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

Xiaowen Liu

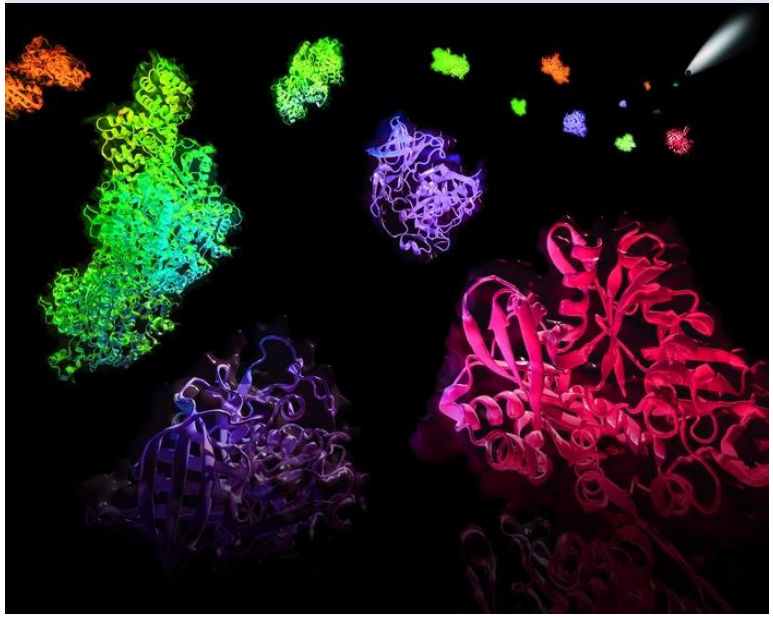
Department of BioHealth Informatics, Department of Computer and Information Sciences, Indiana University-Purdue University Indianapolis

Center for Computational Biology and Bioinformatics, Indiana University School of Medicine

SCHOOL OF INFORMATICS AND COMPUTING
INDIANA UNIVERSITY
IUPUI

Top-Down Proteomics Becomes Reality

C&EN
CHEMICAL & ENGINEERING NEWS



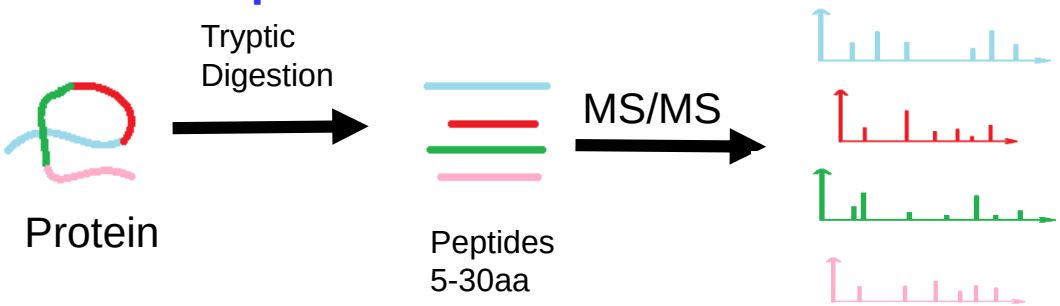
"Early proteomics methods **used enzymes to digest proteins into pieces** that could be easily analyzed by mass spectrometry. Those methods are now mature and routinely detect peptides from thousands of proteins in a single run…

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…

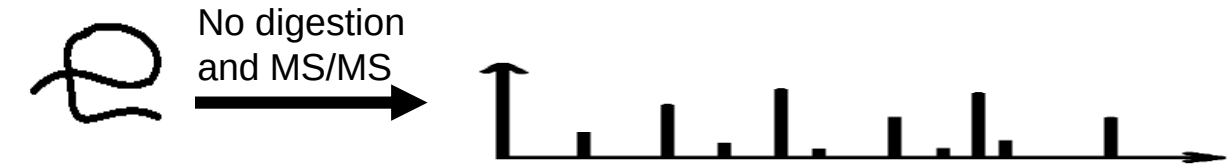
By starting with intact proteins, rather than their pieces, **top-down analysis more accurately reflects the structure and properties of actual biological systems** than does bottom-up proteomics…."

