Deconvolution and Database Search of Complex Top-Down Tandem Mass Spectra

Xiaowen Liu

Department of Computer Science and Engineering, University of California, San Diego
Top-down/Middle-down vs bottom-up

Bottom-up:
- Protein
- Digestion → Peptides 5-30aa
- MS/MS

Middle-down:
- Protein
- Digestion with enzymes targeting less abundant amino acids → Long peptides 3K-20K Da
- MS/MS

Top-down:
- Protein 3K-50K Da
- No digestion and MS/MS
Top-down MS

Top Down MS

upfront protein separation

single protein or moderately complex mixture

static infusion ESI

high accuracy precursor ion measurement

on-line LC-MS

targeted MS/MS

untargeted data-dependent MS/MS

Protein identification and characterization

Garcia BA, JASMS 2010
Middle-down MS

GluC AspN
Why top-down/middle-down is important

• Identify multiple protein species (proteolytically processed protein species).

• Localize multiple PTMs in a coordinated fashion (e.g., combinatorial PTM code).

• 100% protein sequence determination.

http://proteomics.ucsd.edu
Long peptides/proteins vs short peptides

• Identify proteolytically processed protein species

Long peptides:

MGFFDQCLIKTTGSSKAS

Short peptides:

MGFFDQCLIKTTGSSKAS

• Identify multiple PTMs in a coordinated fashion

Long peptides:

AGFMDQCLIKTGMSKASMSHGTTHSAKSQPSMA DPQGSGCFD

AGFMDQCLIKTGMSKASMSHGTTHSAKSQPSMA DPQGSGCFD

Short peptides:

AGFMDQCLIKTGMSHGTTHSAK

http://proteomics.ucsd.edu
Several groups have identified over 100 proteins using top-down MS.

Protein separation:
- Free-flow electrophoresis (FFE)
- Hydrophilic interaction liquid chromatography (HILIC)

Mass spectrometer:
- High resolution mass spectrometer: FT-ICR, Orbitrap
- ECD/ETD fragmentation: better performance for high charge state long peptides than CID

Kellie et al. Molecular BioSystems 2010

http://proteomics.ucsd.edu
## Mass spectrometers for top-down MS

<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</td>
</tr>
<tr>
<td>TOF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ETD</td>
<td>ISD</td>
</tr>
<tr>
<td>TOF-TOF Q-TOF</td>
<td>+</td>
<td>&lt;0.01</td>
<td>10000</td>
<td>5–20</td>
<td></td>
<td>PSD</td>
</tr>
<tr>
<td>FT</td>
<td></td>
<td></td>
<td>60000</td>
<td>3–10</td>
<td>CID</td>
<td>ETD HCD CID ECD IRMPD</td>
</tr>
<tr>
<td>Orbitrap</td>
<td>++</td>
<td>0.1–1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FTICR</td>
<td>++</td>
<td>0.1–1</td>
<td>200000</td>
<td>1–3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Kellie et al. Molecular BioSystems 2010

[http://proteomics.ucsd.edu](http://proteomics.ucsd.edu)
Database search for top-down/middle-down tandem mass spectra
Database search of high resolution tandem mass spectra of long peptides/proteins

Preprocessing → Deconvolved mass list → Compare

Best scoring protein (may contain PTMs)

Protein database
- MASKAV...
- PACSKCA...
- MPASGH...
- MSTVGKBC...

Top-down spectrum

http://proteomics.ucsd.edu
Computational challenges for identifying long peptides

- **Deconvolution algorithms generate many false positives.** Many deconvoluted peaks have +1/-1 Da errors, the issue that remain poorly addressed in all deconvolution tools. In complex spectra, up to 50% of deconvoluted peaks may represent false positives.

- **Most long peptides have multiple modifications.** For example, common N-terminal modifications alone include very frequent NME, acetylation, signal peptide removal, etc. **PTMs may be unexpected requiring blind PTM searches.**

- **Speed becomes a bottleneck** when searching for multiple PTMs, particularly in blind mode.

