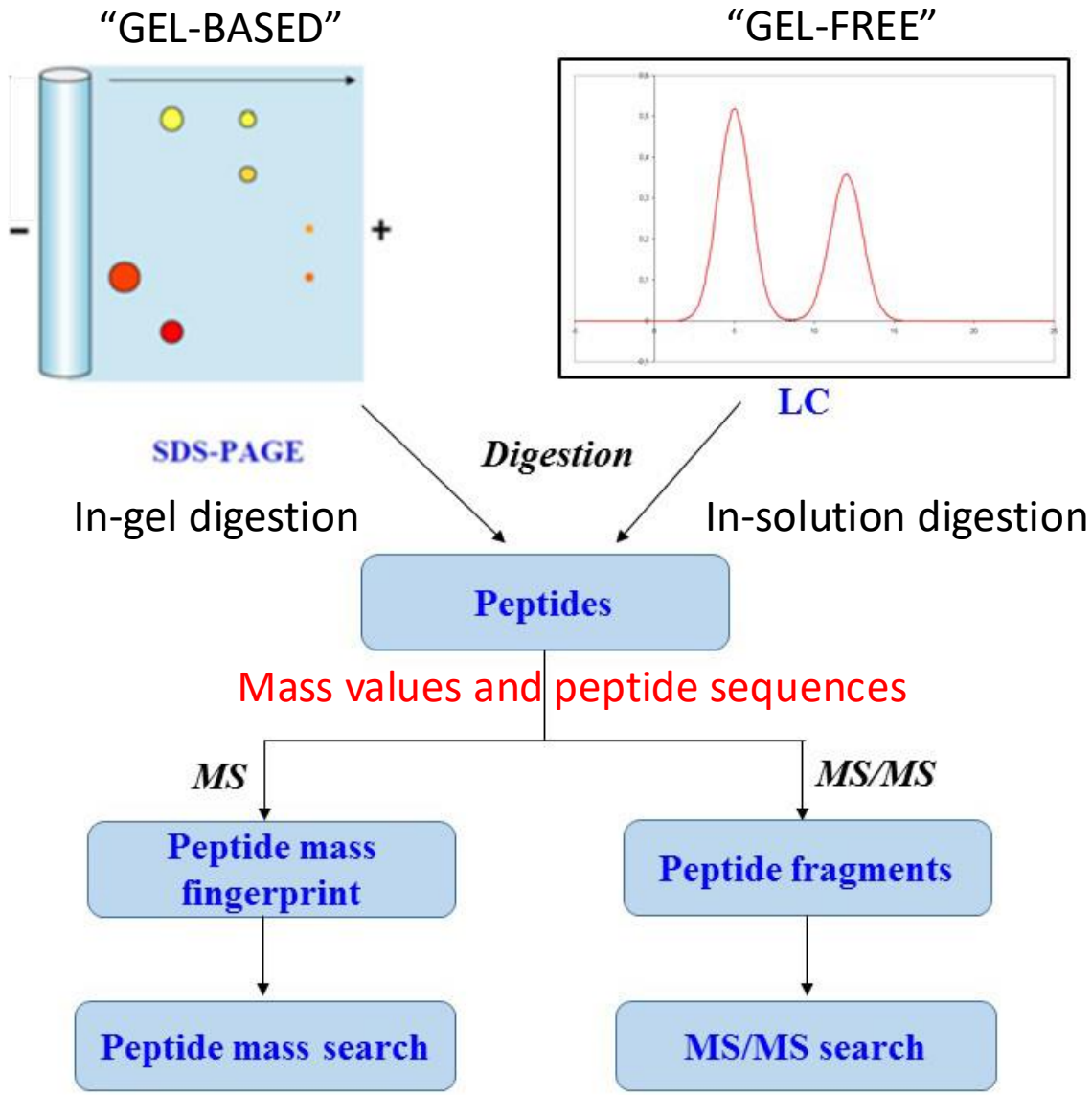
PROTEIN IDENTIFICATION requires protein digestion before analysis

- TWO GENERAL APPROCHES:
  - peptide mass fingerprinting
  - MS/MS or tandem MS (de novo sequencing or database ion search)



"PEPTIDE MASS FINGERPRINTING"

Peptide mass fingerprinting (**PMF**) (also known as **protein fingerprinting**) is an analytical technique for protein identification in which the unknown protein of interest is first cleaved into smaller peptides, whose masses can be accurately measured with a mass spectrometer

The mass spectrometric analysis produces a list of molecular weights of the fragments which is often called "peak list".

The peptide masses are compared to those present in protein databases such as Swissprot, which contains protein sequence information.

Software performs *in silico* digests on proteins in the database with the same enzyme (e.g. trypsin) used in the experimental cleavage reaction.

The mass of these peptide fragments is then calculated and compared to the peak list of measured peptide masses. The results are statistically analyzed and possible matches are returned in a results table.

Protein Identification - Peptide Mass Fingerprinting

https://www.youtube.com/watch?app=desktop&v=4xSUWK_ueWI

Why we call it PEPTIDE MASS FINGERPRINTING?

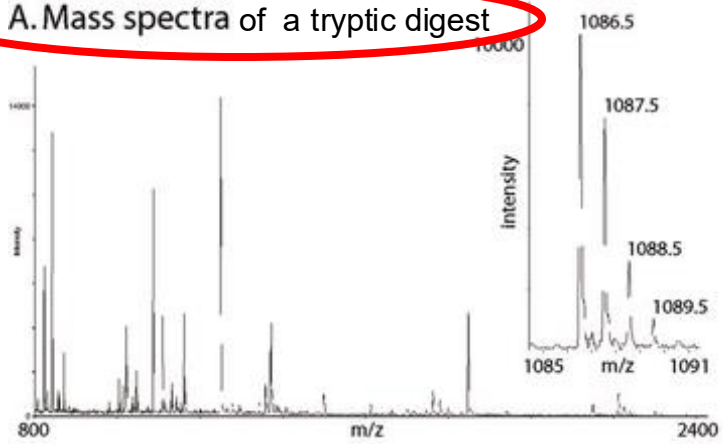
1)the protein must be PURIFIED!

2) the protein must have been previously characterized and loaded into the protein database

UNIQUENESS

The peak list is the protein fingerprint: as no person shares the same fingerprint, no protein shares the same list of peptides obtained from digestion with a specific enzyme (trypsin)

A. Mass spectra of a tryptic digest



B. Peak list

822.4545	1342.7736
830.4705	1358.7519
842.5096	1370.7164
856.5178	1374.6670
858.4764	1384.7440
870.5316	1498.8122
1021.5628	1611.8566
1045.5811	1762.6748
1086.5421	1780.6810
1108.6920	1847.8563
1141.7405	2148.9510
1161.6615	2211.1040
1249.5686	2225.1070

C. Database search

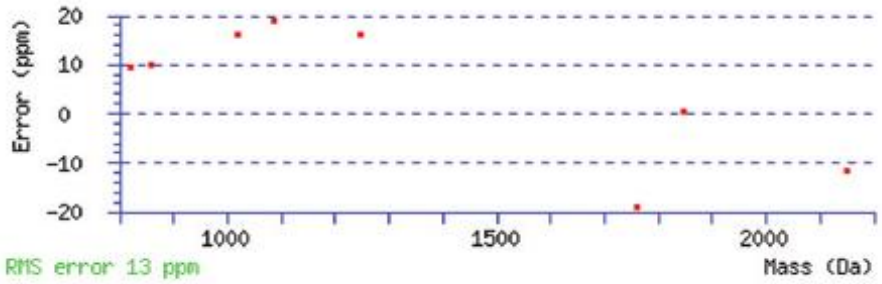
D. Protein ID

```
Match to: OMPF_ECOLI Score: 121 Expect: 9.4e-09
Outer membrane protein F precursor - Escherichia coli

Nominal mass (Mr): 39309; Calculated pI value: 4.76
```

E. Matching peptides

```
Number of mass values matched: 8
Sequence Coverage: 26%

Matched peptides shown in Bold Red
  1 MMKRNILAVI VPALLVAGTA NAAEIYNKDG NKVDLYGKAV GLHYFSKGNG
 51 ENSYGGNGDM TYARLGFKGE TQINSDLTGY GQWEYNFQGN NSEGADAQTG
101 NKTRLAFAGL KYADVGSFDY GRNYGVVYDA LGYTDMLPEF GGDTAYSDDF
151 FVGRVGGVAT YRNSNFFGLV DGLNFAVQYL GKNERDTARR SNGDGVGGSI
201 SYEYEGFGIV GAYGAADRTN LQEAQPLGNG KKAEQWATGL KYDANNIYLA
251 ANYGETRNAT PITNKFTNTS GFANKTQDVL LVAQYQFDFG LRPSIAYTKS
301 KAKDVEGIGD VDLVNYFEVG ATYYFNKNMS TYVDYIINQI DSDNKLGVGS
351 DDTVAVGIVY QF
```

