

Deep and Fast Analysis of High Resolution MS/MS Data



——引擎那些事儿

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

迟浩

中国科学院计算技术研究所

pFind团队

2014-11-12

The Achilles Heel of Proteomics

Data analysis—the Achilles heel of proteomics

Scott D. Patterson

March 2003 · Volume 21

Nature Biotechnology

During the past few years there has been a resurgence of research using parallel protein-based analysis, now commonly referred to as proteomics. However, our ability to generate data now outstrips our ability to analyze it. This occurs even though proteomics is inherently a substrate-limited science and proteins exist over a wide concentration range in biological samples. Therefore, it is not surprising that the entire proteome of any species has yet to be observed. In this article, I address some of the primary issues currently facing researchers in this field, with an emphasis on the computational aspects affecting progress, including the accuracy of matches from mass spectrometric data to sequence databases and the integration of the results of proteomics experiments to yield biological meaning.

Parallel protein-based analysis first came to the fore during the mid-1970s with the introduction of two-dimensional gel electrophoresis, which for the first time allowed a staggering number of different protein species to be revealed in a single experiment and permitted the comparison of expression patterns between samples. At that

Table 1. Proteomics experiments require handling of diverse data sets

Stage of process	Type of data
Preparation for analysis	Project information Sample information Separation, fractionation
Sample processing	Quantitative analysis (LC-MS/MS, 2-DE MS/MS) Identification, MS/MS data analysis
Data analysis	Data capture and validation Data management and integration
Monitoring	All processes require quality assurance and quality control

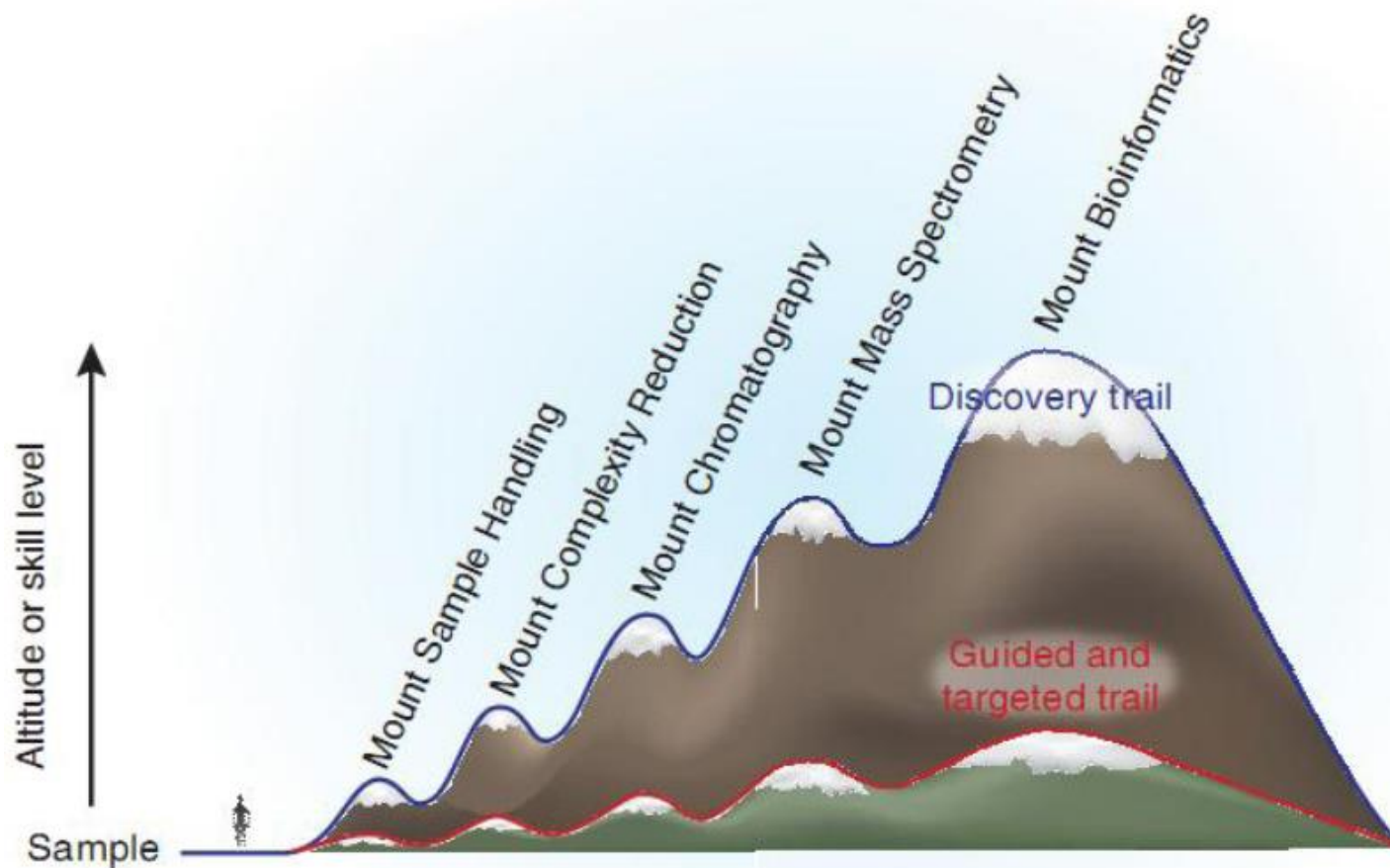
numerous examples in which transcript amounts at steady state, or even after induction, do not correlate with the amount of protein present in the system because of well-known, but currently mostly unpredictable, effects of post-transcriptional and post-translational regulation¹. Therefore, analysis at the level of

sis problems because of the nature of the two main technologies used in studying proteins: first, systems based on two-dimensional gel electrophoresis (i.e., quantification through image analysis, proteolytic digestion of the isolated proteins of interest, and identification by mass spectrometry) and non-gel based systems (i.e., quantification and identification by mass spectrometry of mixtures of proteolytically digested proteins)¹.

Two-dimensional gel software continues to be improved, but after 20 years of development it still requires some manual intervention. Mass spectrometers measure the mass-to-charge ratio of charged molecules,

“...our ability to generate data now outstrips our ability to analyze it”

Ruedi Aebersold, 2009



ID Rate of Low Resolution Data

NATURE METHODS | VOL.2 NO.9 | **SEPTEMBER 2005** | 667

Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations

Joshua E Elias¹, Wilhelm Haas¹, Brendan K Faherty² & Steven P Gygi^{1,2}

- LTQ (Mascot + Sequest): $5366 / 15992 = 33.6 \%$
- QTOF (Mascot + Sequest): $3477 / 15309 = 22.7 \%$



ID Rate of High Resolution Data

Molecular & Cellular Proteomics 2011

Research

Author's Choice

© 2011 by The American Society for Biochemistry and Molecular Biology, Inc.
This paper is available on line at <http://www.mcponline.org>

Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer*

Annette Michalski‡, Eugen Damoc\$, Jan-Peter Hauschild\$, Oliver Lange\$, Andreas Wieghaus\$, Alexander Makarov\$, Nagarjuna Nagaraj‡, Juergen Cox‡, Matthias Mann‡¶, and Stevan Horning\$¶

TABLE II

A, Peptide identification from HeLa lysate triplicate analysis on a Q Exactive (90 min gradient)

	MS spectra	MSMS spectra	Identifications [%]	Unique peptides	Proteins	Isotope clusters
HeLa (1)	5427				3	146138
HeLa (2)	5098				1	143556
HeLa (3)	5274				7	144336
Σ Triplicates						

Q Exactive: 37.9%

B, Peptide identification from HeLa lysate triplicate analysis on an LTQ Orbitrap Velos (90 min gradient)

	MS spectra	MSMS spectra	Identifications [%]	Unique peptides	Proteins	Isotope clusters
HeLa (1)						125738
HeLa (2)						120553
HeLa (3)						126717
Σ Triplicates						

LTQ Orbitrap Velos: 56.2%



Identification Rate

- **Present**
 - What is the current rate?

- **Past**
 - Why low in the past?

- **Prospect**
 - How high in the future?



Identification Rate

- **Present**
 - **What is the current rate?**
- **Past**
 - Why low in the past?
- **Prospect**
 - How high in the future?