Top-Down vs. Bottom-Up MS

Bottom-up:



Top-down:



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

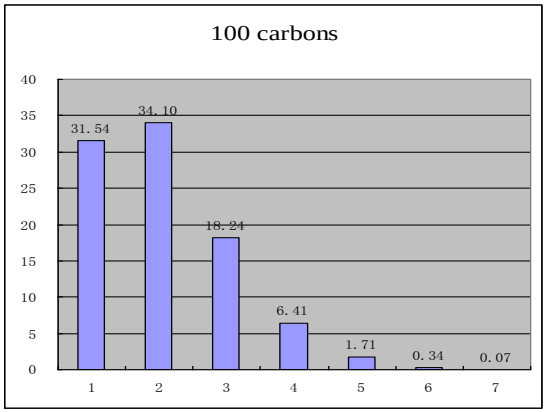
- Isotopes

C^{12}: Mass: 12.000, frequency: 98.93%

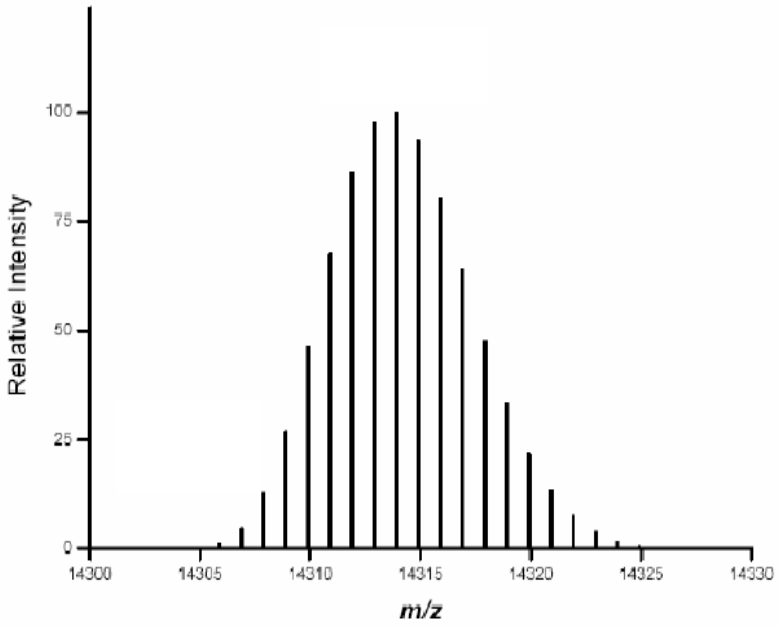
C^{13}: Mass: 13.003, frequency: 1.07%

•100-carbon molecules

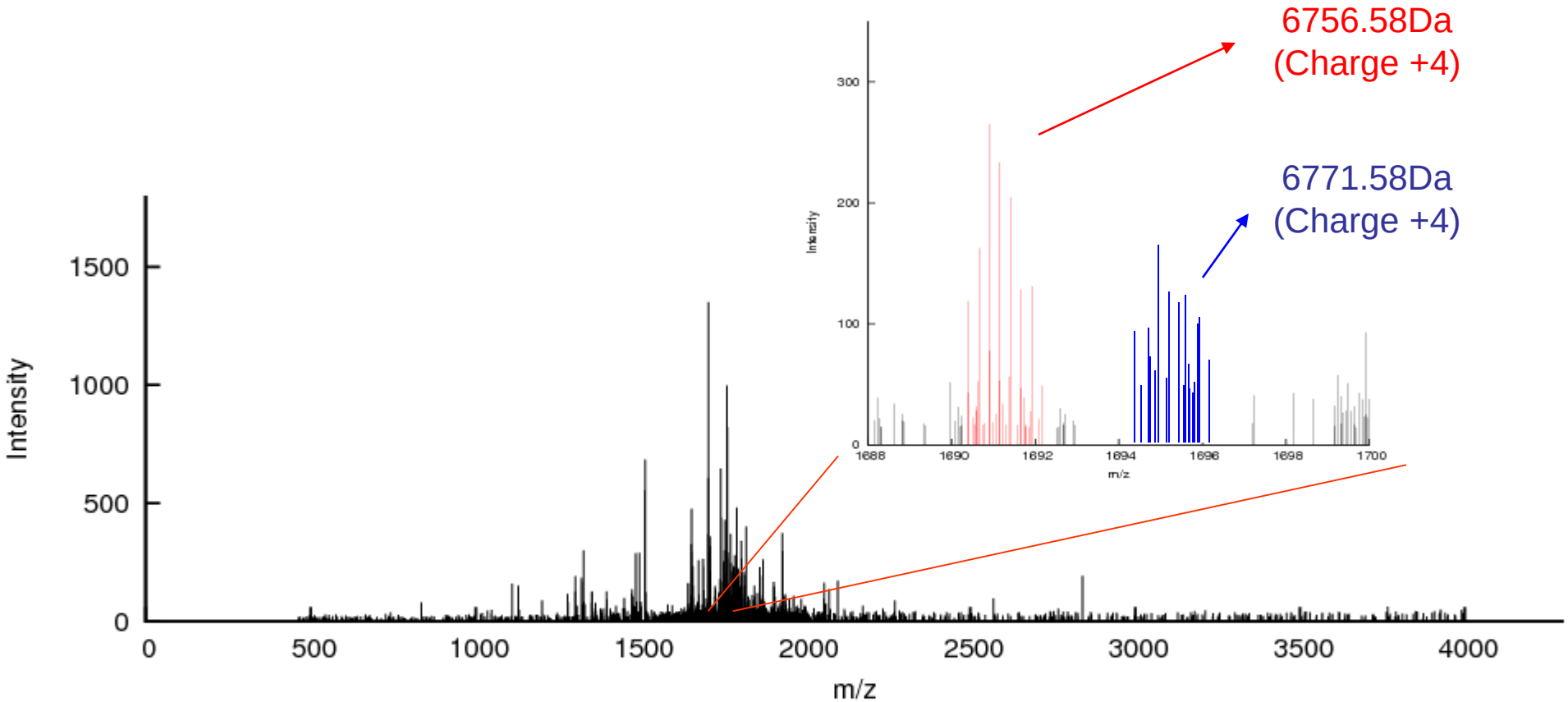
The proportion of molecules is with all 100 carbons being C^{12} is $0.9893^{100} \approx$ 31.54%