F. Error distribution



The database search compares the theoretical peak list obtained by in silico digestion with trypsin with the experimental one

HOW TO perform a DATABASE SEARCH for PMF

In **silico digestion** can be done by usigng different software:

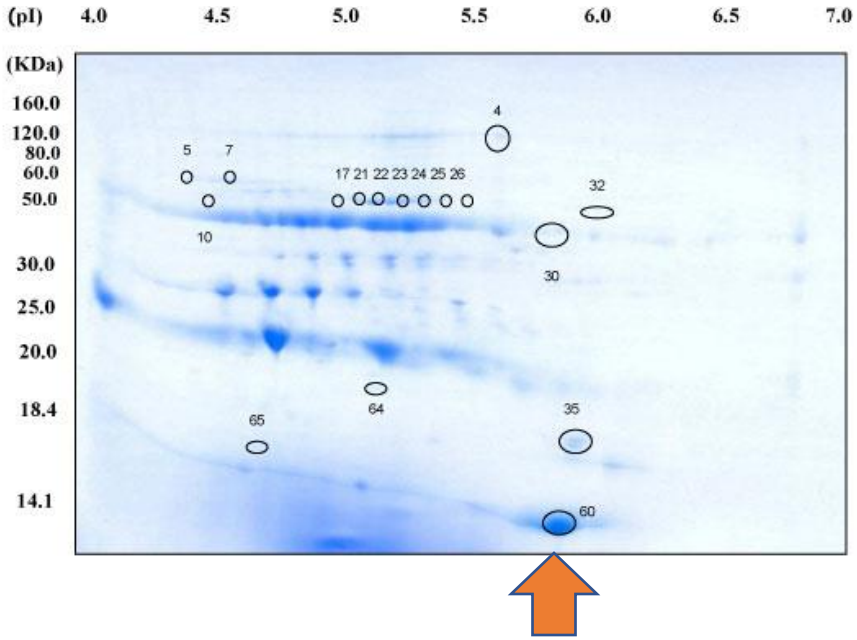
- **PeptideMass**
  - cleaves a protein sequence from the UniProt Knowledgebase (Swiss-Prot and TrEMBL)
  - **or** a user-entered protein sequence with a chosen enzyme, and computes the masses of the generated peptides.

→ Comparison whit a single known protein

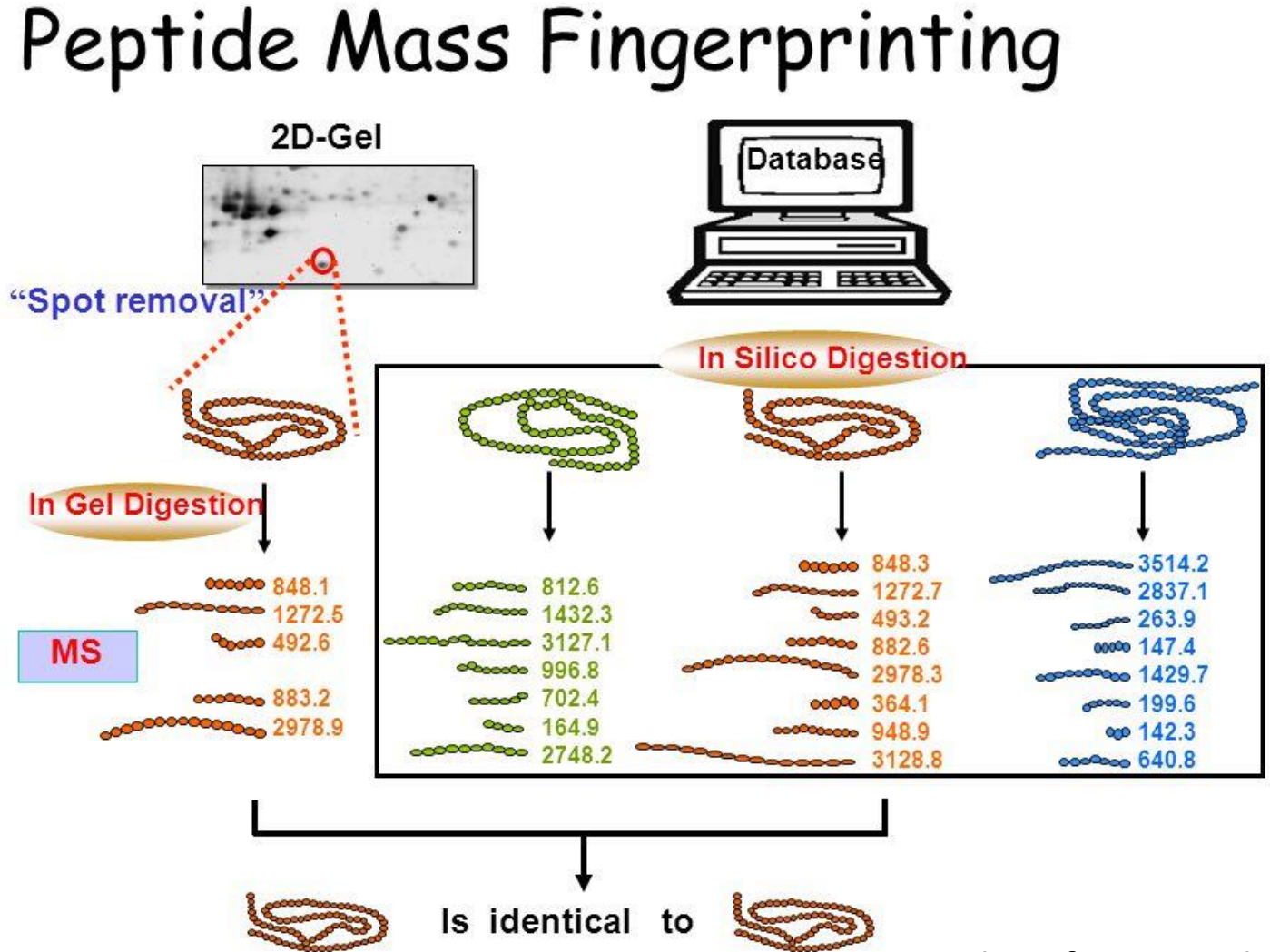
The **ExPASy** (the Expert Protein Analysis System) World Wide Web server (http://www.expasy.org),

While useful, there are some disadvantages to peptide mass fingerprinting. These include the need for pure proteins or relatively simple mixtures. Peptide mass fingerprinting also relies on the identification of multiple peptides before a protein can be characterised.

Workflow for PMF- first hypothesis: 1) We need to confirm if the selected protein is MYOGLOBIN

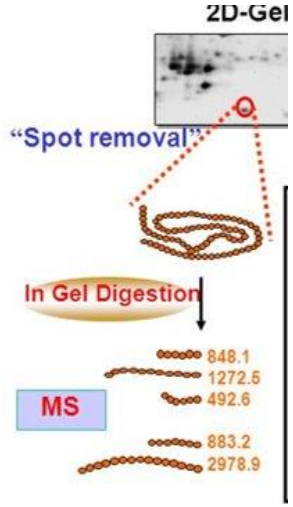


Is this spot myoglobin?



Looking for a match!

IN SILICO DIGESTION OF THE SELECTED PROTEIN AND EXPERIMENTAL IN-GEL DIGESTION OF THE PROTEIN FOLLOWED BY MS



From HR-MS analisys
Monoisotopic/monocharged
experimental mass values

1931.9
1913.0
1853.9
1632.8
1515.6
1350.8
910.4
828.3
748.4
738.2

https://www.uniprot.org/

>sp|P02144|MYG_HUMAN Myoglobin
OS=Homo sapiens OX=9606 GN=MB PE=1 SV=2
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFK
GHPETLEKFDKFKHLKSEDEMKASEDLKKHGATVLTA
LGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECI
IQVLQSKHPGDFGADAQGAMNKALELFRKDMASNY
KELGFQG

From in silico digestion
Monoisotopic/monocharged
experimental mass values

1931.9683
1913.0088
1853.9616
1632.8703
1515.6645
1350.8103
910.4628
828.3556
748.4352
738.2974

https://web.expasy.org/peptide_mass/

PeptideMass

PeptideMass [references] cleaves a protein sequence from the UniProt Knowledgebase (Swiss-Prot a theoretical isoelectric point and mass values for the protein of interest. If desired, PeptideMass can retu polymorphisms or splice variants.

Instructions are available.

Enter a UniProtKB protein identifier, ID (e.g. ALBU_HUMAN), or accession number, AC (e.g. P04406),

Reset the fields. Perform the cleavage of the protein.

The peptide masses are
with cysteines treated with: nothing (in reduced form)
with acrylamide adducts
with methionines oxidized
$[M+H]^+$ or $[M]$ or $[M-H]^-$ or $[M+2H]^{2+}$ or $[M+3H]^{3+}$
average or monoisotopic.

Select an enzyme: Trypsin

Allow for 0 missed cleavages.
Display the peptides with a mass bigger than 500 and smaller than unlimited Dalton

sorted by peptide masses or in chronological order in the protein.

