



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by **Battelle** Since 1965

MSPathFinder

*Open Source Software Package
for Top-Down Proteomics*

Sangtae Kim, Christopher S. Wilkins, Jungkap Park,
Paul D. Piehowski, Anil K. Shukla, Yufeng Shen,
Samuel H. Payne and Richard D. Smith

Pacific Northwest National Laboratory

Why is Top-Down Proteomics Important?

Proteoforms: Protein isoforms and species that arise from four major sources:

1. Multigene families
2. Alternative splicing
3. Coding polymorphisms
4. Post-translational modifications (PTMs)

Top-Down proteomics directly analyzes the actual biological actors!

Existing Tools for Proteoform Identification

ProSightPC, **ProSight PTM**, **ProSight Lite**

(Neil Kelleher @ Northwestern,
Thermo Scientific)

MS-Align+, **MS-Align-E**, **TopPIC**

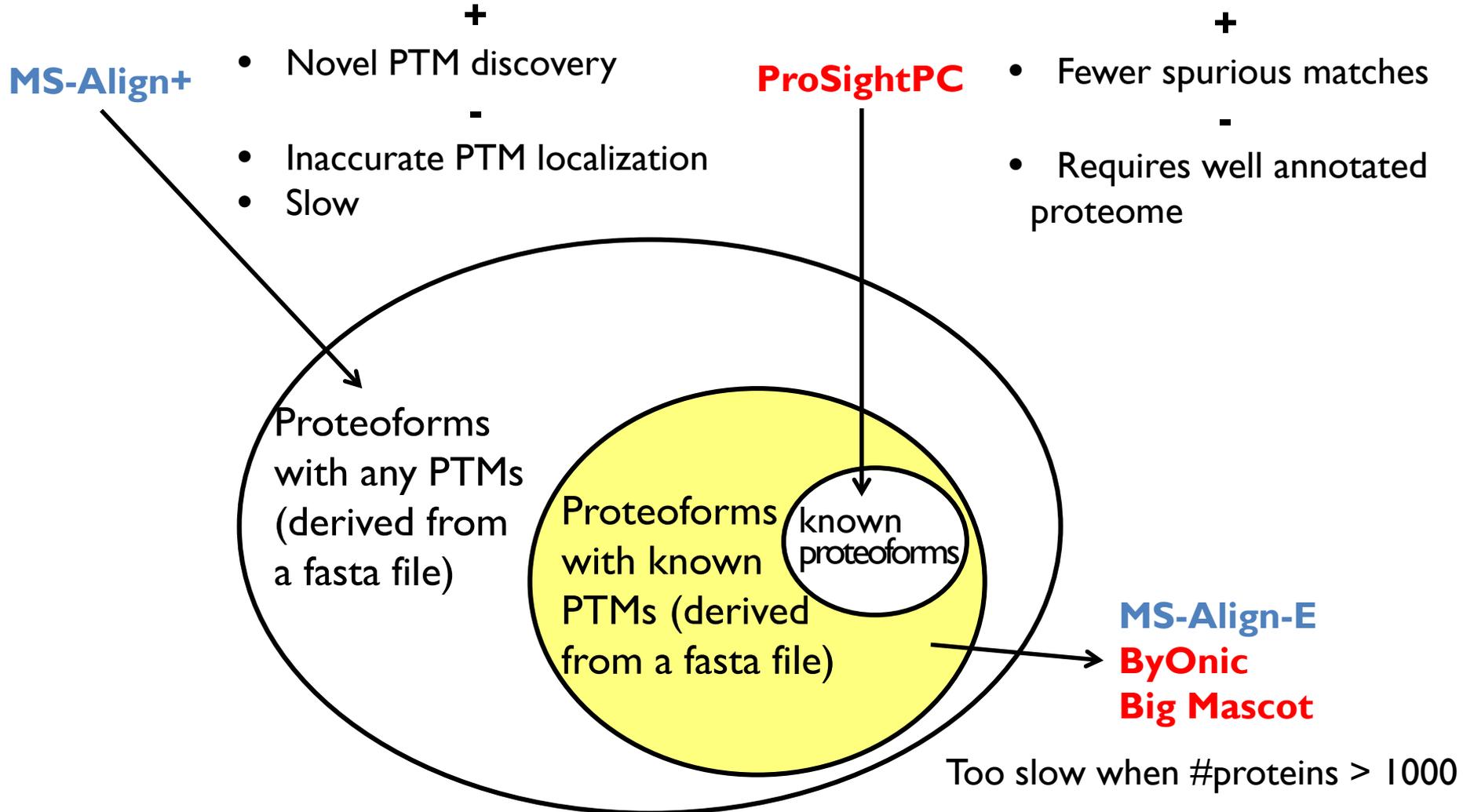
(Xiaowen Liu @ IUPUI, Pavel Pevzner @ UCSD)

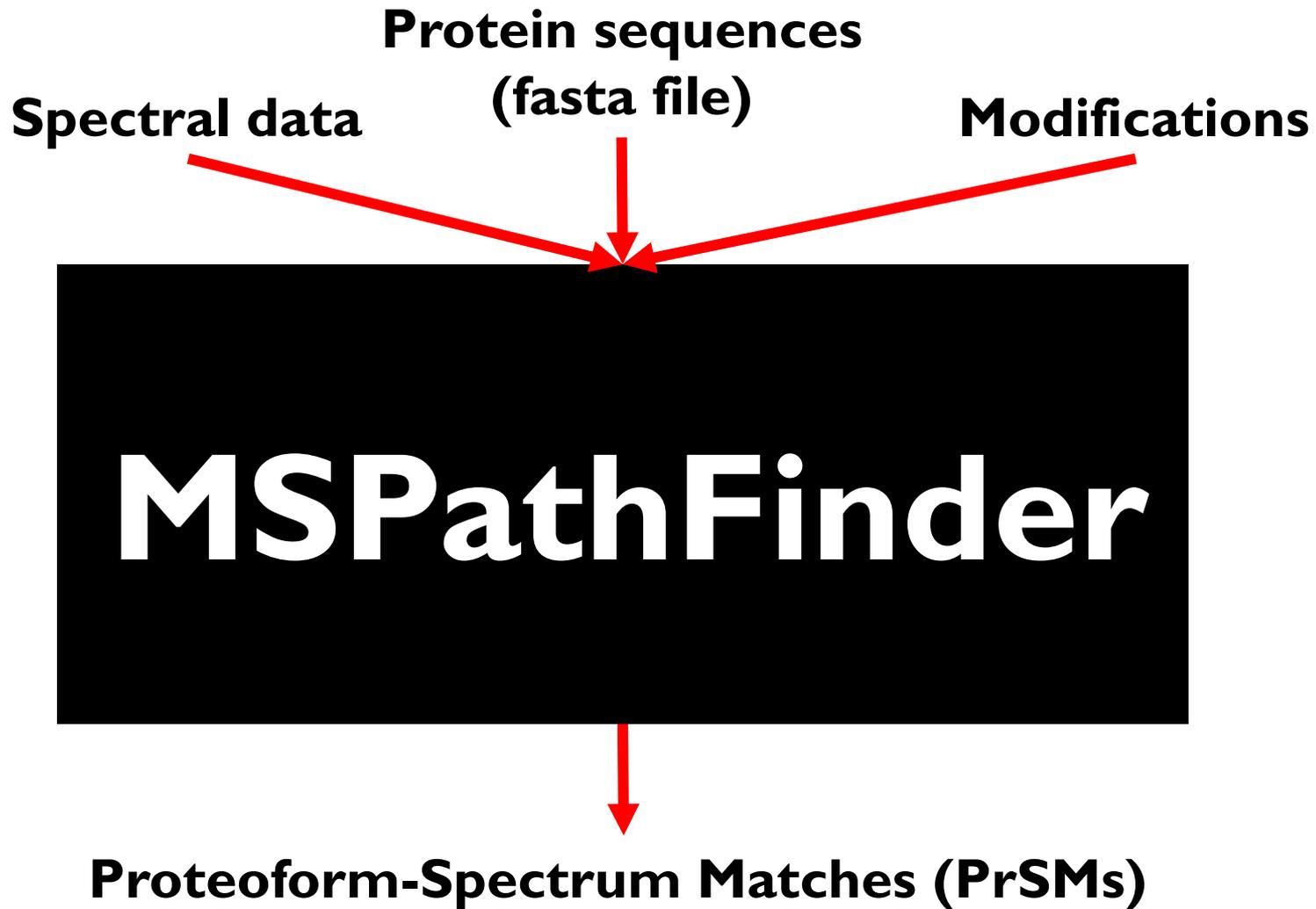
ByOnic (Protein Metrics)

Big Mascot (Matrix Science)

