# Mass spectrometry-based proteomics

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2020

Daifeng Wang

daifeng.wang@wisc.edu

# Goals for lecture

Key concepts

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

# Mass spectrometry uses

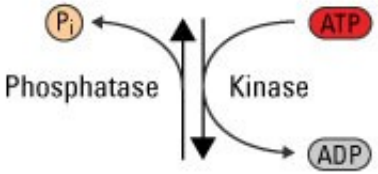
- Mass spectrometry is like the protein analog of RNA-seq
  - Quantify abundance or state of all (many) proteins
  - No need to specify proteins to measure in advance

- Other applications in biology
  - Targeted proteomics
  - Metabolomics
  - Lipidomics

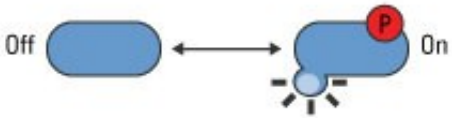
# Advantages of proteomics

- Proteins are functional units in a cell
  - Protein abundance directly relevant to activity

- Post-translational modifications
  - Change protein state

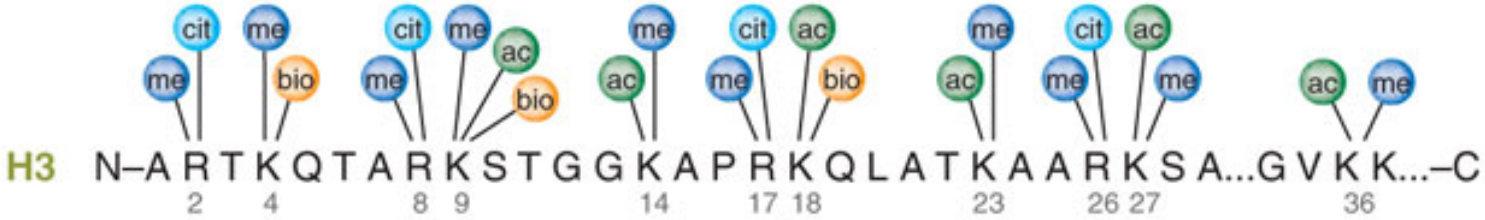



Phosphorylation in signaling




Thermo Fisher Scientific


Histone modifications

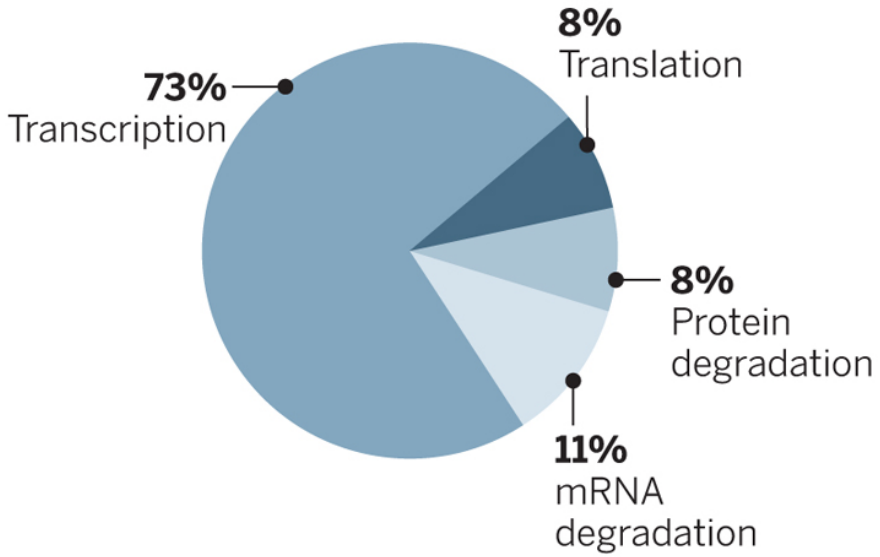



Latham *Nature Structural & Molecular Biology* 2007; Katie Ris-Vicari

# Estimating protein levels from gene expression

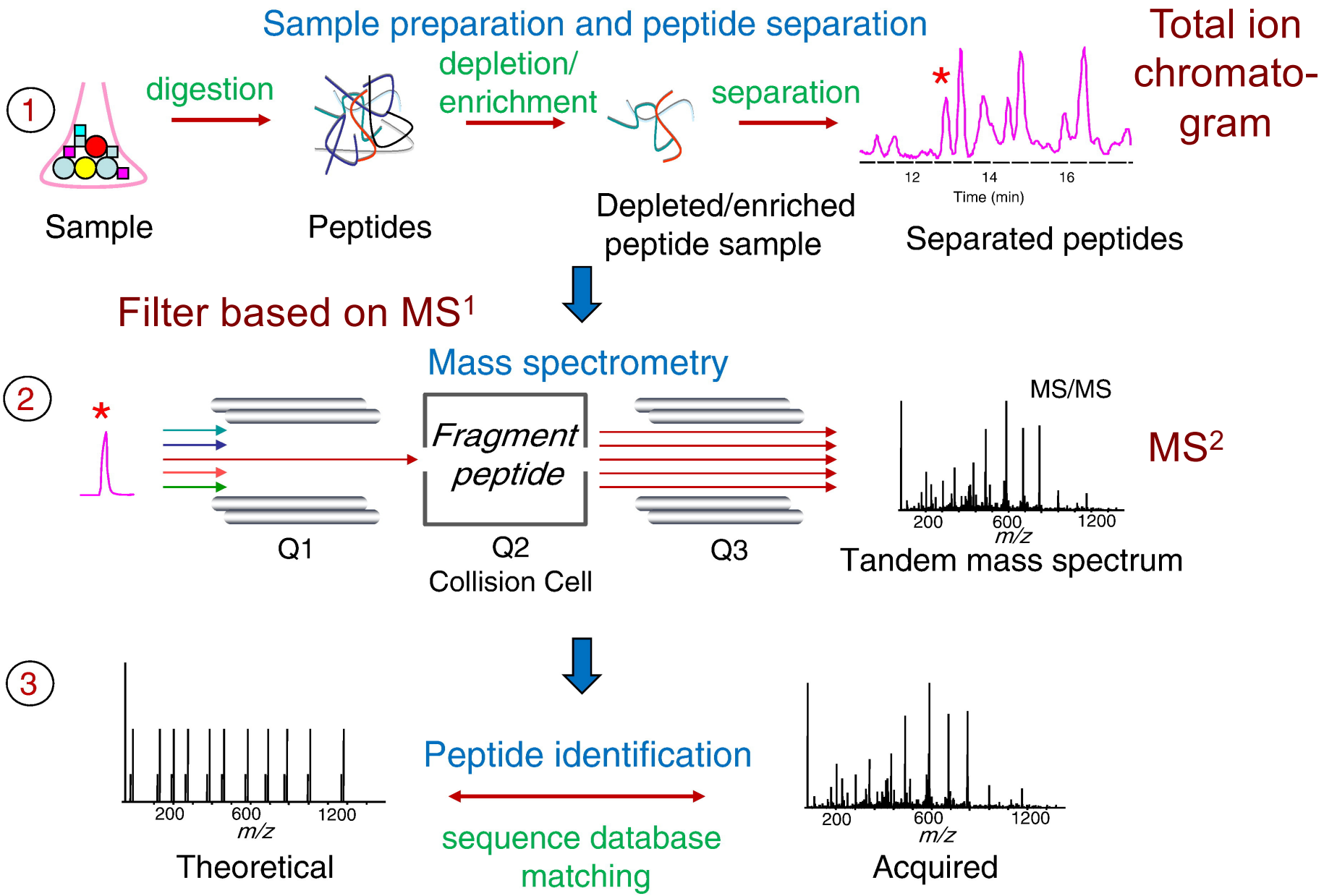
- Correlation between gene expression and protein abundance has been debated
- Gene expression tells us nothing about post-translational modifications

