

# Mass spectrometry-based proteomics

BMI/CS 776

[www.biostat.wisc.edu/bmi776/](http://www.biostat.wisc.edu/bmi776/)

Spring 2016

Anthony Gitter

[gitter@biostat.wisc.edu](mailto:gitter@biostat.wisc.edu)

# Goals for lecture

## Key concepts

- Benefits of mass spectrometry
- Generating mass spectrometry data
- Computational tasks
- Matching spectra and peptides

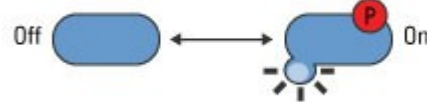
# Mass spectrometry uses

- Mass spectrometry is protein analog of microarrays or RNA-seq
  - Quantify abundance or state of all (many) proteins
  - No need to specify proteins to measure in advance
- Other applications in biology
  - Targeted proteomics
  - Metabolomics

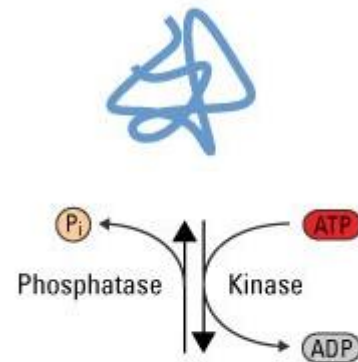
# Advantages of proteomics

- Proteins are functional units in a cell
  - Protein abundance directly relevant to activity
- Post-translational modifications
  - Change protein state

Phosphorylation  
in signaling



[Thermo Fisher Scientific](#)



