Mass spectrometry-based proteomics

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Spring 2020
Daifeng Wang
daifeng.wang@wisc.edu
Goals for lecture

Key concepts
- Benefits of mass spectrometry
- Generating mass spectrometry data
- Computational tasks
- Matching spectra and peptides
Mass spectrometry uses

• Mass spectrometry is like the protein analog of RNA-seq
  – Quantify abundance or state of all (many) proteins
  – No need to specify proteins to measure in advance

• Other applications in biology
  – Targeted proteomics
  – Metabolomics
  – Lipidomics
Advantages of proteomics

• Proteins are functional units in a cell
  – Protein abundance directly relevant to activity

• Post-translational modifications
  – Change protein state

Phosphorylation in signaling

Thermo Fisher Scientific

Histone modifications

Latham *Nature Structural & Molecular Biology* 2007; Katie Ris-Vicari
Estimating protein levels from gene expression

- Correlation between gene expression and protein abundance has been debated
- Gene expression tells us nothing about post-translational modifications

Contribution to protein levels

Li and Biggin *Science* 2015
Mass spectrometry workflow

1. **Sample preparation and peptide separation**
   - Digestion
   - Depletion/enrichment
   - Separation

2. **Filter based on MS**
   - Mass spectrometry
     - Fragment peptide
     - Q1
     - Q2 Collision Cell
     - Q3

3. **Peptide identification**
   - Theoretical
   - Acquisition
   - Sequence database matching

---

Nesvizhskii *Journal of Proteomics* 2010
### Amino Acids

- 20 amino acids
- Building blocks of proteins
- Known molecular weight

**Common template**

<table>
<thead>
<tr>
<th>Amino-terminal</th>
<th>Carboxy-terminal</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Amino-terminal" /></td>
<td><img src="image2.png" alt="Carboxy-terminal" /></td>
</tr>
</tbody>
</table>

---

**Nonpolar, Hydrophobic**

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>R Groups</th>
<th>Molecular Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>-</td>
<td>MW = 89</td>
</tr>
<tr>
<td>Valine</td>
<td>-</td>
<td>MW = 117</td>
</tr>
<tr>
<td>Leucine</td>
<td>-</td>
<td>MW = 131</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>-</td>
<td>MW = 131</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>-</td>
<td>MW = 204</td>
</tr>
<tr>
<td>Methionine</td>
<td>-</td>
<td>MW = 149</td>
</tr>
<tr>
<td>Proline</td>
<td>-</td>
<td>MW = 115</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>-</td>
<td>MW = 133</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>-</td>
<td>MW = 147</td>
</tr>
</tbody>
</table>

**Polar, Uncharged**

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>R Groups</th>
<th>Molecular Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycine</td>
<td>-</td>
<td>MW = 75</td>
</tr>
<tr>
<td>Serine</td>
<td>-</td>
<td>MW = 105</td>
</tr>
<tr>
<td>Threonine</td>
<td>-</td>
<td>MW = 119</td>
</tr>
<tr>
<td>Cysteine</td>
<td>-</td>
<td>MW = 121</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>-</td>
<td>MW = 181</td>
</tr>
<tr>
<td>Asparagine</td>
<td>-</td>
<td>MW = 132</td>
</tr>
<tr>
<td>Glutamine</td>
<td>-</td>
<td>MW = 146</td>
</tr>
<tr>
<td>Lysine</td>
<td>-</td>
<td>MW = 146</td>
</tr>
<tr>
<td>Arginine</td>
<td>-</td>
<td>MW = 174</td>
</tr>
<tr>
<td>Histidine</td>
<td>-</td>
<td>MW = 155</td>
</tr>
</tbody>
</table>

---

Peptide fragmentation

- Select similar peptides from MS\(^1\)
- Fragment with high energy collisions
- Break peptide bonds

Wikipedia, Yassine Mrabet

Charge on amino-terminal (b) or carboxy-terminal fragment (y)
Subscript = # R groups retained

Steen and Mann Nat Rev Mol Cell Biol 2004
Mass spectra

Mass-to-charge ratio

Fragment and analyze one precursor ion

Spectrum contains information about amino acid sequence, fragment at different bonds

Steen and Mann Nat Rev Mol Cell Biol 2004
From spectra to peptides

Peptide

From spectra to peptides

Acquired spectrum

uninterpreted

MS/MS

partial sequence
(short sequence tag)

VDL / LDV

De novo sequencing
(full (extended) sequence)

Sequence DB search
Theoretical spectrum

200 400 600 800 1000 1200

m/z

Spectral library search
Library spectrum

Sequence tag-assisted search
Theoretical spectrum

peptides with VDL or LDV tag only

Output: ranked peptide list

peptide score

VSTPNVSVVDLTCR 5.6
ISLLDAQSAPLR 1.3
CDVVSNTIIAE 1.1

best match

VSTPNVSVVDLTCR

Nesvizhskii Journal of Proteomics 2010 10
Sequence database search

- Need to define a scoring function
- Identify peptide-spectrum match (PSM)

Steen and Mann *Nat Rev Mol Cell Biol* 2004
SEQUEST

- Cross correlation (xcorr)
- Similarity between theoretical spectrum (x) and acquired spectrum (y)
- Correction for mean similarity at different offsets

