Mass spectrometry-based proteomics

Lukas Käll

Overview

Mass spectrometry

- Ionization sources
- MS Technologies
- Fragmentation

Proteomics

- Dynamic range
- Complexity
- Limitations

Peptide Identification

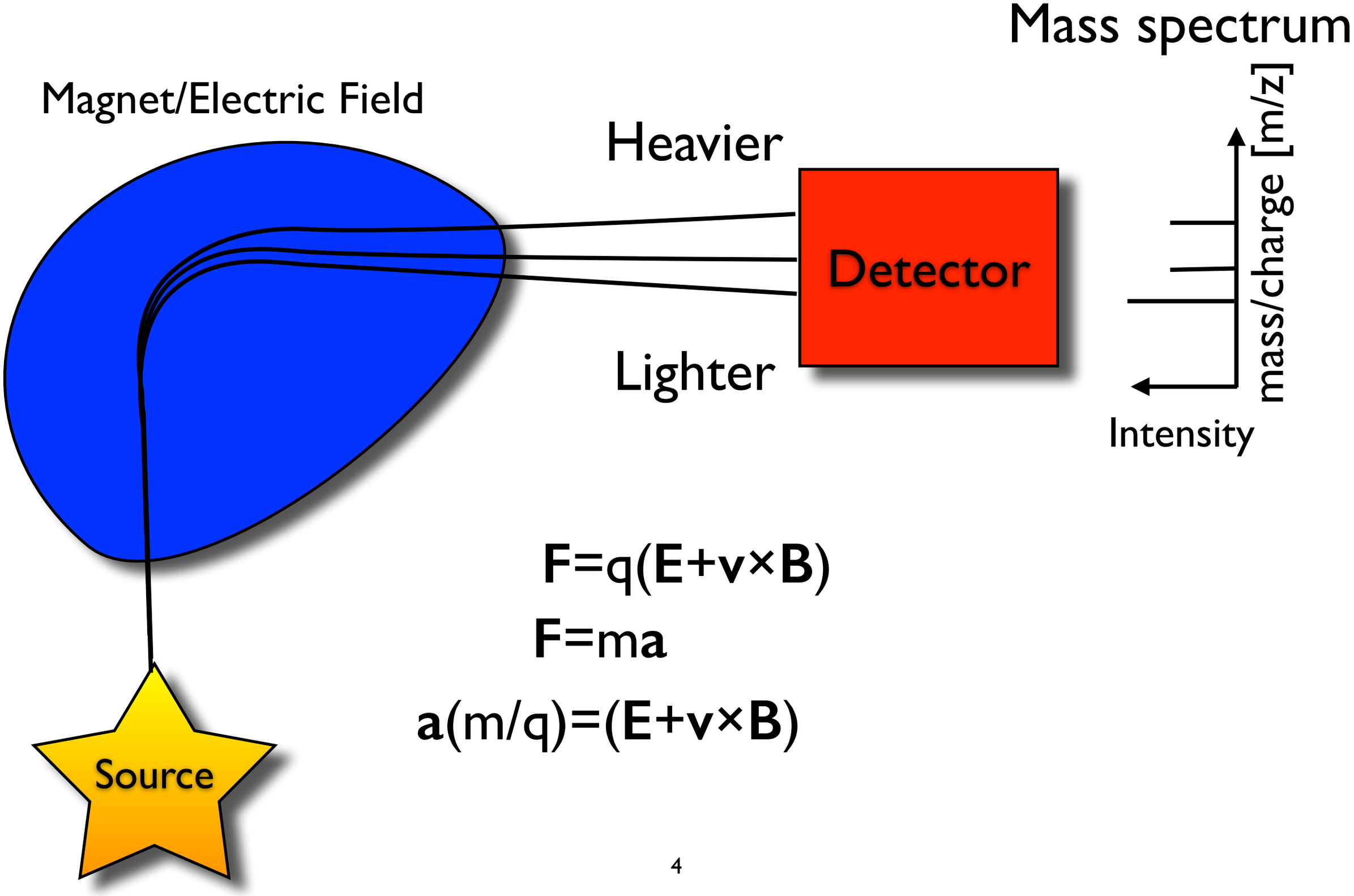
- Search engines
- Statistics

- Post-translational Modifications (PTMs)

Protein inference

Mass spectrometry

Mass spectrometry



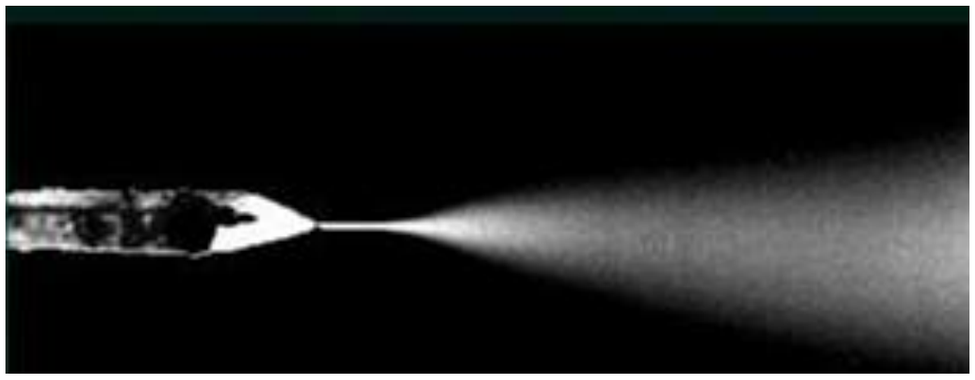
$$\mathbf{F}=q(\mathbf{E}+\mathbf{v}\times\mathbf{B})$$

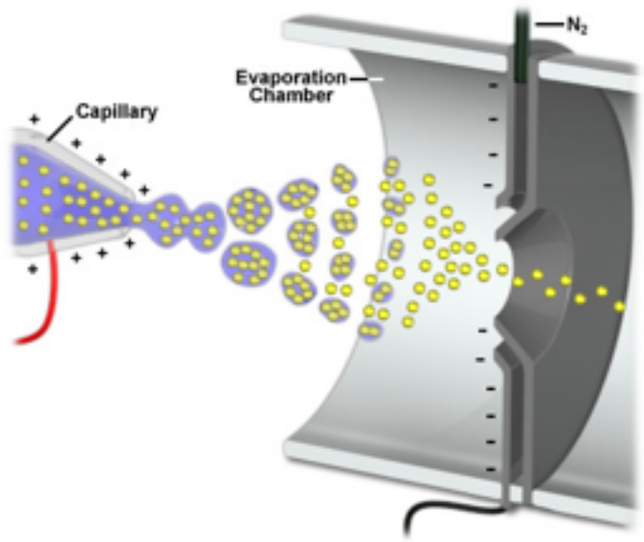
$$\mathbf{F}=m\mathbf{a}$$

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

Ion-Sources

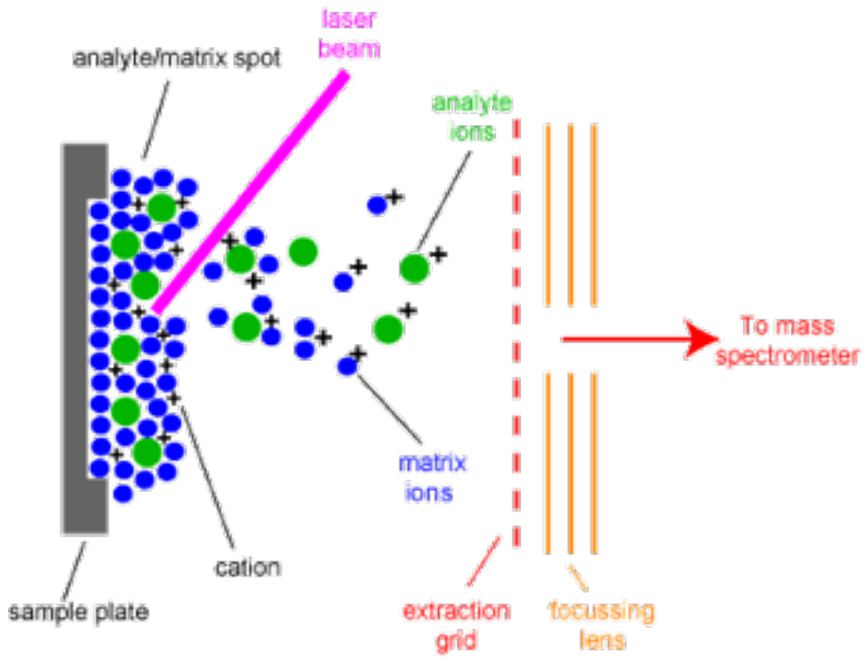
Electro spray ionization





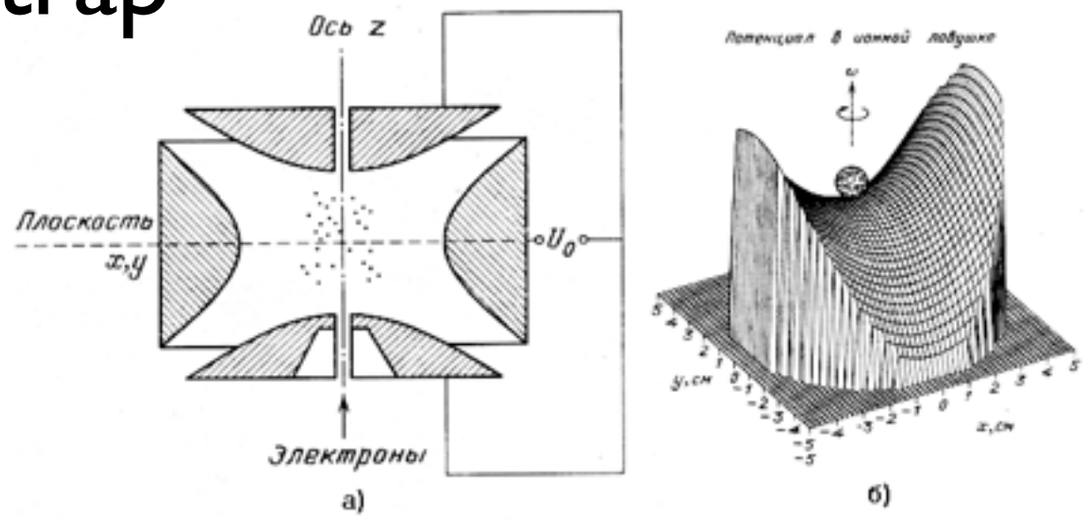
Nobel Prize 2002
John Fenn

MALDI



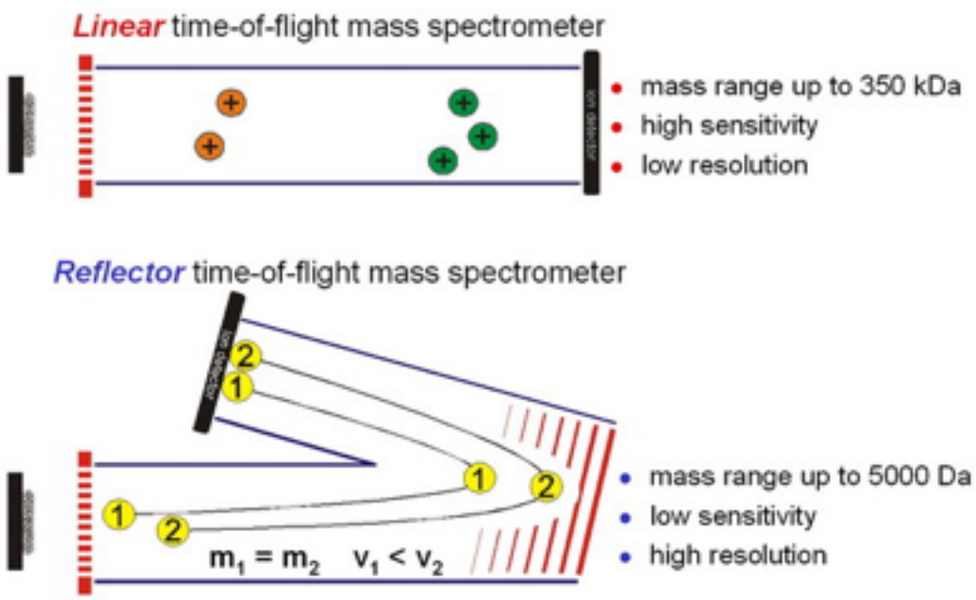
Separation

Ion trap

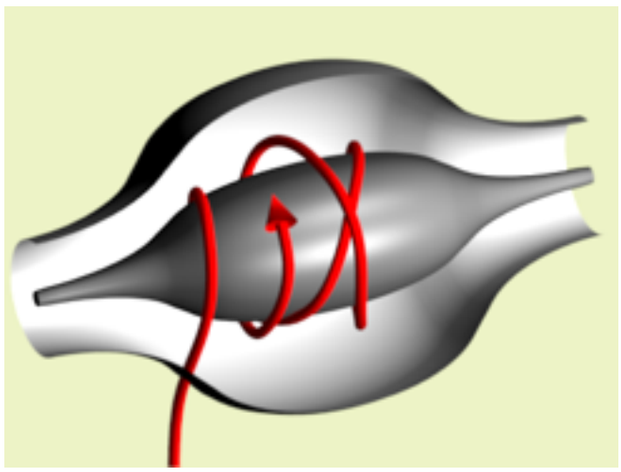


- Ions kept in place by AC current generating an electric field
- The ions will leave the trap as we increase the voltage of the control in order of their m/z (lowest first)

Time of Flight (TOF)



OrbiTrap

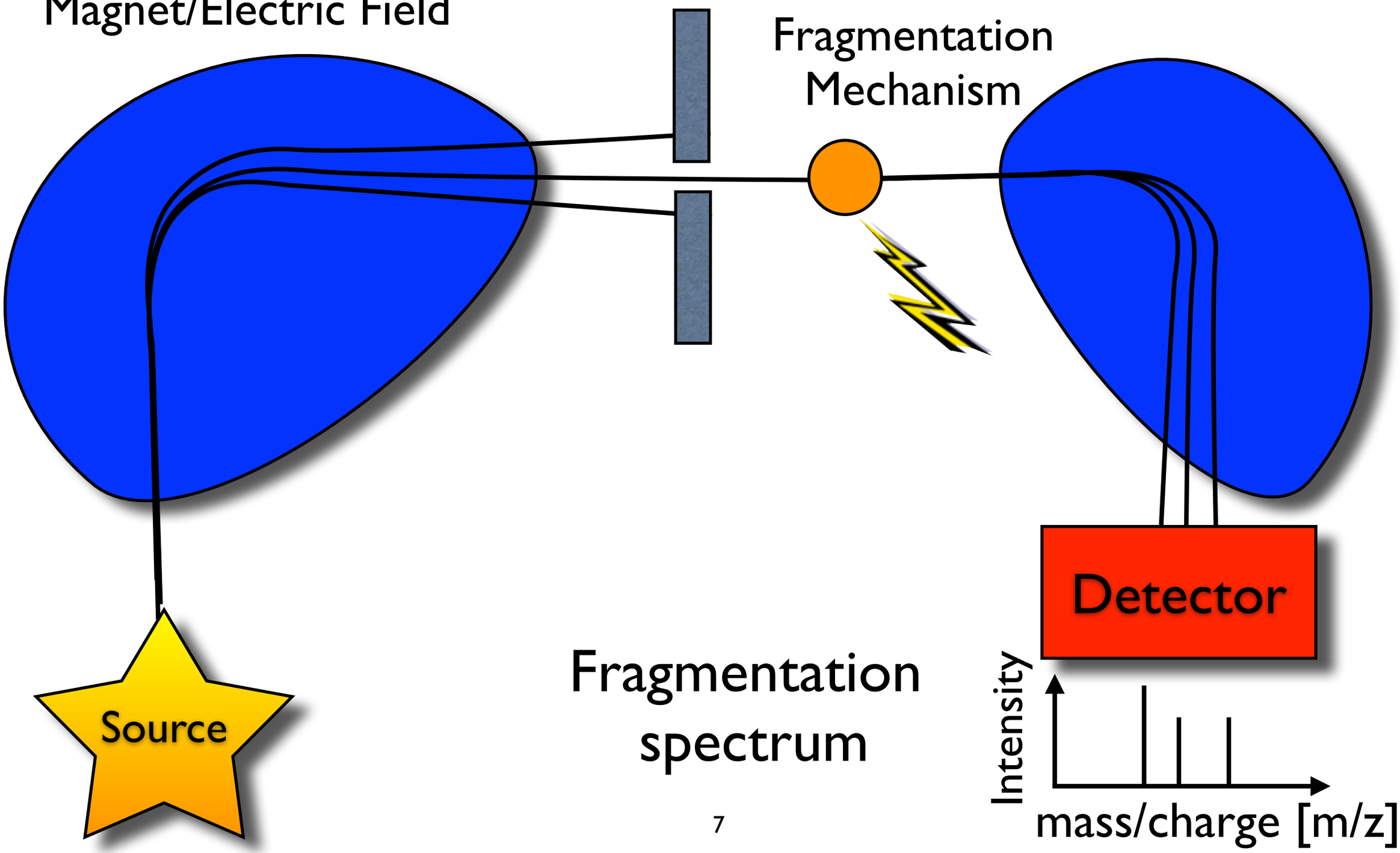


- Electrostatic attraction to the inner electrode is balanced by centrifugal forces.
- Detect the ions by their movement along the axis of the electrode.
- High mass accuracy (1–2 ppm)

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?

