# Mass spectrometry-based proteomics

Lukas Käll

# Overview

Mass spectrometry

- Ionization sources
- MS Technologies
- Fragmentation

Proteomics

- Why?
- Dynamic range
- Complexity
- Limitations

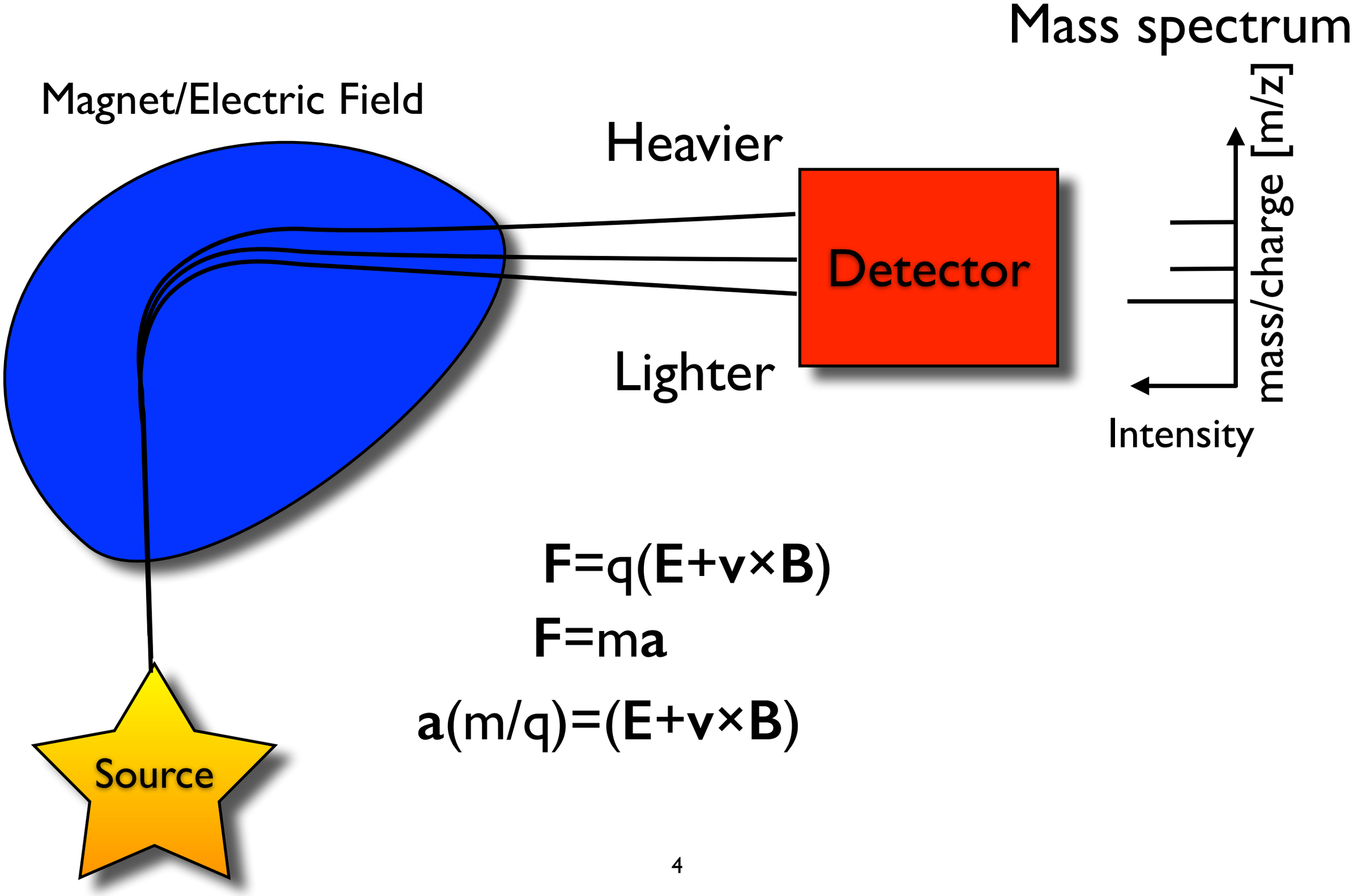
Peptide Identification

- Search engines
- Statistics
- Post-translational Modifications (PTMs)

Protein inference

# Mass spectrometry

# Mass spectrometry



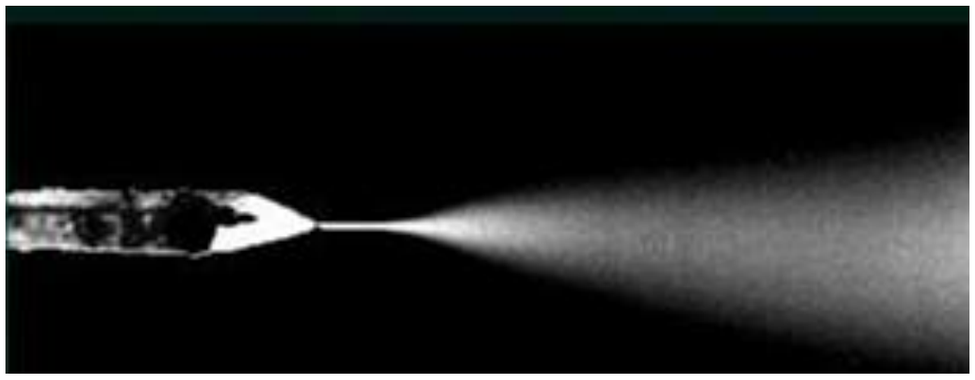

$$\mathbf{F}=q(\mathbf{E}+\mathbf{v}\times\mathbf{B})$$

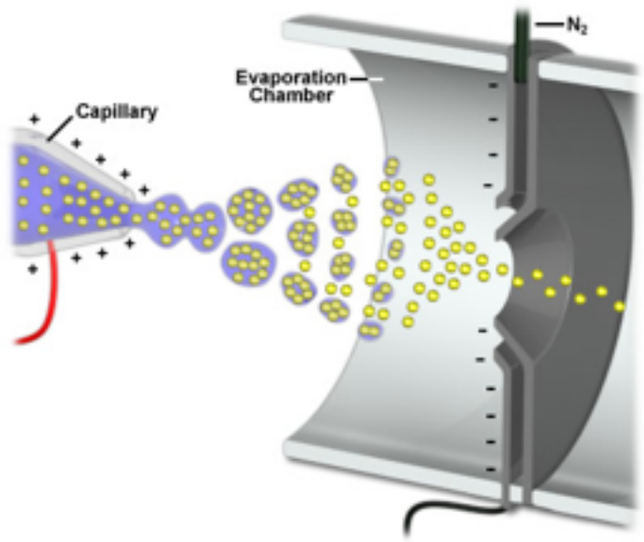
$$\mathbf{F}=m\mathbf{a}$$

$$\mathbf{a}(m/q)=(\mathbf{E}+\mathbf{v}\times\mathbf{B})$$

# Ion-Sources

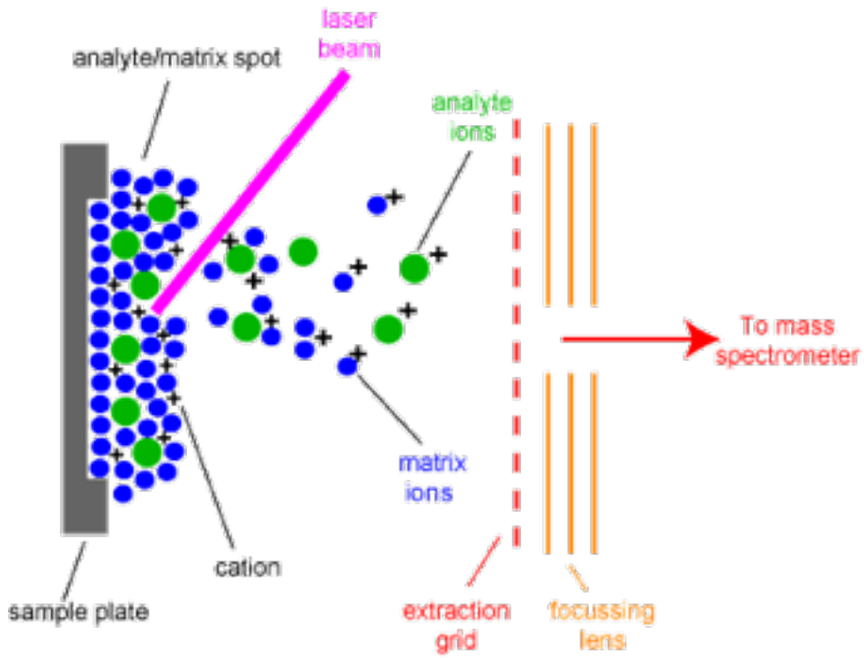
## Electro spray ionization





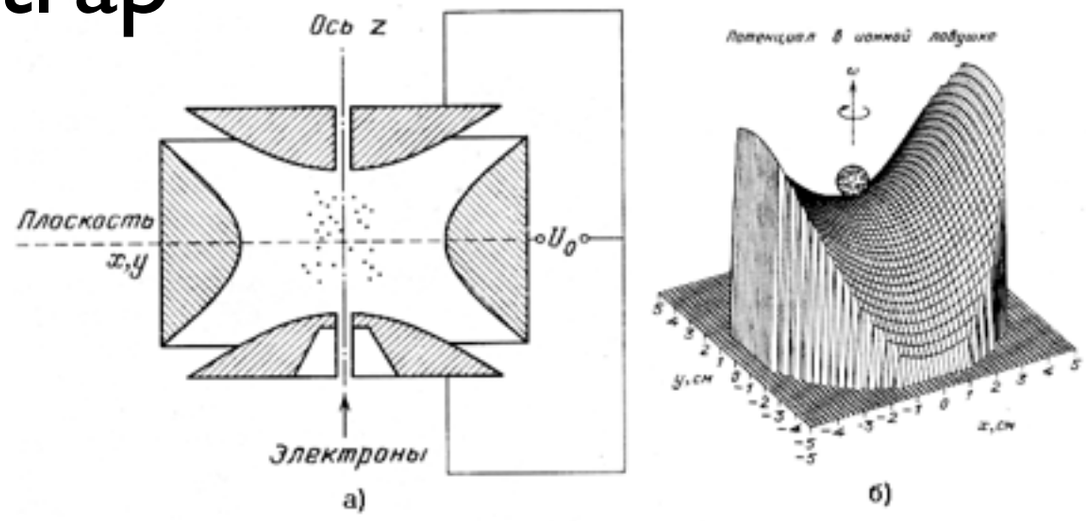

Nobel Prize 2002
John Fenn

## MALDI



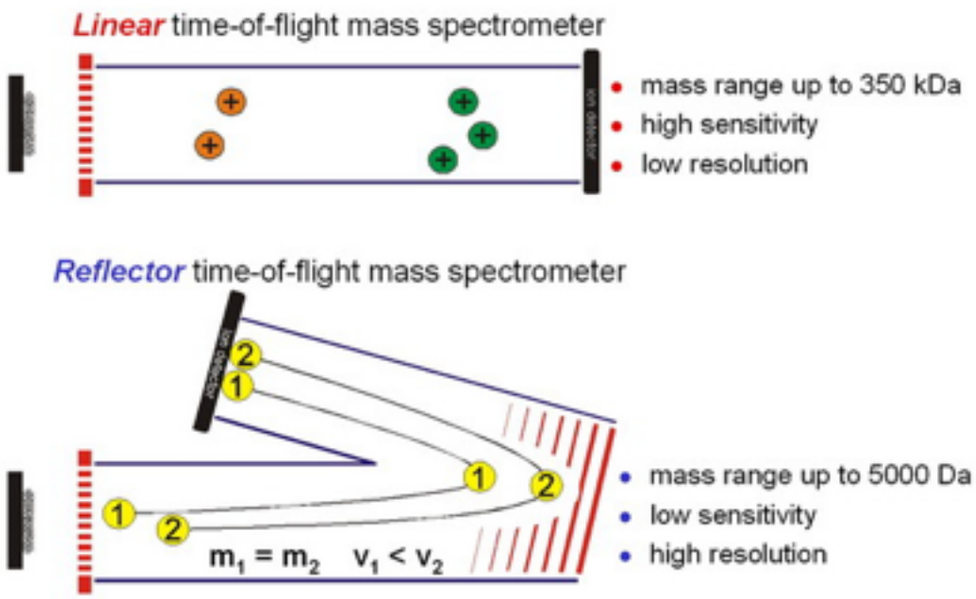
# Separation

## Ion trap

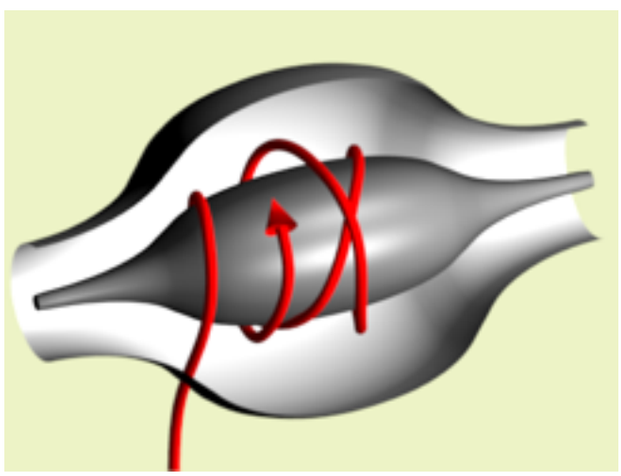



- Ions kept in place by AC current generating an electric field
- The ions will leave the trap as we increase the voltage of the control in order of their m/z (lowest first)

## Time of Flight (TOF)




## OrbiTrap

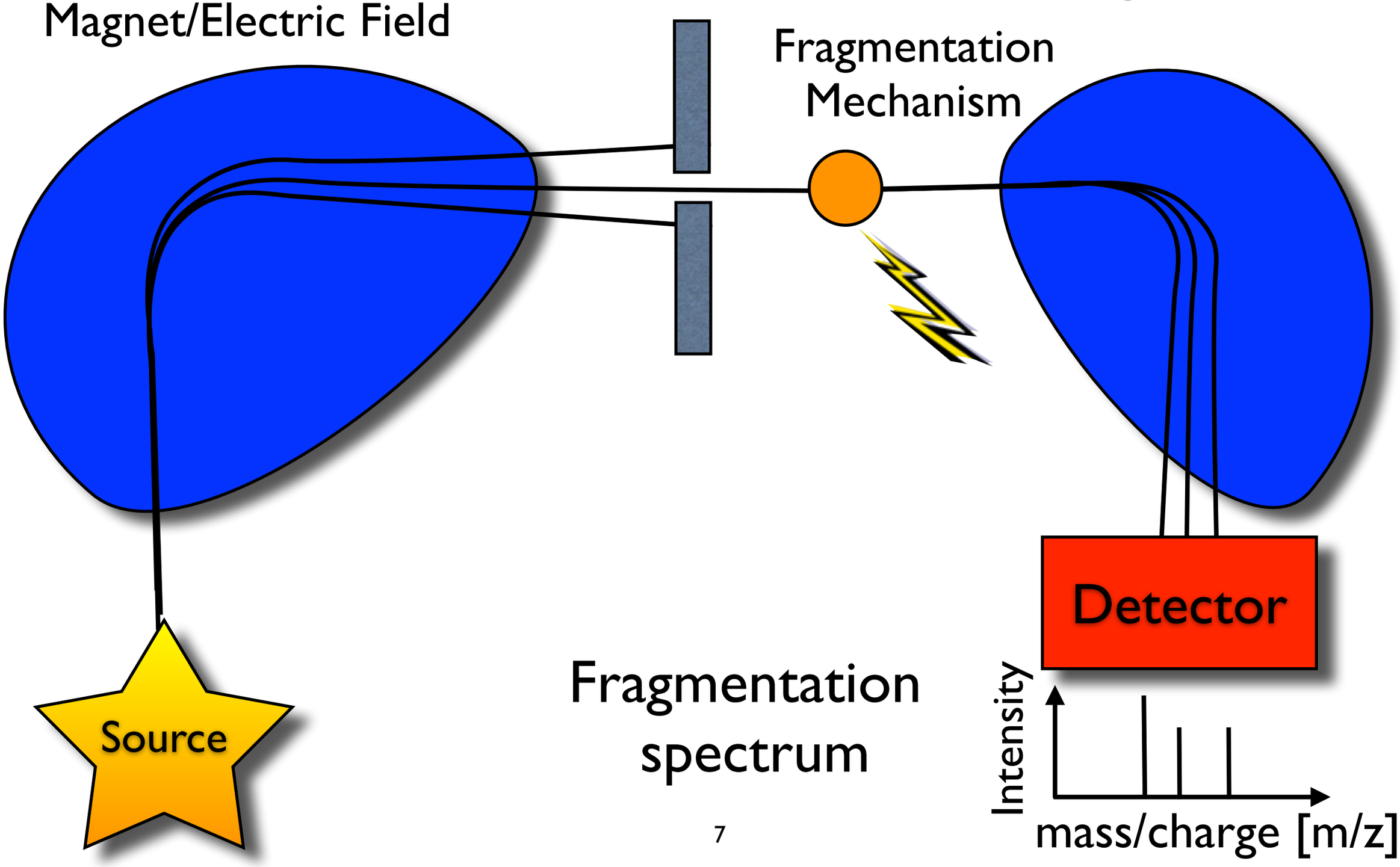


- Electrostatic attraction to the inner electrode is balanced by centrifugal forces.
- Detect the ions by their movement along the axis of the electrode.
- High mass accuracy (1–2 ppm)

http://www.youtube.com/watch?v=KjUQYuy3msA

# Tandem mass spectrometry

a.k.a MS/MS or $MS^2$




Fragmentation spectrum

# What do they look like?

Time of Flight

Orbitrap

# Why do we need proteomics?

- We already know the sequence of the genome!
- We know how to measure transcription!
- What else do we need to know?

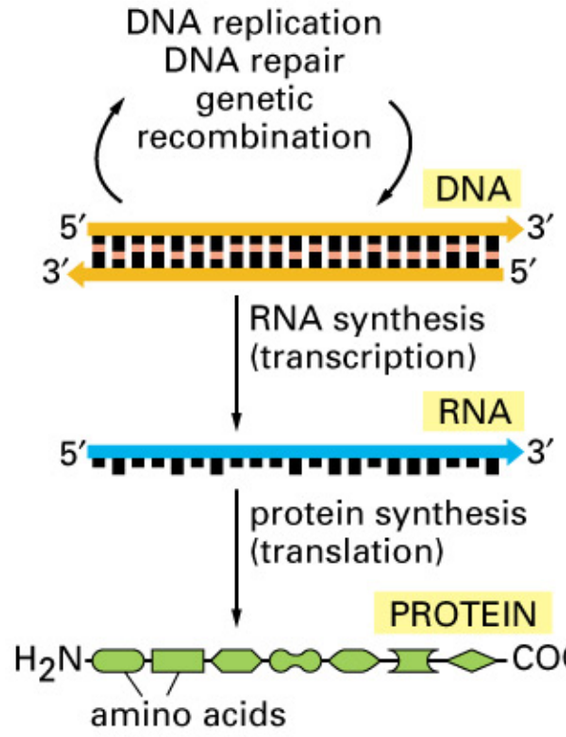
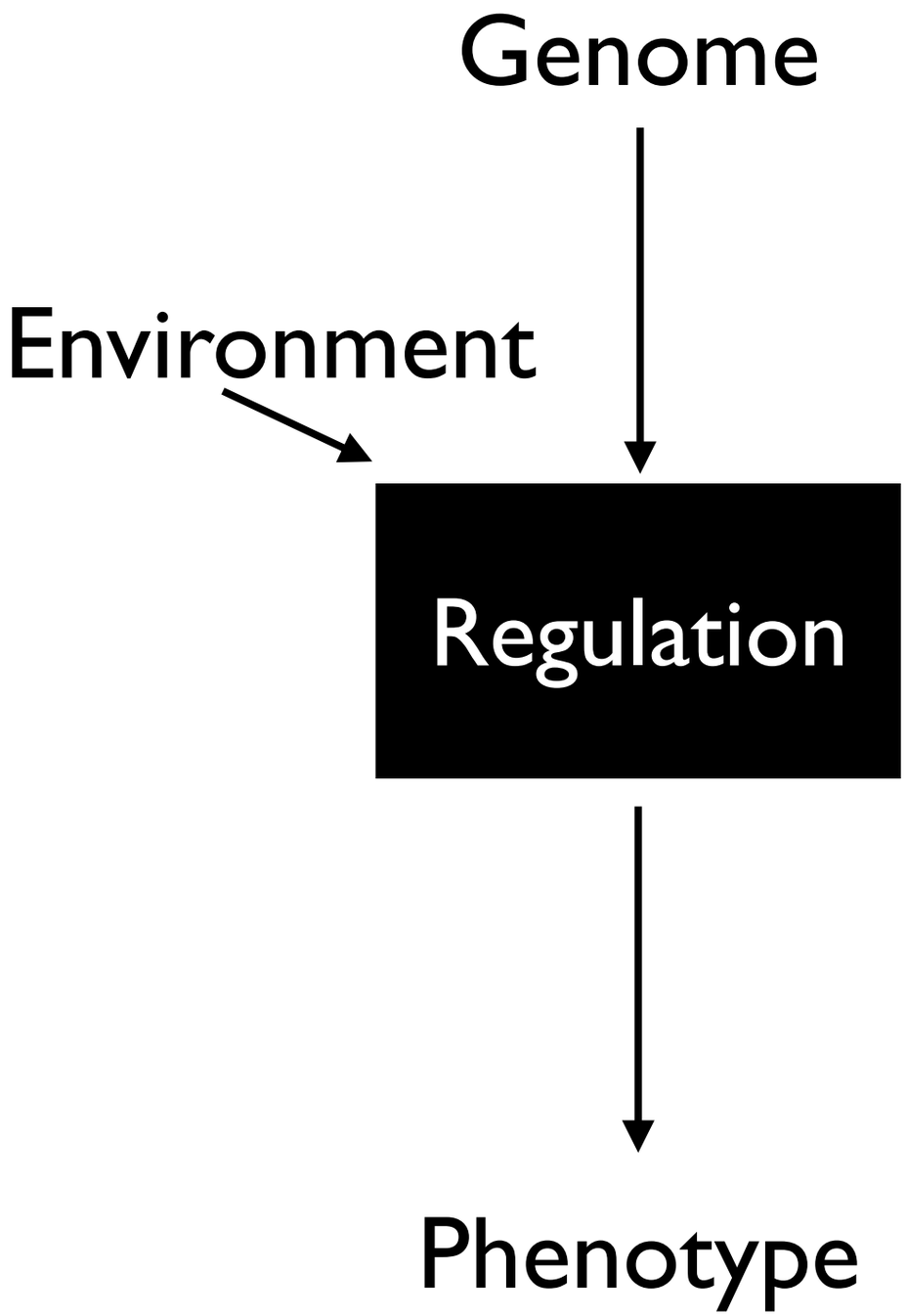