— Commercial
— Free

Top-Down Software Tools





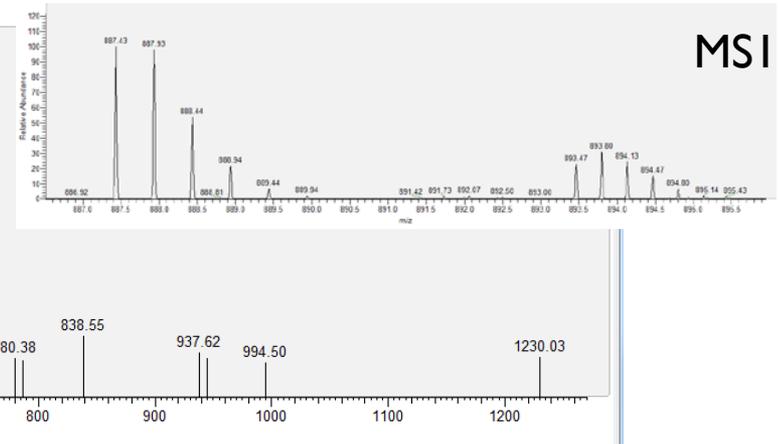
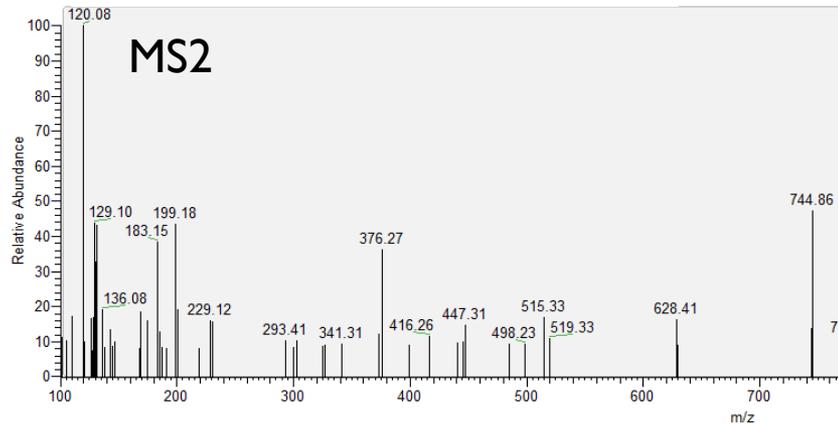
Challenges

Compared to bottom-up

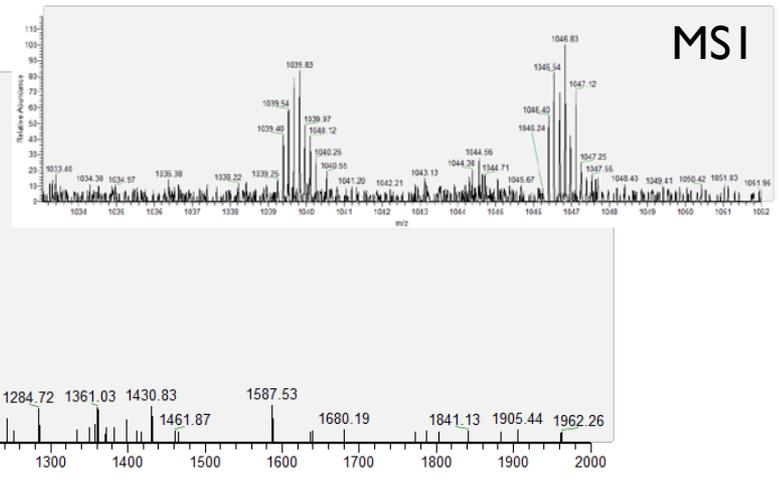
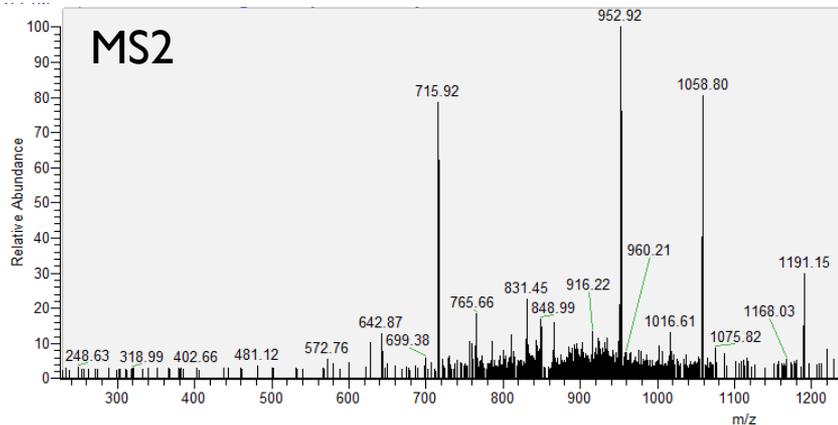
1. Spectra are more complex
2. Spectral information for proteoforms are distributed over many peaks
3. Many possible proteoforms

Challenge: Complexity of Spectra Increases with Molecular Weight

Bottom-Up: Less complex

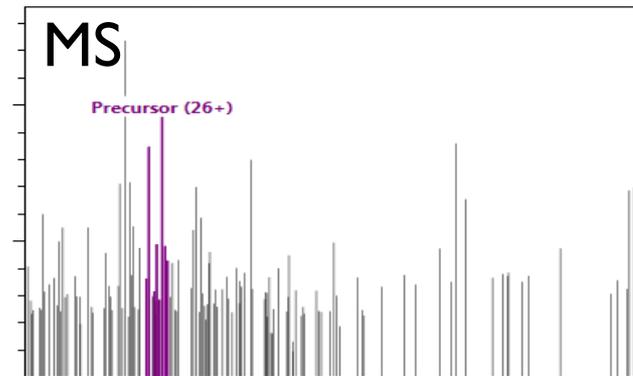
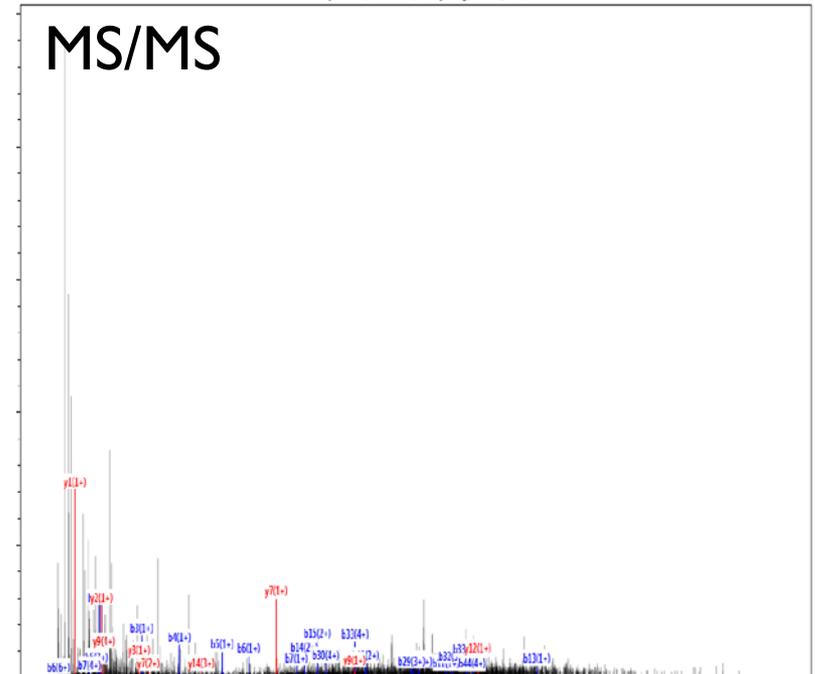
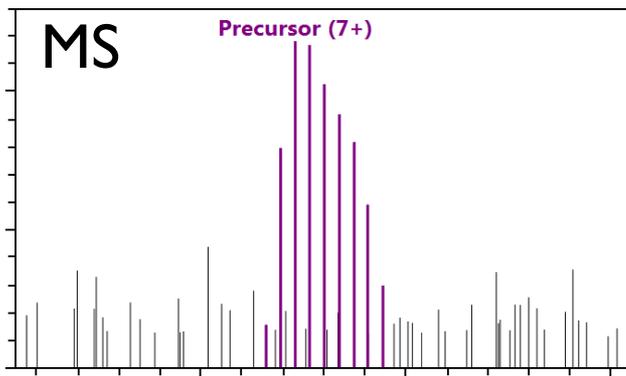
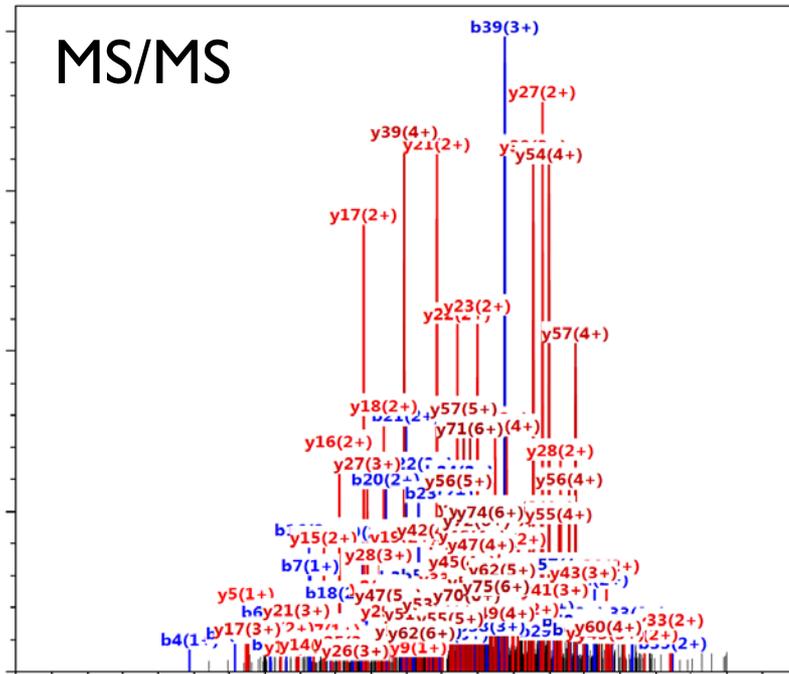


Top-Down: More complex
(highly charged ions)



Small Proteins (8,480 Da)

Larger Proteins (26,116 Da)



ThP 438

ProMex (Protein Mass Extractor)