For UniProtKB (Swiss-Prot/TrEMBL) entries only:
For each peptide display
all known post-translational modifications,
all database conflicts,
all variants (polymorphisms),
all mRNA variants (due to alternative splicing, initiation or promoter usage).

The experimental and theoretical peak listes must be identical!!

HOW TO perform a DATABASE SEARCH for PMF

In **silico digestion** can be done by usigng different software:

- **PeptideMass**
  - cleaves a protein sequence from the UniProt Knowledgebase (Swiss-Prot and TrEMBL)
  - **or** a user-entered protein sequence with a chosen enzyme, and computes the masses of the generated peptides.

Comparison whit a single known protein

The **ExPASy** (the Expert Protein Analysis System) World Wide Web server (http://www.expasy.org),

If you don't know the protein sequence OR you need to confirm the previously obatained results

- **MATRIX** The tool searches in the swissprot database all the human proteins whose in silico digestion with trypsin provides an theoretical peak list identical the the experimental one.

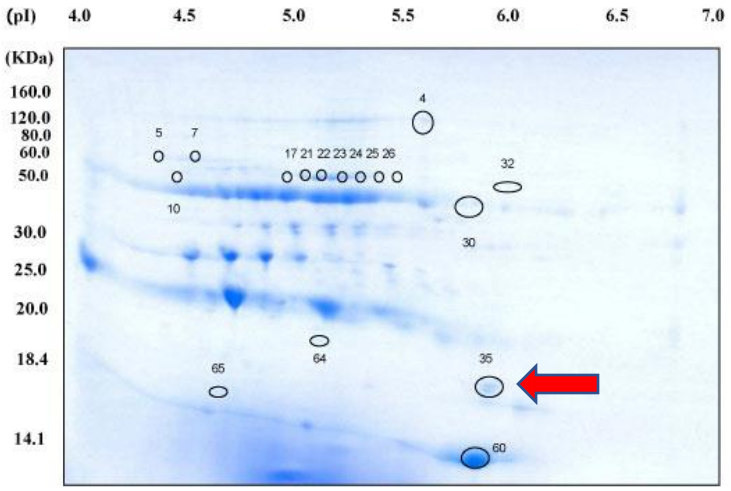
While useful, there are some disadvantages to peptide mass fingerprinting. These include the need for pure proteins or relatively simple mixtures. Peptide mass fingerprinting also relies on the identification of multiple peptides before a protein can be characterised.

PMF- second hypothesis: 2) If we do not know the protein sequence or if we neen to better characterize the protein sequence:

We need to confirm the sequence by comparing the experimental peptide masses using a protein databases such as Swissprot, which contain protein sequence information.

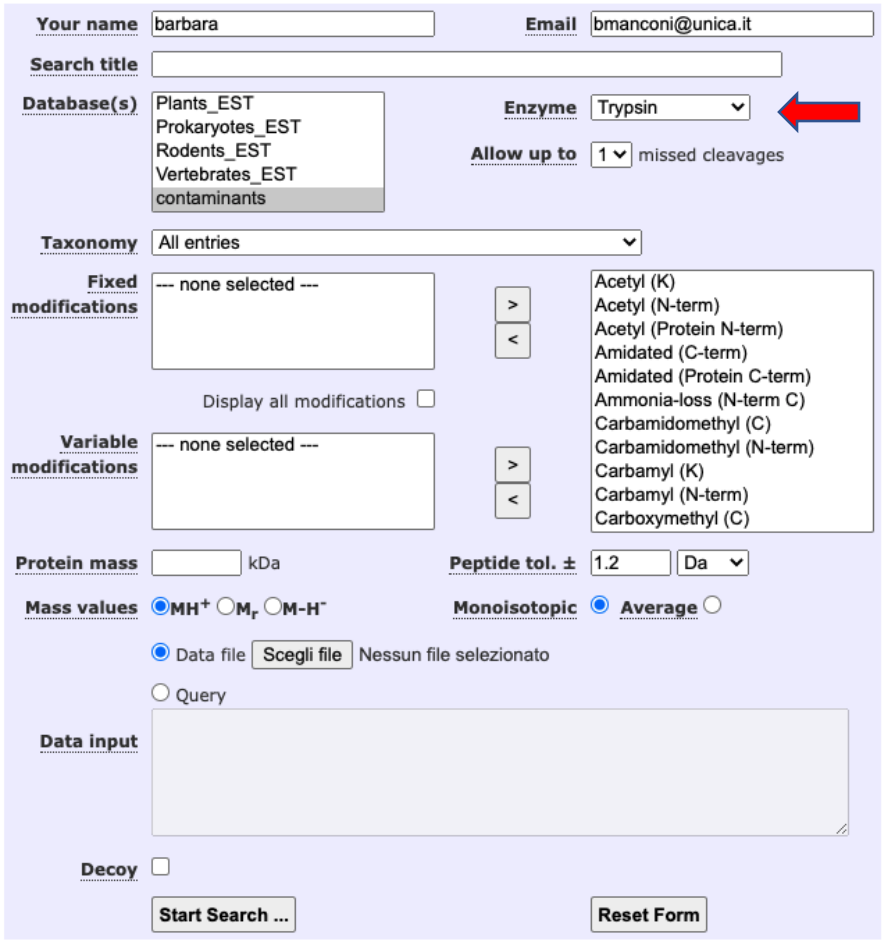
From MS analisys
Monoisotopic/monocharged experimental mass values

1793.9432
1767.9388
1749.8840
921.4676
903.4934
890.4618
871.3978
860.4876
858.4719
844.4233
840.4574
830.4076
826.4669
700.3083



Website for both Peptide Mass Fingerprint and MS/MS database searches.:MASCOT

https://www.matrixscience.com/



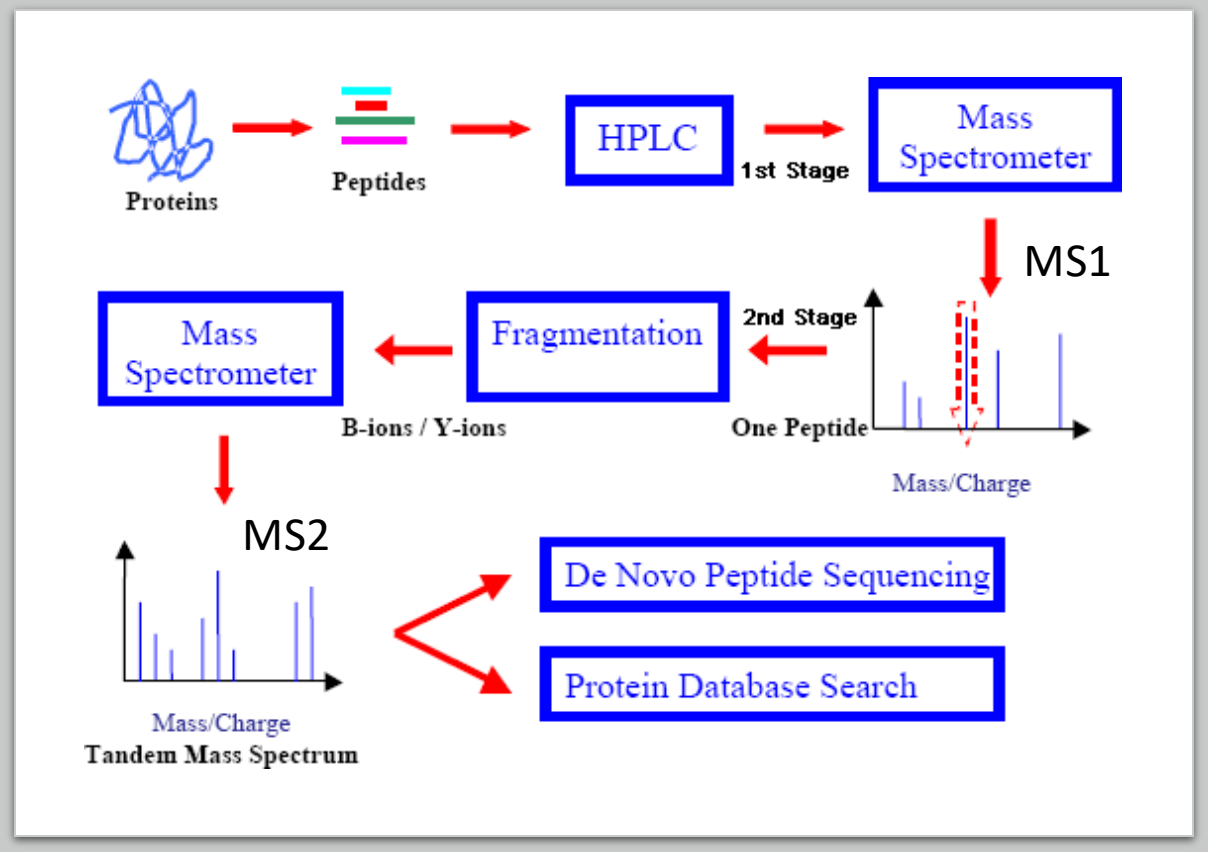
The tool searches in the swissprot database all the human proteins whose in silico digestion with trypsin provides an theoretical peak list identical the the experimental one.