ID Rate of High Resolution Data

Molecular & Cellular Proteomics 2011

Research

Author's Choice

© 2011 by The American Society for Biochemistry and Molecular Biology, Inc.
This paper is available on line at <http://www.mcponline.org>

Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer*

Annette Michalski‡, Eugen Damoc\$, Jan-Peter Hauschild\$, Oliver Lange\$, Andreas Wieghaus\$, Alexander Makarov\$, Nagarjuna Nagaraj‡, Juergen Cox‡, Matthias Mann‡¶, and Stevan Horning\$¶

TABLE II

A, Peptide identification from HeLa lysate triplicate analysis on a Q Exactive (90 min gradient)

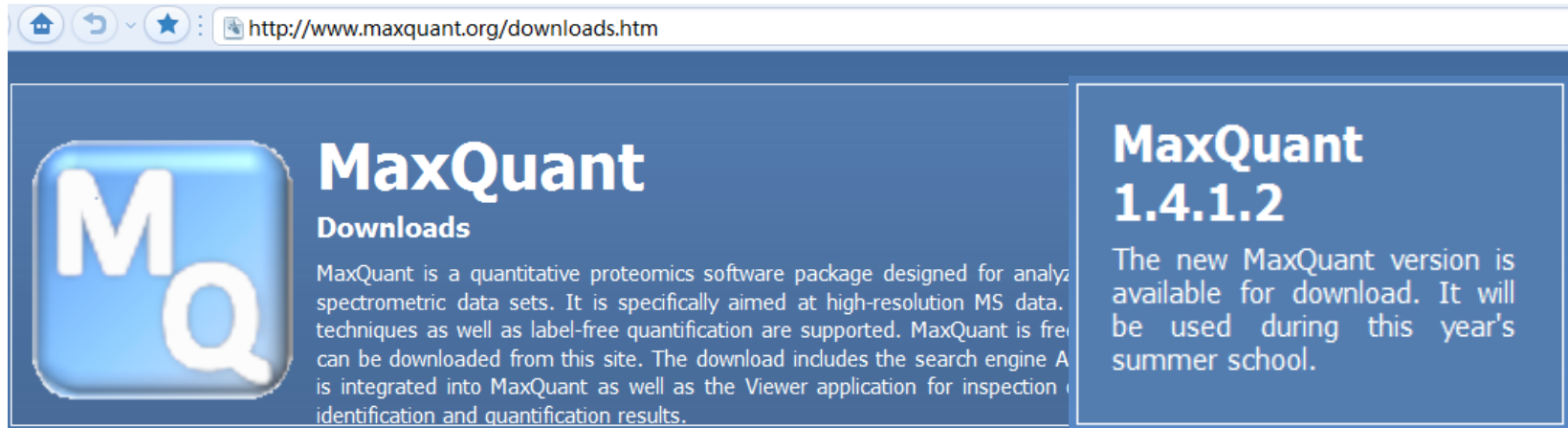
	MS spectra	MSMS spectra	Identifications [%]	Unique peptides	Proteins	Isotope clusters
HeLa (1)	5427	35203	37.23	12298	2513	146138
HeLa (2)	5098	35911	38.35	12830	2601	143556
HeLa (3)	5274	35348	38.23	12560	2557	144336
Σ Triplicates			37.94	16255	2864	

B, Peptide identification from HeLa lysate triplicate analysis on an LTQ Orbitrap Velos (90 min gradient)


	MS spectra	MSMS spectra	Identifications [%]	Unique peptides	Proteins	Isotope clusters
HeLa (1)						125738
HeLa (2)						120553
HeLa (3)						126717
Σ Triplicates			56.21	14401	4474	

LTQ Orbitrap Velos: 56.2%

Software: MaxQuant and pFind



http://www.maxquant.org/downloads.htm



MaxQuant

Downloads


MaxQuant is a quantitative proteomics software package designed for analyzing spectrometric data sets. It is specifically aimed at high-resolution MS data. Techniques as well as label-free quantification are supported. MaxQuant is free and can be downloaded from this site. The download includes the search engine Andromeda, which is integrated into MaxQuant as well as the Viewer application for inspection of identification and quantification results.

MaxQuant 1.4.1.2

The new MaxQuant version is available for download. It will be used during this year's summer school.



http://pfind.ict.ac.cn/downloads.html



pFind Studio version 2.8 beta2 for Windows

Dec. 31st, 2012

pFind Studio 2.8 including [pFind](#), [pBuild](#), [pLabel](#), [pXtract](#), [pParse](#), and [pScan](#) is available now.



Search Parameters

Parameters	Values
Target Database	Uniprot_Human + 286 Contaminants
Decoy Database	MQ: Reversal + Swapping pFind: Reversal
Digestion	Trypsin, Specific, Up to 2 Missing Cleavage Sites
Fixed Modifications	Carbamidomethyl (C)
Variable Modifications	Acetyl (Protein N-term), Oxidation (M)
Raw Extraction	MQ: MaxQuant pFind: pXtract
Mixture Spectra Extraction	MQ: "Second Peptide" closed pFind: pParse closed
MS1 Precursor Tolerance	±20 ppm
MS2 Fragment Tolerance	±20 ppm
Validation	FDR ≤ 1% at Spectrum Level



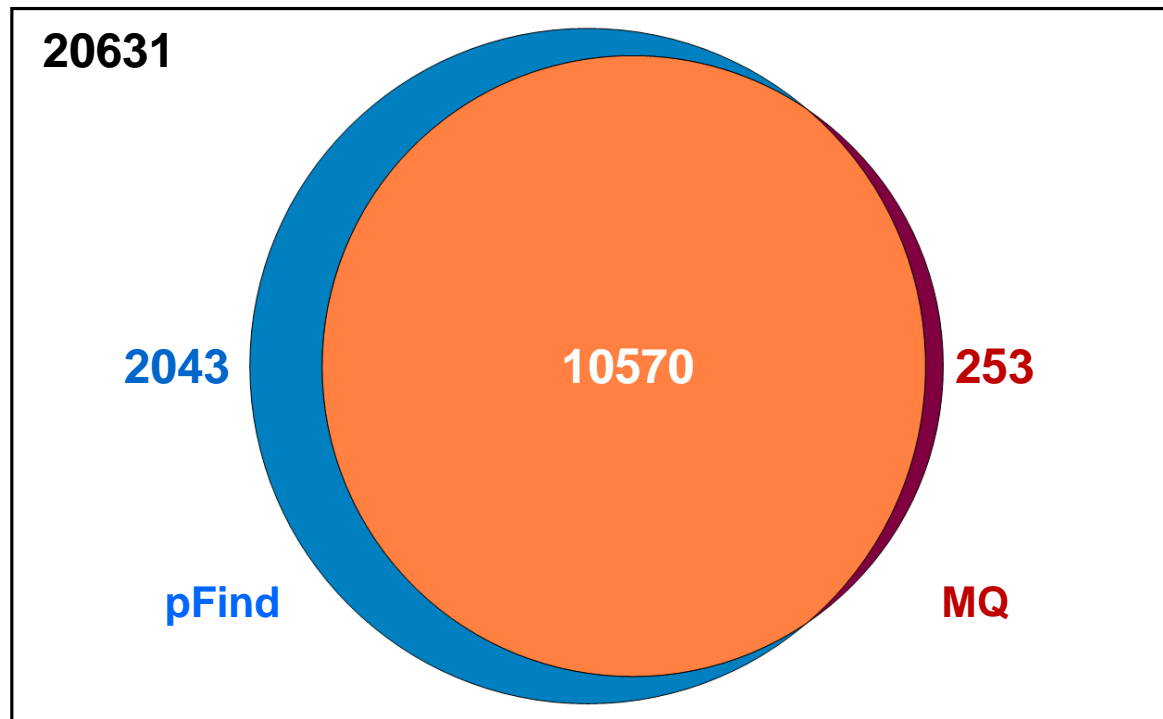
MS1 = ± 20 ppm | MS2 = ± 20 ppm

MQ = 10873 | 52.7 %

MQ \cap pFind = 10570 | 51.2 %

pFind = 12663 | 61.4 %

MQ \cup pFind = 12916 | 62.6 %



Identification Rate

- **Present**
 - **Q: What is the current rate?**
 - **A: 50% ~ 60%** ✓
- **Past**
 - **Why low in the past?**
- **Prospect**
 - **How high in the future?**



Identification Rate

- Present
 - What is the current rate?
- **Past**
 - **Why low in the past?**
- Prospect
 - How high in the future?



- Use HCD data to simulate CID data
- Use HCD results as benchmarks

Instrument	Fragmentation	Mode	MS1	MS2
Orbitrap Velos	HCD	High-High	± 20 ppm	± 20 ppm
Orbitrap XL	CID	High-Low	± 20 ppm	± 0.5 Da
Q-TOF	CID	Low-Low	± 0.2 Da	± 0.2 Da
LTQ	CID	Low-Low	± 2.0 Da	± 0.8 Da



ID Rate: MaxQuant

Instrument	Dissociation	Mode	MS1	MS2	MaxQuant
Orbitrap Velos	HCD	High-High	± 20 ppm	± 20 ppm	52.7 %
Orbitrap XL	CID	High-Low	± 20 ppm	± 0.5 Da	47.8 %
Q-TOF	CID	Low-Low	± 200 ppm	± 0.2 Da	37.0 %
LTQ	CID	Low-Low	± 2000 ppm	± 0.8 Da	24.6 %



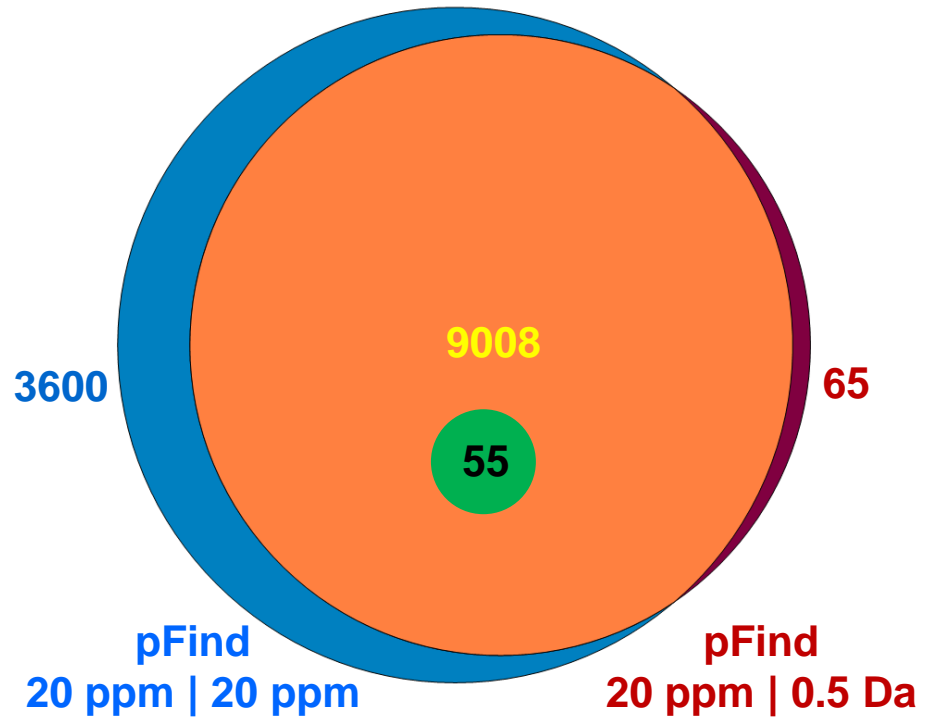
ID Rate: pFind

Instrument	Dissociation	Mode	MS1	MS2	pFind
Orbitrap Velos	HCD	High-High	± 20 ppm	± 20 ppm	61.4%
Orbitrap XL	CID	High-Low	± 20 ppm	± 0.5 Da	44.2%
Q-TOF	CID	Low-Low	± 0.2 Da	± 0.2 Da	40.5%
LTQ	CID	Low-Low	± 2.0 Da	± 0.8 Da	33.3%