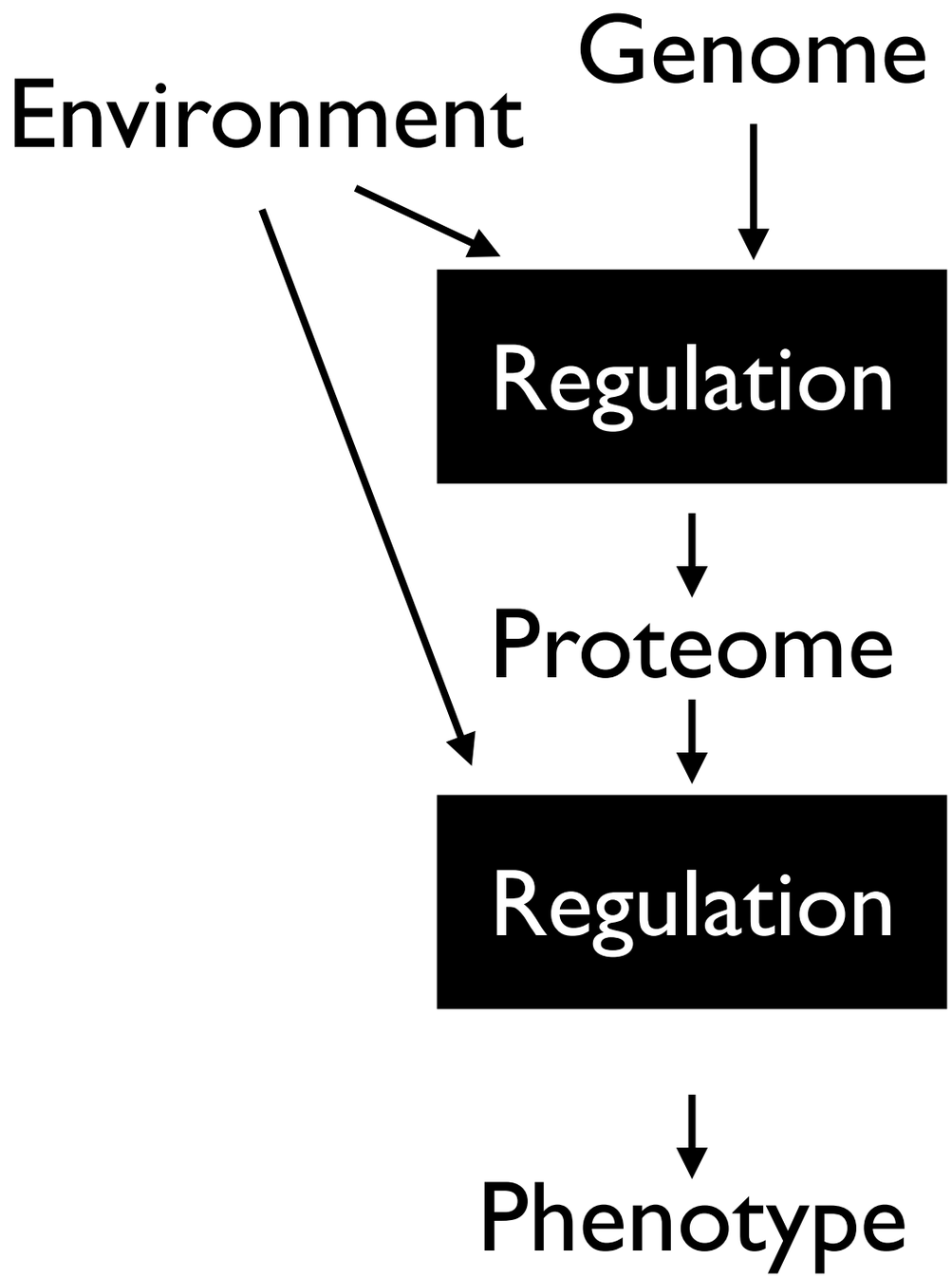

Figure 6–2. Molecular Biology of the Cell, 4th Edition.

# Same DNA, different configuration of proteins


https://youtu.be/jEtaqmW3ZK4

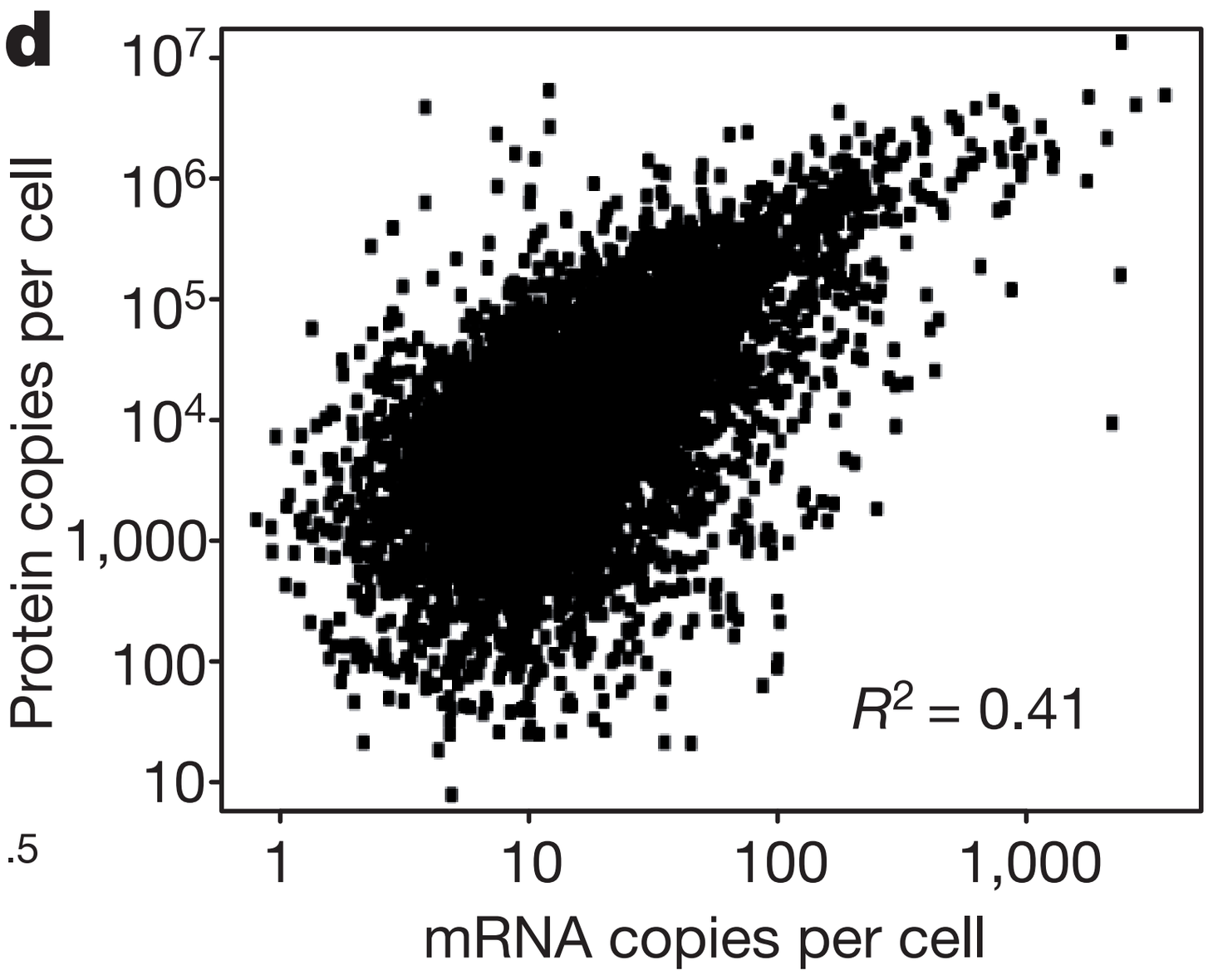
# Proteins are closer to the Phenotype



# mRNA levels correlate only weakly with protein levels

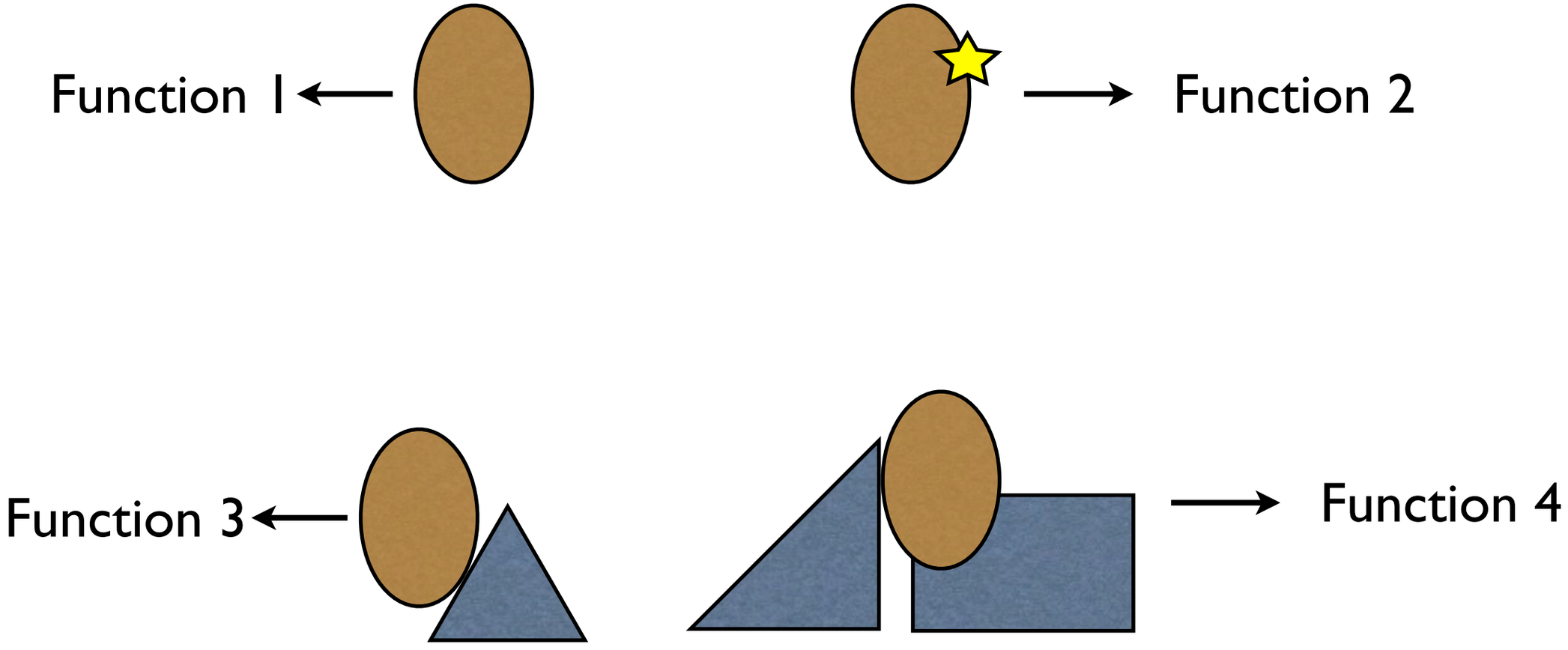
[Schwanhäusser *et al.* Nature 2011]



# One protein - Many functions

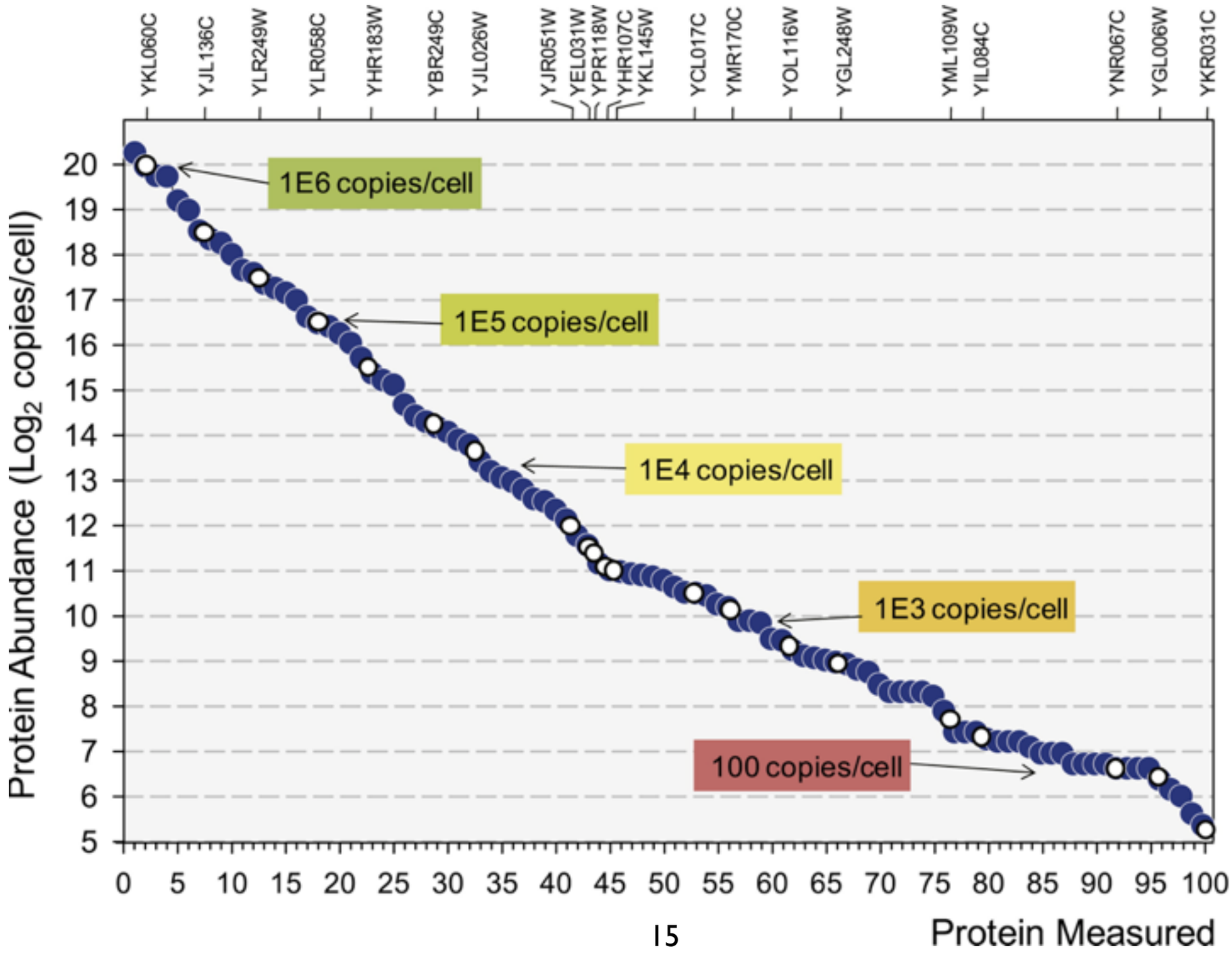
Proteins undergo post-translational modifications

Proteins may have different functions in different complexes or compartments



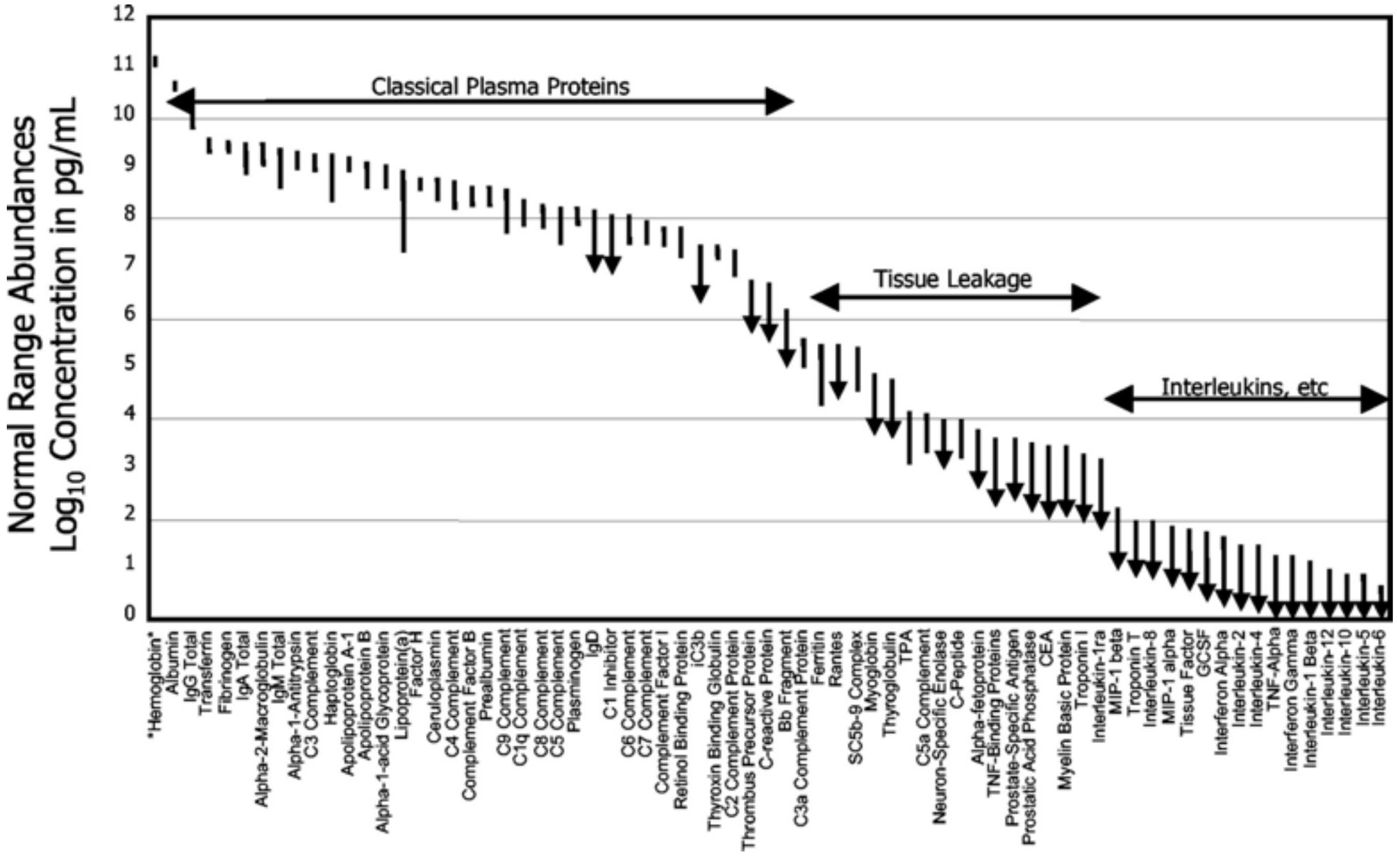
# Why don’t everybody do proteomics?

# Proteins concentration in yeast range >4 orders of magnitude




[Picotti et al Cell 2009]

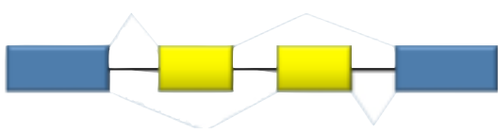
# Protein concentration in blood plasma range >10 orders of magnitude




[Andersson & Anderson, MCP 2002]

Difference between earth diameter and 1 mm is about 10 orders of magnitude

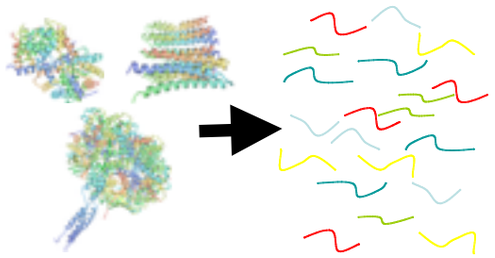
# How to define a protein?

| *Definition* | | *Occurrence in Human* |
|---|---|---|
| Protein coding ORF | | 21,257 [EnsembI] |
|  | Splice variant | 148,792 [Ensembl] |
| PTMs | Protein species | > $10^6$ |
| Sequence rearrangements, mutations | Cell specific protein species | > $10^7$ |

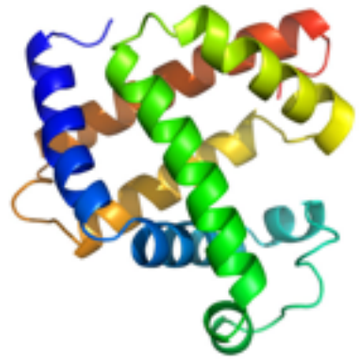
# Proteomic techniques

Mass spectrometry-based approaches

e.g.