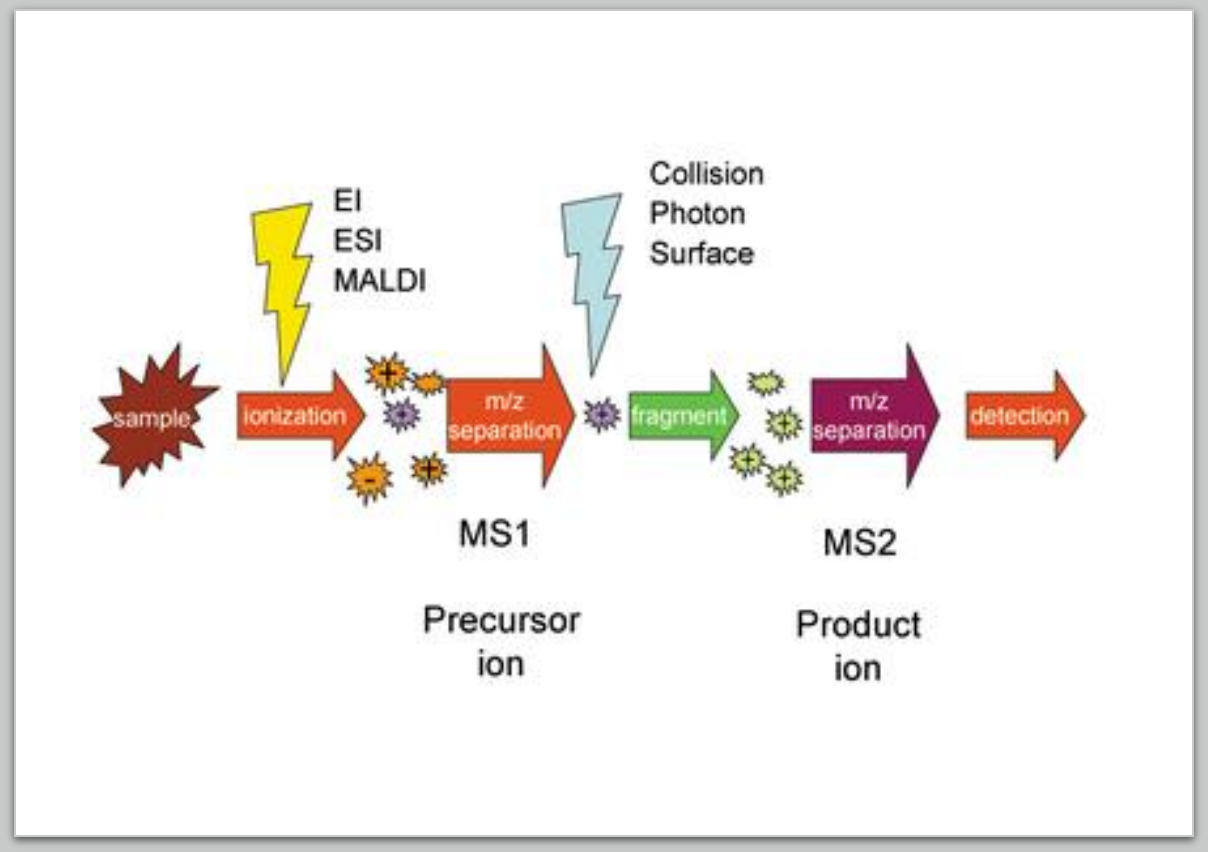
PROTEIN IDENTIFICATION BY

Tandem mass spectrometry or MS-MS or MS^2

If we don't get a protein identification by PMF....

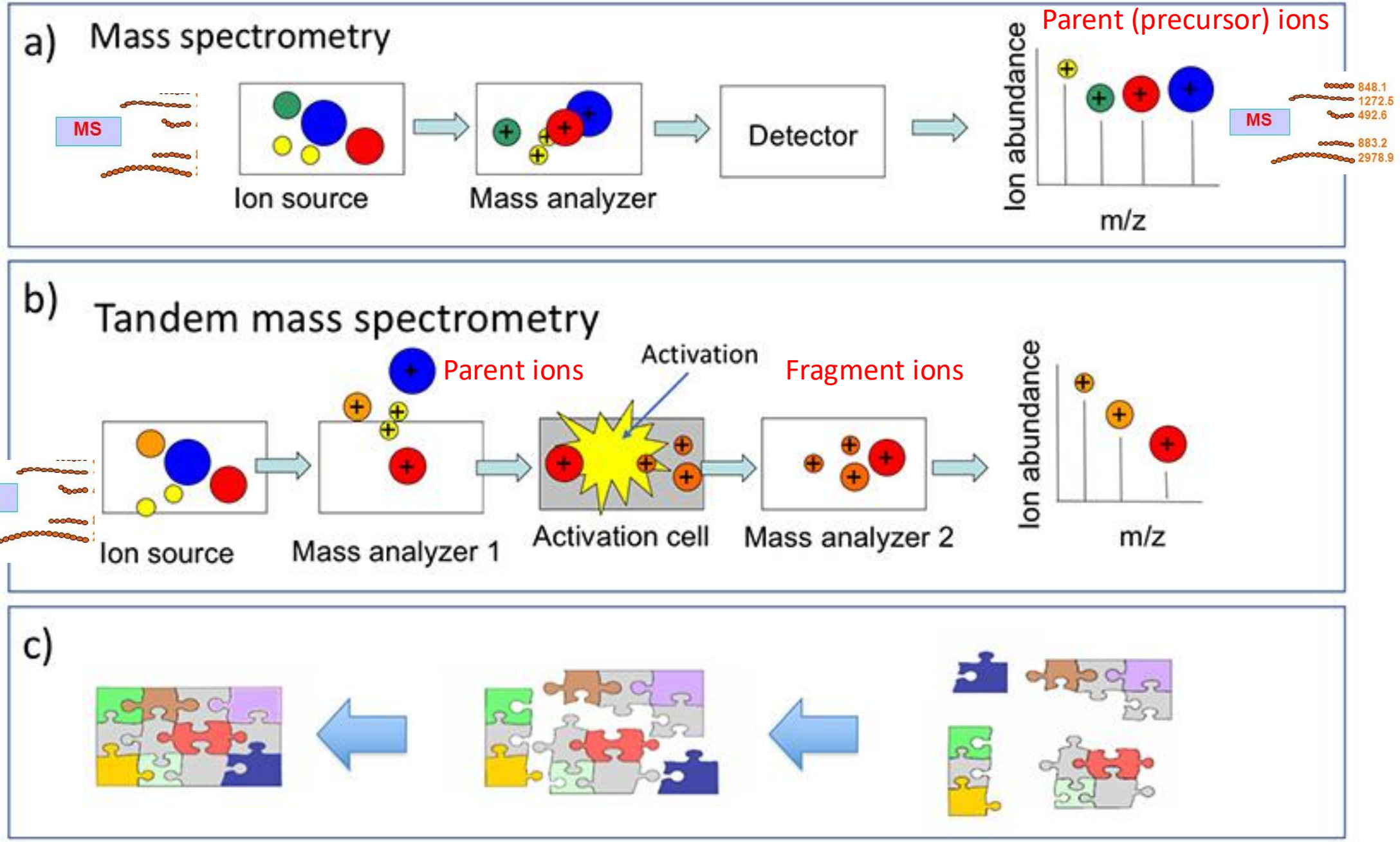
- MS/MS means using two mass analyzers to select an ion from a mixture, then fragment it to give structural information.
- Tandem mass spectrometry selects one of the intense peaks observed in the mass spectrum and further fragments all peptides with the selected mass to charge ratio.
- The tandem mass spectrum typically contains mass to charge ratio information about fragments of a a single peptide.