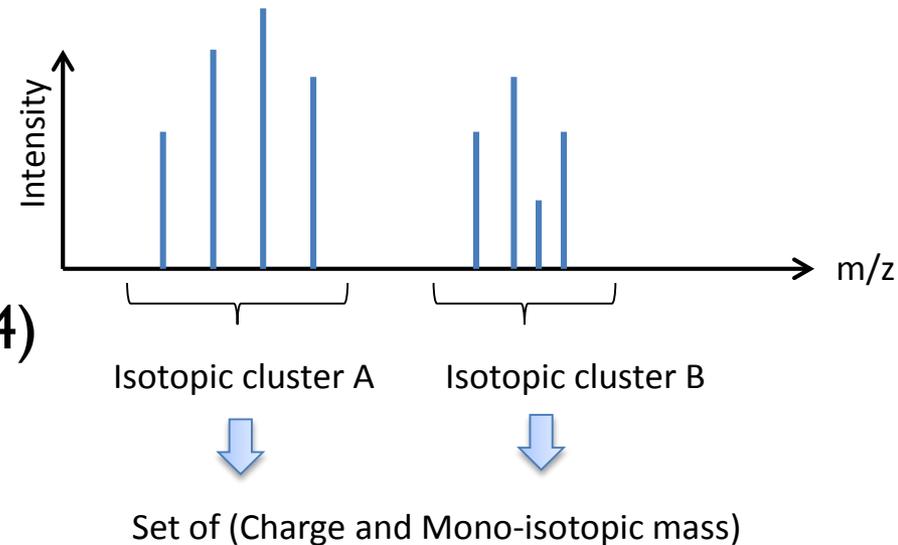
Jungkap Park

Feature Extraction from LC/MS data

- MSI Feature : (Elution Time, Mono-isotopic Mass, Charges)

- Existing tools

- THRASH, (Horn et al., 2000)
- MS-Deconv, (Liu et al., 2010)
- MS-Deconv+, (Kou et al., 2014)

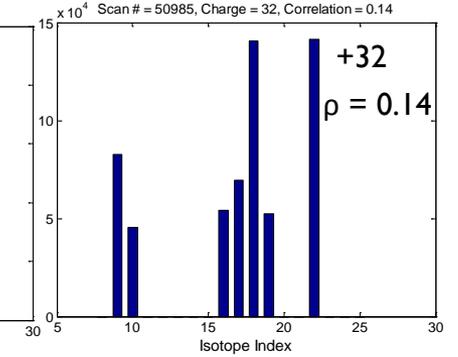
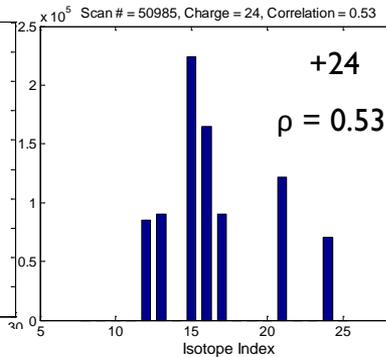
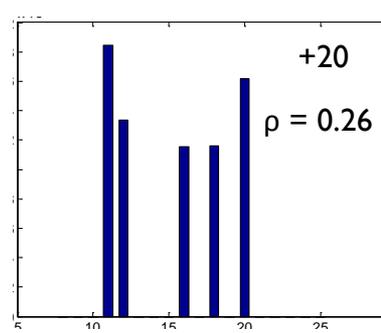
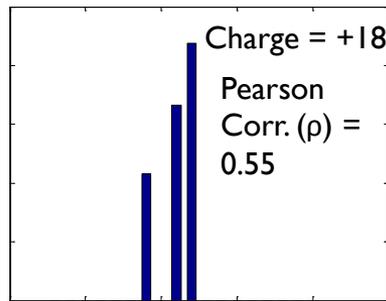


- Challenges

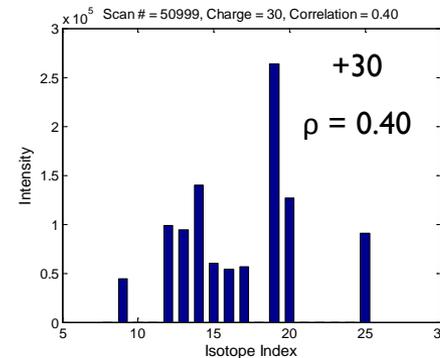
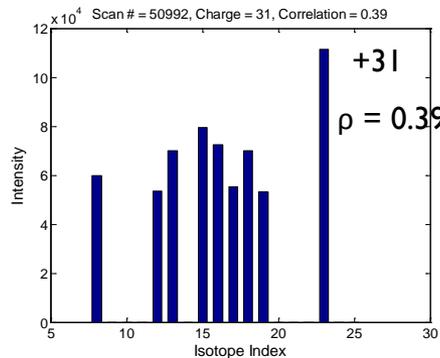
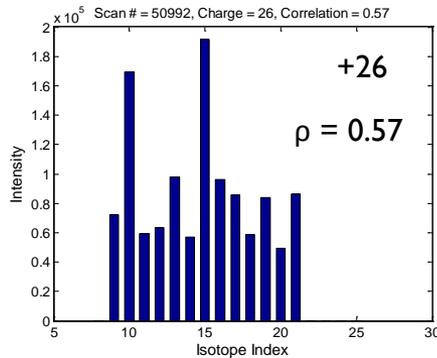
- Do not work well for proteins with large masses
- Signal is **scattered across different charges/scans**
- Too few ions often leading to poor definition of isotope envelopes (Orbitrap)

Observed Peaks for a 26 KDa Protein

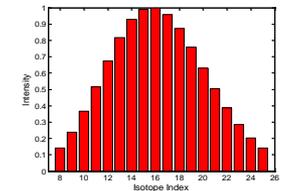
Scan# = 50985



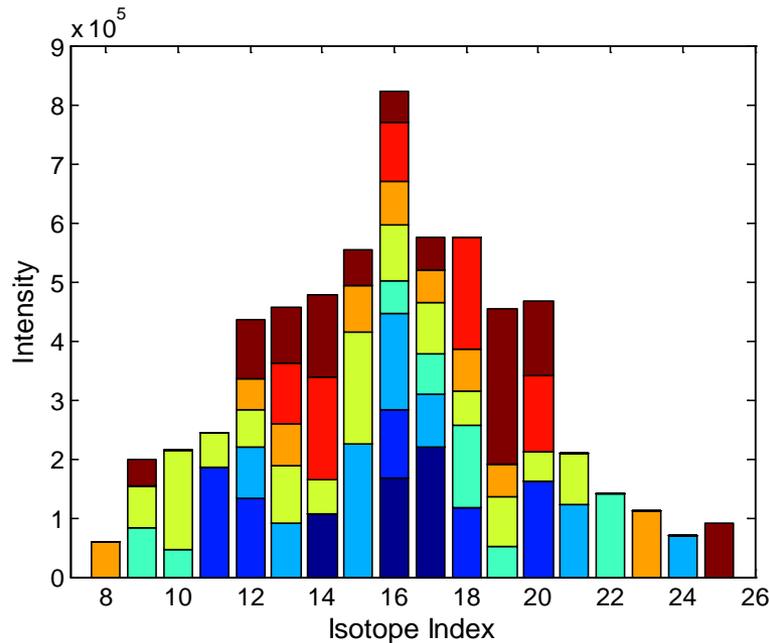
Scan# = 50992



Theoretical Envelope



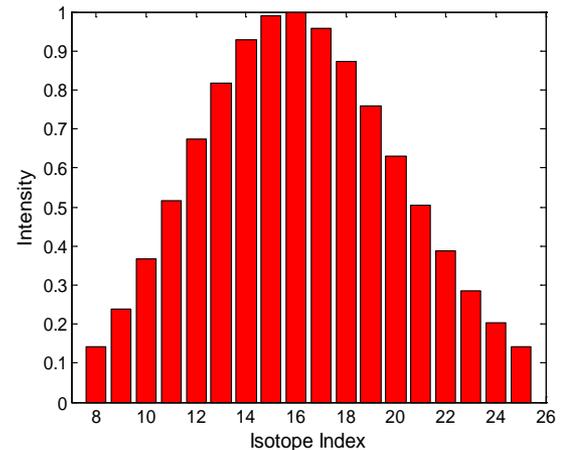
Summed Envelope across Different Charges/Scans