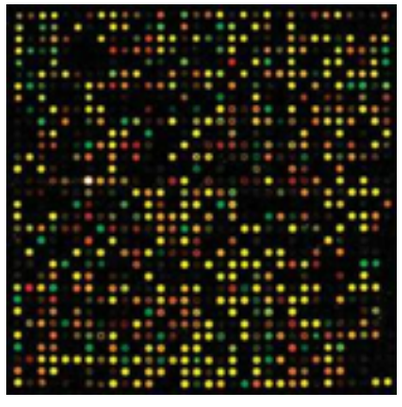
Shotgun proteomics



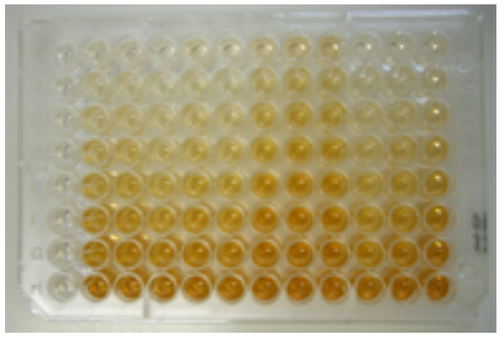
Top-down proteomics

Antibody-based approaches

e.g.

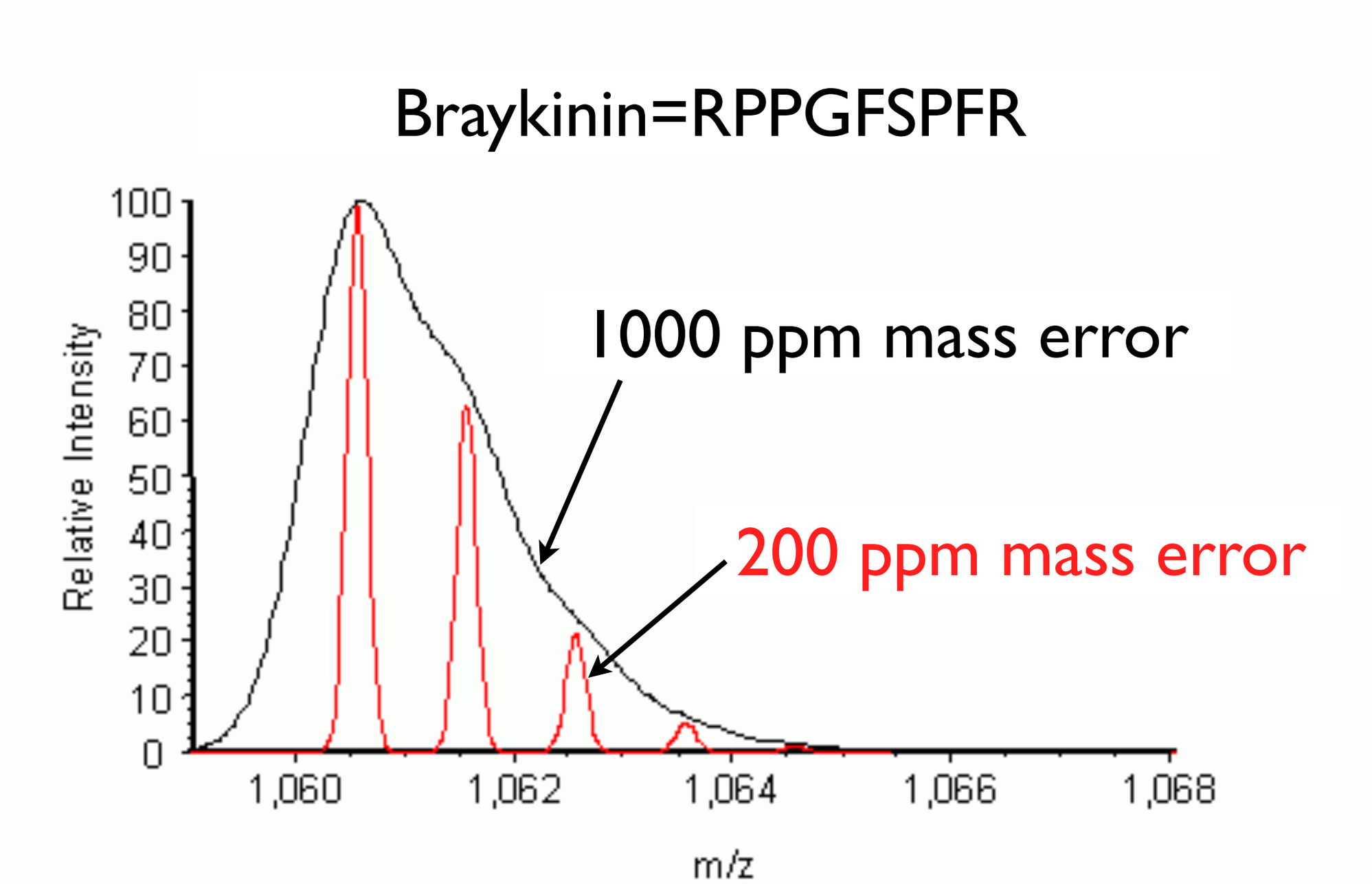


Protein Arrays

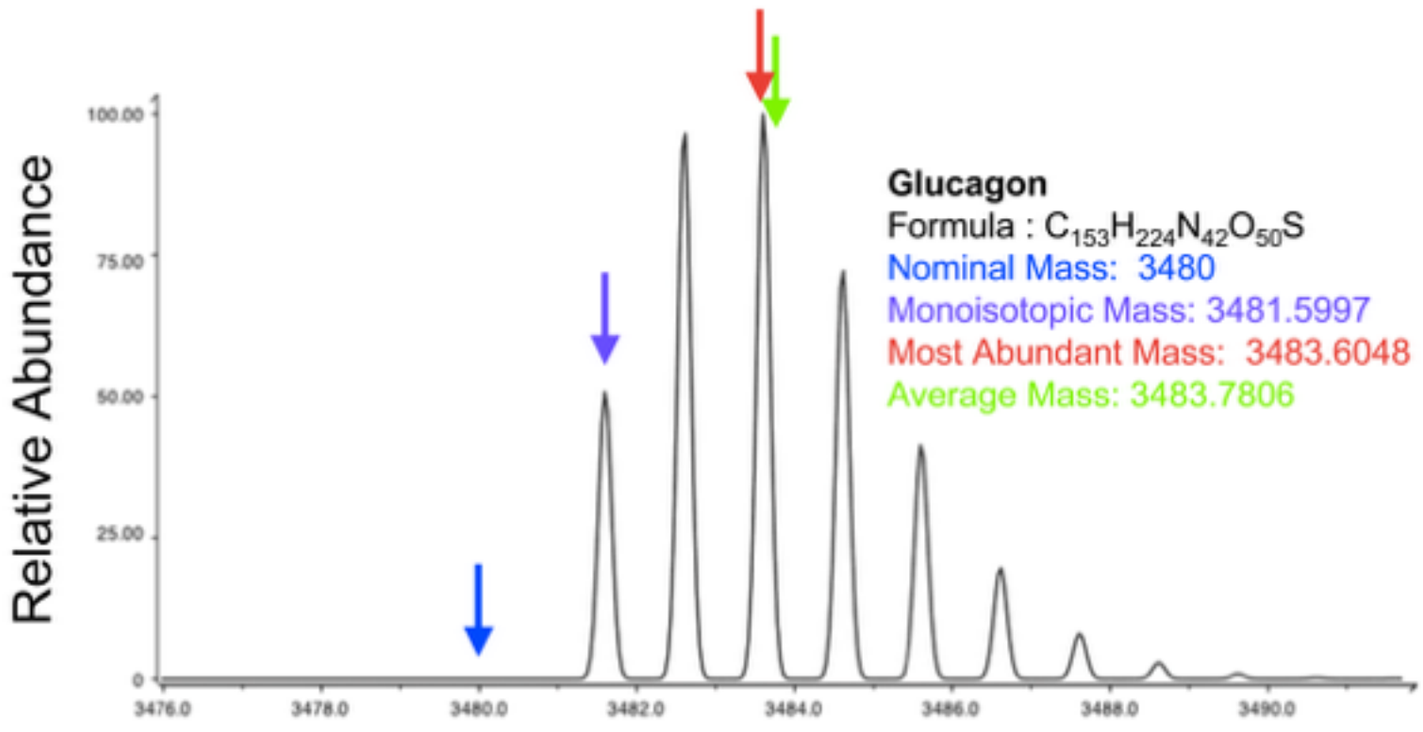


Enzyme-linked immunosorbent assay (ELISA)

# Resolving isotopes



# Different definitions of Masses




| Mass Type | The sum of the molecules' atoms': |
|---|---|
| Nominal Mass | Integer mass of the most abundant isotope |
| Monoisotopic Mass | Masses of unbound, ground-state, isotope<br>*The preferred measure for high-resolution MS* |
| Average Mass | Average Mass given the isotopes and their natural abundance<br>The preferred measure for low-resolution MS |

# Mass of Elements

| Element | Average mass | Monoisotopic Mass |
|---|---|---|
| C | 12.0107 | 12 |
| N | 14.00674 | 14.00307 |
| H | 1.00794 | 1.00782 |
| O | 15.9994 | 15.99492 |
| S | 32.066 | 31.97207 |
| Pepide ALLETYCATPAKSE, $C_{65}H_{105}N_{15}O_{23}S$ | 1496.682 | 1495.72281 |

**monoisotopic mass** is the mass of the principal (most abundant) isotope.

**average mass** is the average mass of all isotopes, normalized for natural abundance.

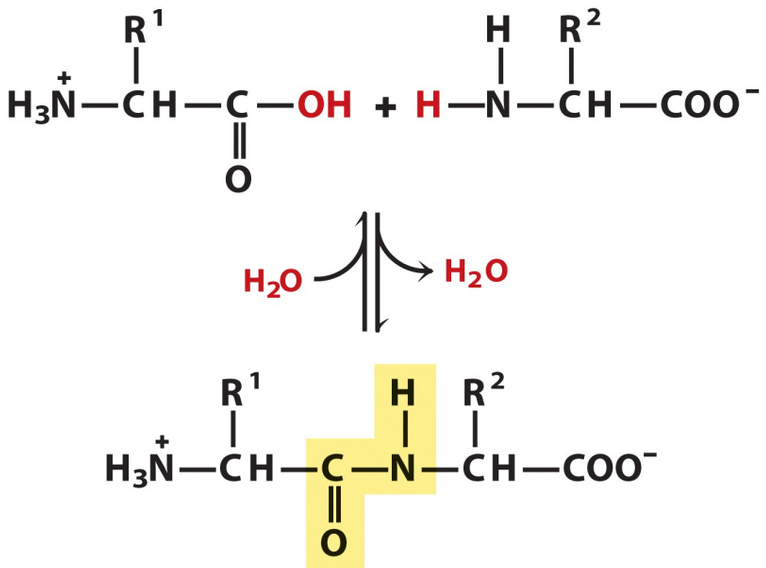
# Residue mass of amino acids

| Amino Acid | Short | Abbrev. | Formula | Mon. Mass§ (Da) | Avg. Mass (Da) |
|---|---|---|---|---|---|
| Alanine | A | Ala | $C_3H_5NO$ | 71.03711 | 71.0788 |
| Cysteine | C | Cys | $C_3H_5NOS$ | 103.00919 | 103.1388 |
| Aspartic acid | D | Asp | $C_4H_5NO_3$ | 115.02694 | 115.0886 |
| Glutamic acid | E | Glu | $C_5H_7NO_3$ | 129.04259 | 129.1155 |
| Phenylalanine | F | Phe | $C_9H_9NO$ | 147.06841 | 147.1766 |
| Glycine | G | Gly | $C_2H_3NO$ | 57.02146 | 57.0519 |
| Histidine | H | His | $C_6H_7N_3O$ | 137.05891 | 137.1411 |
| Isoleucine | I | Ile | $C_6H_{11}NO$ | 113.08406 | 113.1594 |
| Lysine | K | Lys | $C_6H_{12}N_2O$ | 128.09496 | 128.1741 |
| Leucine | L | Leu | $C_6H_{11}NO$ | 113.08406 | 113.1594 |
| Methionine | M | Met | $C_5H_9NOS$ | 131.04049 | 131.1986 |
| Asparagine | N | Asn | $C_4H_6N_2O_2$ | 114.04293 | 114.1039 |
| Pyrrolysine | O | Pyl | $C_{12}H_{21}N_3O_2$ | 255.15829 | 255.3170 |

## etc...

The free form of the amino acids are a the equivalent of a water molecule heavier (~18 Da) than its residue mass

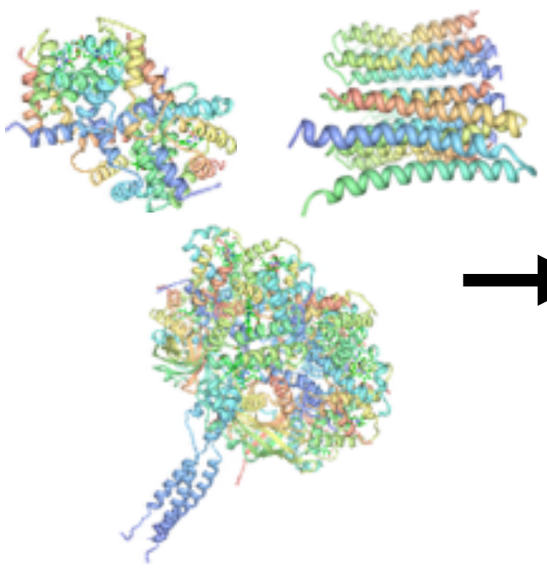
# The mass of a peptide

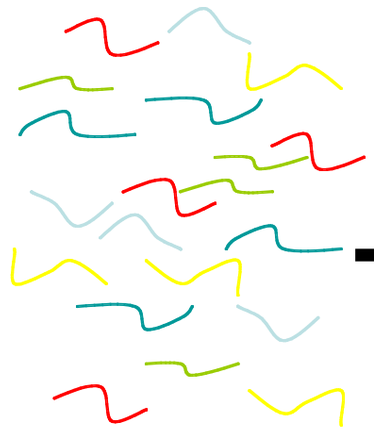


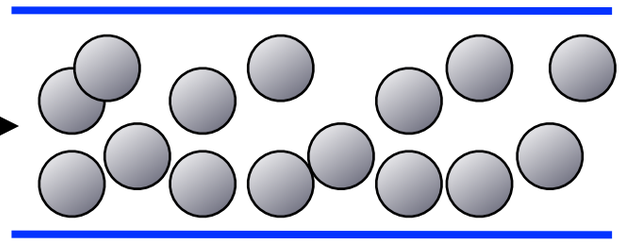

The mass $m(p)$ of peptide $p$ can be calculated as the residue mass of its constituent amino acids, $a_1 \ldots a_n$, and the mass of a water molecule

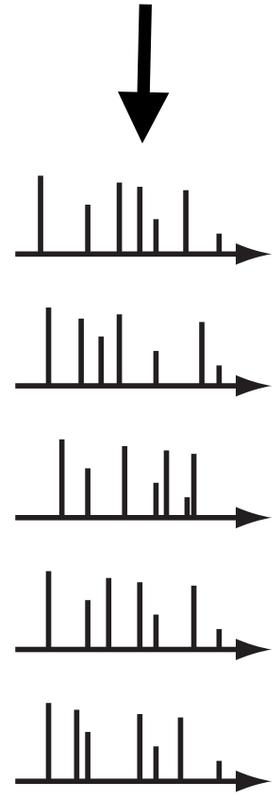
$$m(p) = m(H_2O) + \sum_{i=1 \ldots n} m(a_i)$$

