BROAD INSTITUTE

Fundamentals of Biological Mass Spectrometry and Proteomics

Steve Carr
Broad Institute of MIT and Harvard

Modern Mass Spectrometer (MS) Systems

Orbitrap

Q-Exactive

Triple Quadrupole

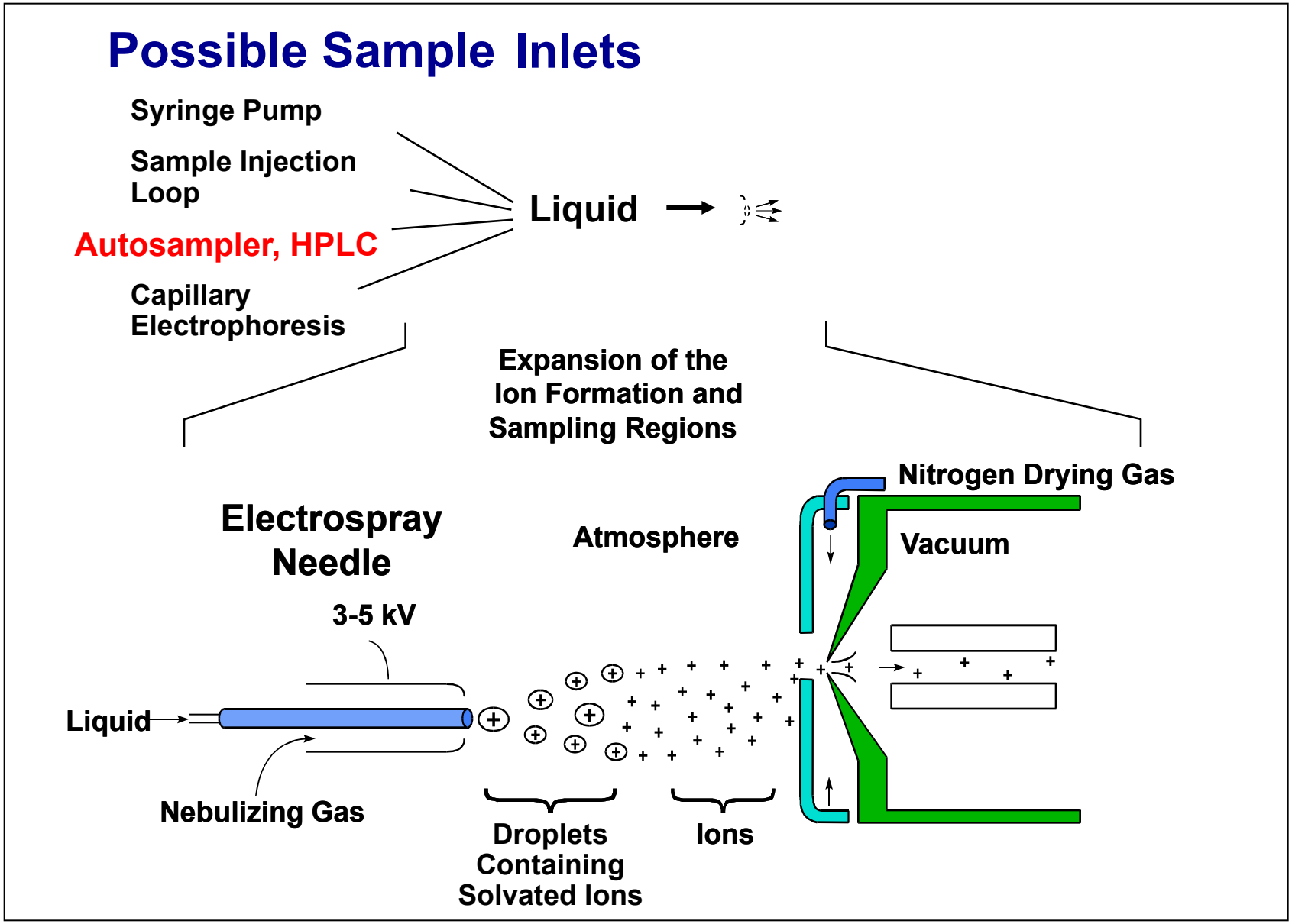
Discovery/Global Experiments

Targeted MS

MS systems used for proteomics have 4 tasks:

- Create ions from analyte molecules
- Separate the ions based on charge and mass
- Detect ions and determine their mass-to-charge
- Select and fragment ions of interest to provide structural information (MS/MS)

Electrospray MS: ease of coupling to liquid-based separation methods has made it the key technology in proteomics



Isotopes

Most elements have more than one stable isotope.

For example, most carbon atoms have a mass of 12 Da, but in nature, 1.1% of C atoms have an extra neutron, making their mass 13 Da.

Why do we care?

Mass spectrometers "see" the isotope peaks provided the resolution is high enough.

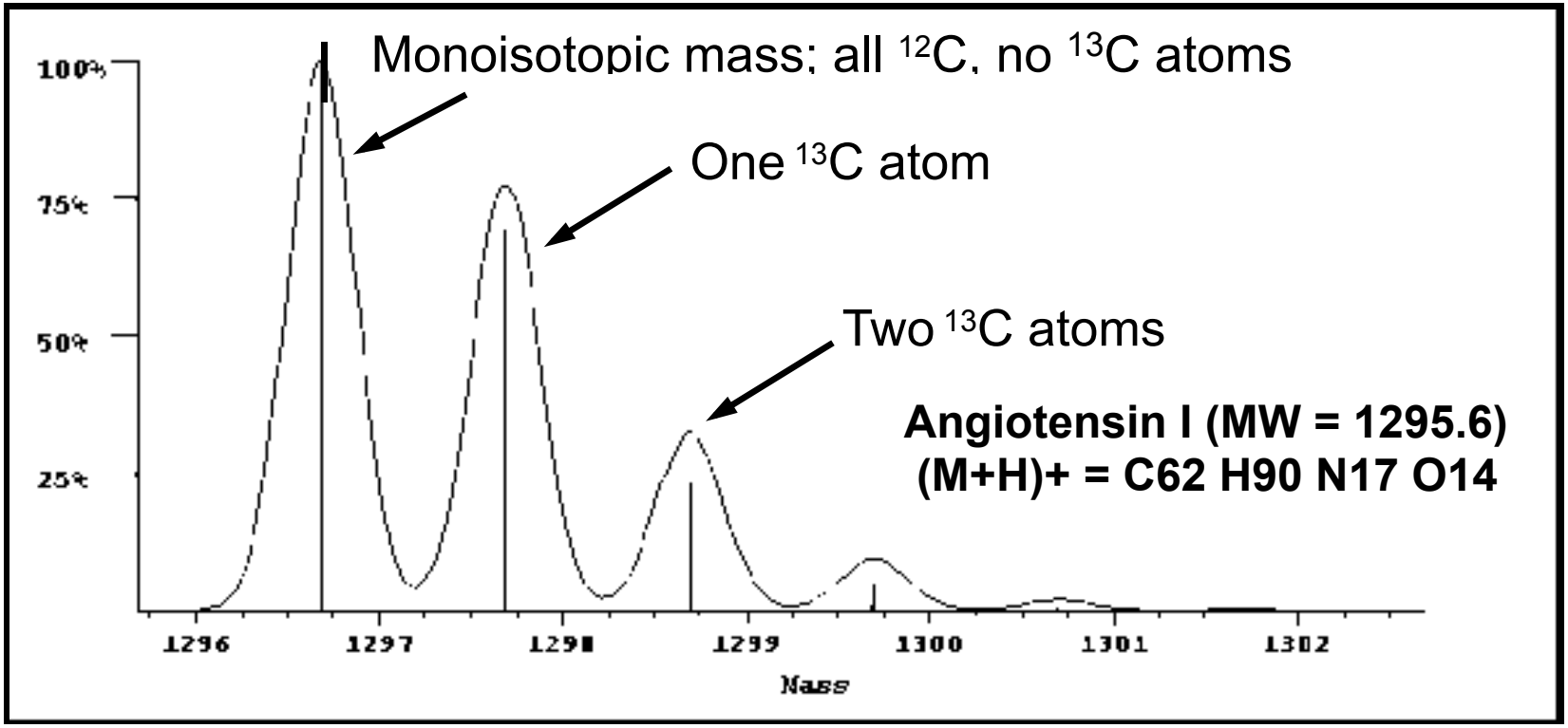
If an MS instrument has resolution high enough to resolve these isotopes, better mass accuracy is achieved.

Stable isotopes of most abundant elements of peptides

Element	Mass	Abundance
H	1.0078 2.0141	99.985% 0.015
C	12.0000 13.0034	98.89 1.11
N	14.0031 15.0001	99.64 0.36
O	15.9949 16.9991 17.9992	99.76 0.04 0.20

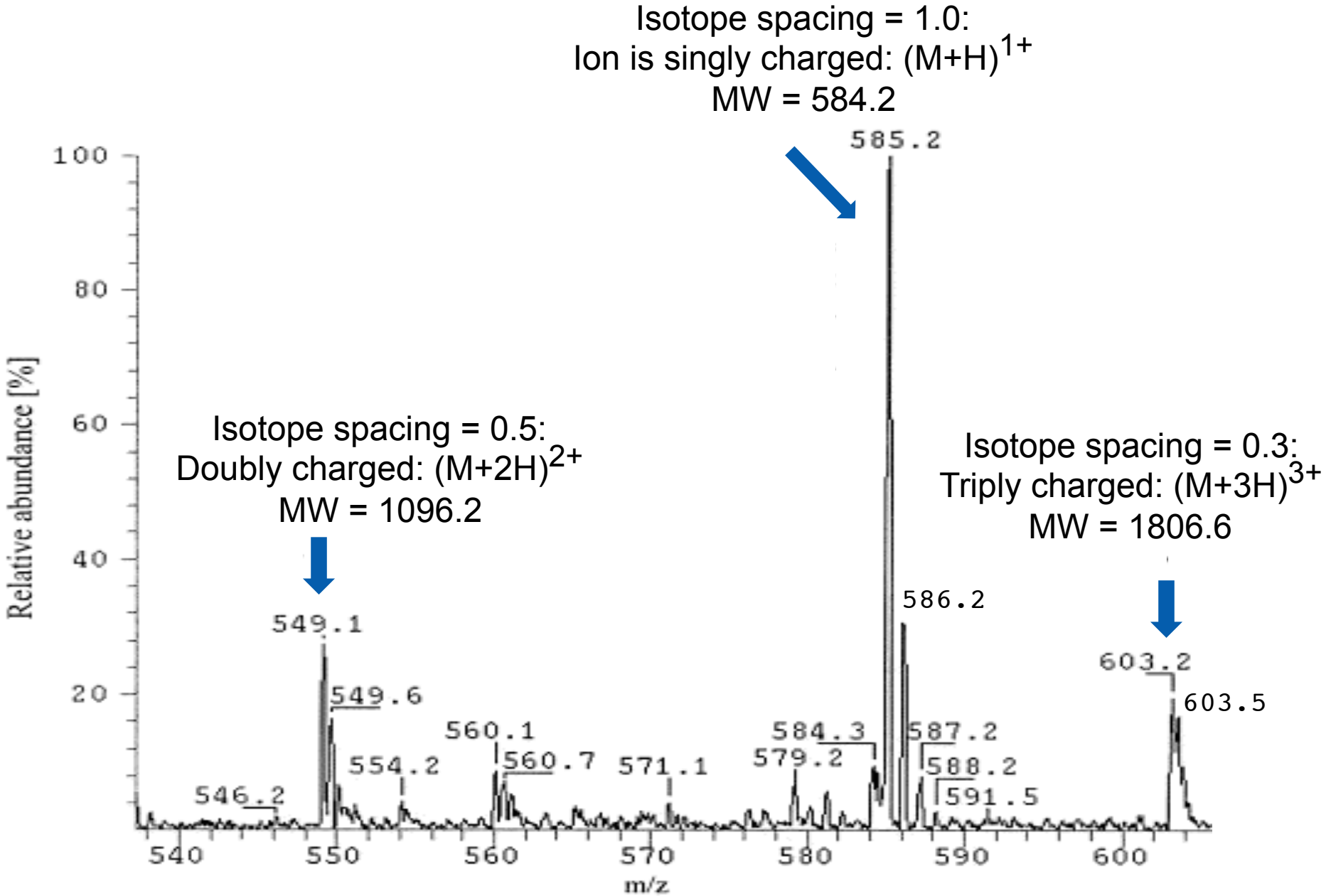
Monoisotopic mass and isotopes

We use instruments that resolve the isotopes enabling us to accurately measure the monoisotopic mass



The **monoisotopic mass** of a molecule is the sum of the accurate masses for the most abundant isotope of each element present. As the number of atoms of any given element increases, the percentage of the population of molecules having one or more atoms of a heavier isotope of this element also increases. The most significant contributors to the isotopic peak pattern for peptides is the 13C isotope of carbon (1.1%) and 15N peak of nitrogen (0.36%).