Theoretical isotopomer envelope for Lysozyme (14303.88 Da)



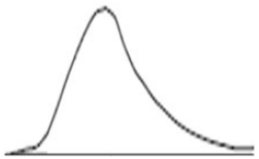
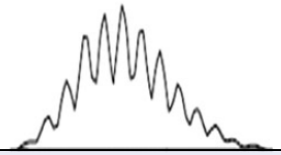
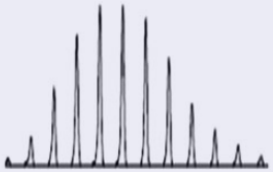
An Example of Top-Down Mass Spectra



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

Why Top-Down Proteomics Becomes Reality?

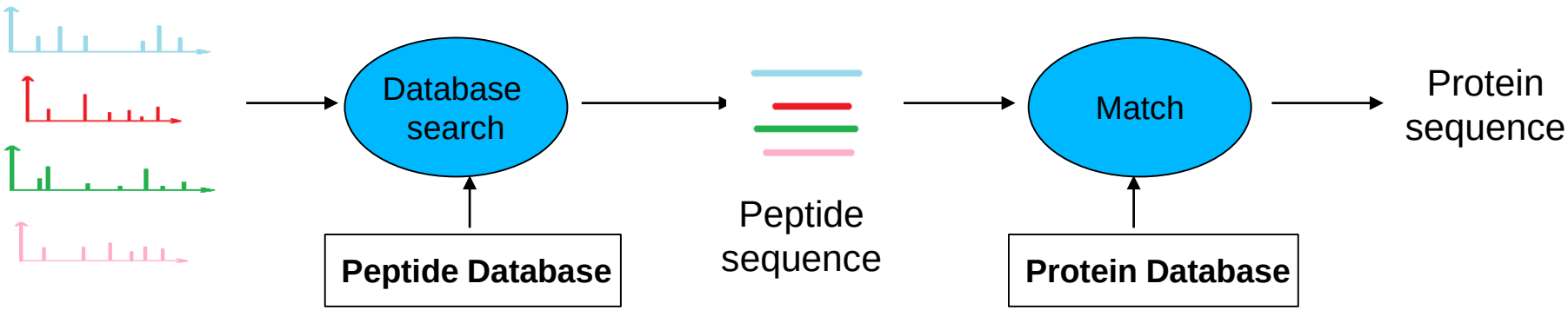
- High accuracy, high resolution, and high-throughput mass spectrometers: Orbitrap, FTICR.