**Pearson
correlation
= 0.9246**



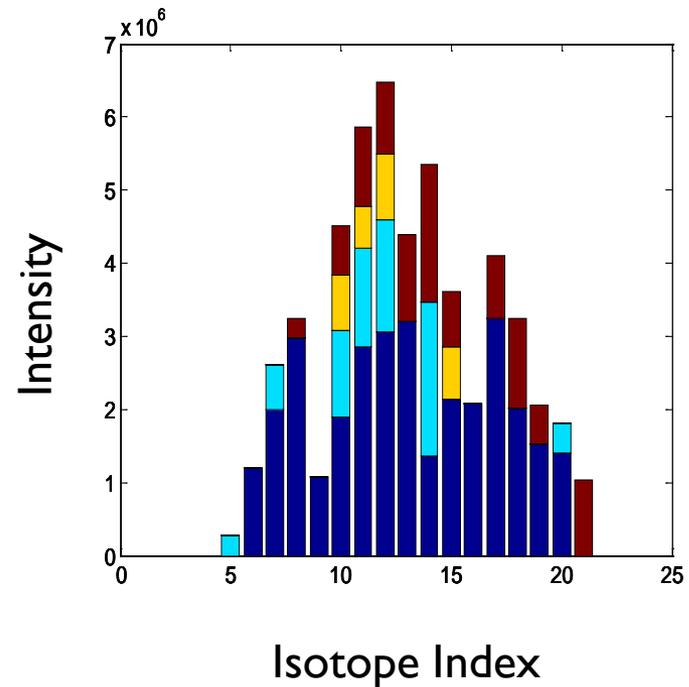
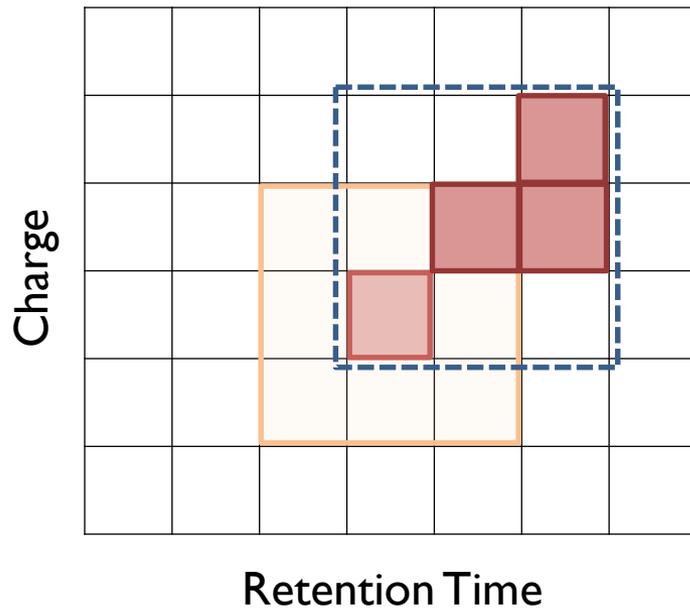
Theoretical Envelope



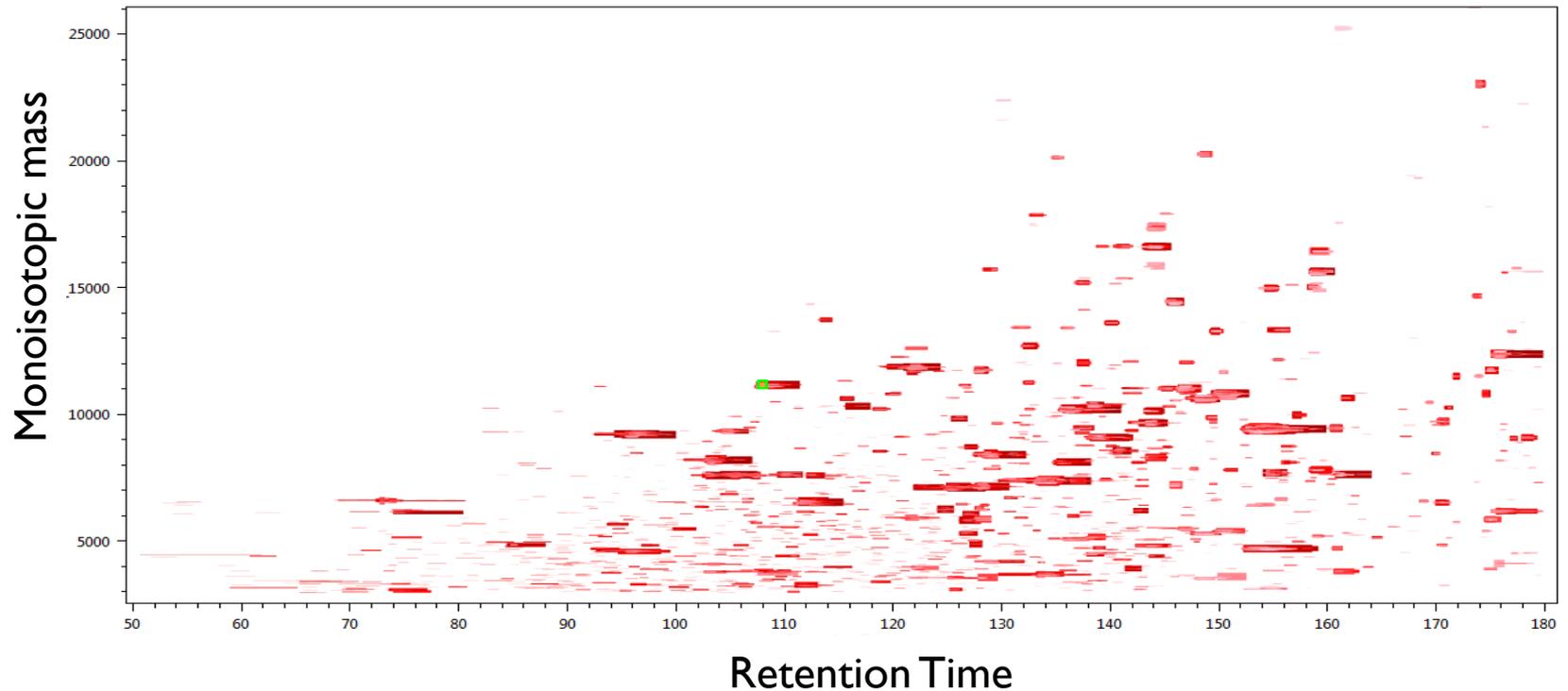
How many spectra? Which charges?

ProMex Algorithm Overview

- **Hill-Climbing Optimization**



ProMex Generates “Feature Map”



Data from LTQ-Orbitrap Elite

1. Spectra are more complex
2. Spectral information for proteoforms are distributed over many peaks
3. **Too many possible proteoforms**

Observation

Many proteoforms
have the same
elemental compositions
(#C,#H,#N,#O,#S)

#Compositions << #Proteoforms

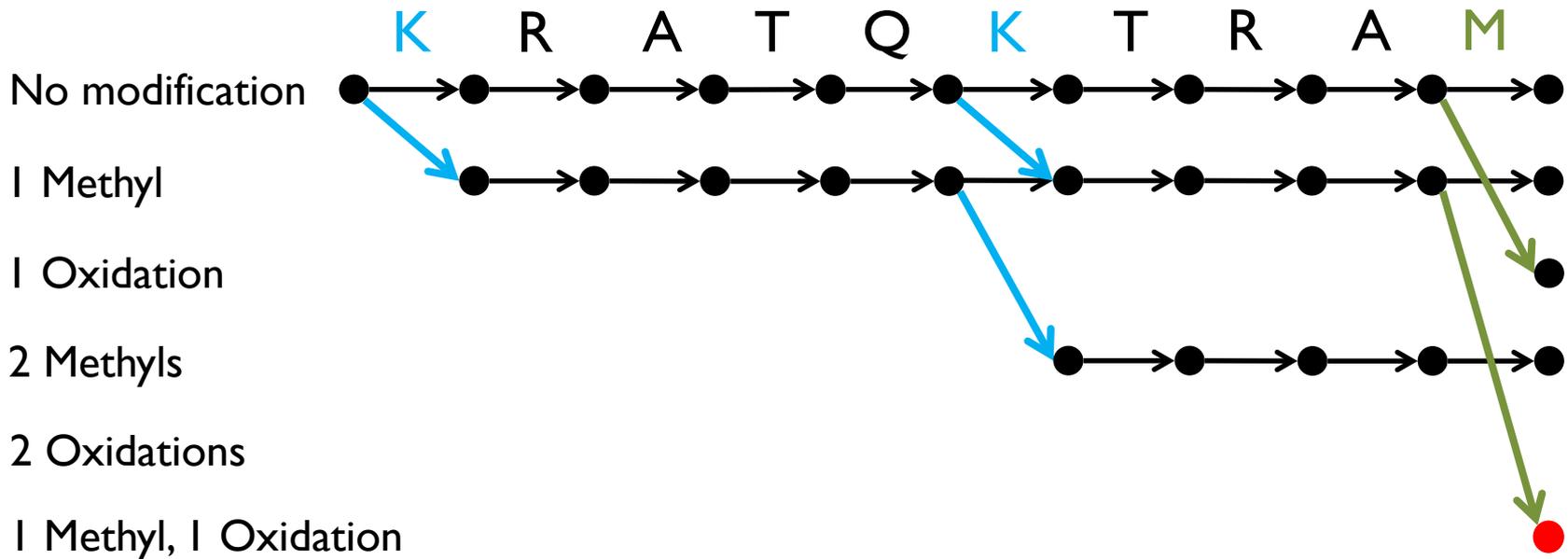
Histone H4



#Proteoforms? $5 \cdot 10^{13}$

#Unique Elemental Compositions? 4,368

Sequence Graphs are used to interrogate many proteoforms simultaneously



All proteoforms are represented as paths

Interrogating > 50 trillion different proteoforms
in less than a minute

Similar to “Spectral Alignments” (Pevzner et al., J. Comp. Biol. 2000)

Possible Proteoforms Due to Internal Cleavages

Protein in a database

	% in total ID	#Sequences derived from a database
No cleavage or N-term single residue cleavage	25%	112K
Single internal cleavage (+ N-term single residue cleavage)	60%	3M
Multiple internal cleavages	15%	223M