- **Statistical methods** for evaluating Protein-Spectrum-Matches (particularly with PTMs!) remain poorly developed. **Noncalibrated p-values reduce the number of reliably identified Protein-Spectrum Matches.**
Top-down spectral deconvolution

http://proteomics.ucsd.edu
Profile and centroid data

Profile data

Centroid data

Peak picking

Yan et al. BMC Bioinformatics 2009

http://proteomics.ucsd.edu
An example of top-down tandem mass spectra

Top-down spectra usually have order(s) of magnitude more peaks and complex pattern of isotopomer envelopes.

http://proteomics.ucsd.edu
Isotopomer envelopes

Isotopes

\[ \text{C}^{12} \quad \text{Mass: 12.000} \quad \text{Freq: 98.93\%} \]

\[ \text{C}^{13} \quad \text{Mass: 13.003} \quad \text{Freq: 1.07\%} \]

10-carbon molecules: the proportion of molecules is with all 10 carbons being \( \text{C}^{12} \) is \( 0.9893^{10} \approx 89.8\% \). The proportion of molecules with 9 carbons being \( \text{C}^{12} \) and 1 carbon being \( \text{C}^{13} \) is \( 10 \cdot 0.9893^9 \cdot 0.0107 \approx 9.7\% \).

100-carbon molecules: the proportion of molecules is with all 100 carbons being \( \text{C}^{12} \) is \( 0.9893^{100} \approx 31.54\% \). The proportion of molecules with 99 carbons being \( \text{C}^{12} \) and 1 carbon being \( \text{C}^{13} \) is \( 100 \cdot 0.9893^{99} \cdot 0.0107 \approx 34.1\% \).
Isotopomer envelopes

Theoretical envelope for Lysozyme (14303.88 Da)

The monoisotopic mass (14303.88 Da for Lysozyme) of a chemical formula is the sum of the masses of the atoms using the most abundant isotope for each element.

http://proteomics.ucsd.edu
Spectral deconvolution problem

- Top-down spectra are characterized by complex envelopes that often overlap and may share peaks.
- Spectral deconvolution focuses on grouping spectral peaks into isotopomer envelopes, and deriving a list of monoisotopic masses.

```
spectrum  Envelope detection  Candidate envelopes  Envelope selection  Output Envelopes  Monoisotopic mass extraction  Mass list
```

http://proteomics.ucsd.edu
Existing tools for spectral deconvolution

- Spectral deconvolution is a complex problem first addressed by Horn et al. in Thrash [Horn et al., JASMS 2000]
- **Xtract** [Zabrouskov et al., JASMS 2005]
  Thrash and Xtract are part of ProsightPC
- **Decon2LS** [Jaitly et al., BMC Bioinformatics, 2009]
- **RAPID** [Park et al., Anal. Chem. 2008]
- **Hardklör** [Hoopmann et al., Anal. Chem. 2007]
- **MS-Deconv** [Liu et al. MCP, 2010, UCSD]

http://proteomics.ucsd.edu
Envelope detection

1. Generate theoretical envelope with a specific charge
2. Find real peaks in the envelope
3. Scale intensities of the peaks in the theoretical envelope
4. Scoring function

Score of the envelope

http://proteomics.ucsd.edu
Generating theoretical envelopes


Real envelope:

Mass: 4000 Da  
Charge: 4

Estimate average composition

\[ \text{C}_{154}\text{H}_{306}\text{N}_{42}\text{O}_{77}\text{S} \]
Charge: 4

Emass

Theoretical envelope:

Senko et al., JASMS 1995
Rockwood and Haimi, JASMS 2006

http://proteomics.ucsd.edu
Scoring functions

- **Thrash**: Squared error between the intensities of theoretical and real peaks.
  \[
  \text{Figure-of-merit} = \frac{\text{Number of comparisons}}{\sum [(A_{n} - N_{p})^2 + (N_{V})^2]}
  \]

- **Rapid**: intensity ratio between neighboring peak pairs
  \[
  \text{Score} = \sum_{k=0}^{\text{#-2}} \text{scoreR}(k, p, m) + \sum_{k=0}^{\text{#-3}} \text{scoreRP}(k, p, m), \quad 0 \leq p \leq 3
  \]

- **Hardklör**: dot product

- **MS-Deconv**: intensity similarity and m/z similarity.
  \[
  \text{sim}(p, \tilde{p}) = \sqrt{h(\tilde{p})} \cdot s_x(p, \tilde{p}) \cdot s_i(p, \tilde{p})
  \]

Real envelope:

```
  m/z 1001 1002
```

Theoretical envelope:

```
  m/z 1001 1002
```

http://proteomics.ucsd.edu
Previous deconvolution algorithms were greedy: selection of the highest-scoring envelope and removal of all envelopes it conflicts with (e.g. sharing peaks).