# Shotgun proteomics





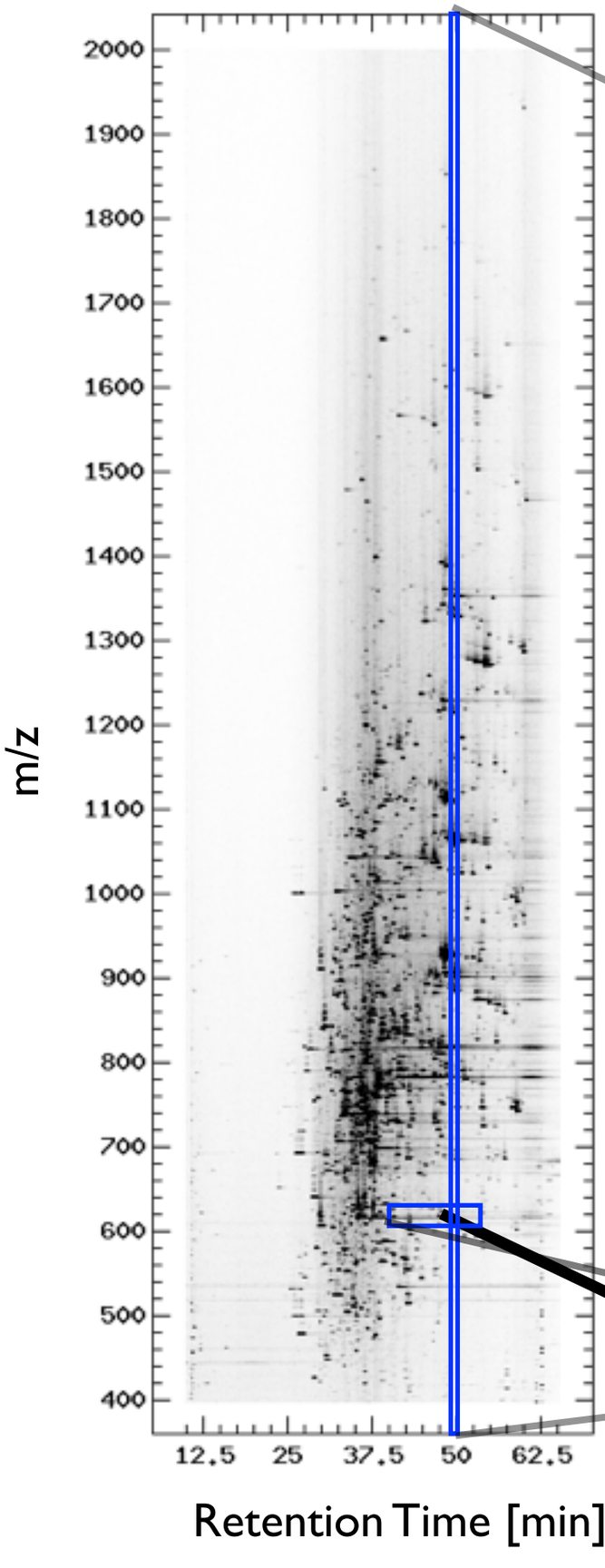


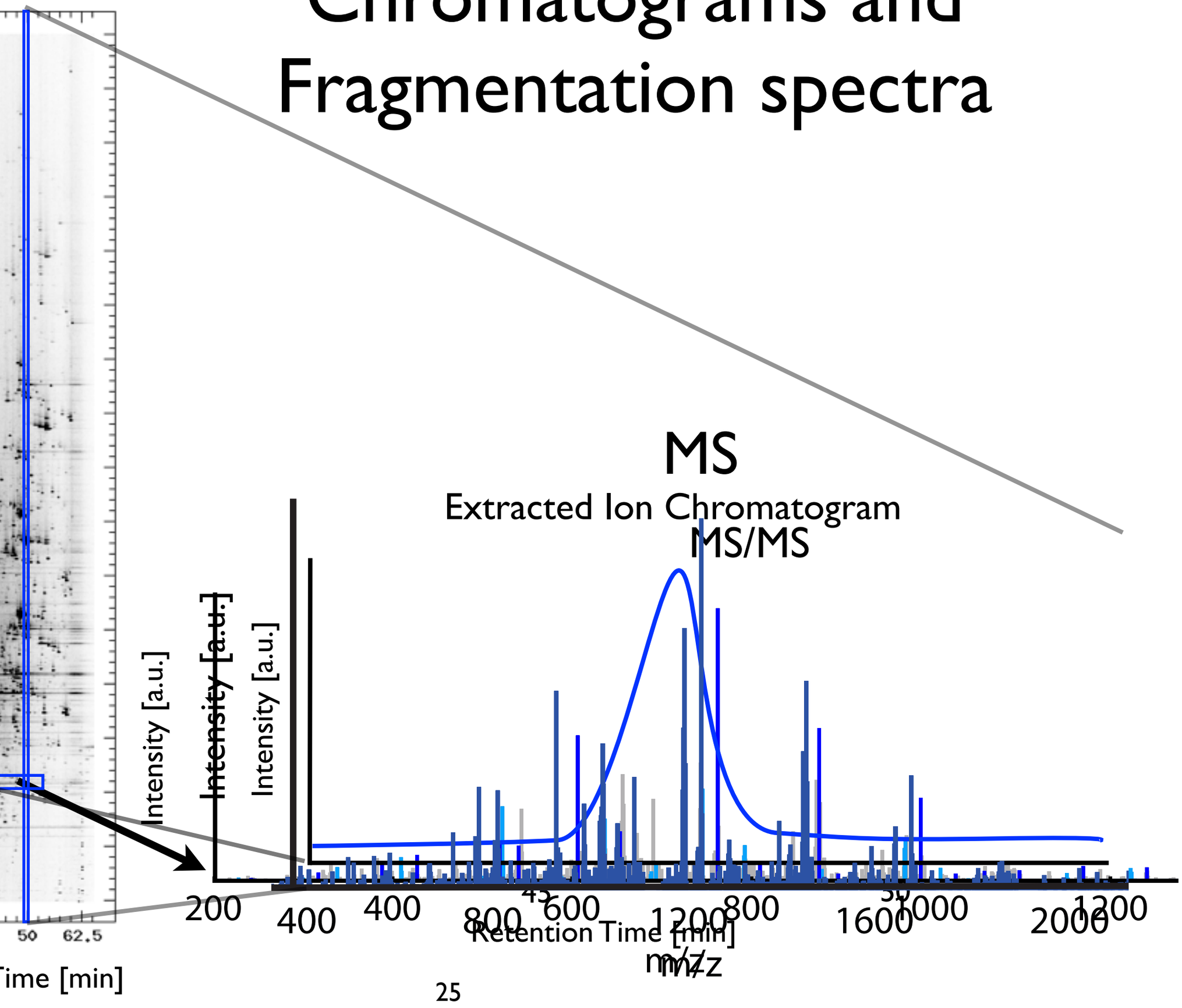


IAMAPEPTIDER
MGREATMATCHK
ATRYIDENTIFYMEK
LMAKEMYDAYR

GLB4_LUMTE
COX1_LUMTE
ATP6_LUMTE

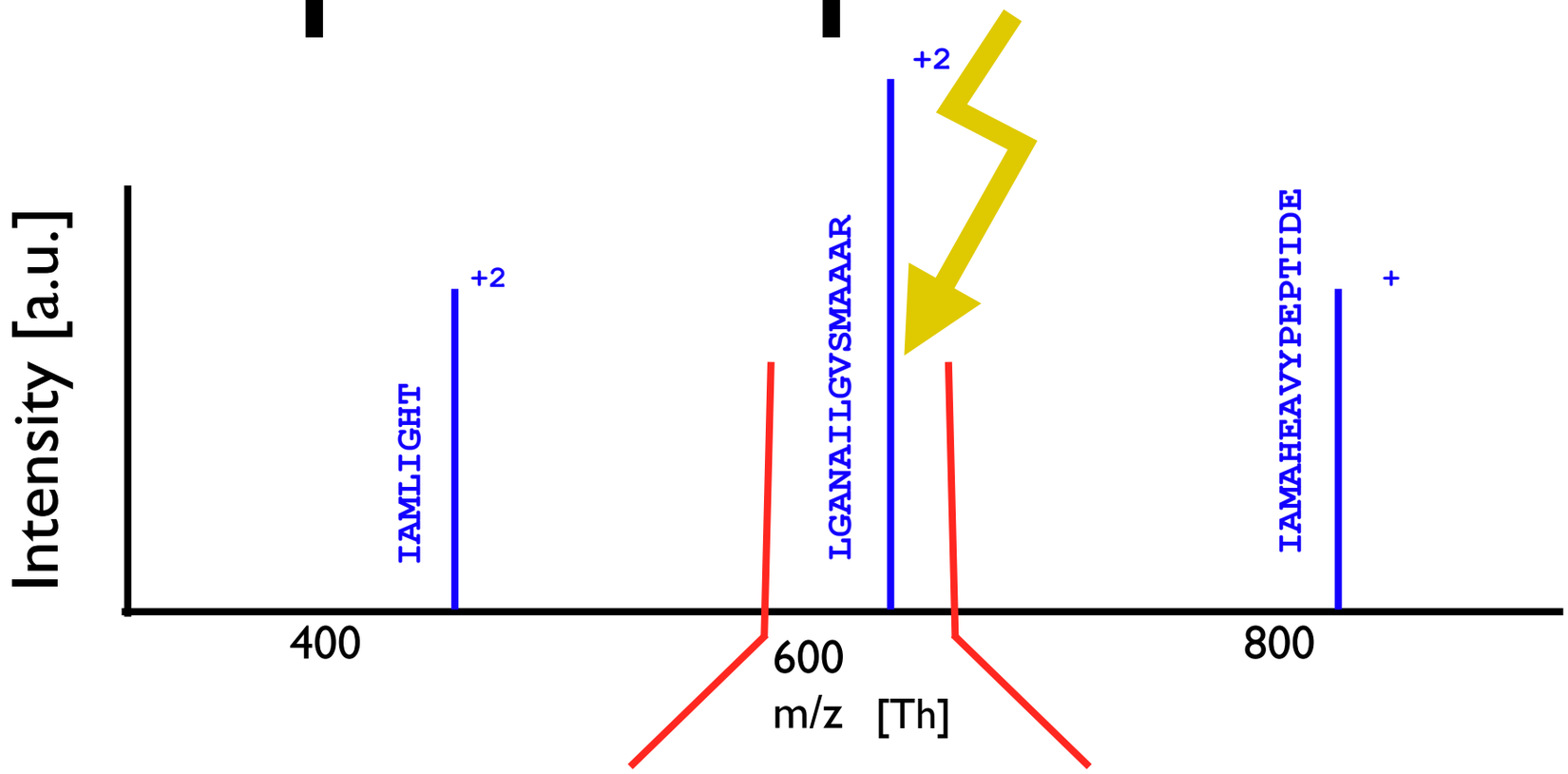
# Chromatograms and Fragmentation spectra



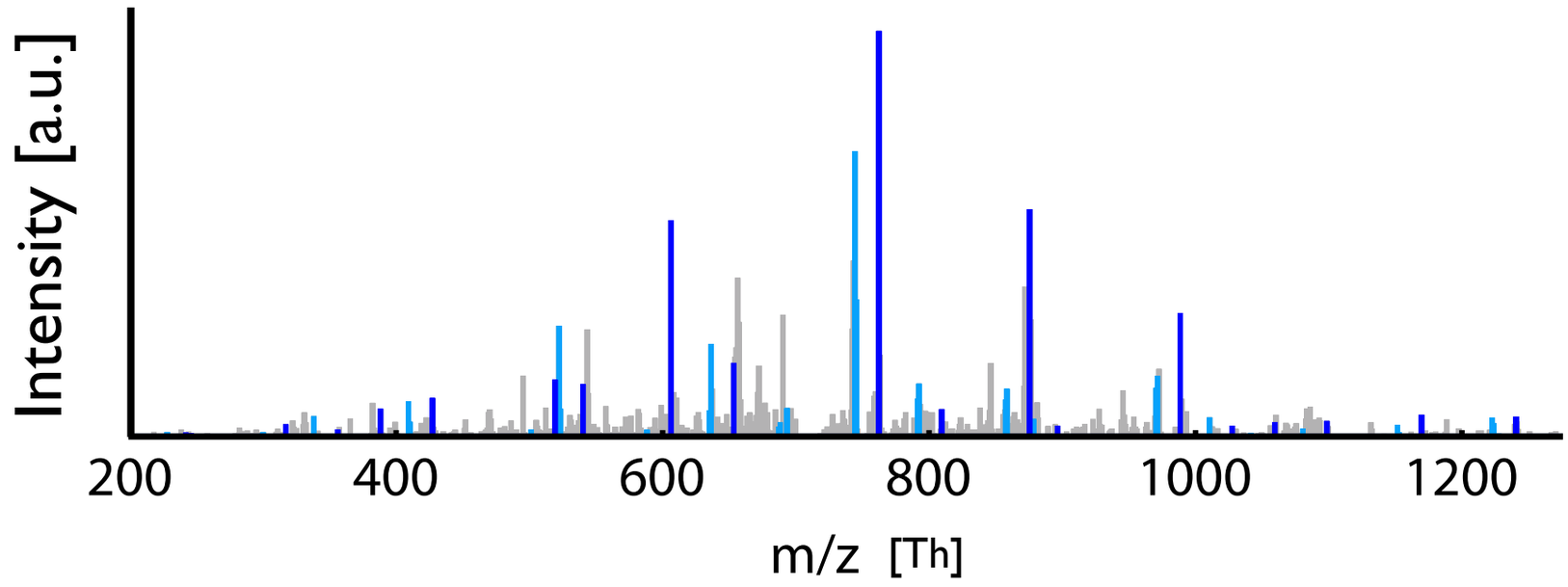



# Peptide spectra

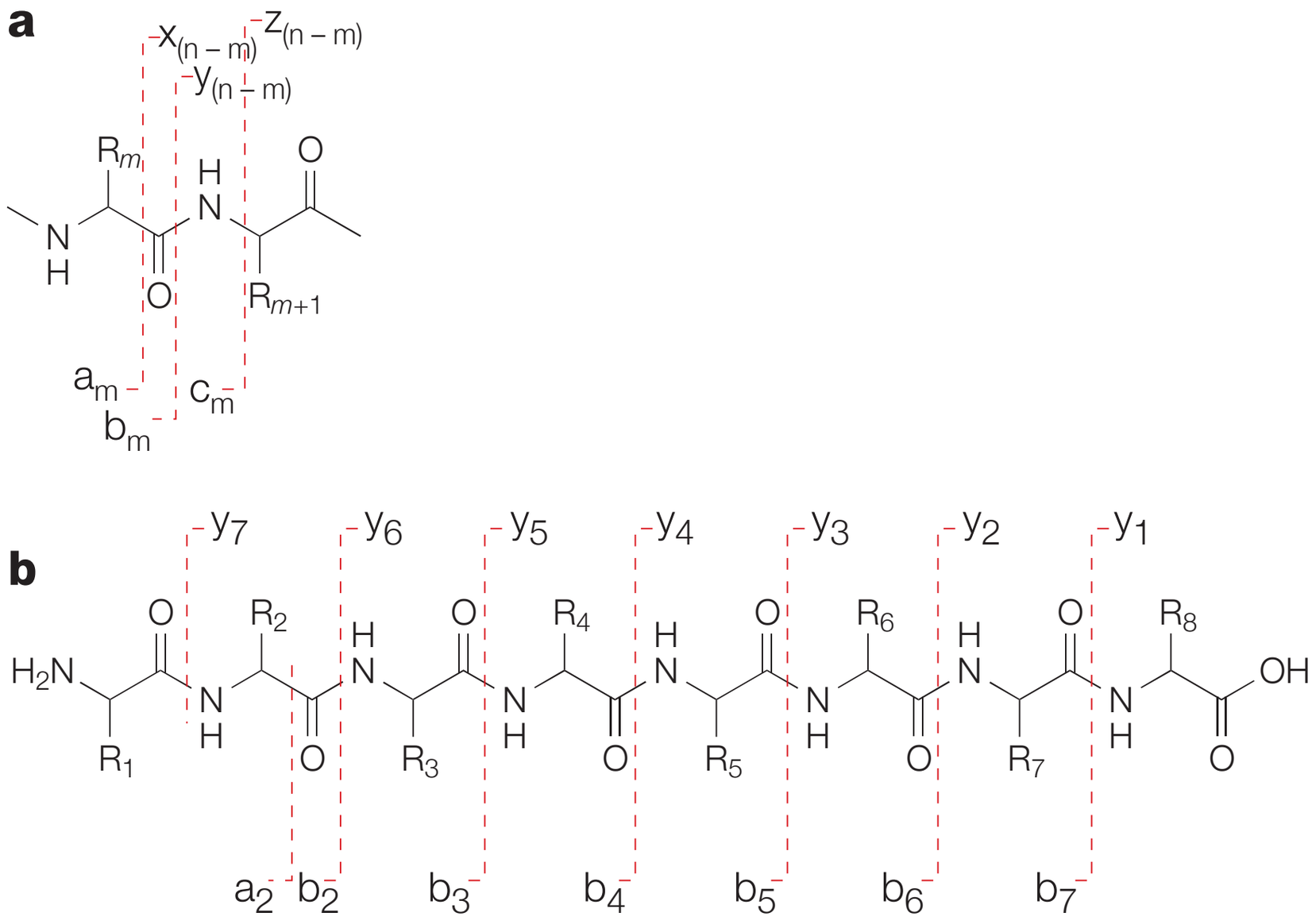
MS¹




MS²



# Peptide Fragmentation

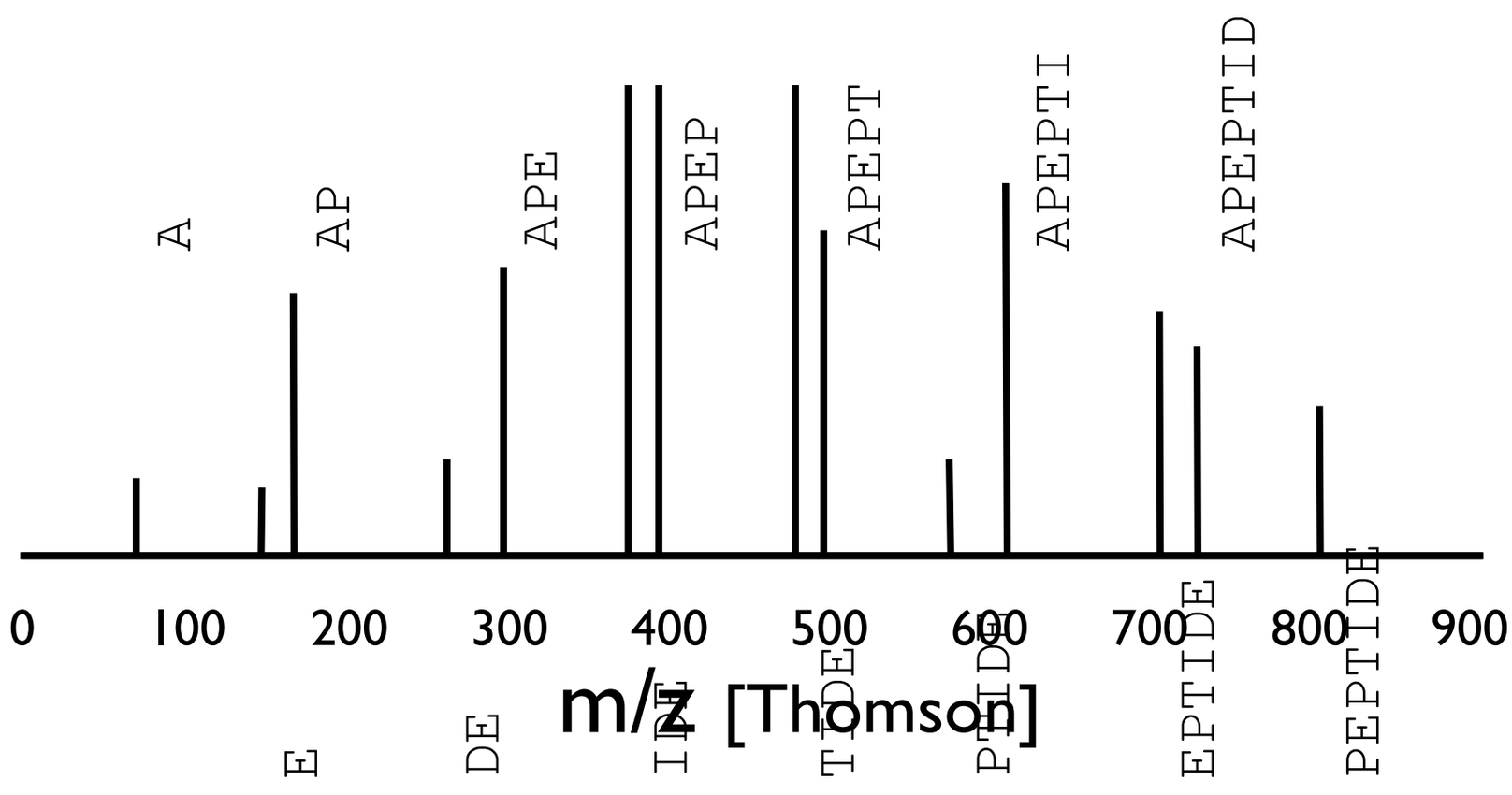


# Fragmentation Spectrum

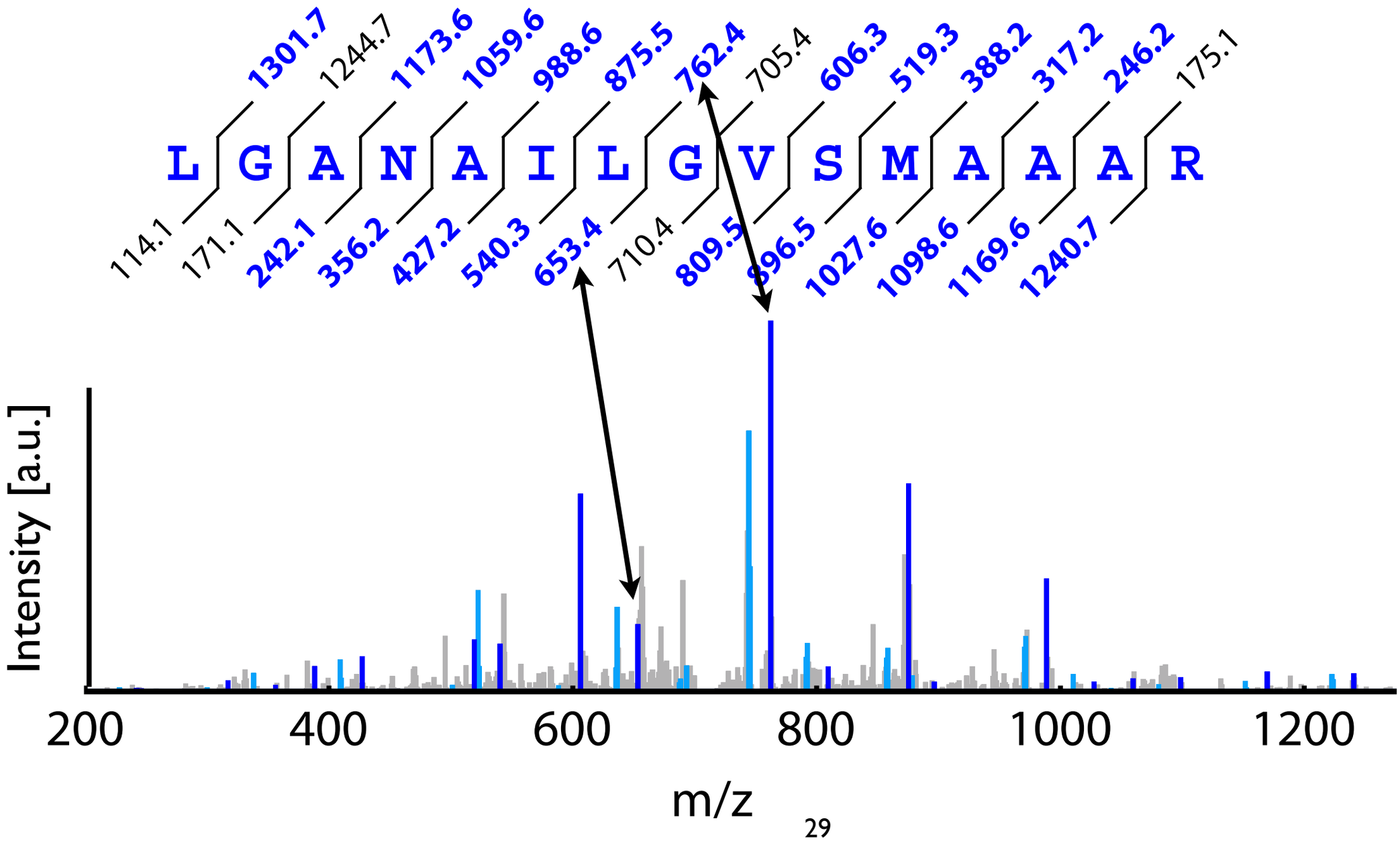
A|P|E|P|T|I|D|E

b:

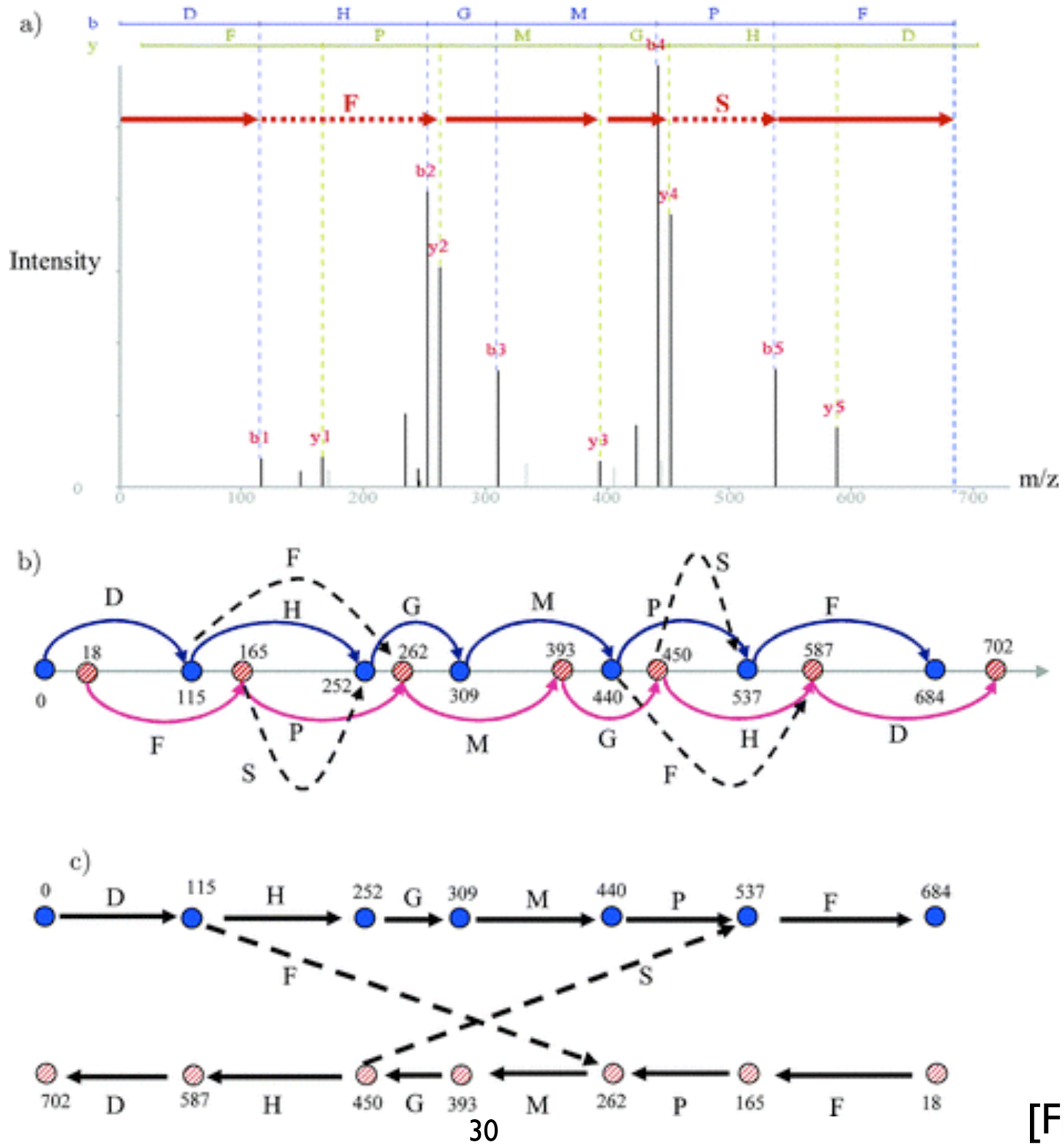
y:



# Peptide fragmentation spectrum

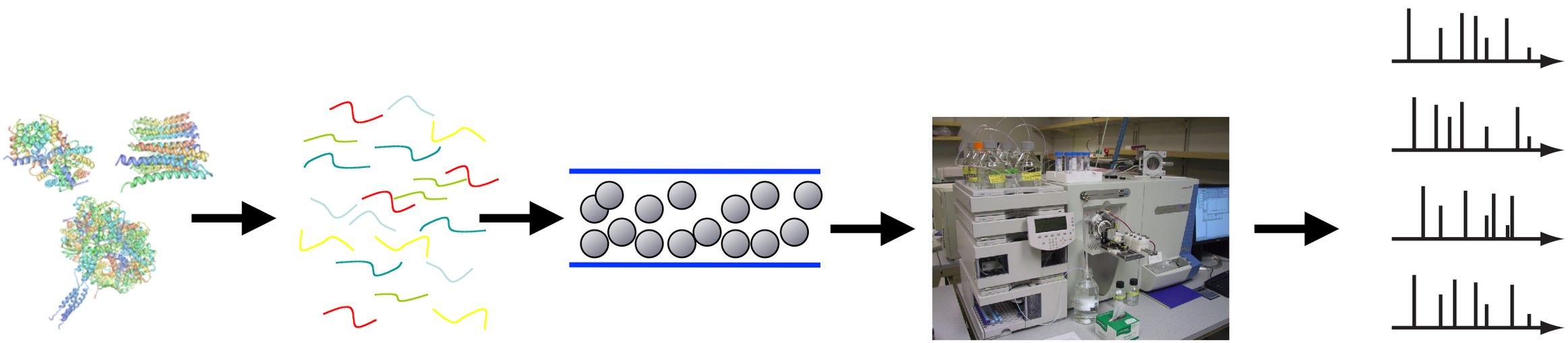


# *de Novo* sequencing

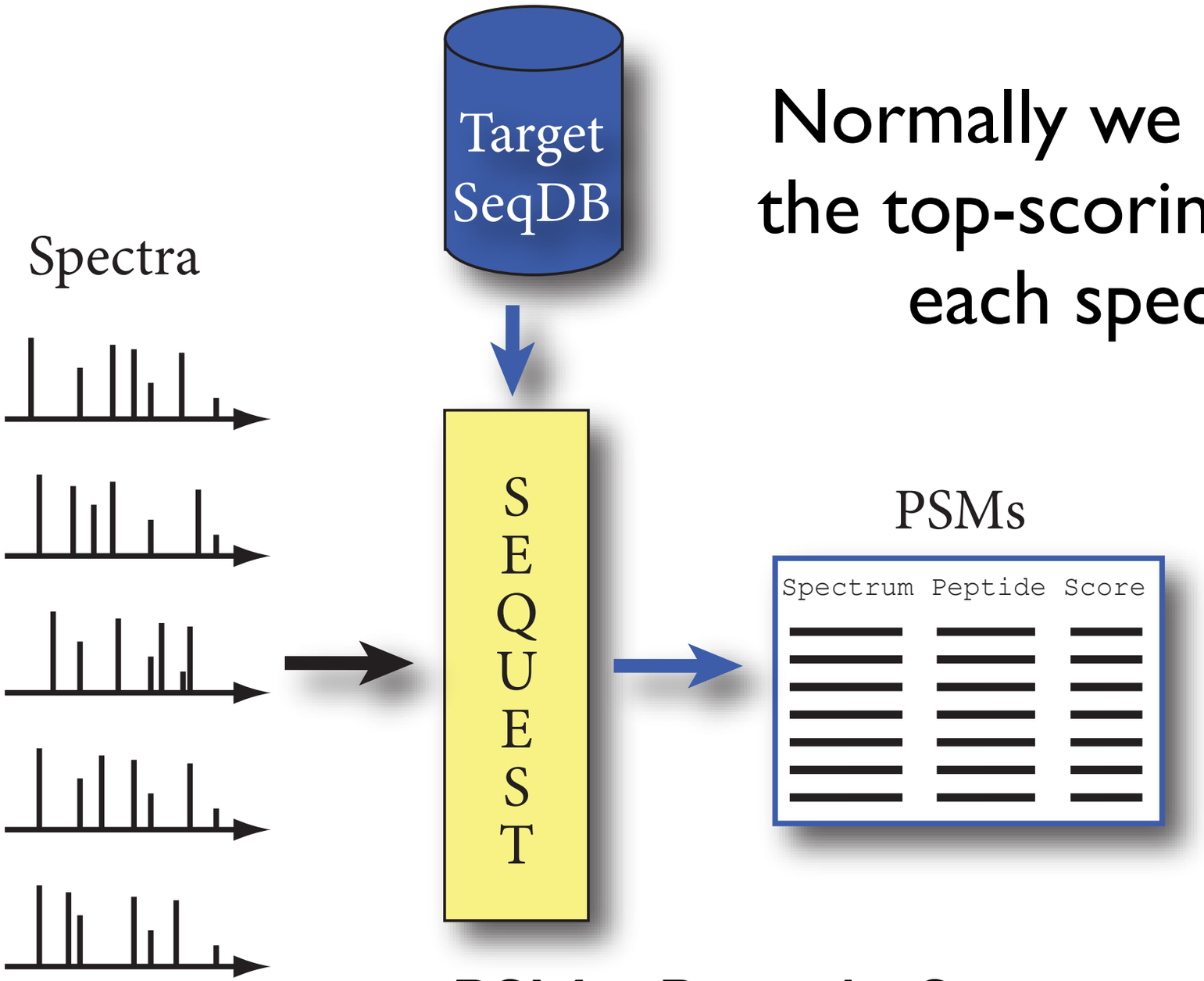



[Frank *et al.* JPR 2007]

# Peptide Identification



# Peptide identification




Normally we keep only the top-scoring PSM for each spectrum

PSM - Peptide Spectrum Match

# Four popular search engines

- SEQUEST (Scripps, Thermo Fisher Scientific)
  http://fields.scripps.edu/sequest
- MASCOT (Matrix Science)
  http://www.matrixscience.com
- X! Tandem (The Global Proteome Machine Organization)
  http://www.thegpm.org/TANDEM
- MS-GFDB
  http://proteomics.ucsd.edu/Software/MSGFDB.html

# Sequest

- First published automated spectral search engine
- Published but patented algorithm [Eng *et al.* JASMS 1994]
- For each spectrum *x* the top 500 candidate peptides are selected by a fast calculated preliminary score *Sp*.
- The theoretical spectra *y* are calculated for each of these top candidates and a background normalized cross correlation score, *Xcorr*, is calculated

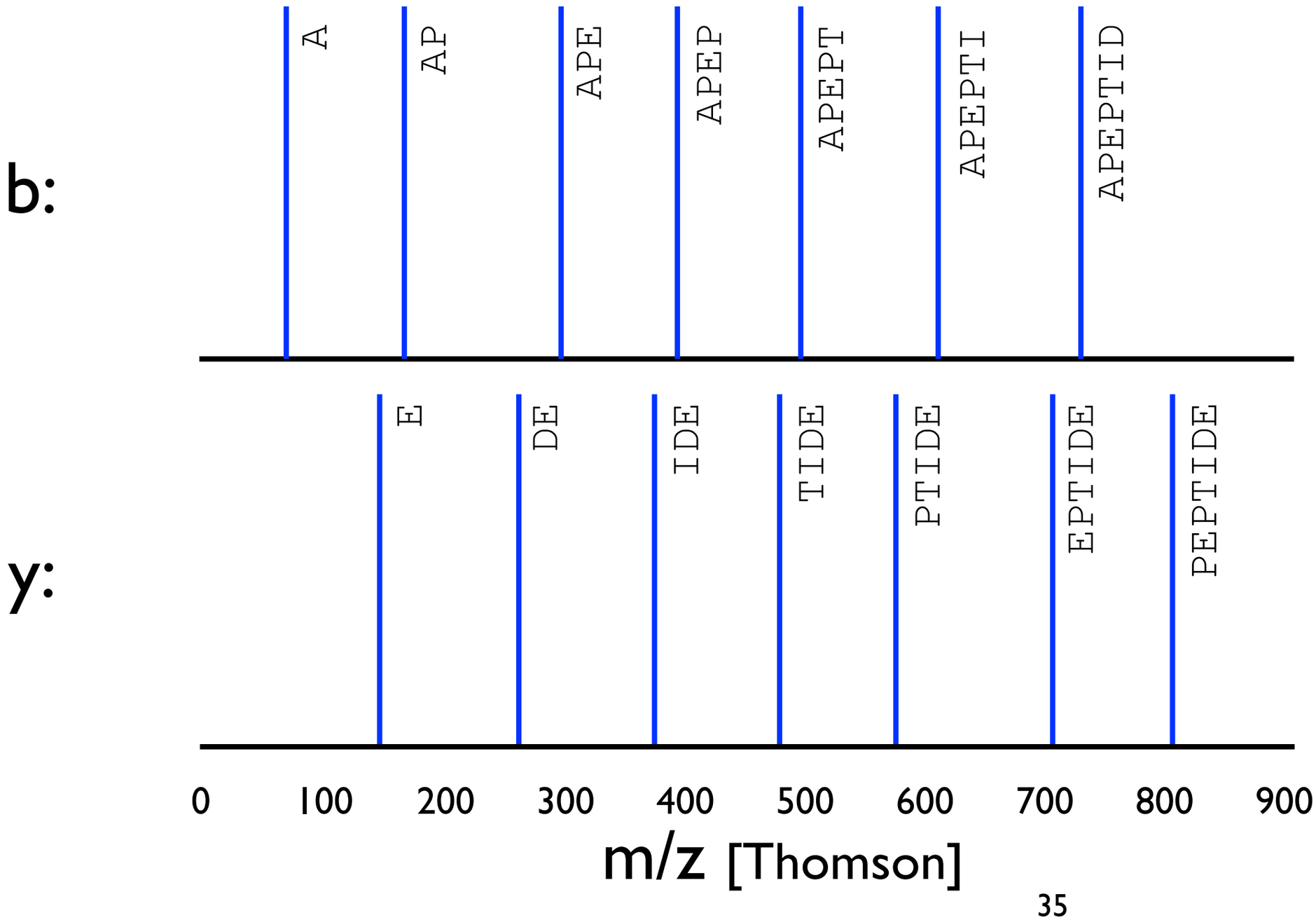
$$R_i = \sum_{j=1}^{n} x_j y_{j+i}$$

$$X = R_0 - \frac{1}{151} \sum_{i=-75}^{75} R_i$$

- A score *deltaCn* is provided which gives the relative difference between the first and second best *Xcorr*
- Re-implementations free for Academic users: Crux and Tide

# Theoretical Spectrum of a peptide

A|P|E|P|T|I|D|E



# Search engine

SEQUEST:

Observed Spectrum

Scoring Function

$$\text{matched\_peptide}(s,D) = \underset{p \in D}{\text{argmax}}\ f(s,T(p))$$

Database of peptides

peptide

Theoretical spectrum of $p$

other:

$$\text{matched\_peptide}(s,D) = \underset{p \in D}{\text{argmax}}\ f(s,p)$$

# Mascot

- Probably the most spread commercial spectral search engine
- Unpublished scoring function (Trade secret),
- Reports Rank, score and E-value for each PSM
- Predicts a *homology threshold* from database size and instrument accuracy which each PSM should pass
- Provides a fancy web report

# X! Tandem

- Open source, published algoritm
  [Craig & Beavis Rapid Commun. Mass Spectrom 2003]
- Scoring function, HyperScore, is build around the hypergeometric distribution (of number of matched b- and y-ions)
- Provides hyperscore and E-value for each PSM
- Relatively fast

Intensity

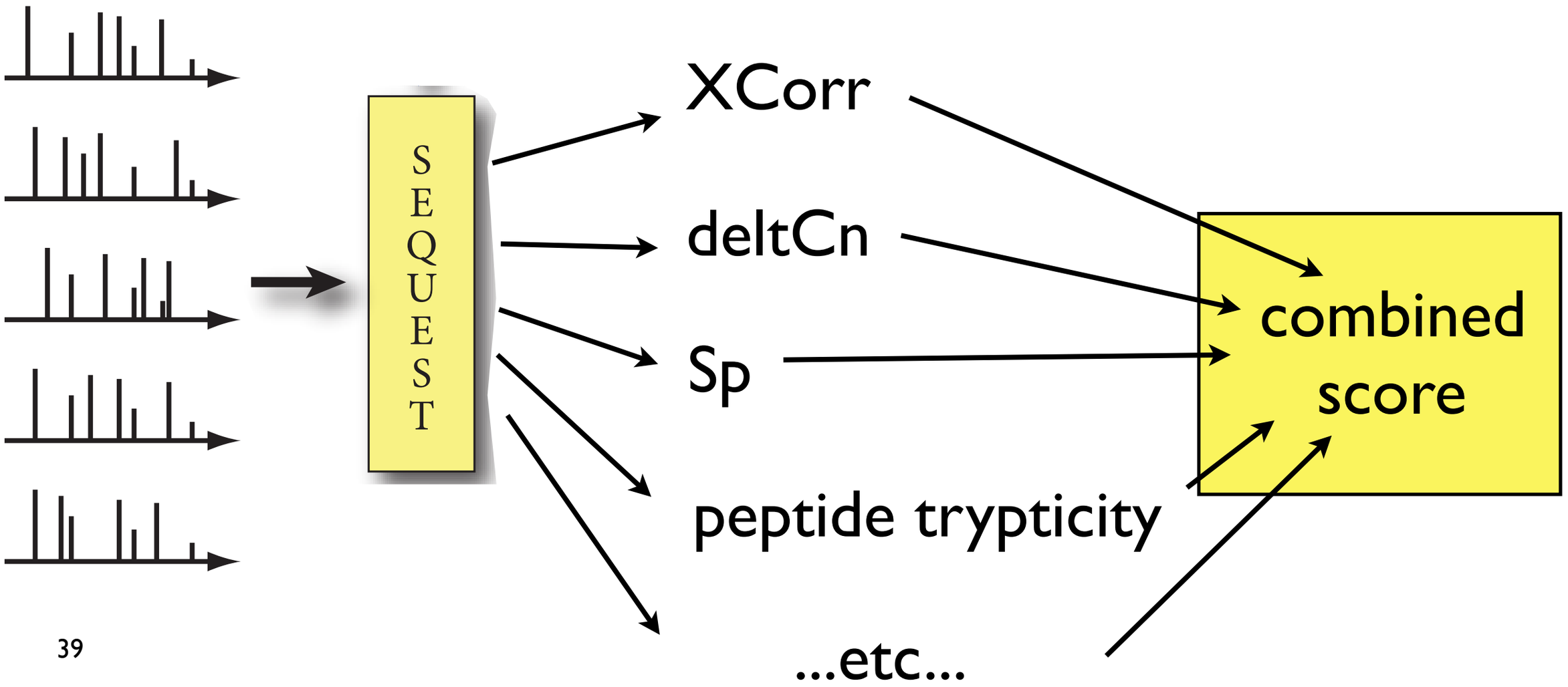
Present 0/1

$$HyperScore = \left( \sum_{i=0}^{n} I_i * P_i \right) * N_b! * N_y!$$

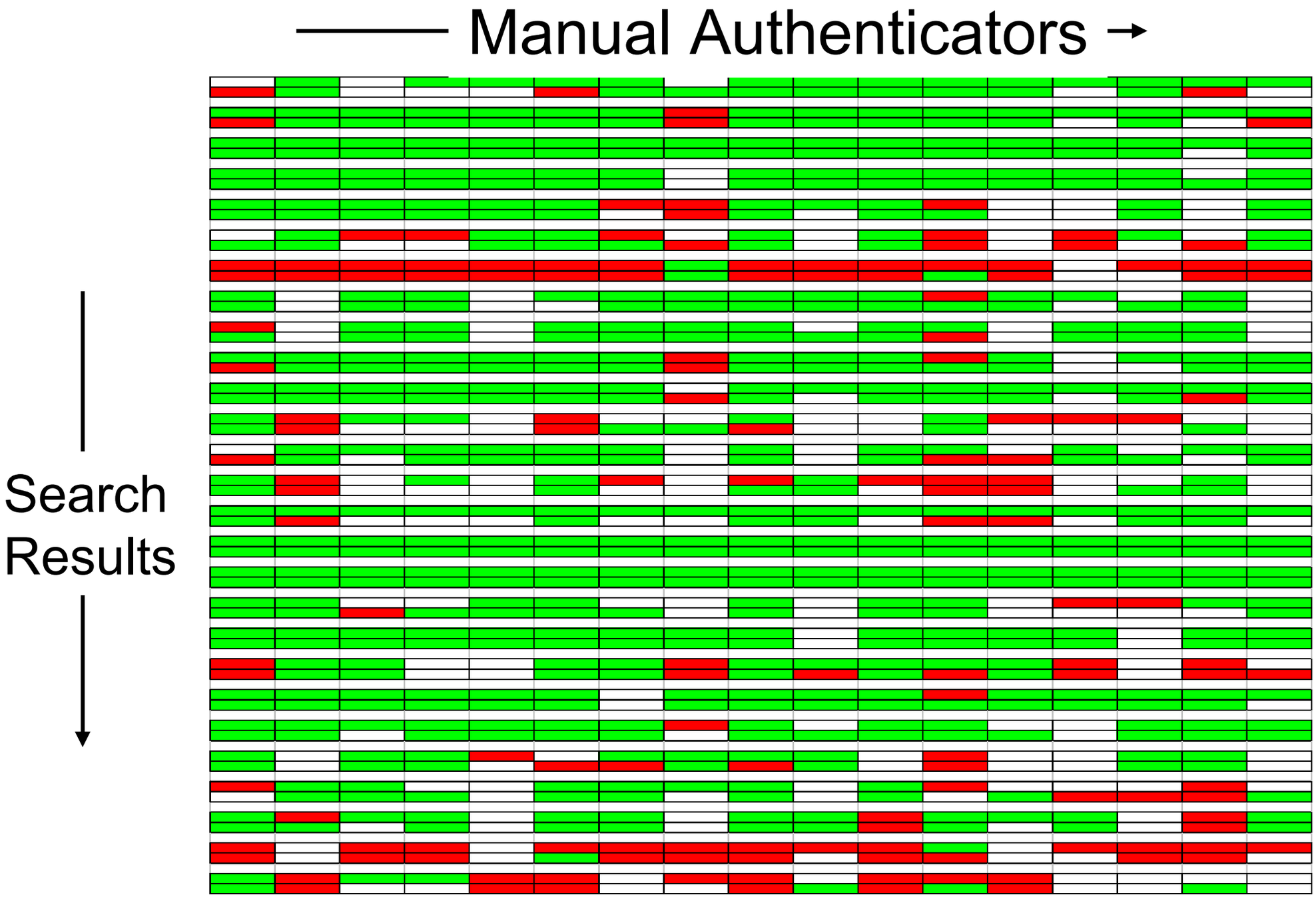
# Post Processors

Combinations of scores are known to give better yield than individual scores. Two examples are:

- PeptideProphet (LDA) [Keller *et al.* 2002 Anal Chem]
- Percolator (semi-supervised SVM) [Käll *et al.* Nat Methods 2007]