Example of electrospray mass spectrum of mixture of 3 peptides

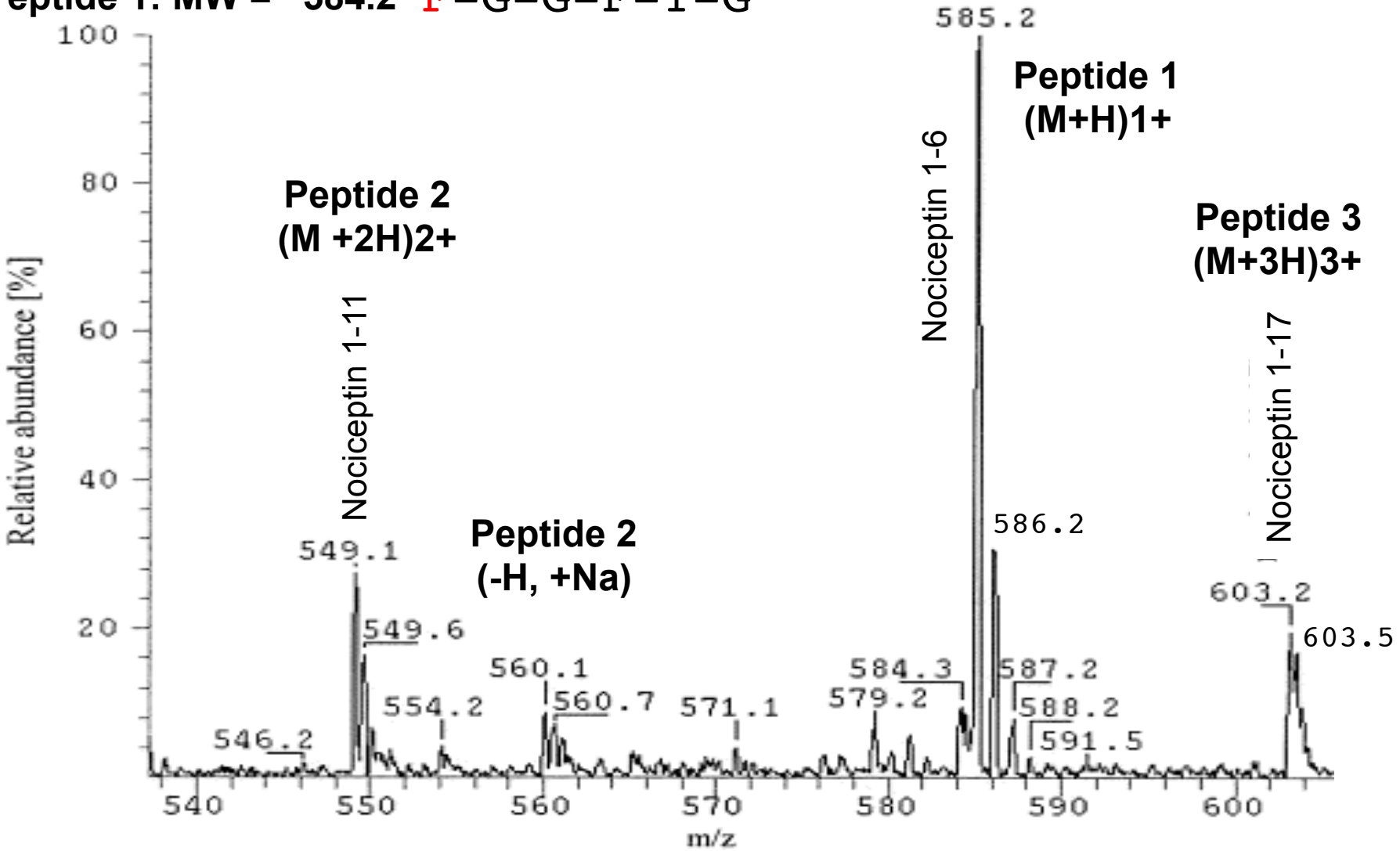


Example of electrospray mass spectrum of mixture of 3 peptides

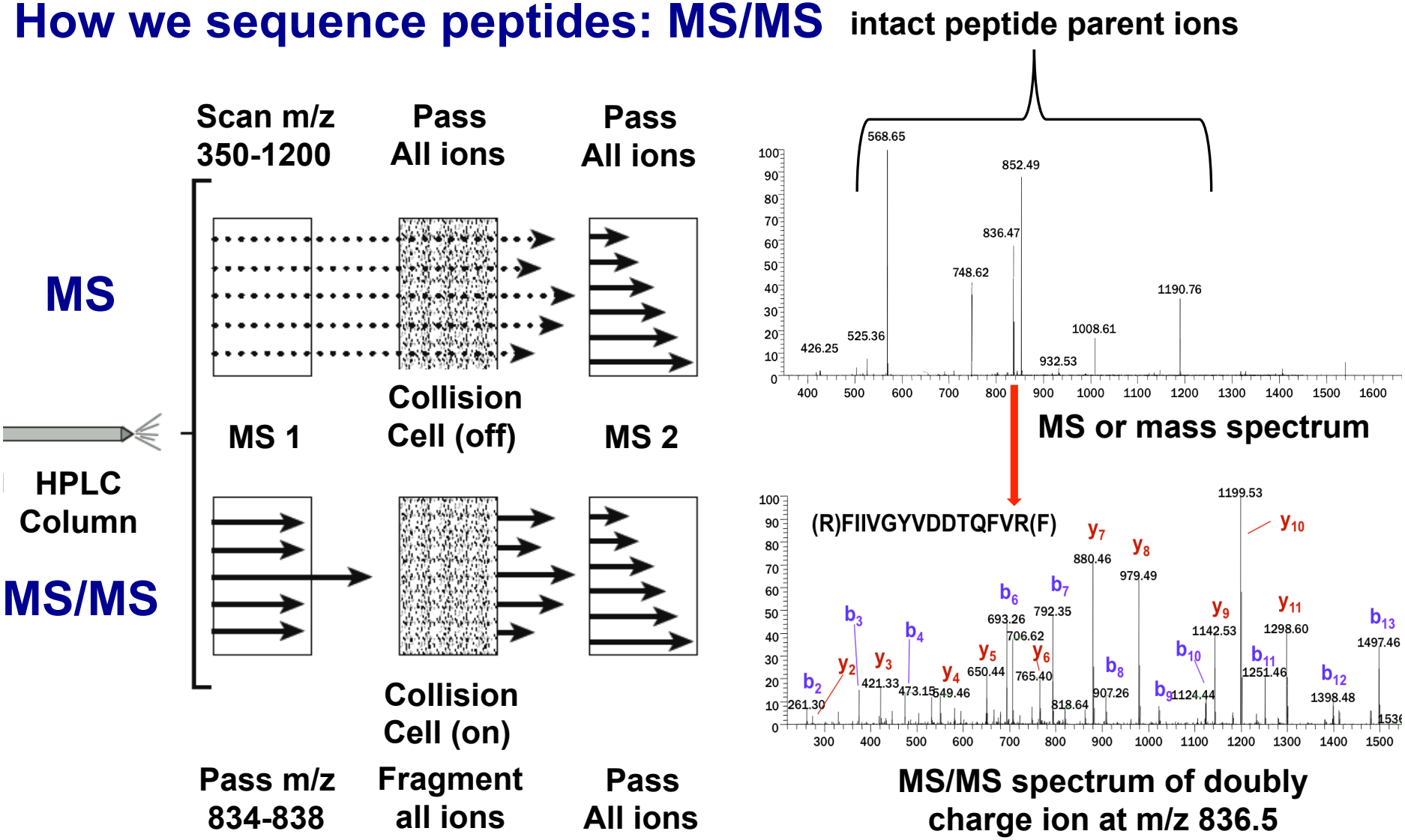
Peptide 3: MW = 1806.6 F–G–G–F–T–G–A–R–K–S–A–R–K–L–A–N–Q

Peptide 2: MW = 1096.2 F–G–G–F–T–G–A–R–K–S–A

Peptide 1: MW = 584.2 F–G–G–F–T–G

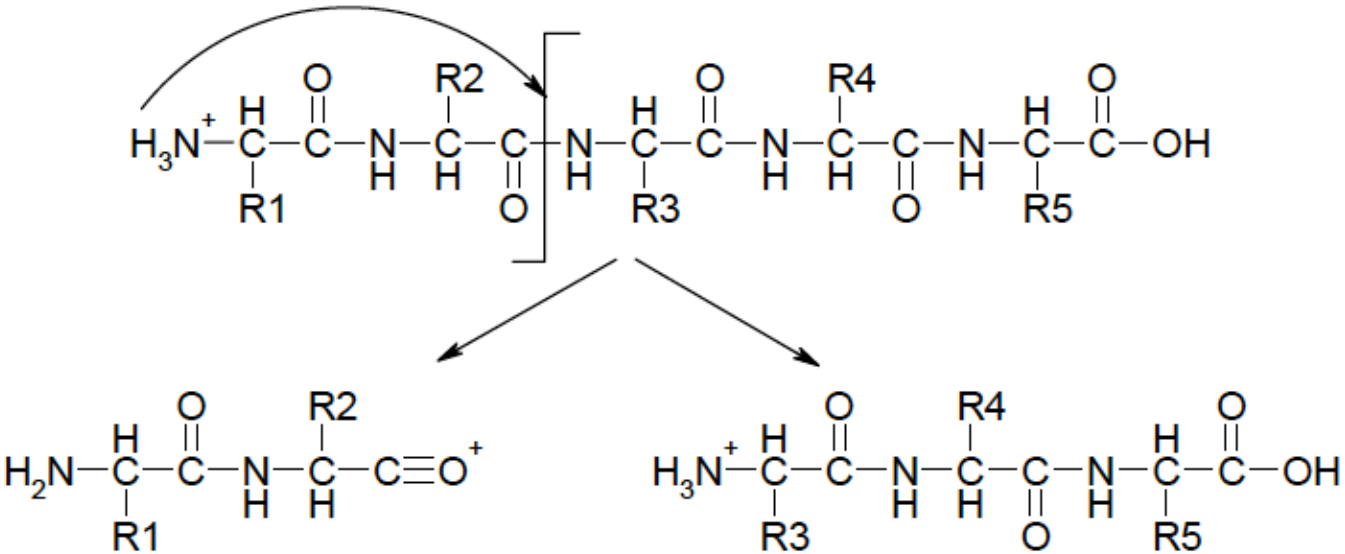


How we sequence peptides: MS/MS



MS/MS means using two mass analyzers (combined in one instrument) to select an analyte (ion) from a mixture, then generate fragments from it to give structural information.

Dominant fragment ions observed by collision-induced dissociation (CID) of peptides

