De novo Protein Sequencing by Combining Top-Down and Bottom-Up Tandem Mass Spectra

Xiaowen Liu

Department of BioHealth Informatics, Department of Computer and Information Sciences, Indiana University-Purdue University Indianapolis

Center for Computational Biology and Bioinformatics, Indiana University School of Medicine
Top-Down Proteomics Becomes Reality

“Early proteomics methods used enzymes to digest proteins into pieces that could be easily analyzed by mass spectrometry. Those methods are now mature and routinely detect peptides from thousands of proteins in a single run...

But the great strength of those methods is also their greatest weakness. What’s being analyzed is no longer the actual biological actors but the pieces left after they’ve been broken apart...

By starting with intact proteins, rather than their pieces, top-down analysis more accurately reflects the structure and properties of actual biological systems than does bottom-up proteomics....”
Top-Down vs. Bottom-Up MS

**Bottom-up:**

- Tryptic Digestion
- Peptides 5-30aa

**Top-down:**

- No digestion and MS/MS
- Protein 3K-50K Da, 30-400aa
Top-Down vs Bottom-Up MS

• Measurable $m/z$ values
  – Commercial iontrap/ Orbitrap mass spectrometers: up to 4000 $m/z$

• Bottom-up mass spectra
  – Small masses: 500 Da – 4000 Da
  – Low charge

• Top-down mass spectra
  – Large masses, i.e., 20k Da
  – High charge ions
Large Masses Make Spectra Complex

- Isotopes
  - $^{12}\text{C}$: Mass: 12.000, frequency: 98.93%
  - $^{13}\text{C}$: Mass: 13.003, frequency: 1.07%
- 100-carbon molecules
  - The proportion of molecules is with all 100 carbons being $^{12}\text{C}$ is $0.9893^{100} \approx 31.54\%$

Theoretical isotopomer envelope for Lysozyme (14303.88 Da)
Top-down spectra usually have order(s) of magnitude more peaks and complex pattern of isotopomer envelopes.
Why Top-Down Proteomics Becomes Reality?

• High accuracy, high resolution, and high-throughput mass spectrometers: Orbitrap, FTICR.

<table>
<thead>
<tr>
<th>Mass analyzer</th>
<th>Suitable for Top Down</th>
<th>Spectral acquisition time/s</th>
<th>Resolution/Da</th>
<th>Mass accuracy (ppm)</th>
<th>Performance at 8 kDa</th>
<th>Available fragmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ion trap</td>
<td>+</td>
<td>0.05–0.3</td>
<td>1000</td>
<td>100–200</td>
<td></td>
<td>CID, ETD</td>
</tr>
<tr>
<td>TOF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CID, ISD</td>
</tr>
<tr>
<td>TOF-TOF</td>
<td>++</td>
<td>&lt;0.01</td>
<td>10 000</td>
<td>5–20</td>
<td></td>
<td>PSD</td>
</tr>
<tr>
<td>Q-TOF</td>
<td>++</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FT</td>
<td>++</td>
<td>0.1–1</td>
<td>60 000</td>
<td>3–10</td>
<td></td>
<td>CID, ETD, HCD, CID, ECD, IRMPD</td>
</tr>
<tr>
<td>Orbitrap</td>
<td>+++</td>
<td>0.1–1</td>
<td>200 000</td>
<td>1–3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FTICR</td>
<td>+++</td>
<td>0.1–1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Protein Sequencing: Database Search

• Bottom-up MS

- Database search
  - Peptide Database
  - Peptide sequence
  - Match
  - Protein sequence

• Top-down MS

- Top-down spectrum
  - Deconvolution
  - Mass list
  - Database Search
  - Protein sequence
  - Protein Database
Antibodies

- An antibody is a large Y-shaped protein produced by B-cells.
- The antibody recognizes and binds to an antigen, a unique part of the foreign target.
- The variable domains of antibodies are highly mutated.
- Indispensable reagents for biomedical research and as diagnostic and therapeutic agents.
- The sequences of most antibodies are unknown.
De Novo Peptide Sequencing

• Bottom-up MS

- Tryptic Digestion
- MS/MS
- De novo sequencing

The order of peptides is missing.
- Which sequence is correct?
  - Candidate 1: DIQMR PDSSLK MCDSEFK VTITCKR
  - Candidate 2: PDSSLK DIQMR VTITCKRMCDSEFK

Available tools
- PEAKS Ma et. al. RCMS 2003
- PepNovo Frank et al. JPR 2005
- pNovo Chi et. al. JPR 2010
De Novo Protein Sequencing by Bottom-Up MS

- Multiple enzyme digestion
  - Trypsin: after residues R and K
  - GluC: after residues D and E

- Example

  Target protein (unknown): D I Q M R Q K P S D L S K S V G D R V T I T C K R S Q

  Bottom-up spectra (trypsin):

  Bottom-up spectra (GluC):

  De novo result: D I Q M R P S D L S K S V G D V T I T C K R

- Challenges
  - Overlaps may be short
  - Very short peptides

Bandeira et al. Nature Biotechnology 2008
De Novo Protein Sequencing by Top-Down MS

- Top-down tandem mass spectra cover whole proteins.