Mass analyzer	Suitable for Top Down	Spectral acquisition time/s	Resolution/Da	Mass accuracy (ppm)	Performance at 8 kDa	Available fragmentation
Ion trap	+	0.05–0.3	1000	100–200		CID ETD
TOF						
TOF-TOF Q-TOF	+ +	<0.01	10 000	5–20		CID ISD PSD
FT						
Orbitrap	+ + +	0.1–1	60 000	3–10		CID ETD HCD
FTICR	+ + +	0.1–1	200 000	1–3		CID ECD IRMPD

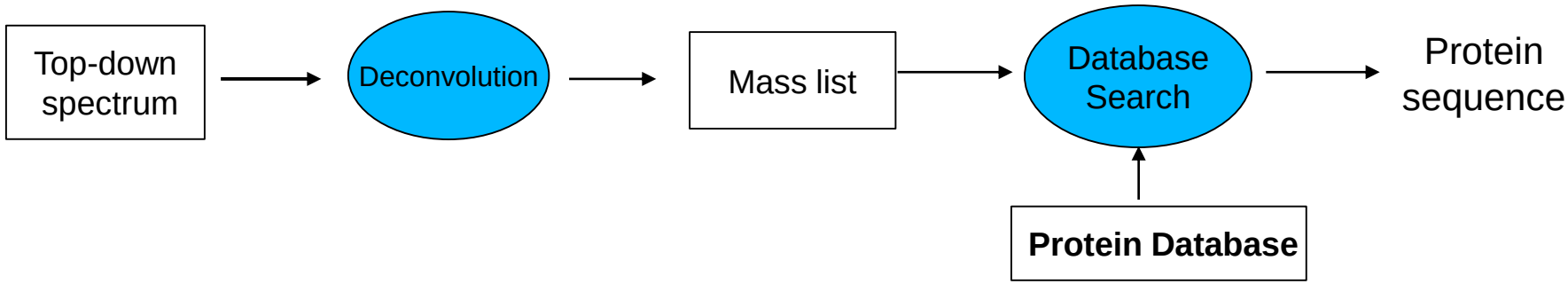
Kellie et al. Molecular BioSystems 2010