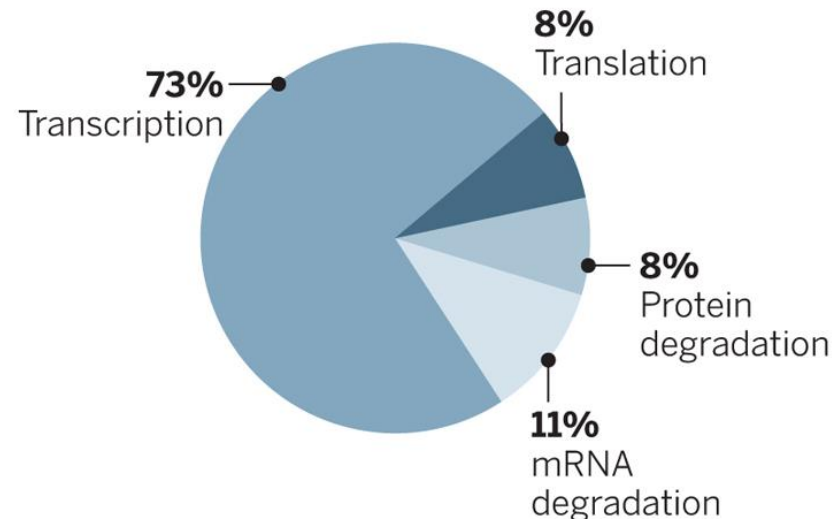
Histone  
modifications



# Estimating protein levels from gene expression

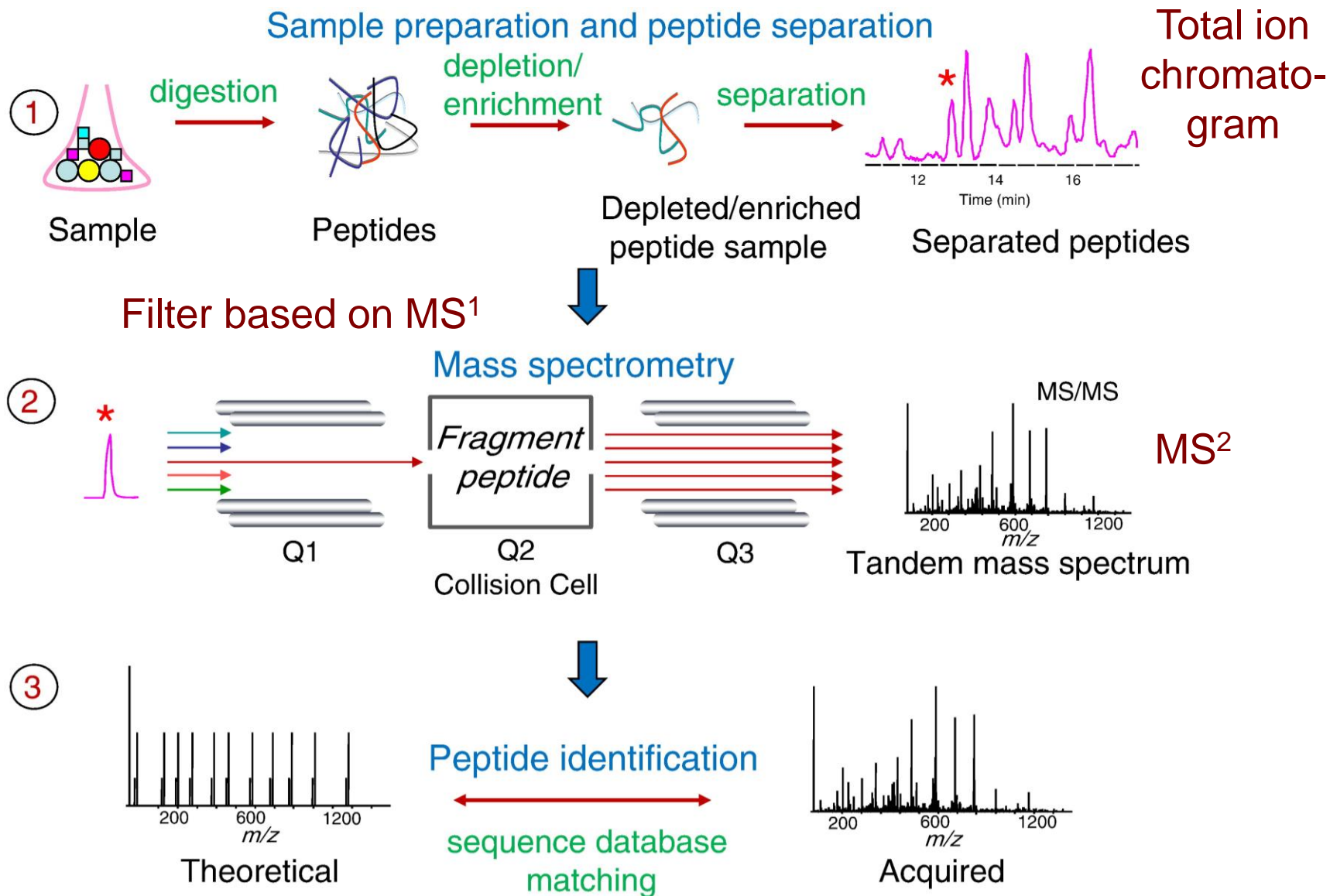
- Correlation between gene expression and protein abundance has been debated
- Gene expression tells us nothing about post-translational modifications

## Contribution to protein levels



Li and Biggin *Science* 2015

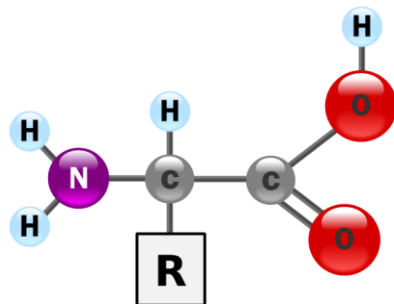
# Mass spectrometry workflow



# Amino Acids

- 20 amino acids
- Building blocks of proteins
- Known molecular weight
- Common template