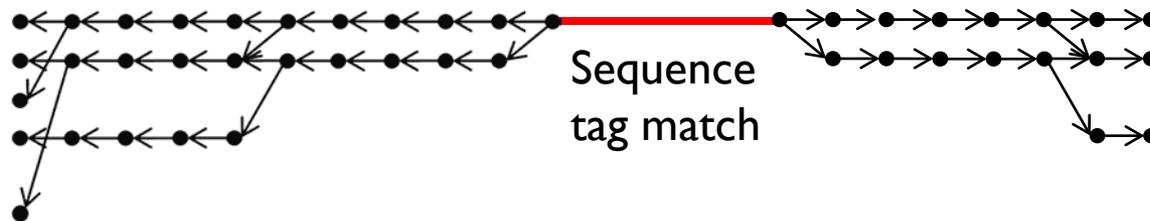
99% search time

Salmonella database containing 5,634 proteins

Sequence Tag Based Search

- Used to find proteoforms having multiple cleavages

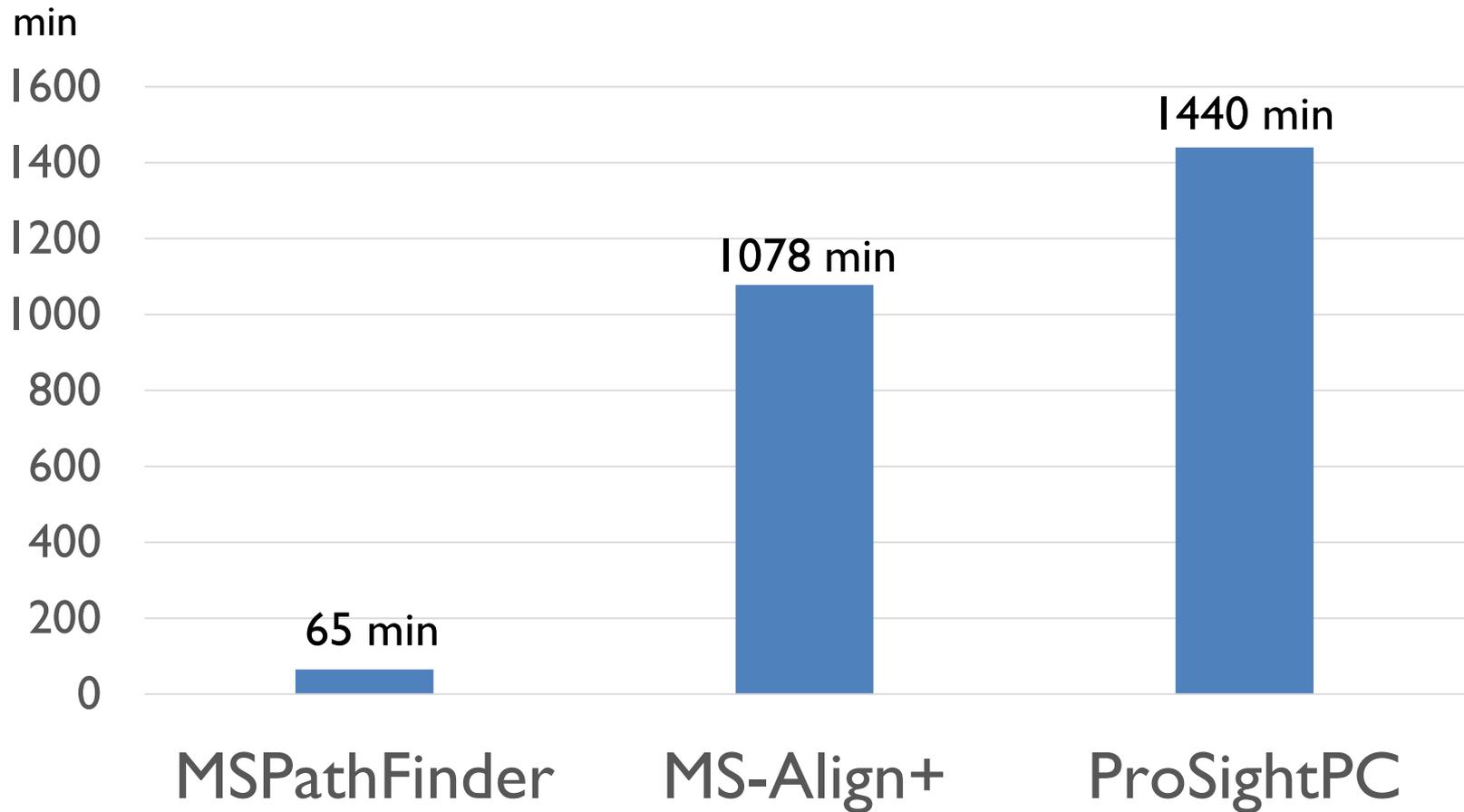
Protein



- Generate short de novo sequence tags
- Find proteins matching the sequence tags
- Extend sequence tag matches using sequence graphs