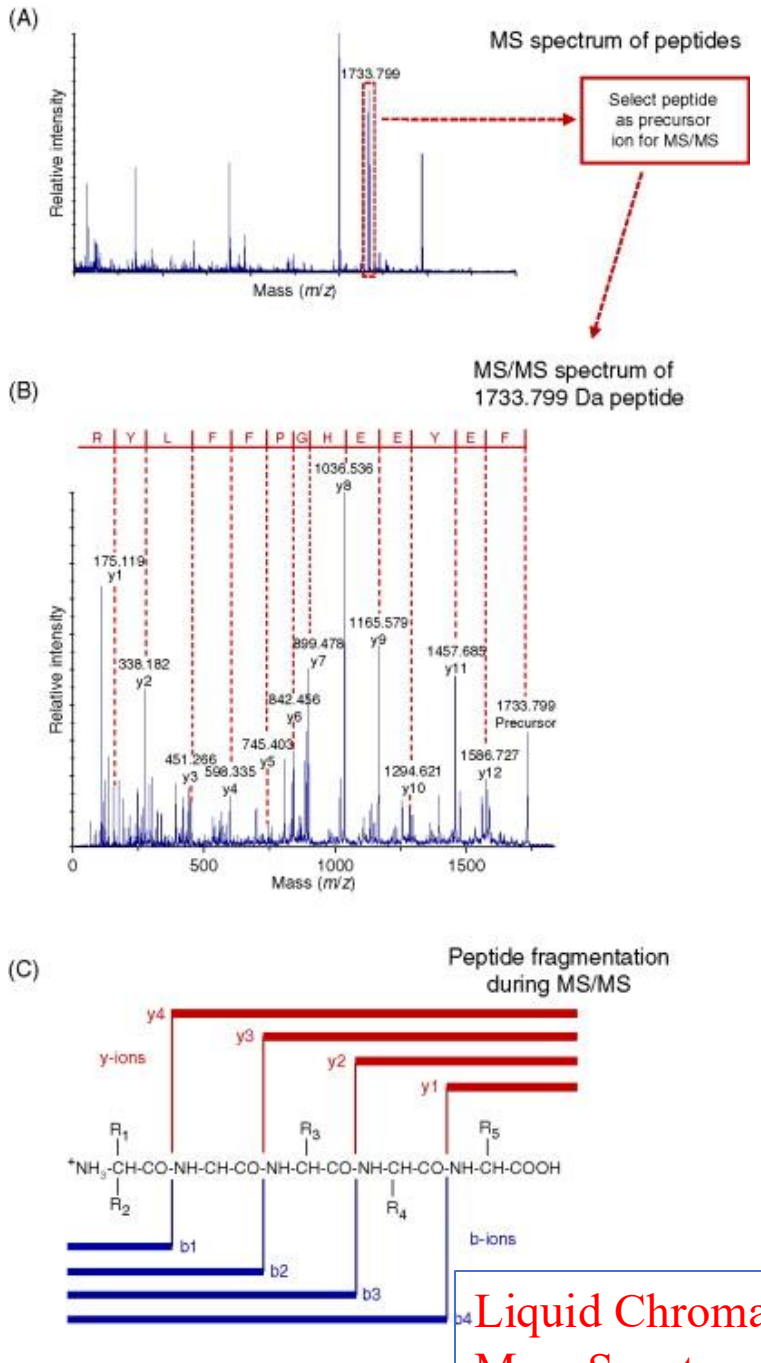
MS¹

MS²

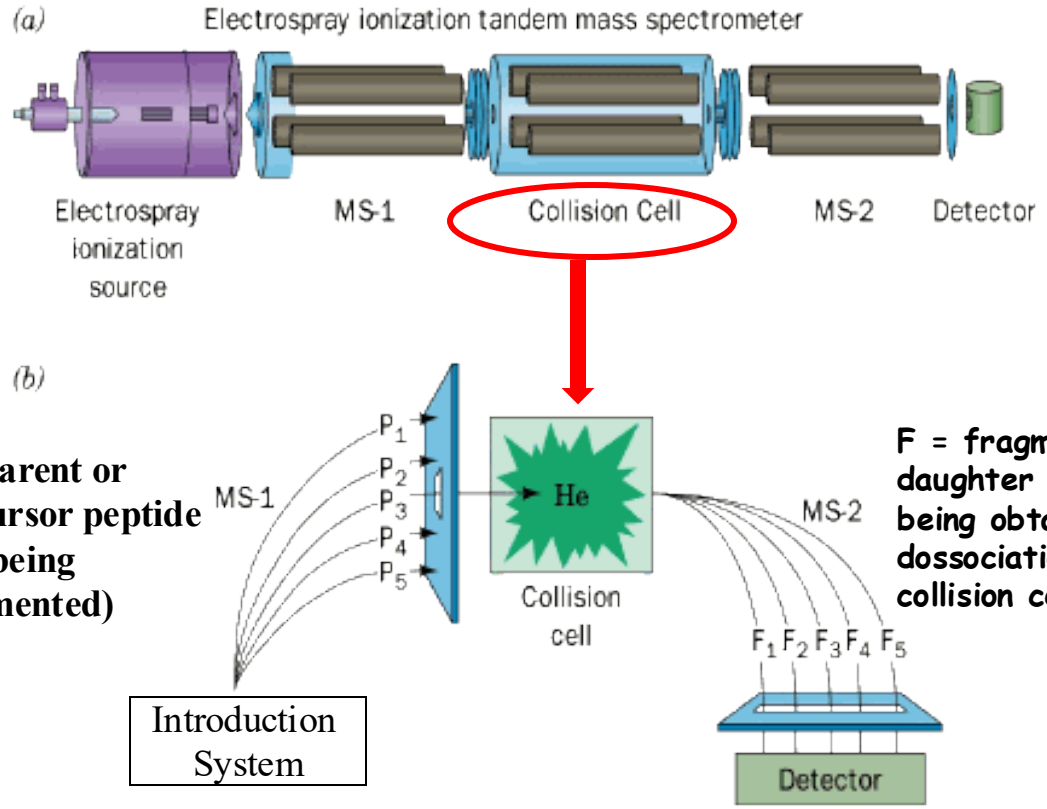

a) Mass spectrometry
MS
Ion source
Mass analyzer
Detector
Ion abundance
m/z
Parent (precursor) ions
MS
848.1
1272.5
492.6
883.2
2978.9
b) Tandem mass spectrometry
Parent ions
Activation
Fragment ions
MS
Ion source
Mass analyzer 1
Activation cell
Mass analyzer 2
Ion abundance
m/z
c)

MS1 spectrum

MS2 spectrum



Fragmentation of the selected parent peptides occurs into the COLLISION CELL in the presence of an inert gas (Ar, He)



P = parent or precursor peptide (ion being fragmented)

F = fragment or daughter peptide (ion being obtained after dossociation into the collision cell)

Introduction System

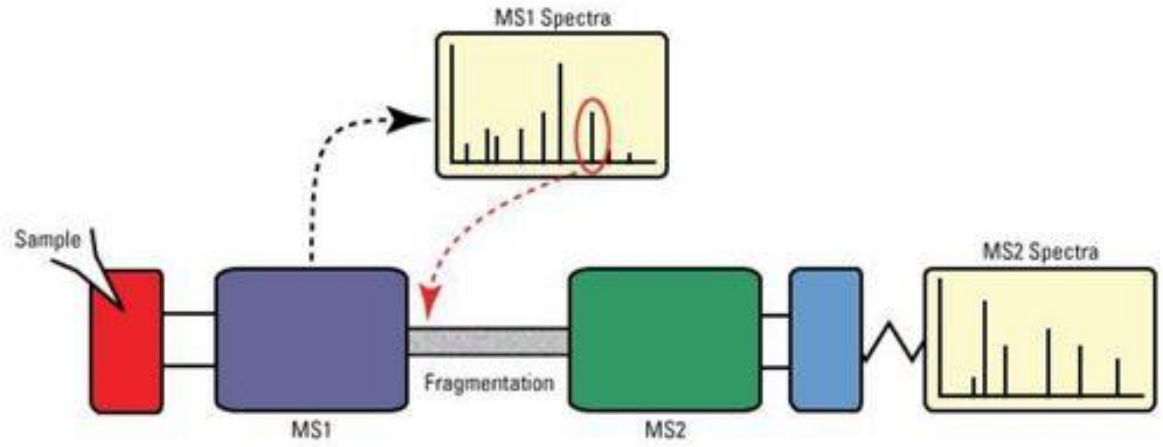
The use of a tandem mass spectrometer (MS/MS) in amino acid sequencing

Liquid Chromatography-Tandem Mass Spectropmetry (LC-MS/MS)