We propose a dynamic programming approach for envelope selection.
Dynamic programming algorithm

Step 1. Divide m/z axis into intervals

Span of $E_1 = \{p_1, p_2, p_3, p_4\}$

Span of $E_2 = \{p_5, p_6, p_7, p_8\}$

Span of $E_3 = \{p_9, p_{10}, p_{11}\}$

http://proteomics.ucsd.edu
Dynamic programming algorithm

Step 2. Represent the set of envelopes as a graph. Each interval will correspond to one or more vertices in the graph.

Connect vertices \([I, x]\) and \([I+1, x']\) separated by a start/end of an envelope \(A\) if

1. \(X = X'\); 2. \(A\) is the start of \(E\) and \(X' = X \cup \{E\}\); 3. \(A\) is the end of \(E\) and \(X' = X - \{E\}\).

The weight of a vertex \([I, x]\) is the score of envelope \(E\) if \(I\) is the first interval of \(E\) and \(E \in X\). Otherwise, the weight is zero.

http://proteomics.ucsd.edu
Dynamic programming algorithm

Step 3. Find the heaviest path in the graph

http://proteomics.ucsd.edu
Peak sharing model

A peak can be shared by two or more envelopes

MS-Deconv has a more complex peak sharing model.

Liu et al. MCP 2010

http://proteomics.ucsd.edu
Time complexity

• How large are DP graphs?
  • A graph for a spectrum from Bacteriorhodopsin contains 19,219 nodes and 61,927 edges.

• Time Complexity $O(n \cdot 2^m)$
  • $n$: the number of envelopes
  • $m$: the maximum number of envelopes covering an interval (usually $m \leq 10$)
Deconvolution of complex spectra

• Six CAD spectra (FT-ICR) of intact proteins
  • Three from Bacteriorhodopsin (BR)
    • 248 amino acids, 26766.12 Dalton
    • Charge states 10, 11 and 16.
  • Three from Apolipoprotein A-I (Apo-A1)
    • 241 amino acids, 27586.22 Dalton
    • Charge states 23, 25, and 26.

• Desktop PC with a 2.2G CPU and 3.0G RAM
MS-Deconv vs. Thrash

- **Average running time**
  - MS-Deconv 9 seconds per spectrum
  - Thrash 302 second per spectrum

- **Force MS-Deconv and Thrash to output the same number monoisotopic peaks** (using $r$-value parameter in Thrash)

- **True positive masses**
  - Generate theoretical monoisotopic mass list (from known protein sequence using theoretical b and y-ions).
  - Match recovered masses (by MS-Deconv and Thrash) to theoretical masses
  - PPM $\leq$ 3, 5, 10
  - Report the numbers of true positive masses