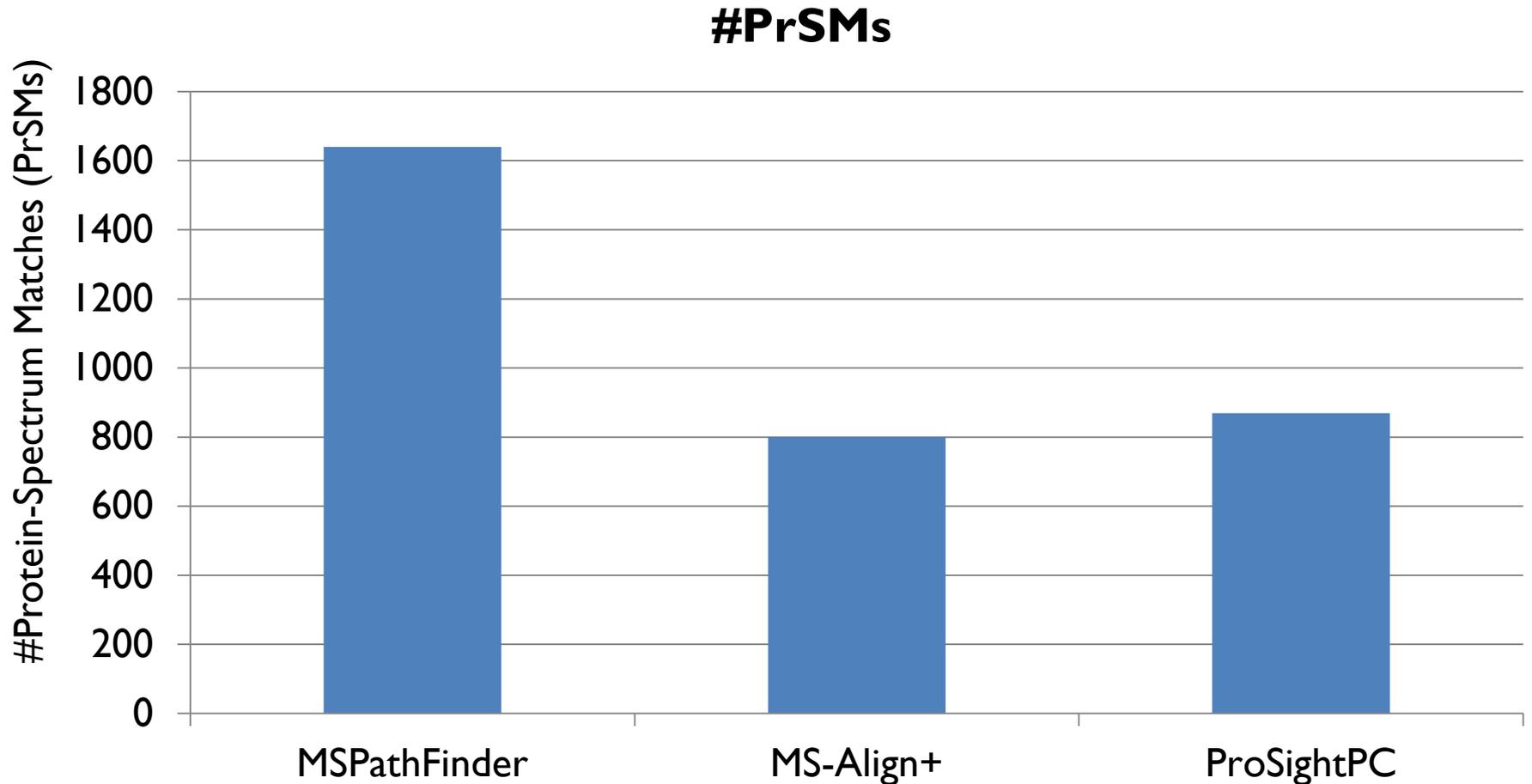
Results

Running Time – No modification



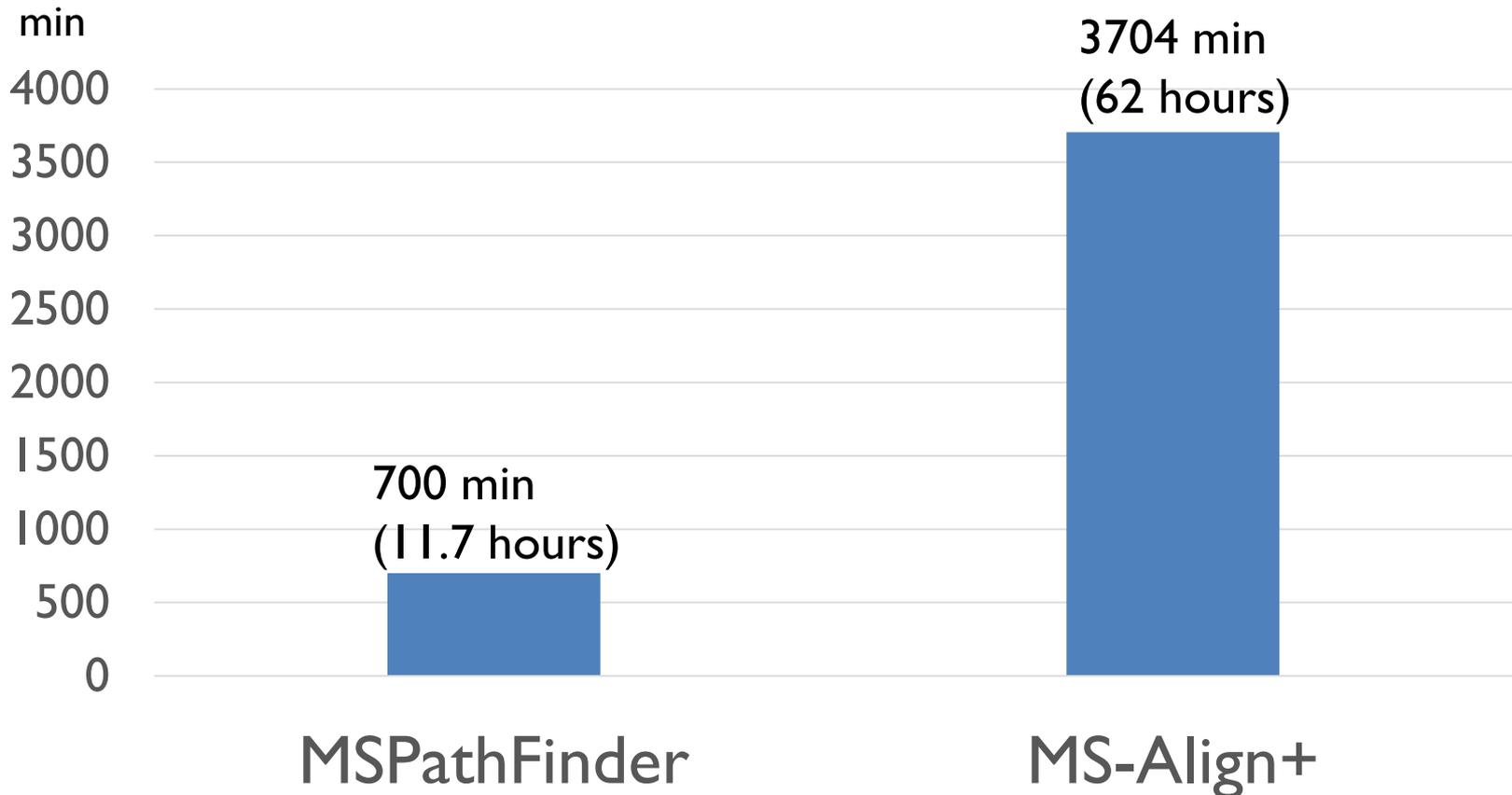
*Salmonella dataset (Ansong et al., PNAS 2013)
7703 spectra, LTQ-Orbitrap Elite*

#Protein Spectrum Matches (PrSMs)



No modification search,
FDR 1% (Target-Decoy Approach)