Contribution to protein levels




Li and Biggin *Science* 2015

# Mass spectrometry workflow




Nesvizhskii *Journal of Proteomics* 2010

m/z is mass-to-charge raio (unit: Dalton or Da)

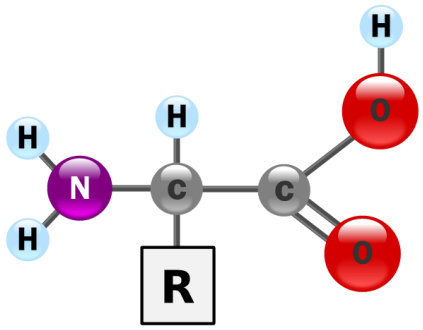
# Amino Acids

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

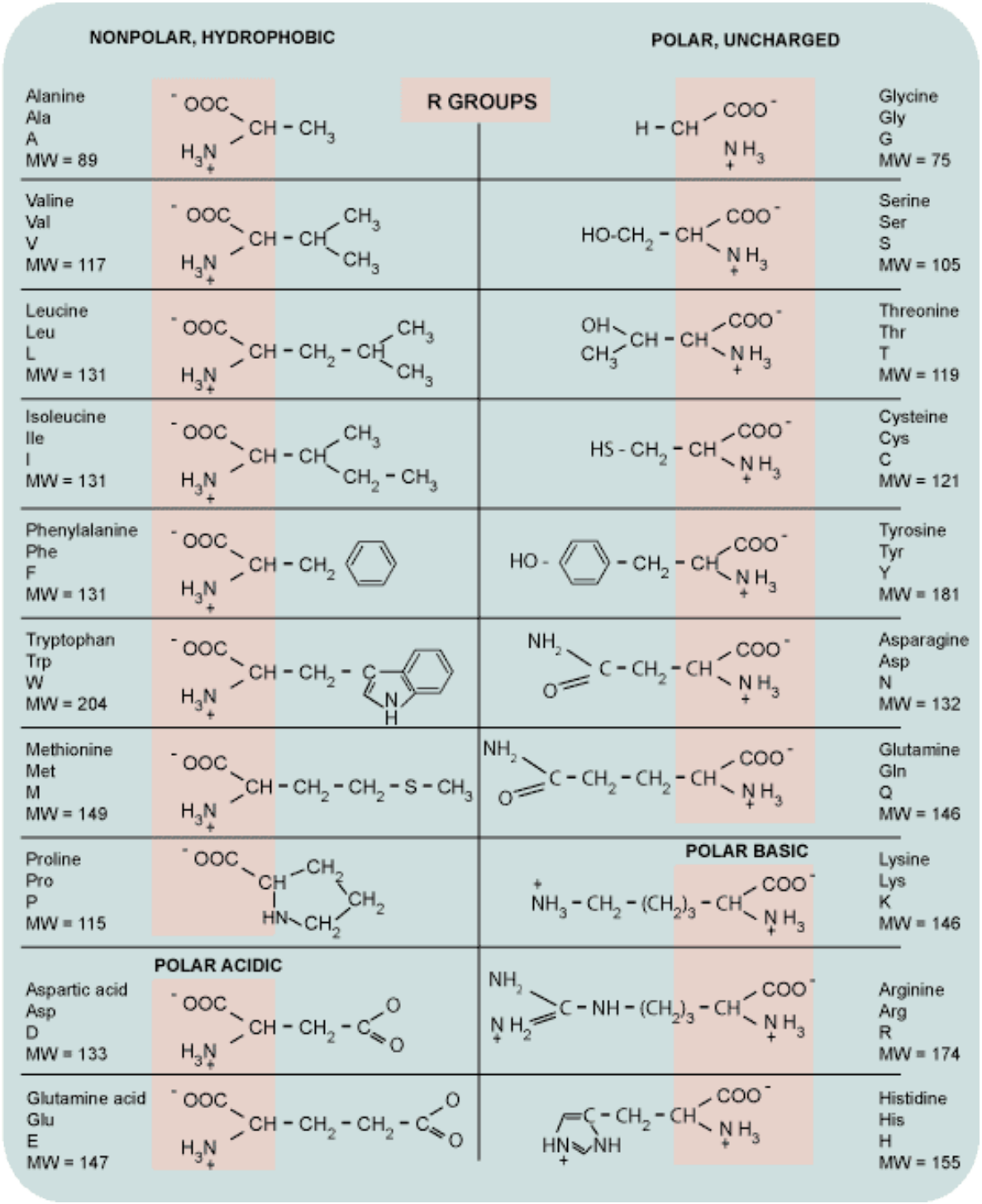
- Common template

Amino-terminal

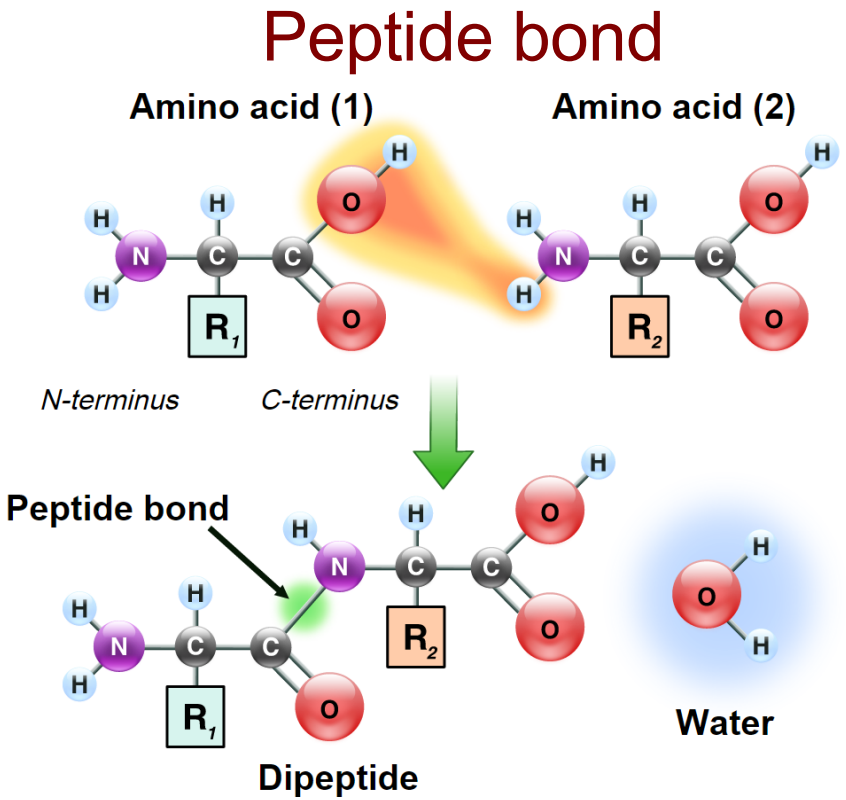
Carboxy-terminal




Wikipedia, Yassine Mrabet



# Peptide fragmentation

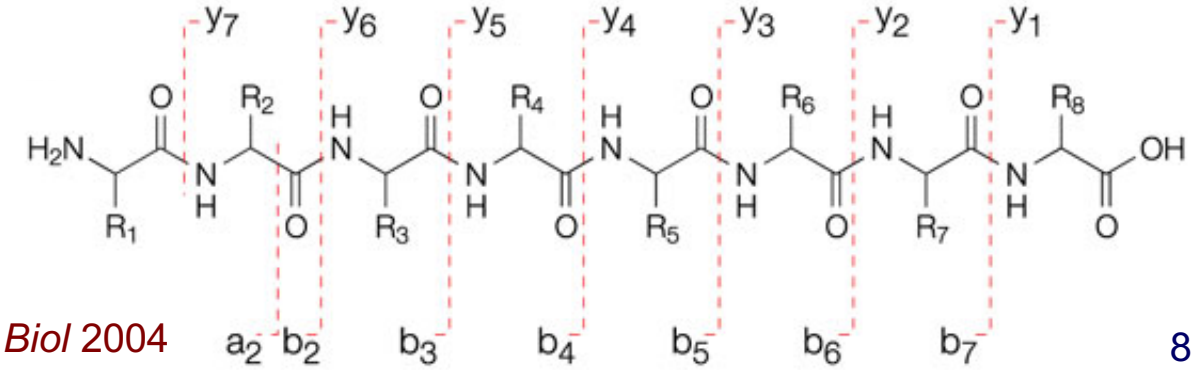



Wikipedia, Yassine Mrabet

- Select similar peptides from $MS^1$
- Fragment with high energy collisions
- Break peptide bonds