Why: 20 ppm vs 0.5 Da

- **65 + 55** distinct so few
- **9008** consistent so many
- **3600** unidentified so many

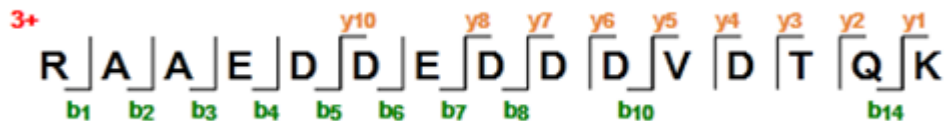
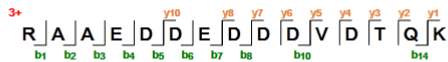




Low Resolution Result

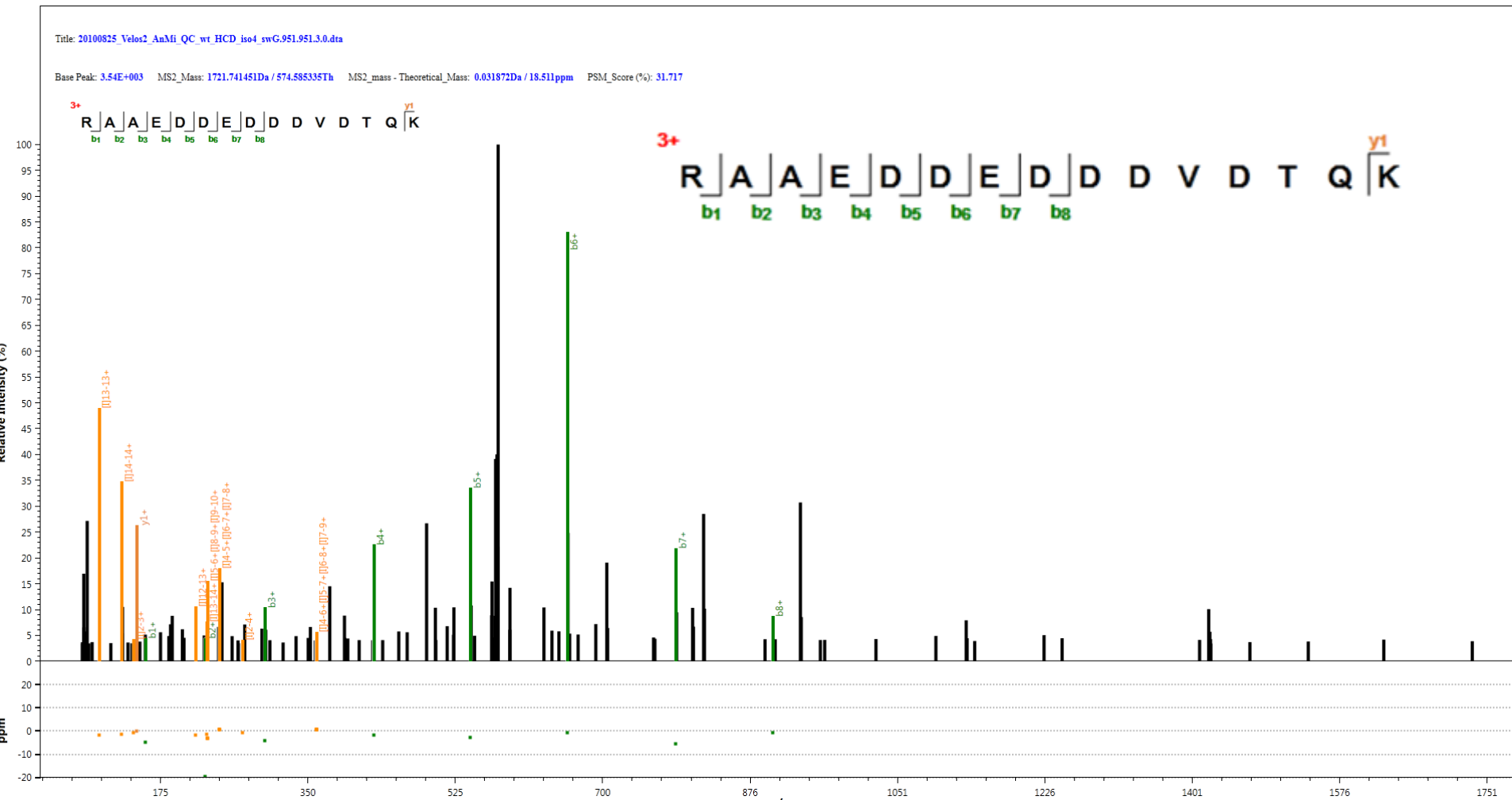
Title: 20100825_Velos2_AnMi_QC_wt_HCD_iso4_swG.951.951.3.0.dta

Base Peak: 3.54E+003 MS2_Mass: 1721.741451Da / 574.585335Th MS2_mass - Theoretical_Mass: 0.031872Da / 18.511ppm PSM_Score (%): 30.466





The Same Peptide in High Resolution Mode

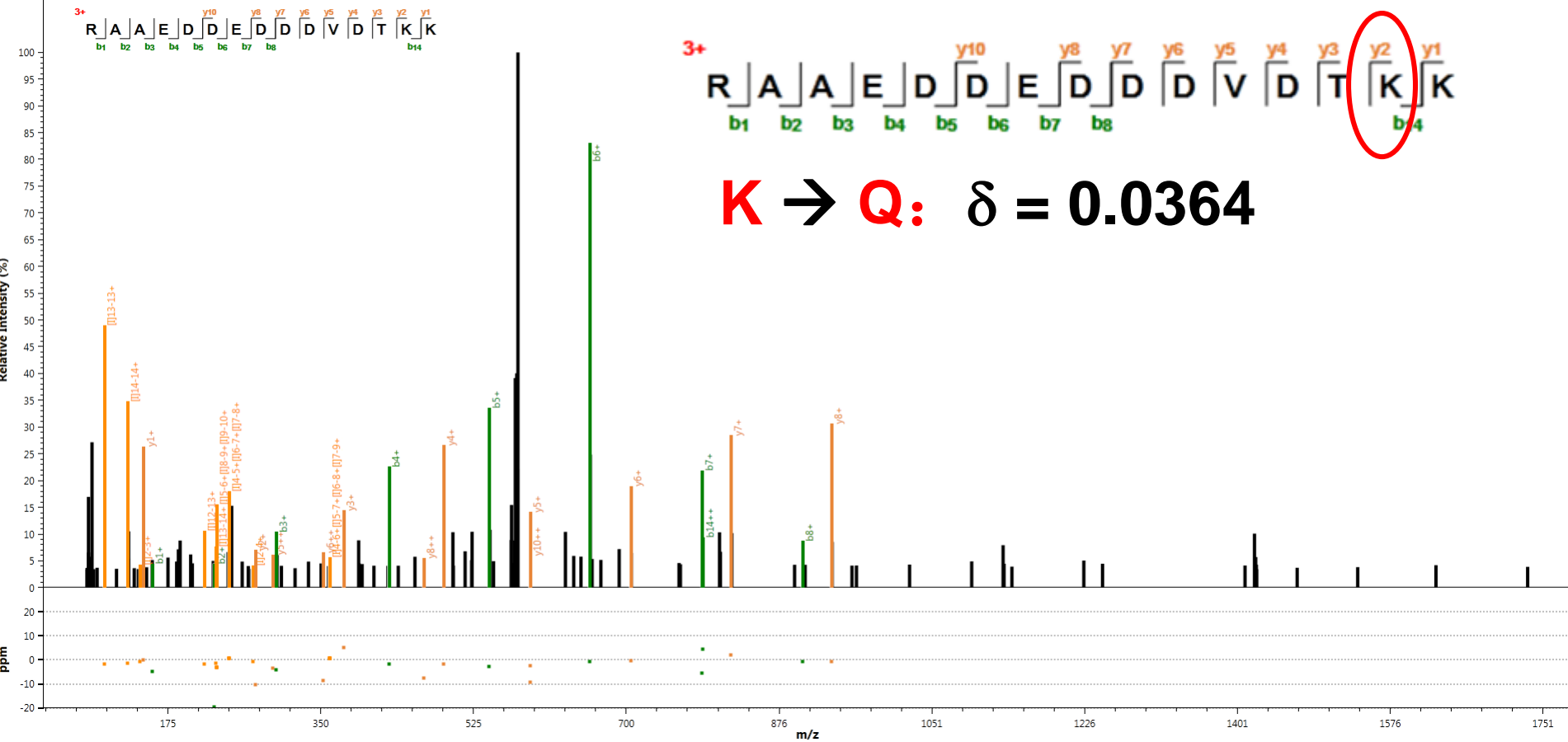




Actually...