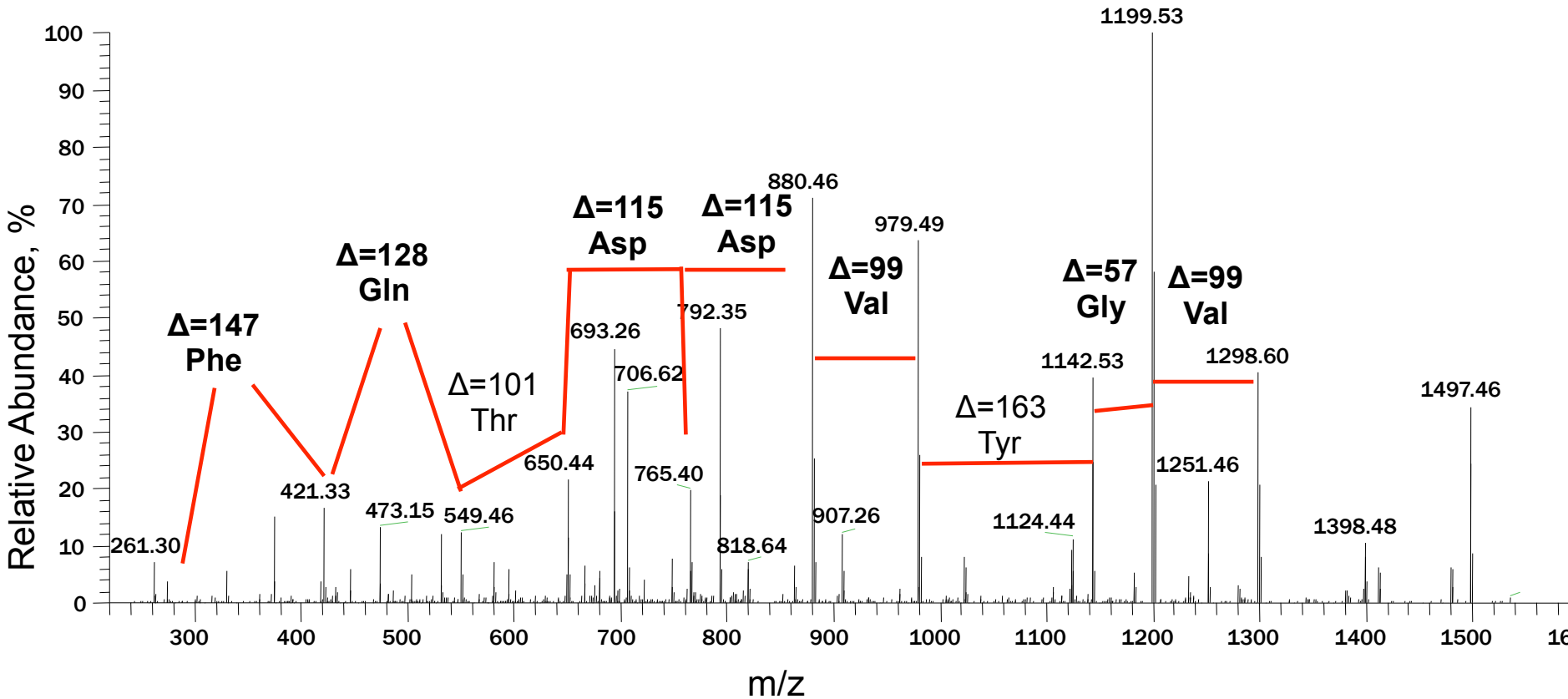


Direct cleavage of peptide bond

Rearrangement of mobile proton

Example electrospray MS/MS spectrum of a peptide

F-I-I-V-G-Y-V-D-D-T-Q-F-V-R

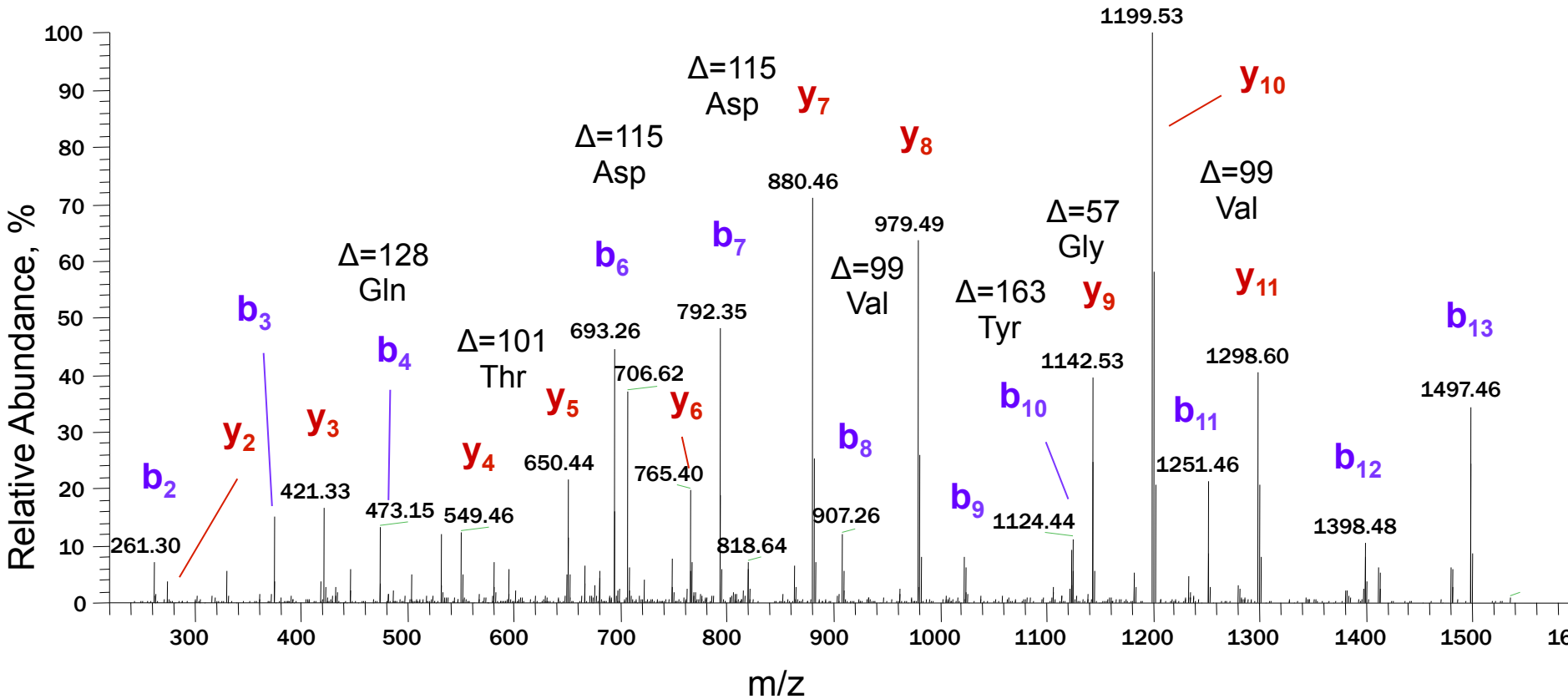


Example electrospray MS/MS spectrum of a peptide

y ions: 11 10 9 8 7 6 5 4 3 2

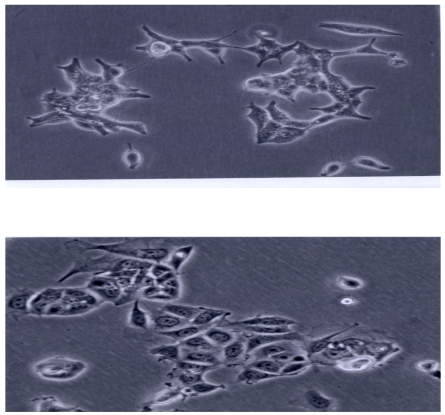
F-I-I-V-G-Y-V-D-D-T-Q-F-V-R

b ions: 2 3 4 6 7 8 9 10 11 12 13



Discovery Proteomics: differential expression profiling by MS

Biological Samples (case vs. control)



Protein Mixtures
- Biofluids
- Tissue lysates

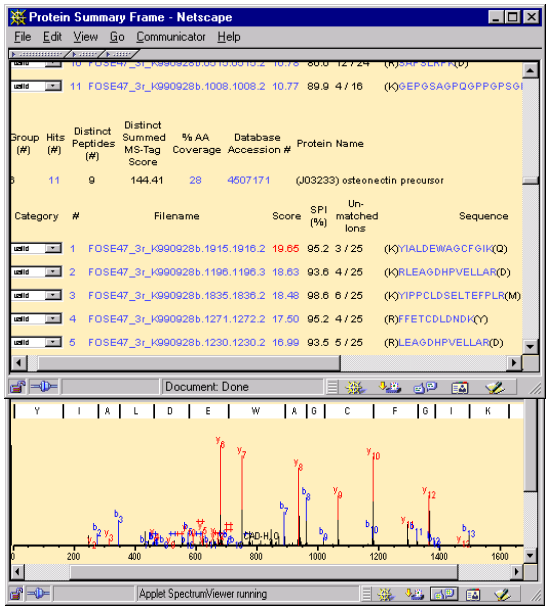
- digest to peptides
- fractionate peptides

LC-MS/MS

Separate and Analyze Peptides by LC-MS/MS

- m/z and intensity of peptides
  - rich ***pattern***
- Fragment ions for sequence

Data Analysis



Search DB using peptide m/z and sequence

- Peptide **identity**
- Protein **identity**
- Relative abundance

Most analyses of proteins are done by digestion of proteins to peptides ("bottom-up" proteomics)

Advantages:

- Data acquisition easily automated
- Fragmentation of tryptic peptides well understood
- Reliable software available for analysis
- Separation of peptides to create less complex subsets of the proteome for MS analysis is far easier than for proteins (relates to breadth and depth of coverage)

Disadvantages:

- Simple relationship between peptide and protein lost
- Took highly complex mixture and made it 20-100x more complex
    - Puts high analytical demands on instrumentation

Obtaining sequence information on intact proteins: “top-down” proteomics

- Most useful for single proteins or relatively simple mixtures (1)
- Can distinguish sequence variants
- Enables deciphering of combinatorial modification “codes” on proteins like histones (2).

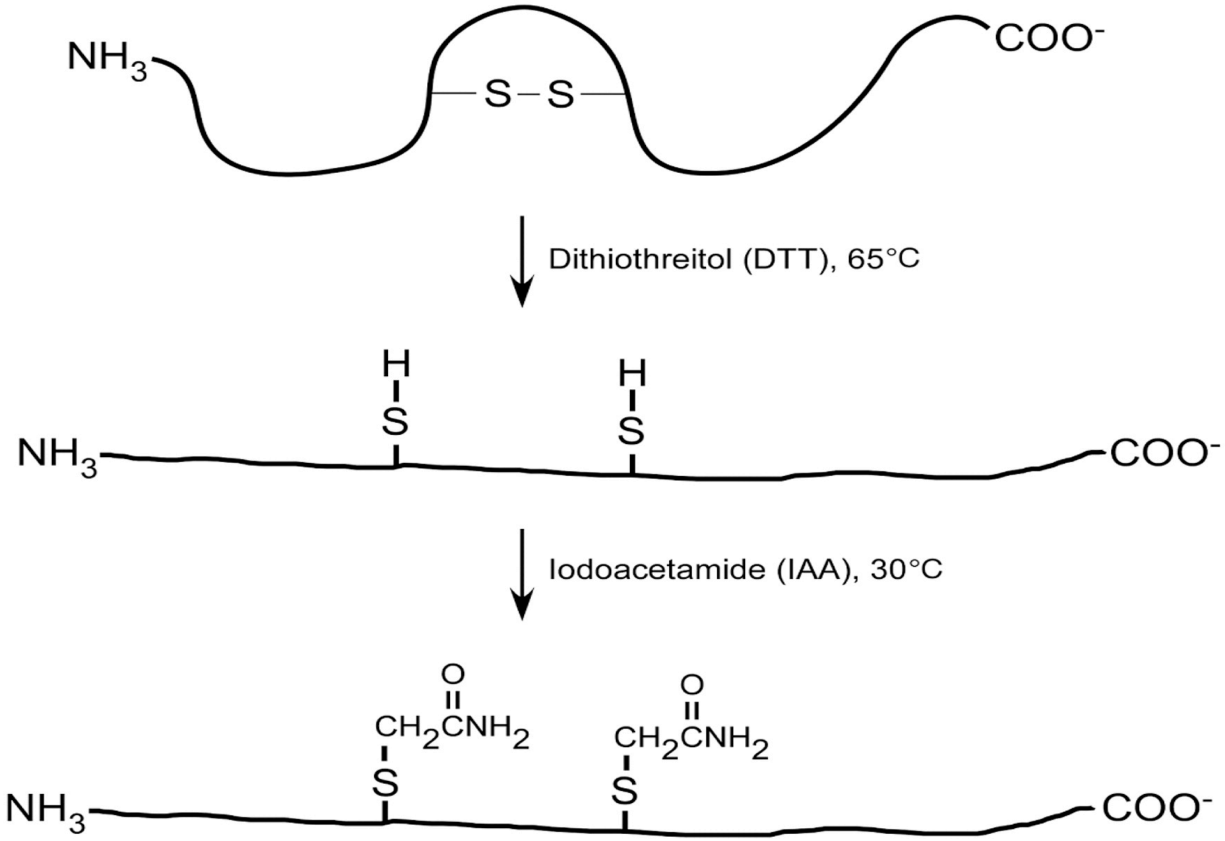
- While useful, its not suitable for most biomedical applications yet:
  - requires highly specialized instrumentation
  - cannot be easily applied to complex biological samples
  - Data interpretation is far more difficult and less automated
  - Breadth and depth of coverage of the proteome is orders of magnitude less than for bottom-up proteomics

1. Tran et al.Nature, 2011, 480(7376) p. 254-8
2. Tian et al. Genome Biology 2012, 13:R86

A selective look into the proteomic "tool chest"

Reduction and Alkylation

- Routinely done prior to enzymatic digestion to break disulfide bonds, unfolding proteins to make them more susceptible to enzymatic cleavage



A selective look into the proteomic "tool chest"

Highly Specific Proteases

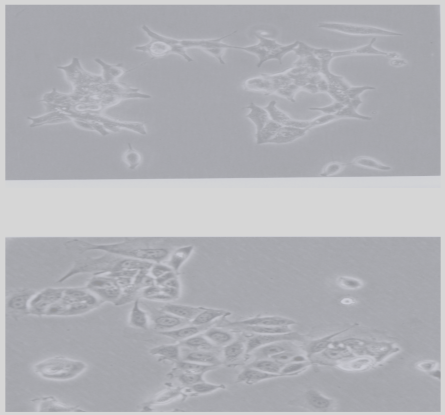
• Trypsin	C-terminal to Arg and Lys
• Lys-C	C-terminal to Lys
• Staph. V8	C-terminal to Glu and Asp
• Asp-N	N-terminal to Asp

Non-Specific Proteases

• Chymotrypsin	C-terminal to aromatic, aliphatic (e.g., Tyr, Trp, Phe, Leu)
• Proteinase K, Thermolysis	C-terminal to aromatic, aliphatic

Discovery Proteomics: differential expression profiling by MS

Biological Samples (case vs. control)