Protein Sequencing: Database Search

- Bottom-up MS

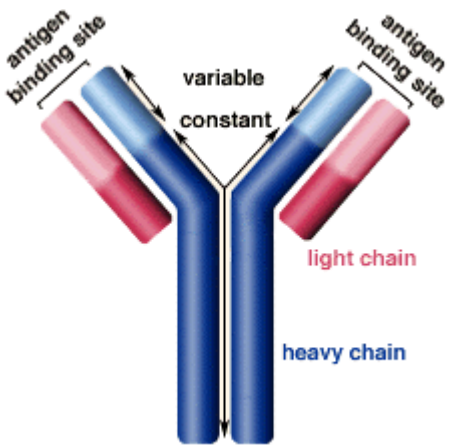


- Top-down MS



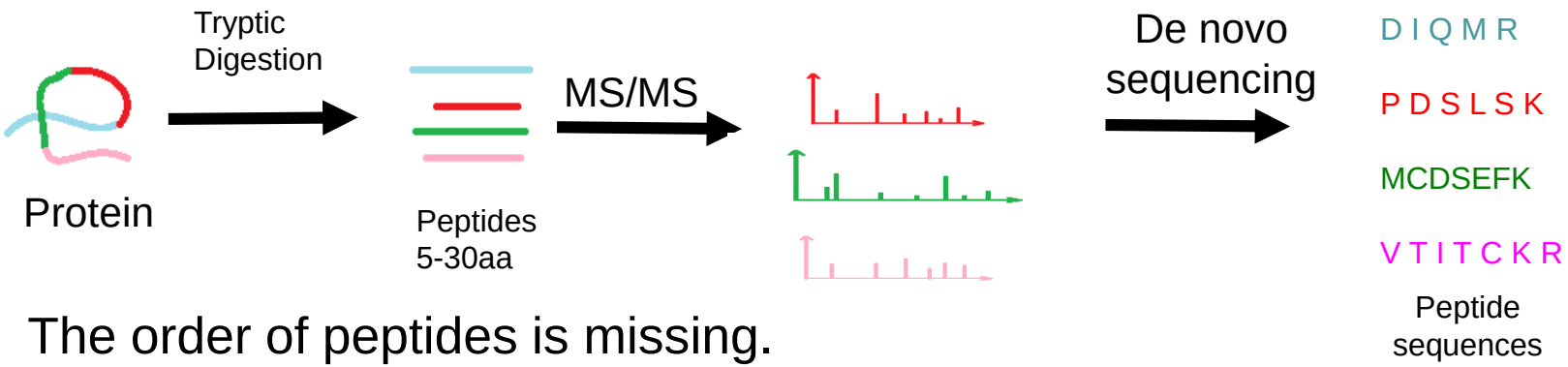
Antibodies

- 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



- The order of peptides is missing.
  - Which sequence is correct?

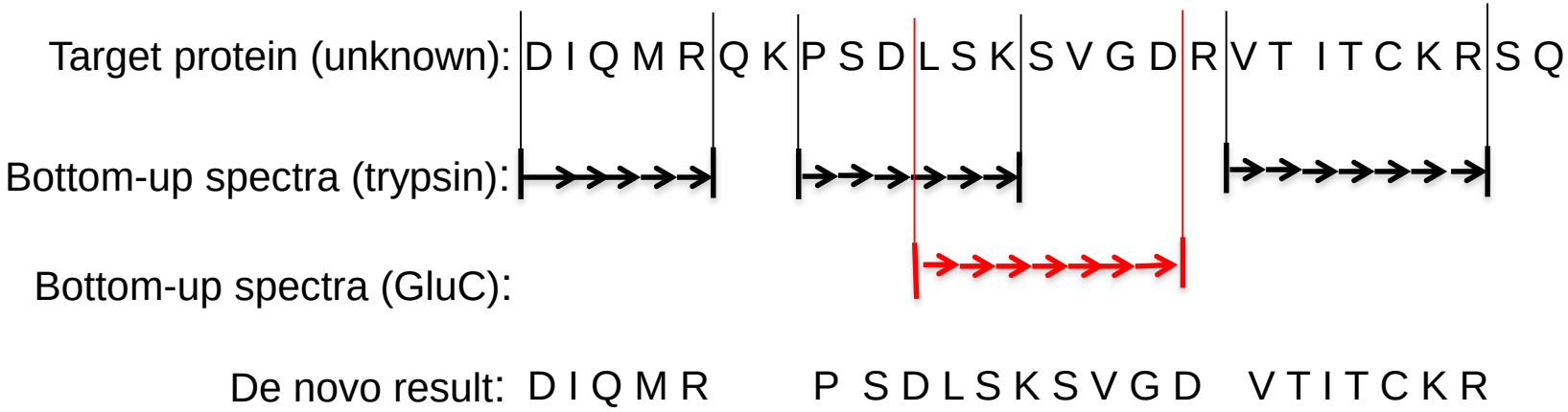
    Candidate 1: D I Q M R P D S L S K MCDSEFK V T I T C K R