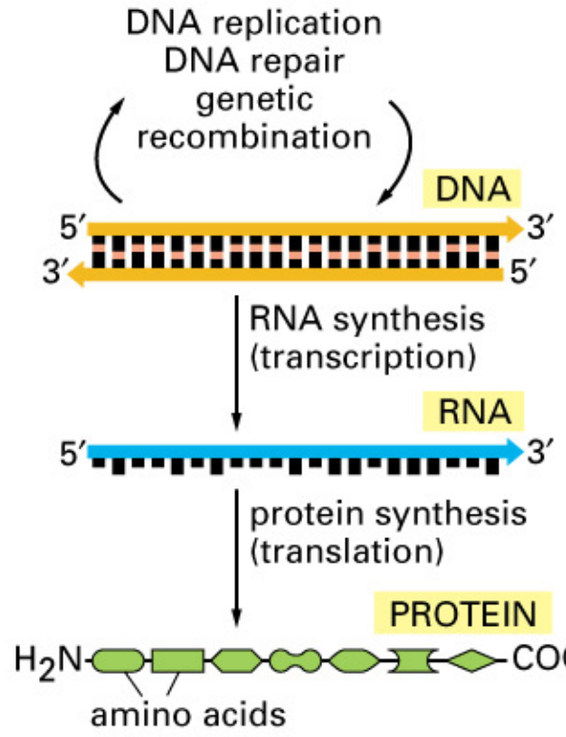
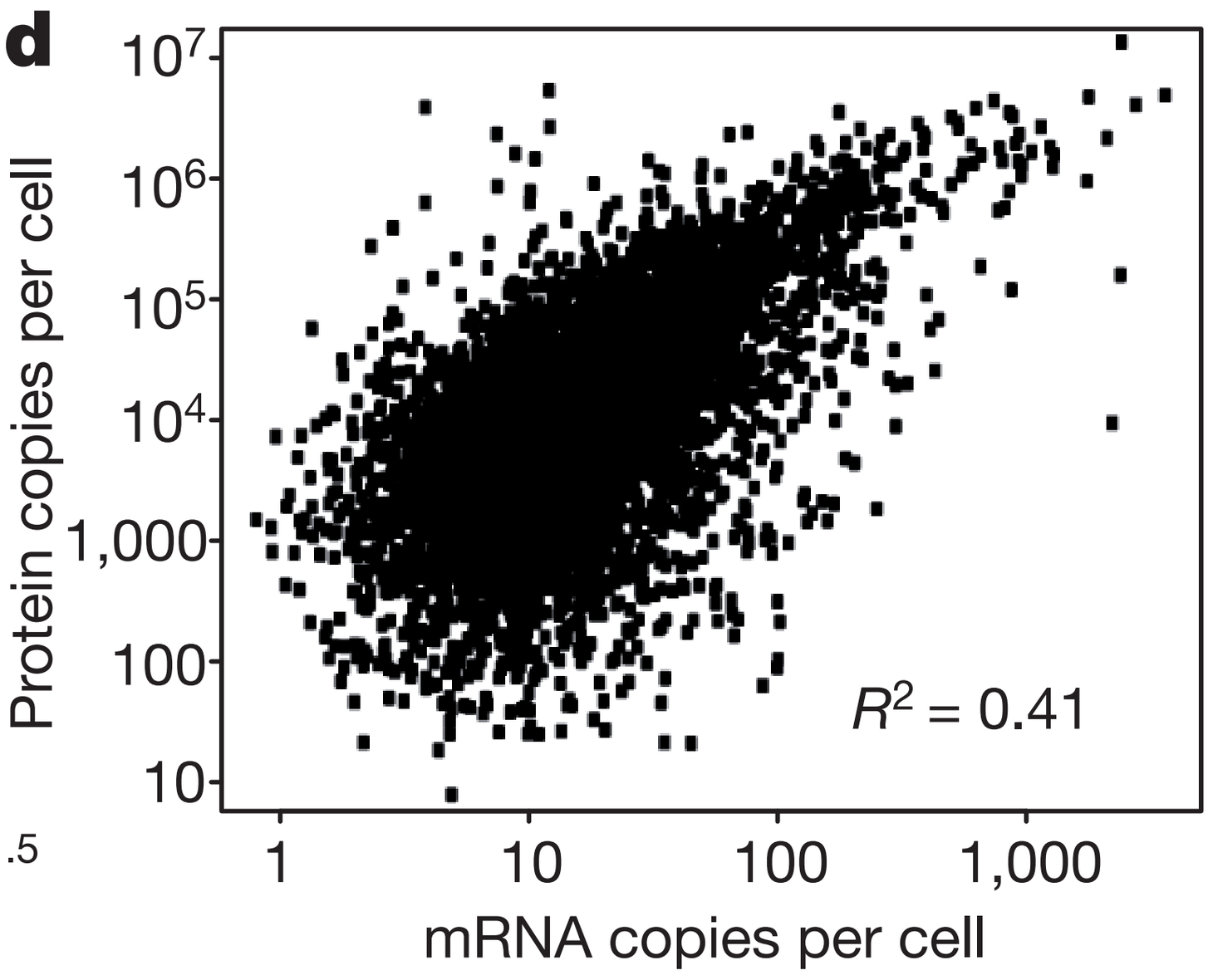


Figure 6–2. Molecular Biology of the Cell, 4th Edition.

mRNA levels correlate only weakly with protein levels

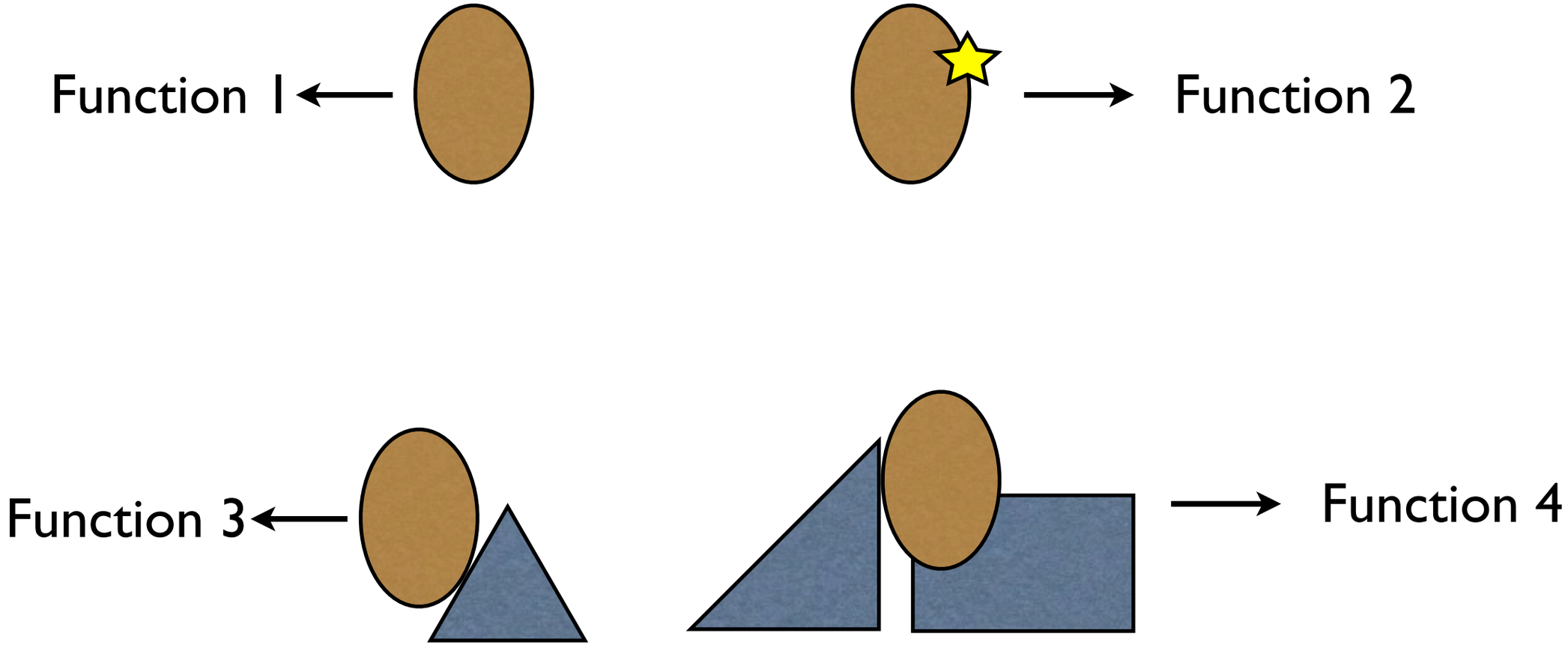
[Schwanhäusser *et al.* Nature 2011]



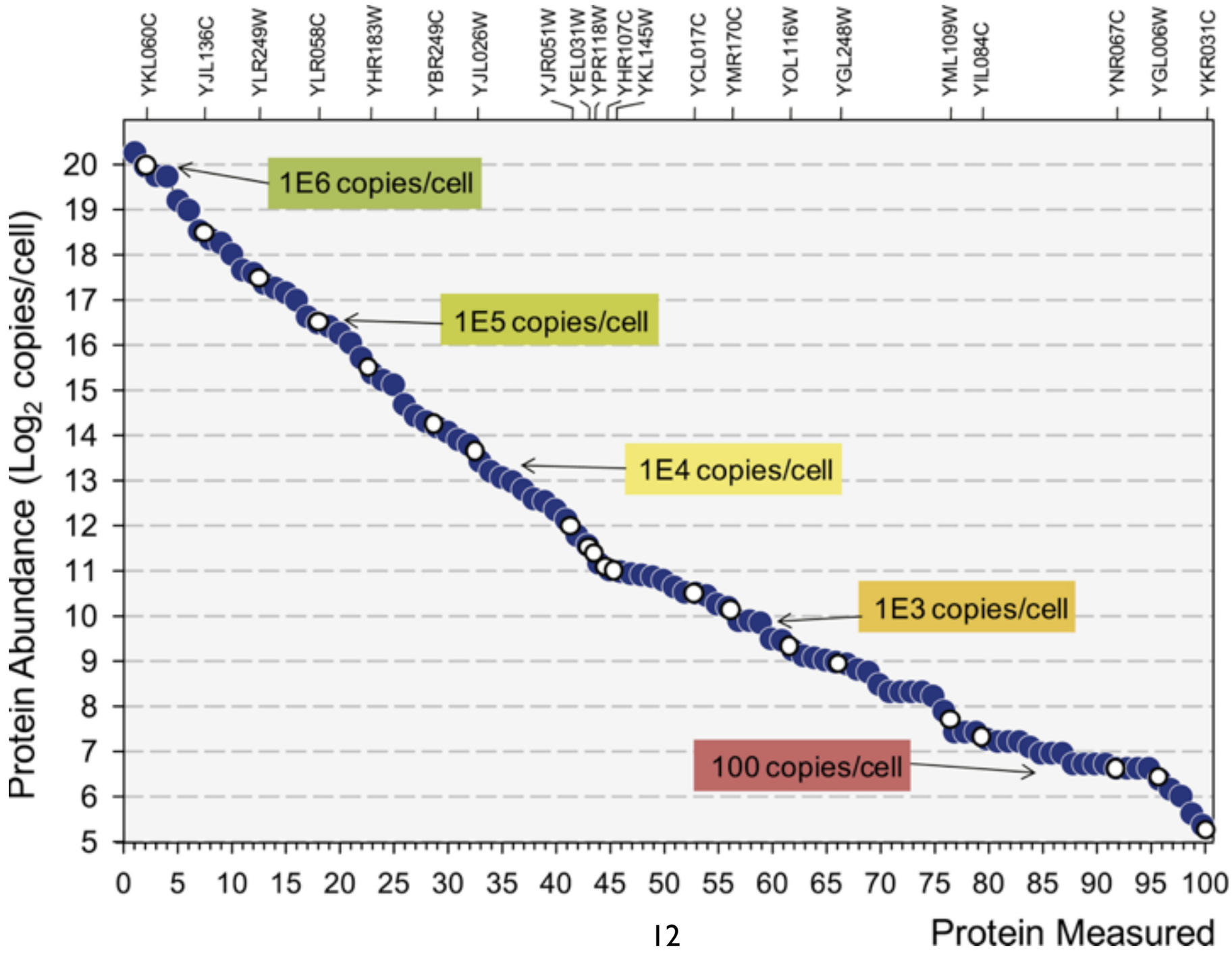
One protein - Many functions

Proteins undergo post-translational modifications

Proteins may have different functions in different complexes or compartments

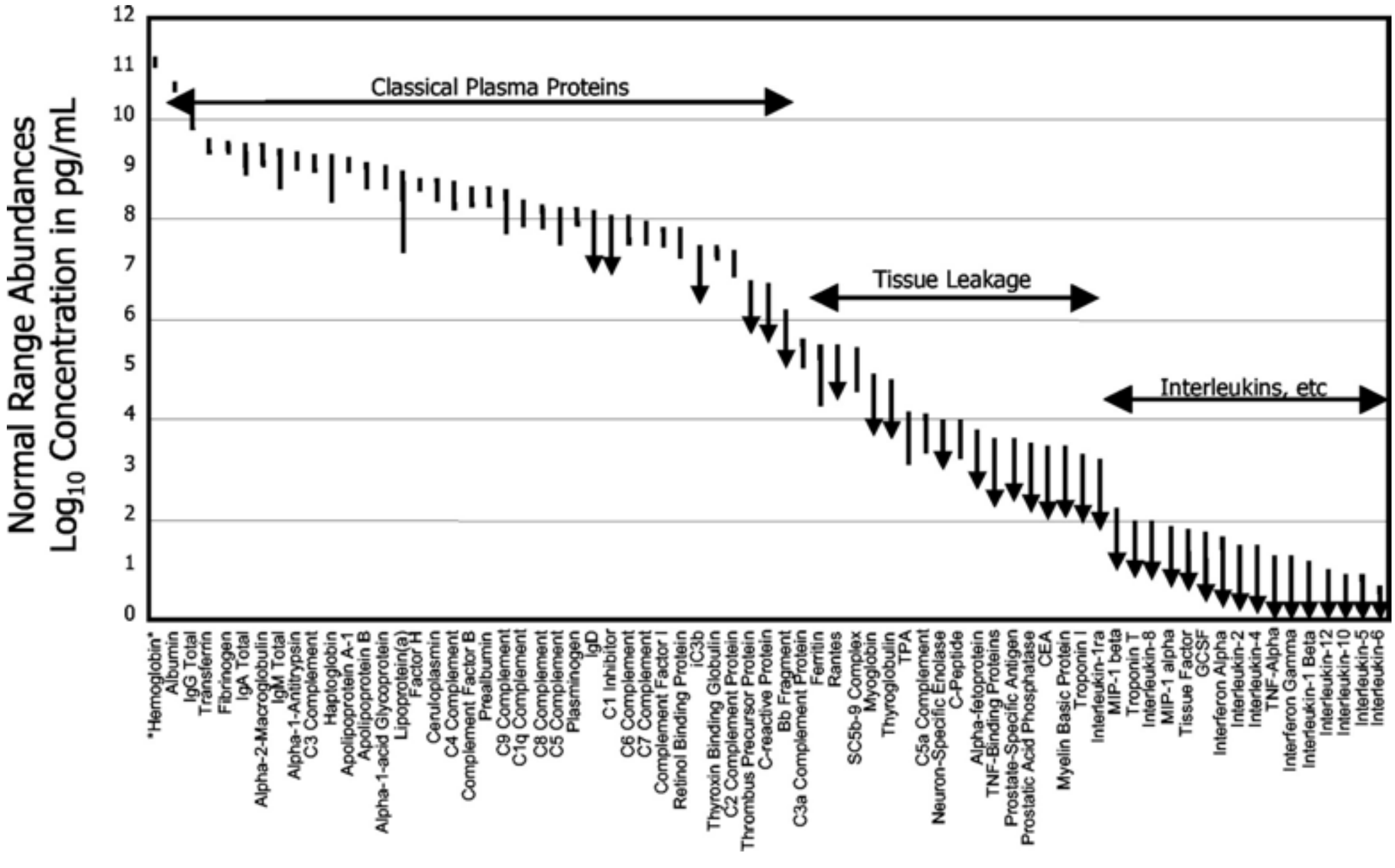


Proteins concentration in yeast range >4 orders of magnitude



[Picotti et al Cell 2009]

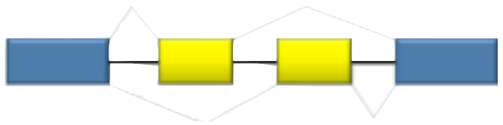
Protein concentration in blood plasma range >10 orders of magnitude



[Andersson & Anderson, MCP 2002]

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

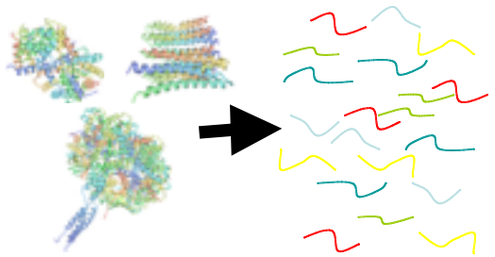
How to define a protein?

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

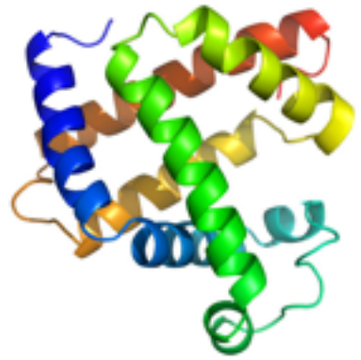
Proteomic techniques

Mass spectrometry-based approaches

e.g.



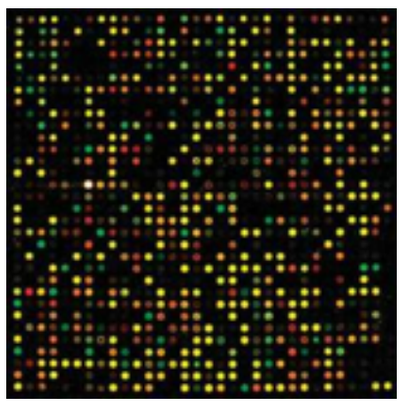
Shotgun proteomics



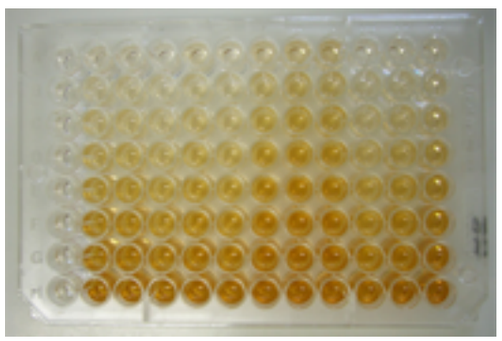
Top-down proteomics

Antibody-based approaches

e.g.