http://proteomics.ucsd.edu
# MS-Deconv vs Thrash: True positive masses

<table>
<thead>
<tr>
<th>Sp.</th>
<th>Protein</th>
<th>Charge</th>
<th>$r$-value</th>
<th>Number of output envelopes</th>
<th>Number of true positive masses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Thrash/MS-Deconv</td>
</tr>
<tr>
<td>1</td>
<td>BR</td>
<td>10</td>
<td>0.9999</td>
<td>150</td>
<td>70/107 63/106 62/105</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.9</td>
<td>272</td>
<td>131/147 114/143 110/140</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.5</td>
<td>342</td>
<td>151/166 127/156 122/152</td>
</tr>
<tr>
<td>2</td>
<td>BR</td>
<td>11</td>
<td>0.9999</td>
<td>238</td>
<td>108/184 107/184 89/178</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.9</td>
<td>597</td>
<td>259/312 244/300 204/278</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.5</td>
<td>734</td>
<td>306/335 285/316 236/290</td>
</tr>
<tr>
<td>3</td>
<td>BR</td>
<td>16</td>
<td>0.9999</td>
<td>41</td>
<td>23/27 21/21 21/21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.9</td>
<td>106</td>
<td>59/66 49/51 44/50</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.5</td>
<td>159</td>
<td>83/75 63/58 58/57</td>
</tr>
<tr>
<td>4</td>
<td>Apo-AI</td>
<td>23</td>
<td>0.9999</td>
<td>77</td>
<td>46/44 45/43 37/40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.9</td>
<td>180</td>
<td>95/97 90/94 73/83</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.5</td>
<td>269</td>
<td>128/119 116/114 95/103</td>
</tr>
<tr>
<td>5</td>
<td>Apo-AI</td>
<td>25</td>
<td>0.9999</td>
<td>55</td>
<td>31/45 31/44 26/42</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.9</td>
<td>132</td>
<td>68/70 63/66 55/64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.5</td>
<td>223</td>
<td>89/96 82/90 71/85</td>
</tr>
<tr>
<td>6</td>
<td>Apo-AI</td>
<td>26</td>
<td>0.9999</td>
<td>38</td>
<td>24/30 24/29 22/28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.9</td>
<td>75</td>
<td>41/47 38/43 31/40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.5</td>
<td>143</td>
<td>57/65 52/59 43/54</td>
</tr>
</tbody>
</table>

http://proteomics.ucsd.edu
MS-Deconv vs. Thrash: True positive masses

Deconvolution results of Spectrum from BR with charge 10

Number of true positive masses vs. Number of output masses

Percentage of true positive masses vs. Number of output masses

http://proteomics.ucsd.edu
Annotated protein sequence
Bacteriorhodopsin

Q A Q I T G R|P E W I W L A L G T A L M G L G T L Y F L V K
G M|G V S D|P D|A K K F Y A I T T L|V|P A I A F T|M Y L S M
L L|G Y|G|L|T M V|P F G G E Q N P I Y W A R Y A D|W L F T T|
P L L L D|L|A|L|L|V D A|D|Q G T|I|L|A|L V|G A D|G I M I G
T G L V G A L T K V Y S Y R|F V|W W A I|S T A A M|L Y I L Y
V L F F G F T S K A E S M R P E V A S T|F K V L R N V T V V
L D V S A K V|G F G I L|L R S R A I F G E A E A P E|P S A|
G D|G A A|A T S

Number of envelopes supporting b-ion

Number of envelopes supporting y-ion
Availability

proteomics.ucsd.edu/Software.html

Spectral Networks

Main Page, Download, Copyright Notice, Documentation
Spectral networks are based on the idea of performing an MS/MS database search without comparing a spectrum against a database. Spectral neworks capitalize on spectral pairs, which allow for the identification of prefix and suffix ladders and greatly reduce noise.

MS Top Down

Download, Copyright Notice
Recent advances in mass spectrometry instrumentation, such as FT-ICR and Orbitrap, have made it possible to generate high resolution spectra of entire proteins. While these methods offer new opportunities for performing "top-down" studies of proteins, the computational tools for analyzing top-down data are still scarce. MS-TopDown is a new algorithm for sequencing such data. It implements a version of the Spectral Alignment algorithm specially suited for the problem of identifying protein forms in top-down mass spectra (i.e., identifying the modifications, mutations, insertions and deletions). MS-TopDown can efficiently discover protein forms even in the presence of numerous modifications, and it can also recover positional isoforms from spectra of mixtures of isobaric protein forms.

MS-Deconv

Main Page
MS-Deconv is a software tool for top-down spectral deconvolution. MS-Deconv uses a combinatorial algorithm. The algorithm first generates a large set of candidate isotopomer envelopes for a spectrum, then represents the spectrum as a graph, and finally selects its highest scoring subset of envelopes as the heaviest path in the graph. In contrast with other approaches, the algorithm scores sets of envelopes rather than individual envelopes.

http://proteomics.ucsd.edu
Top-down database search

ProSightPTM  [Zamdborg et al., Nucleic Acids Res., 2007]
PIITA  [Tsai et al., JASMS., 2009]
USTag  [Shen et al., Anal. Chem., 2008, PNNL]
MS-TopDown  [Frank et al., Anal. Chem., 2008, UCSD]
MS-Align+  [Liu et al. submitted, UCSD]
## Existing software tools for Precision top-down mass spectrometry