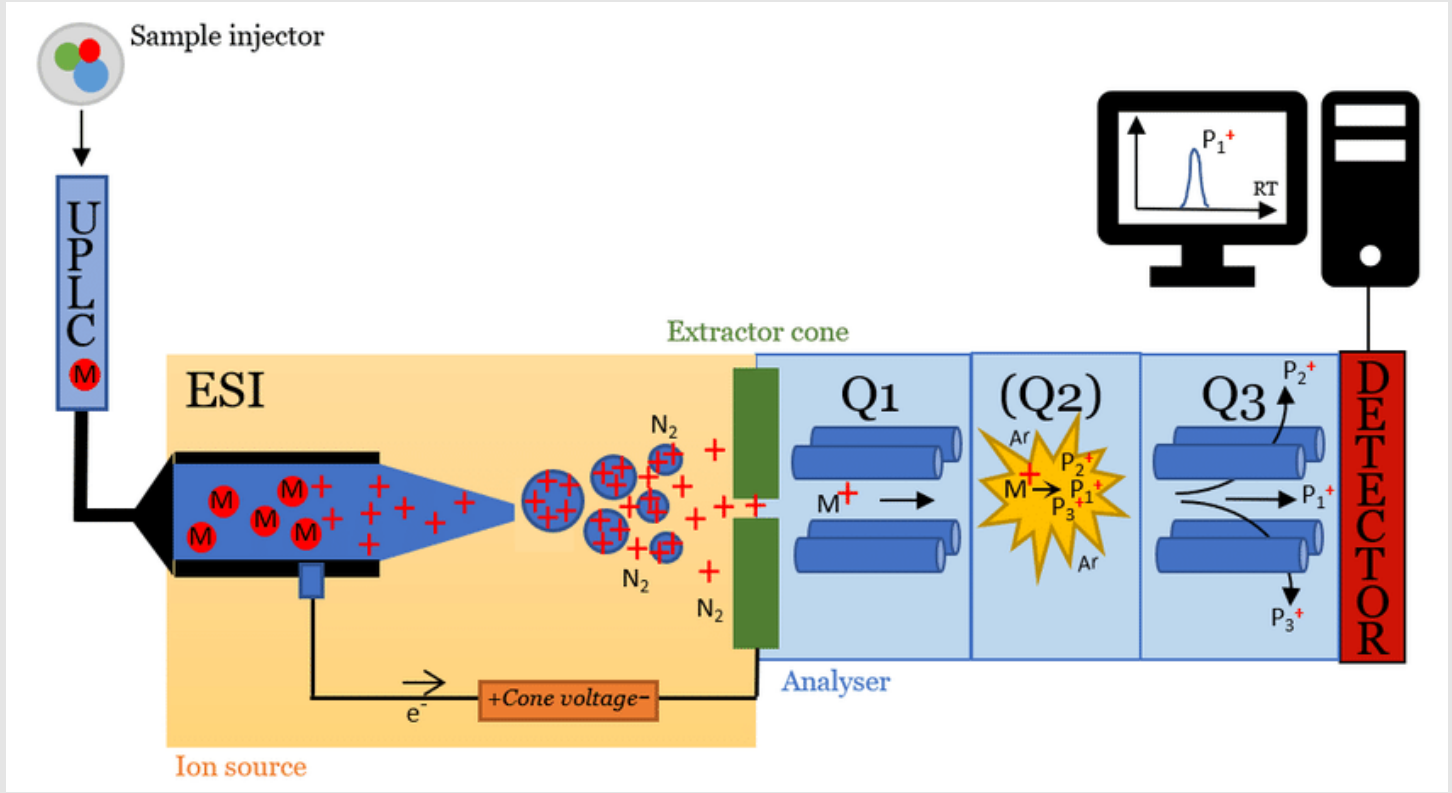
https://www.youtube.com/watch?v=Jc1uC6EbMCs

Tandem MS (MS/MS)

- Overview
  - **MS1:** Samples ionised, separated by *m/z* and then detect
  - **Isolation:** Isolate an ion of interest
  - **Fragmentation:** Create fragment ions from collision induced dissociation (CID) or other methods
  - **MS2:** For the fragment ions separate by *m/z* and then detect
- Tandem in space
  - Quadrupole
  - Time of Flight (TOF)
- Tandem in time
  - Ion trap
  - Allows MS^n

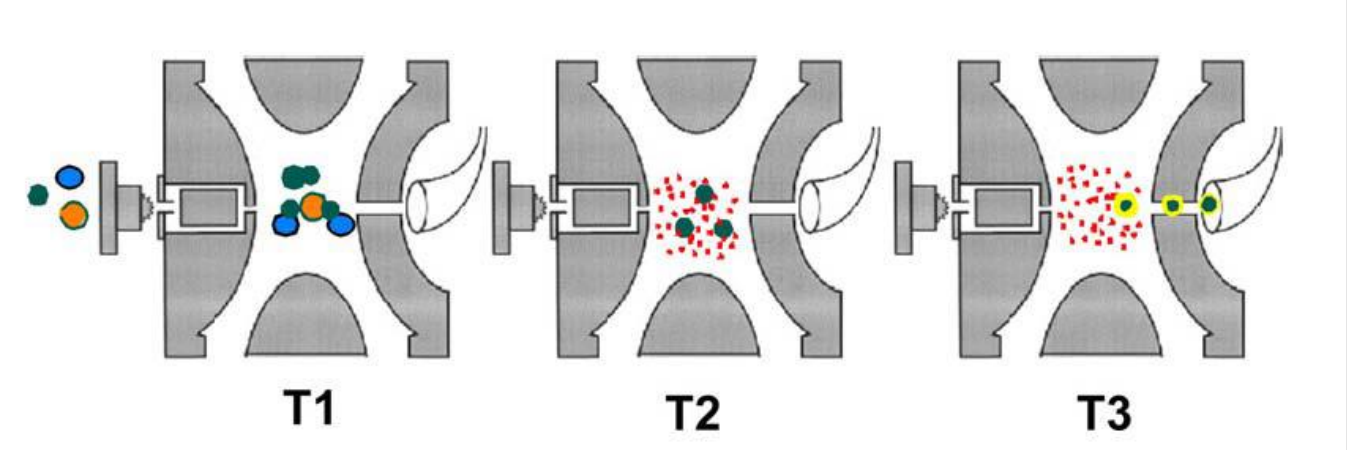


Typical Tandem MS in space instruments include QqQ, QTOF, and hybrid ion trap/FTMS, etc.



- Three Quadrupoles (Q 1, Q 2, and Q 3) are lined up in a row. Precursor ions are selected by the first mass analyzer (Q1), fragmented in the next Q2 and subsequently separated according to *m/z (Q3)*.

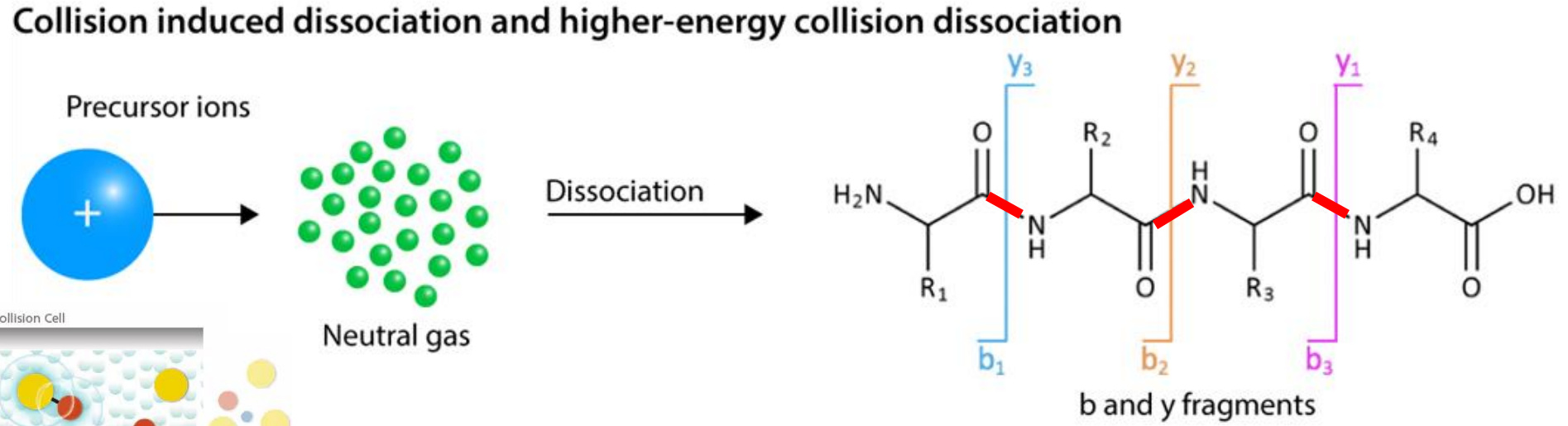
Typical Tandem MS in TIME instruments include ion trap.



Tandem-in-time repeatedly uses a single mass analyzer which of course must be able to perform precursor ion selection, ion activation, storage for the period of activation, fragmentation, and subsequent mass analysis of the fragments.

1. All ions are trapped (T1)
2. The RF potential is set to trap only the ion to be fragmented
3. Fragmentation occurs by collision of the ions with the inert gas (T2)
4. Perform m/z scan of product ions (T3)

There are many methods used to fragment the ions during a MS2 analysis and these result in different types of fragmentation and thus different information about the structure and composition of the molecule.



