Proteomics

November 13, 2007
Acknowledgement

• Slides presented here have been borrowed from presentations by:
  – Dr. Mark A. Knepper (LKEM, NHLBI, NIH)
  – Dr. Nathan Edwards (Center for Bioinformatics and Computational Biology, University of Maryland, College Park)
  – Dr. Catherine Fenselau (Dept. of Chemistry, University of Maryland, College Park)
What Is “Proteomics”?

Simultaneous analysis of all members of a defined set of proteins.
Identification & Analysis of Proteins

- Sample preparation and handling
- Determination of amino acid sequence
- Protein identification and quantification
- Cell mapping
What is Mass Spectrometry?

- Measures masses of individual molecules
- Molecules are converted into ions in the gas phase
- Molecules must be charged (+ or -)
- Measure mass-to-charge ratio (m/z) of ions and its charged fragments
- Mass-to-charge ratio (m/z) of ions detected
Mass Spectrometry Components

Ion Source → Mass Analyzer → Detector

Inlet
Mass Spectrometry Components

Ion Source → Mass Analyzer → Detector

Vacuum

Inlet
Mass Spectrometry Components

Ion Source → Mass Analyzer → Detector

Vacuum

Electronics
Example Mass Spectrum
Ionization Methods

• Matrix-assisted laser desorption/ionization (MALDI)
• Electrospray Ionization (ESI)
MALDI

- Creates ions by excitation with a laser of a sample that is mixed with an excess amount of a matrix component
- Singly charged ions are guided to mass analyzer and detector
- Used with time-of-flight (TOF) analyzer
Protein Identification by MALDI-TOF Mass Spectrometry

Trypsin Digestion (cleaves after arginine or lysine) → MALDI-TOF Mass Spectrometry (measures peptide masses) → Peptide Mass Fingerprint

List of measured peptide masses is matched against theoretically predicted tryptic peptides of all proteins to identify sample protein.

Limitations of 2-Dimensional Electrophoresis

1. Does not work well for hydrophobic membrane proteins.

2. Does not work well for extremes of molecular weights.

3. Difficult to detect non-abundant proteins.
ESI

- Creates gas-phase ions by applying a potential to a flowing liquid that contains the analyte and solvent molecules.
- Yields multiply charged ions
- Used with quadrupole or ion-trap mass analyzers in tandem mass spectrometry
Protein Identification in Mixtures Using LC-MS/MS

Mixture of Proteins

Trypsin Digestion (cleaves after arginine or lysine)

Mixture of Tryptic Peptides

HPLC Separation of Peptides; Outflow to Electrospray MS/MS

$\text{MS}^1$

$\text{MS}^2$

Peptide sequence from CID spectrum is used to search protein sequence databases to identify protein in original mixture.
Sample Preparation for Mass Spectrometry

Enzymatic Digest and Fractionation

Trypsin cleaves at Arg & Lys
Single Stage MS
Tandem Mass Spectrometry (MS/MS)
Peptide Fragmentation

Peptides consist of amino-acids arranged in a linear backbone.

H…-HN-CH-CO-NH-CH-CO-NH-CH-CO-...OH

<table>
<thead>
<tr>
<th>R_{i-1}</th>
<th>R_i</th>
<th>R_{i+1}</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA residue_{i-1}</td>
<td>AA residue_i</td>
<td>AA residue_{i+1}</td>
</tr>
</tbody>
</table>

N-terminus

C-terminus
Peptide Fragmentation
Peptide Fragmentation

-\text{HN-CH-CO-NH-CH-CO-NH-}

$\ y_{n-i} \quad \ y_{n-i-1} \$

$\ R_i \quad \ CH-R'_{i+1} \quad \ CH-R''_{i+1} \quad \ R_i'' \quad \ R_i'''