    Candidate 2: P D S L S K D I Q M R V T I T C K R MCDSEFK

- 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



- 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

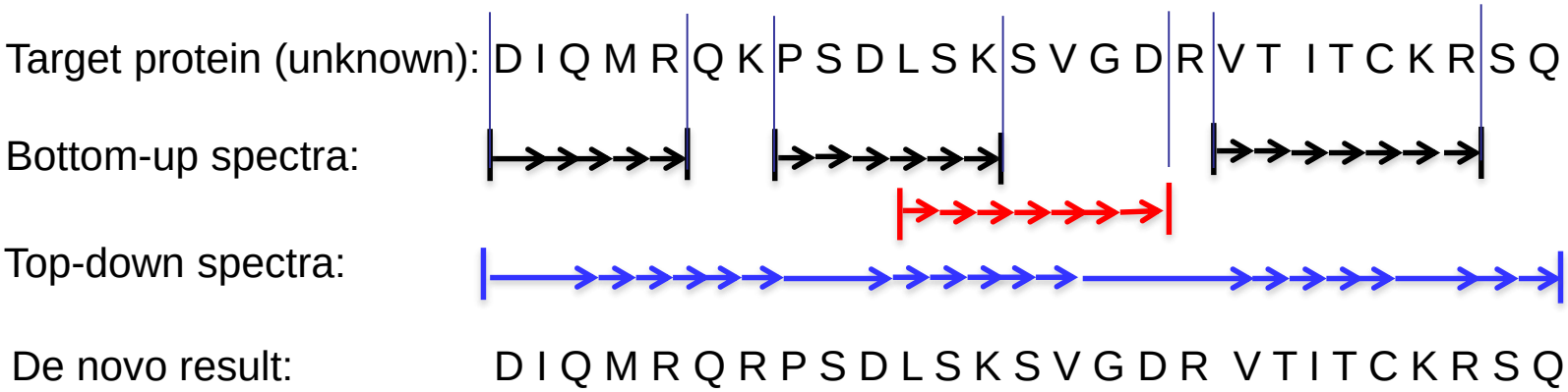
Top-down spectra:

De novo result: [ ]Q M R Q K [ ] D L S K S V[ ] V T I T C[ ]S Q

- 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



Data Sets

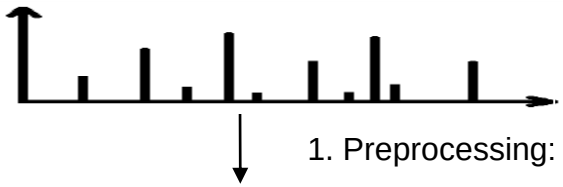
- Light chain of alemtuzumab (MabCampath)
  - Top-down
    - Thermo LTQ Orbitrap Velos and Thermo Q-Exactive
    - ETD: 12134 spectra; CID: 7686; and HCD: 4931
  - Bottom-up
    - Thermo LTQ Orbitrap XL
    - HCD spectra
    - Trypsin: 2716 spectra, chymotrypsin: 4328, proteinase K: 1616 and pepsin: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



Tandem mass spectrum from peptide
PRTEINSTRING

1. Preprocessing:

parent mass: **M=1111 Da** → PRTEINSTRING

Neutral mass list:

253 Da → PR
457 Da → RING
483 Da → PRTE
…
569 Da → PRTEINS

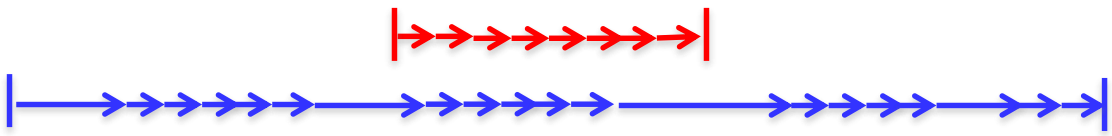
Add complementary masses

M-253 Da → TEINSTRING
M-457 Da → PRTEINST
M-483 Da → INSTRING
…
M-596 Da → TRING

Prefix residue masses: 253 483, M-457

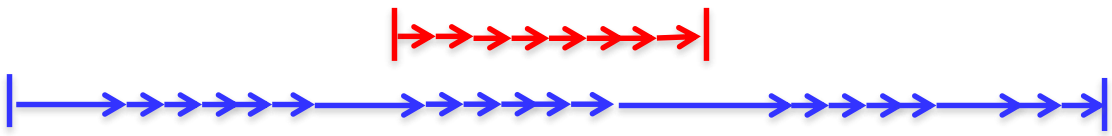
Candidate prefix residue masses: 253, 457, 483, …, 569, M-253,…, M-596