Amino-terminal      Carboxy-terminal

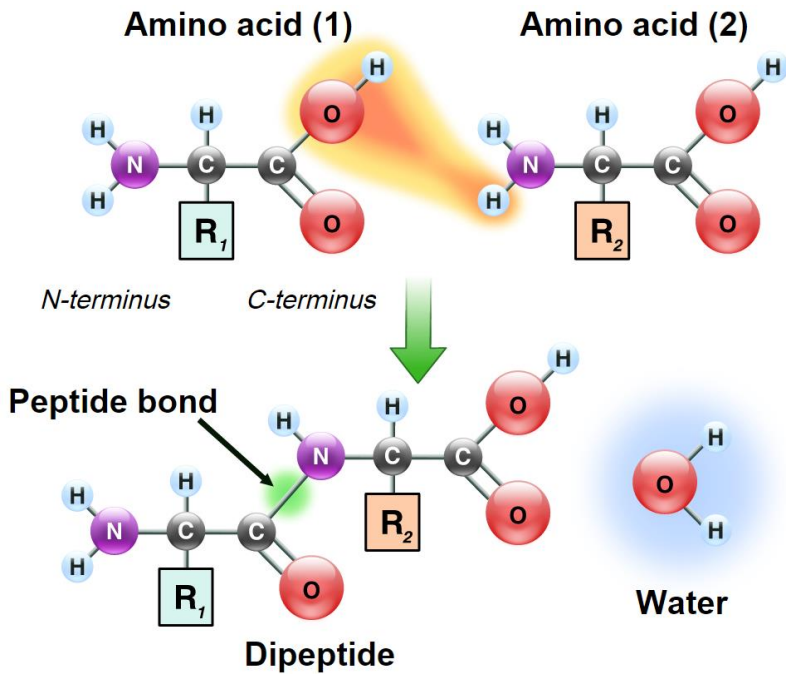


Wikipedia, Yassine Mrabet

NONPOLAR, HYDROPHOBIC		POLAR, UNCHARGED	
		R GROUPS	
Alanine Ala A MW = 89	$\begin{array}{c} ^- \text{OOC} \\   \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_3 \end{array}$		$\begin{array}{c} \text{H} - \text{CH} - \text{COO}^- \\   \\ \text{N}^+ \text{H}_3 \end{array}$ Glycine Gly G MW = 75
Valine Val V MW = 117	$\begin{array}{c} ^- \text{OOC} \\   \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}(\text{CH}_3)_2 \end{array}$		$\begin{array}{c} \text{HO} - \text{CH}_2 - \text{CH} - \text{COO}^- \\   \\ \text{N}^+ \text{H}_3 \end{array}$ Serine Ser S MW = 105
Leucine Leu L MW = 131	$\begin{array}{c} ^- \text{OOC} \\   \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{CH}(\text{CH}_3)_2 \end{array}$		$\begin{array}{c} \text{OH} \\   \\ \text{CH}_3 - \text{CH} - \text{CH} - \text{COO}^- \\   \\ \text{N}^+ \text{H}_3 \end{array}$ Threonine Thr T MW = 119
Isoleucine Ile I MW = 131	$\begin{array}{c} ^- \text{OOC} \\   \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}(\text{CH}_3) - \text{CH}_2 - \text{CH}_3 \end{array}$		$\begin{array}{c} \text{HS} - \text{CH}_2 - \text{CH} - \text{COO}^- \\   \\ \text{N}^+ \text{H}_3 \end{array}$ Cysteine Cys C MW = 121
Phenylalanine Phe F MW = 131	$\begin{array}{c} ^- \text{OOC} \\   \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{C}_6\text{H}_5 \end{array}$		$\begin{array}{c} \text{HO} - \text{C}_6\text{H}_4 - \text{CH}_2 - \text{CH} - \text{COO}^- \\   \\ \text{N}^+ \text{H}_3 \end{array}$ Tyrosine Tyr Y MW = 181
Tryptophan Trp W MW = 204	$\begin{array}{c} ^- \text{OOC} \\   \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{C}_8\text{H}_6\text{N}_2 \end{array}$		$\begin{array}{c} \text{NH}_2 \\   \\ \text{O}=\text{C} - \text{CH}_2 - \text{CH} - \text{COO}^- \\   \\ \text{N}^+ \text{H}_3 \end{array}$ Asparagine Asp N MW = 132
Methionine Met M MW = 149	$\begin{array}{c} ^- \text{OOC} \\   \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{CH}_2 - \text{S} - \text{CH}_3 \end{array}$		$\begin{array}{c} \text{NH}_2 \\   \\ \text{O}=\text{C} - \text{CH}_2 - \text{CH}_2 - \text{CH} - \text{COO}^- \\   \\ \text{N}^+ \text{H}_3 \end{array}$ Glutamine Gln Q MW = 146
Proline Pro P MW = 115	$\begin{array}{c} ^- \text{OOC} \\   \\ \text{CH} - \text{CH}_2 - \text{CH}_2 \\   \quad \quad   \\ \text{HN} \quad \quad \text{CH}_2 \end{array}$		<b>POLAR BASIC</b> $\begin{array}{c} ^+ \text{NH}_3 - \text{CH}_2 - (\text{CH}_2)_3 - \text{CH} - \text{COO}^- \\   \\ \text{N}^+ \text{H}_3 \end{array}$ Lysine Lys K MW = 146
Aspartic acid Asp D MW = 133	$\begin{array}{c} ^- \text{OOC} \\   \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{C}(=\text{O})\text{O}^- \end{array}$		$\begin{array}{c} \text{NH}_2 \\   \\ \text{N}^+ \text{H}_2 = \text{C} - \text{NH} - (\text{CH}_2)_3 - \text{CH} - \text{COO}^- \\   \\ \text{N}^+ \text{H}_3 \end{array}$ Arginine Arg R MW = 174
Glutamine acid Glu E MW = 147	$\begin{array}{c} ^- \text{OOC} \\   \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{CH}_2 - \text{C}(=\text{O})\text{O}^- \end{array}$		$\begin{array}{c} \text{HN} = \text{C} - \text{CH}_2 - \text{CH} - \text{COO}^- \\   \quad \quad   \\ \text{H} \quad \quad \text{N}^+ \text{H}_3 \end{array}$ Histidine His H MW = 155