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

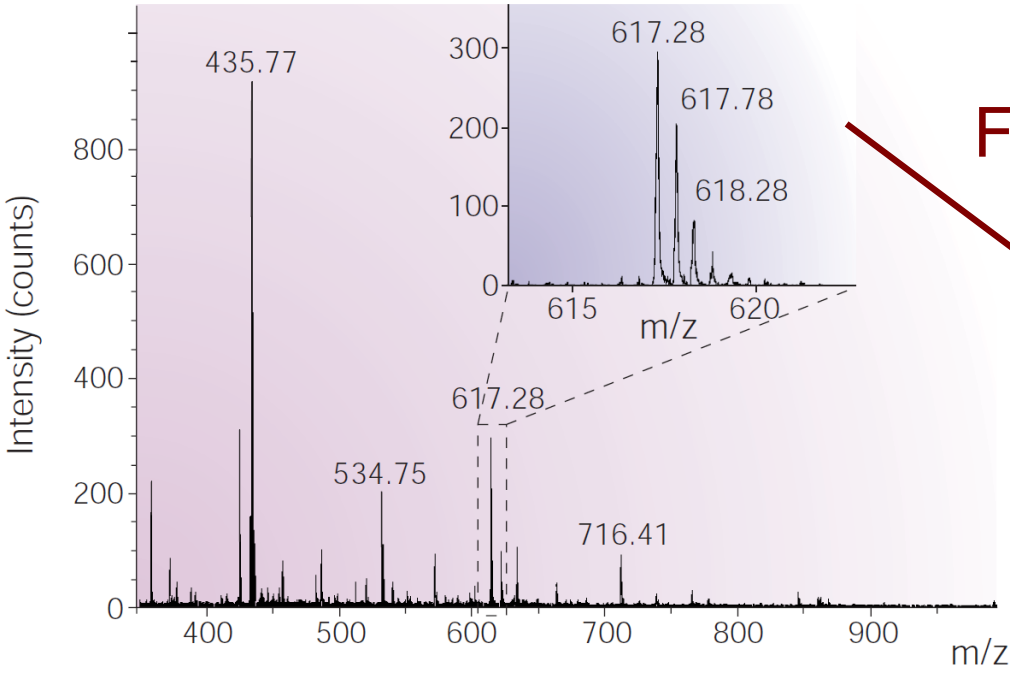
Subscript = # R groups retained




Steen and Mann *Nat Rev Mol Cell Biol* 2004

# Mass spectra

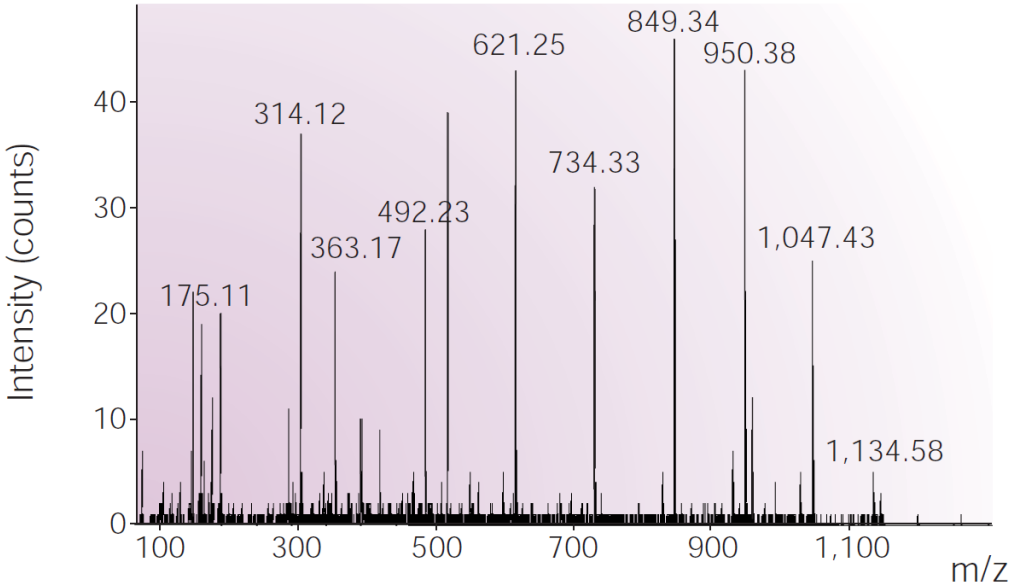
$MS^1$