<table>
<thead>
<tr>
<th>Software</th>
<th>Identification of unexpected mods</th>
<th>Search 6-frame translation</th>
<th>Speed</th>
<th>Estimation of statistical significance</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>ProSightPTM</td>
<td>No</td>
<td>Yes</td>
<td>Fast</td>
<td>Yes (inaccurate)</td>
<td>Commercial</td>
</tr>
<tr>
<td>PIITA</td>
<td>Yes/No</td>
<td>No</td>
<td>Fast</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>USTag</td>
<td>Yes</td>
<td>Yes</td>
<td>Fast</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>MS-TopDown</td>
<td>Yes</td>
<td>No</td>
<td>Slow</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>MS-Align+</td>
<td>Yes</td>
<td>Yes</td>
<td>Fast</td>
<td>Yes (accurate)</td>
<td>Yes</td>
</tr>
</tbody>
</table>

[http://proteomics.ucsd.edu](http://proteomics.ucsd.edu)
“Shotgun annotated” protein database (ProSightPTM)

• Generate all know protein species with PTMs
  
  • Short peptide with 3 Serine residues and 2 phosphorylated sites (at unknown positions)
    
    • 3 variants
      
      ABSCDSEFS
      ABSCDSEFS
      ABSCDSEFS

  • Long peptide with 12 Serine residues and 6 phosphorylation sites (at unknown positions)
    
    • 924 variants
      
      ABSCDSEFSGPASQDSTFSGMBSIDSLFSGWBSADSGFSG
      ABSCDSEFSGPASQDSTFSGMBSIDSLFSGWBSADSGFSG
      ....
      ABSCDSEFSGPASQDSTFSGMBSIDSLFSGWBSADSGFSG

• Explicitly adding all these configurations to a protein sequence database results in a combinatorial explosion

http://proteomics.ucsd.edu
Procuror ion independent algorithm (PIITA)

Experimental MS2 spectrum
Deconvoluted MS2 spectrum

\[ \Delta S_c = \frac{(FIM_{1st} - FIM_{2nd})}{FIT} \]

- \( FIM_{1st} \): the highest fragment ion match score among all proteins.
- \( FIM_{2nd} \): the 2\textsuperscript{nd} highest fragment ion match score among all proteins.
- \( FIT \): the total number of deconvoluted fragment ions

http://proteomics.ucsd.edu
Long tags (US-tags, ProSight sequence tag)

Shen et al. JPR 2010
1. Preprocessing:

**Parent mass:** \( M = 1225 \text{ Da} \)

**Fragment mass list:**
- 253 Da
- 457 Da
- 483 Da
- ...
- 683 Da

**Prefix residue mass (PRM) list:**
- \( M - 253 \text{ Da} \)
- \( M - 457 \text{ Da} \)
- \( M - 483 \text{ Da} \)
- ...
- \( M - 683 \text{ Da} \)
Protein-Spectrum Matches = Diagonal Paths in the Protein-Spectrum Grid

PRM Spectrum: \{0, 97, 253, 484, 569, 683, 784, 941, 1054, 1168, 1225\}
Protein: PRTEINSTRING

- **Spectral Alignment**
  Spectral alignment with \( F \) modifications is a diagonal path from the top left node to the bottom right node with at most \( F \) breaks.

- **Spectral alignment score**
  Number of 2-D points \((a_i, b_j)\) that the path passes through.

http://proteomics.ucsd.edu
How to find the best-scoring alignment?

Spectrum: {0, 97, 253, 354, 483, 596, 683, 784, 1054, 1168, 1225}
Protein: {0, 97, 253, 354, 483, 596, 790, 797, 898, 1168, 1272, 1339}

Spectral alignment between a protein and a spectrum has time complexity:

\[ O(\text{ProteinLength} \times \#\text{Peaks}) \]

with a rather large constant hidden in “O.” This is too slow for searching entire proteomes.


http://proteomics.ucsd.edu
Speed up

- **Diagonal score**: the maximum number of matched mass pairs on a single diagonal line (among all diagonals).
- Time complexity $O(nm)$.
Speed up