Title: 20100825_Velos2_AnMi_QC_wt_HCD_iso4_swG.951.951.3.0.dta

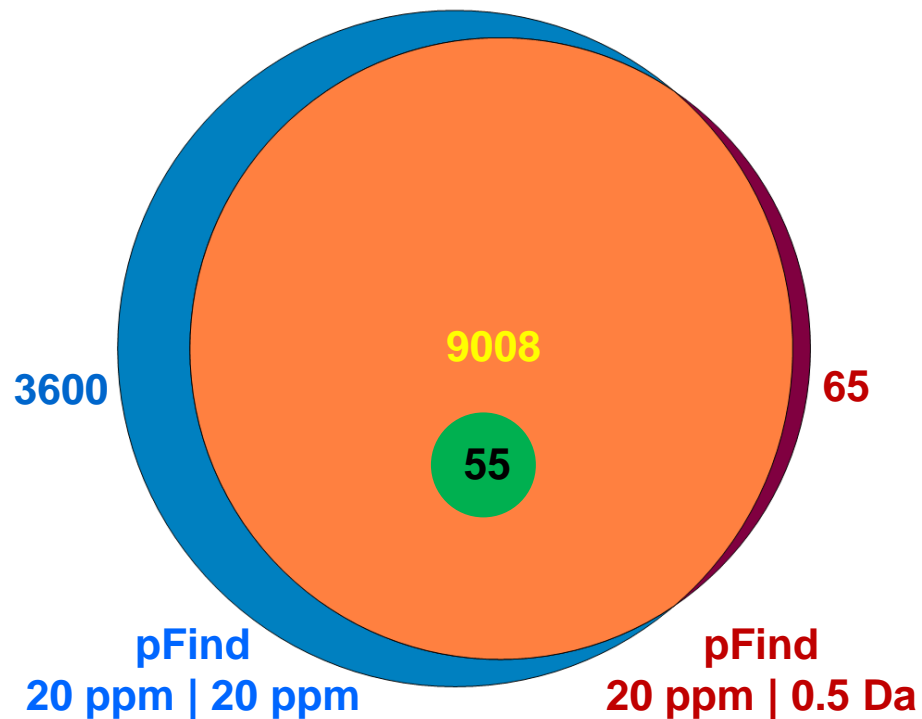
Base Peak: 3.54E+003 MS2_Mass: 1721.741451Da / 574.585335Th MS2_mass - Theoretical_Mass: -0.004513Da / -2.621ppm PSM_Score (%): 44.015





Why: 20 ppm vs 0.5 Da

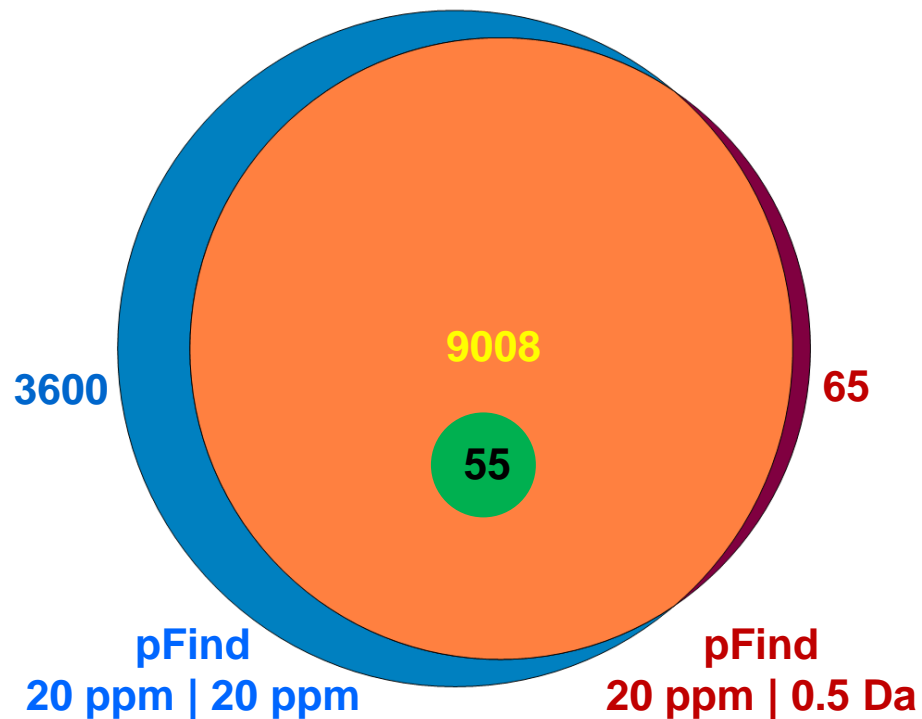
- **65 + 55** distinct
so few
- **9008** consistent
so many
- **3600** unidentified
so many



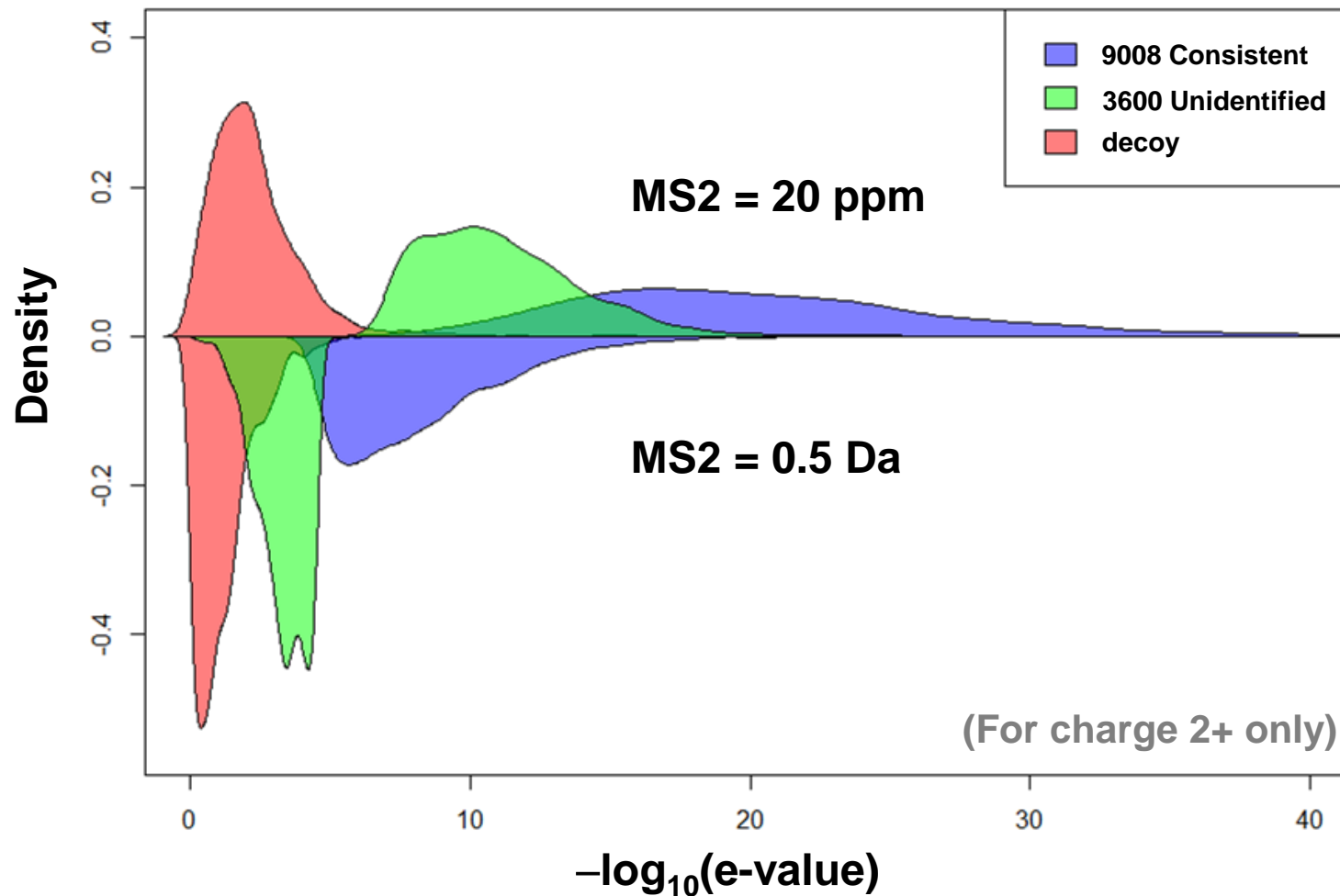


Why: 20 ppm vs 0.5 Da

- **65 + 55** distinct
so few
- **9008** consistent
so many
- **3600** unidentified
so many



Why Unidentified?