Steen and Mann *Nat Rev Mol Cell Biol* 2004

Fragment and analyze one precursor ion

$MS^2$

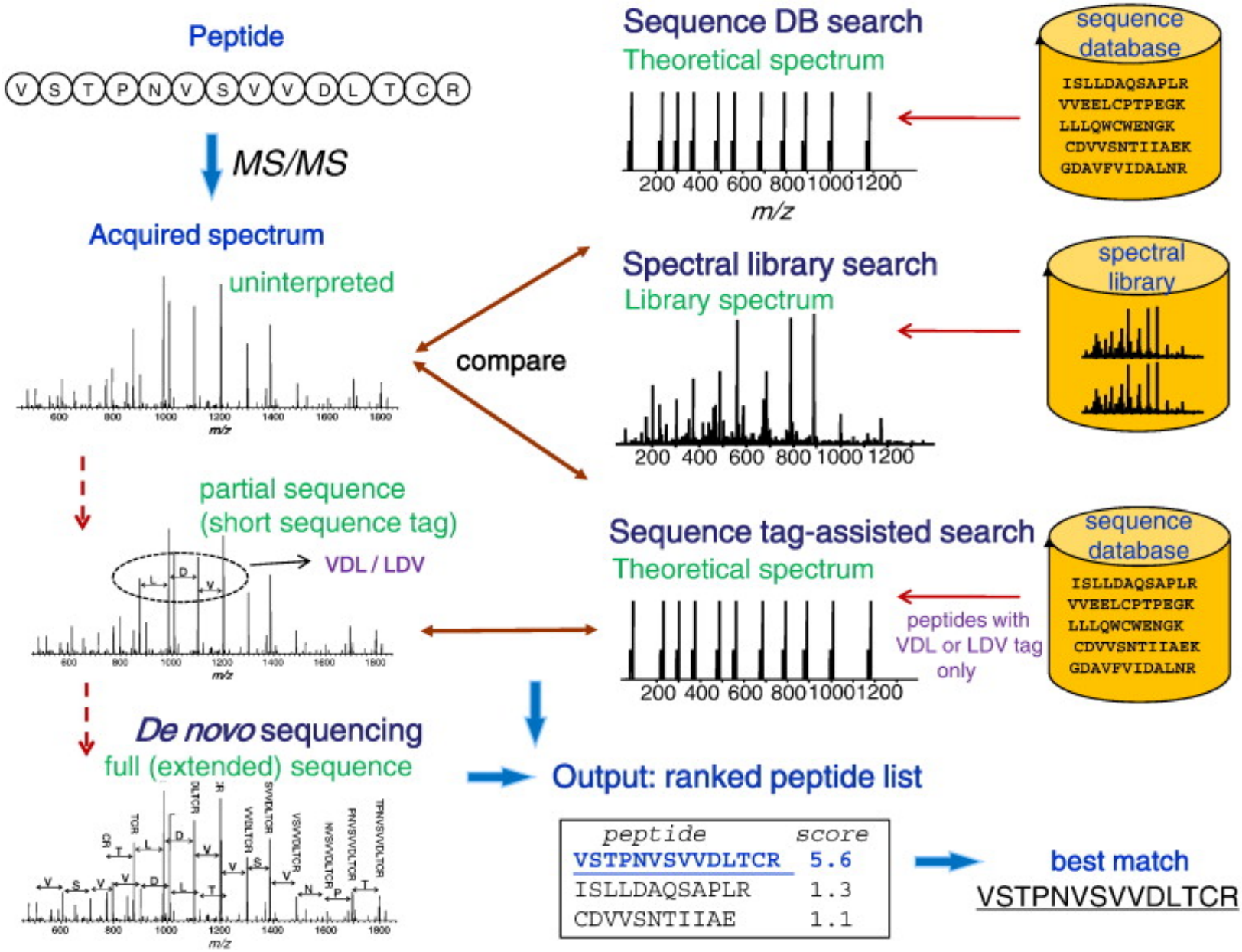




Mass-to-charge ratio

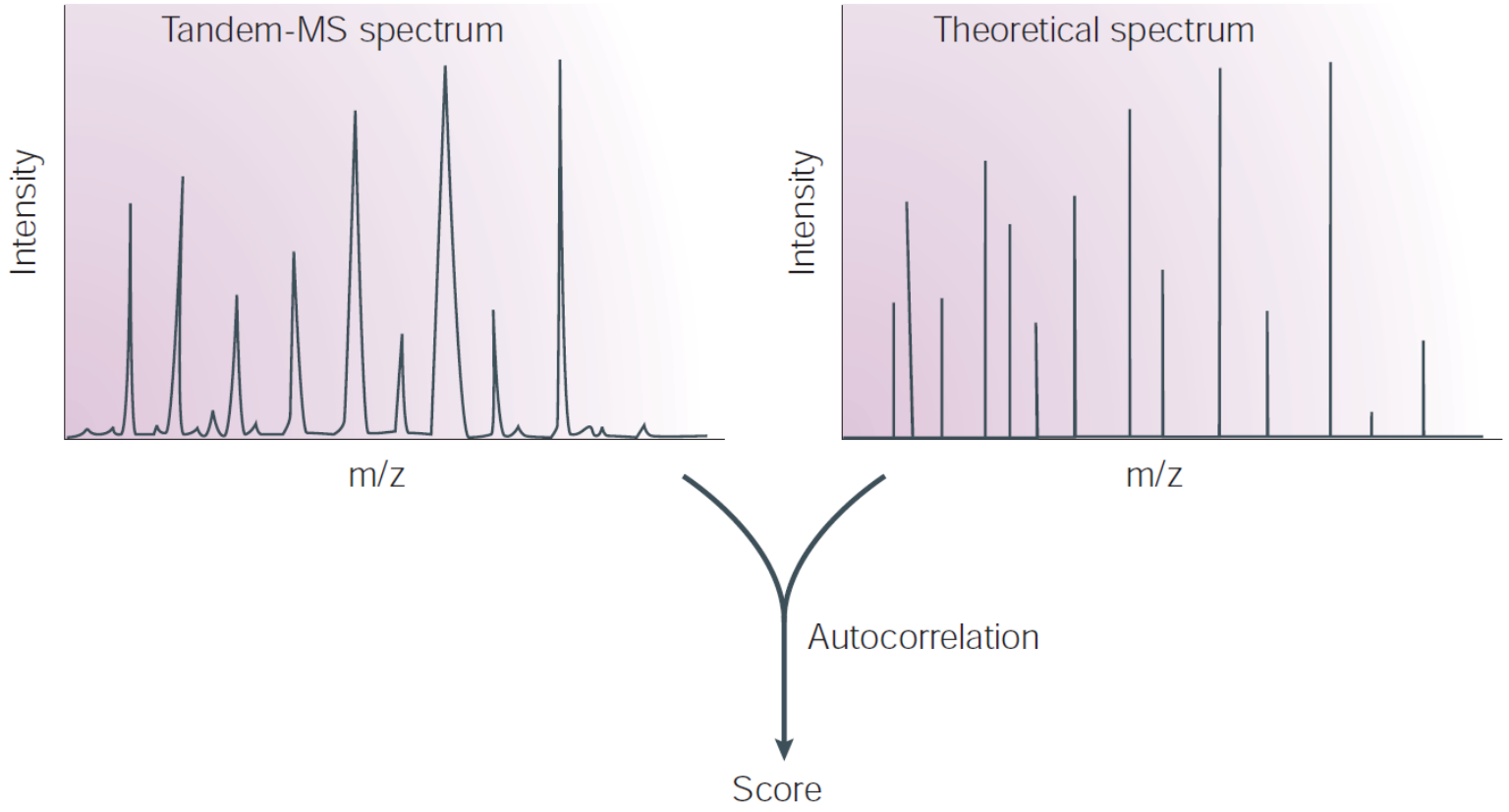
Spectrum contains information about amino acid sequence, fragment at different bonds