- Example

  Target protein (unknown): D I Q M R Q K P S D L S K S V G D R V T I T C K R S Q

  Top-down spectra: 

  De novo result: 

  - Missing peaks
    - Resulting sequences contain gaps
De Novo Protein Sequencing by Combining Top-Down and Bottom-Up MS (TBNovo)

• Complementary information
  – Use bottom-up spectra to fill gaps in top-down spectra
  – Use top-down spectra to find the order of bottom-up spectra

Target protein (unknown): D I Q M R Q K P S D L S K S V G D R V T I T C K R S Q

Bottom-up spectra: |---------------|---------------|---------------|

Top-down spectra: |---------------|---------------|---------------|

De novo result:   D I Q M R Q R P S D L S K S V G D R V T I T C K R S Q
Data Sets

• Light chain of alemtuzumab (MabCampath)
  – Top-down
    • Thermo LTQ Orbitrap Velos and Thermo Q-Exactive
    • ETD: 12134 spectra; CID: 7686; and HCD: 4931
  – Bottom-up
    • Thermo LTQ Orbitrap XL
    • HCD spectra
    • Trypsin: 2716 spectra, chymotrypsin: 4328, proteinase K: 1616 and pepsin: 1910

• Carbonic anhydrase 2 (CAH2 BOVIN)
  – Top-down
    • ETD: 3045; CID: 3363; HCD: 3437.
  – Bottom-up
    • Trypsin: 47536 spectra
Preprocessing

- **Prefix residue masses** corresponds to neutral b-ion masses.
- Convert all spectra to lists of candidate prefix residue masses.
- **Bottom-up spectra**
  - De novo peptide sequencing (PEAKS)
  - Represented by prefix residue masses of the peptides
- **Top-down spectra**
  - Spectral deconvolution (MS-Deconv)
  - Convert neutral masses to candidate prefix residue masses.
  - Merge multiple top-down spectra to one.

Ma et al. RCMS 2003, Liu et al. MCP 2010
Candidate Prefix Residue Masses

1. Preprocessing:

parent mass: \( M = 1111 \text{ Da} \)

Neutral mass list:
- 253 Da
- 457 Da
- 483 Da
- ...
- 569 Da

Add complementary masses
- M-253 Da
- M-457 Da
- M-483 Da
- ...
- M-596 Da

Prefix residue masses: 253, 483, M-457

Candidate prefix residue masses: 253, 457, 483, ..., 569, M-253, ..., M-596
Spectral Mapping

- Mass count score: number of prefix residue masses shared by a top-down spectrum and a bottom-up spectrum,

- Shifted bottom-up spectra: adding a shift value to each prefix residue mass

- Optimal shift: the shift that maximizes the mass count score between a top-down spectrum and a bottom-up spectrum.

- Shifted mass count score: the best mass count score between a top-down spectrum and a shift bottom-up spectrum.
Spectral Mapping

• Keep only bottom-up spectra with a shifted mass count score $\geq 7$.

• Keep only prefix residue masses supported by two bottom-up spectra or the top-down spectrum + a bottom-up spectrum

• Result: combined prefix residue mass list
Gap Filling

- Shift bottom-up spectra to possible cleavage sites.
- Map bottom-up spectra to the combined prefix residue mass list.

Combined prefix residue mass list

Shifted bottom-up spectrum

Masses used to fill the gap
**Gap Filling**

- Compute possible peptide masses
- Find bottom-up spectra with similar precursor masses

Combined prefix residue mass list

Bottom-up spectrum

Masses used to fill the gap
• Compute best shift for mapping bottom-up spectrum to the combined prefix residue masses.
• Update the list of combined prefix residue masses.
• Convert the list of prefix residue masses to a spectral graph.
• Find a heaviest path corresponding a protein sequence that best explains the experimental spectra.
Results

• Light chain of alemtuzumab (MabCampath)
  – 214 amino acids
  – Tbnovo reported 188 prefix residue masses, 184 were correct.
  – Coverage 86.9%, accuracy 97.8%

• Carbonic anhydrase 2 (CAH2 BOVIN)
  – 258 amino acids
  – Tbnovo reported 229 prefix residue masses, 194 were correct.
  – Coverage 75.2%, accuracy 84.7%
De novo sequencing result of MabCampath light chain

```
DIQMTQSPSS LSASVGDRVT ITCKASQNID KYLNWYYQKPGKAPKLI

1 TNNLQTGVPS RFSGSGSGTD FTFTISSLQP EDIATYYCLQ HISRPRTF

51 DNAEQSGNSQ ESVTEQDSKD STYSISSSTLT LSKADYEKHK VYACEVTH
32 DNAIQSGNSQ ESVTEQDSKD STYSISSSTLT ISQADYEKHO VYACEVTH

31 LSSPVTKSFN RGEC 214
32 ISSPVTKS[456.04] 191
```
Software Tools

- **TBNovo**: Protein sequencing by combing top-down and bottom-up tandem mass spectra.

- **MS-Deconv**: Top-down spectral deconvolution.

- **MS-Align+/TopPIC**: Protein identification by top-down tandem mass spectra.

- [http://mypage.iu.edu/~xwliu/](http://mypage.iu.edu/~xwliu/)
Acknowledgements

UCSD

Pavel A. Pevzner

Erasmus MC, Netherlands

Lennard J. M. Dekker,
Martijn M. Vanduijn
Theo M. Luider

PNNL

Si Wu

Saint Petersburg Academic University

Mikhail Dvorkin
Sonya Alexandrova
Kira Vyatkina

Ljiljana Paša-Tolić

Nikola Tolić