Protein Mixtures
- Biofluids
- Tissue lysates

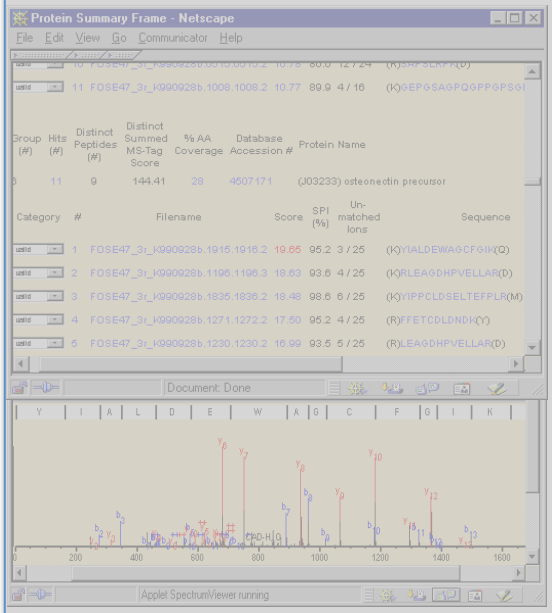
- digest to peptides
- fractionate peptides

LC-MS/MS

Separate and Analyze Peptides by LC-MS/MS

- m/z and intensity of peptides
  - rich ***pattern***
- Fragment ions for sequence

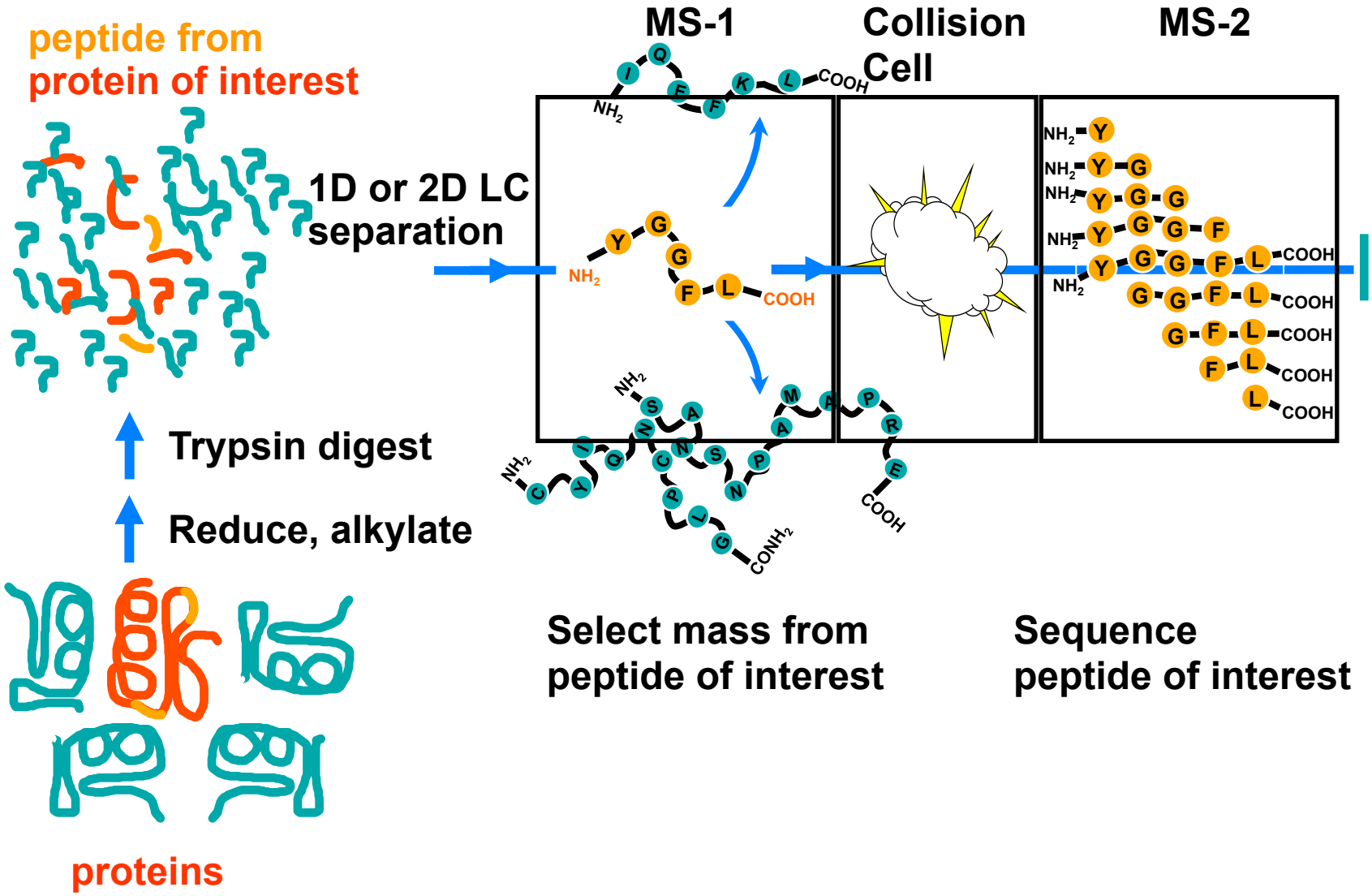
Data Analysis



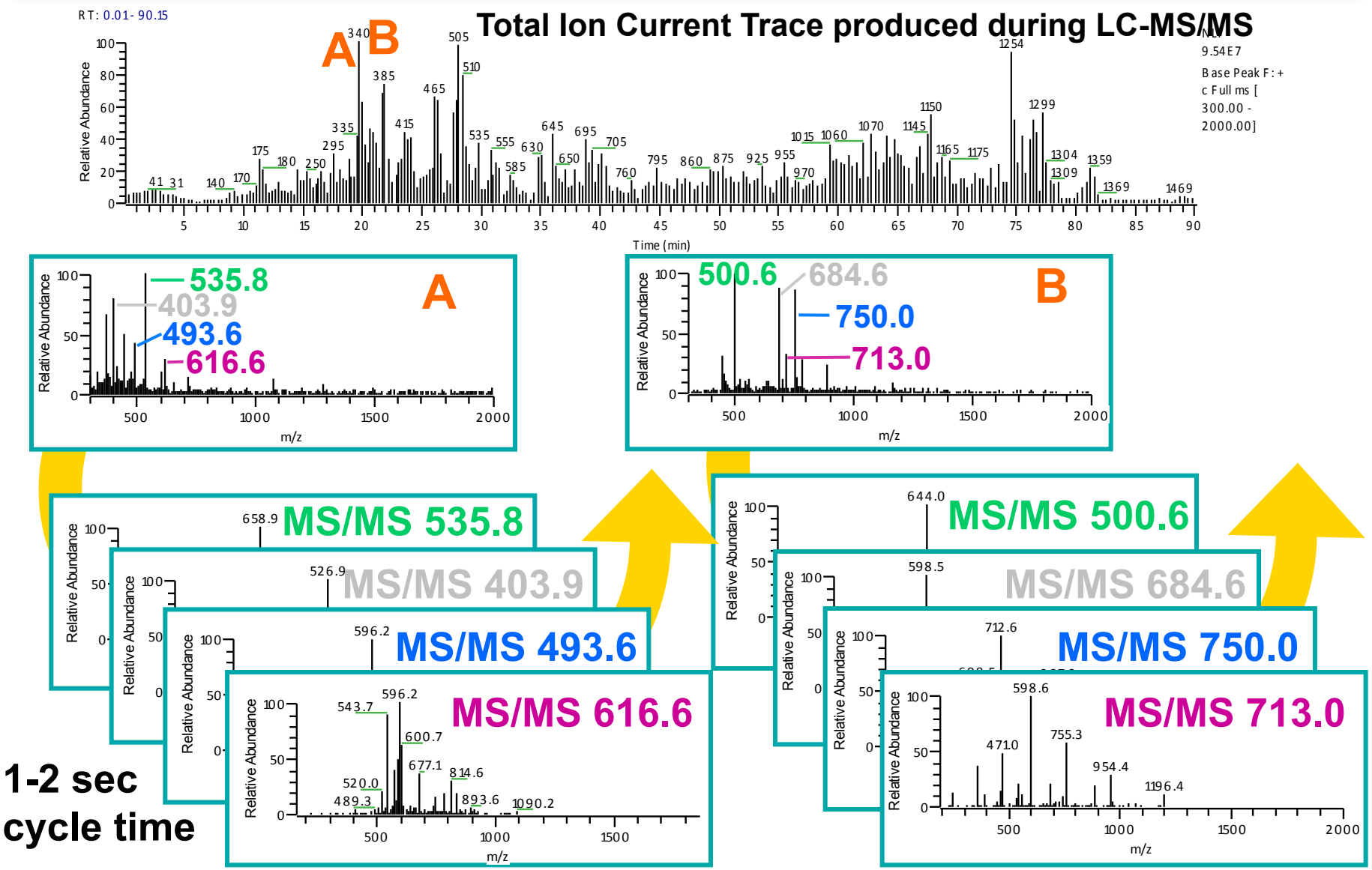
Search DB using peptide m/z and sequence

- Peptide **identity**
- Protein **identity**
- Relative abundance

Peptide Sequencing by LC/MS/MS



Automated Peptide Sequencing by LC/MS/MS



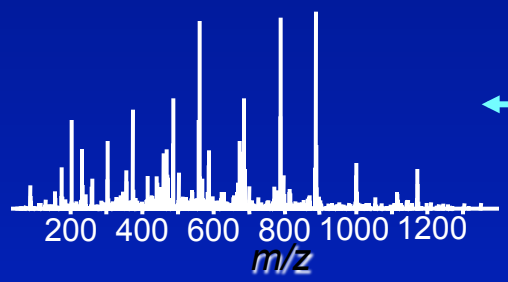
1-2 sec cycle time

"Top 4 Method" (modern MS systems can do up to "top 20")

MS/MS Search Engines: looking up the answer in the back of the book

Acquired MS/MS spectrum

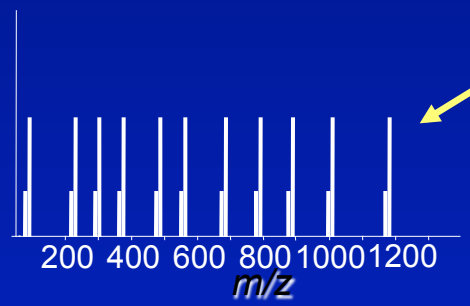
Sequence Database
(translation of transcriptome)



correlate

similarity score

Theoretical spectrum



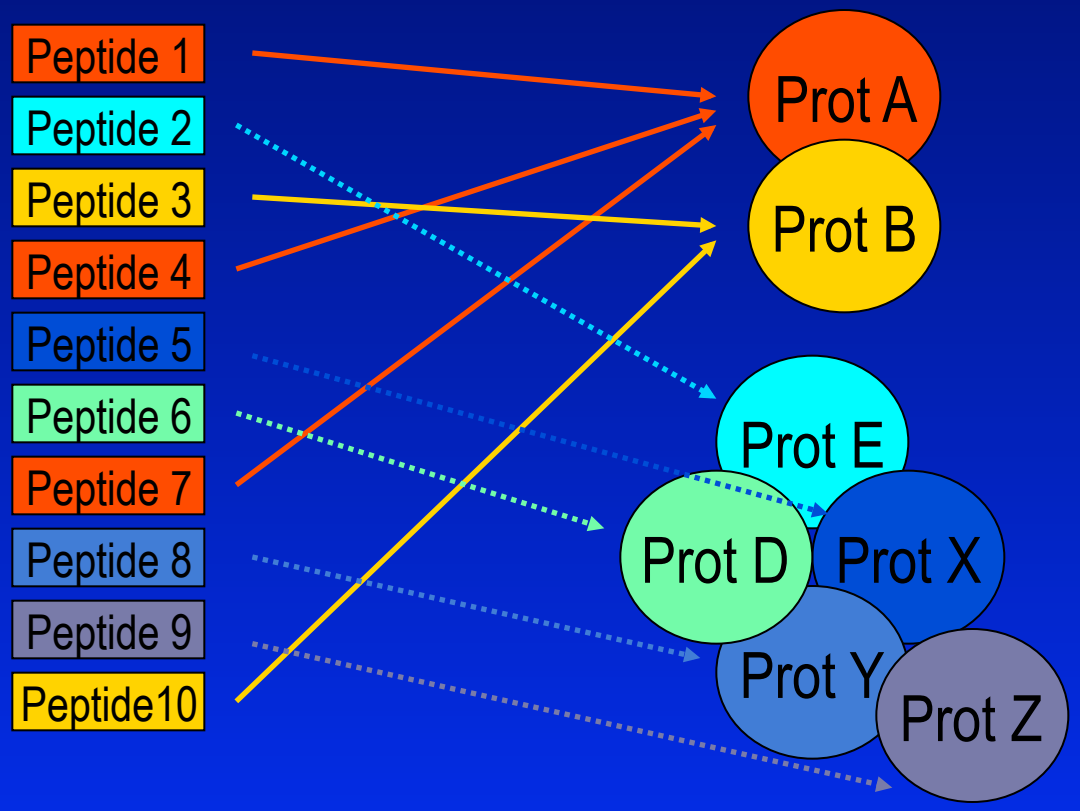
ISLLDAQSAPLR
VVEELCPTPEGK
DLLLQWCWENGK
ECDVVSNTIIAEK
GDAVFVIDALNR
VPTPNVSVVDLTNR
SYLFCMENSAEK
PEQSDLRSWTAK