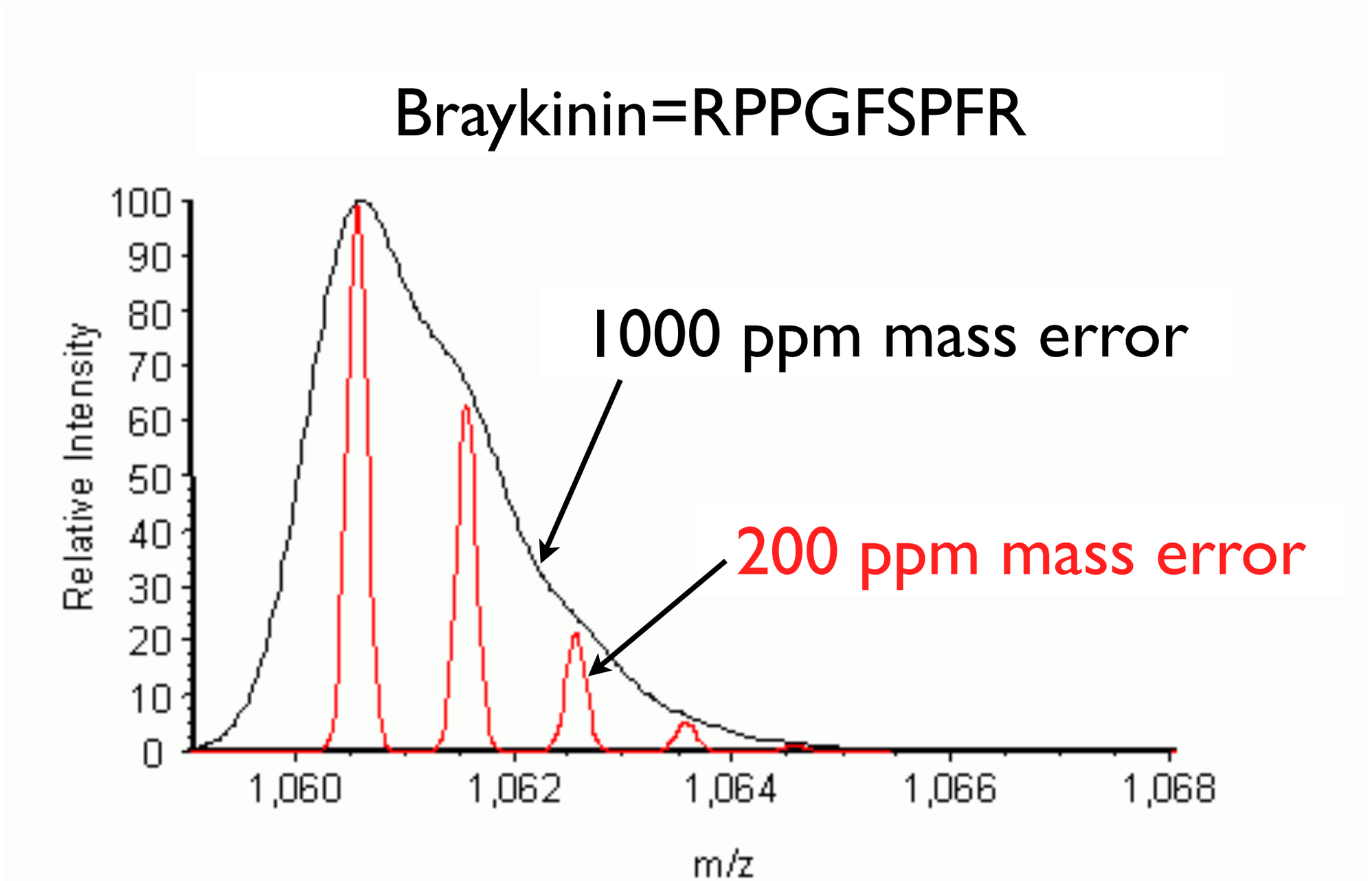


Protein Arrays

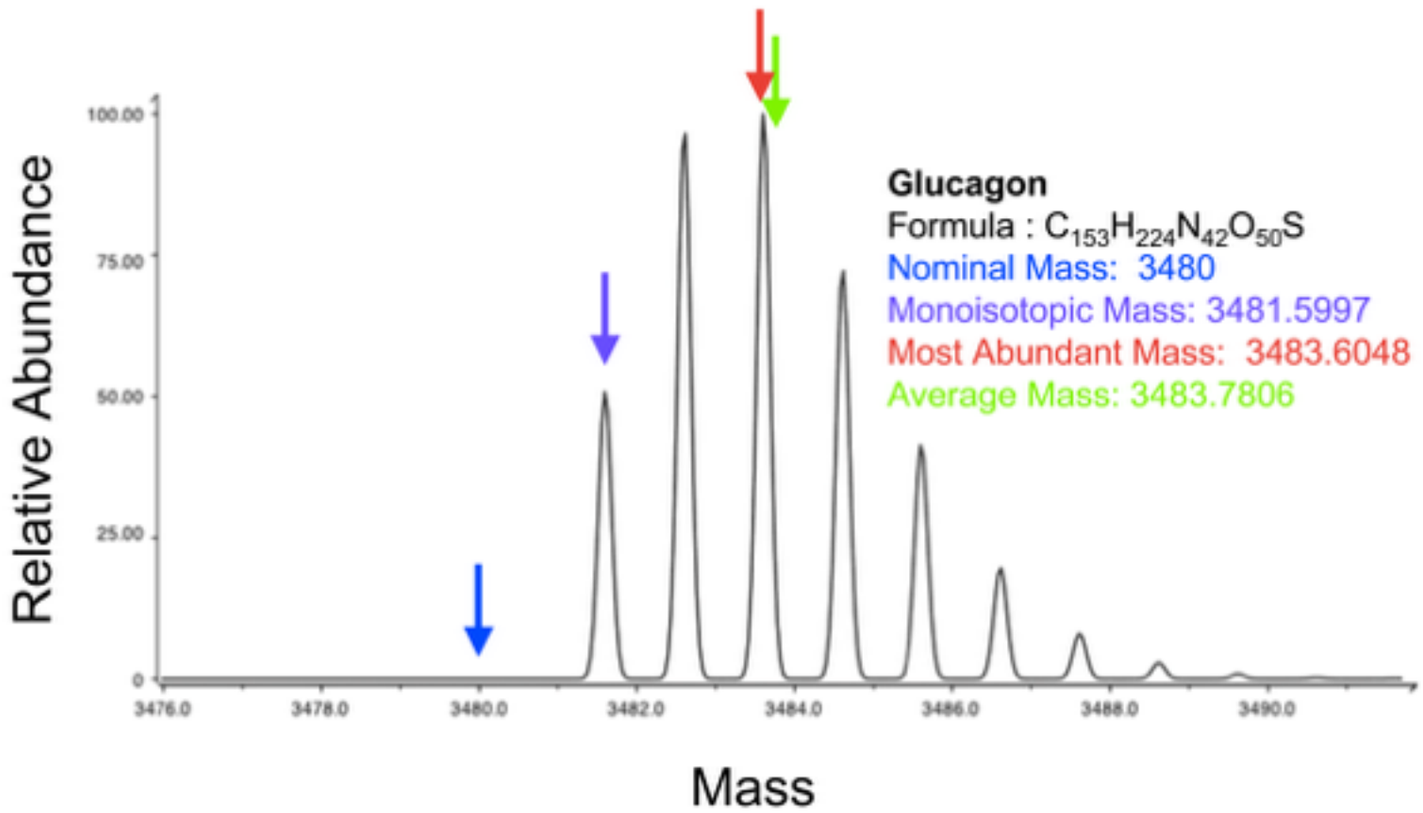


Enzyme-linked immunosorbent assay (ELISA)

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 *The preferred measure for high-resolution MS*
Average Mass	Average Mass given the isotopes and their natural abundance 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.

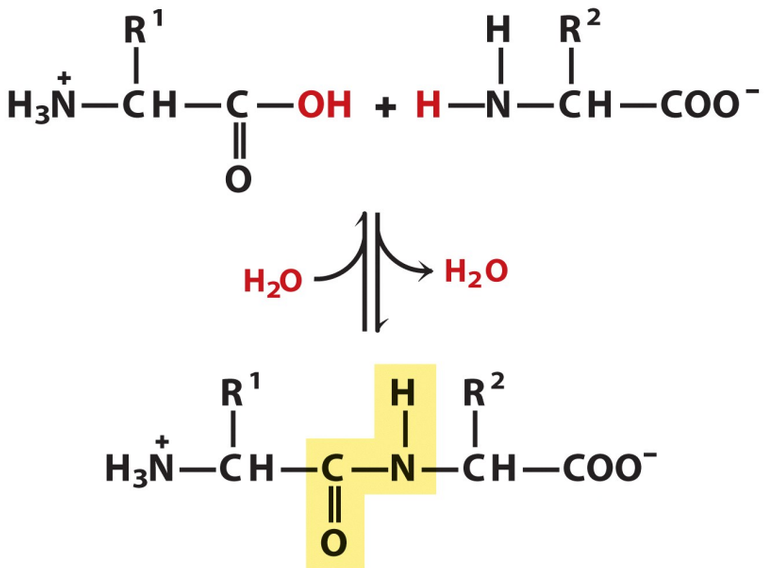
Residue mass of amino acids

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	

etc...

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

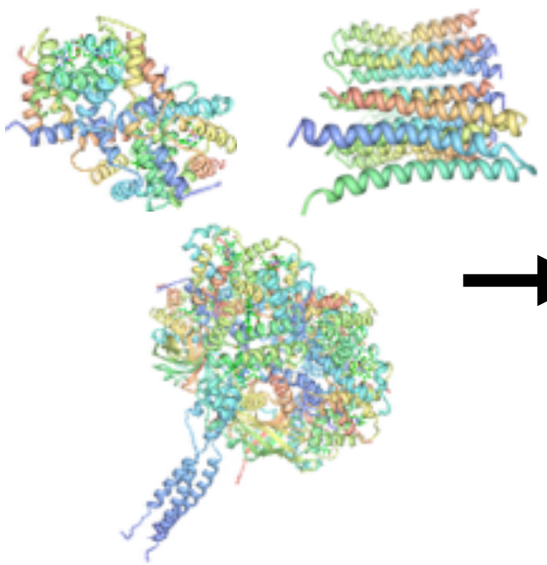
The mass of a peptide

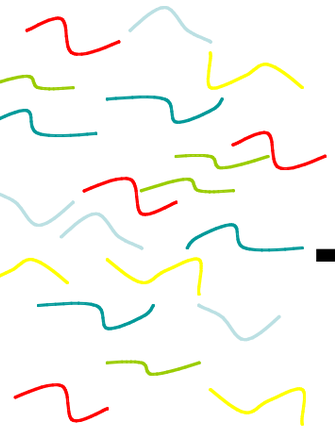


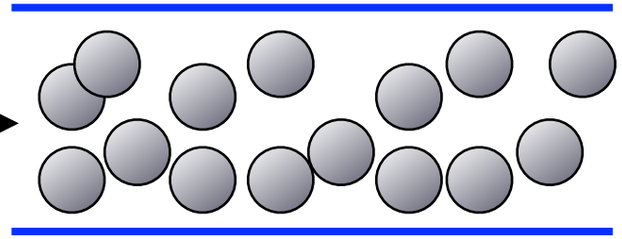
The mass $m(p)$ of peptide p can be calculated as the residue mass of its constituent amino acids, $a_1 \ldots a_n$, and the mass of a water molecule

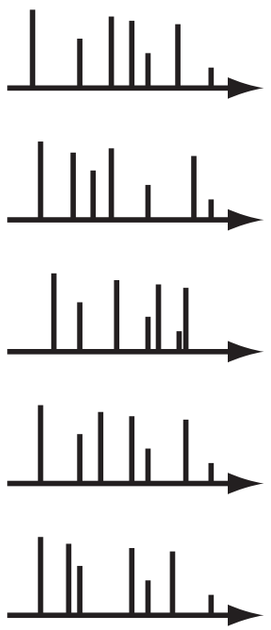
$$m(p) = m(H_2O) + \sum_{i=1 \ldots n} m(a_i)$$