# Peptide fragmentation

## Peptide bond

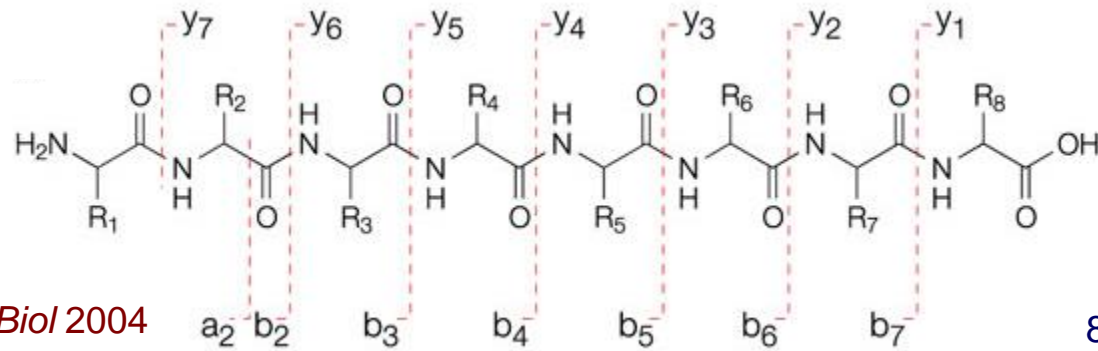


Wikipedia, Yassine Mrabet

- Select similar peptides from MS<sup>1</sup>
- Fragment with high energy collisions
- Break peptide bonds

Charge on amino-terminal (b) or carboxy-terminal fragment (y)