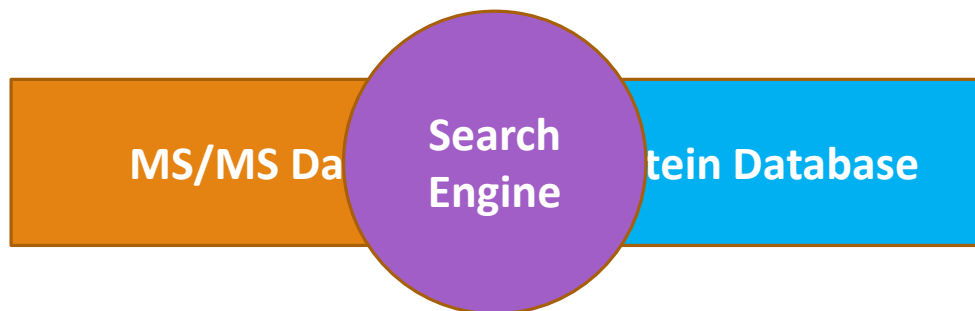


Identification Rate

- Present
 - What is the current rate?
- **Past**
 - **Q: Why low in the past?**
 - **A: Low discrimination power.** ✓
- Prospect
 - How high in the future?

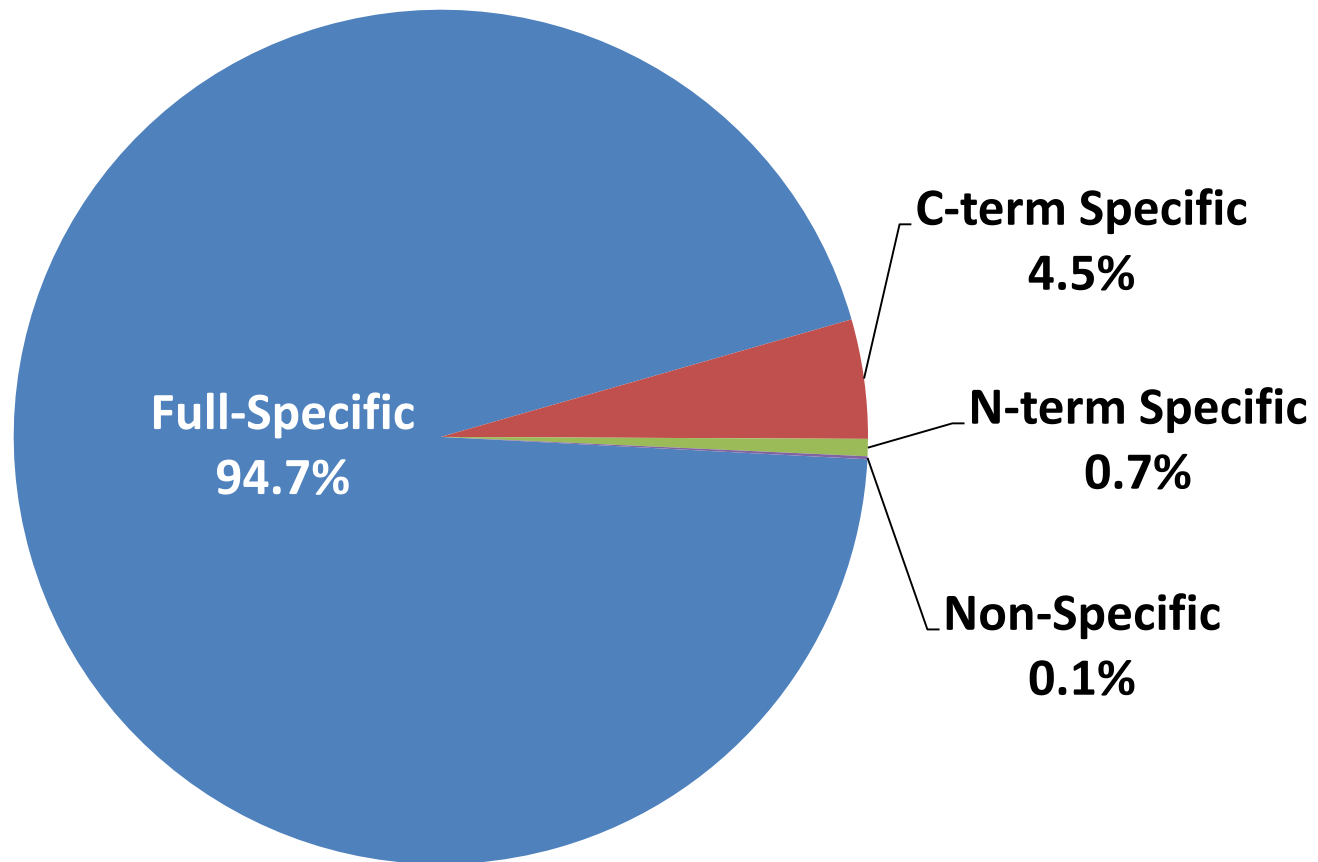
Identification Rate

- Present
 - What is the current rate?
- Past
 - Why low in the past?
- **Prospect**
 - **How high in the future?**

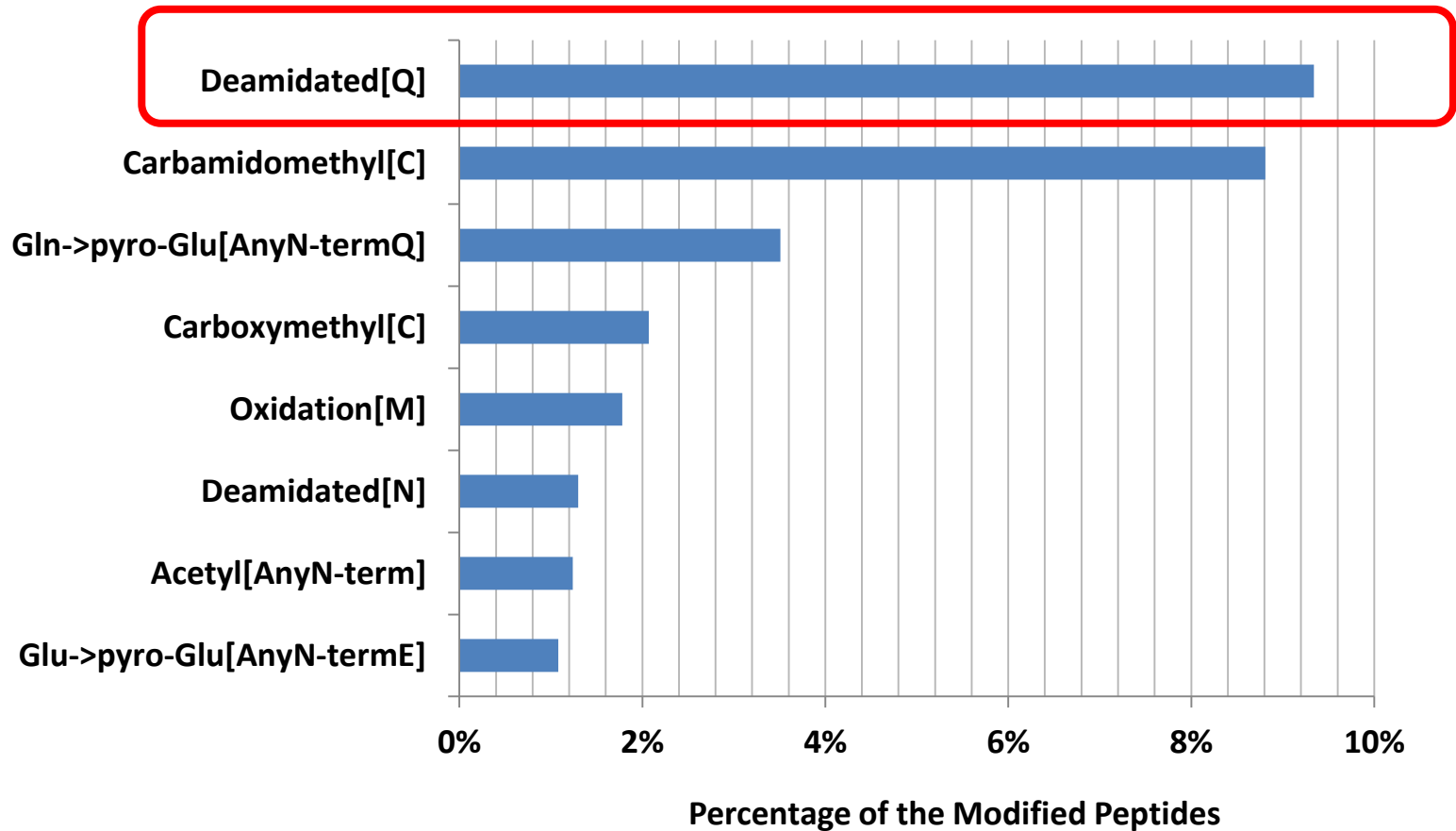




1. Different Digestion Manners



2. Modifications



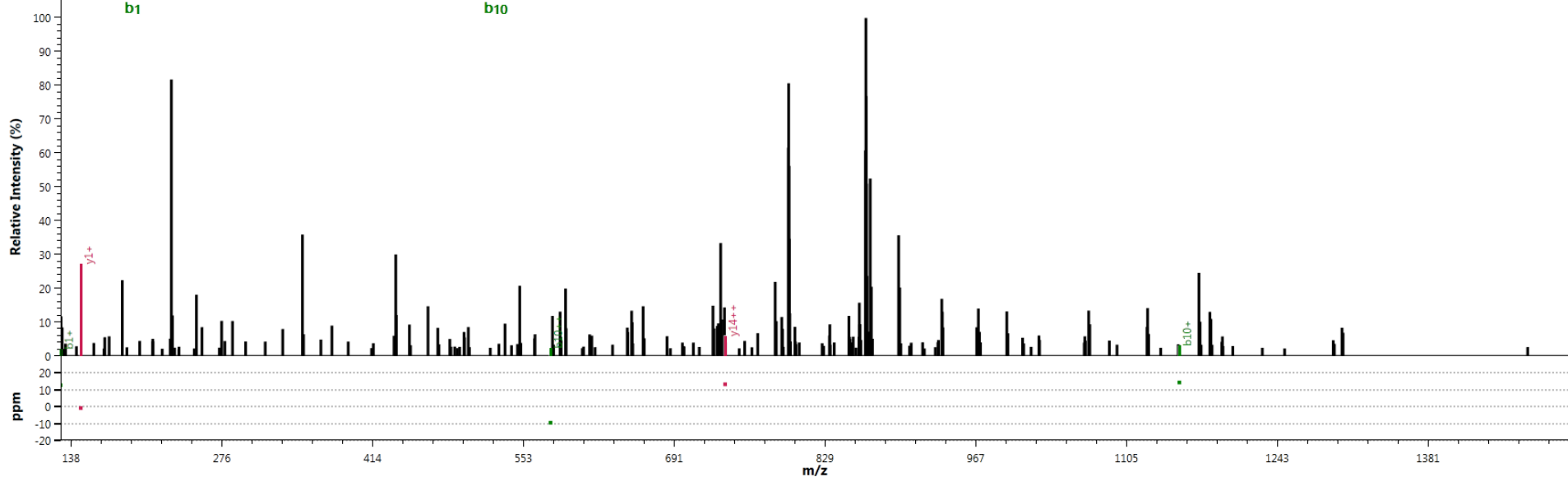


Deamidated[Q]