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



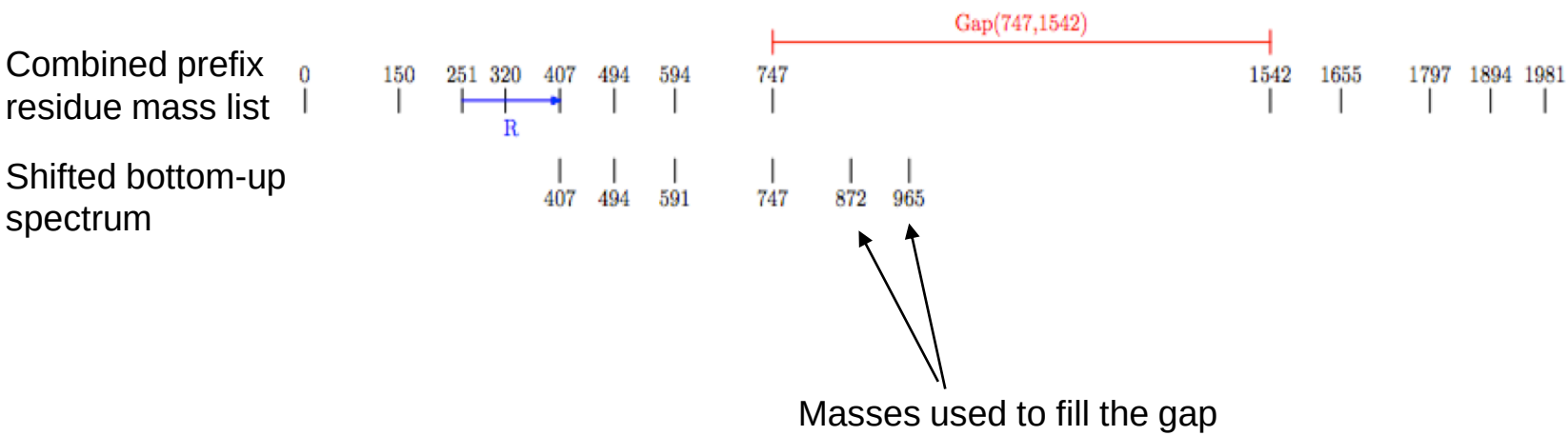
- Keep only bottom-up spectra with a shifted mass count score >= 7.