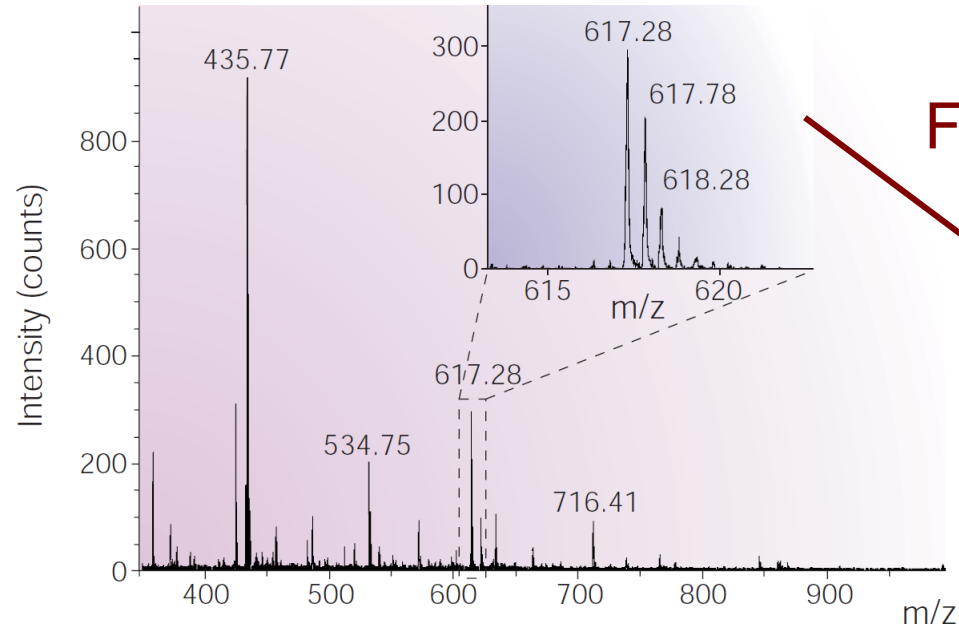
Subscript = R groups retained





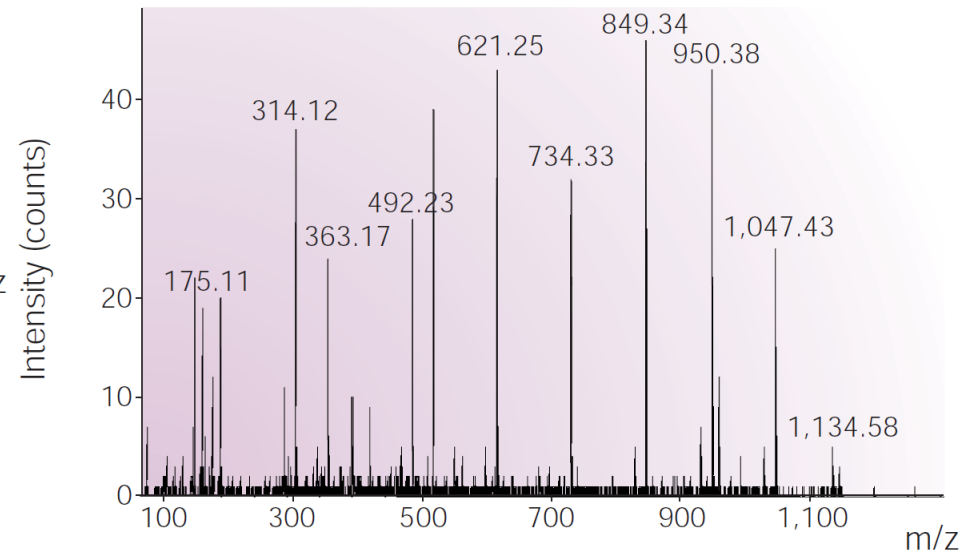
# Mass spectra

MS<sup>1</sup>



Fragment and analyze  
one precursor ion

MS<sup>2</sup>



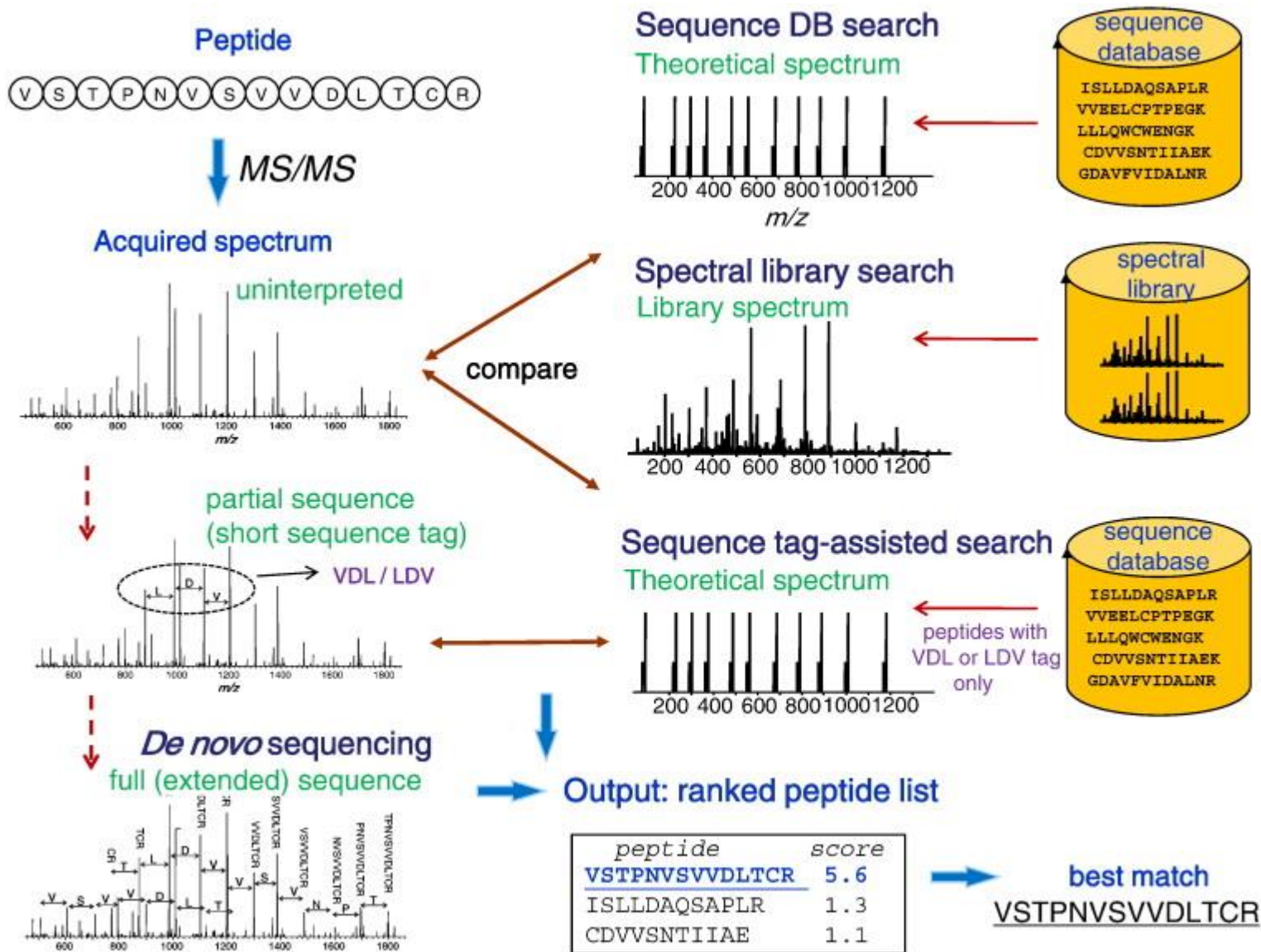
Steen and Mann *Nat Rev Mol Cell Biol* 2004