Fast SEQUEST

- SEQUEST originally only applied to top 500 peptides based on coarse filtering score

\[
\text{xcorr} = x_0 \cdot y_0 - \left( \sum_{\tau=-75}^{\tau=+75} x_0 \cdot y_\tau \right) / 151
\]

\[
\text{xcorr} = x_0 \cdot \left( y_0 - \left( \sum_{\tau=-75}^{\tau=+75} y_\tau \right) / 151 \right)
\]

\[
\text{xcorr} = x_0 \cdot y' \quad \text{where} \quad y' = y_0 - \left( \sum_{\tau=-75, \tau \neq 0}^{\tau=+75} y_\tau \right) / 150
\]

Eng et al. J Proteome Res 2008
PSM significance

• E-value: expected number of null peptides with score ≥ observed score

• Compute FDR from E-value distribution

• Add decoy peptides to database
  – Reversed peptide sequences
  – Used to estimate false discoveries
**Target-decoy strategy**

Entire dataset, N spectra

Best match for each spectrum

<table>
<thead>
<tr>
<th>spec</th>
<th>peptide</th>
<th>score</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ISLLDAQSAPLR</td>
<td>4.5</td>
<td>target</td>
</tr>
<tr>
<td>2</td>
<td>VVEELCTPEGK</td>
<td>3.9</td>
<td>target</td>
</tr>
<tr>
<td>3</td>
<td>GDAVFPVIDALNR</td>
<td>3.6</td>
<td>target</td>
</tr>
<tr>
<td>4</td>
<td>VNSPMKVVVPTPK</td>
<td>1.7</td>
<td>decoy</td>
</tr>
<tr>
<td>5</td>
<td>ECDVVSNTIIAEK</td>
<td>1.5</td>
<td>target</td>
</tr>
<tr>
<td>6</td>
<td>LIHSVFGIGEKE</td>
<td>1.1</td>
<td>decoy</td>
</tr>
</tbody>
</table>

(sorted by score)

Apply score threshold $S_T$

Calculate $N_t(S_T)$ and $N_d(S_T)$:
- number of target/decoy PSM with $S \geq S_T$

Estimate FDR

\[
FDR(S_T) = \frac{N_d(S_T)}{N_t(S_T)}
\]

Select threshold $S_T$ to achieve desired FDR

**Nesvizhskii Journal of Proteomics 2010**

target PSMs above score threshold = $N_t(S_T)$
decoy PSMs above score threshold = $N_d(S_T)$
Identifying proteins

- Even after identifying PSM, still need to identify protein of origin

Serang and Noble Stat Interface 2012
Mass spectrometry versus RNA-seq

• RNA-seq
  – Transcript → RNA fragment → paired-end read

• Mass spectrometry
  – Protein → peptides → ions → spectrum

• Mapping spectra to proteins more ambiguous than mapping reads to genes
• Spectra state space is enormous
Protein-protein interactions

- Affinity-purification mass spectrometry
- Purify protein of interest, identify complex members

Smits and Vermeulen Trends in Biotechnology 2016
Protein-protein interactions

- Mass spectrometry identifies proteins in the complex
- Must control for contaminants

Post-translational modifications (PTMs)

- Shift the peptide mass by a known quantity
## Phosphoproteomics example

<table>
<thead>
<tr>
<th>Gene</th>
<th>Modified Site</th>
<th>Peptide</th>
<th>Phosphorylation (Treatment / Control)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAMTS10</td>
<td>S74</td>
<td>RGTGATAAES[167.00]R</td>
<td>0.30</td>
</tr>
<tr>
<td>CABYR</td>
<td>T16</td>
<td>T[181.01]LLEGISR</td>
<td>0.37</td>
</tr>
<tr>
<td>TTC7B</td>
<td>T152</td>
<td>VIEQDET[181.01]R</td>
<td>5.97</td>
</tr>
</tbody>
</table>

Sychev et al *PLoS Pathogens* 2017
Phosphoproteomics interpretation

- Predict kinases/phosphatases for phospho sites

Linding et al Cell 2007
Mass spectrometry replicates

- Doesn’t identify all proteins in the sample
  - Data dependent acquisition has low overlap across replicates
  - Partly due to biological variation
  - New protocols to overcome this

- Phosphorylation PTMs are especially variable
  - Grimsrud et al *Cell Metabolism* 2012
    - 5 biological replicates
    - 9,558 phosphoproteins identified
    - 5.6% in all replicates
Data independent acquisition

- Not dependent on most abundance signals in MS\(^1\)
- Sliding \(m/z\) window

Doerr Nature Methods 2015
Mass spectrometry summary

• Incredibly powerful for looking at biological processes beyond gene expression
  – Protein abundance
  – Post-translational modifications
  – Metabolites
  – Protein-protein interactions

• Typically reports relative abundance

• Labeling strategies for comparative analysis
  – Compare relative abundance in multiple conditions

• Missing data was a big problem, but improving

• Fully probabilistic analysis pipelines are not the most popular tools
  – Arguably greater diversity in software than RNA-seq