>10 5</td>
<td>10 5</td>
</tr>
<tr>
<td>Plasminogen</td>
<td>2.5 25</td>
<td>2.5 25</td>
<td>2.5 25</td>
<td>2.5 25</td>
<td>2.5 25</td>
</tr>
<tr>
<td>beta-Galactosidase</td>
<td>1 10</td>
<td>1 10</td>
<td>1 10</td>
<td>1 10</td>
<td>1 10</td>
</tr>
</tbody>
</table>

All concentrations in fmol/ul (nM)  
Inject 1 ul x 5 replicates each

### Complex Variability Mixture:

**Mix α + Mitochondrial Protein from 2 wk. mouse liver**

**Mix β + Mitochondrial Protein from 6 wk. mouse liver**

1 prep each sample, 6 injections each
Discovery of Novel Markers with PEPPeR

- Designed accurate mass ‘inclusion lists’ to hit these targets
  - Confident IDs of previously identified peptides agree 100% of the time (59/59)
  - 60 novel confident peptide IDs
    - 25 belong to proteins in the mix
      - 24/25 are changing
    - 35 are from proteins not designed to be in the mixture

<table>
<thead>
<tr>
<th>gi Number</th>
<th>Species</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>223424</td>
<td>E. coli</td>
<td>RNA polymerase β'</td>
</tr>
<tr>
<td>38491462</td>
<td>E. coli</td>
<td>GroEL</td>
</tr>
<tr>
<td>42144</td>
<td>E. coli</td>
<td>NusA</td>
</tr>
<tr>
<td>42818</td>
<td>E. coli</td>
<td>RNA polymerase β</td>
</tr>
<tr>
<td>42900</td>
<td>E. coli</td>
<td>Ribosomal protein S1</td>
</tr>
<tr>
<td>26249756</td>
<td>E. coli</td>
<td>Argininosuccinate synthase</td>
</tr>
<tr>
<td>8099322</td>
<td>B. taurus</td>
<td>κ-casein</td>
</tr>
</tbody>
</table>

B-Galactosidase had 1:10 ratio!
Casein had 10:1 ratio!
GenePattern is a suite of tools originally developed for microarray analysis
- AIM: reproducible research through well-defined processing pipelines

Many analysis modules available
- PEPPeR: Landmark Matching and Peak Matching
- Daisy-chainable into pipelines
- Feed into statistical tools
PEPPeR in GenePattern

**LandmarkMatching**

- **peakList filename:** C:\PepperExample\peakData.zip (Browse...)
  - Zip file containing peak lists
- **retentionTime filename:** C:\PepperExample\rtData.zip (Browse...)
  - Zip file containing retention time (in this case from mapquant)
- **sampleInfo filename:** C:\PepperExample\smallSample.csv (Browse...)
  - CSV file with header: experiment, sample, class
- **globalID filename:** C:\PepperExample\varmixGlobalID.txt (Browse...)
  - Global list of identified peptides
- **accurateMass filename:**
- **p-reflCoef filename:**
  - OPTIONAL: Specify the accurate mass table
  - OPTIONAL: Specify the p-refl coefficients in a file
- **Bootstrap:**
  - No (Use bootstrapping to calculate landmark match statistics)

**Buttons:**
- run
- reset
- help
Landmark Matching is platform agnostic

- Need to get your data into a few simple flat-file formats and then zip them up together
- Search engines i.e. SEQUEST, SpectrumMill, Mascot, etc.
- Peak Pickers: MAPQUANT, msInspect, DeCyder, etc.
- Some helper apps can be found with the PEPPeR bundle on the GenePattern website

All works via web-client interface

- Just press go
Landmark Matching Output

The main output is a zipped directory of all the processed files. This can be used as input into the PeakMatch module.

It is a good idea to check the error log to make sure that everything was processed correctly.
Peak Matching Interface
GenePattern Downstream Tools

- **Differential analysis.marker selection**
  - Gene/Class neighbors
  - Comparative marker selection
  - Gene Set Enrichment Analysis

- **Class Prediction – supervised learning – with cross-validation**
  - Regression trees
  - K-nearest neighbors
  - Neural networks
  - Support Vector Machine

- **Class Discovery – unsupervised learning**
  - Hierarchical clustering
  - Self-organizing maps
  - Principal Component Analysis

- **Data Visualization**
  - Heat Maps, etc.
Summary – what I hope you learned

- One basis for peak picking based on image processing techniques (MAPQUANT)
- PEPPeR: Landmark Matching and Peak Matching
  - Keep track of all of those pesky peaks that you picked!
- GenePattern: A web-based tool to coordinate reproducible research
- An entrée into downstream discovery methods in an automated pipeline (more GenePattern)
Acknowledgements

- **Broad Proteomics:**
  - Steve Carr
  - **D.R. Mani**
  - Vincent Fusaro

- **Broad GenePattern Team:**
  - **Michael Reich**
  - Josh Gould

- **Church Lab, Harvard Medical School**
  - George Church
  - **Kyriacos Leptos**
URLs:

● PEPPeR / GenePattern:
  ● http://www.broad.mit.edu/cancer/software/genepattern/
  ● http://www.broad.mit.edu/cancer/software/genepattern/desc/proteomics.html

● MAPQUANT:
  ● http://arep.med.harvard.edu/MapQuant/
Course Outline

- Introduction
- Feature discovery in LC-MS datasets
  - Feature discovery in individual spectra
  - Feature definition over elution time
- Identifying LC-MS Features using an AMT tag DB
- Break
- AMT tag Pipeline Demo
- MAPQUANT, PEPPeR and GenePattern
- Panel Discussion