$\ b_i \quad \ b_{i+1} \$
# Peptide Fragmentation

Peptide: S-G-F-L-E-E-D-E-L-K

<table>
<thead>
<tr>
<th>MW</th>
<th>ion</th>
<th>Peptide</th>
<th>ion</th>
<th>MW</th>
</tr>
</thead>
<tbody>
<tr>
<td>88</td>
<td>(b_1)</td>
<td>S GFLEEDELK</td>
<td>(y_9)</td>
<td>1080</td>
</tr>
<tr>
<td>145</td>
<td>(b_2)</td>
<td>SG FLEEDELK</td>
<td>(y_8)</td>
<td>1022</td>
</tr>
<tr>
<td>292</td>
<td>(b_3)</td>
<td>SGF LEEDELK</td>
<td>(y_7)</td>
<td>875</td>
</tr>
<tr>
<td>405</td>
<td>(b_4)</td>
<td>SGFL EEDELK</td>
<td>(y_6)</td>
<td>762</td>
</tr>
<tr>
<td>534</td>
<td>(b_5)</td>
<td>SGFLE EDELK</td>
<td>(y_5)</td>
<td>633</td>
</tr>
<tr>
<td>663</td>
<td>(b_6)</td>
<td>SGFLEE DELK</td>
<td>(y_4)</td>
<td>504</td>
</tr>
<tr>
<td>778</td>
<td>(b_7)</td>
<td>SGFLEED ELK</td>
<td>(y_3)</td>
<td>389</td>
</tr>
<tr>
<td>907</td>
<td>(b_8)</td>
<td>SGFLEEDE LK</td>
<td>(y_2)</td>
<td>260</td>
</tr>
<tr>
<td>1020</td>
<td>(b_9)</td>
<td>SGFLEEDEL K</td>
<td>(y_1)</td>
<td>147</td>
</tr>
</tbody>
</table>
Peptide Fragmentation

<table>
<thead>
<tr>
<th>m/z</th>
<th>% Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>88</td>
<td></td>
</tr>
<tr>
<td>145</td>
<td></td>
</tr>
<tr>
<td>292</td>
<td></td>
</tr>
<tr>
<td>405</td>
<td></td>
</tr>
<tr>
<td>534</td>
<td></td>
</tr>
<tr>
<td>663</td>
<td></td>
</tr>
<tr>
<td>778</td>
<td></td>
</tr>
<tr>
<td>907</td>
<td></td>
</tr>
<tr>
<td>1020</td>
<td></td>
</tr>
<tr>
<td>1166</td>
<td></td>
</tr>
<tr>
<td>1166</td>
<td></td>
</tr>
<tr>
<td>1080</td>
<td></td>
</tr>
<tr>
<td>1022</td>
<td></td>
</tr>
<tr>
<td>875</td>
<td></td>
</tr>
<tr>
<td>762</td>
<td></td>
</tr>
<tr>
<td>633</td>
<td></td>
</tr>
<tr>
<td>504</td>
<td></td>
</tr>
<tr>
<td>389</td>
<td></td>
</tr>
<tr>
<td>260</td>
<td></td>
</tr>
<tr>
<td>147</td>
<td></td>
</tr>
</tbody>
</table>

b ions: S, G, F, L, E, E, D, E, L, K

y ions: 1166, 1080, 1022, 875, 762, 633, 504, 389, 260, 147
Peptide Fragmentation

<table>
<thead>
<tr>
<th>88</th>
<th>145</th>
<th>292</th>
<th>405</th>
<th>534</th>
<th>663</th>
<th>778</th>
<th>907</th>
<th>1020</th>
<th>1166</th>
<th>b ions</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>G</td>
<td>F</td>
<td>L</td>
<td>E</td>
<td>E</td>
<td>D</td>
<td>E</td>
<td>L</td>
<td>K</td>
<td></td>
</tr>
<tr>
<td>1166</td>
<td>1080</td>
<td>1022</td>
<td>875</td>
<td>762</td>
<td>633</td>
<td>504</td>
<td>389</td>
<td>260</td>
<td>147</td>
<td>y ions</td>
</tr>
</tbody>
</table>