Best matching database peptide

Determine peptide FDR by searching reversed DB

Algorithms: Mascot, MaxQuant, SpectrumMill, X-Tandem…

Rolling peptides up to the protein level



Slide courtesy of Alexey Nesvizhskii

Examples of a Protein Centric Table (MaxQuant)

Protein IDs	Gene Names	Unique Peptides Rep01	Sequence Coverage [%]	Mol. Weight [kDa]	Ratio H/L	Ratio H/L Count	p-value
A0ELI5	Edc3	3	10.2	55.9	1.2	4	0.96
A0MNP4	mCG_9684	1	3	33.7	1.0	1	0.01
A1A549	Tcf3	3	5.2	64.0	1.0	5	0
A1L013	2510012J08Rik	5	7.8	90.6	0.78	8	0
A1L329	mCG_20206	9	10.9	109.9	0.86	16	0
A1L3B6	mCG_19432	8	37.4	28.9	0.73	19	0

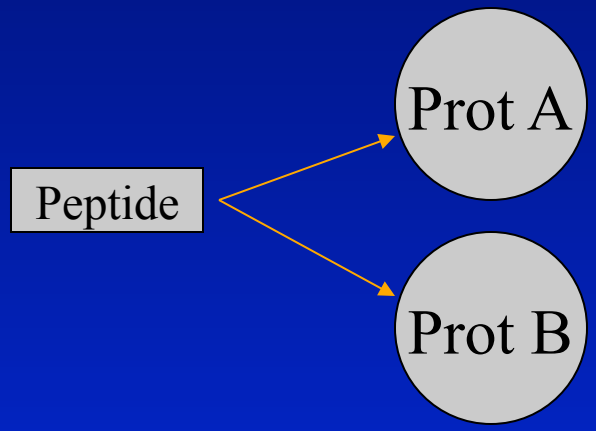
Protein X

Peptide	Ratio H/L
APEPTIDEK	1.0
YKPSTELLIR	1.2
EWERTHEFAASLR	1.6
IAMAPEPTIDER	0.9
GWQIMNCSTYK	0.5
YHTLSSVTYEHLK	1.5
ISEEALARGEPEPTIDEK	1.2
Median	1.2

- Table of values organized around proteins
- A ratio that indicates a fold-change vs. a control condition
- We generate a false discovery rate or p-value statistic for each protein ratio to indicate how different from the null hypothesis (unchanged)
- ***A prioritized list of candidates for follow-up studies***

Protein Inference Problem

Shared peptides: map to more than a single entry in protein database



protein A or protein B ??

Or both?

In bottom-up proteomics the connectivity between peptides and proteins is lost

Shared peptides are more prevalent with databases of higher eukaryotes due to the presence of:

- related protein family members
- alternative splice forms
- partial sequences

Slide courtesy of Alexey Nesvizhskii

Peptide Quant to Protein Quant

Peptide	Log_2 SILAC Ratio
APEPTIDEK	0.12
YKPSTELLIR	0.15
EWERTHEFAASLR	0.07
IAMAPEPTIDER	0.21
GWQIMNCSTYK	0.14
YHTLSSVTYEHLK	0.29
ISEEALARGEPEPTIDEK	0.23
EITHERWAYK	***0.22***
SIMPLESEQK	0.77
LITTLEPEPTIDER	0.99

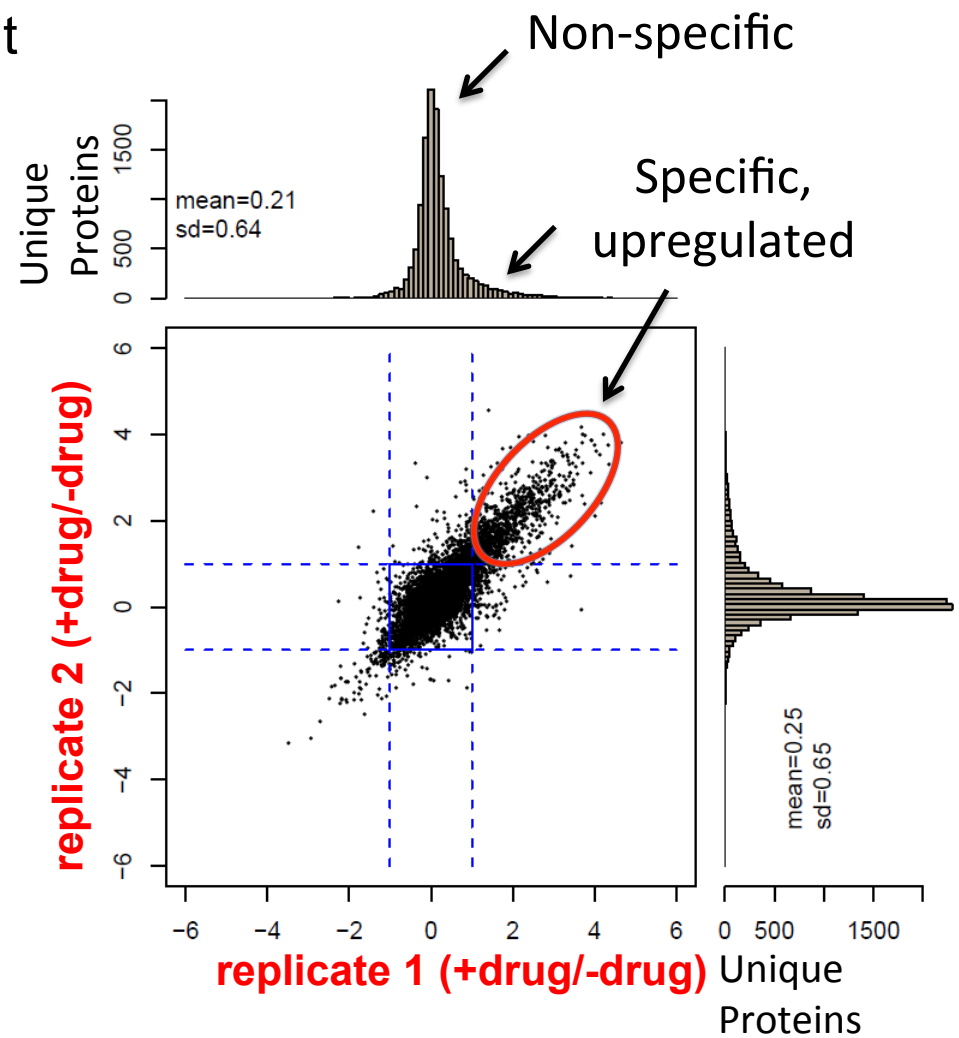
Protein X

Protein Y

- A peptide could belong to more than one protein
- Go with preponderance of the evidence to assign peptide
  - Occam's *razor* principle
- In this case, peptide is assigned to ***Protein X*** because there are more peptides supporting it

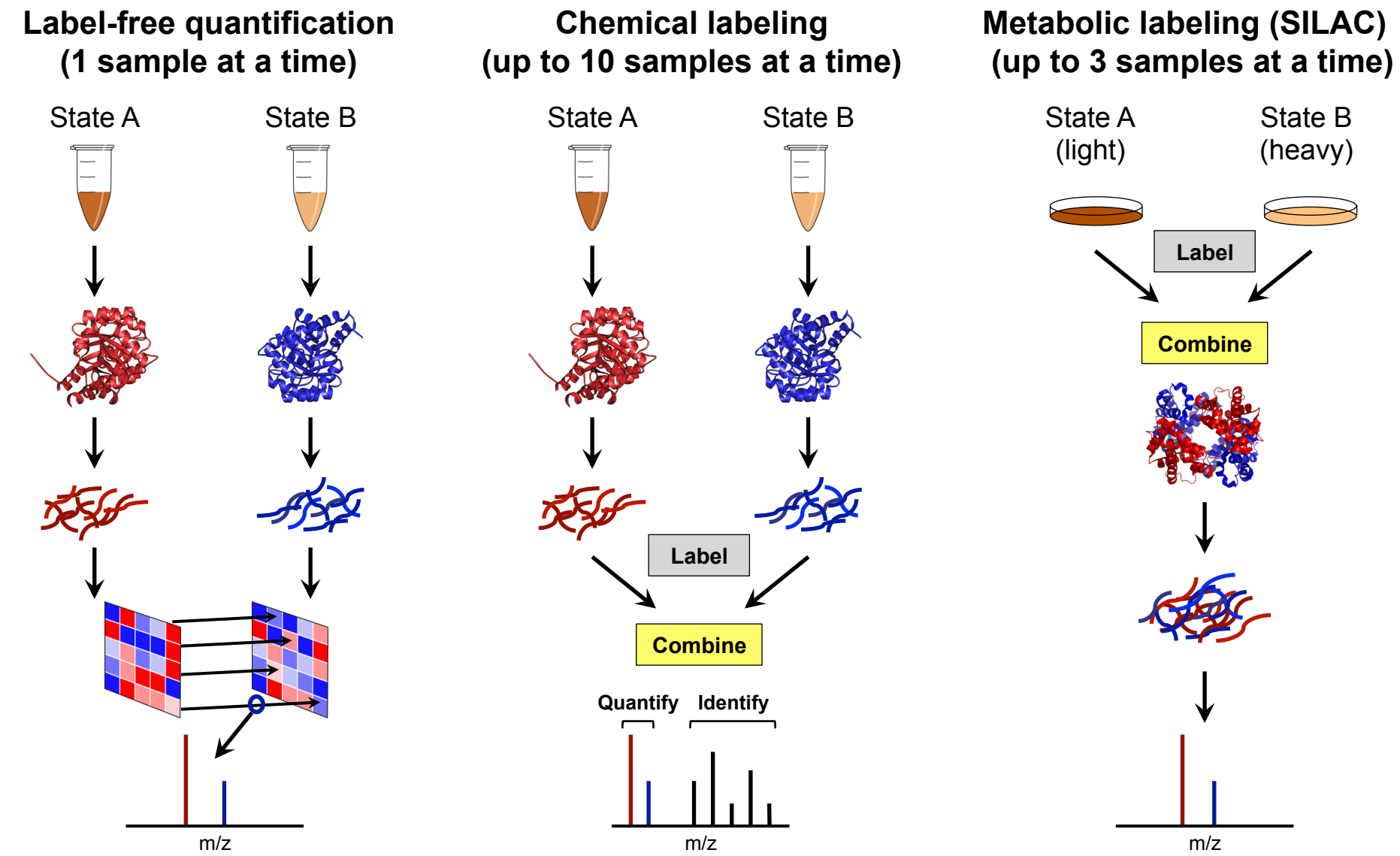
Quantitative Data Drives Modern Proteomics

- Analyze biological replicates of state comparisons
- The end result is always a ratio:
  - WT expression vs. mutant
  - Drug vs. no drug
  - Bait vs. control



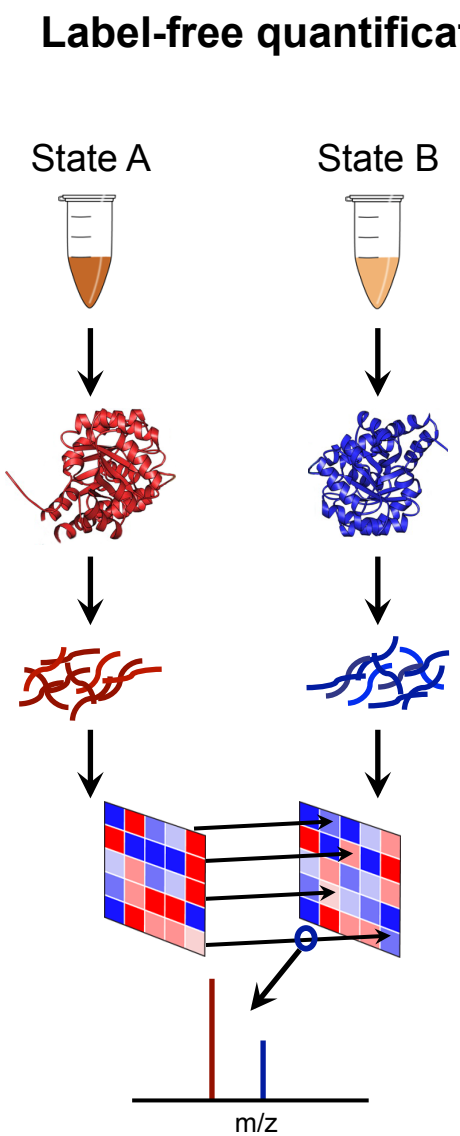
- T-statistics or Gaussian Modeling drives calling of regulation or specificity