• Diagonal score as a filter

PRM spectrum: \{0, 97, 253, 484, 569, 683, 784, 941, 1054, 1168, 1225\}

<table>
<thead>
<tr>
<th>Protein database</th>
<th>diagonal score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRTEINSTRING</td>
<td>5</td>
</tr>
<tr>
<td>RPOS........</td>
<td>1</td>
</tr>
<tr>
<td>GIRNS........</td>
<td>1</td>
</tr>
</tbody>
</table>

• Calculate the diagonal score for each protein in DB
• Keep only top 20 best-scoring proteins for spectral alignment.
• Time complexity $O(Cnm) \rightarrow O(nm)$

http://proteomics.ucsd.edu
Statistical significance estimation

- **ProSightPTM: Possion distribution**

\[ p(n) = 1 - \sum_{i=0}^{n-1} \frac{e^{-xf}(xf)^i}{i!} \]

where:

- \( n \) is the number of matching fragments.
- \( x \) is the probability of an observed fragment ion matching a random theoretical fragment ion by chance.
- \( f \) is the total number of fragment ions observed.

[http://proteomics.ucsd.edu](http://proteomics.ucsd.edu)
Statistical significance estimation

MS-Align+: generating function approach

Amino acids
A: mass 2
B: mass 3

NodeScore: 0 1 1 0 1 0 1 0 0 0

<table>
<thead>
<tr>
<th>NodeScore</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score=0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Score=1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Score=2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Score=3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

- Compute the score distribution of all peptides.
- Each node stores a score distribution instead of a maximum score.
- MS-GeneratingFunction (MS-GF)

Kim et al. JPR 2008

http://proteomics.ucsd.edu
Precision top-down MS: from spectra of single proteins (2007-2009) to large datasets of endogenous peptides/proteins (2010-2011)

- *Saccharomyces cerevisiae* dataset from PNNL (*Yeast dataset*)
  - 30,760 LTQ-Orbitrap spectra
- *Salmonella typhimurium* dataset from U. of Wash. (*Salmonella dataset*)
  - 14,041 LTQ-Orbitrap (with gas phase fractionation) spectra

- Charge states: up to 30,
- Precursor masses: up to 20 KDa
- *Thousands spectra and protein species identified from a single run!*
- **DOES PRECISION MS APPROACH THE POWER OF BOTTOM UP MS WHILE PROVIDING NEW APPLICATIONS?**
MS-Align+ generates reliable identifications

- Target and decoy database approach for estimating FDR
- Yeast identifications (> 2500 Da)

<table>
<thead>
<tr>
<th>E-value</th>
<th>Target database</th>
<th>Decoy database</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#spectra</td>
<td>#species</td>
</tr>
<tr>
<td>≤ 0.0001</td>
<td>2.455</td>
<td>1.307</td>
</tr>
<tr>
<td>≤ 0.001</td>
<td>2.751</td>
<td>1.442</td>
</tr>
<tr>
<td>≤ 0.01</td>
<td>3.255</td>
<td>1.803</td>
</tr>
</tbody>
</table>

- Salmonella identifications (> 2500 Da)

<table>
<thead>
<tr>
<th>E-value</th>
<th>Target database</th>
<th>Decoy database</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#spectra</td>
<td>#species</td>
</tr>
<tr>
<td>≤ 0.0001</td>
<td>1.400</td>
<td>241</td>
</tr>
<tr>
<td>≤ 0.001</td>
<td>1.667</td>
<td>300</td>
</tr>
<tr>
<td>≤ 0.01</td>
<td>1.915</td>
<td>440</td>
</tr>
</tbody>
</table>

- Surprising result: top-down approach identified almost as many proteins as bottom up for Salmonella dataset.

http://proteomics.ucsd.edu
Comparison with PIITA (Salmonella dataset)

- **MS-Deconv** and **MS-Align+** vs. **Hardklör** and **PIITA** (Univ. of Washington)

<table>
<thead>
<tr>
<th>E-value</th>
<th>MS-Align+</th>
<th>PIITA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># proteins identified by MS-Align+ only</td>
<td># proteins identified by PIITA only</td>
</tr>
<tr>
<td>≤ 0.0001</td>
<td>165</td>
<td>35</td>
</tr>
<tr>
<td>≤ 0.001</td>
<td>183</td>
<td>49</td>
</tr>
<tr>
<td>≤ 0.01</td>
<td>286</td>
<td>148</td>
</tr>
</tbody>
</table>

Hoopmann et al., Anal. Chem. 2007, Tsai et al., JASMS, 2009

[http://proteomics.ucsd.edu](http://proteomics.ucsd.edu)
Comparison with ProSightPTM (Salmonella dataset)

- MSDeconv and MS-Align+ vs. Thrash and ProSightPTM (Northwest Univ.)