Mass-to-charge ratio



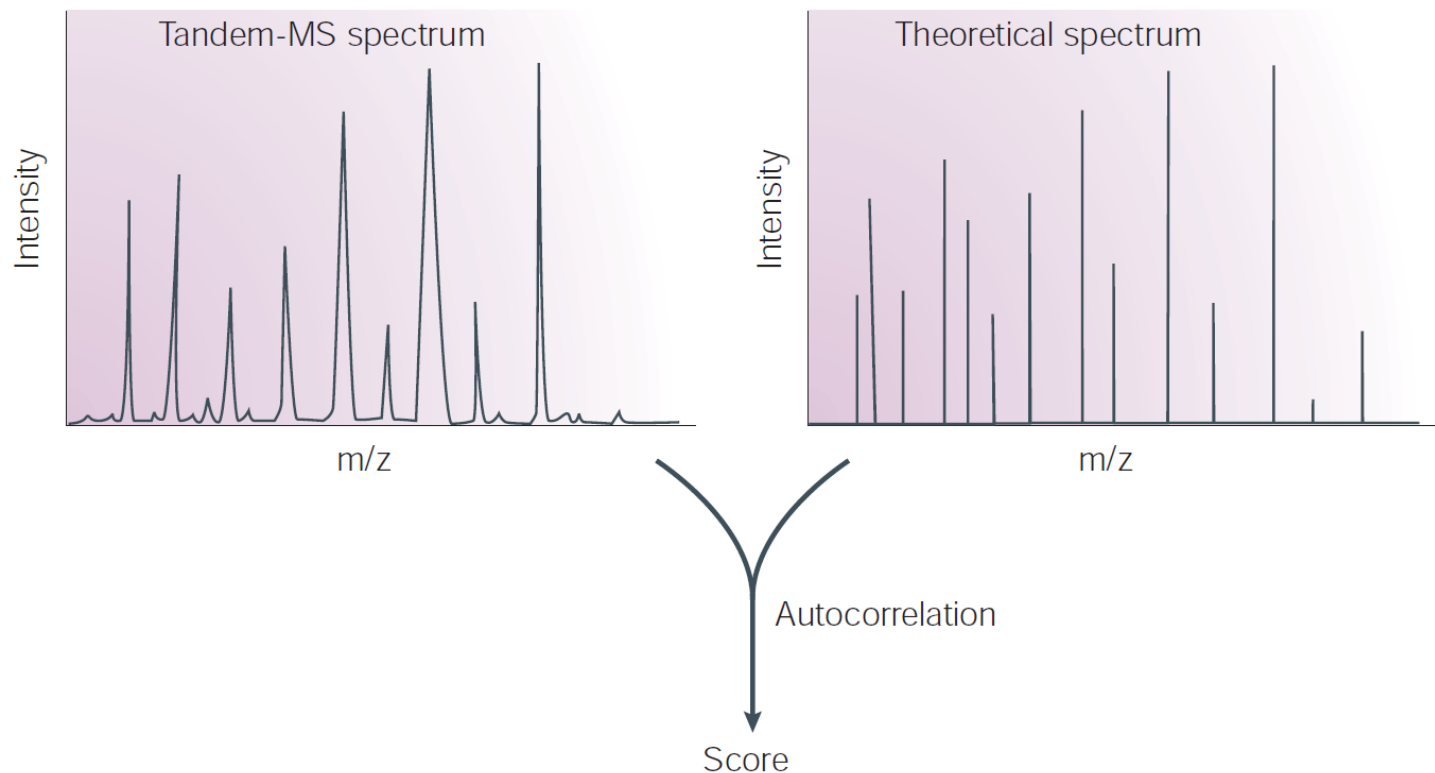
Spectrum contains information about amino  
acid sequence, fragment at different bonds

# From spectra to peptides



# Sequence database search

- Need to define a scoring function
- Identify peptide-spectrum match (PSM)



# SEQUEST

- Cross correlation (xcorr)
- Similarity between theoretical spectrum (x) and acquired spectrum (y)
- Correction for mean similarity at different offsets

$$\text{xcorr} = R_0 - \left( \sum_{\tau=-75}^{\tau=+75} R_{\tau} \right) / 151$$