Relative Quantification Methods for Discovery Proteomics



Increasing precision

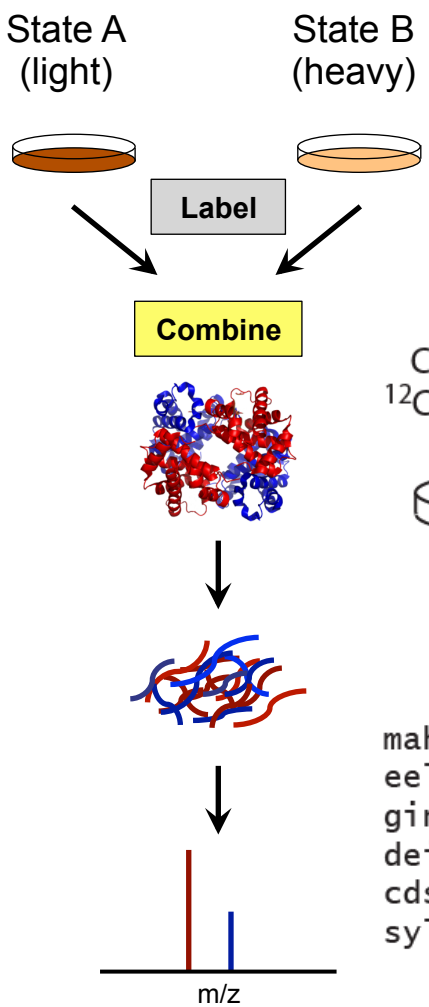
Label-free quantification: spectral counting or peak area



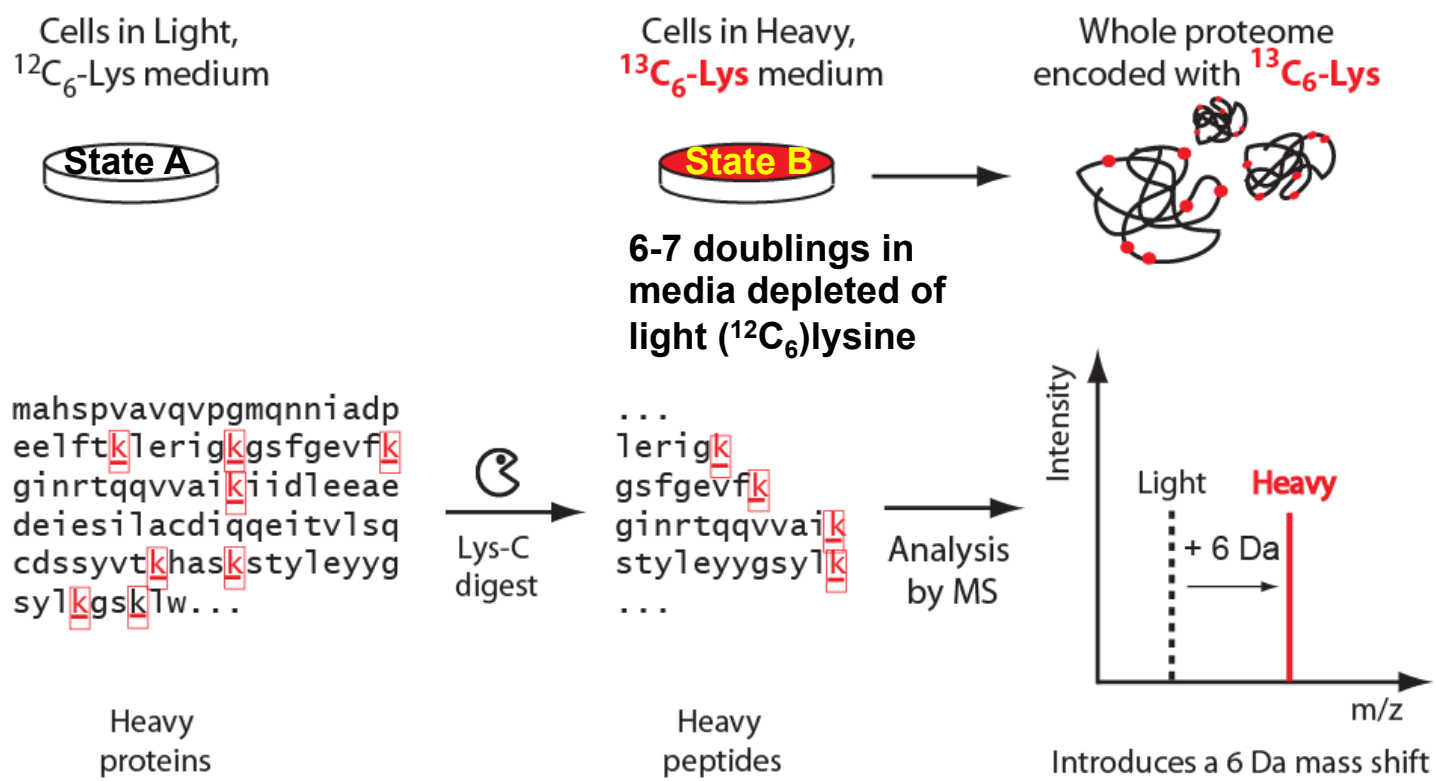
- One spectrum with peptide ID that can be linked to a protein = 1 count for that protein
  - Basis of spectral counting
- Detection likelihood is tied to abundance
  - Results vary depending on Instrument settings and number of peptides in protein
- Only reliable for moderate to highly abundant proteins
  - Lots of missing data, especially for lower abundance proteins
  - Poor precision leads to high FDR
- Low throughput
  - Every sample run separately
  - Triplicate analyses required for stat. confidence
  - Instrument time = $; not inexpensive!

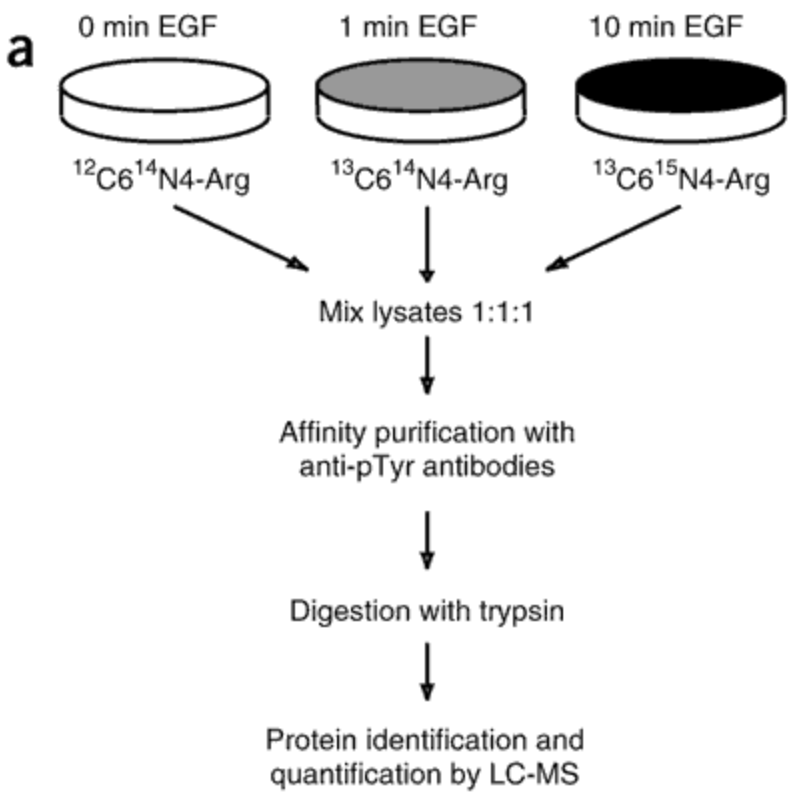
SILAC: Stable Isotope Labeling by Amino acids in Cell culture

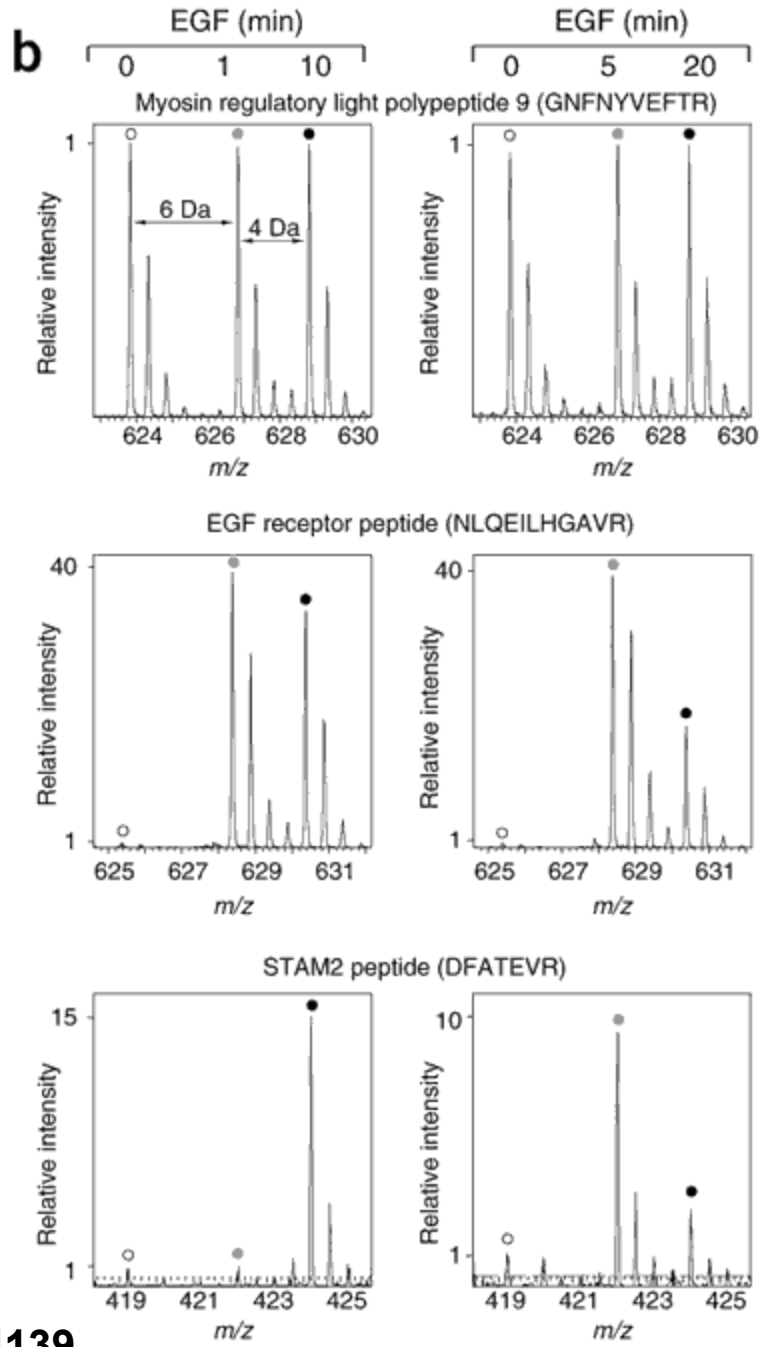
Metabolic labeling (SILAC)
(up to 3 samples at a time)



Pros	Cons
Deep, highly precise quant.	Limited plex level (3 max)
Works well in most cell lines	Not practical for most model systems
Works with all PTMs	Can't label humans
Relatively inexpensive	