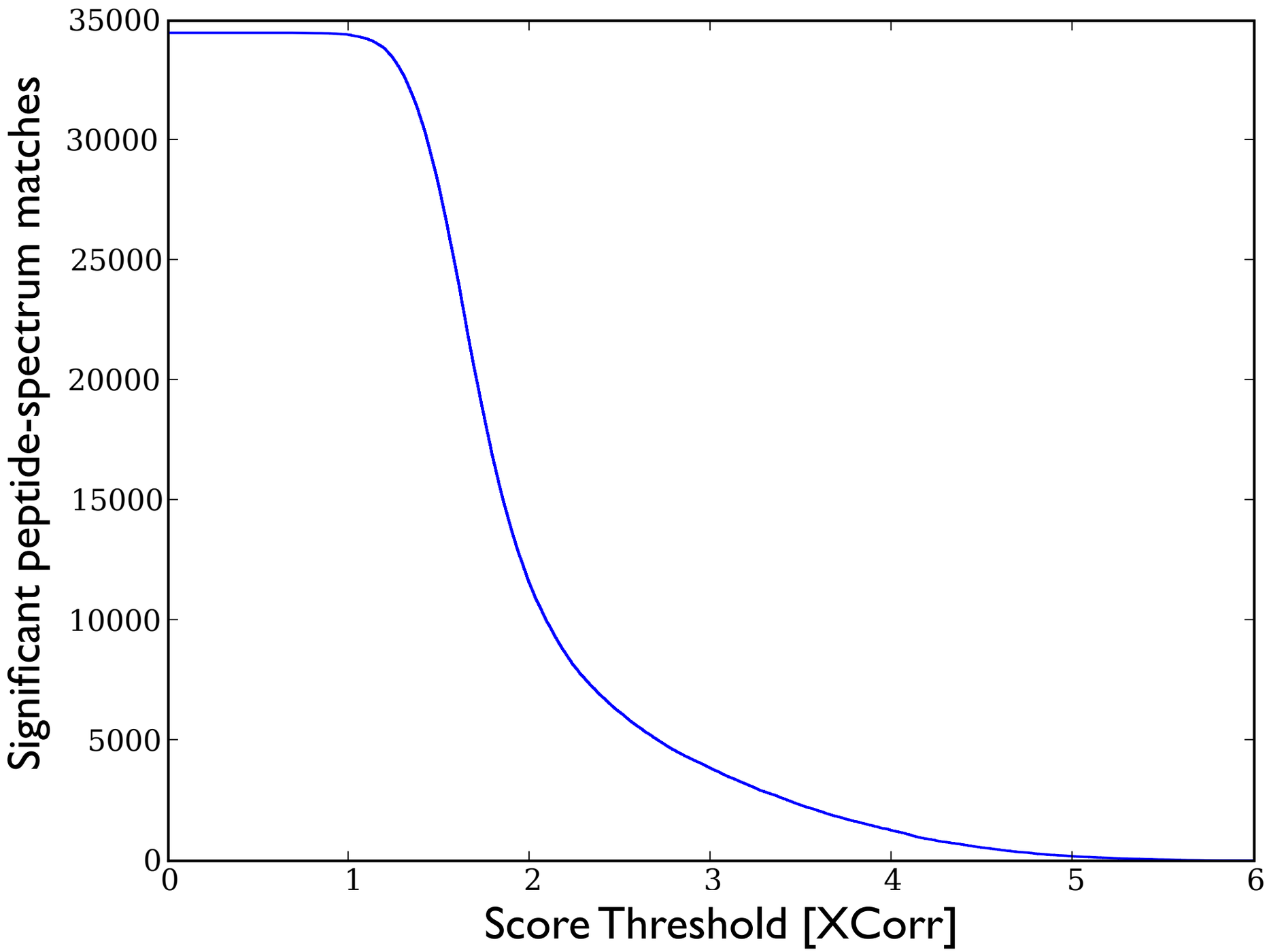
# (Un)reliability of manual validation






[www.systemsbiology.org]

# Score thresholds



# correct/incorrect target PSMs

| score | type |
|---|---|
| 7.5 | correct |
| 7.2 | correct |
| 6.9 | correct |
| 6.8 | correct |
| 6.7 | incorrect |
| 6.5 | correct |
| 6.4 | correct |
| 6.3 | incorrect |
| 6.1 | correct |
| 6 | incorrect |
| 5.9 | correct |
| 5.7 | incorrect |
| ... | ... |

threshold

$$\frac{2}{10}$$

*FDR*(x) is the expectation value of the fraction of PSMs above threshold *x* that are incorrect

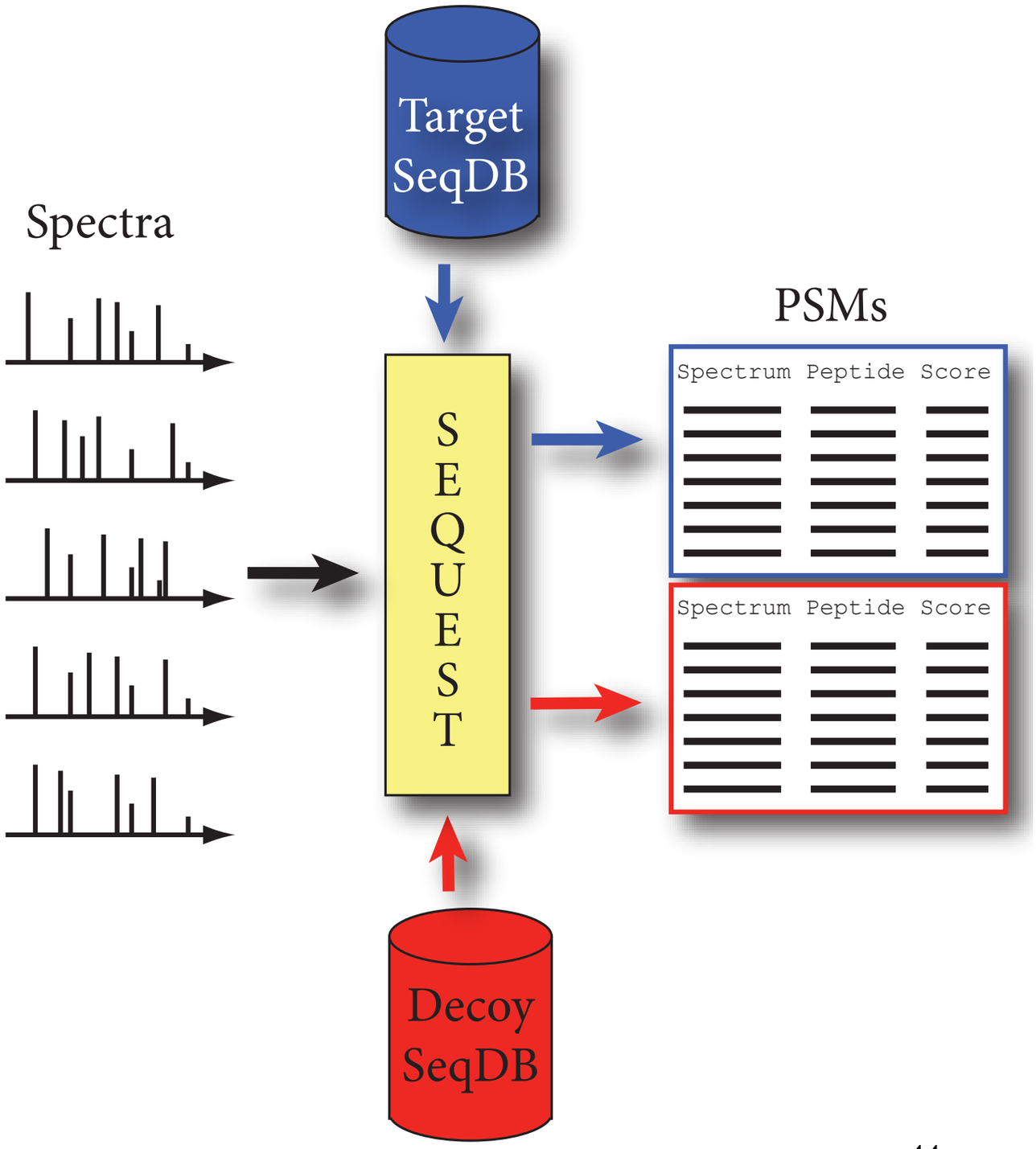
# control for ...

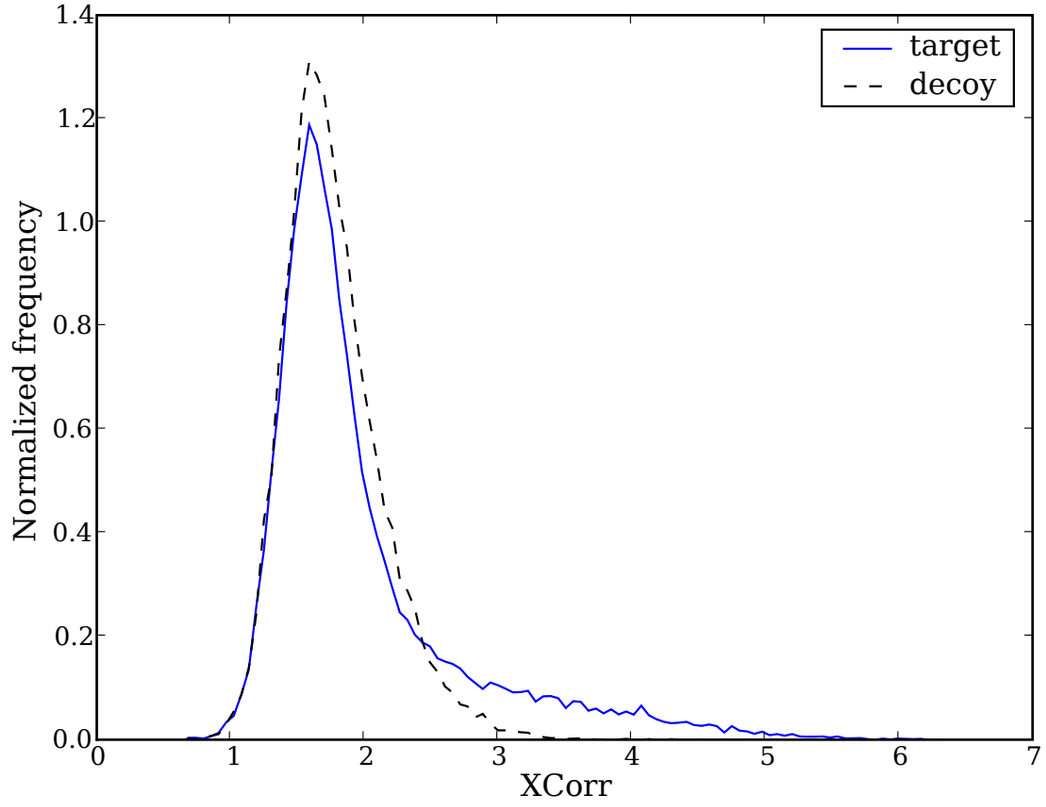
... FDR or q value when you are interested in identifying a set of PSMs

... PEP when you are interested in assessing the quality of a particular PSM.

... p or E value in an experiment rendering one single spectrum.

# Target-decoy analysis



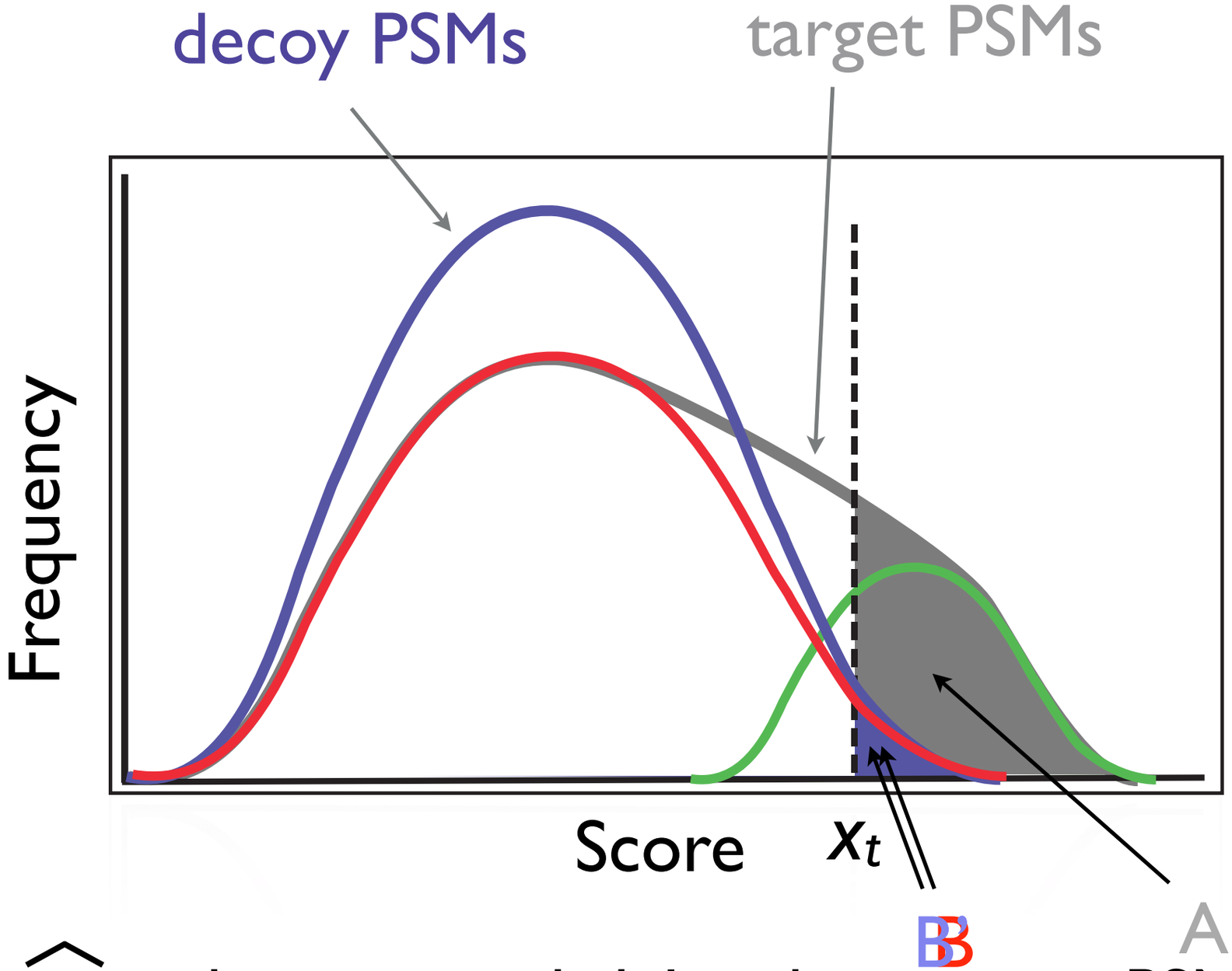

[Moore et al. JASMS 2002]

# Methods to generate decoy sequences

1. Shuffling sequences [Klammer *et al.* JPR 2006]
2. Markov models [Colinge *et al.* Proteomics 2003]
3. Reversing sequences [Moore *et al.* JASMS 2002]
4. Pseudo-Reversing sequences
   [Elias&Gygi *NMeth* 2007]

Its essential that the decoy PSMs are good proxies for incorrect target PSMs, which makes the first two methods less suitable

# Using decoy PSMs to estimate false discovery rate




$$FDR(x_t) = \frac{Pr(x \geq x_t, H=0)}{Pr(x \geq x_t)}$$

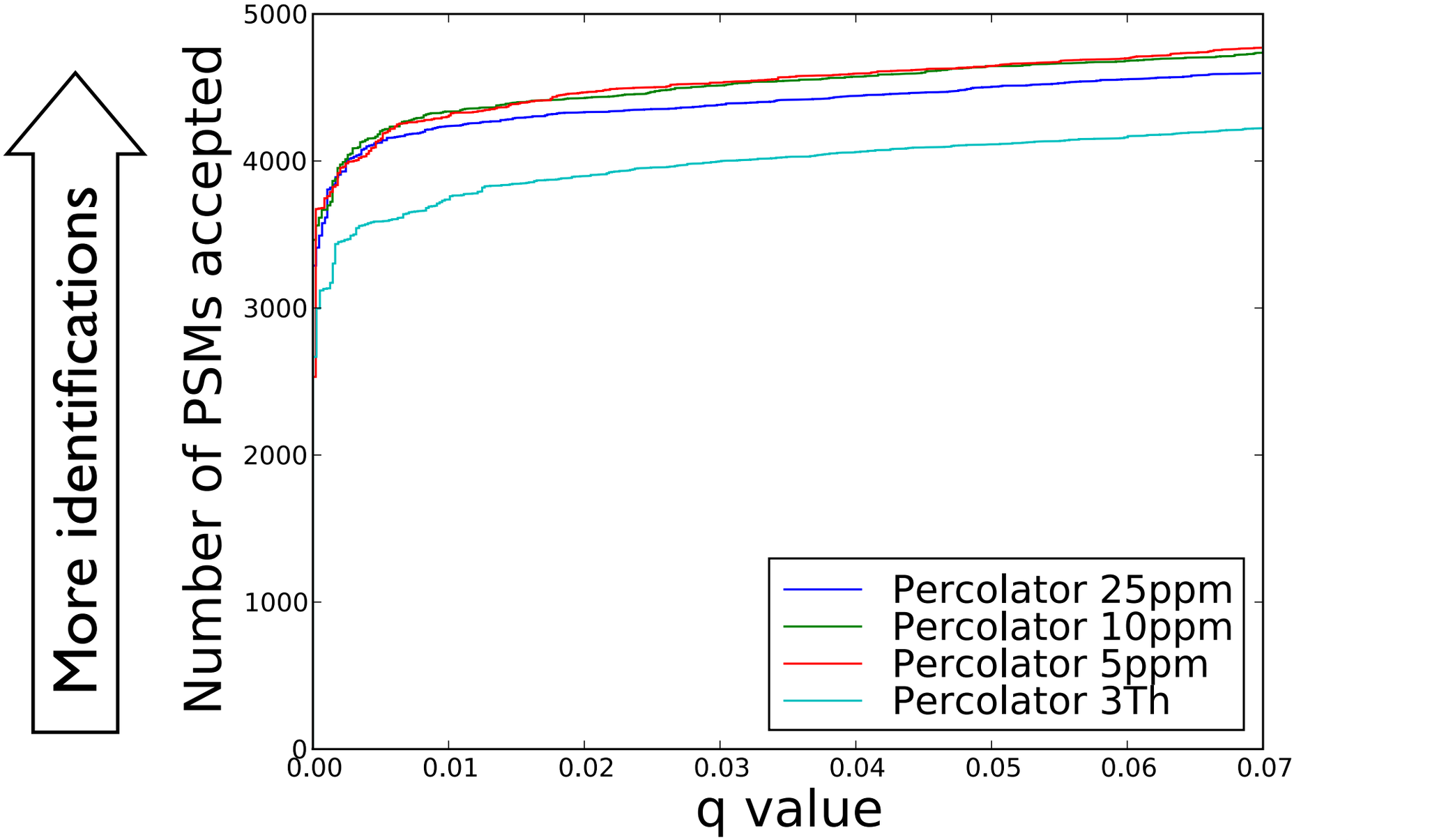
$$\widehat{FDR} = \frac{B}{A} = \frac{\widehat{\pi_0} \cdot B'}{A}$$

$$\widehat{q(x_t)} = \inf_{x \leq x_t} \{\widehat{FDR}(x)\}$$

[Käll *et al.* JPR 2008]

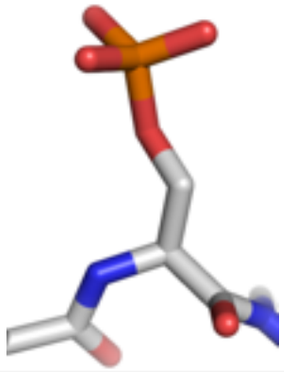
$\widehat{\pi_0}$ is the prior probability that a target PSM is incorrectly matched

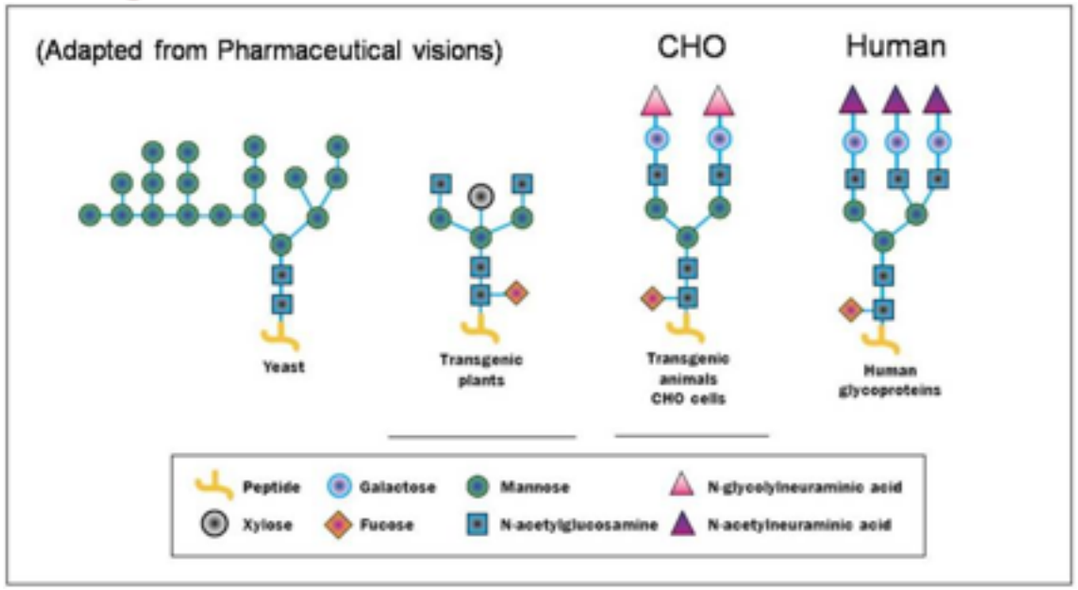
# q-p plot

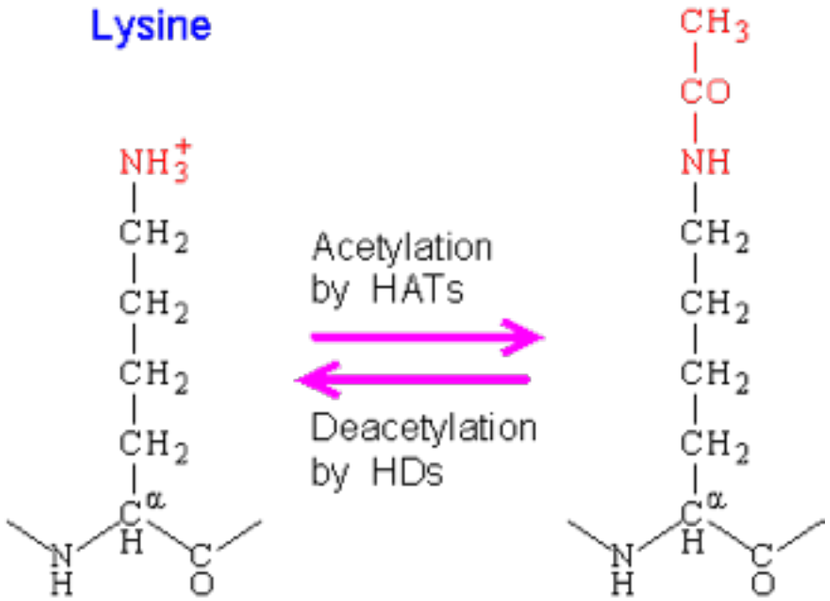

More identifications
Number of PSMs accepted
5000
4000
3000
2000
1000
0
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
q value
Percolator 25ppm
Percolator 10ppm
Percolator 5ppm
Percolator 3Th
Higher fraction incorrect identifications