# From spectra to peptides



# Sequence database search

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




Steen and Mann *Nat Rev Mol Cell Biol* 2004

# SEQUEST

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

Offsets

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

Actual similarity

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

Theoretical

Acquired

Eng, McCormack, Yates *J Am Soc Mass Spectrom* 1994

# Fast SEQUEST

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

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

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

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

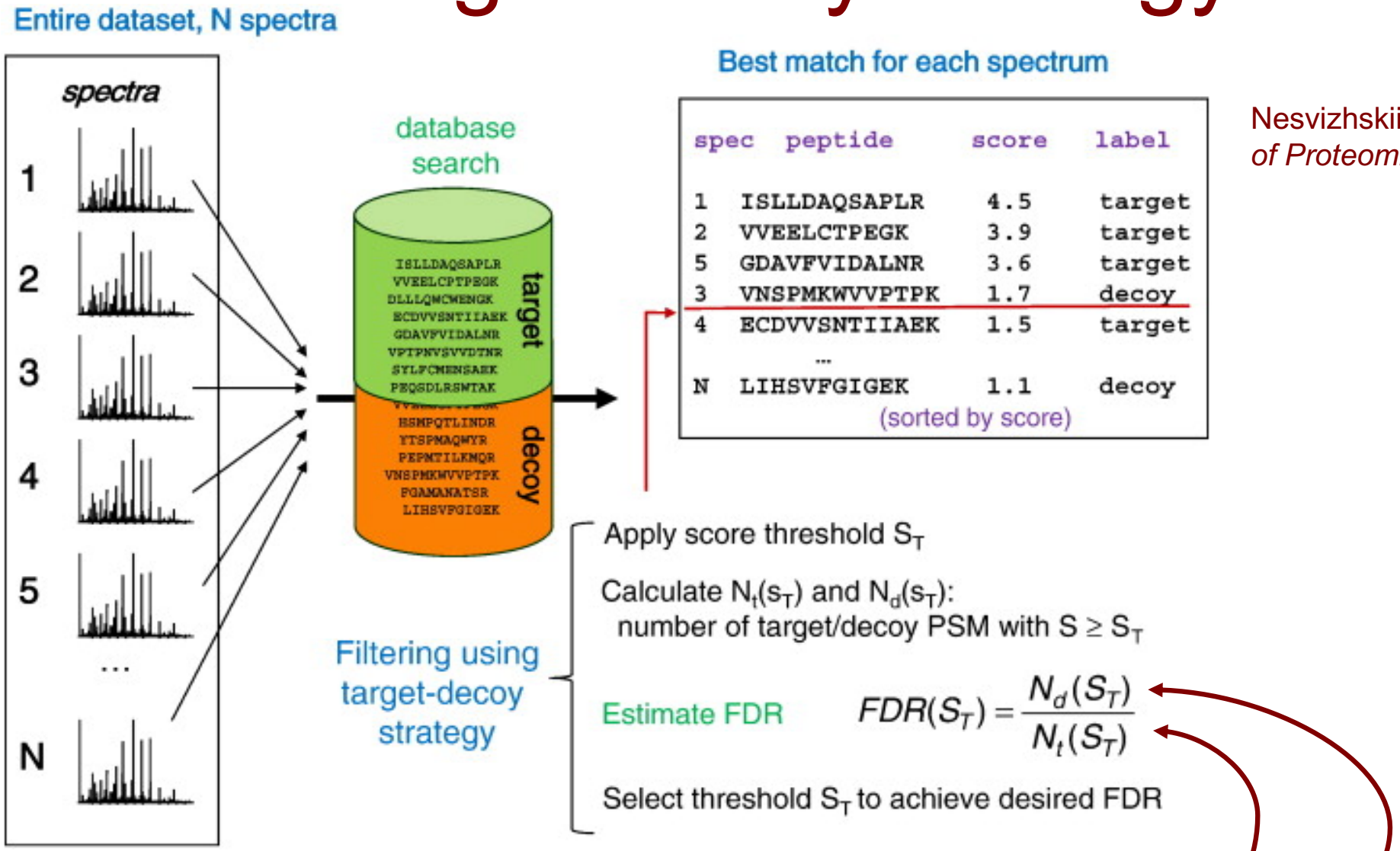
Skip the 0 offset

Eng et al *J Proteome Res* 2008

# PSM significance

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