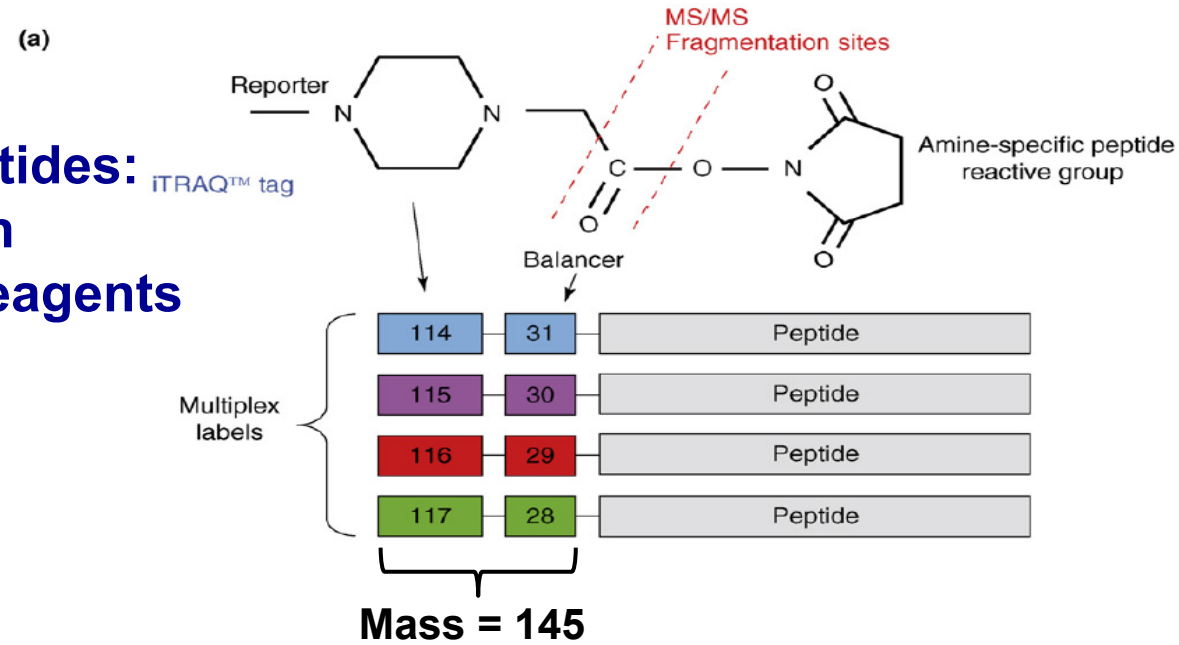




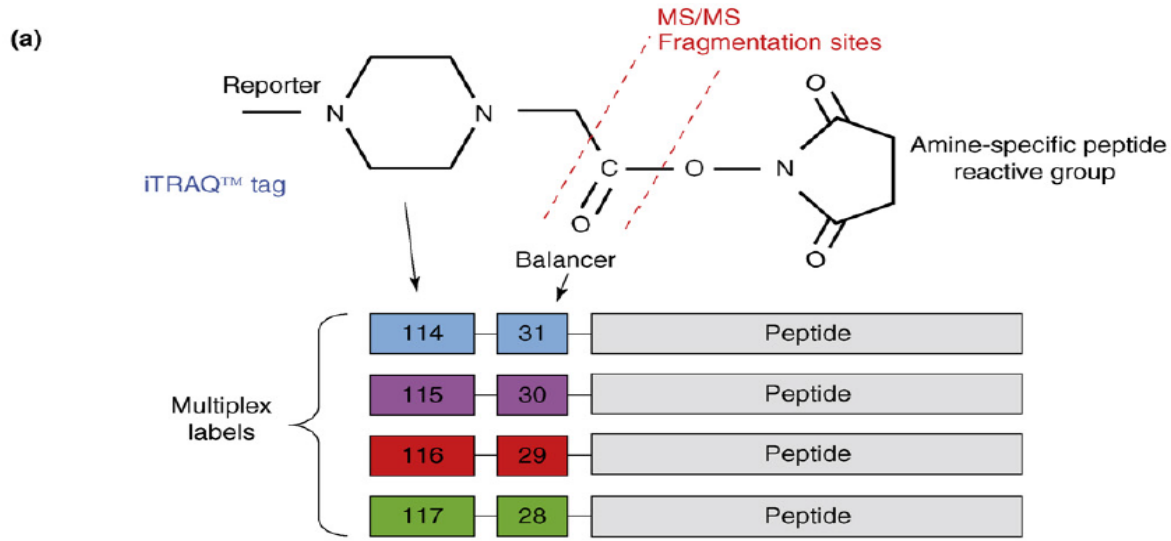
- **Time course of activation**
  - **Mixing samples improves data and saves instrument time**
- **ID of p-sites requires MS/MS**
- **Detects some proteins associated with pY-proteins**

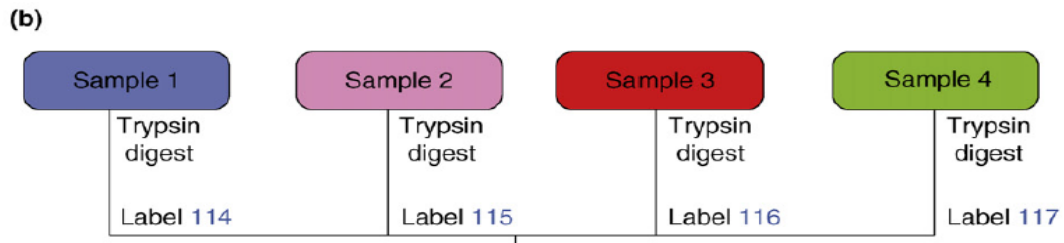
Blagoev, Ong et al. (2004) Nature Biotech. 22: 1139

Chemical Labeling of Peptides: Multiplexed Quantification with Isobaric Mass Tag Reagents



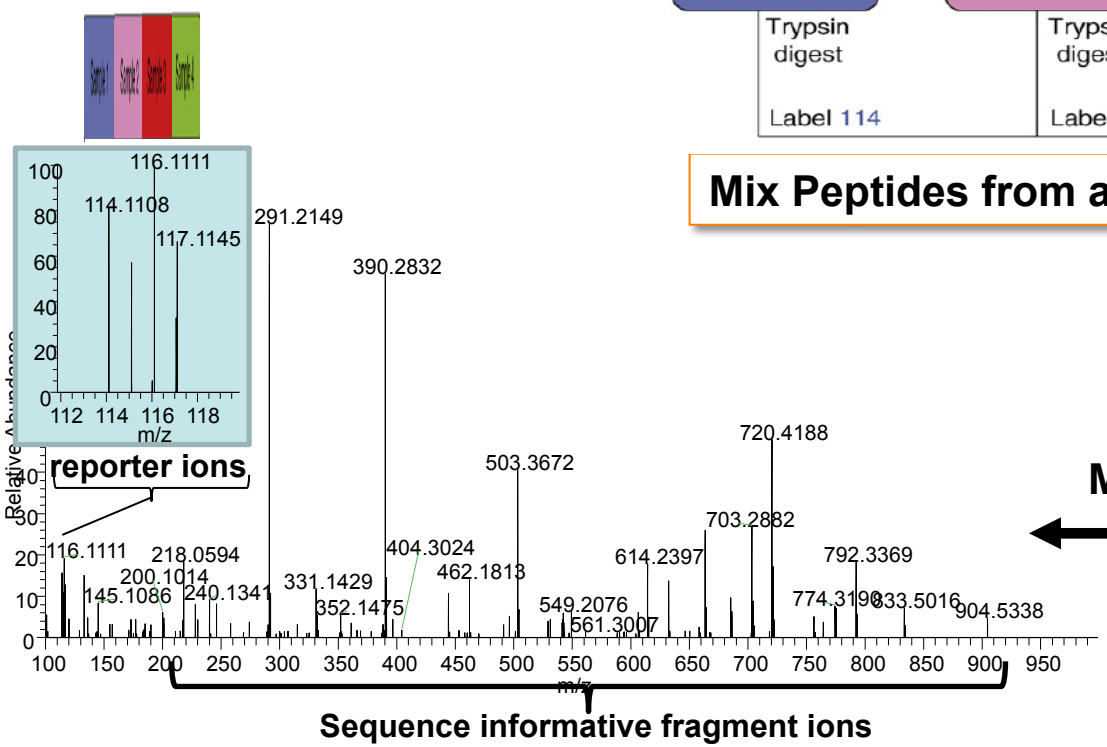
Chemical Labeling of Pep
Multiplexed Quantificatio
with Isobaric Mass Tag R



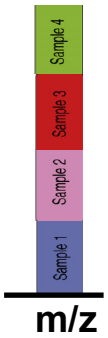


Mix Peptides from all 4 Samples: analyze by MS

MS

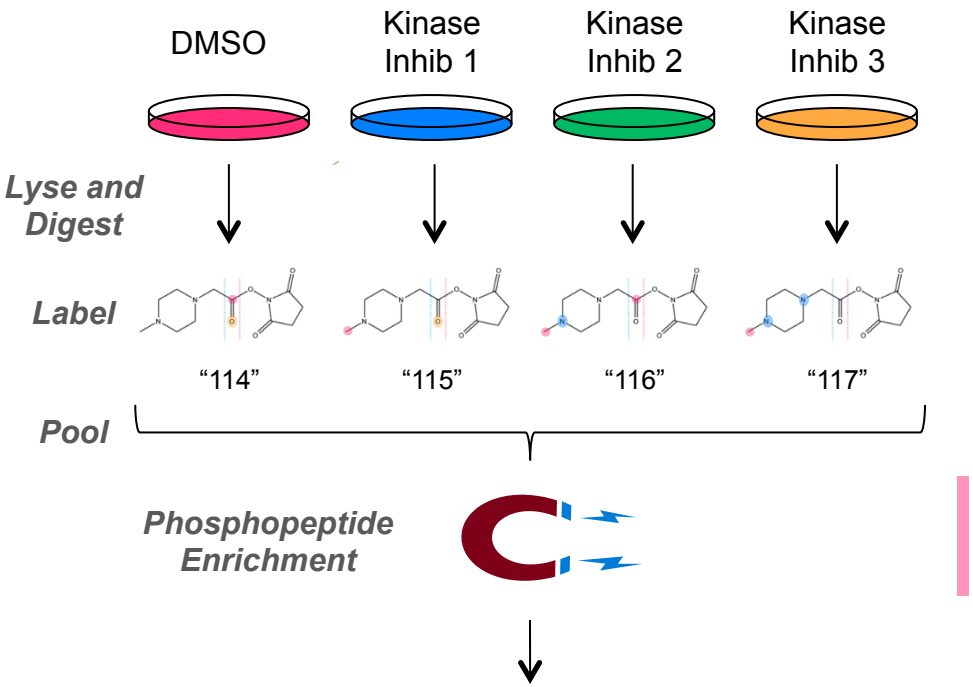


MS/MS

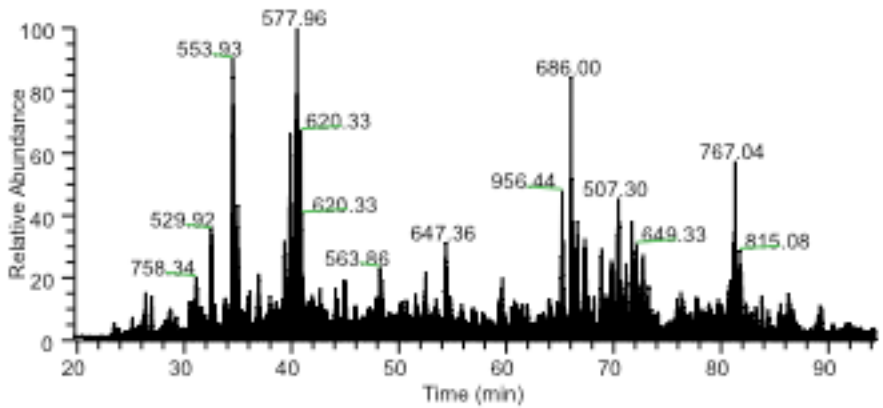


same peptide from 4 different samples: Observed precursor intensity = Σ of all labeled versions

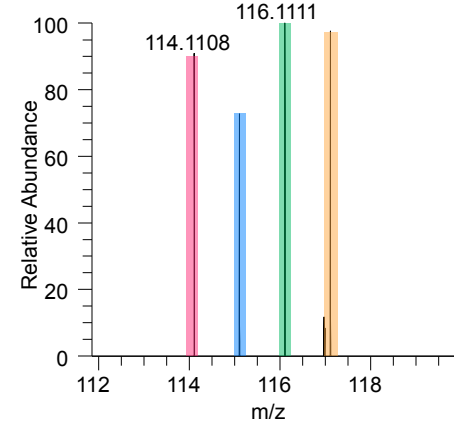
iTRAQ Experimental Example



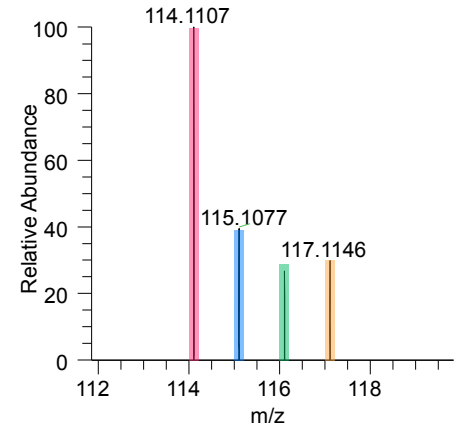
LCMS



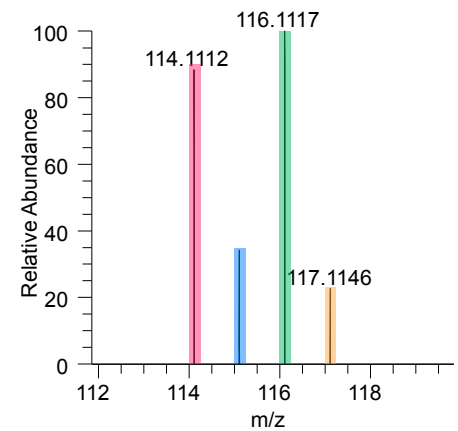
Peptide #1: No effect



Peptide #2: Sensitive to all inhibitors



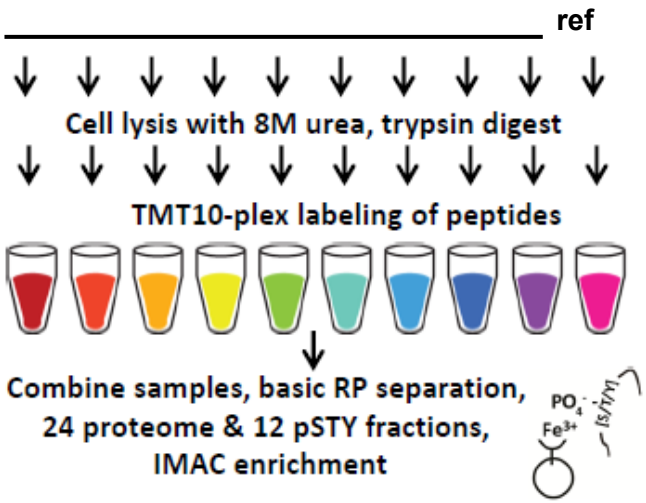
Peptide #3: Sensitive to inhibitors 1 & 3