Actual similarity

$$R_{\tau} = \sum x[i] \cdot y[i + \tau]$$

Theoretical

Acquired

Offsets

# Fast SEQUEST

- SEQUEST originally only applied to top 500 peptides based on coarse filtering score

$$\text{xcorr} = x_0 \cdot y_0 - \left( \sum_{\tau=-75}^{\tau=+75} x_0 \cdot y_{\tau} \right) / 151$$

$$\text{xcorr} = x_0 \cdot \left( y_0 - \left( \sum_{\tau=-75}^{\tau=+75} y_{\tau} \right) / 151 \right)$$

$$\text{xcorr} = x_0 \cdot y' \quad \text{where} \quad y' = y_0 - \left( \sum_{\tau=-75, \tau \neq 0}^{\tau=+75} y_{\tau} \right) / 150$$

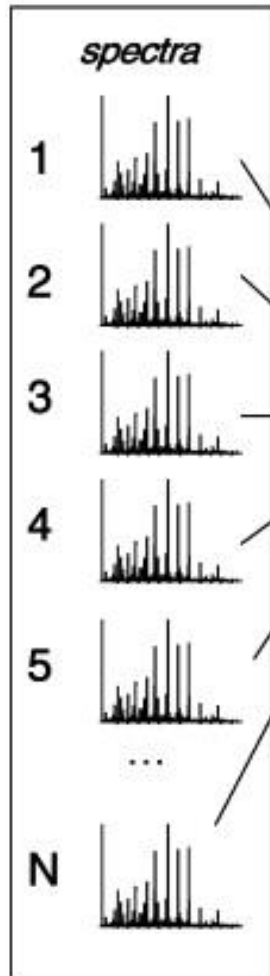
 Skip the 0 offset

# PSM significance

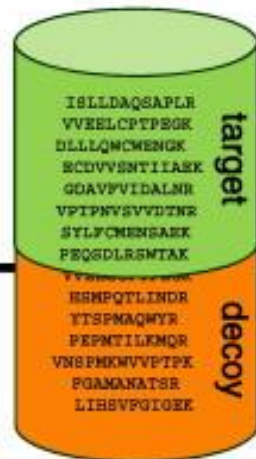
- E-value: expected number of null peptides with score  $\geq$  observed score
- Compute FDR from E-value distribution
- Add decoy peptides to database
  - Reversed peptide sequences
  - Used to estimate false discoveries

# Target-decoy strategy

Entire dataset, N spectra



database  
search



Filtering using  
target-decoy  
strategy

Best match for each spectrum

spec	peptide	score	label
1	ISLLDAQSAPLR	4.5	target
2	VVEELCTPEGR	3.9	target
5	GDAVFVIDALNR	3.6	target
3	VNSEPMKVVPK	1.7	decoy
4	ECDVVSNTIIAEK	1.5	target
...	...	...	...
N	LIHSVFGIGEK	1.1	decoy

(sorted by score)

Apply score threshold  $S_T$

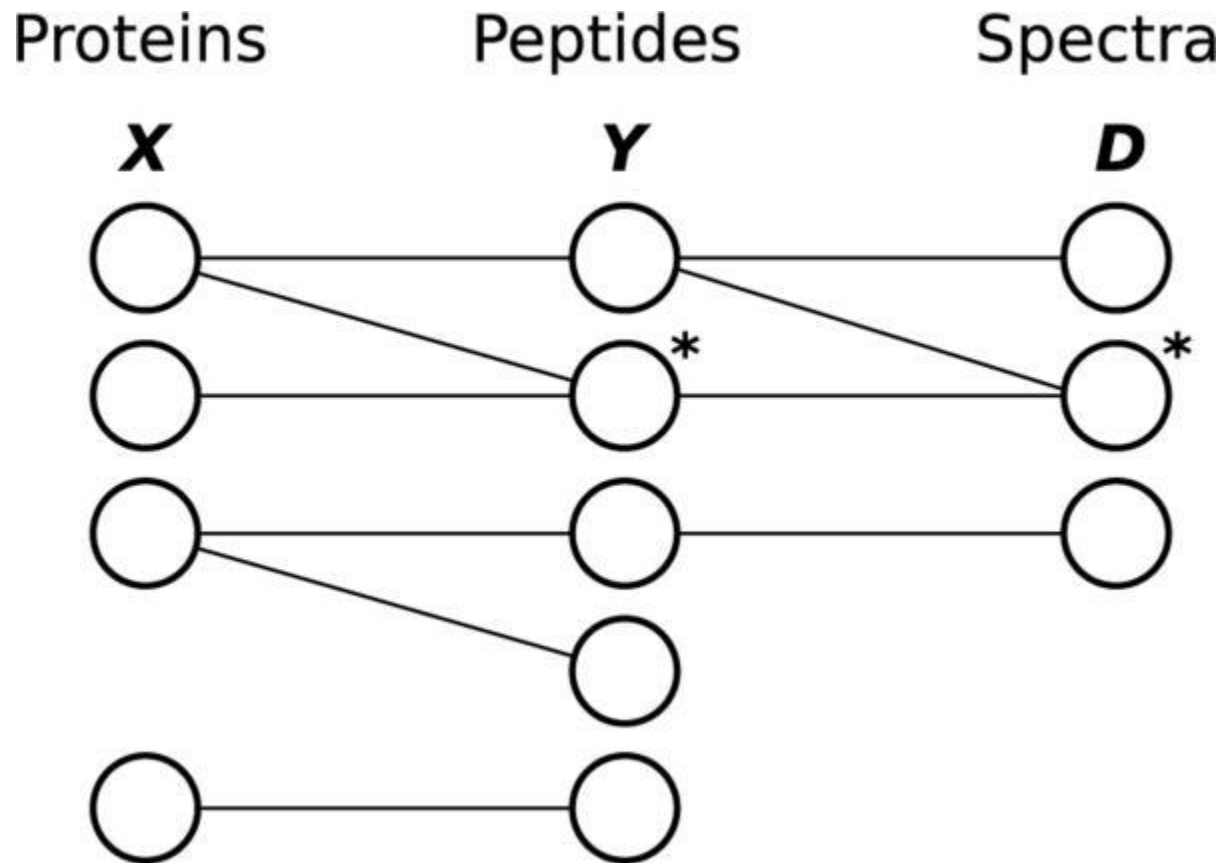
Calculate  $N_t(S_T)$  and  $N_d(S_T)$ :  
number of target/decoy PSM with  $S \geq S_T$