# Target-decoy strategy



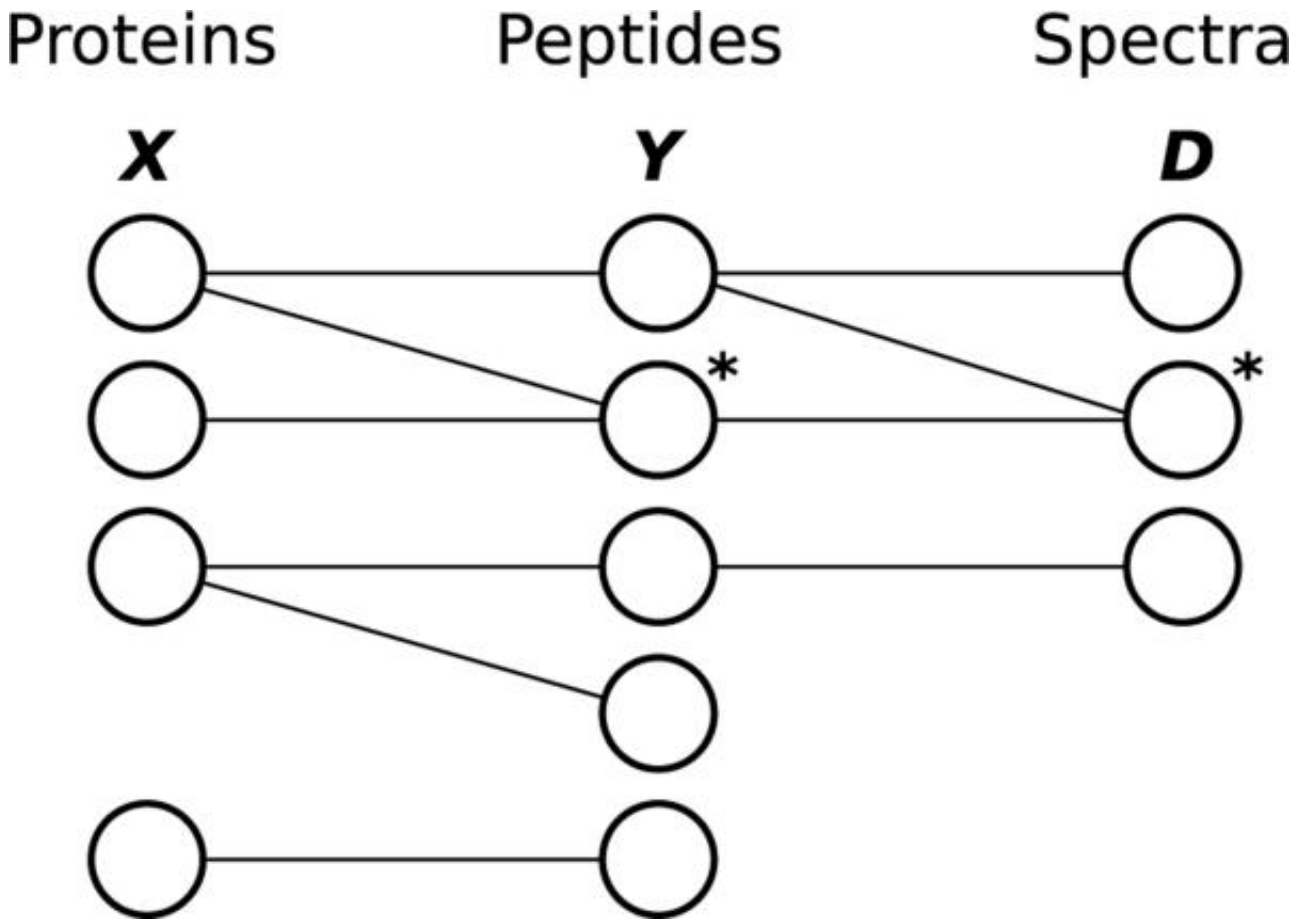

Nesvizhskii *Journal of Proteomics* 2010

target PSMs above score threshold = $N_t(S_T)$

decoy PSMs above score threshold = $N_d(S_T)$

# Identifying proteins

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

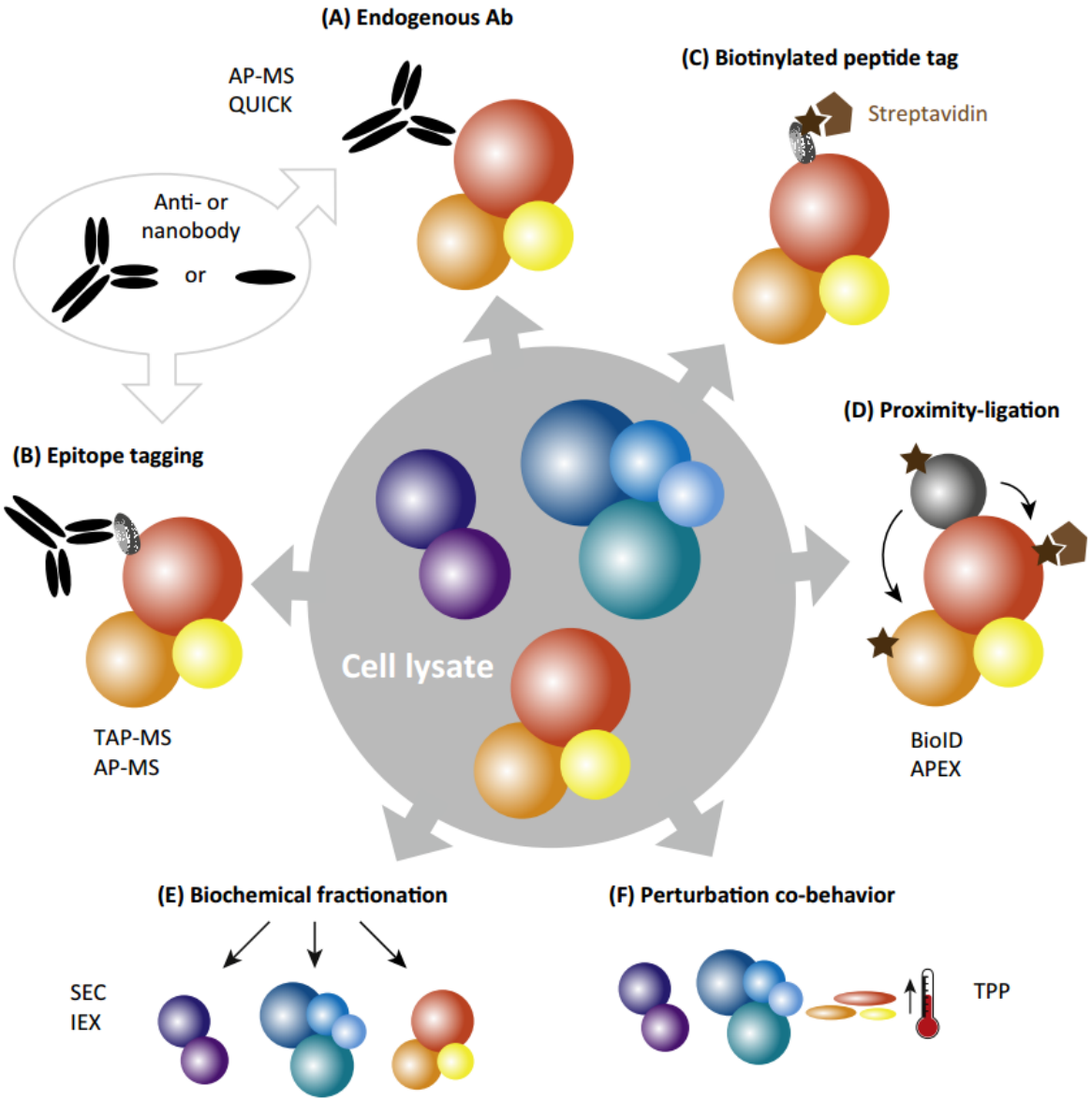



Serang and Noble *Stat Interface* 2012

# 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 genes
- Spectra state space is enormous

# Protein-protein interactions

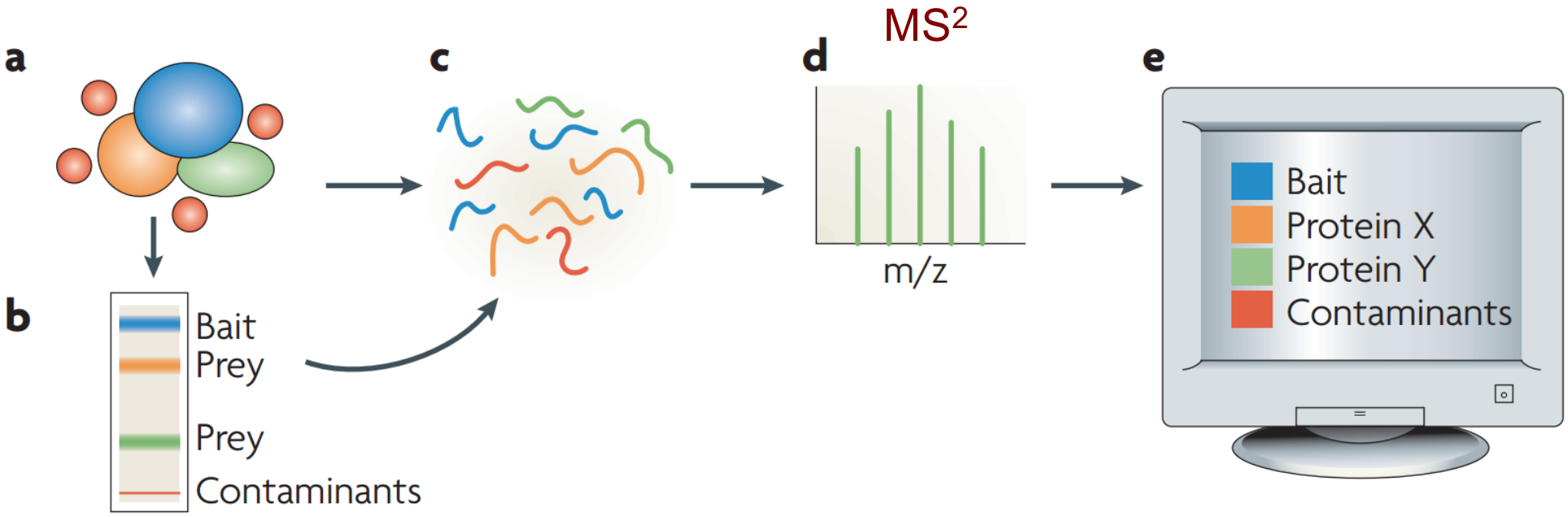



- Affinity-purification mass spectrometry
- Purify protein of interest, identify complex members

Smits and Vermeulen T*rends in Biotechnology* 2016

# Protein-protein interactions

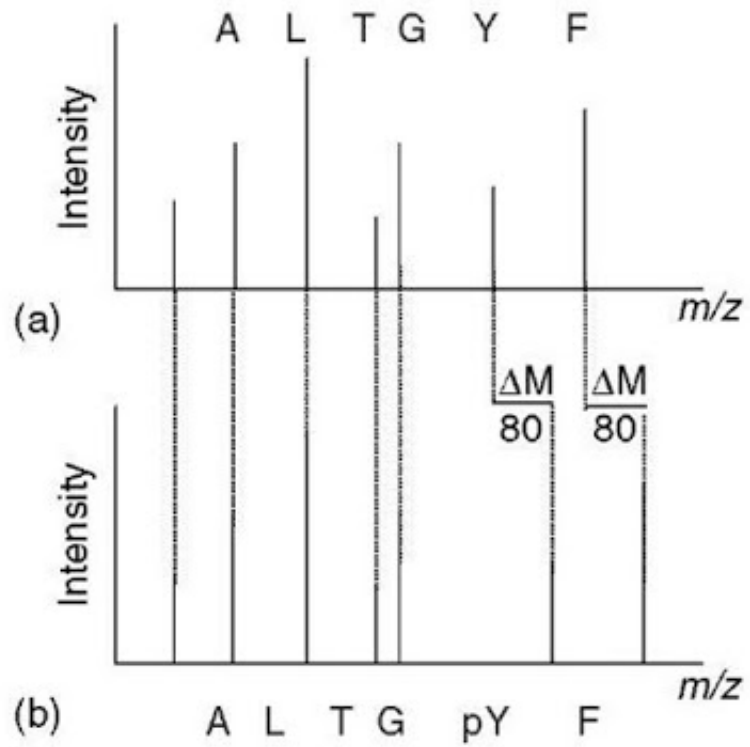
- Mass spectrometry identifies proteins in the complex
- Must control for contaminants