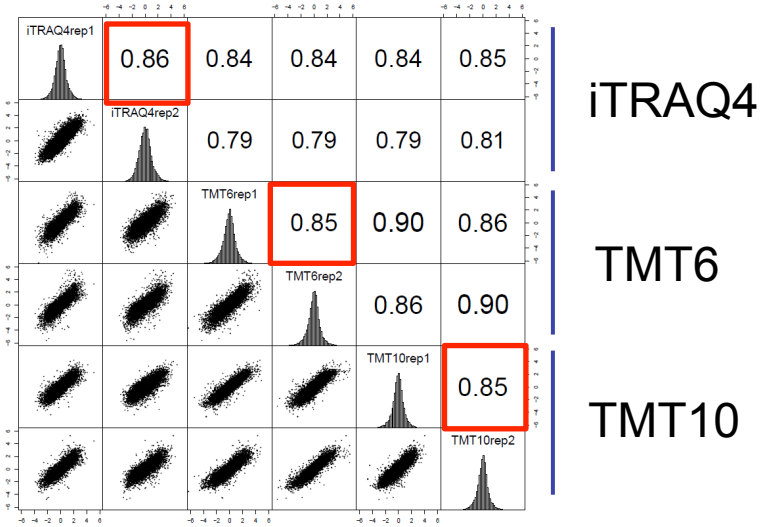
Isobaric tag reagents with higher multiplex levels now available: increased sample throughput with high sensitivity and good quantitative fidelity

3x increased throughput

9 tumor samples (4 basal; 4 luminal; 1 reference)



Highly consistent quantification results

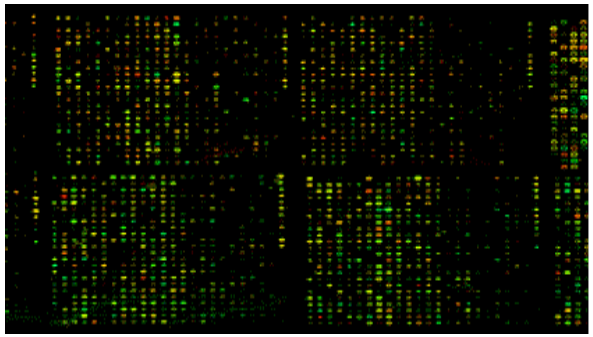


Log2 basal/luminal tumors

		Rep1/Rep2	
		Proteins/sites	Distinct Peptides
Proteome Coverage	iTRAQ4	13,201/13,101	198953/196484
	TMT6	12,839/13,839	174590/196521
	TMT10	12,624/12,908	170190/168828
Phosphoproteome Coverage	iTRAQ4	45,495/45,815	60,945/58,005
	TMT6	33,131/32,261	39,090/42,543
	TMT10	33,523/31,119	39,044/34,958

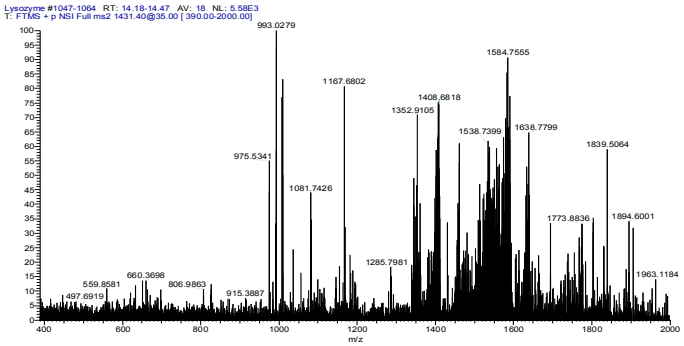
Analytical challenges of proteomics differ in important ways from transcriptional analysis

Transcriptional Profiling



- All possible features known
- Sample is static during analysis
- All features measured
- Robust means to amplify low numbers DNA or RNA (PCR)
- Signal not detected means feature not present

MS-based Proteomics



- All possible features <u>not</u> known
- Sample is dynamic during analysis
- 20-50% of features measured
- No protein PCR (analytics have to deal with enormous dynamic range)
- Signal not detected means <u>either</u> that feature not present <u>or</u> feature present but not detected

Discovery defines a reduced set of "sentinel" marks that need to be repeatedly measured in a range perturbations

Discovery:

- Disease
- Development
- Drug
- KO/KI

Currently: Westerns, immunoassays

Future: Targeted MS, ImmunoMS

Assays:

- Highly specific
- Sensitive
- Highly precise
- Multiplexed
- Interference-free

Not all proteins and (especially) PTMs observed in all experiments

Precisely measure selected analytes in all experiments – no missing data

How do you start a proteomics project?

We meet to discuss your project (scarr@broadinstitute.org)

- Project proposals are reviewed for scientific merit, technical feasibility and alignment with our interests and the Broad mission
  - Discussion of the science and experimental design
  - Sample preparation discussed in detail - what, how and by whom
  - All projects are collaborative

Funding:

- Platforms are largely self-supporting and must charge the work performed. If projects are reviewed favorably but lack funding, we will help investigators explore options for support, including consideration for collaborative funding through the Broad.

Tutorials

These helpful guides are meant to educate potential collaborators about some of the technologies and methodologies utilized by the Proteomics Platform.

Proteomic Mass Spectrometry: An overview of our core technology and how we use it to identify proteins.

SILAC (Stable Isotope Labeling of Amino Acids in Culture): A quantitative technique based on metabolic labeling of cellular proteins prior to sample preparation.

iTRAQ (Isobaric Tags for Relative and Absolute Quantification): A quantitative technique based on chemical labeling of proteins after sample preparation.

Phosphoproteomics: Specific methodologies to focus on signalling and phosphorylation events.

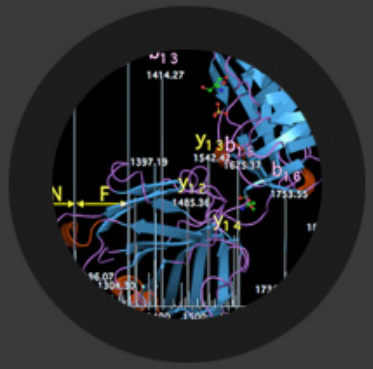
Target Identification: Techniques to probe interactions of small molecules (like drugs) with the proteins that bind to them.

MRM (Multiple Reaction Monitoring): A focused quantitative method that uses synthetic peptides as standards for quantification of specific proteins in pre-clinical samples.

SISCAPA: An antibody-based technique used to enrich specific protein or peptide targets prior to MS-based quantification.

Proteome Fractionation: The proteome is extremely complex. Sometimes it's best to divide and conquer!

www.broadinstitute.org/proteomics



The Broad Institute Proteomics Group

Suggested additional reading

Carr and Annan, 2001. Overview of Peptide and Protein Analysis by Mass Spectrometry. *Current Protocols in Molecular Biology* 10: 10.21.1–10.21.27.

Aebersold, R. and Mann, M. 2003. Mass spectrometry-based proteomics. *Nature* 422:198-207.

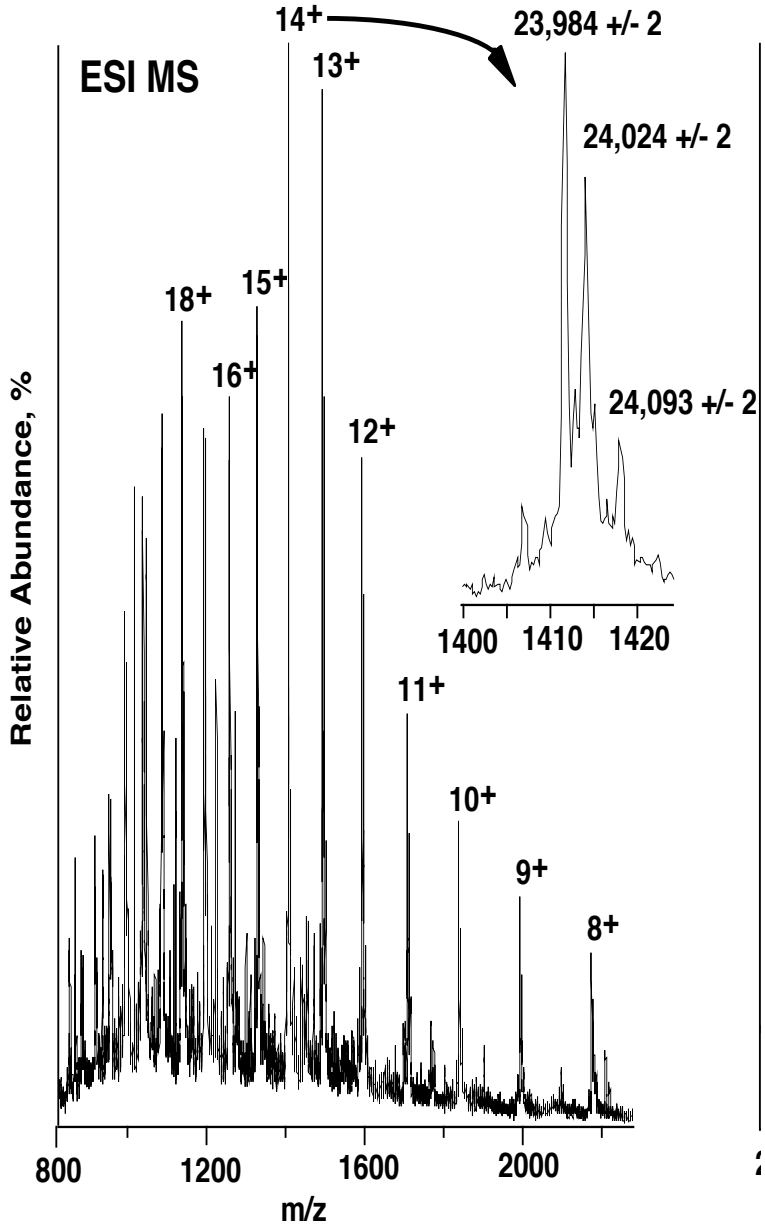
Cravatt et al. 2007. The biological impact of mass-spectrometry-based proteomics. *Nature* 450: 991-1000.

Additional resources for MS data interpretation

- Manual de novo tutorials
  - Don Hunt and Jeff Shabanowitz
    - http://www.ionsource.com/tutorial/DeNovo/DeNovoTOC.htm
  - Rich Johnson
    - http://www.abrf.org/ResearchGroups/MassSpectrometry/EPosters/ms97quiz/SequencingTutorial.html
- Automated de novo - PEAKS
  - http://www.bioinformaticssolutions.com/peaks/tutorials/denovo.html
  - http://www.youtube.com/watch?v=lyhpRu6s7Ro
- *De Novo* Sequencing and Homology Searching Tutorial
  - Ma B, Johnson R. *Mol Cell Proteomics 11: O111.014902, 1–16,* ***2012****..*
- Modification Site Localization Scoring: Strategies and Performance - Review
  - Chalkley, RJ and Clauser, KR. *Mol Cell Proteomics* 11, 3-14, **2012** .
- Target/Decoy FDR - Tutorial
  - Elias & Gygi, *Nature Methods*, 4, 207-214, **2007**.
- Protein Inference - Tutorial
  - Nesvizhskii, *Mol Cell Proteomics*, 4, 1419-1440, **2005**.

BACKUPS

Example of electrospray mass spectrum of intact protein (beta-Casein)



If a positive ion series is assumed to represent different protonation states, then the mass/charge ratios, x_1 and x_2, of adjacent members of the ion series are given by

$$x_1 = (M + n)/n$$

and

$$x_2 = (M + n + 1)/(n + 1)$$

where M is the molecular mass. Solving these equations gives

$$n = (x_2 - 1)/(x_1 - x_2)$$

and allows the estimation of M. In practice, such conversion of m/z data to a 'true' mass spectrum is carried out by the mass spectrometer data system; the redundancy of data allows a concomitant estimate of the precision of determination of molecular mass.[47]

47. M. Mann, C. K. Meng and J. B. Fenn, Anal . Chem. **61**, 1702 (1989).