# Some common PTMs

- Phosphorylations
  - Phosphate attached to serine or threonine
- Glycosylations
  - Glycans attached to a nitrogen (N-linked) of asparagine or arginine side-chains
- Accetylations
  - Accetyl group attached to lysine or N-terminus

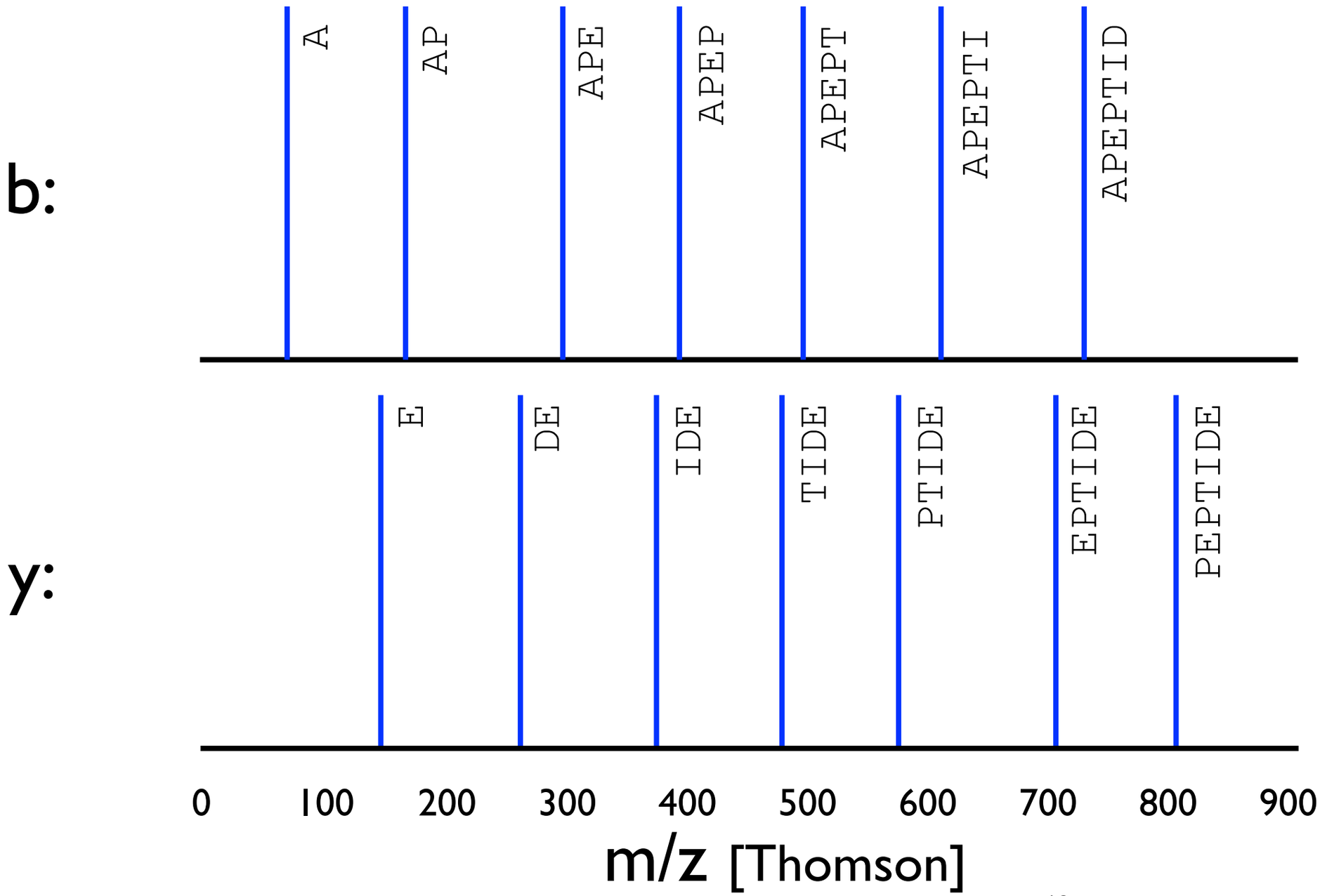




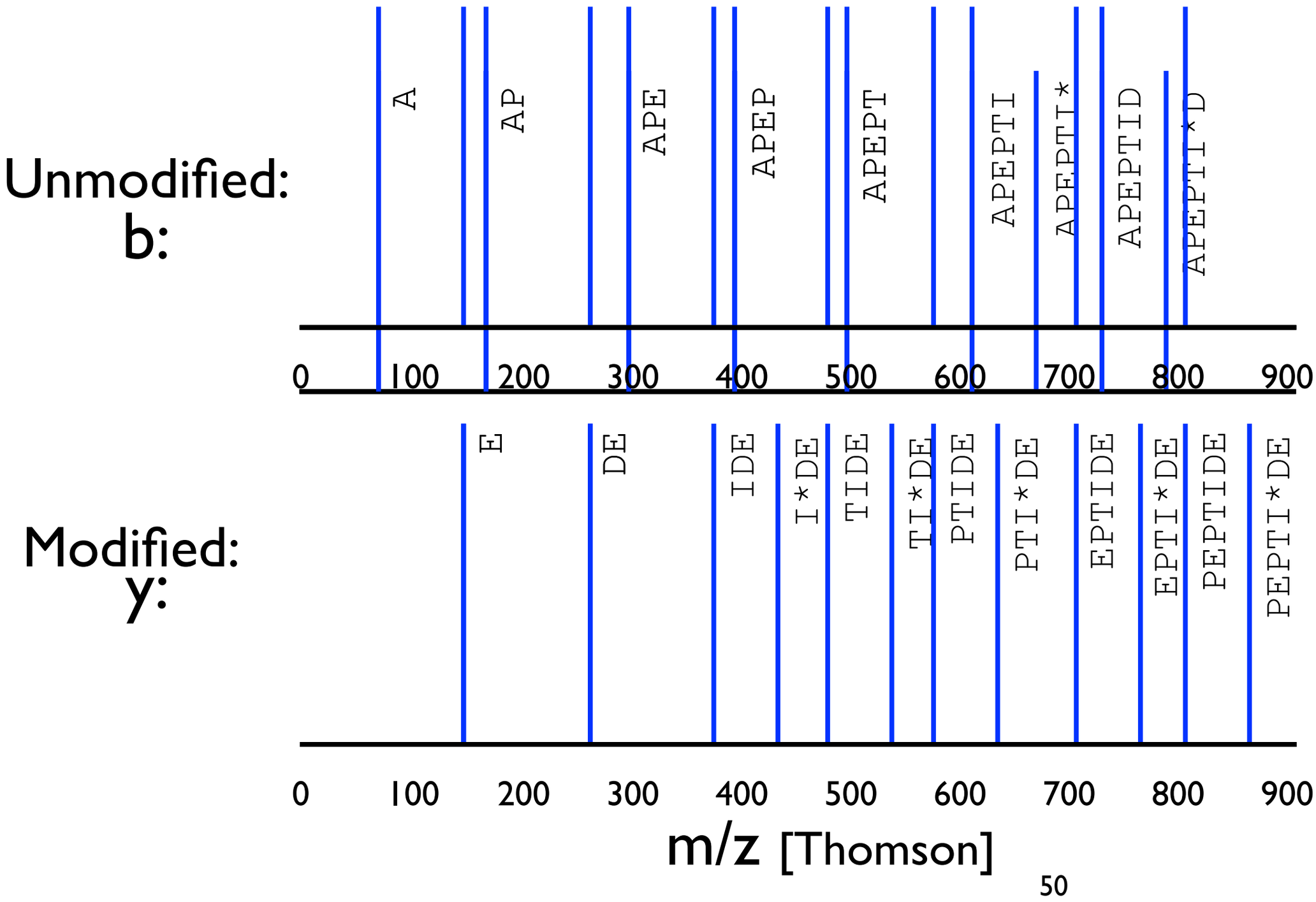
# Theoretical Spectrum of a peptide

A|P|E|P|T|I|D|E

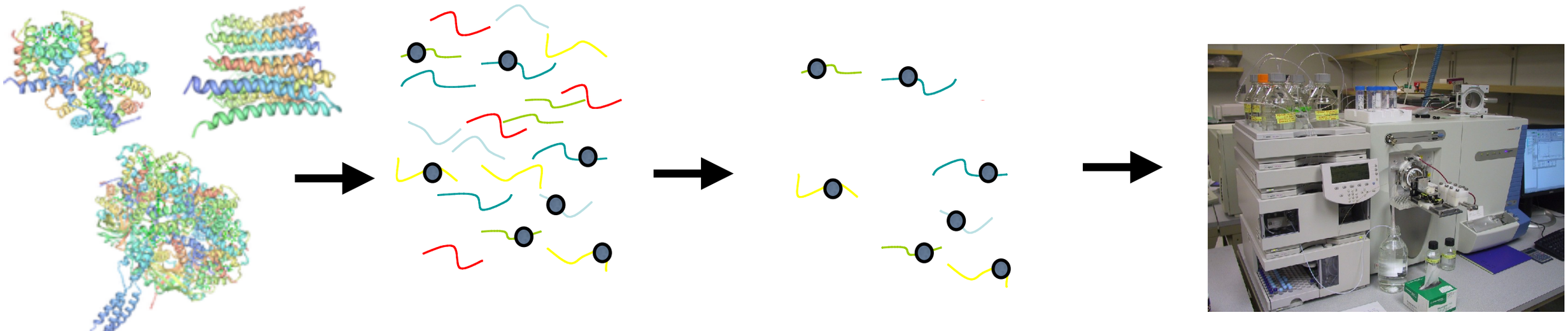


# Theoretical Spectrum of a PTM Peptide

A|P|E|P|T|I*|D|E



# Identification Post-Translational Modifications



- Large-scale identification of PTMs normally involve digestion, PTM enrichment and identification by MS/MS

# File format for spectral data


Spectral data

- XML-based: .mzXML, .mzData
- tab delimited: .ms2

```
S   45894  45894  2   1   maccoss007    2038.59    9199.5 147.1  153628
M     1  27 2040.244  0.0000  1.5881     245.6  11 34      V.YKCAADKQDATVVELTNL.T U
L   YCR102C
M     2  68 2038.265  0.0116  1.5698     208.4  11 36      S.TQSGIVAEQALLHSLNENL.S   U
L   YGR080W
M     3  34 2039.247  0.1582  1.3369     239.3  11 36      I.NEKTSPALVIPTPDAENEI.S   U
L   YLR035C
M     4 322 2040.365  0.1699  1.3183     160.0   9 36      I.LKESKSVQPGKAIPDIIES.P   U
L   YJL126W
```

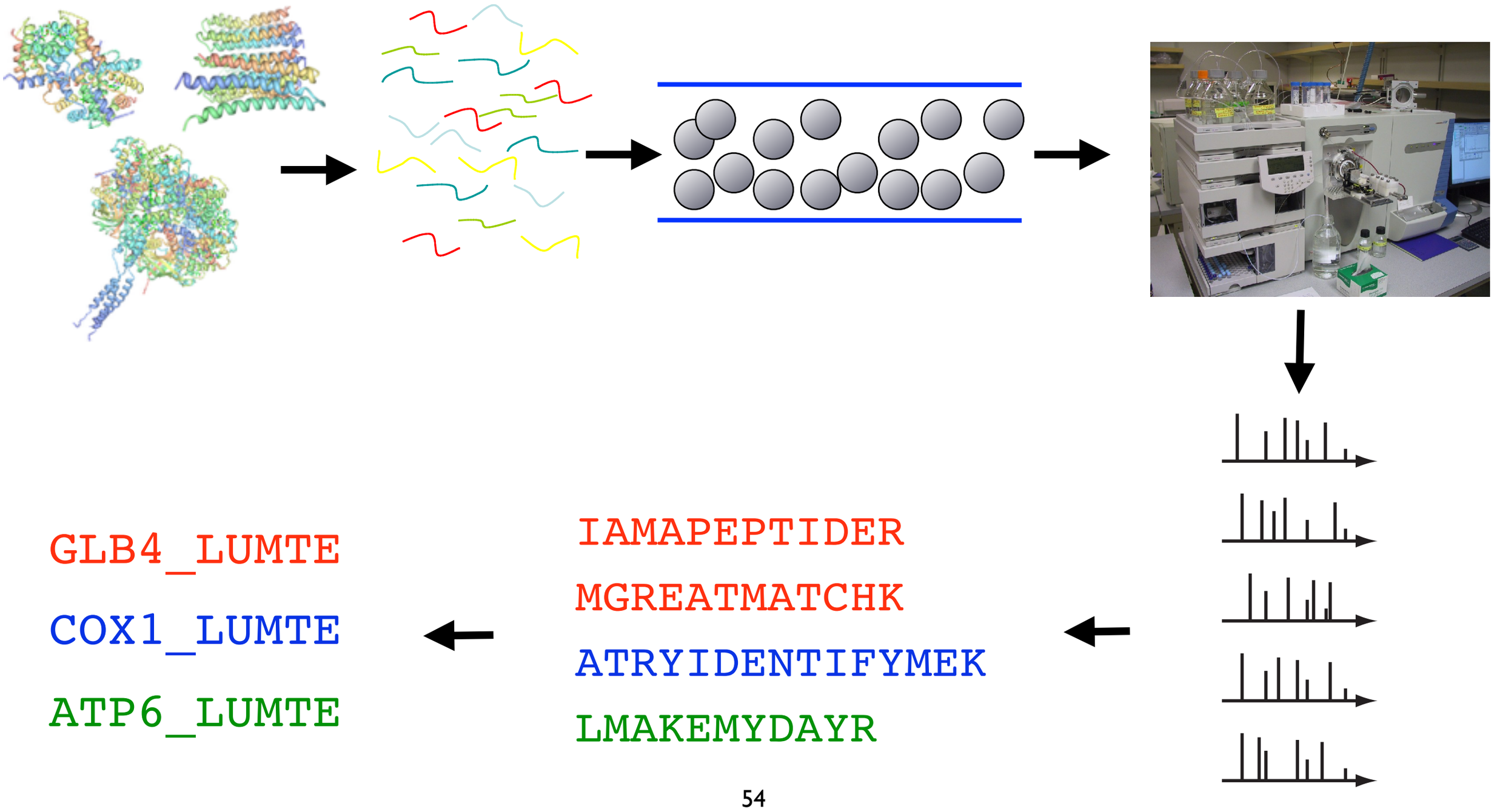
Peptide Spectrum Matches

- XML-based: pepXML, mzIdentML
- tab delimited: .sqt

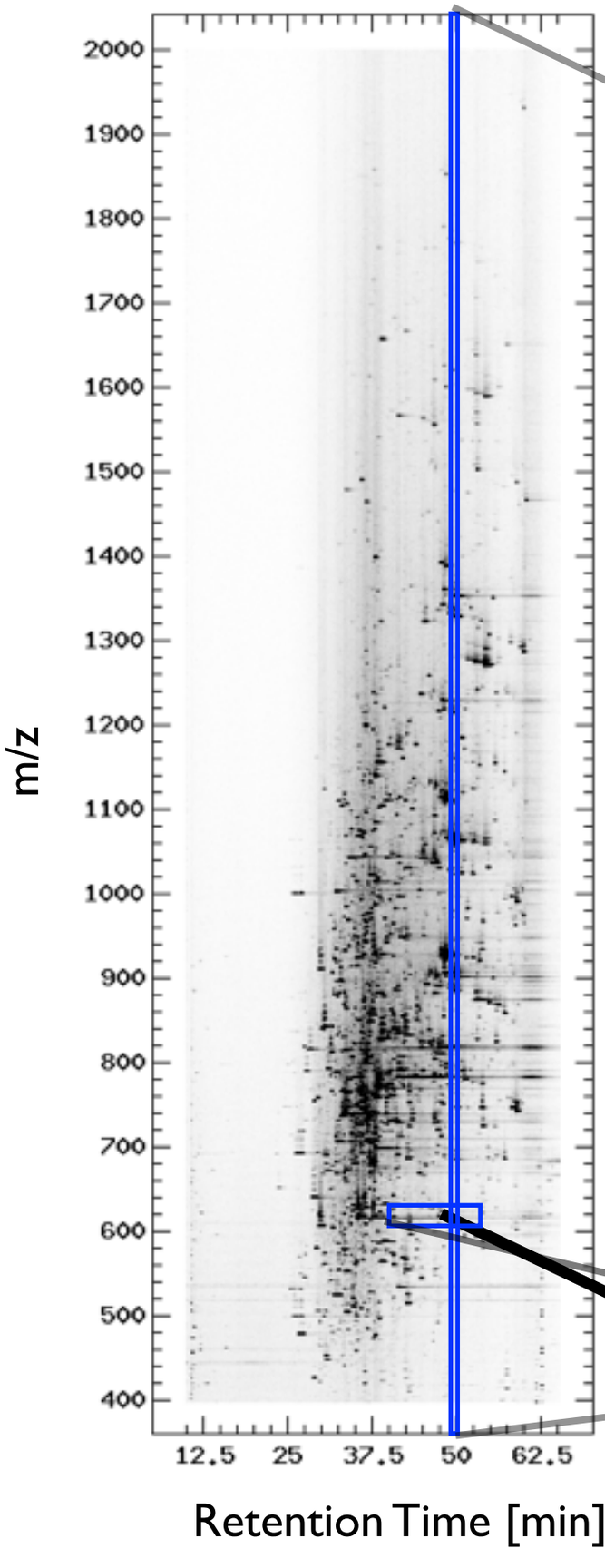
A nice file format converter - Proteowizard