Title: 20100825_Velos2_AnMi_QC_wt_HCD_iso4_swG.21355.21355.3.0.dia

Base Peak: 5.10E+005 MS2_Mass: 2596.332338Da / 866.11563Th MS2_mass - Theoretical_Mass: 73.004205Da / 28118.205ppm PSM_Score (%): 1.802

3+ Q I H S M I A R I C P N D T G G L R S L V N K y14 b1 b10 y1



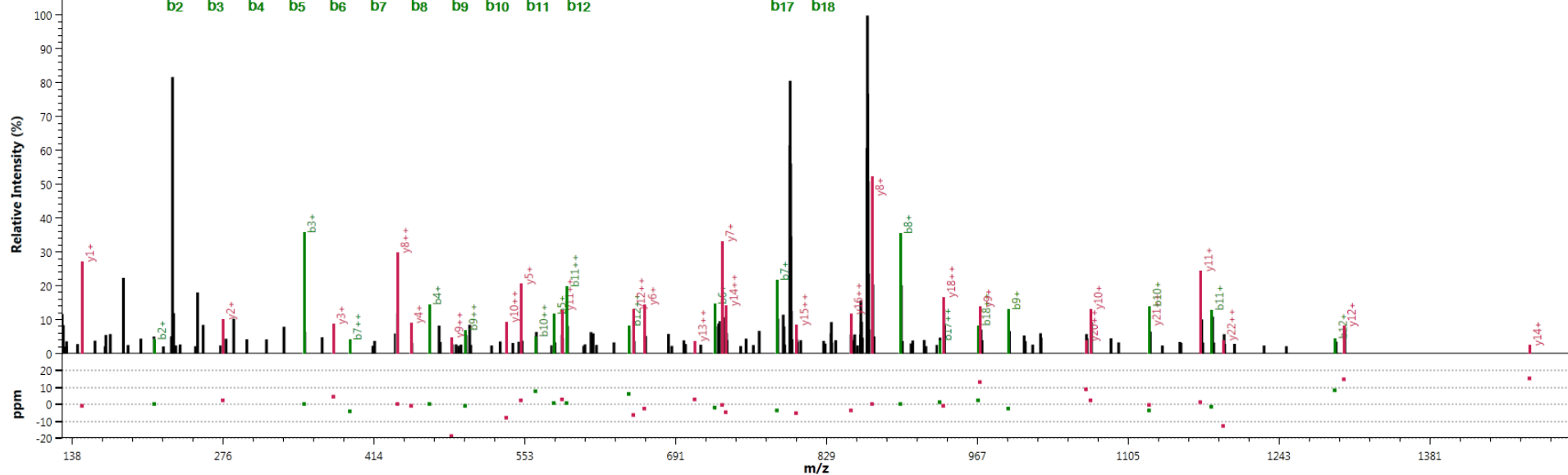


Deamidated[Q]

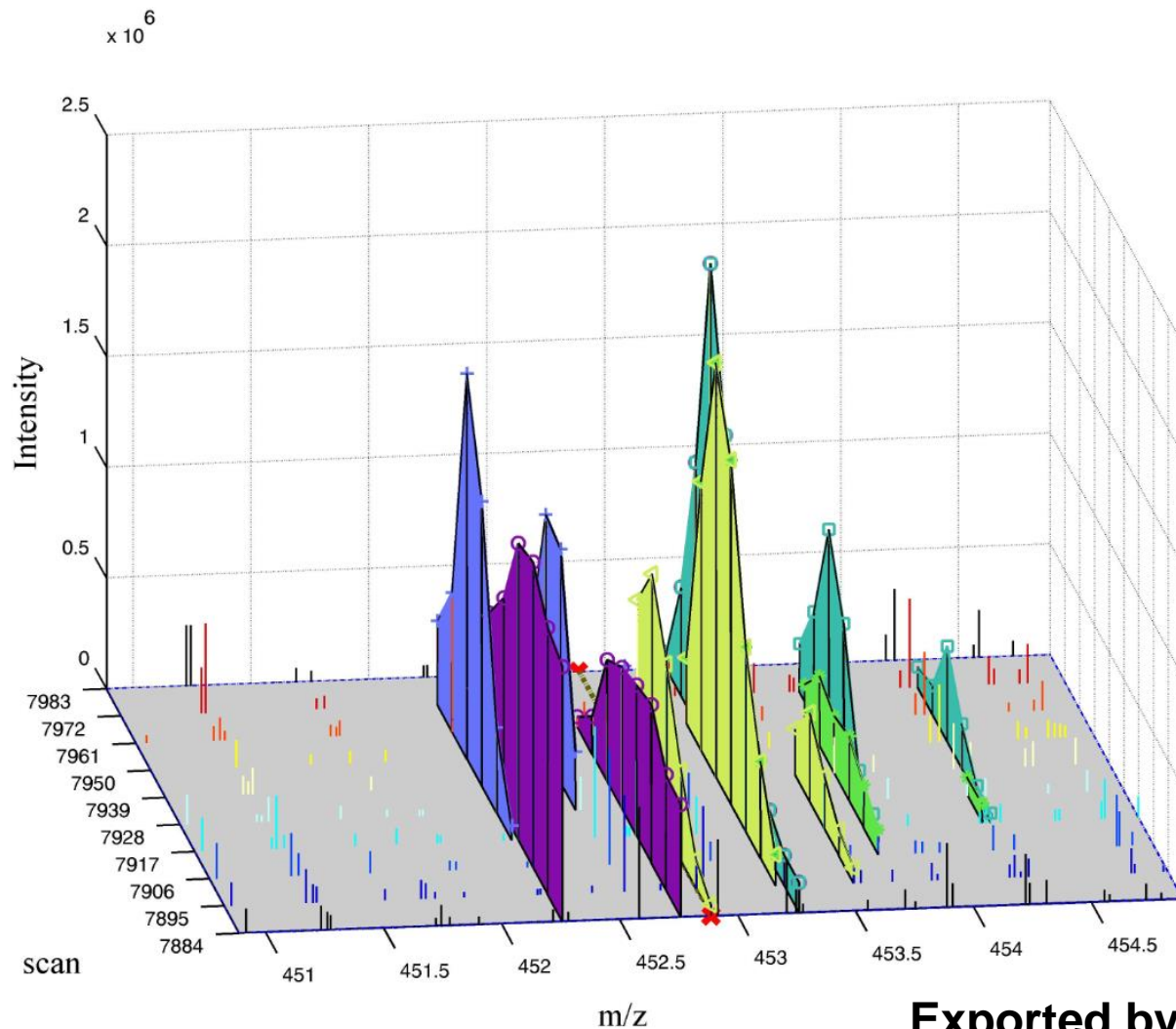
Title: 20100825_Velos2_AnMi_QC_wt_HCD_iso4_swG.21355.21355.3.0.data Mods: 23,Deamidated[Q](None);

Base Peak: 5.10E+005 MS2_Mass: 2596.332338Da / 866.11563Th MS2_mass - Theoretical_Mass: -0.011489Da / -4.425ppm PSM_Score (%): 27.249

3+ V I H D N F G I V E G L M T T V H A I T A T Q K
b2 b3 b4 b5 b6 b7 b8 b9 b10 b11 b12 b17 b18



3. Mixture Spectra



Exported by pParse 2.0



3. Mixture Spectra

Title: 20100825_Velos2_AnMi_QC_wt_HCD_iso4_swG.7934.7934.3.4.dta

Base Peak: 2.41E+005 MS2_Mass: 1354.671894Da / 452.228816Th MS2_mmass - Theoretical_Mass: 451.157364Da / 333038.108ppm PSM_Score (%): 41.152