Gingras et al *Nature Reviews Molecular Cell Biology* 2007

# Post-translational modifications (PTMs)

- Shift the peptide mass by a known quantity




what-when-how

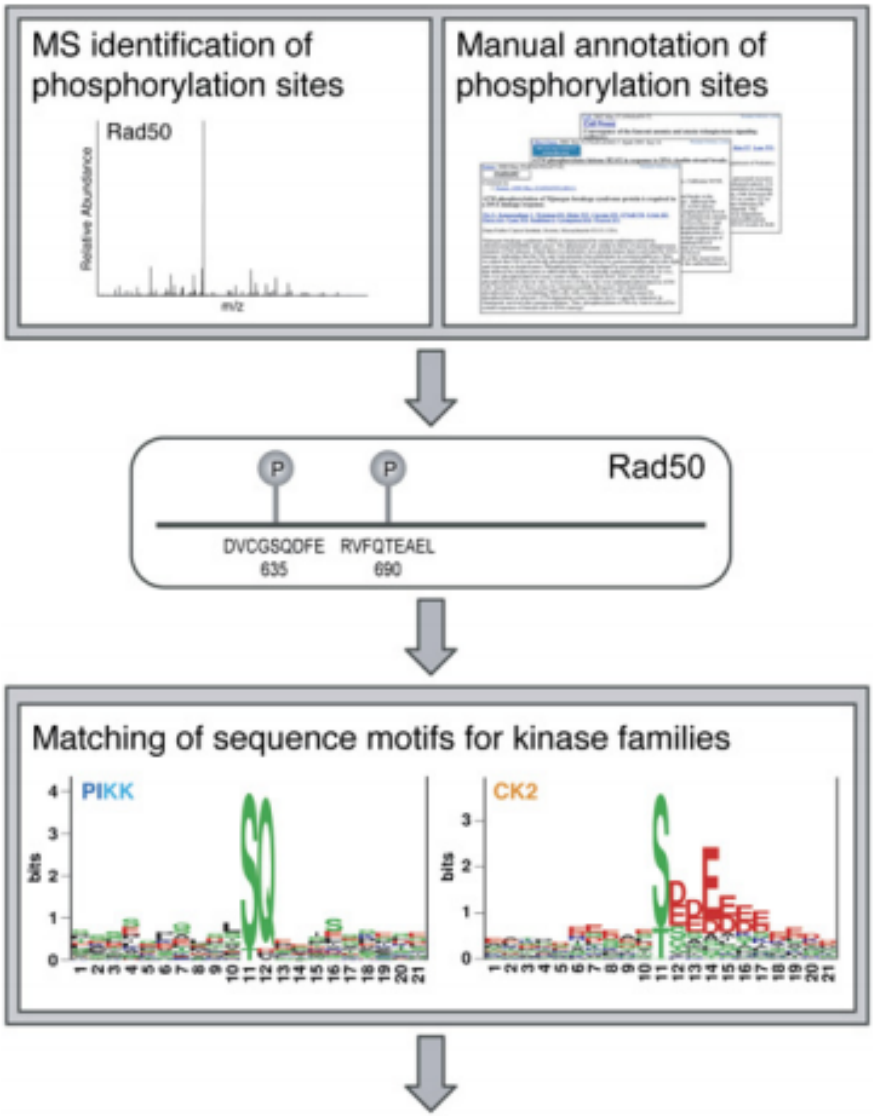
# Phosphoproteomics example

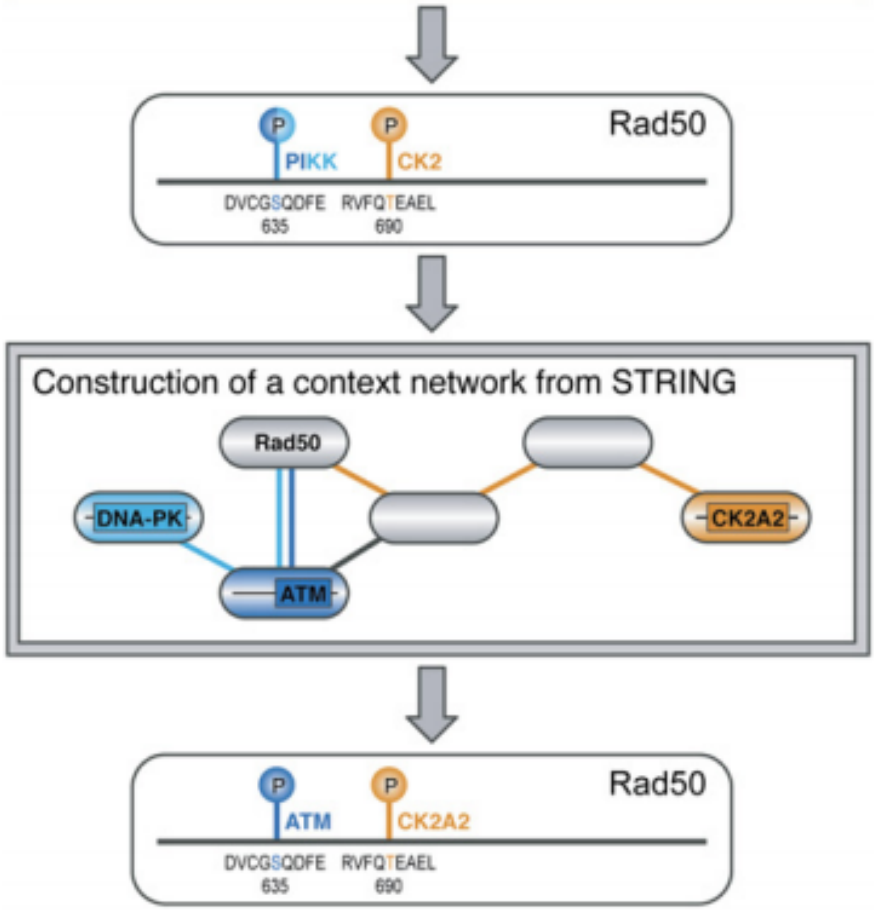
| Gene | Modified Site | Peptide | Phosphorylation (Treatment / Control) |
|---|---|---|---|
| AGRN | S671 | AGPC[160.03]EQAEC[160.03]GS[167.00]GGSGSGEDGDC[160.03]EQELC[160.03]R | 4.54 |
| ADAMTS10 | S74 | RGTGATAES[167.00]R | 0.30 |
| CABYR | T16 | T[181.01]LLEGISR | 0.37 |
| TTC7B | T152 | VIEQDET[181.01]R | 5.97 |
| STAT3 | Y705 | K.n[305.21]YC[160.03]RPESQEHPEADPGSAAPY[243.03]LK[432.30].T | 4.50 |

Sychev et al *PLoS Pathogens* 2017

# Phosphoproteomics interpretation

- Predict kinases/phosphatases for phospho sites