Shotgun proteomics









IAMAPEPTIDER

MGREATMATCHK

ATRYIDENTIFYMEK

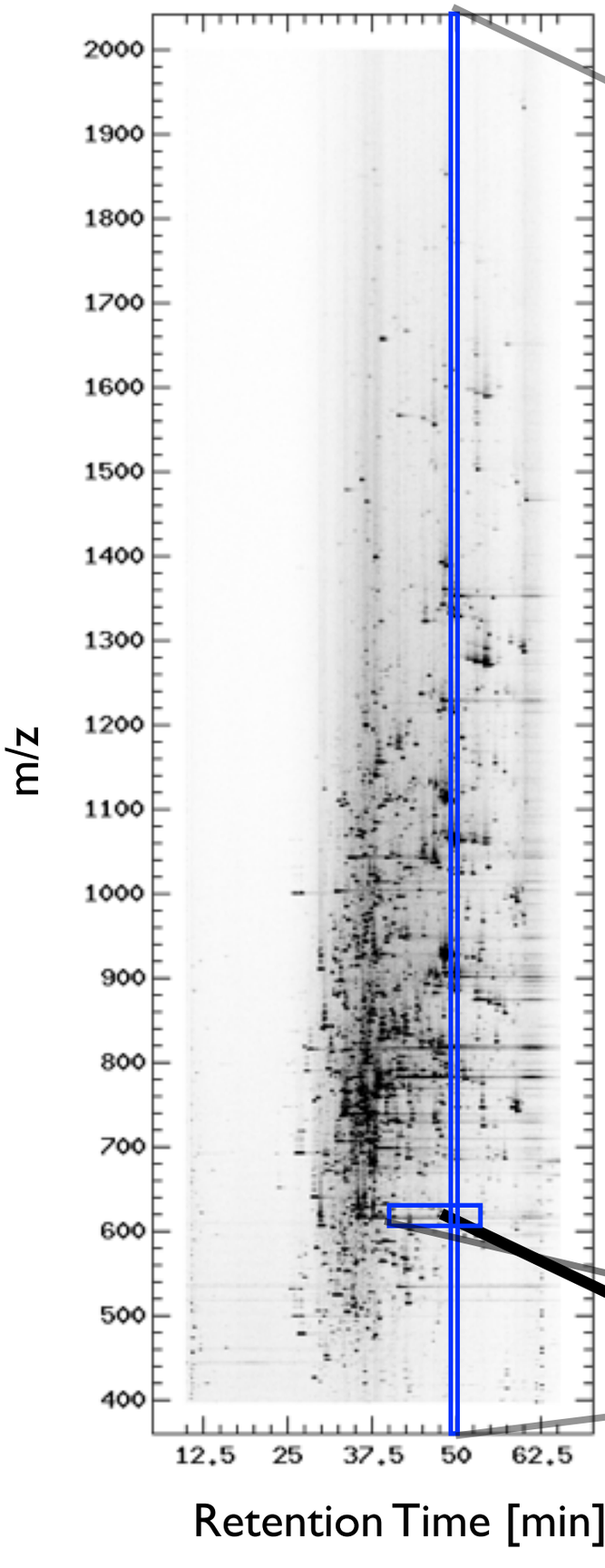
LMAKEMYDAYR

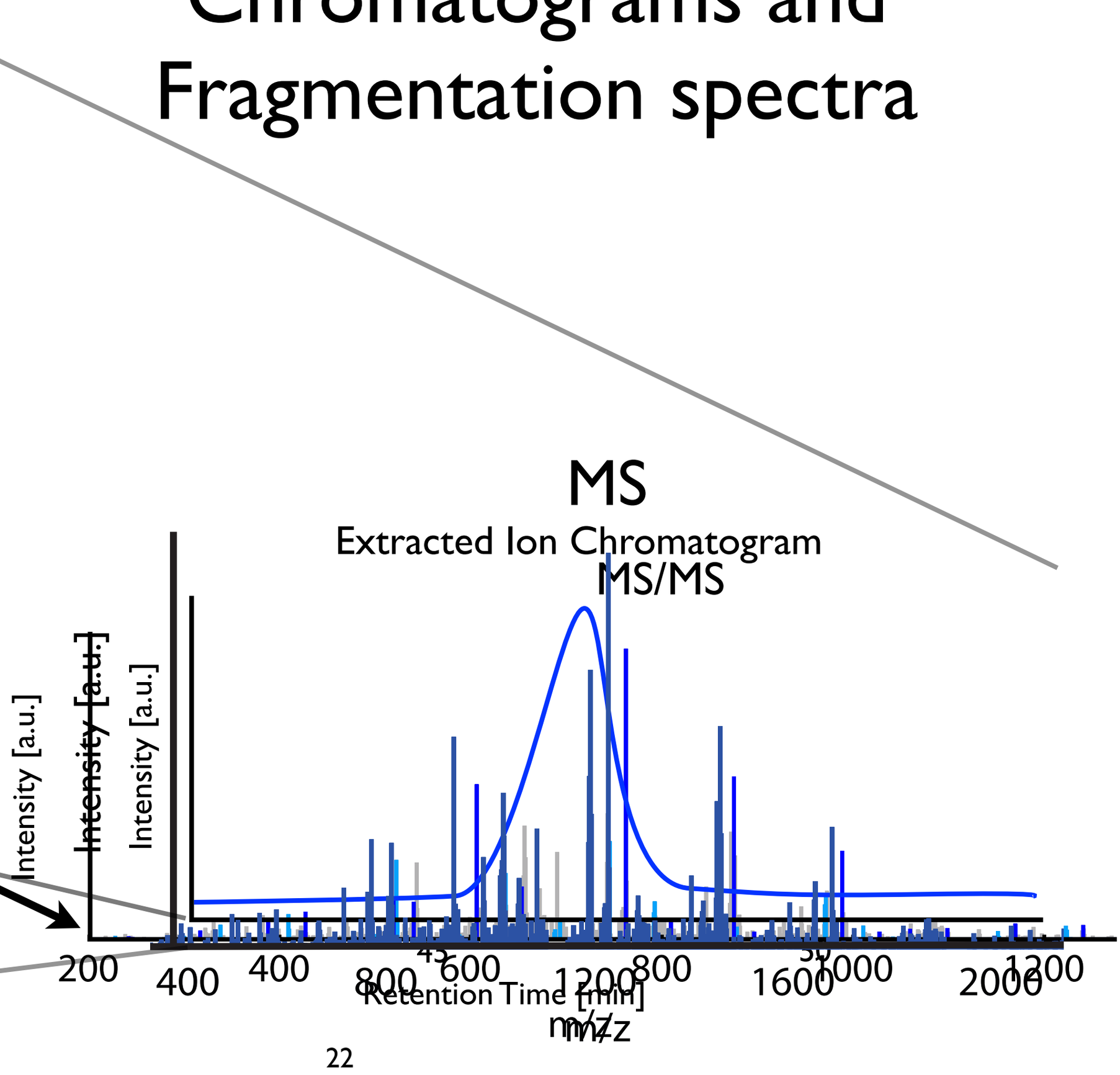
GLB4_LUMTE

COX1_LUMTE

ATP6_LUMTE

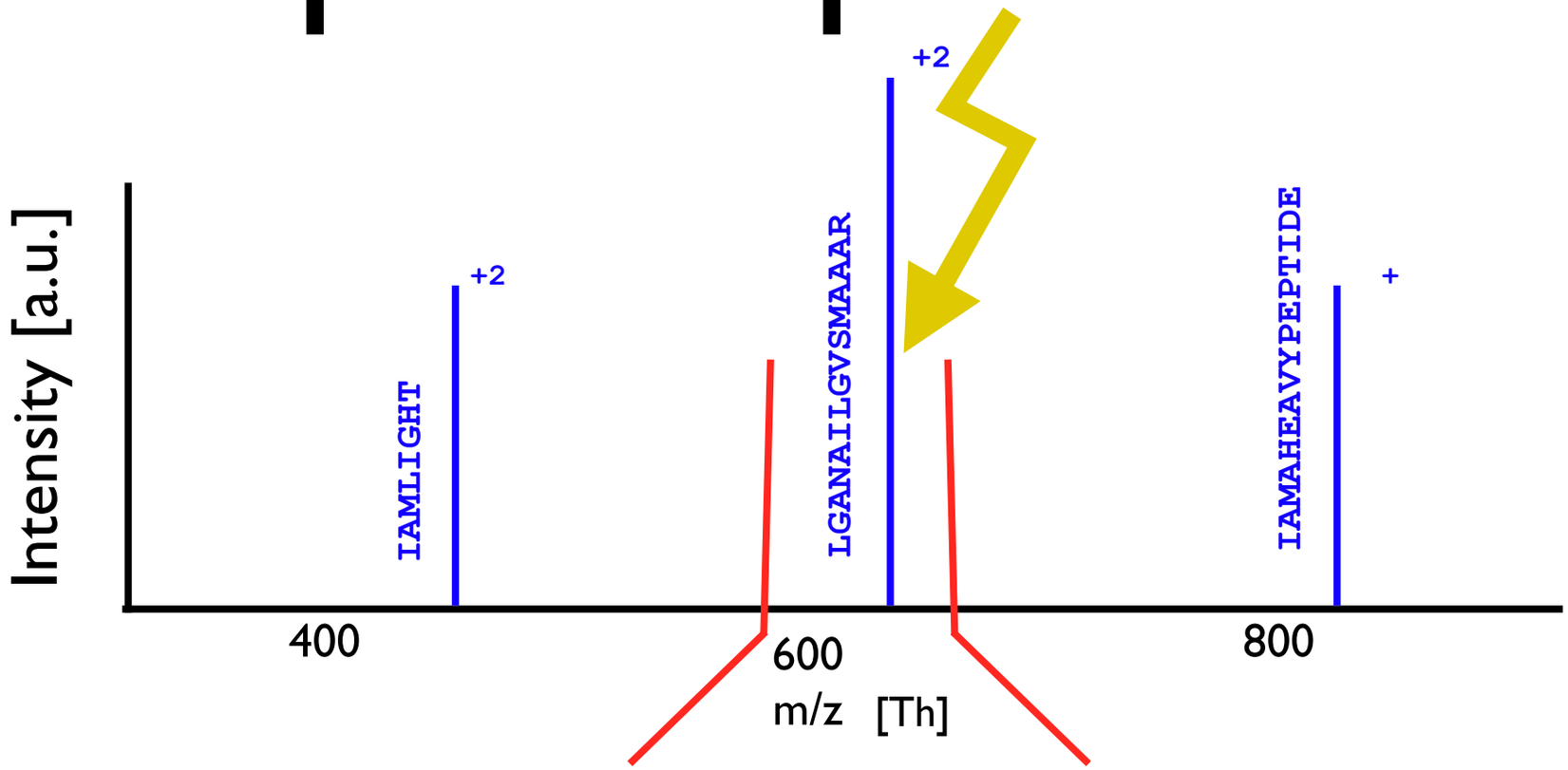
Chromatograms and Fragmentation spectra



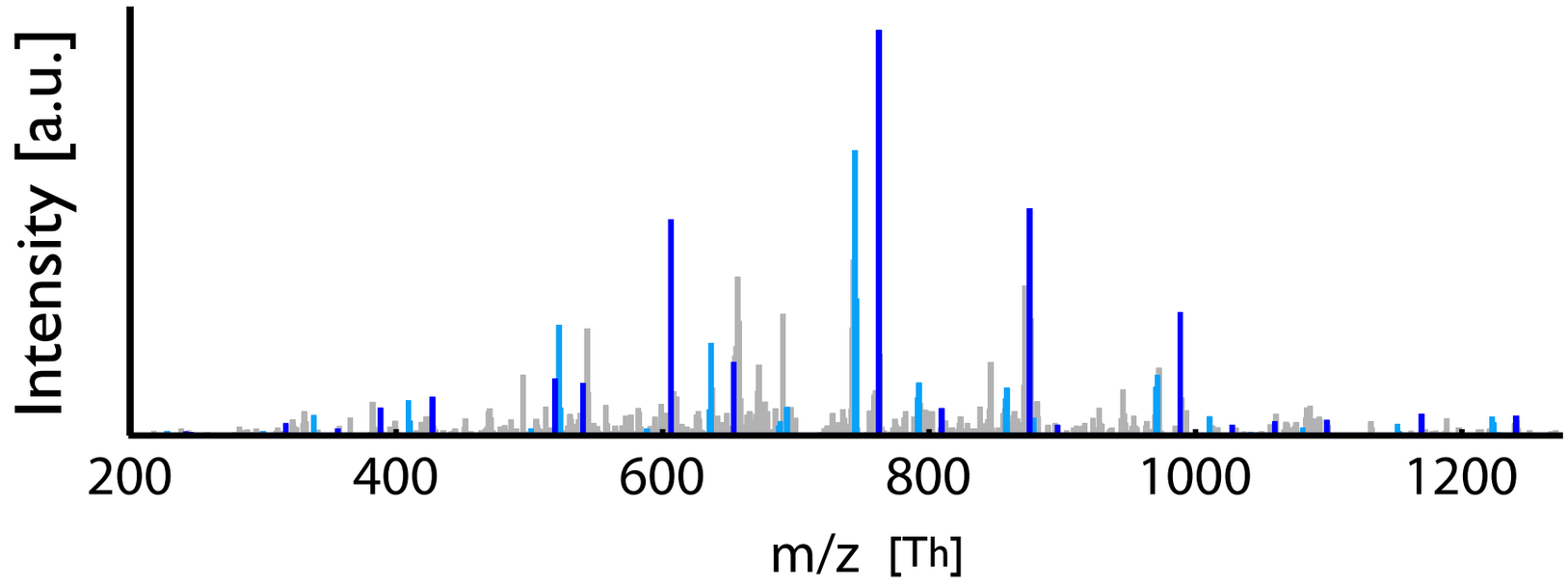


Peptide spectra

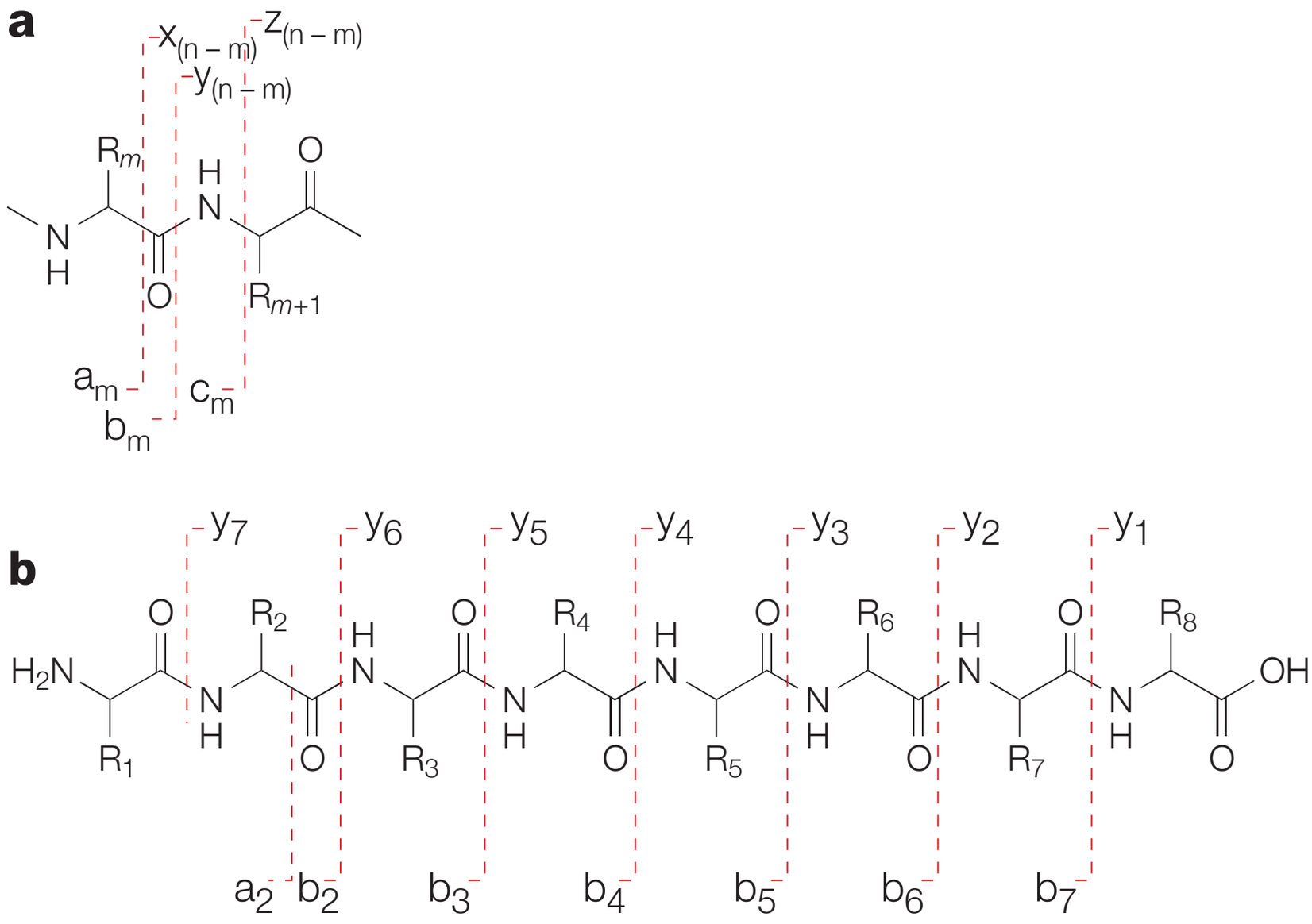
MS¹



MS²



Peptide Fragmentation

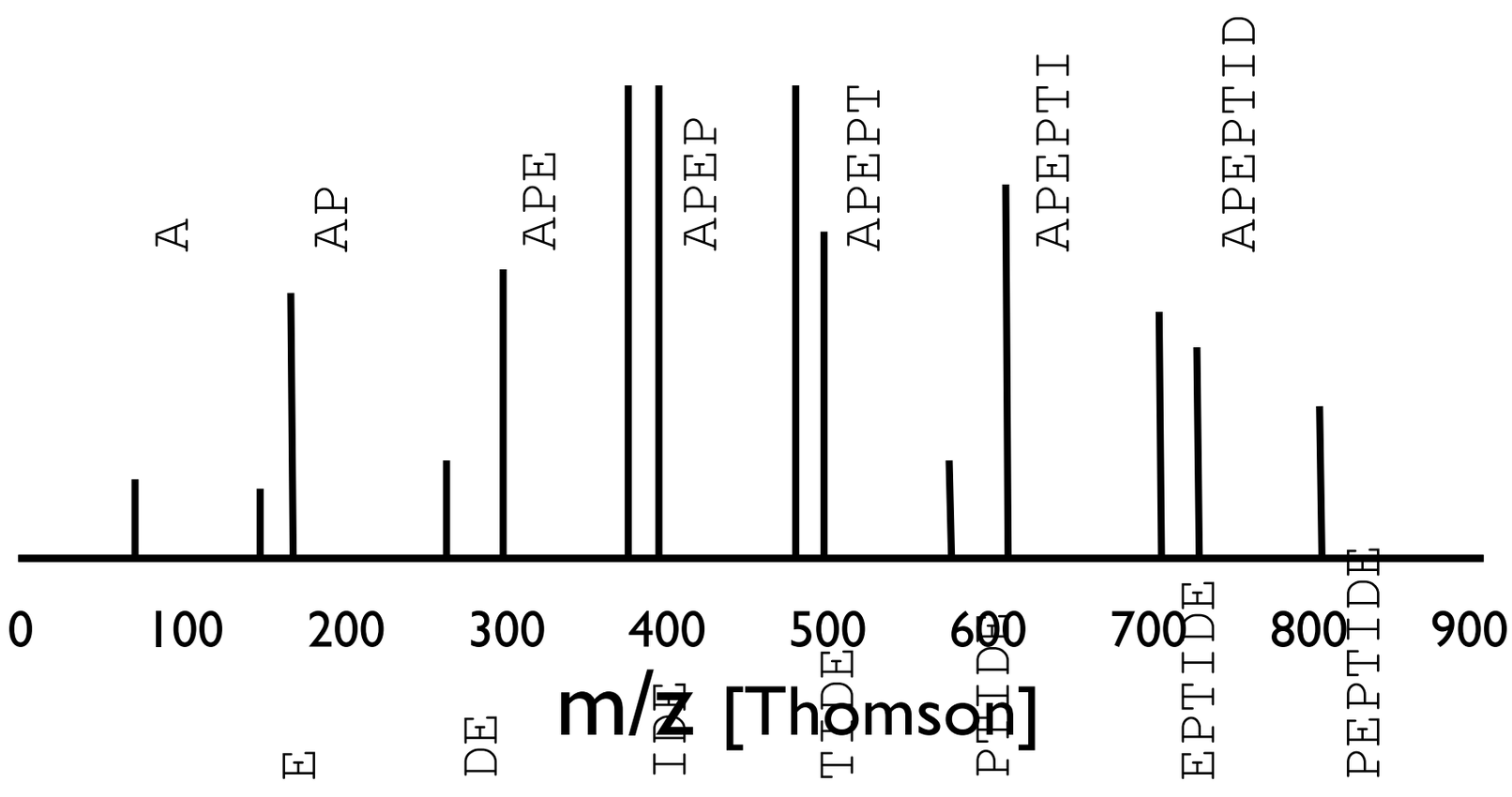


Fragmentation Spectrum

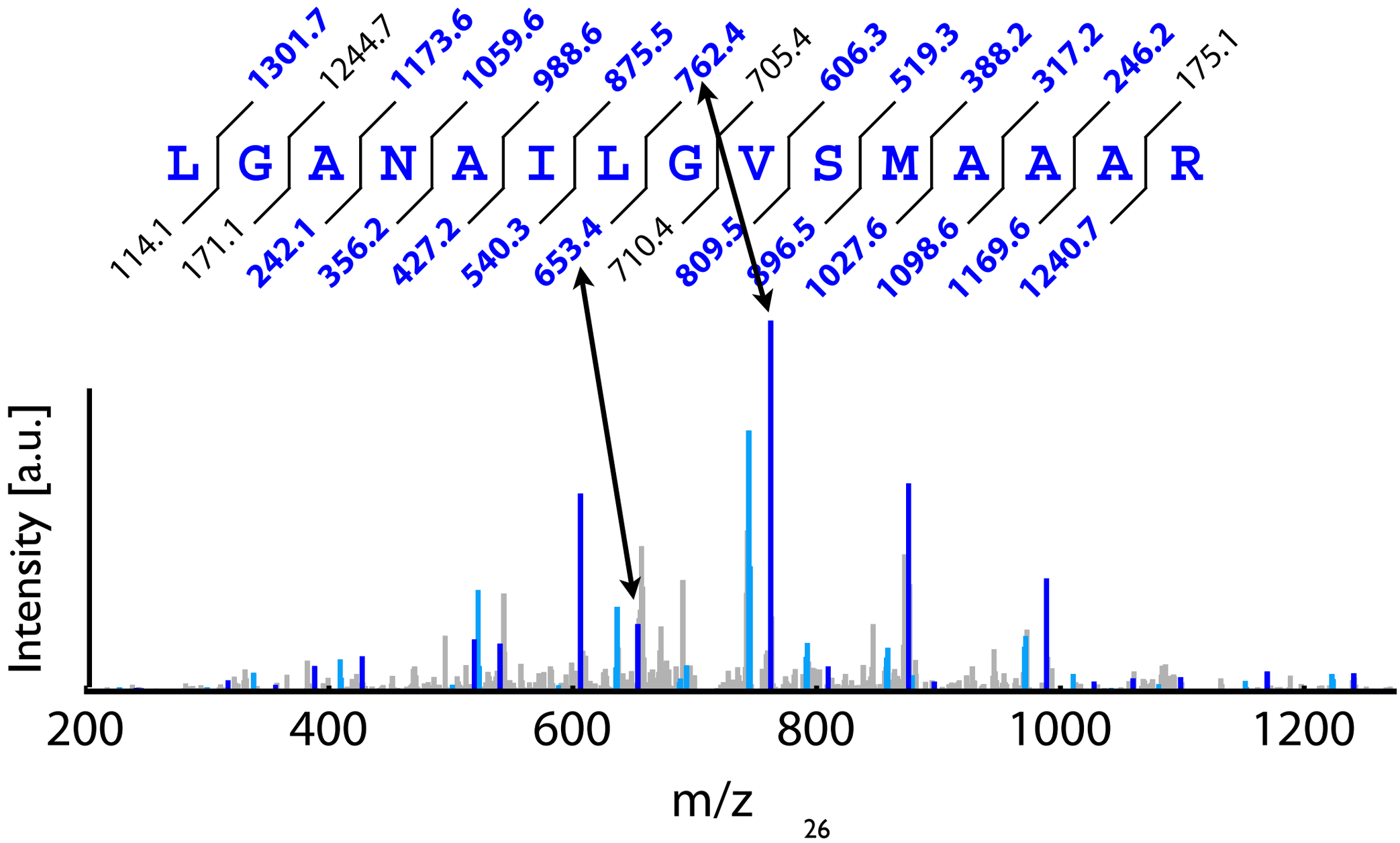
A|P|E|P|T|I|D|E

b:

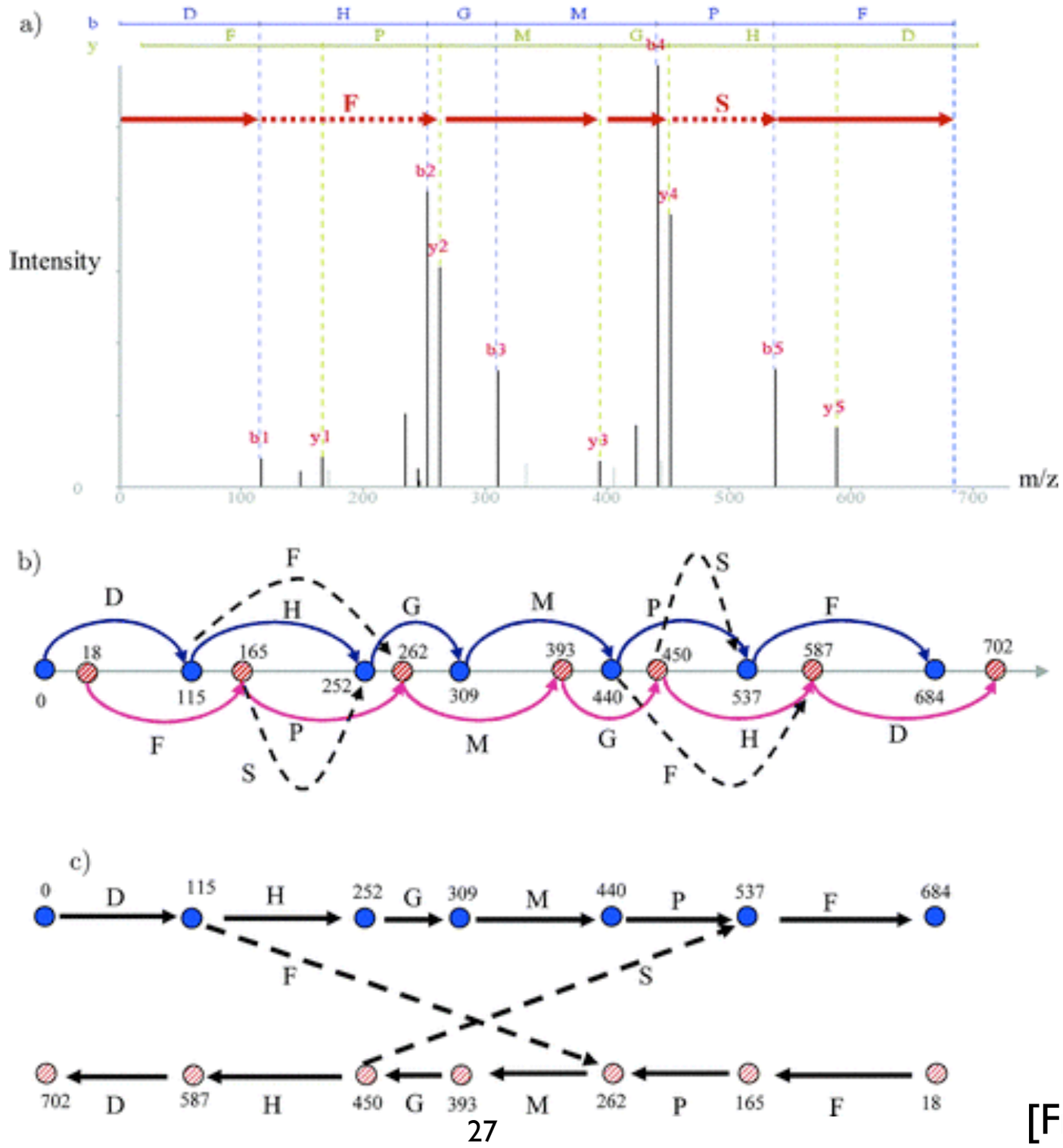
y:



Peptide fragmentation spectrum

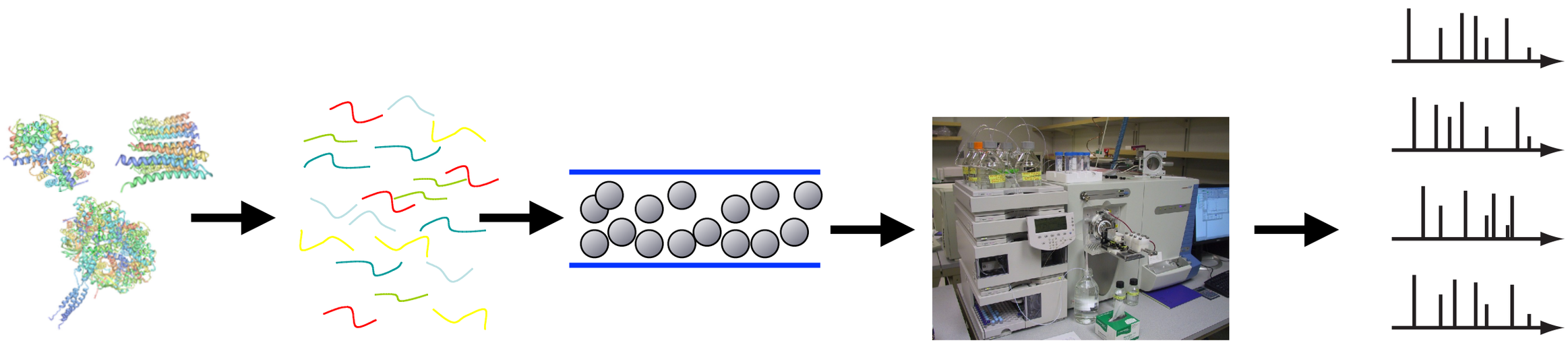


de Novo sequencing

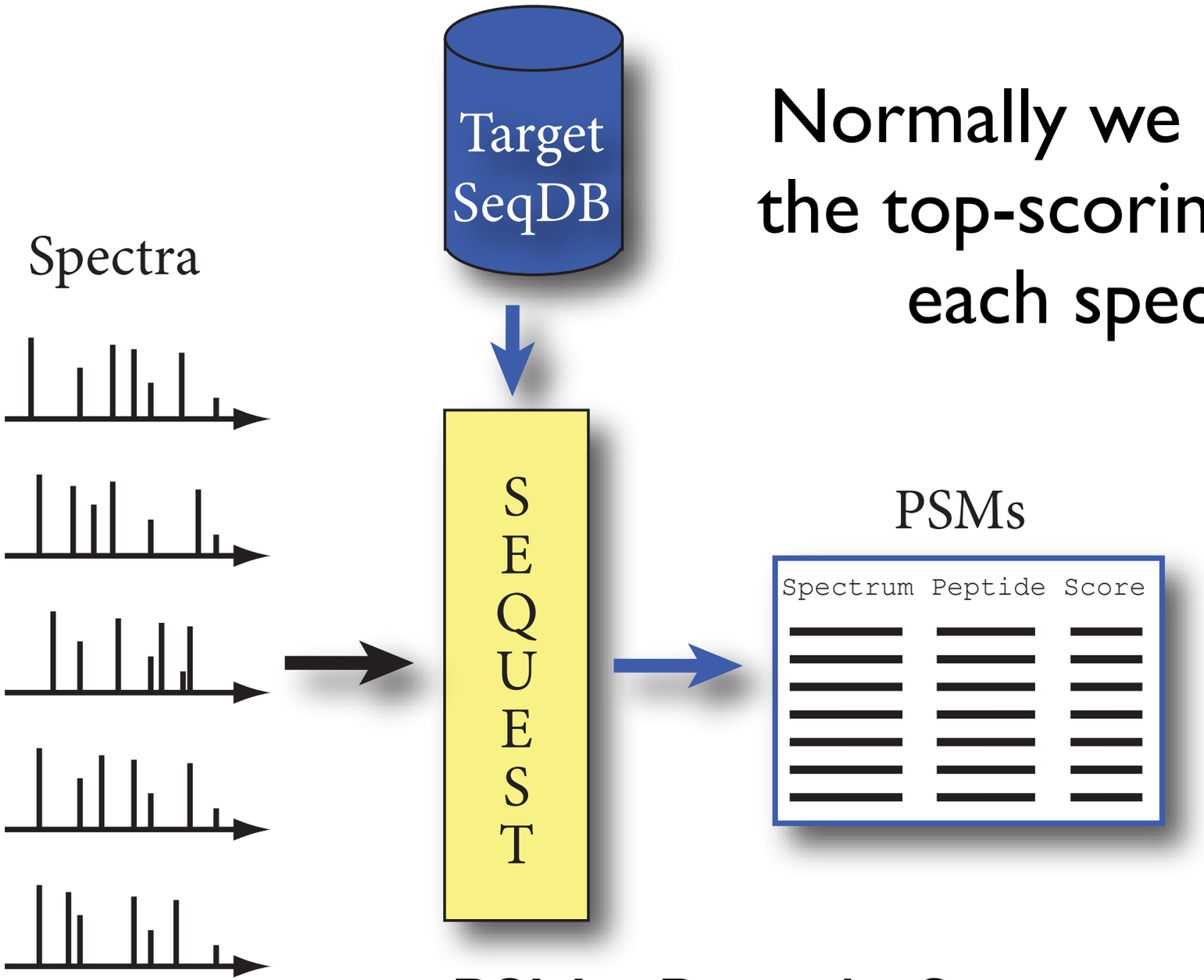


[Frank *et al.* JPR 2007]

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

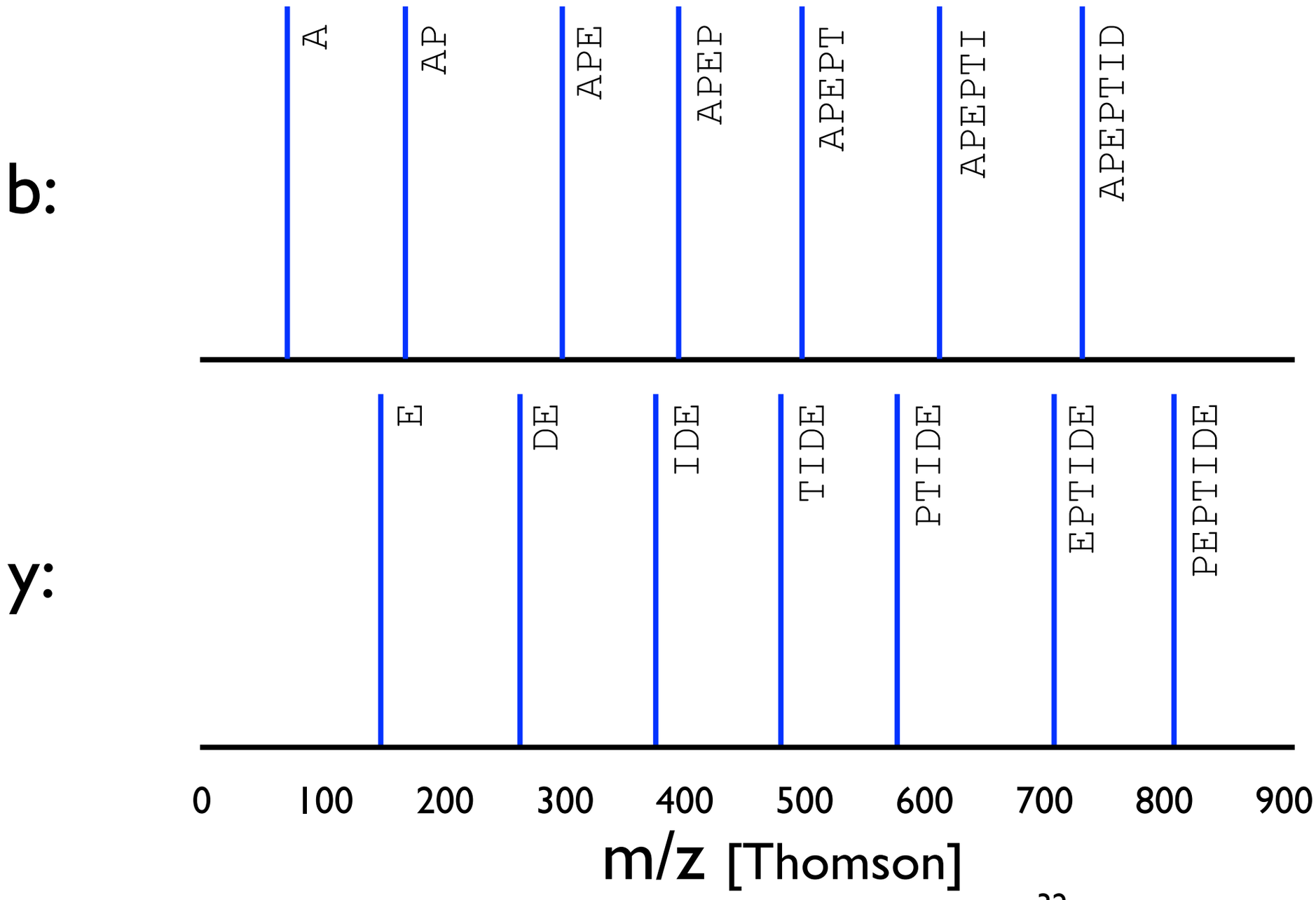
$$R_i = \sum_{j=1}^{n} x_j y_{j+i}$$

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

- 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

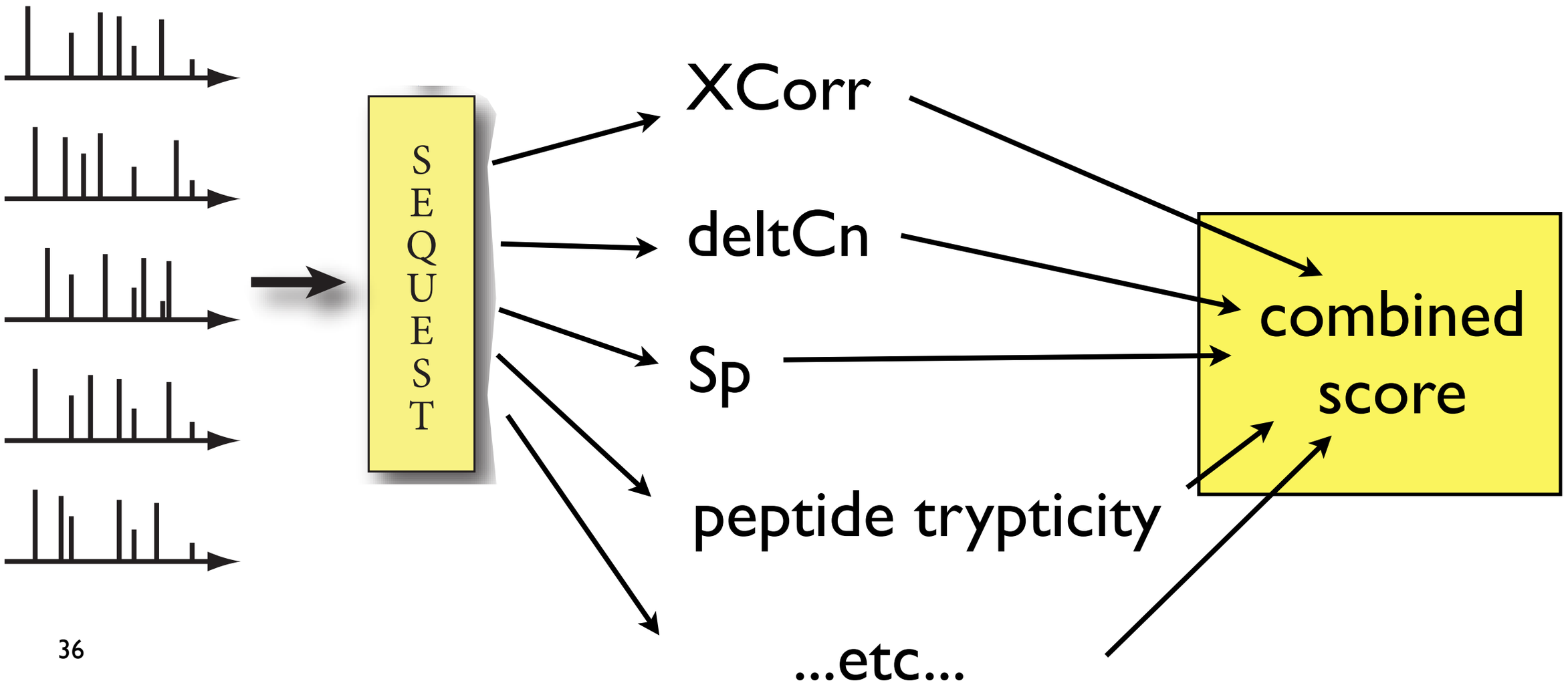
Present 0/1

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

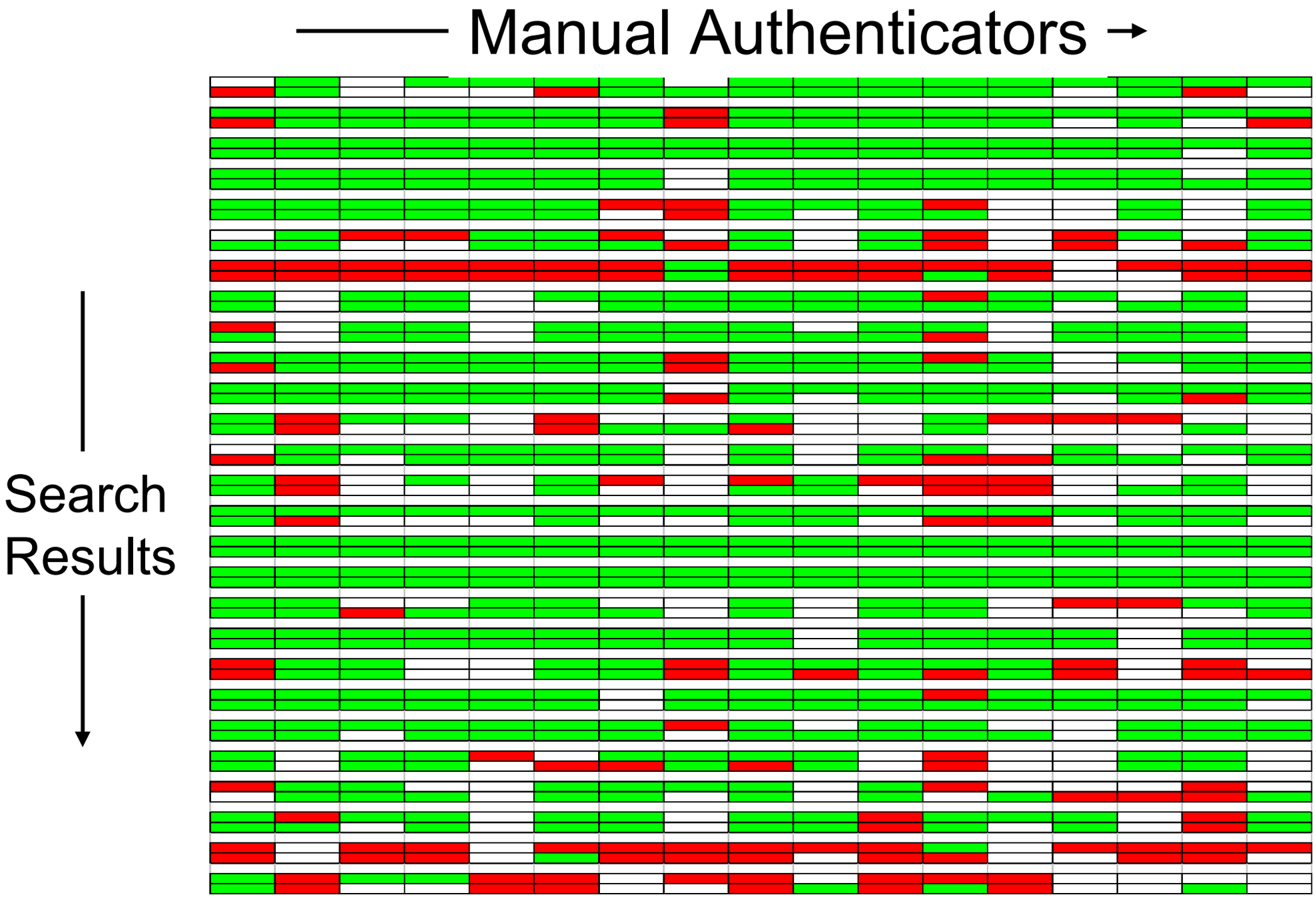
Post Processors

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]