<table>
<thead>
<tr>
<th>E-value</th>
<th>MS-Align+ # proteins</th>
<th>ProSightPTM # proteins identified by only</th>
<th>ProSightPTM # proteins identified by only</th>
<th>Both approaches # proteins identified by</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 0.0001</td>
<td>165</td>
<td>86</td>
<td>83</td>
<td>79</td>
</tr>
<tr>
<td>≤ 0.001</td>
<td>183</td>
<td>104</td>
<td>83</td>
<td>79</td>
</tr>
<tr>
<td>≤ 0.01</td>
<td>286</td>
<td>206</td>
<td>83</td>
<td>80</td>
</tr>
</tbody>
</table>

Both PIITA and MS-Align+ significantly improve on ProsightPTM (roughly doubling the number of IDs)


http://proteomics.ucsd.edu
Example: proteins with short truncated prefixes

- Proteins with short truncated prefixes (15-35 aa)
- MS-Align+ identified 25 proteins with short truncated prefix in S. typhimurium dataset (potential signal peptides)
  - 15 proteins have the canonical `AXA` signal peptide motif
  - 10 proteins do not have `AXA` motif
  - but 9 truncated prefixes are the same as the signal peptides predicted by SignalP (consistent with SignalP predictions)
- 50S ribosomal subunit protein L32 (correcting SignalP prediction)

SignalP prediction

http://proteomics.ucsd.edu
Example: correcting gene annotations

- Protein species starting with methionine or preceding by a methionine
- Salmonella: 12 proteins

<table>
<thead>
<tr>
<th>Protein</th>
<th>Length of truncated prefix</th>
<th>First 10 amino acids</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP_459993.1 hypothetical protein STM018[Phage Gifsy-2]</td>
<td>7</td>
<td>-MNYEDKLMYQRMIPEEG</td>
</tr>
<tr>
<td>NP_459378.1 putative cytoplasmic protein</td>
<td>2</td>
<td>-MVLMKNLIAELL</td>
</tr>
<tr>
<td>NP_461253.1 putative inner membrane protein</td>
<td>3</td>
<td>-MRMSYQFGESRVD</td>
</tr>
</tbody>
</table>

- Gene NP_462331.2 has four more amino acids “AKQS” compared to the sequence in database (identified by searching against 6-frame translation database)
Mods found by MS-Align+ (in blind mode) make sense

- **Yeast**
  - Oxidation
  - +183 Da AEBSF (serine protease inhibitor)
  - +38 Da (one possible explanation is a replacement of proton by potassium)
  - many others

- **Salmonella**
  - Oxidation
  - Methylation
  - +152 Da Carboxidomethyl DTT
  - -116 Da disulfide bond compared to cysteines with carboxidomethylation.

- **Gene NP_462977.1**, 50S ribosomal subunit protein L31 with 152 Da modification

  No internal modification: poor fragmentation pattern

  One internal modification: rich fragmentation pattern

  ![Fragmentation Pattern]

http://proteomics.ucsd.edu
Availability

• Release MS-Align+ web server in two months.

http://proteomics.ucsd.edu
Open problems

• Speed up the protein identification using faster and more sensitive protein database filtering methods.

• Identify multiple PTMs using limited number of matched mass pairs and localize PTMs.

http://proteomics.ucsd.edu
Acknowledgements

Vineet Bafna
Pavel A. Pevzner
Yuval Inbar
Phil Bouchard
Yufeng Shen
Gordon A. Anderson
Richard D. Smith
Yihhsuan Tsai
Ting Ying
Dave Goodlett
Nathan Edwards
Colin Wynne
Pieter C. Dorrestein
Puneet Souda
Julian P. Whitelegge

NIH P-41-RR024851, S10 RR023045.

http://proteomics.ucsd.edu