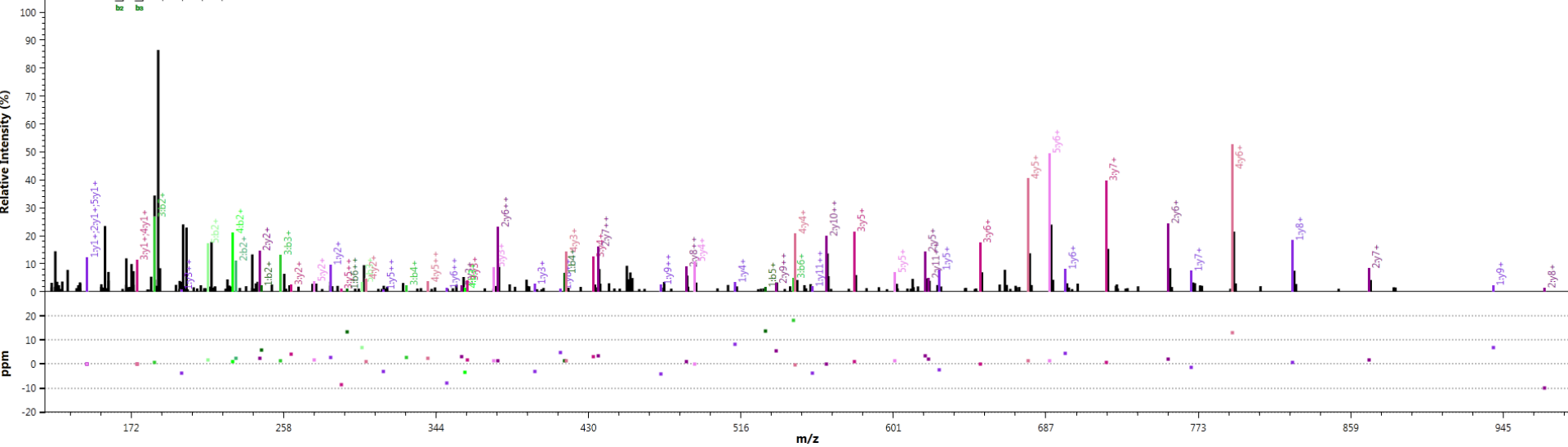
3+ I M_{bz}G_{bz} N_{bz}L_{bz}G_{bz}A_{bz} A_{bz}D_{bz}I_{bz} D_{bz}H_{bz}K_{bz}

3+ D D_{bz}G_{bz}T_{bz} V_{bz}I_{bz}H_{bz}F_{bz} N_{bz}P_{bz}K_{bz}

2+ L A_{bz}A_{bz}A_{bz}F_{bz} A_{bz}V_{bz}S_{bz}R_{bz}

2+ L D_{bz}M_{bz}E_{bz}I_{bz}E_{bz}R_{bz}

2+ D V_{bz}S_{bz}L_{bz} L_{bz}T_{bz}Q_{bz}K_{bz}





Improve the ID rate

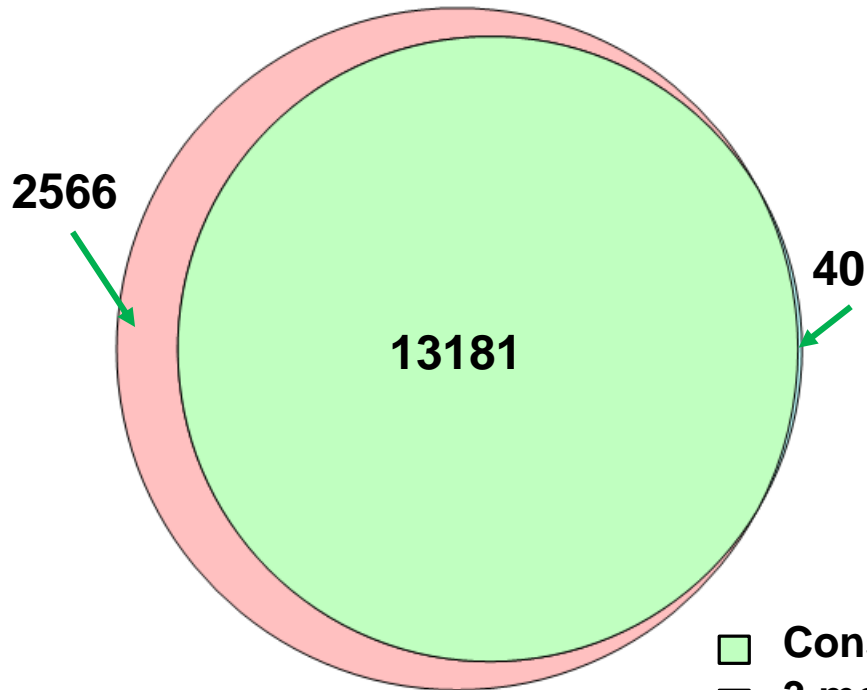
Parameters	Values
Target Database	Uniprot_Human + 286 Contaminants
Decoy Database	MQ: Reversal + Swapping pFind: Reversal
Digestion	Trypsin, Specific, Up to 2 Missing Cleavage Sites
Fixed Modifications	NULL
Variable Modifications	Acetyl(Protein N-term), Carbamidomethyl(C), Oxidation(M), Carboxymethyl(C), Deamidation(NQ), Gln→pyroGlu(AnyN-termQ)
Raw Extraction	MQ: MaxQuant pFind: pXtract
Mixture Spectra Extraction	MQ: "Second Peptide" open pFind: pParse open
MS1 Precursor Tolerance	±7 ppm [To differentiate 0.984 from 1.003]
MS2 Fragment Tolerance	±20 ppm
Validation	FDR ≤ 1% at Spectrum Level



Higher ID Rates

pFind 2.8

ID Rate: 76%



MaxQuant

ID Rate: 65%



- Consistent
- 3 modifications
- 6 modifications



Search Speed

- **Dell Precision T1500 (8-Core PC)**
- **Routine Search**
 - pFind = 10 min, MQ = 45 min
- **Open pFind Search**
 - 4 hours
- **Re-Search**
 - pFind = 25 min, MQ = 115 min



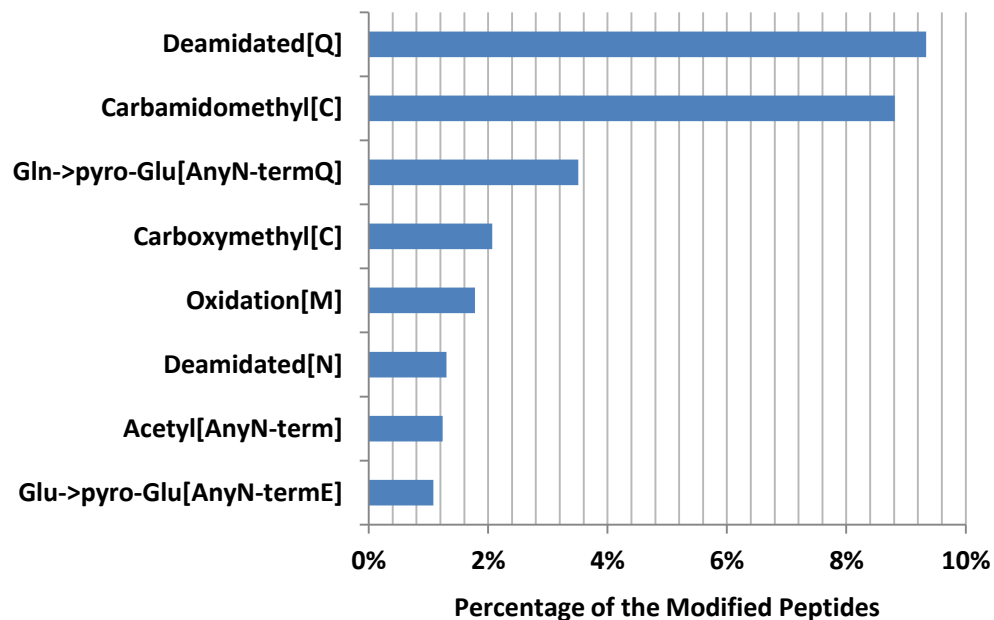
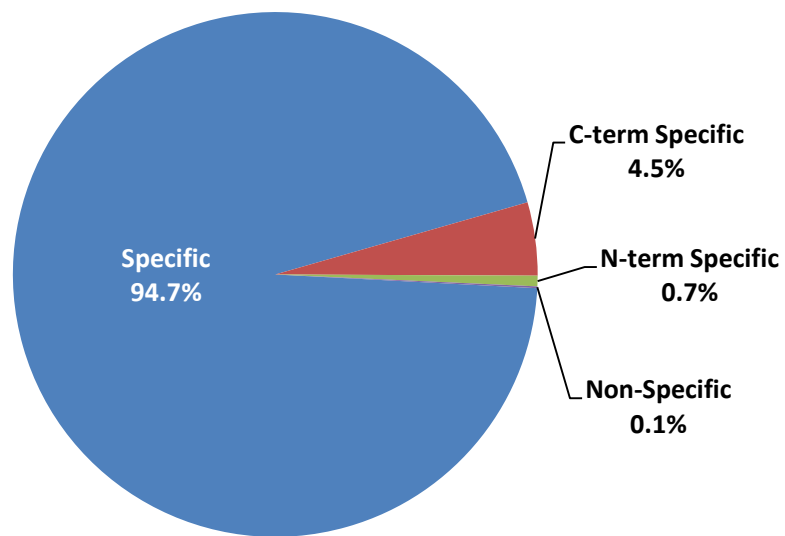
Space and Speed

	Avg. Peptides	Amplification	Time Used	
			pFind 2.8	MaxQuant
Full-Specific	349	1	/	
Semi-Specific	5,925	17		
Non-Specific	26,895	77		
Regular Search	625	2	10 min	45 min
Complex Modifications	2298	7	25 min (+240 min)	115 min