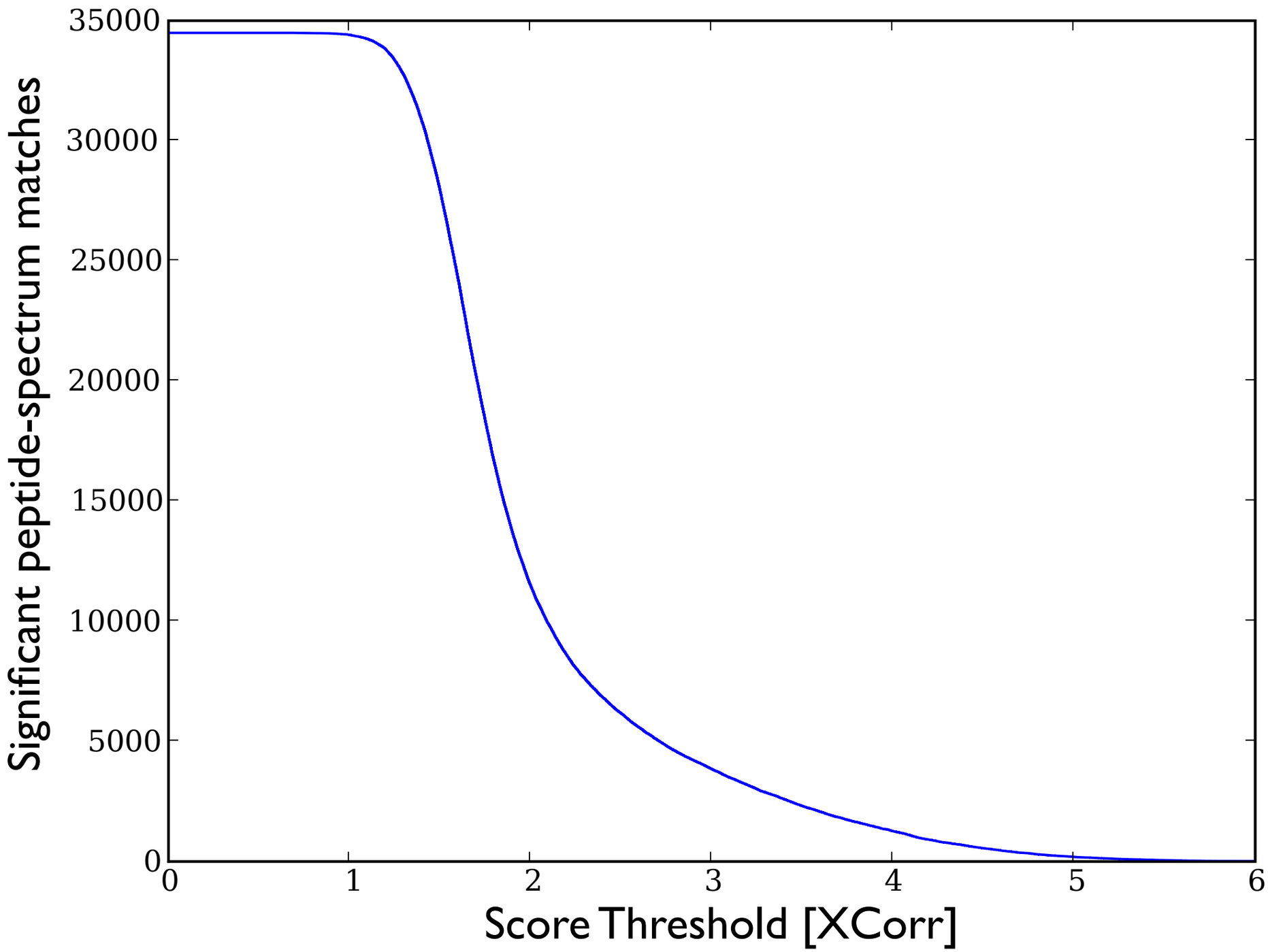


(Un)reliability of manual validation



[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

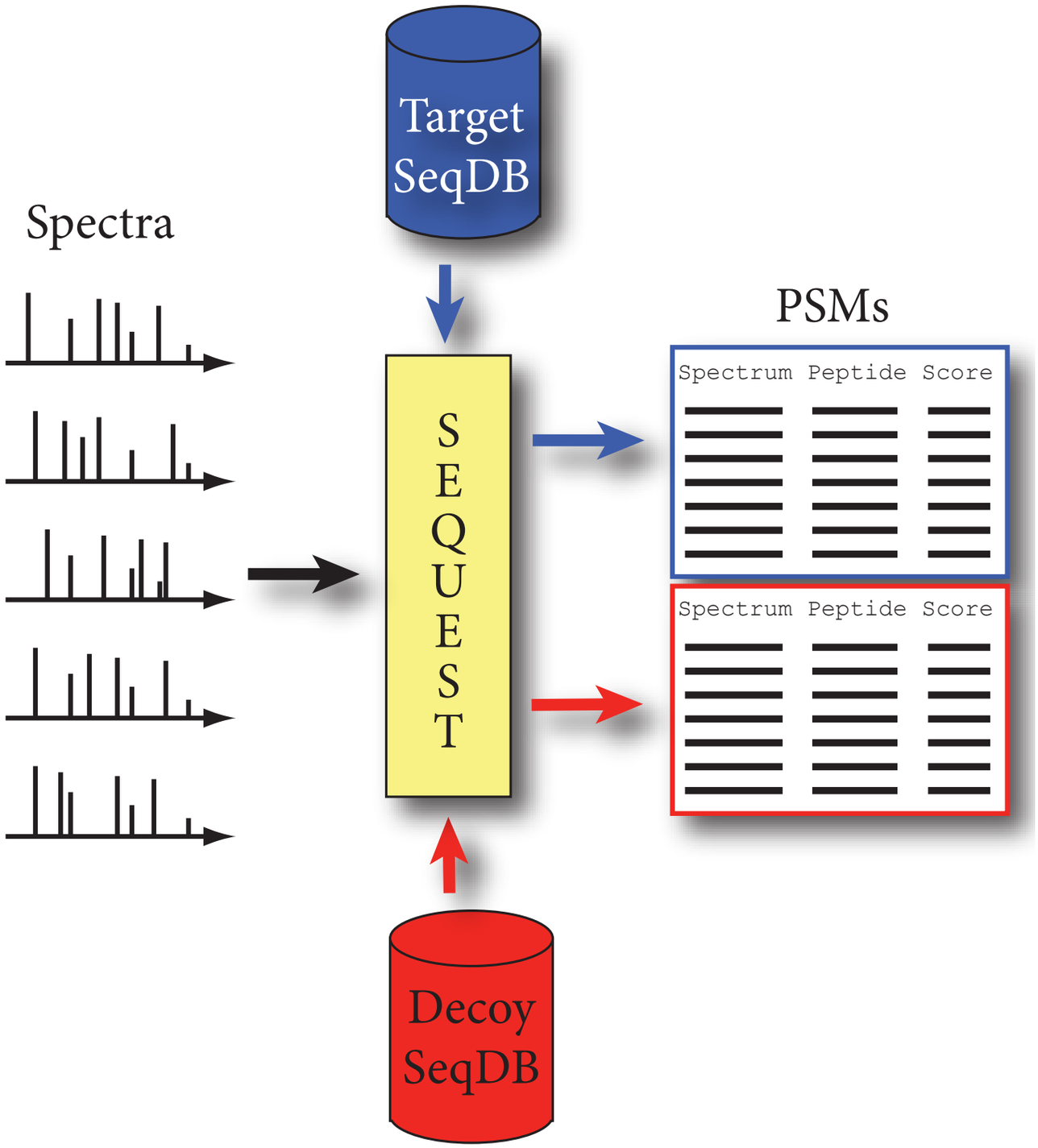
control for ...

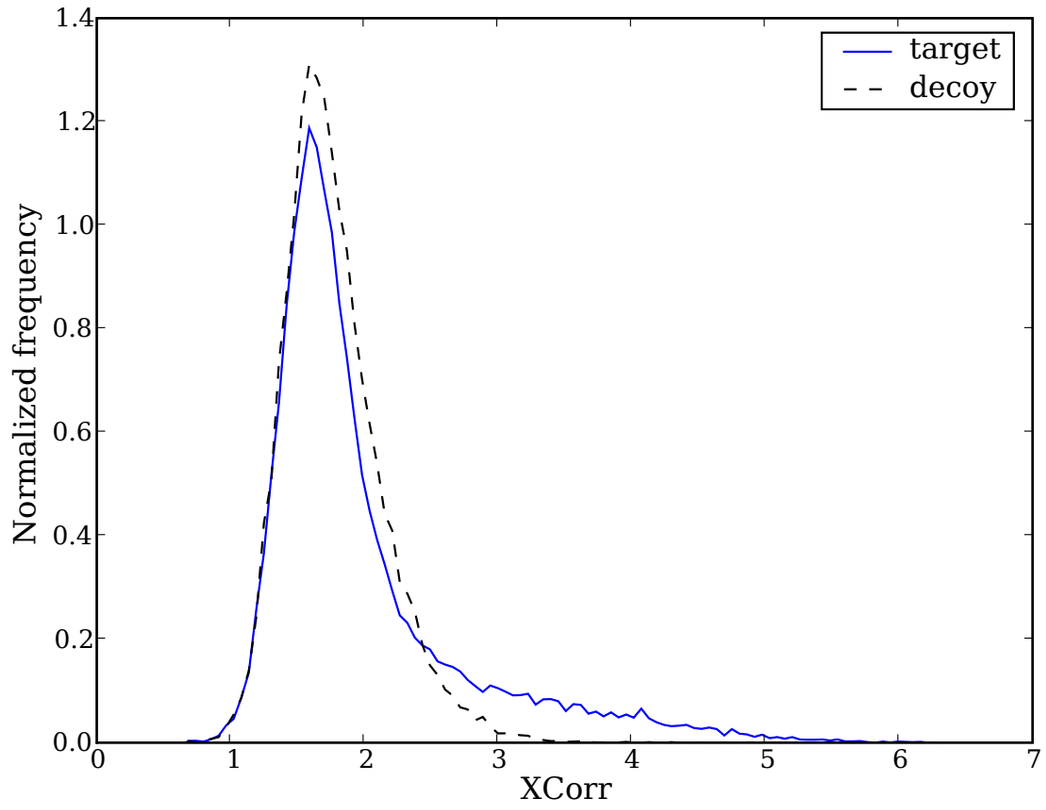
... FDR or q value when you are interested in identifying a set of PSMs

... 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





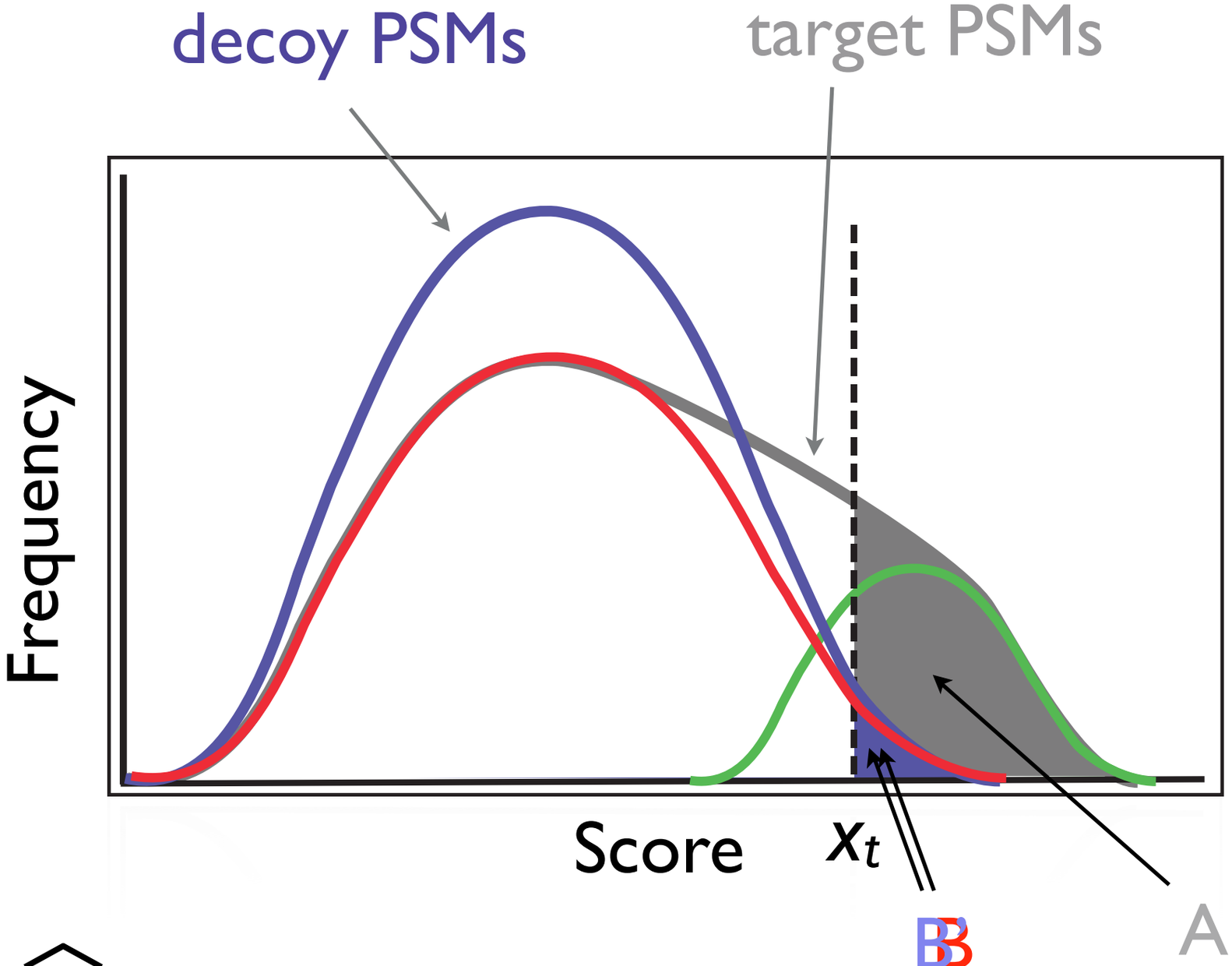
[Moore et al. JASMS 2002]

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



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

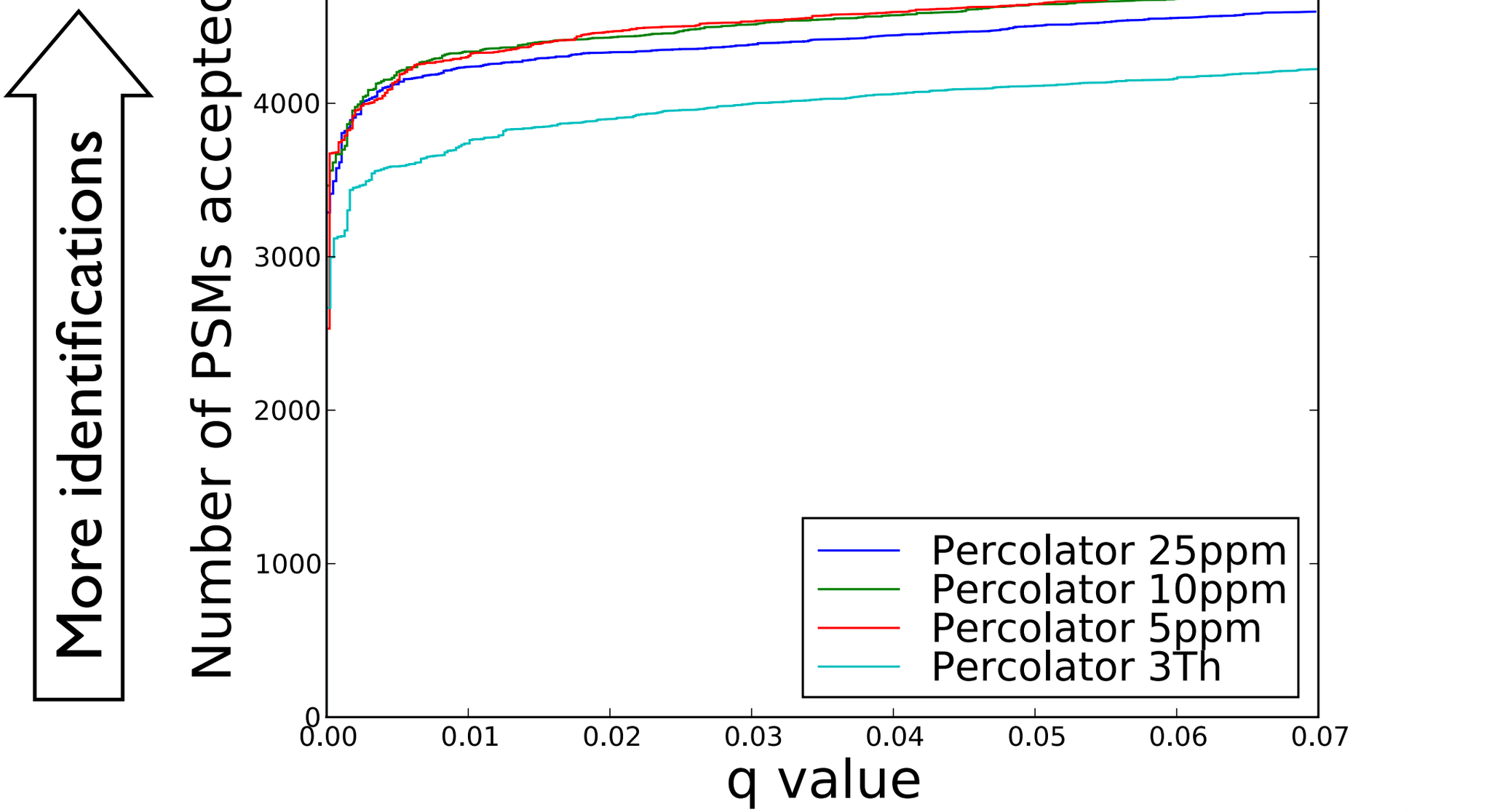
$$\widehat{FDR} = \frac{B}{A} = \frac{\widehat{\pi_0} \cdot B'}{A}$$

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

[Käll *et al.* JPR 2008]

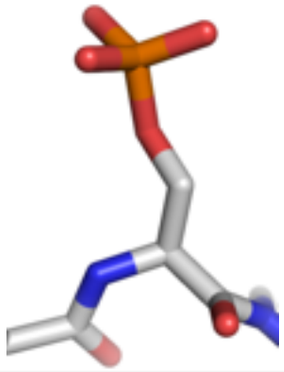
$\widehat{\pi_0}$ is the prior probability that a target PSM is incorrectly matched

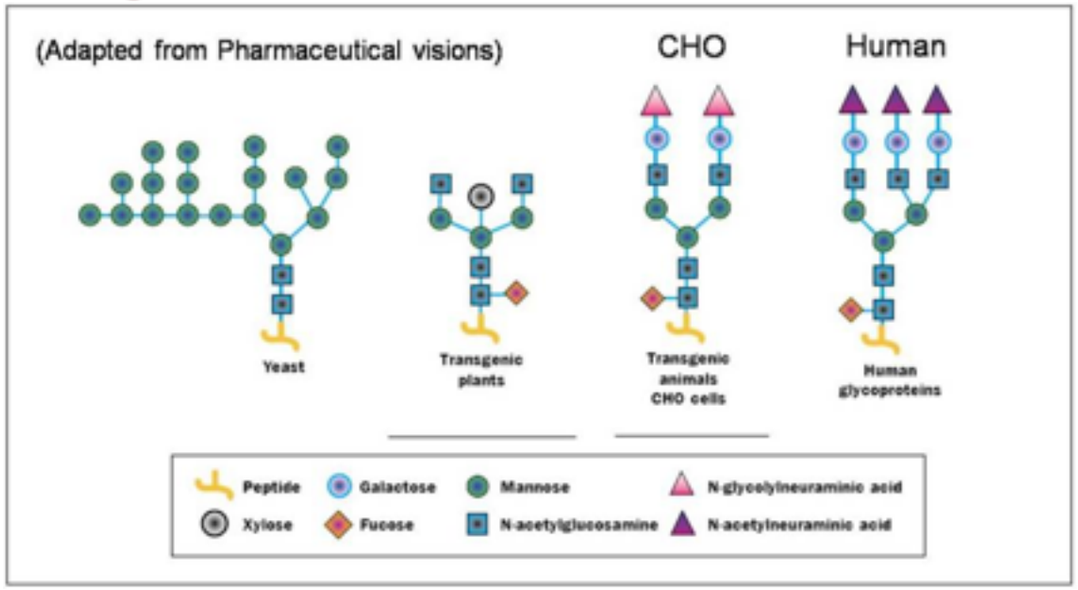
q-p plot

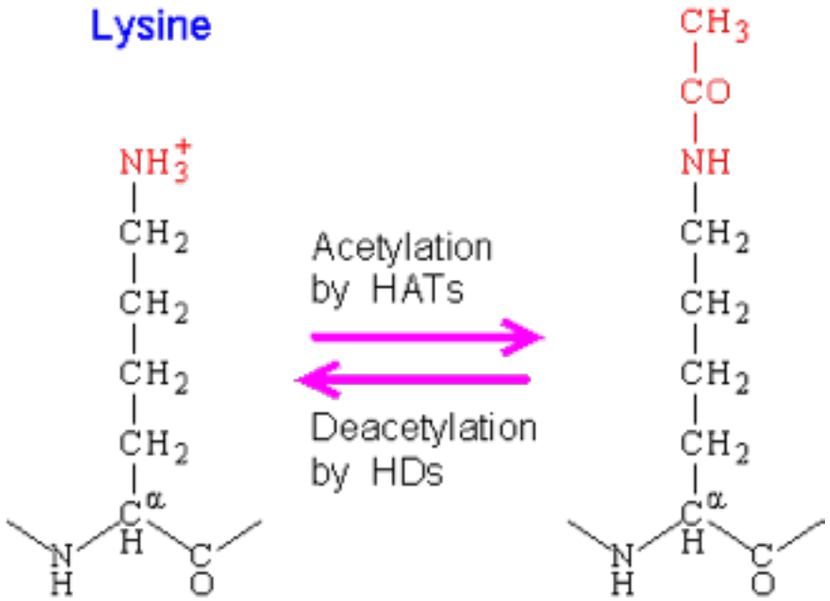

More identifications
Number of PSMs accepted
5000
4000
3000
2000
1000
0
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
q value
Percolator 25ppm
Percolator 10ppm
Percolator 5ppm
Percolator 3Th
Higher fraction incorrect identifications