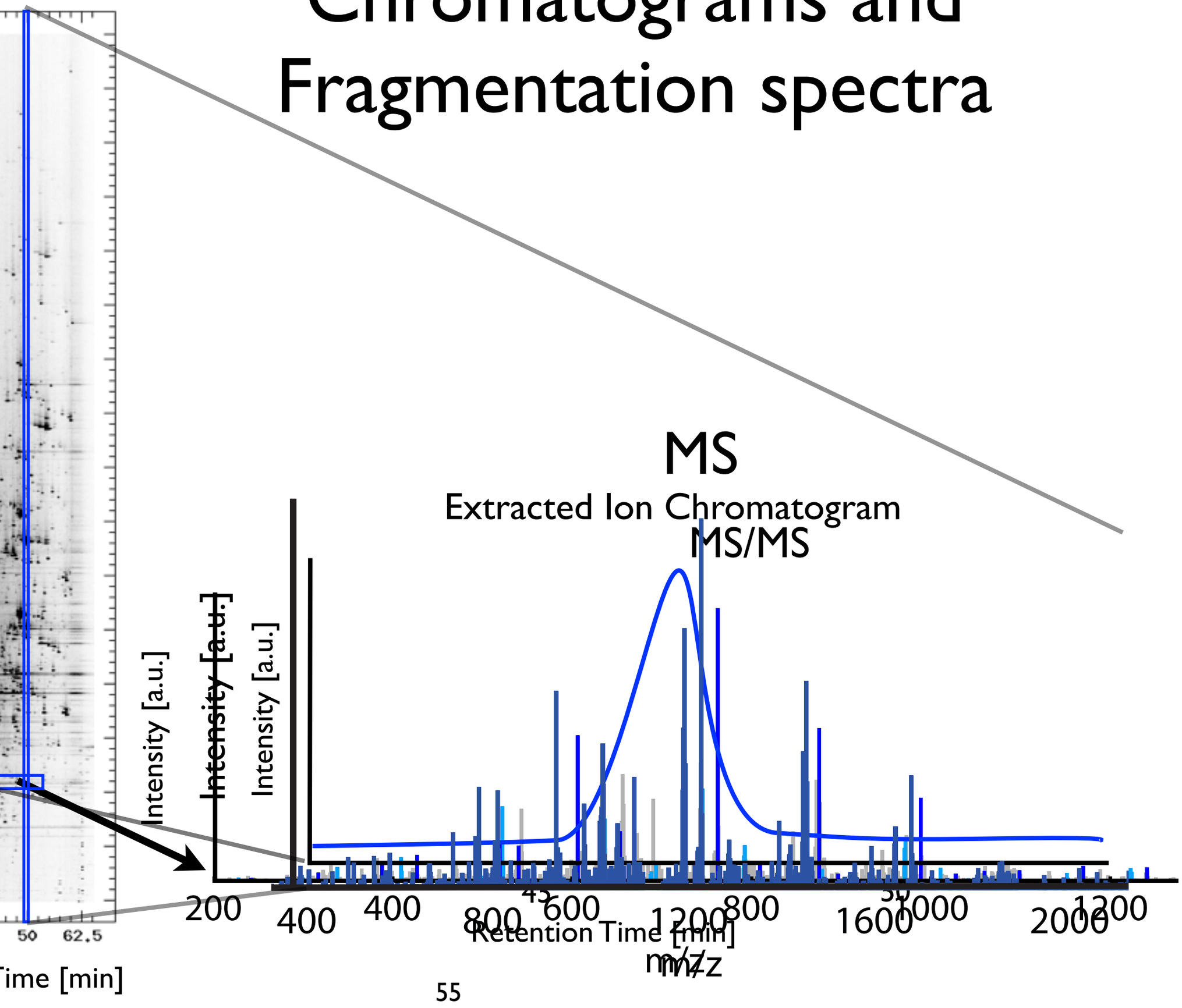
# Inferring proteins

# Shotgun proteomics

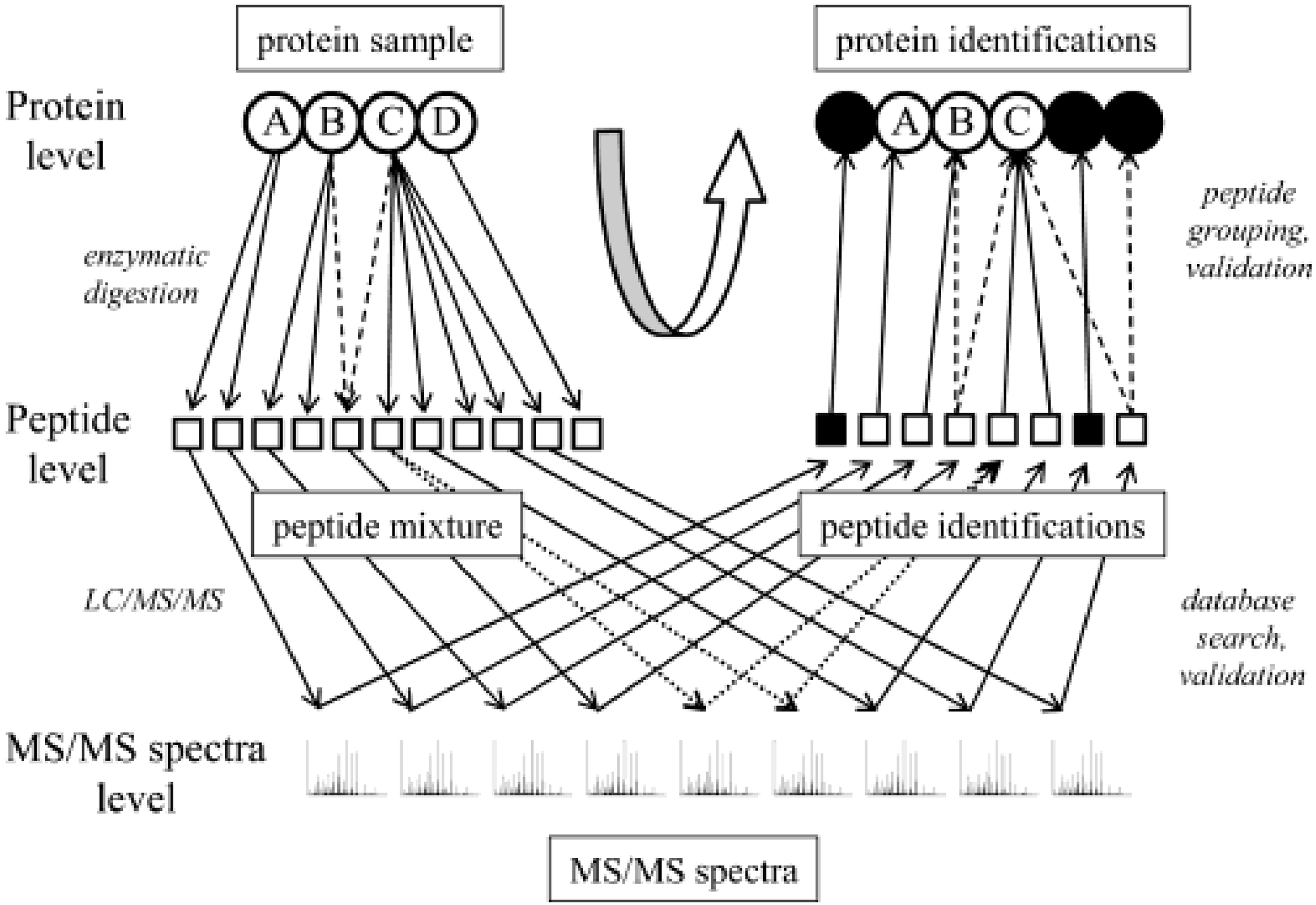


# Chromatograms and Fragmentation spectra



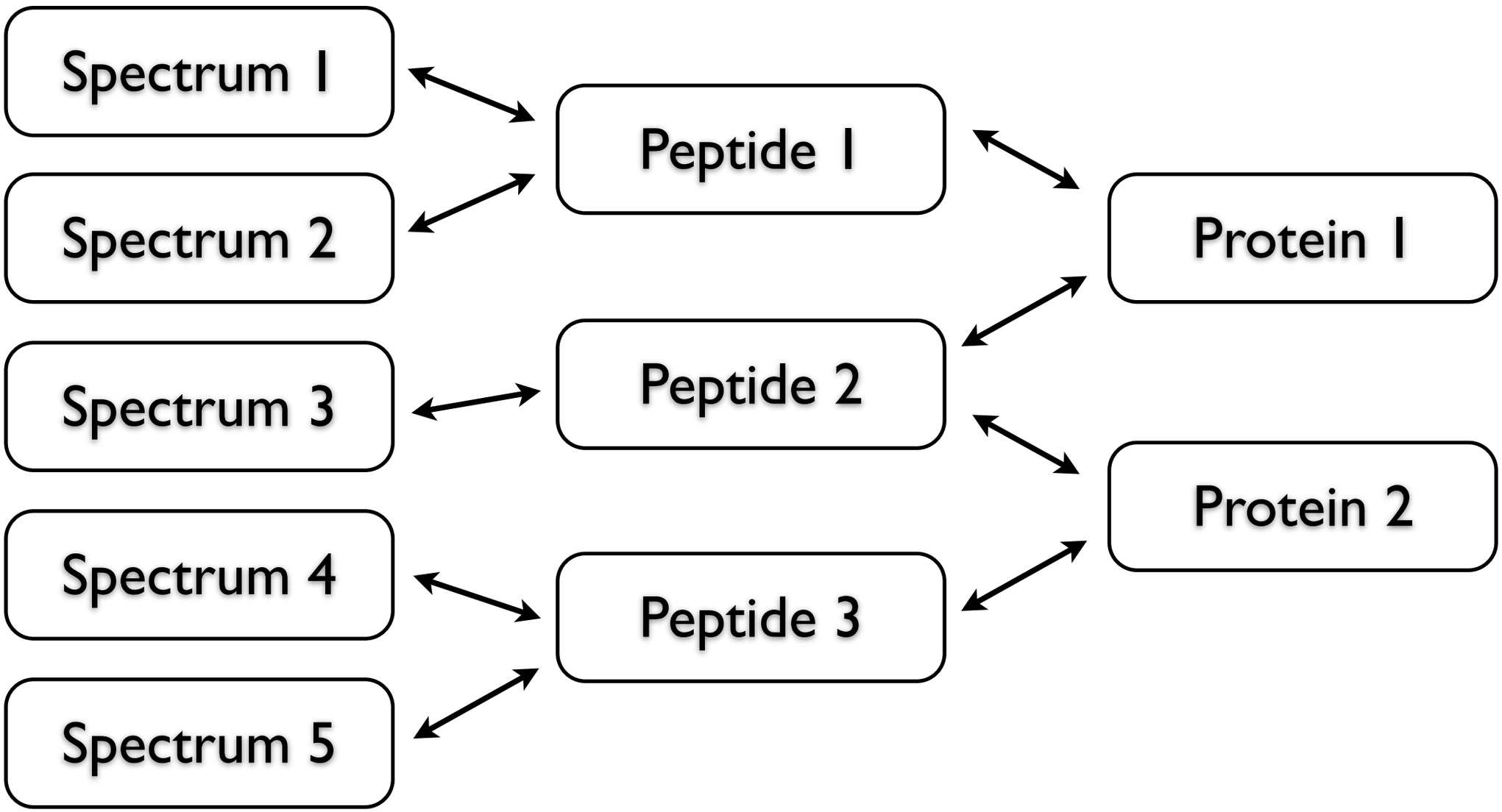



# protein identification

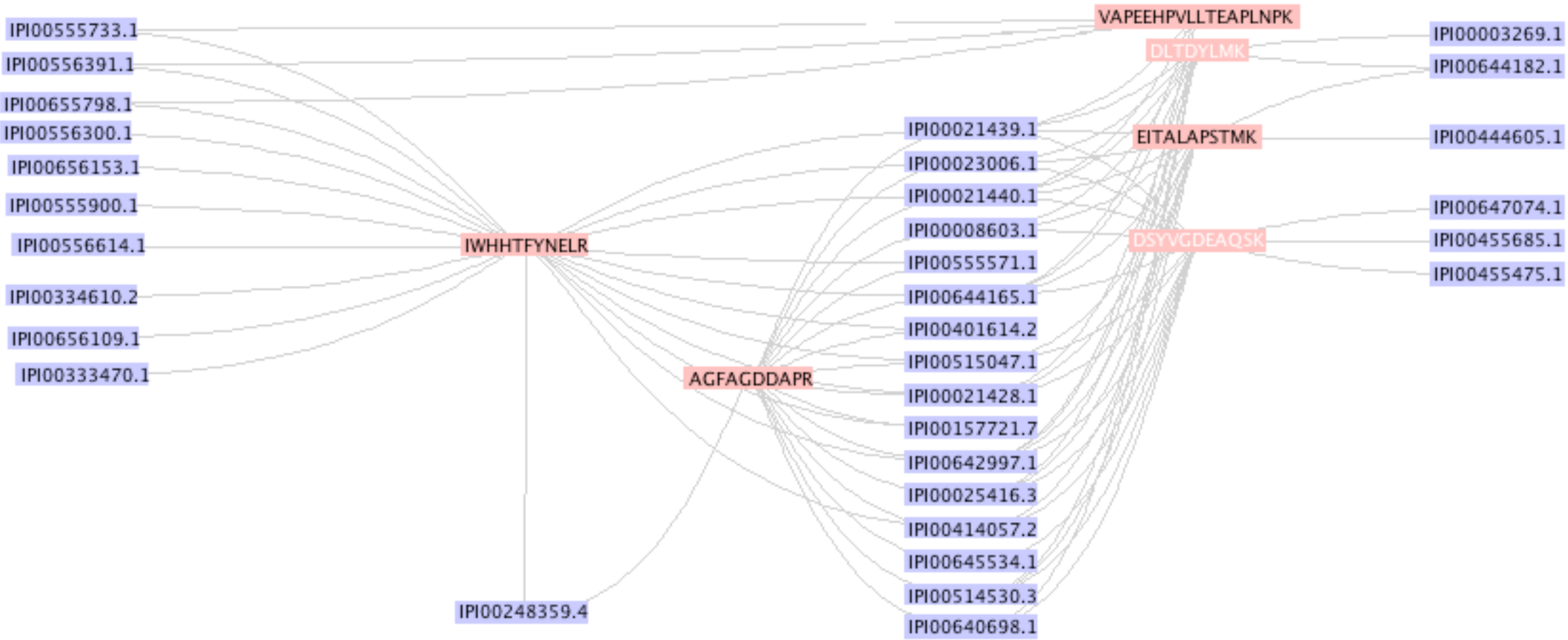



[Nesvizhskii *et al.* Anal Chem 2003]

# PSM/Peptide/Protein level



# Shared peptides



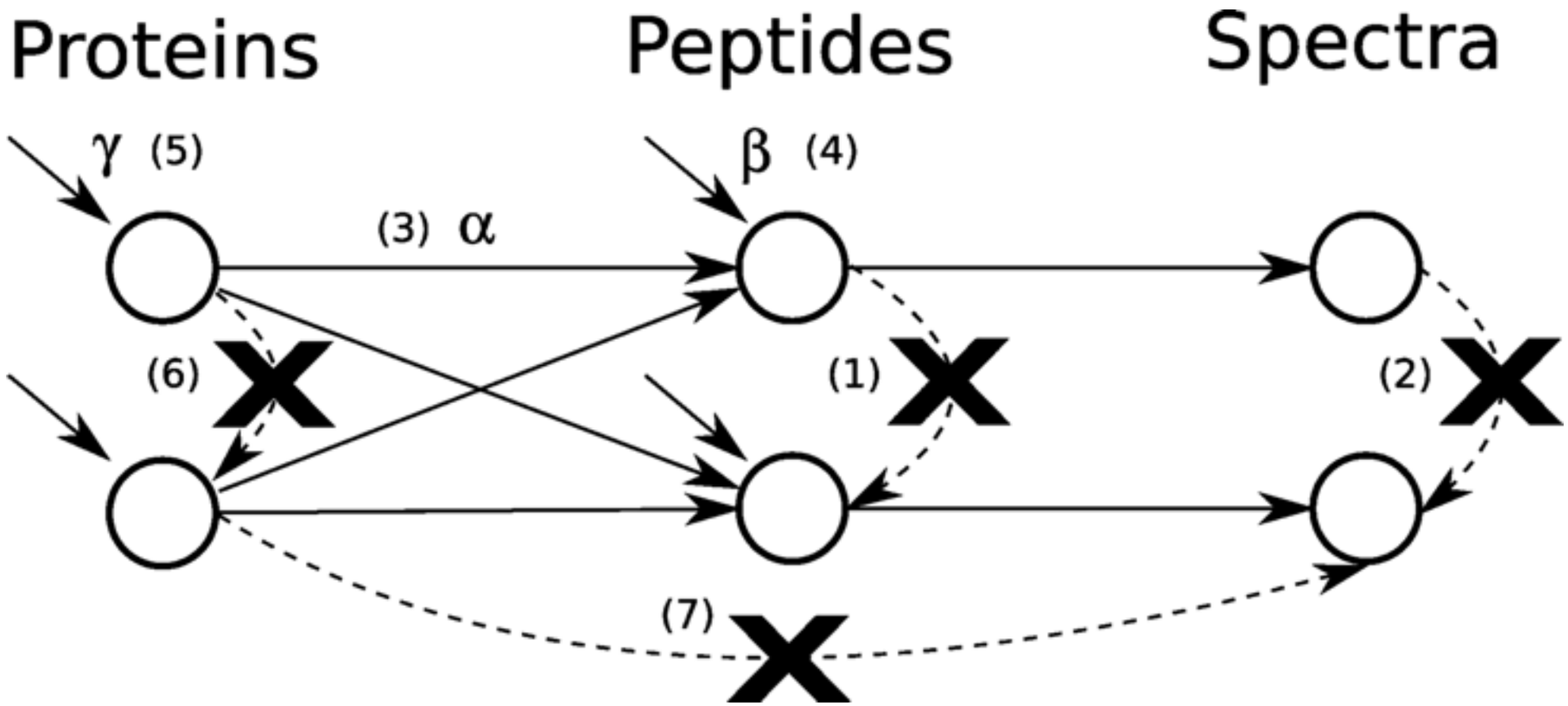

[MassSieve]

# Quality assessment of identified proteins

Two strategies:

1. Design a probabilistic models from which we may infer protein level probabilities
2. Target-decoy competition on protein level

# Bayesian Approach




1 Conditional Independence of Peptides Given Proteins
2 Conditional Independence of Spectra Given Peptides
3 Emission of a Peptide Associated with a Present Protein
4 Creation of a Peptide from Noise
5 Prior Belief a Protein Is Present in the Sample
6 Independence of Prior Belief between Proteins
7 Dependence of a Spectrum Only on the Best-Matching Peptide

Find MAP estimate protein set by evaluating
Pr(Proteins|Spectra)

[Serang *et al* JPR 2010]

# Papers for L9

perspectives Journal of proteome research

## Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases

Lukas Käll,[†] John D. Storey,[†,‡] Michael J. MacCoss,[†] and William Stafford Noble*[,†,§]

*Departments of Genome Sciences, Biostatistics, and Computer Science and Engineering, University of Washington, Seattle, Washington 98195*

**Received September 18, 2007**

Automated methods for assigning peptides to observed tandem mass spectra typically return a lis
peptide−spectrum matches, ranked according to an arbitrary score. In this article, we describe meth
for converting these arbitrary scores into more useful statistical significance measures. These meth
employ a decoy sequence database as a model of the null hypothesis, and use false discovery
(FDR) analysis to correct for multiple testing. We first describe a simple FDR inference method
then describe how estimating and taking into account the percentage of incorrectly identified spe
n the entire data set can lead to increased statistical power.

**Keywords:** *q*-value • decoy database • false discovery rate • statistical significance • peptide identifica

### ion

e problem in the analysis of tandem mass spectra is
the peptide that gave rise to an observed fragmenta-
rum. The most commonly used tools for solving this
such as SEQUEST,[1] Mascot,[2] or X!Tandem,[3] search a
uence database for the peptide whose theoretical
best matches the observed spectrum. The output of
of the analysis is a collection of peptide−spectrum
PSMs), each with an associated score (Table 1). The
ubsequent question is, "Which of these PSMs are

hese algorithms are very powerful, the problem is that

**Table 1.** Terminology

| | |
|---|---|
| PSM | A peptide−spectrum match, with a associated score |
| target PSM | A PSM created by searching the or database |
| decoy database | A shuffled or reversed version of t database |
| decoy PSM | A PSM created by searching a dec database |
| accepted PSM | A PSM whose score is above some threshold |
| correct PSM | A PSM whose peptide corresponds peptide that generated the obser |

ARTICLE

doi:10.1038

## A draft map of the human proteome

Min-Sik Kim[1,2], Sneha M. Pinto[3], Derese Getnet[1,4], Raja Sekhar Nirujogi[3], Srikanth S. Manda[3], Raghothama Chaerkad
Anil K. Madugundu[3], Dhanashree S. Kelkar[3], Ruth Isserlin[5], Shobhit Jain[5], Joji K. Thomas[3], Babylakshmi Muthusamy
Pamela Leal-Rojas[1,6], Praveen Kumar[3], Nandini A. Sahasrabuddhe[3], Lavanya Balakrishnan[3], Jayshree Advani[3], Bijesh
Santosh Renuse[3], Lakshmi Dhevi N. Selvan[3], Arun H. Patil[3], Vishalakshi Nanjappa[3], Aneesha Radhakrishnan[3], Samarje
Tejaswini Subbannayya[3], Rajesh Raju[3], Manish Kumar[3], Sreelakshmi K. Sreenivasamurthy[3], Arivusudar Marimuthu[3]
Gajanan J. Sathe[3], Sandip Chavan[3], Keshava K. Datta[3], Yashwanth Subbannayya[3], Apeksha Sahu[3], Soujanya D. Yelam
Savita Jayaram[3], Pavithra Rajagopalan[3], Jyoti Sharma[3], Krishna R. Murthy[3], Nazia Syed[3], Renu Goel[3], Aafaque A. Kha
Sartaj Ahmad[3], Gourav Dey[3], Keshav Mudgal[7], Aditi Chatterjee[3], Tai-Chung Huang[1], Jun Zhong[1], Xinyan Wu[1,2], Patrick
Donald Freed[1], Muhammad S. Zahari[2], Kanchan K. Mukherjee[8], Subramanian Shankar[9], Anita Mahadevan[10,11], Henry
Christopher J. Mitchell[1], Susarla Krishna Shankar[10,11], Parthasarathy Satishchandra[13], John T. Schroeder[14], Ravi Sirdes
Anirban Maitra[15,16], Steven D. Leach[1,17], Charles G. Drake[16,18], Marc K. Halushka[15], T. S. Keshava Prasad[3], Ralph H. H
Candace L. Kerr[19]†, Gary D. Bader[5], Christine A. Iacobuzio-Donahue[15,16,17], Harsha Gowda[3] & Akhilesh Pandey[1,2,3,4,

**The availability of human genome sequence has transformed biomedical research over the past decade. However, a alent map for the human proteome with direct measurements of proteins and peptides does not exist yet. Here we a draft map of the human proteome using high-resolution Fourier-transform mass spectrometry. In-depth pr profiling of 30 histologically normal human samples, including 17 adult tissues, 7 fetal tissues and 6 purified prim matopoietic cells, resulted in identification of proteins encoded by 17,294 genes accounting for approximately 84 total annotated protein-coding genes in humans. A unique and comprehensive strategy for proteogenomic enabled us to discover a number of novel protein-coding regions, which includes translated pseudogenes, non RNAs and upstream open reading frames. This large human proteome catalogue (available as an interactive we resource at http://www.humanproteomemap.org) will complement available human genome and transcriptom accelerate biomedical research in health and disease.**

Analysis of the complete human genome sequence has thus far led to the identification of approximately 20,687 protein-coding genes[1], although the annotation still continues to be refined. Mass spectrometry has revolutionized proteomics studies in a manner analogous to the impact of next-generation sequencing on genomics and transcriptomics[2–4]. Several groups, including ours, have used mass spectrometry to catalogue complete proteomes of unicellular organisms[5–7] and to explore proteomes

sets—PeptideAtlas[11], GPMDB[12] and neXtProt[13] (which in
tations from the Human Protein Atlas[14]).

A general limitation of current proteomics methods is
dence on predefined protein sequence databases for ide
teins. To overcome this, we also used a comprehensive pro
analysis strategy to identify novel peptides/proteins that
not part of annotated protein databases. This approach re