% Intensity

m/z

y ions

y5

y6

y7

y8

y9
Peptide Fragmentation

K 1166
L 1020
E 907
D 778
E 663
E 534
L 405
F 292
G 145
S 88
b ions

100
% Intensity

m/z

1166 1080 1022 875 762 633 504 389 260 147

y ions

y6 y7 y8 y9
b3 b4 b5 b6 b7 b8 b9

y2 y3 y4 y5
Peptide Identification

Given:
- The mass of the parent ion, and
- The MS/MS spectrum

Output:
- The amino-acid sequence of the peptide
Peptide Identification

Two methods:

• *De novo* interpretation
• Sequence database search
De Novo Interpretation
De Novo Interpretation
De Novo Interpretation
## De Novo Interpretation

<table>
<thead>
<tr>
<th>Amino-Acid</th>
<th>Residual MW</th>
<th>Amino-Acid</th>
<th>Residual MW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>71.03712</td>
<td>Methionine</td>
<td>131.04049</td>
</tr>
<tr>
<td>Cysteine</td>
<td>103.00919</td>
<td>Asparagine</td>
<td>114.04293</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>115.02695</td>
<td>Proline</td>
<td>97.05277</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>129.04260</td>
<td>Glutamine</td>
<td>128.05858</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>147.06842</td>
<td>Arginine</td>
<td>156.10112</td>
</tr>
<tr>
<td>Glycine</td>
<td>57.02147</td>
<td>Serine</td>
<td>87.03203</td>
</tr>
<tr>
<td>Histidine</td>
<td>137.05891</td>
<td>Threonine</td>
<td>101.04768</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>113.08407</td>
<td>Valine</td>
<td>99.06842</td>
</tr>
<tr>
<td>Lysine</td>
<td>128.09497</td>
<td>Tryptophan</td>
<td>186.07932</td>
</tr>
<tr>
<td>Leucine</td>
<td>113.08407</td>
<td>Tyrosine</td>
<td>163.06333</td>
</tr>
</tbody>
</table>
De Novo Interpretation

…from Lu and Chen (2003), JCB 10:1
De Novo Interpretation

…from Lu and Chen (2003), JCB 10:1
De Novo Interpretation

- Find *good* paths in spectrum graph
- Can’t use same peak twice
  - Forbidden pairs
  - “Nested” forbidden pairs
- Simple peptide fragmentation model
- Usually many apparently good solutions
- Needs better fragmentation model
- Needs better path scoring
De Novo Interpretation

- Amino-acids have duplicate masses!
- Incomplete ladders create ambiguity.
- Noise peaks and unmodeled fragments create ambiguity
- “Best” de novo interpretation may have no biological relevance
- Current algorithms cannot model many aspects of peptide fragmentation
- Identifies relatively few peptides in high-throughput workflows
Sequence Database Search

• Compares peptides from a protein sequence database with spectra
• Filter peptide candidates by
  – Parent mass
  – Digest motif
• Score each peptide against spectrum
  – Generate all possible peptide fragments
  – Match putative fragments with peaks
  – Score and rank
Sequence Database Search

KLEDEELFGS
Sequence Database Search

<table>
<thead>
<tr>
<th>b ions</th>
<th>y ions</th>
</tr>
</thead>
<tbody>
<tr>
<td>88</td>
<td>1166</td>
</tr>
<tr>
<td>145</td>
<td>1080</td>
</tr>
<tr>
<td>292</td>
<td>1022</td>
</tr>
<tr>
<td>405</td>
<td>875</td>
</tr>
<tr>
<td>534</td>
<td>762</td>
</tr>
<tr>
<td>663</td>
<td>633</td>
</tr>
<tr>
<td>778</td>
<td>504</td>
</tr>
<tr>
<td>907</td>
<td>389</td>
</tr>
<tr>
<td>1020</td>
<td>260</td>
</tr>
<tr>
<td>1166</td>
<td>147</td>
</tr>
</tbody>
</table>