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

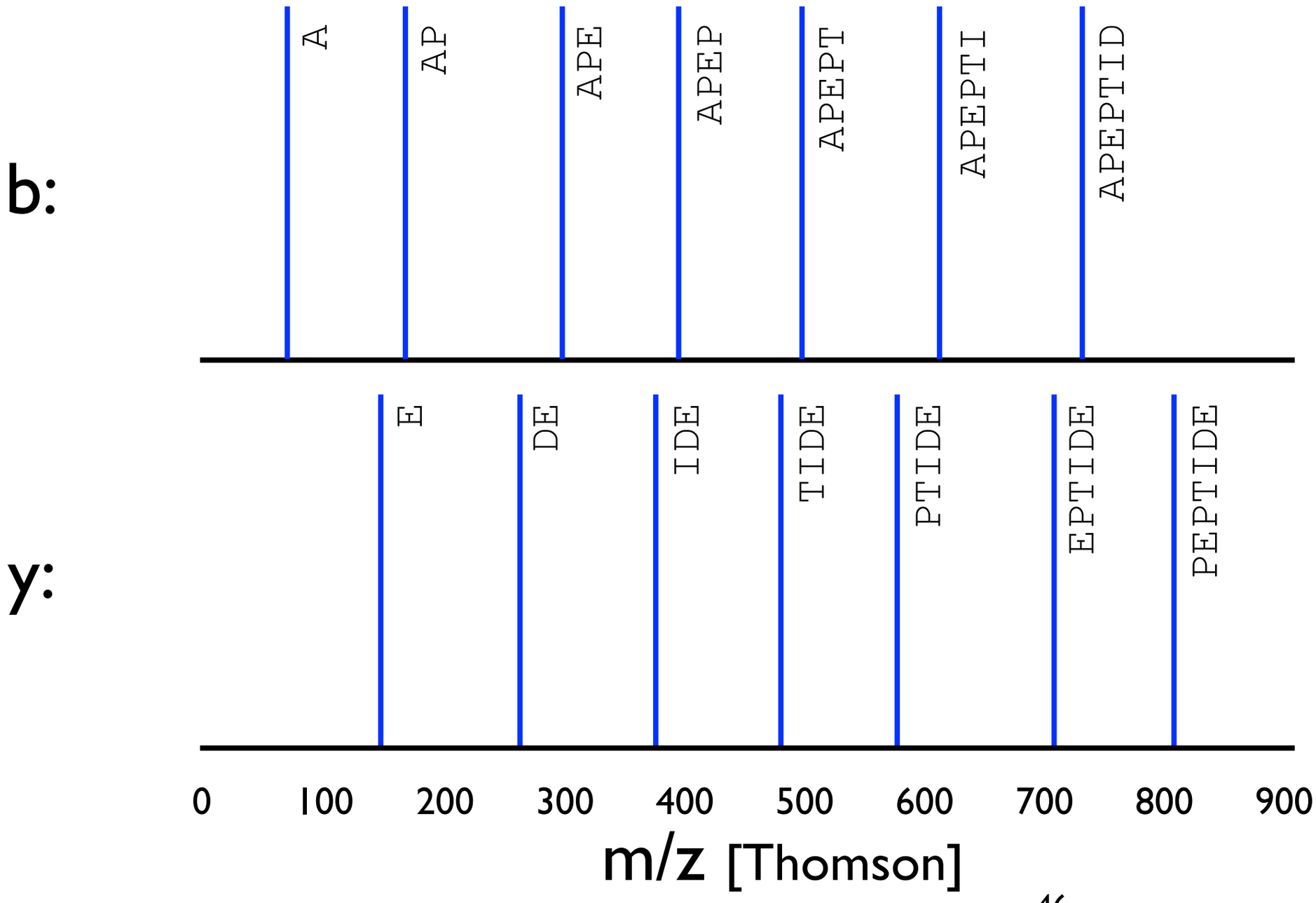






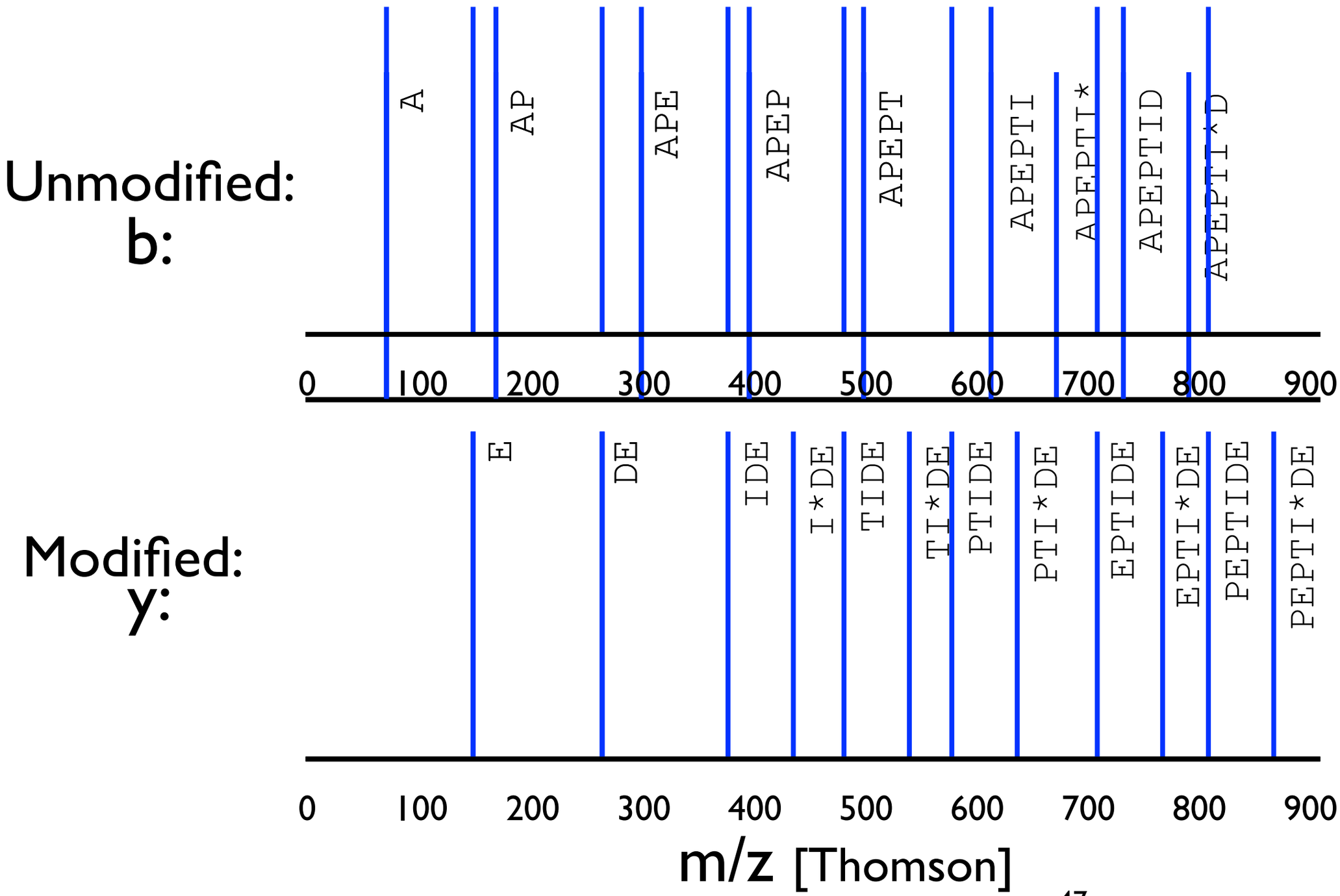
Theoretical Spectrum of a peptide

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

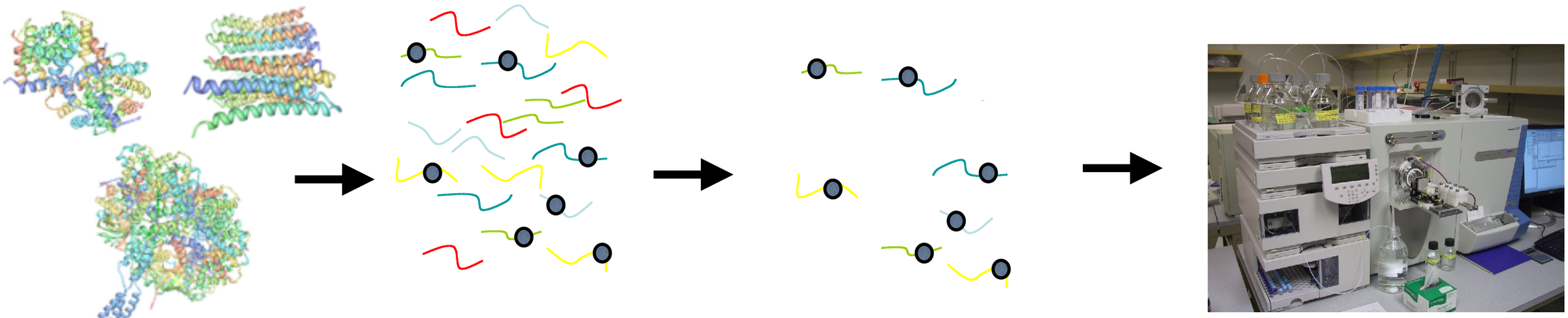


Theoretical Spectrum of a PTM Peptide

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
```

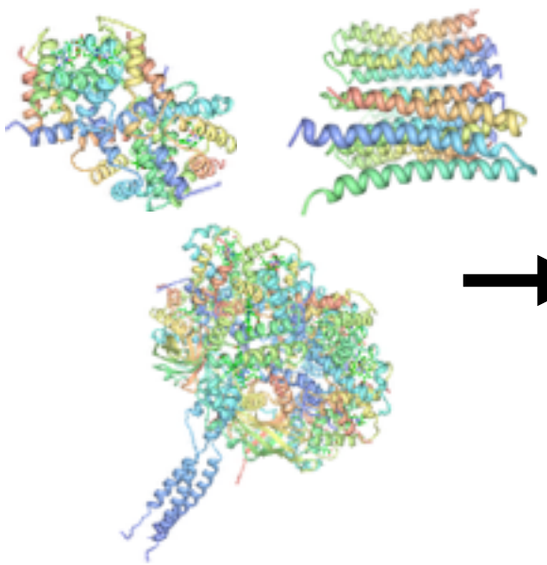
Peptide Spectrum Matches

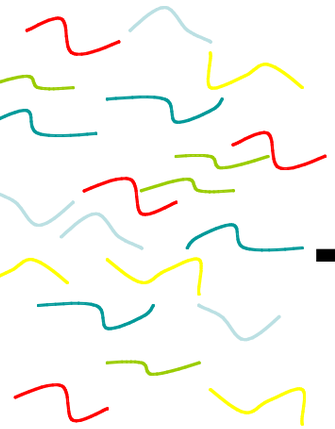
- XML-based: pepXML, mzIdentML
- tab delimited: .sqt

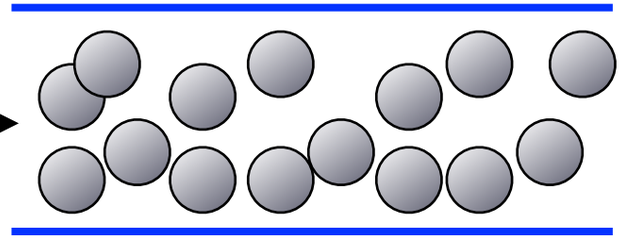
A nice file format converter - Proteowizard

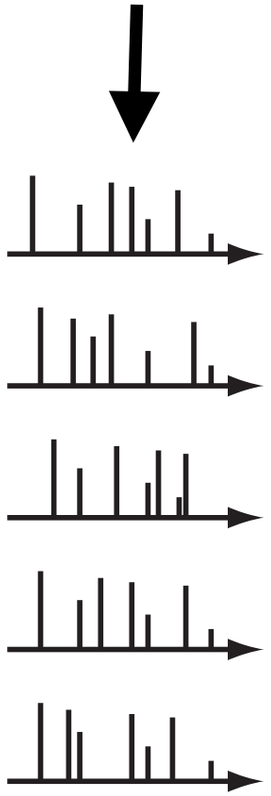
Inferring proteins

Shotgun proteomics





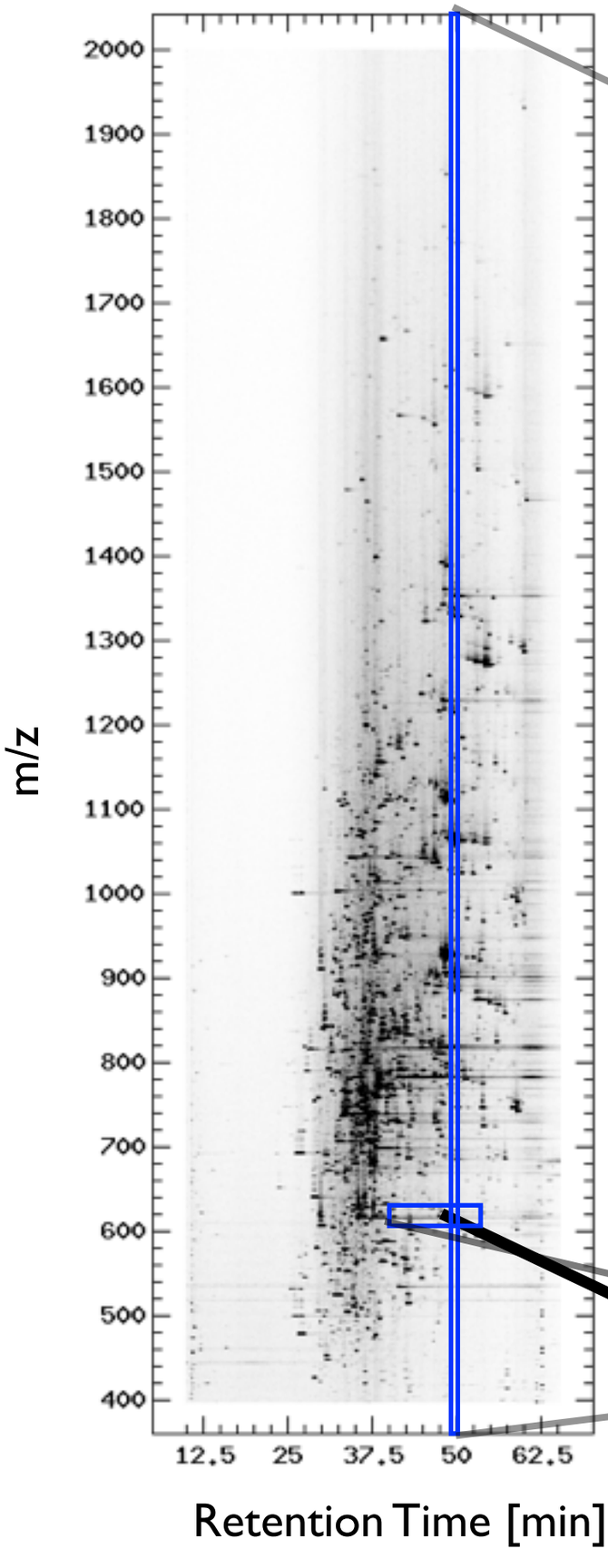


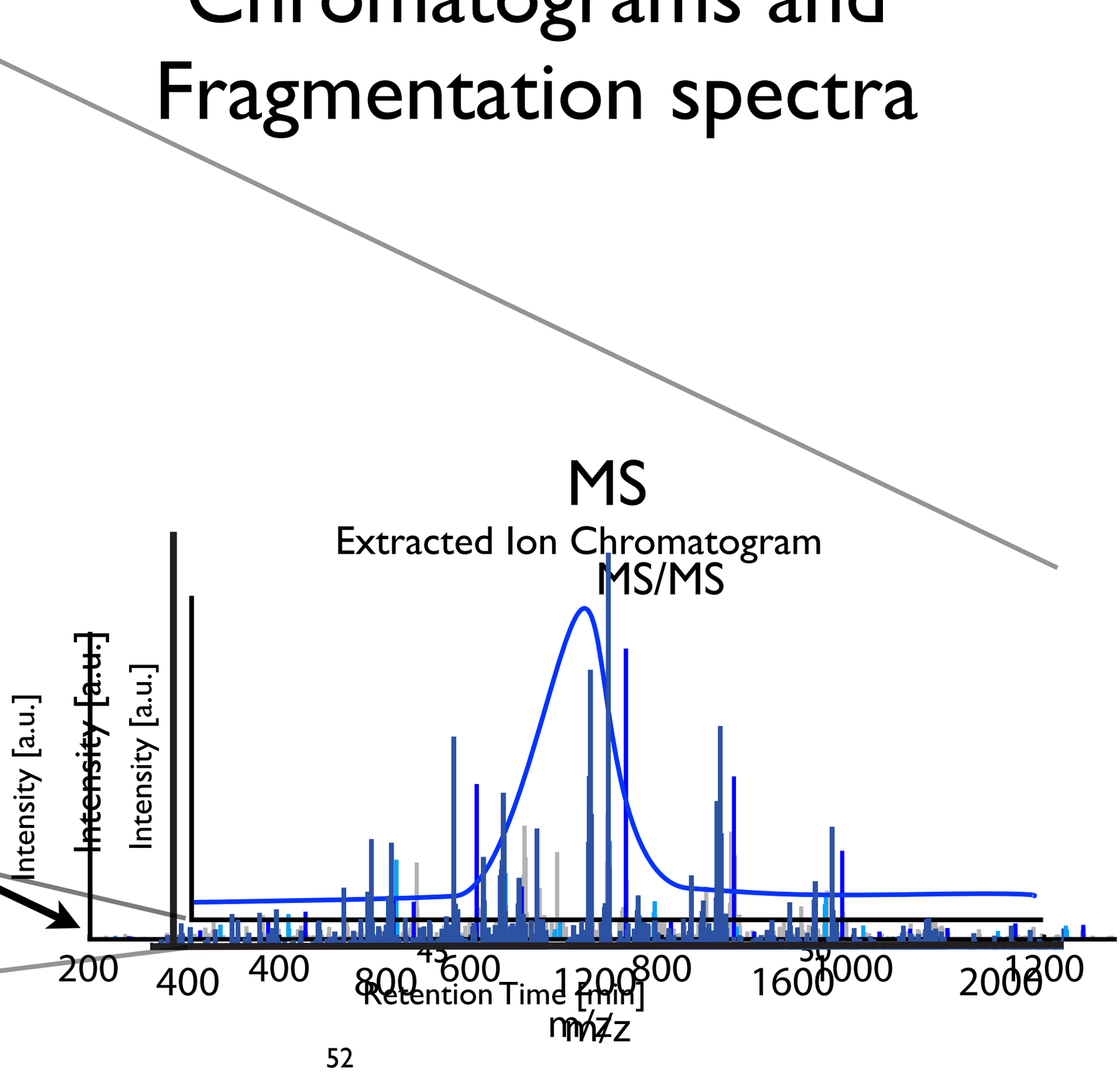


IAMAPEPTIDER
MGREATMATCHK
ATRYIDENTIFYMEK
LMAKEMYDAYR

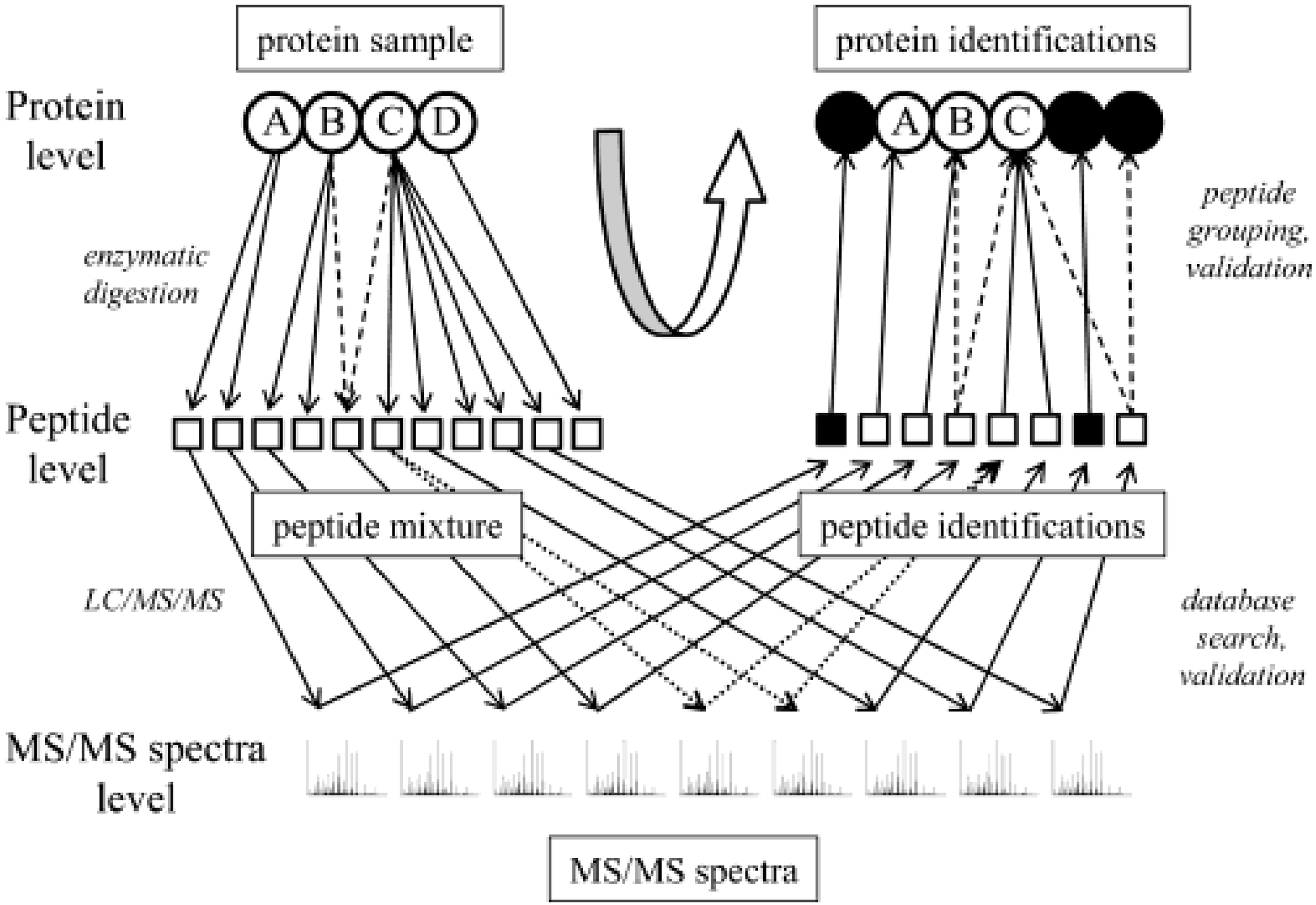
GLB4_LUMTE
COX1_LUMTE
ATP6_LUMTE

Chromatograms and Fragmentation spectra



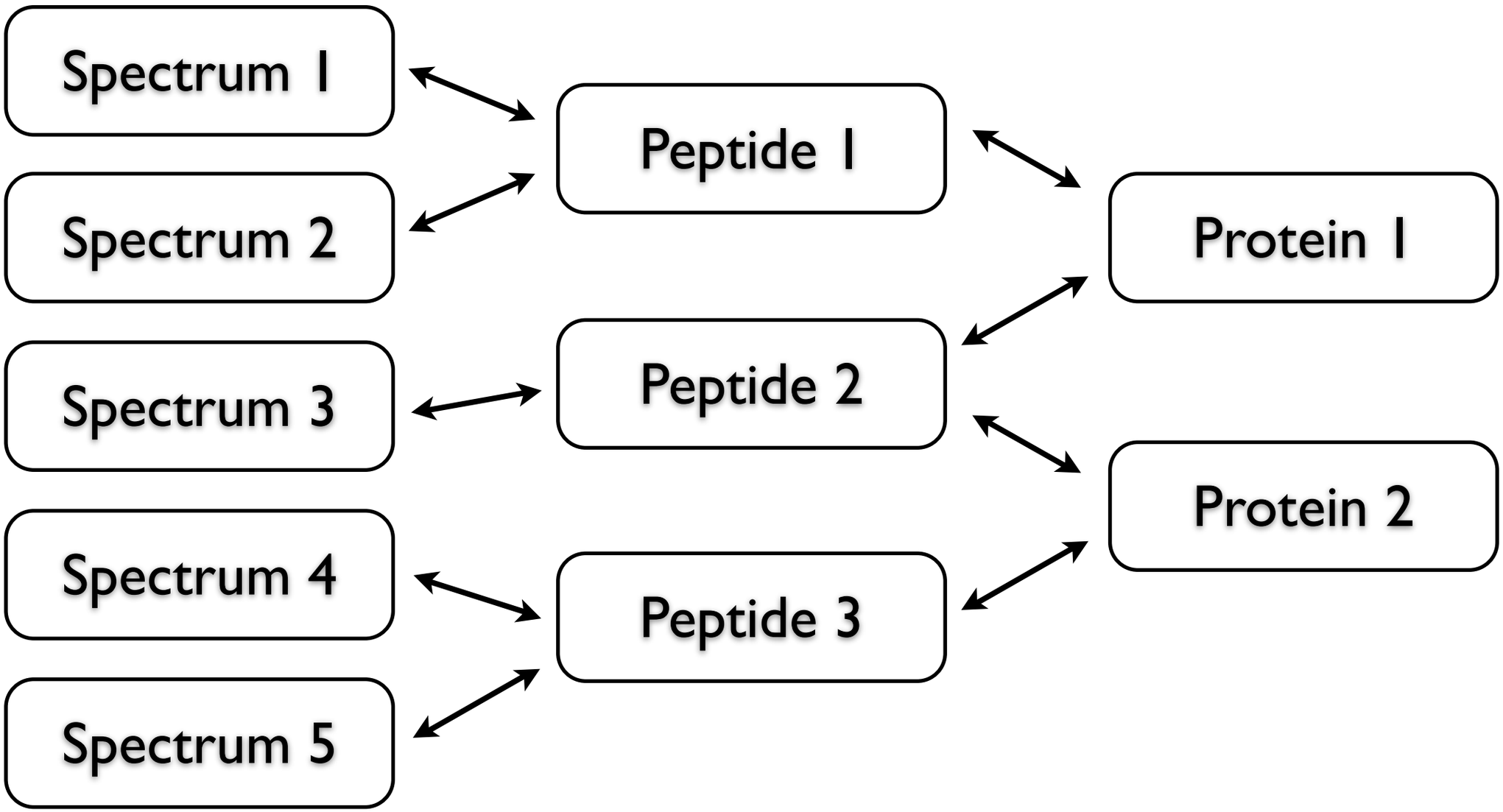


protein identification

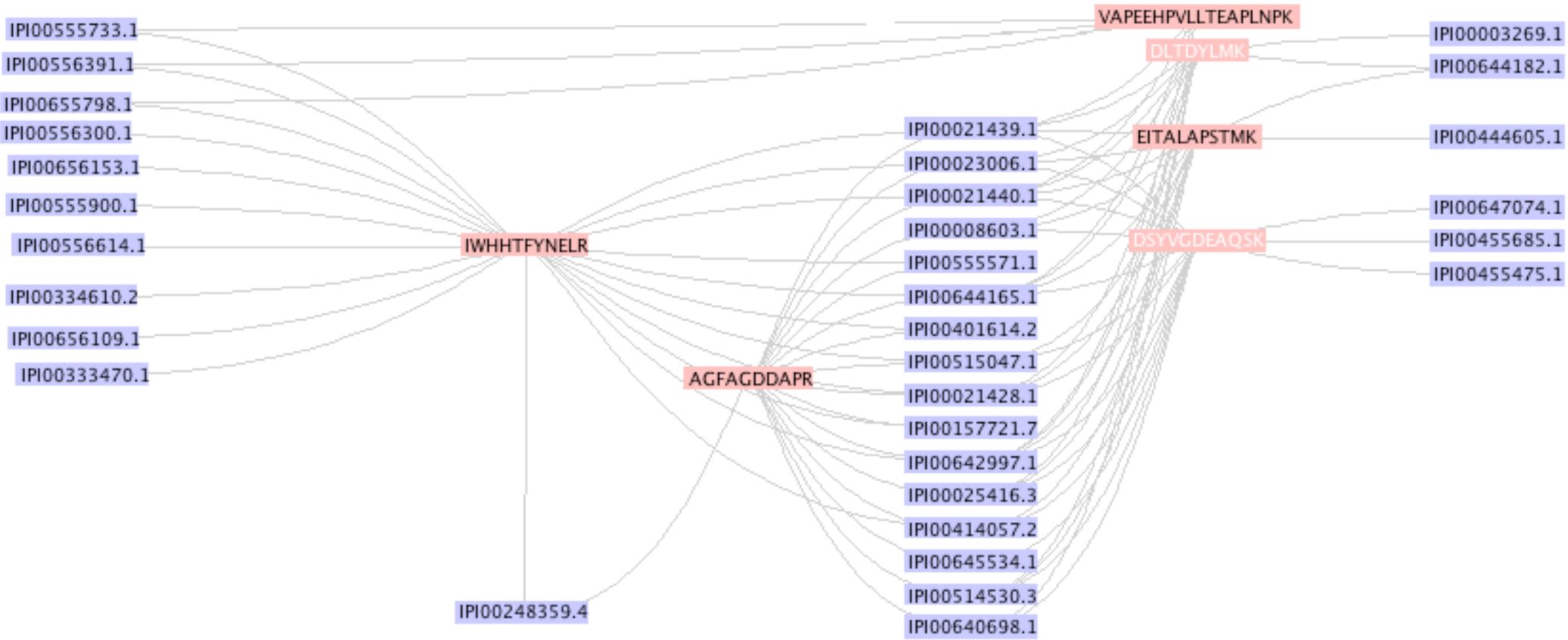


[Nesvizhskii *et al.* Anal Chem 2003]

PSM/Peptide/Protein level



Shared peptides



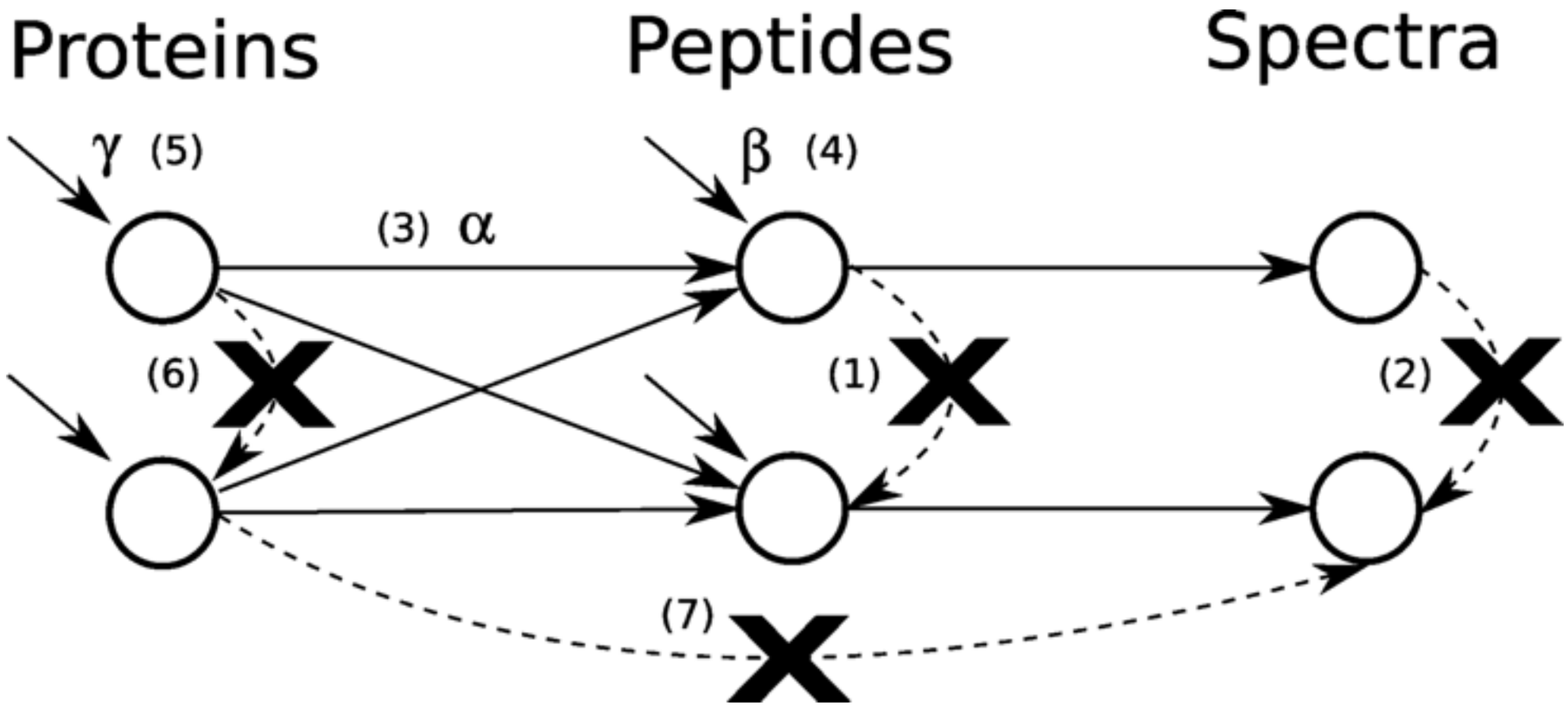
[MassSieve]

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

Technological Innovation and Resources

Rapid and Deep Human Proteome Analysis by Single-dimension Shotgun Proteomics*

Mohammad Pirmoradian‡§, Harshavardhan Budamgunta‡, Konstantin Chingin‡, Bo Zhang‡, Juan Astorga-Wells‡§, and Roman A. Zubarev‡¶‖

Multiparameter optimization of an LC-MS/MS shotgun proteomics experiment was performed without any hardware or software modification of the commercial instrument. Under the optimized experimental conditions, with a 50-cm-long separation column and a 4-h LC-MS run (including a 3-h optimized gradient), 4,825 protein groups and 37,550 peptides were identified in a single run and 5,354 protein groups and 56,390 peptides in a triplicate analysis of the A375 human cell line, for approximately 50% coverage of the expressed proteome. The major steps enabling such performance included optimization of the cell lysis and protein extraction, digestion of even insoluble cell debris, tailoring the LC gradient profile, and choosing the optimal dynamic exclusion window in data-dependent MS/MS, as well as the optimal *m/z* scan window. ***Molecular & Cellular Proteomics 12: 10.1074/mcp.O113.028787, 3330–3338, 2013.***

LC-MS-based proteomics has by now become an analytical method of choice in biological studies that demand deep proteome coverage (1–3). In order to increase the number of identified proteins, LC-MS analysis is commonly preceded by sample fractionation on the level of proteins or proteolytic peptides, or both (*e.g.* using two-dimensional gel electrophoresis, strong anion exchange, or isoelectric focusing) (4–7). These multidimensional approaches greatly reduce the complexity of the protein or peptide mixture in each fraction prior to MS detection, which enables comprehensive analysis of nearly the entire human proteome (>10,000 proteins) (6). The reverse side of the coin is the substantial operational cost, sample consumption (up to milligrams), and integral instrument time spent in these analyses (typically several days or longer). This puts severe limitations on high-throughput biological and clinical research.

In recent years, the power of the core analytical methods employed in proteomics, liquid chromatography and mass spectrometry, has sizably increased. Owing to the technological developments in packing materials of analytical columns and coupling interfaces, LC is now entering the era of ultra-high-pressure liquid chromatography (UPLC) characterized by unparalleled peak capacity and speed of separation (8–11). High-resolution MS is progressing at a fast rate with regard to sequencing capabilities and sensitivity of detection (12–15). Apart from that, notable improvements have been achieved in related areas, such as sample preparation methods and MS data processing (16–21).

The improving performance of shotgun LC-MS proteomics reduces the gap between the analytical capabilities of one-dimensional and multidimensional approaches. This trend is likely to continue in the near future, in view of the ongoing rapid technology developments. Considering the evident advantages of one-dimensional proteomics (*i.e.* the ease and speed of operation, lower sample consumption, and lower cost per run), it may regain the dominant position in many biological and clinical applications that it lost with the advent of multidimensional strategies. A wide selection of one-dimensional LC-MS platforms is commercially available nowadays for routine protein analyses with complete automation of the operational workflow, allowing large arrays of biological samples to be screened without attendance. In contrast, multidimensional analyses often involve interruptions in the experimental procedure for important steps that need to be performed manually by experienced personnel.