Running Time – PTM Search

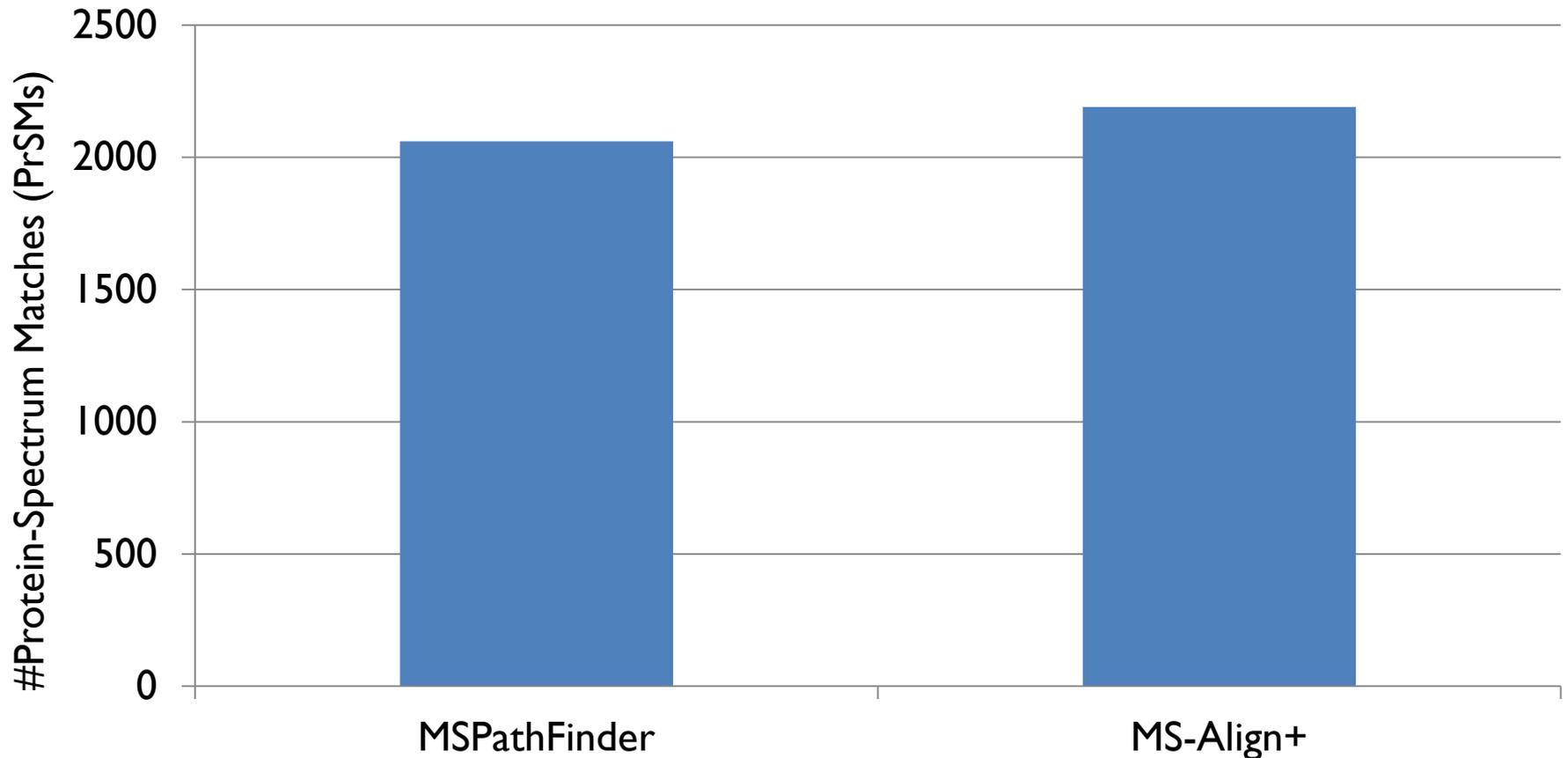


MSPathFinder: Acetyl Prot N-term, Oxidation M, Dehydro C, Glutathione C

MS-Align+: 2 blind modifications

Salmonella dataset (Ansong et al., PNAS 2013)
7703 spectra, LTQ-Orbitrap Elite

#PrSMs – PTM search

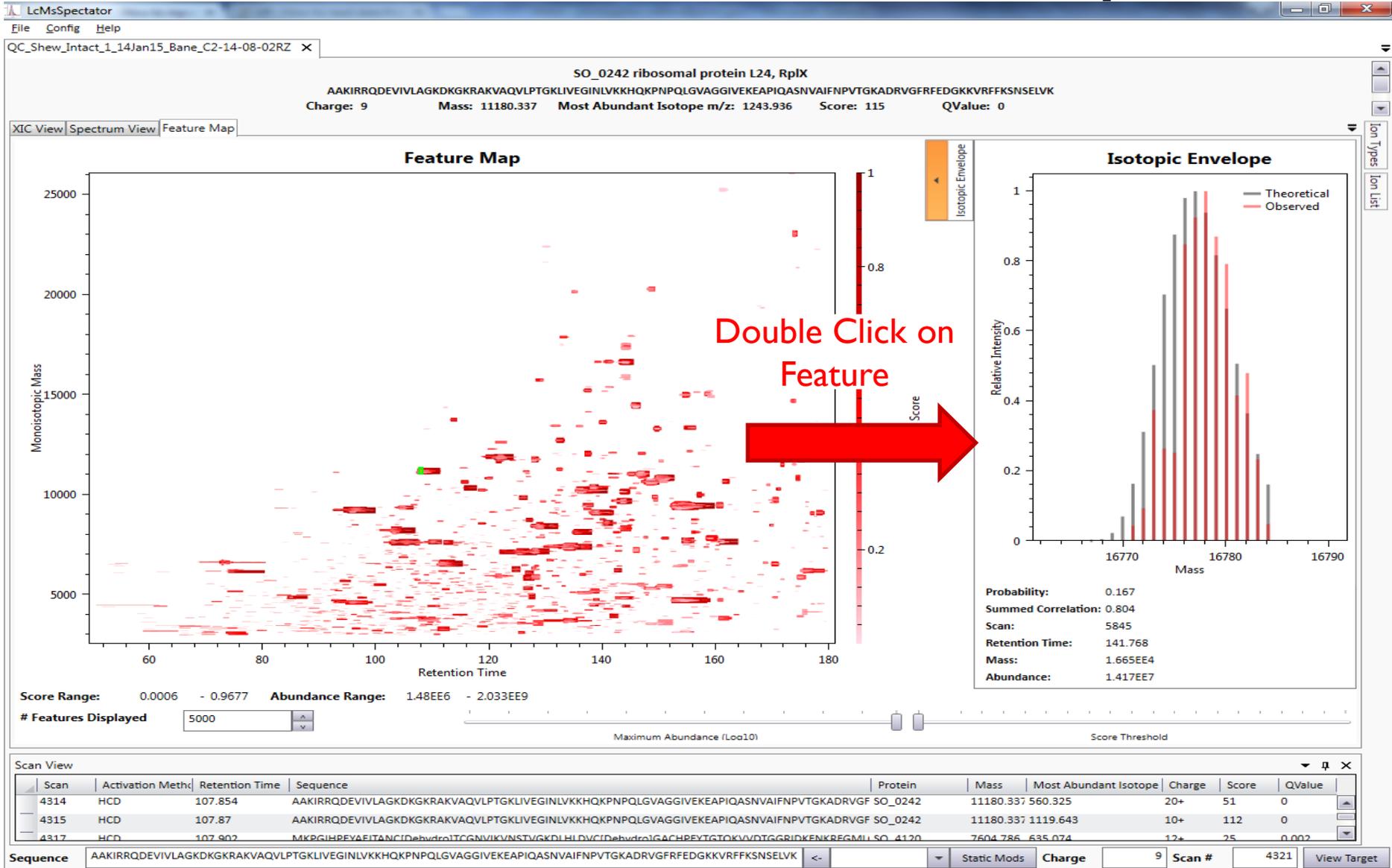


FDR 1%

LcMsSpectator

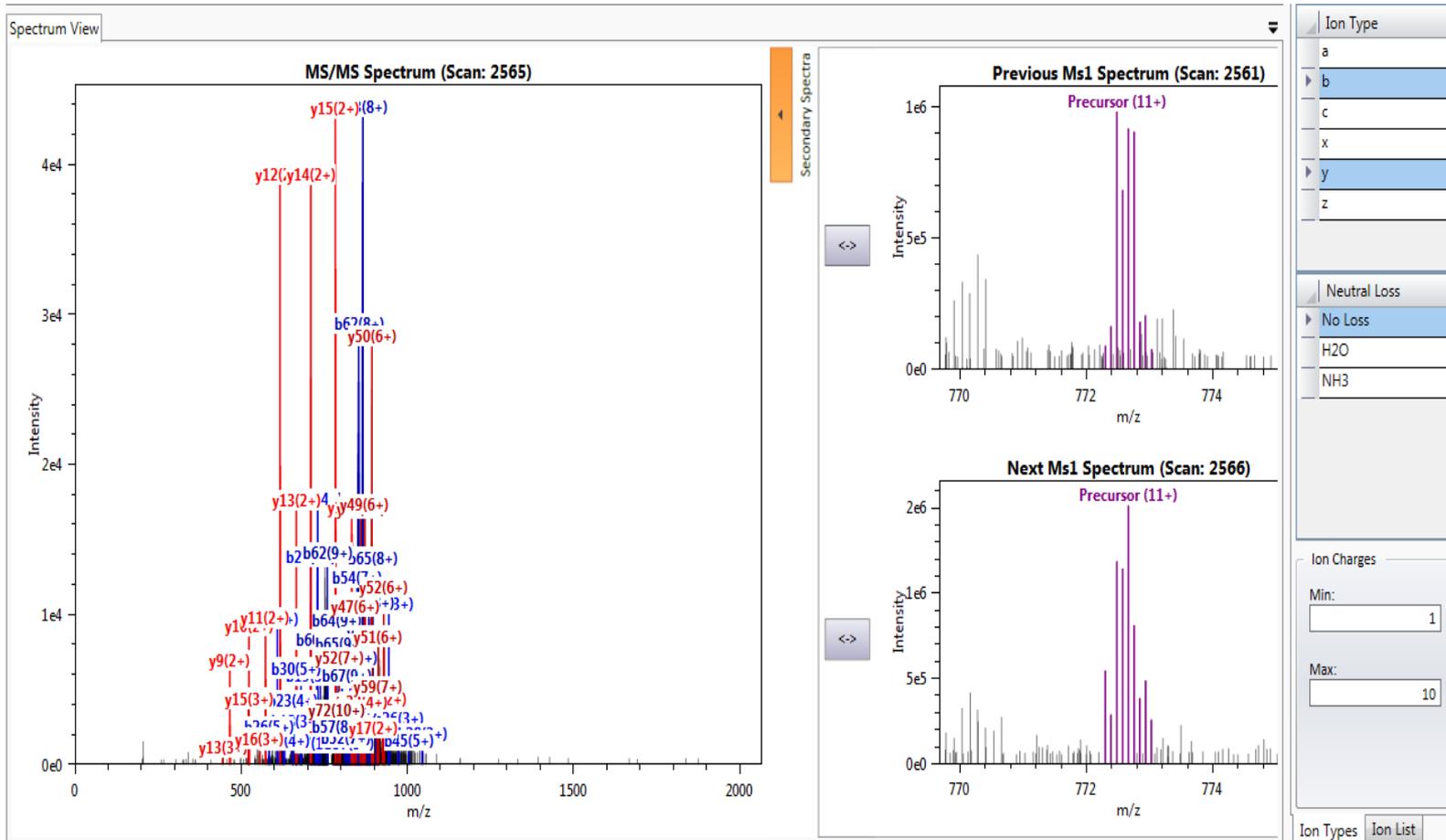
Visualization of MSPathFinder/ProMex Results
Edit identifications