- 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**

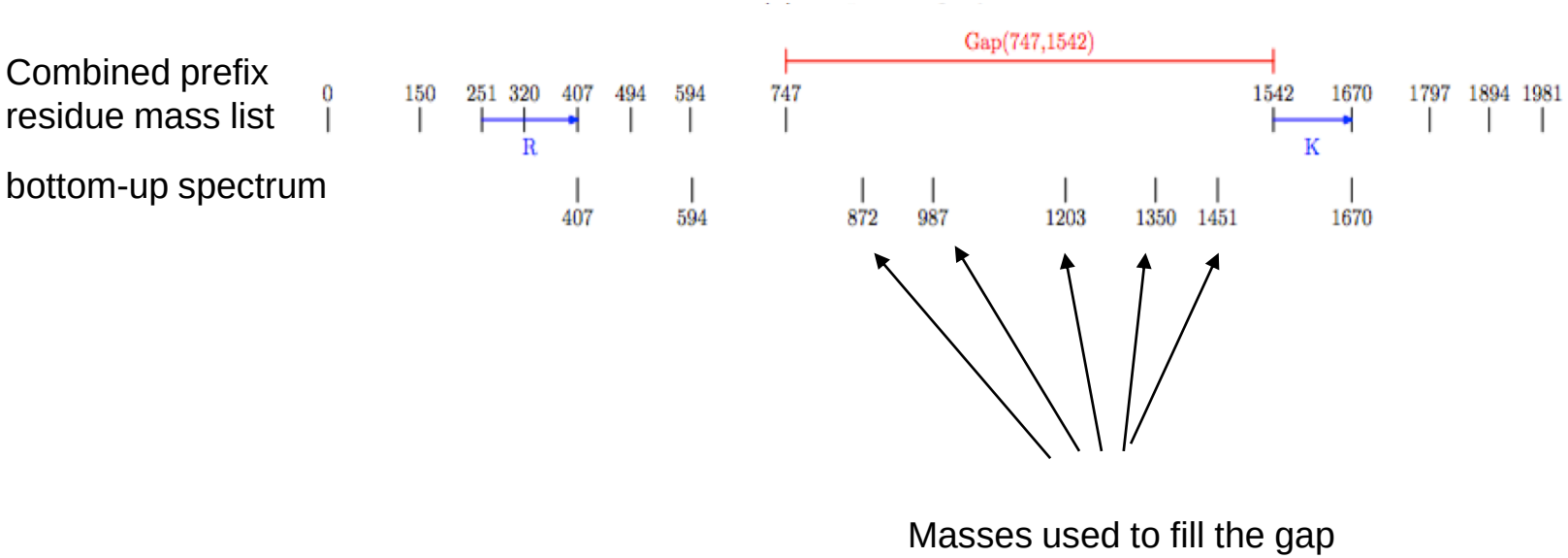
Gap Filling

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



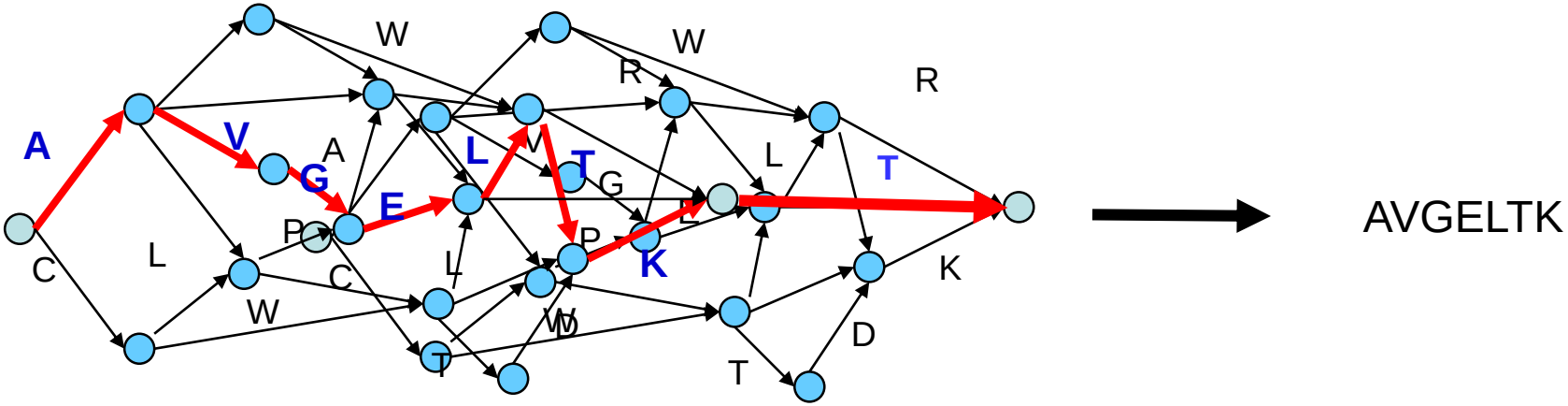
Gap Filling

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



Spectral Graph

- 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  KYLNWYQQKP  GKAPKLLI
  [356.17]MTQSPSS  ISASVGDRVT  ITCK[286.19]NID  KYINWYQQKP  GKAPQIII

1  TNNLQTGVPS  RFSGSGSGTD  FTFTISSLQP  EDIATYYCLQ  HISRPRTF
7  TNNIQTGVPS  RF[231.10]G[ 360.17 ]  FTFTI[      1367.59      ]YCIQ  HISRPRTF

01 GTKVEIKRTV  AAPSVFIFPP  SDEQLKSGTA  SVVCLLNNFY  PREAKVQW
2  GTKVEIKRSI  AAPSVFIFPP  SDEQIKSGTA  SVVCIINNFY  PREAQPRR

51 DNALQSGNSQ  ESVTEQDSKD  STYSLSSTLT  LSKADYEKHK  VYACEVTH
32 DNAIQSGNSQ  ESVTEQDSKD  STYSISSTIT  ISQADYEKHQ  VYACEVTH

01 LSSPVTKSFN  RGEC    214
32 ISSPVTKSF[  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/

Acknowledgements

UCSD

Pavel A. Pevzner

PNNL

Si Wu

Ljiljana Paša-Tolić

Nikola Tolić

Erasmus MC, Netherlands

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

Saint Petersburg Academic University

Mikhail Dvorkin
Sonya Alexandrova
Kira Vyatkina