Most recent one-dimensional proteomics studies employing the combination of UPLC separation and high-resolution MS detection demonstrate remarkable progress in protein coverage, as well as in sensitivity and speed of analysis. In a very recent study, Nagaraj *et al.* reported an average of 3,923 protein groups identified in a single 4-h LC-MS analysis of 4 μg of yeast cell lysate (22). Combined analysis of six single runs increased the number of identifications to more than 4,000, which is close to the total number of proteins expressed in yeast under normal conditions. The median coverage of proteins in pathways with at least 10 members in the

LETTERS

nature biotechnolo

Identification of post-translational modifications by blind search of mass spectra

Dekel Tsur[1,4], Stephen Tanner[2,4], Ebrahim Zandi[3], Vineet Bafna[1] & Pavel A Pevzner[1]

Most tandem mass spectrometry (MS/MS) database search algorithms perform a restrictive search that takes into account only a few types of post-translational modifications (PTMs) and ignores all others. We describe an unrestrictive PTM search algorithm, MS-Alignment, that searches for all types of PTMs at once in a blind mode, that is, without knowing which PTMs exist in nature. Blind PTM identification makes it possible to study the extent and frequency of different types of PTMs, still an open problem in proteomics. Application of this approach to lens proteins resulted in the largest set of PTMs reported in human crystallins so far. Our analysis of various MS/MS data sets implies that the biological phenomenon of modification is much more widespread than previously thought. We also argue that MS-Alignment reveals some uncharacterized modifications that warrant further experimental validation.

PTMs greatly increase the complexity of the proteome, and identifying PTMs is undoubtedly the next big step for proteomics[1,2]. However, reliable computational identification of PTMs remains a formidable challenge. The first approach to PTM identification involved the enumeration and scoring of all possible modifications for each peptide from the database[3]. This exhaustive search approach has limitations because it can take into account only a few modifications and is prohibitively slow for mutation detection. In other words, a researcher must 'guess' in advance which PTMs are present in the sample. As a result, a restrictive search is performed for a small set of PTMs and all others are ignored. The question arises whether one can design an unrestrictive PTM search algorithm that searches for all types of PTMs at once in a blind mode, without knowing which PTMs exist

identification of a modified peptide that best matches bot
de novo interpretation and the database peptide. Although this el
formulation accommodates some *de novo* sequencing erron
approach depends critically on a good *de novo* interpretation.

We emphasize the important difference between the spectral
ment approach[4,5] and these more recent approaches. OpenSea
the heuristic branch-and-bound technique, which, in contra
spectral alignment, (i) does not guarantee the optimal solution
(ii) crucially depends on the quality of *de novo* reconstru
SPIDER[7] uses a rigorous dynamic programming algorithm (
similar to spectral alignment if there are no sequencing errors
only compares a database peptide against a single *de novo* inter
tion of an experimental spectrum. Spectral alignment, in co
compares a database peptide against every possible interpretati
an experimental spectrum, thus eliminating dependence on ac
de novo interpretations.

In this paper, we describe an extension of the spectral alig
algorithm with improved scoring and an order of magnitude imp
ment in speed (code available at http://peptide.ucsd.edu/). Re
two groups[8,9] studied a problem of interpreting peptides with a
modification. In this case, the mass shift is known in advance a
edit distance is 1, allowing one to substitute the dynamic pro
ming with an exhaustive search that analyzes every possible
position. We remark that the time complexity of exhaustive sear
a single modification is quadratic in the length of the peptide (fo
peptide in the database), whereas spectral alignment can be i
mented in linear time (see **Supplementary Methods** online).

Identification of all types of PTMs present in a large collecti
MS/MS spectra is a difficult task. It is important to distin