Proteoform Feature Map

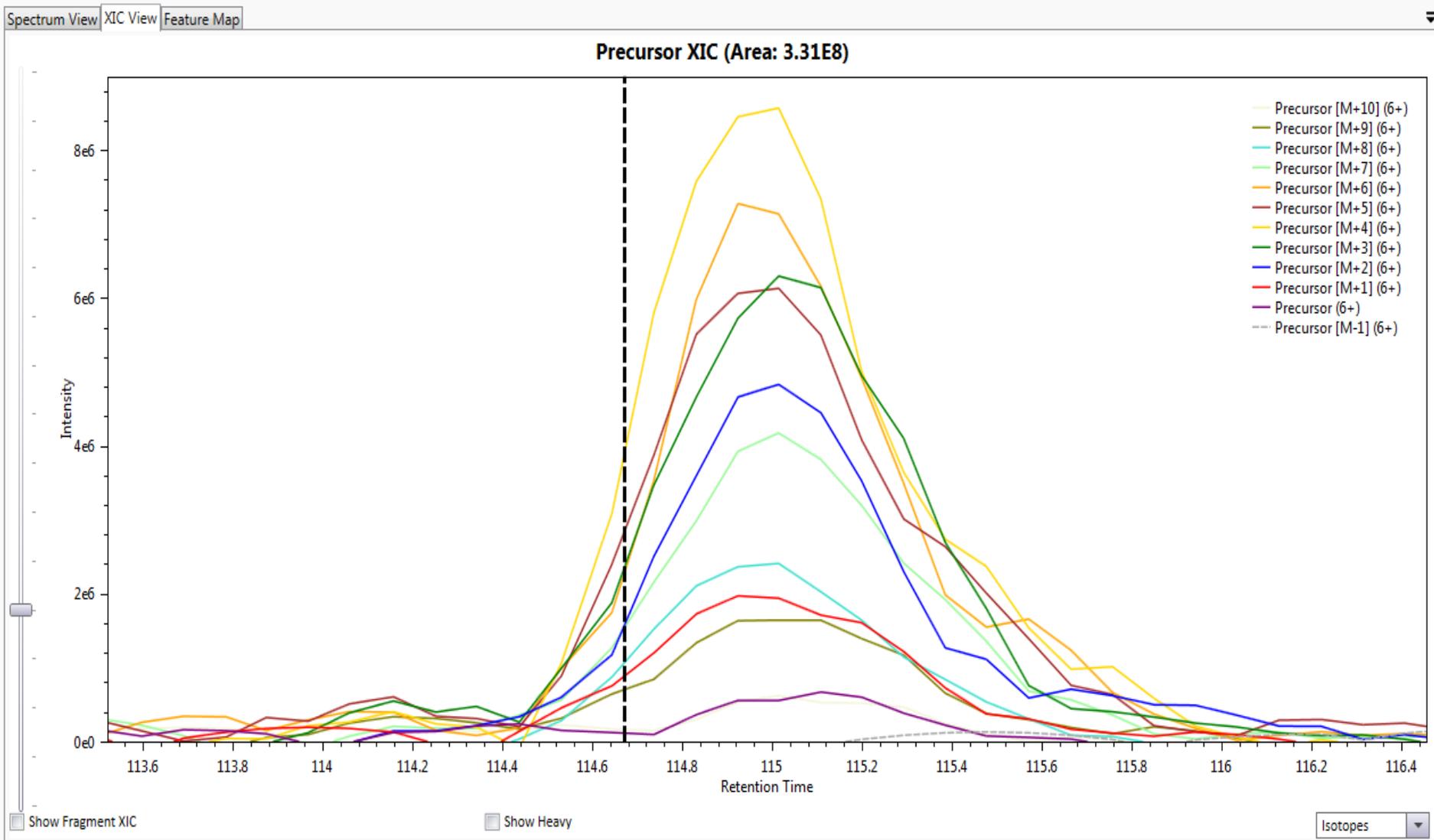


Spectrum View

MS/MS

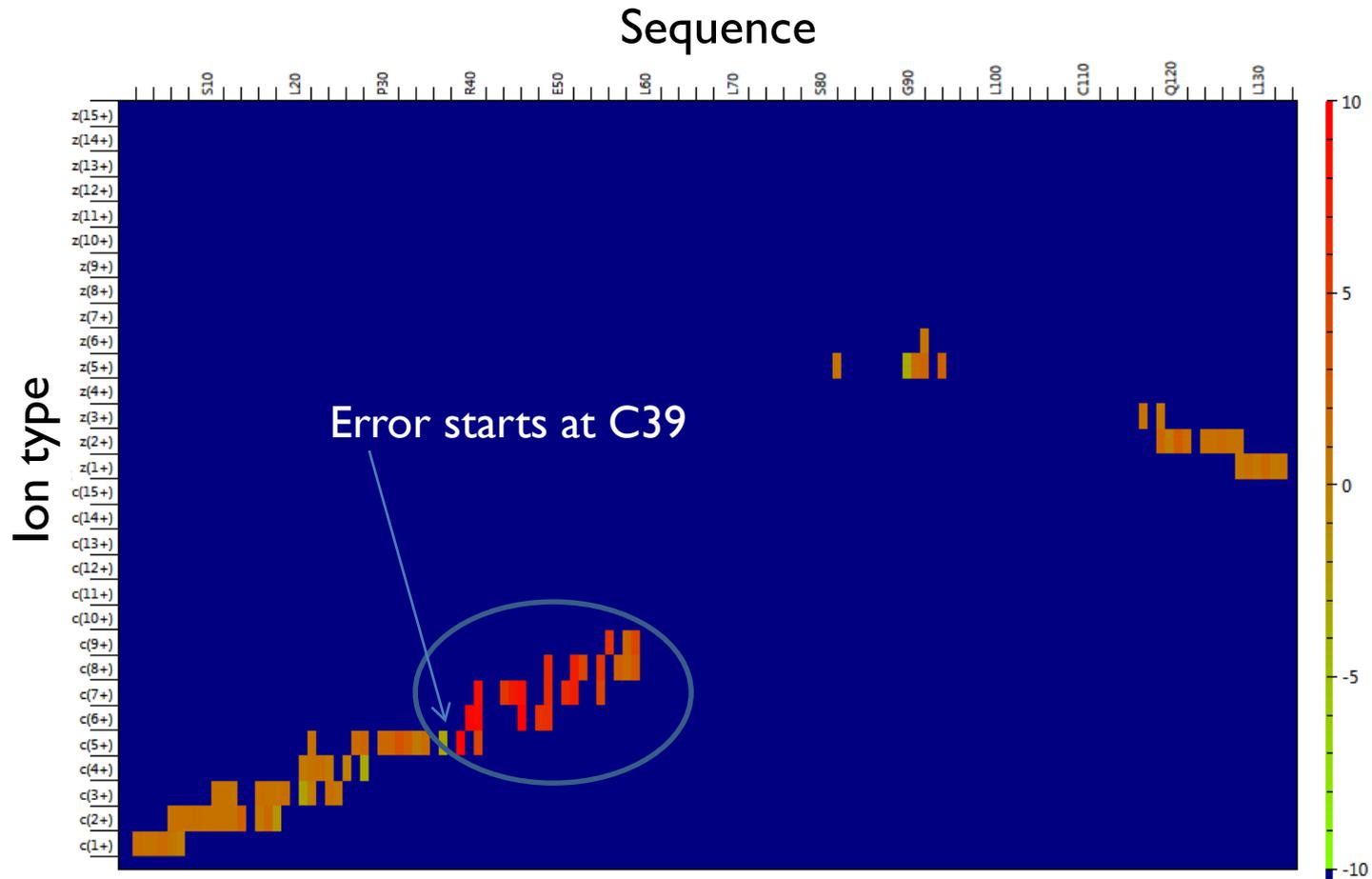


Extracted Ion Chromatogram View

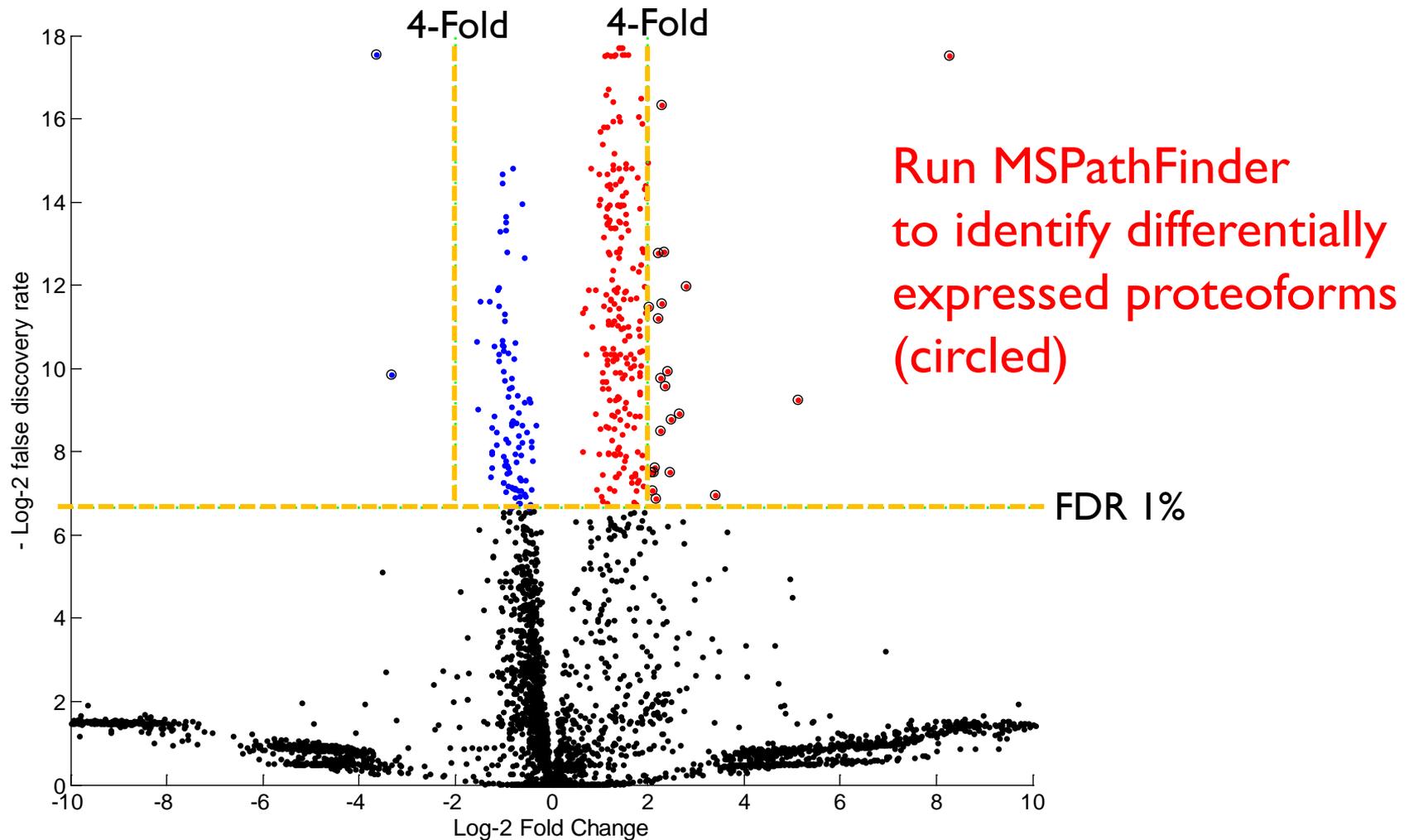


**LcMsSpectator allows
editing of identifications**

ID From MSPathFinder - K36Ac



ProMex+MSPathFinder for Label-Free Quantification



Summary: MSPathFinder

Fast, Free, Effective and Easy-to-use

**Spectral
Data**

**Protein
Sequences (fasta)**

Modifications

MSPathFinder

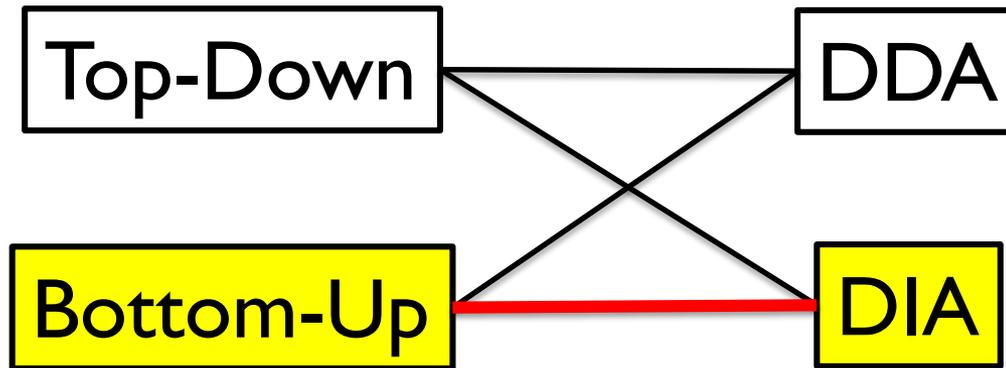
ProMex
(Feature Finding)

Sequence Graphs
(Protein Identification)

LcMsSpectator
(Visualization)

**What proteins were present in the sample?
What are their quantities?**

MSPathFinder



WP 3 | 4 “Improving DIA Peptide Identification via Local Time Profile Similarity”
by Afkham, Kim, and Käll

Acknowledgements

ProMex, Sequence Tag Generation



Jungkap Park (Post-doc)

LcMsSpectator



Chris Wilkins (Undergraduate Intern)

Richard D. Smith

Samuel Payne

Paul Piehowskii

Yufeng Shen

Anil Shukla

Ron Moore

Errol Robinson

Ljiljana Pasa-Tolic

Nikola Tolic

Jered Shaw

Mowei Zhou

Si Wu

Weijun Qian

Aaron Wright

Natalie Sadler

Zhe Xu

Tao Liu

Matthew Monroe

Vlad Petyuk

Support

DOE Pan-omics Program

NIGMS Proteomics Research Resource at PNNL