Estimate FDR 
$$FDR(S_T) = \frac{N_d(S_T)}{N_t(S_T)}$$

Select threshold  $S_T$  to achieve desired FDR

# Identifying proteins

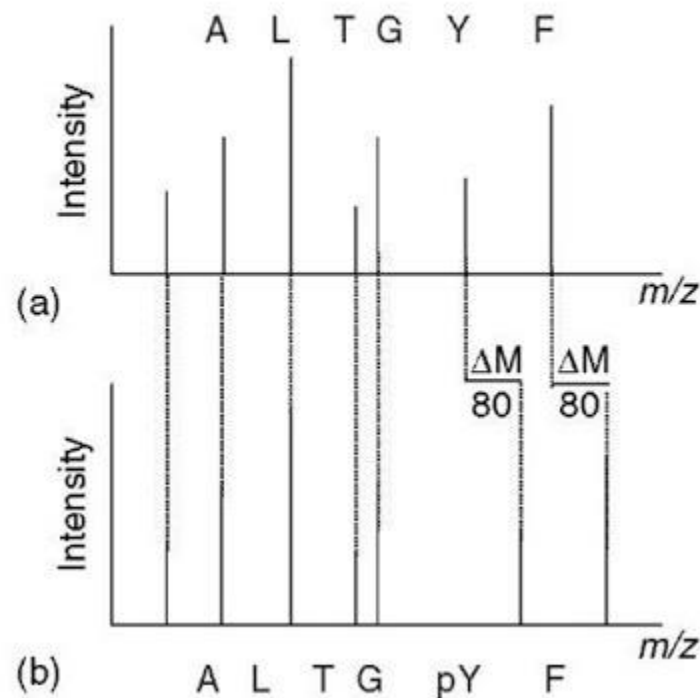
- Even after identifying PSM, still need to identify protein of origin





# Post-translational modifications (PTMs)

- Shift the peptide mass by a known quantity



[what-when-how](#)

# Mass spectrometry versus RNA-seq

- RNA-seq
  - Transcript → RNA fragment → paired-end read
- Mass spectrometry
  - Protein → peptides → ions → spectrum
- Mapping spectra to proteins more ambiguous than mapping reads to transcripts
- Spectra state space is enormous

# Mass spectrometry replicates

- Doesn't identify all proteins in the sample
  - Old technology had low overlap across replicates
  - Partly due to biology variation
- Phosphorylation PTMs are especially variable
  - Wolf-Yadlin lab (unpublished)
    - 3 biological replicates
    - 5,442 phosphopeptides identified
    - 19.6% identified in all replicates
  - Grimsrud et al *Cell Metabolism* 2012
    - 5 biological replicates
    - 9,558 phosphoproteins identified
    - 5.6% in all replicates

# Mass spectrometry summary

- Incredibly powerful for looking at biological processes beyond gene expression
  - Protein abundance
  - Post-translational modifications
  - Metabolites
  - Protein-protein interactions
- Typically reports relative abundance
- Labeling strategies for comparative analysis
  - Compare relative abundance in multiple conditions
- Missing data is a big problem, but improving
- Fully probabilistic analysis pipelines are not the most popular tools
  - Arguably greater diversity in software than RNA-seq