The cleavage occurs at the peptide bonds and generates y- and b-type fragment ions

Collision-induced dissociation

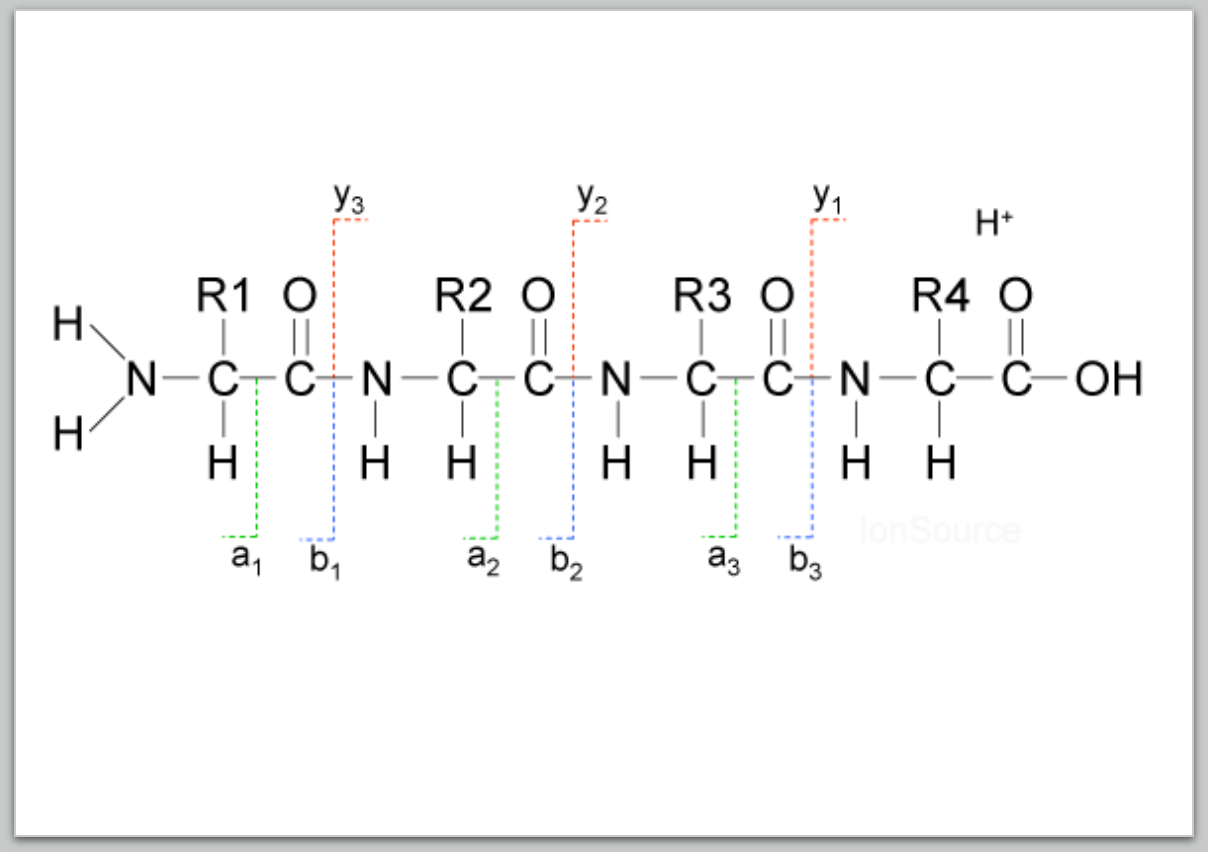
induce fragmentation of selected ions in the gas phase.The selected ions (typically molecular ions or protonated molecules) are usually accelerated by applying an electrical potential to increase the ion kinetic energy and then allowed to collide with neutral molecules (often helium, nitrogen or argon). In the collision some of the kinetic energy is converted into internal energy which results in bond breakage and the fragmentation of the molecular ion into smaller fragments.

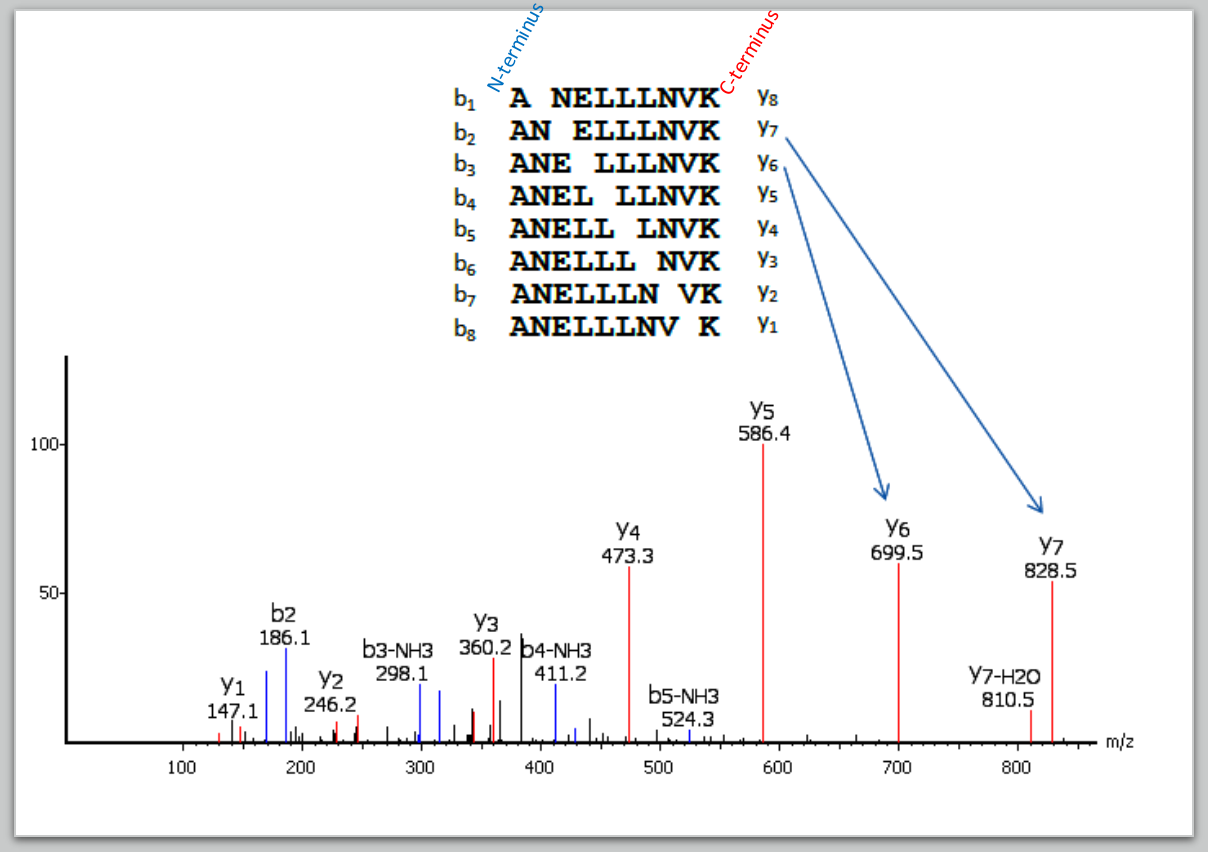
https://www.youtube.com/watch?v=XPR0Tsn3i3I

CID: Peptide Fragmentation Nomenclature

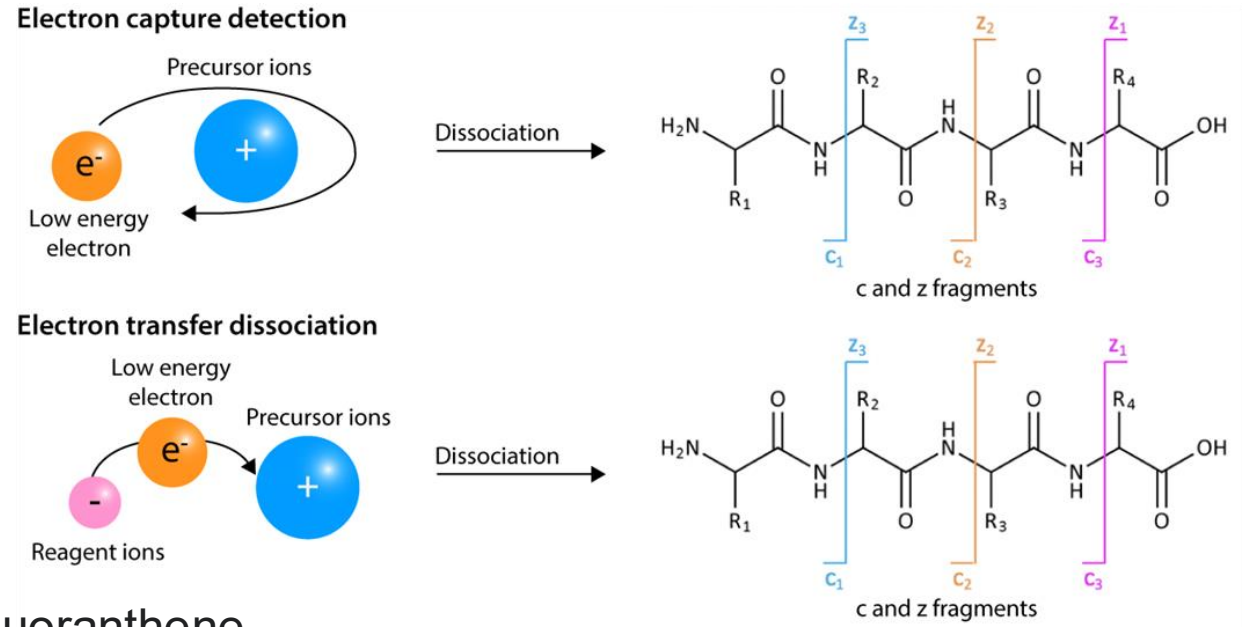
b, y and a ions

- The most common peptide fragments observed in low energy collisions (<100 eV) are **a, b** and **y** ions.
- The **b** ions appear to extend from the amino terminus, called the N-terminus.
- The **a** ions are often used as a diagnostic for **b** ions, such that **a**-**b** pairs are often observed in fragment spectra. The **a**-**b** pairs are separated by 28u, the mass for the carbonyl, C=O.
- The **y** ions appear to extend from the carboxyl terminus, or C-terminus.