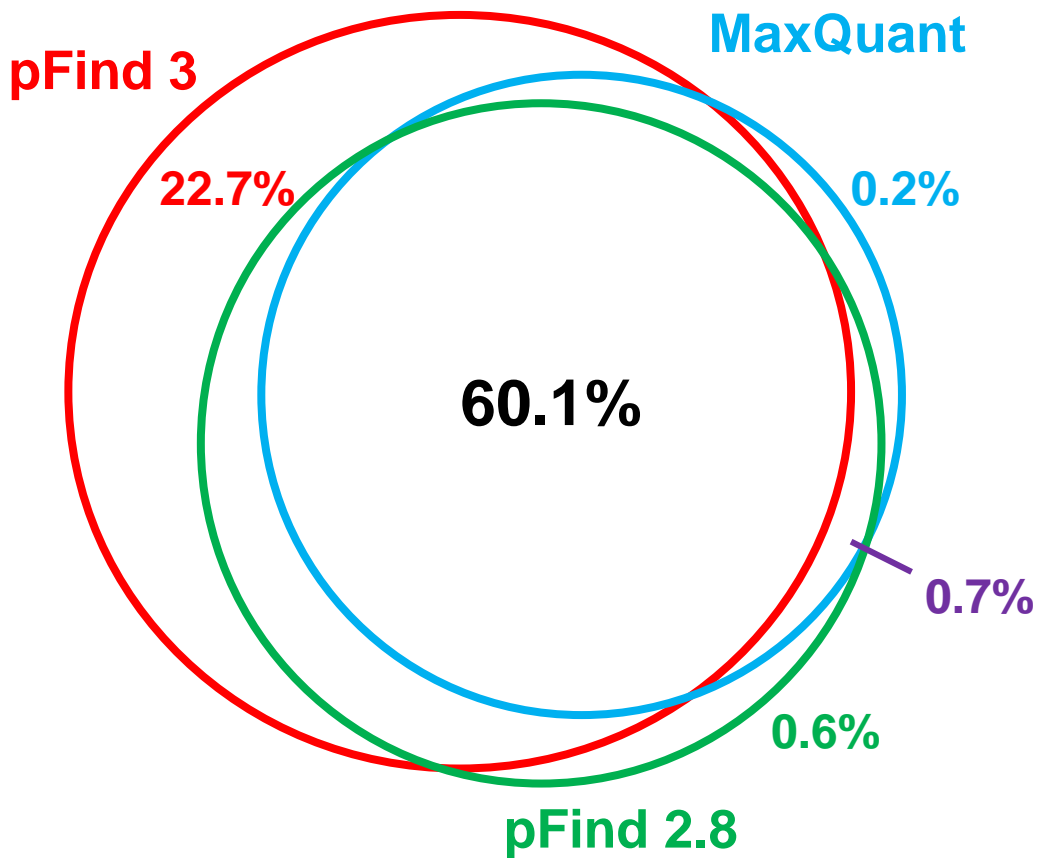
Uniprot-Human, Peptide length: 6 ~ 60, Digested by Trypsin

pFind 3: Deep Analysis

- Identification Rate of MS/MS scans
 - 17630 | **85.5%**



Identified MS2 Scans





Space and Speed

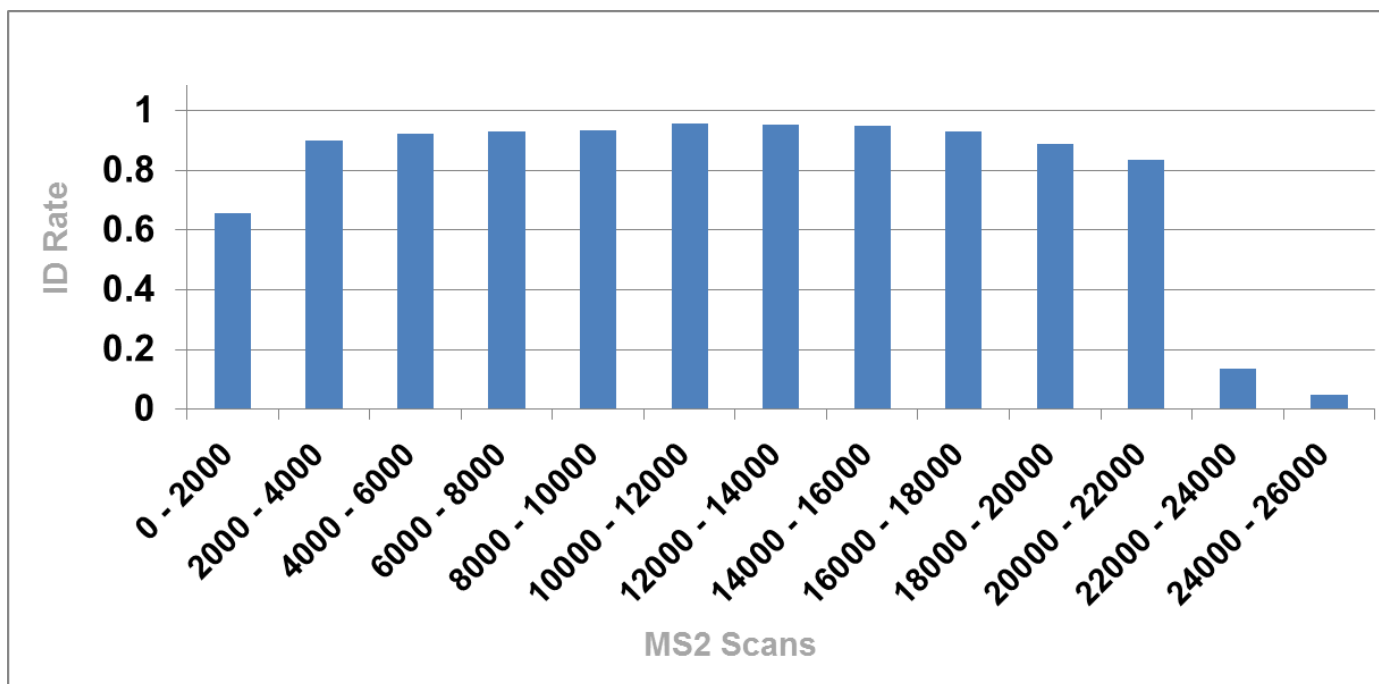
	Avg. Peptides	Amplification	Time Used		
			pFind 2.8	MaxQuant	pFind 3
Full-Specific	349	1	/		
Semi-Specific	5,925	17			
Non-Specific	26,895	77			
Regular Search	625	2	10 min	45 min	/
Complex Modifications	2298	7	25 min (+240 min)	115 min	/

Uniprot-Human, Peptide length: 6 ~ 60, Digested by Trypsin

~ 20 spectra are identified per second

Unidentified Spectra?

- MS2 Scan: 2,000 ~ 20,000
 - 80% of total spectra
- ID Rate: 93.7%





More Datasets

	Instrument	No. Spectra	ID Rate	Speed (spec. / sec.)
1	LTQ Orbitrap Velos	64,112	56% → 86%	17
2	LTQ Orbitrap Velos	486,411	30% → 61%	51
3	Q Exactive	136,560	38% → 71%	16
4	Q Exactive	1,934,361	16% → 70%	54



Identification Rate

- Present
 - What is the current rate?
- Past
 - Why low in the past?
- **Prospect**
 - **Q: How high in the future?**
 - **A: 60% ~ 80%** ✓



Discussion: Deep or Fast?

- **Only Restricted Search**
 - Fast but simple

- **Only Open Search**
 - Deep but slow

- **High Resolution + New Algorithm = Deep and Fast**

Ten Years ago: 2003

Data analysis—the Achilles heel of proteomics

Scott D. Patterson

March 2003 · Volume 21

Nature Biotechnology

During the past few years there has been a resurgence of research using parallel protein-based analysis, now commonly referred to as proteomics. However, our ability to generate data now outstrips our ability to analyze it. This occurs even though proteomics is inherently a substrate-limited science and proteins exist over a wide concentration range in biological samples. Therefore, it is not surprising that the entire proteome of any species has yet to be observed. In this article, I address some of the primary issues currently facing researchers in this field, with an emphasis on the computational aspects affecting progress, including the accuracy of matches from mass spectrometric data to sequence databases and the integration of the results of proteomics experiments to yield biological meaning.

Parallel protein-based analysis first came to the fore during the mid-1970s with the introduction of two-dimensional gel electrophoresis, which for the first time allowed a staggering number of different protein species to be revealed in a single experiment and permitted the comparison of expression patterns between samples. At that

Table 1. Proteomics experiments require handling of diverse data sets

Stage of process	Type of data
Preparation for analysis	Project information Sample information Separation, fractionation
Sample processing	Quantitative analysis (LC-MS/MS, 2-DE MS/MS) Identification, MS/MS data analysis
Data analysis	Data capture and validation Data management and integration
Monitoring	All processes require quality assurance and quality control

numerous examples in which transcript amounts at steady state, or even after induction, do not correlate with the amount of protein present in the system because of well-known, but currently mostly unpredictable, effects of post-transcriptional and post-translational regulation¹. Therefore, analysis at the level of

“...our ability to generate data now outstrips our ability to analyze it”

sis problems because of the nature of the two main technologies used in studying proteins: first, systems based on two-dimensional gel electrophoresis (i.e., quantification through image analysis, proteolytic digestion of the isolated proteins of interest, and identification by mass spectrometry) and non-gel based systems (i.e., quantification and identification by mass spectrometry of mixtures of proteolytically digested proteins)¹.

Two-dimensional gel software continues to be improved, but after 20 years of development it still requires some manual intervention. Mass spectrometers measure the mass-to-charge ratio of charged molecules,



Ten years later: 2014 -

Data analysis—the Achilles heel of proteomics

Greetings from the pFind Team @ ICT. CAS. CHINA