Linding et al *Cell* 2007

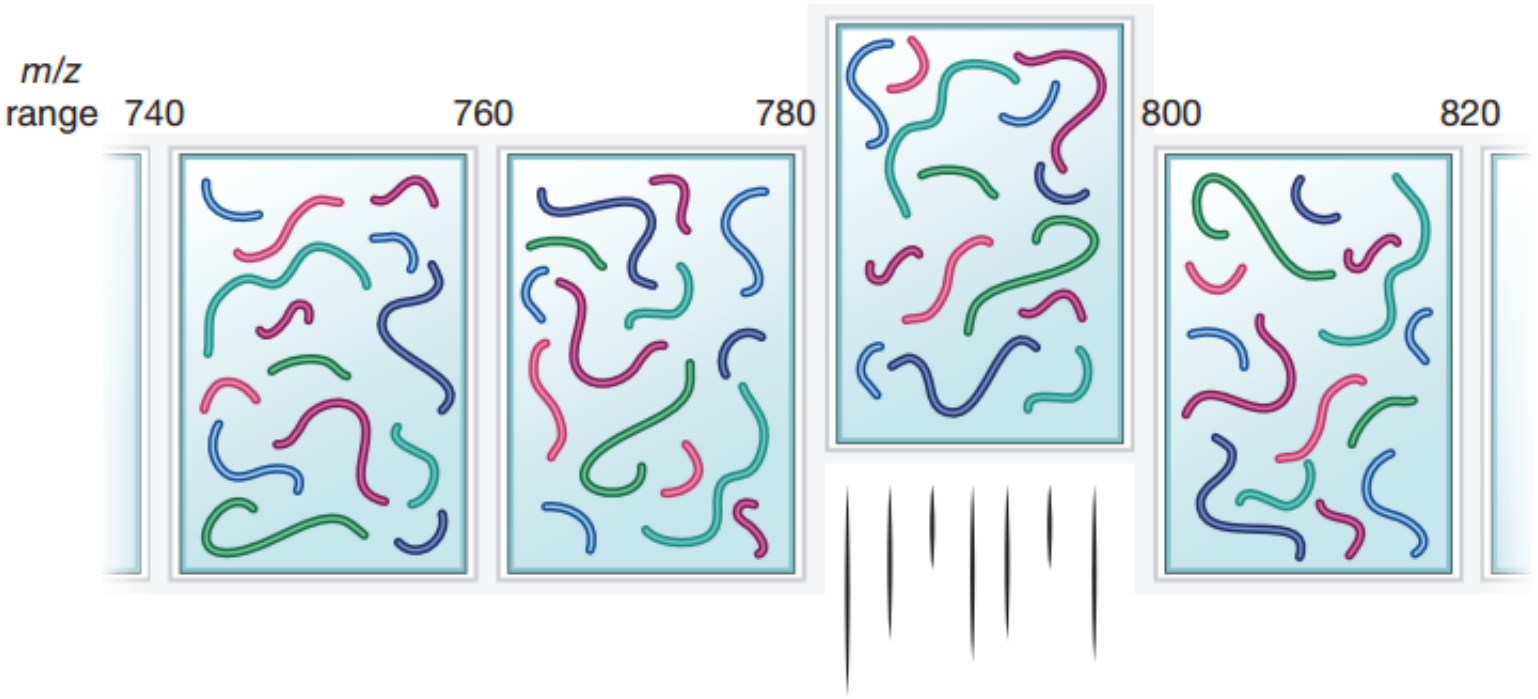
# Mass spectrometry replicates

- Doesn't identify all proteins in the sample
  - Data dependent acquisition has low overlap across replicates
  - Partly due to biological variation
  - New protocols to overcome this

- Phosphorylation PTMs are especially variable
  - Grimsrud et al *Cell Metabolism* 2012
    - 5 biological replicates
    - 9,558 phosphoproteins identified
    - 5.6% in all replicates

# Data independent acquisition

- Not dependent on most abundance signals in $MS^1$
- Sliding *m*/*z* window




Doerr *Nature Methods* 2015

# 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 was a big problem, but improving
- Fully probabilistic analysis pipelines are not the most popular tools
  - Arguably greater diversity in software than RNA-seq