S G F L E E D E L K

% Intensity

m/z
Sequence Database Search

<table>
<thead>
<tr>
<th></th>
<th>88</th>
<th>145</th>
<th>292</th>
<th>405</th>
<th>534</th>
<th>663</th>
<th>778</th>
<th>907</th>
<th>1020</th>
<th>1166</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>G</td>
<td>F</td>
<td>L</td>
<td>E</td>
<td>E</td>
<td>D</td>
<td>E</td>
<td>L</td>
<td>K</td>
</tr>
<tr>
<td>1166</td>
<td>1080</td>
<td>1022</td>
<td>875</td>
<td>762</td>
<td>633</td>
<td>504</td>
<td>389</td>
<td>260</td>
<td>147</td>
<td></td>
</tr>
</tbody>
</table>

b ions

y ions
Sequence Database Search

- No need for complete ladders
- Possible to model all known peptide fragments
- Sequence permutations eliminated
- All candidates have some biological relevance
- Practical for high-throughput peptide identification
- Correct peptide might be missing from database!
Peptide Candidate Filtering

>ALBU_HUMAN
MKWVTFISLLFLFSSAYSRGVFRRDAHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLNVNEVTEFAK...

**No missed cleavage sites**

MK
WVTFISLLFLFSSAYSRGVFR
R
DAHK
SEVAHR
FK
DLGEENFK
ALVLIAFAQYLQQCPFEDHVKLNVNEVTEFAK...
...

Peptide Candidate Filtering

>ALBU_HUMAN
MKWVTFIGLLFLFSSAYSRGVFRRDAHKSEVAHRFKDLGVEENFKALVLIAFAQYLQQCPFEDHVKLNVETEFAK...

**One missed cleavage site**

MKWVTFIGLLFLFSSAYSR
WVTFIGLLFLFSSAYSRGVFRR
GVFRR
RDAHK
DAHKSEVAHR
SEVAHRFK
FKDLGEENFK
DLGVEENFKALVLIAFAQYLQQCPFEDHVKLNVETEFAK
ALVLIAFAQYLQQCPFEDHVKLNVETEFAK...

...
Peptide Candidate Filtering

Peptide molecular weight
• Only have m/z value
  – Need to determine charge state
• Ion selection tolerance
• Mass for each amino-acid symbol?
  – Monoisotopic vs. Average
  – “Default” residual mass
  – Depends on sample preparation protocol
  – Cysteine almost always modified
Peptide Scoring

• Peptide fragments vary based on
  – The instrument
  – The peptide’s amino-acid sequence
  – The peptide’s charge state
  – Etc…

• Search engines model peptide fragmentation to various degrees.
  – Speed vs. sensitivity tradeoff
  – y-ions & b-ions occur most frequently
Sequence Database Search Traps and Pitfalls

Search options may eliminate the correct peptide

- Parent mass tolerance too small
- Fragment m/z tolerance too small
- Incorrect parent ion charge state
- Non-tryptic or semi-tryptic peptide
- Incorrect or unexpected modification
- Sequence database too conservative
Sequence Database Search
Traps and Pitfalls

Search options can cause infinite search times

• Variable modifications increase search times exponentially
• Non-tryptic search increases search time by two orders of magnitude
• Large sequence databases contain many irrelevant peptide candidates
Sequence Database Search
Traps and Pitfalls

Best available peptide isn’t necessarily correct!

• Score statistics are essential
  – What is the chance a peptide could score this well by chance alone?
• The wrong peptide can look correct if the right peptide is missing!
• Need scores that are invariant to spectrum quality and peptide properties
Sequence Database Search
Traps and Pitfalls

Search engines often make incorrect assumptions about sample prep

- Proteins with lots of identified peptides are not more likely to be present
- Peptide identifications do not represent independent observations
- All proteins are not equally interesting to report
Good spectral processing can make a big difference

- Poorly calibrated spectra require large m/z tolerances
- Poorly baselined spectra make small peaks hard to believe
- Poorly de-isotoped spectra have extra peaks and misleading charge state assignments