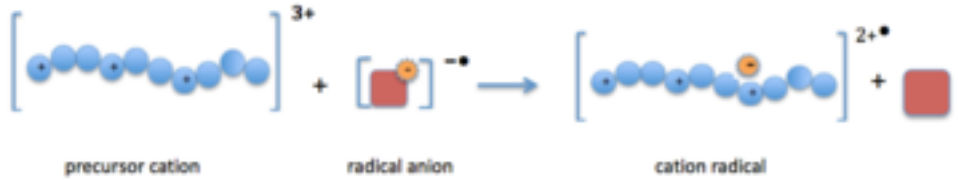


ECD



ETD

Fluoranthene



direct electron attachment to the π* amide

RADICAL-BASED Fragmentation:

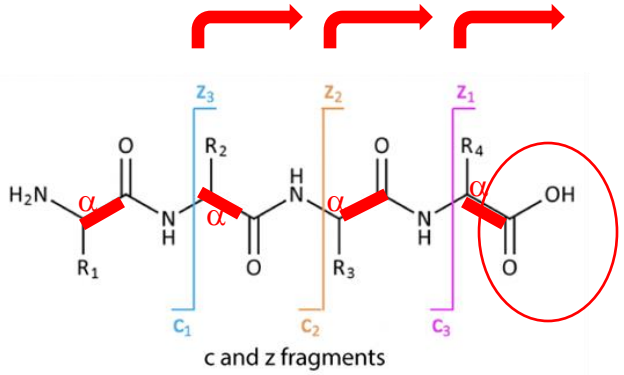
an electron is tranferred to an positive precursor molecule to create a instable cation radical, which spontaneously fragments at sites related to the location of electron capture.

ECD: irradiation of trapped cations with low-energy electrons, typically from a heated filament electron gun or indirectly heated dispenser cathode.

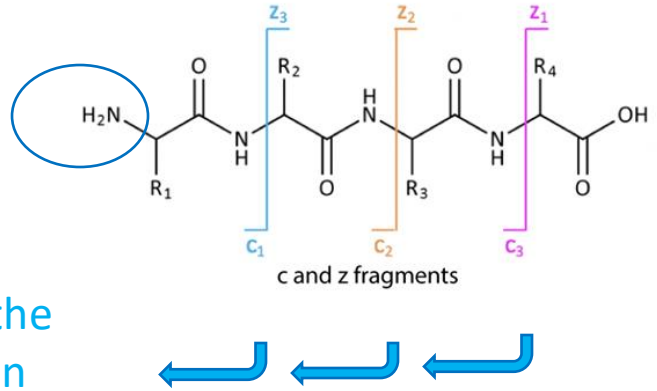
ETD: In order for an electron to be transferred to the positive precursor molecules, radical anions are generated and put into the ion trap with them. During the ion/ion reaction an electron is transferred to the positively-charged protein or peptide, causing fragmentation along the peptide backbone.

RADICAL-BASED Fragmentation:

The cleavage occurs at the N-Cα backbone bonds and generates C- and Z-type fragment ions



Z fragments ions retain the C-terminus of the protein



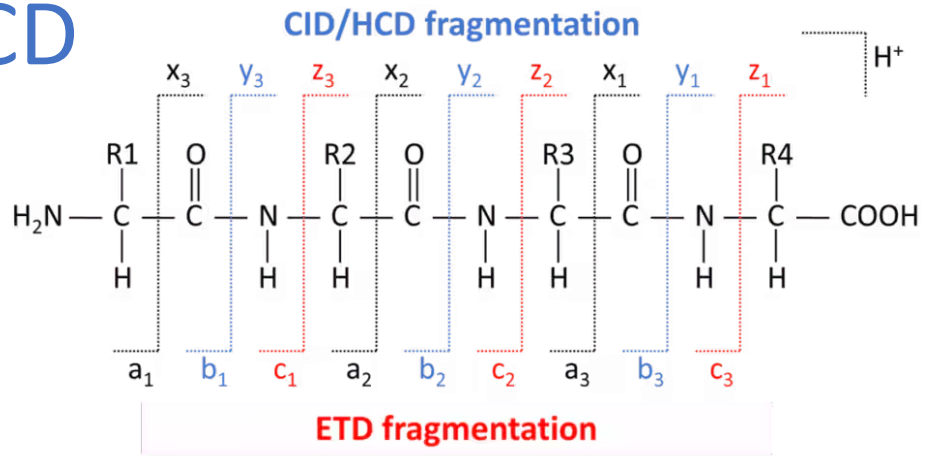
C fragments ions retain the N-terminus of the protein

z frag. C-terminus

```
A NELLLNVK
AN ELLLNVK
ANE LLLNVK
ANEL LLNVK
ANELL LNVK
ANELLL NVK
ANELLLN VK
ANELLLNV K
```

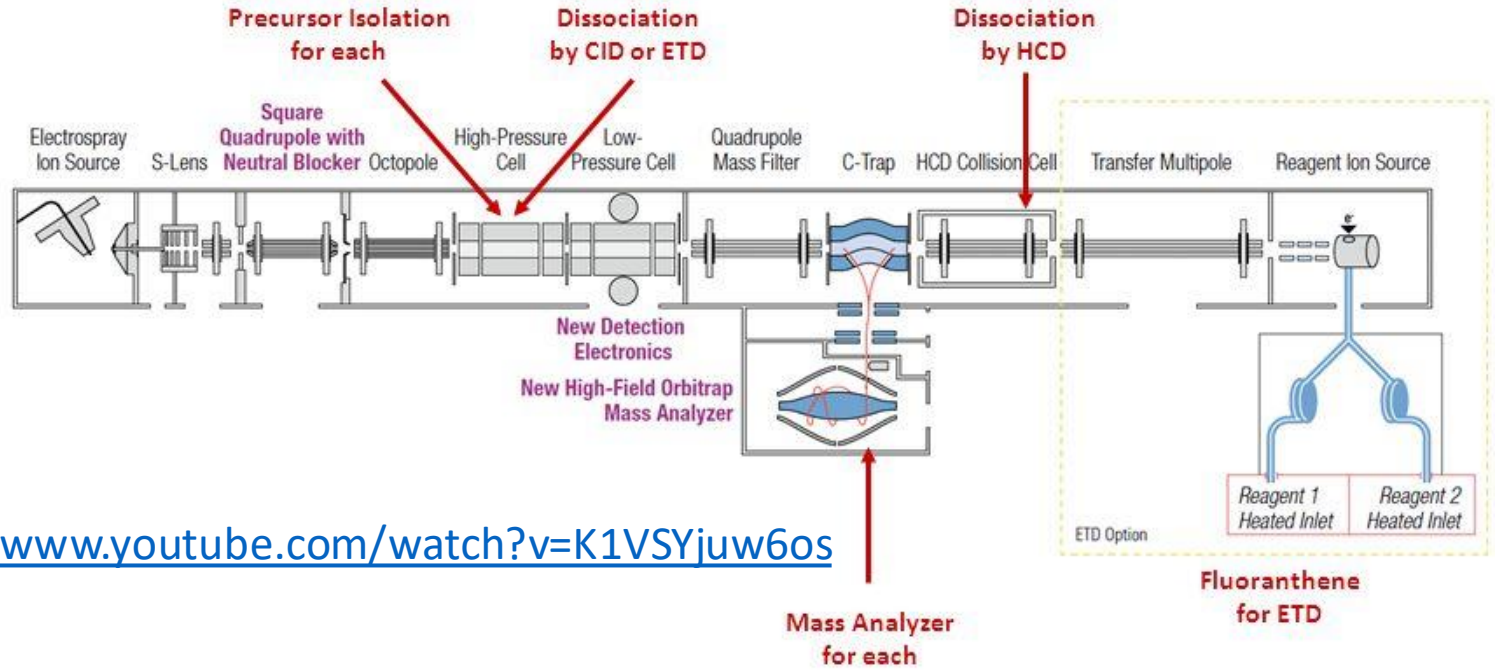
C frag. N-terminus

HCD



Higher-energy C-trap dissociation (HCD) is a **CID technique specific to the orbitrap mass spectrometer** in which fragmentation takes place external to the trap. The ions are then returned to the C-trap before injection into the orbitrap for mass analysis.

Orbitrap Elite, High Resolution MS/MS by CID, HCD, or ETD



https://www.youtube.com/watch?v=K1VSYjuw6os

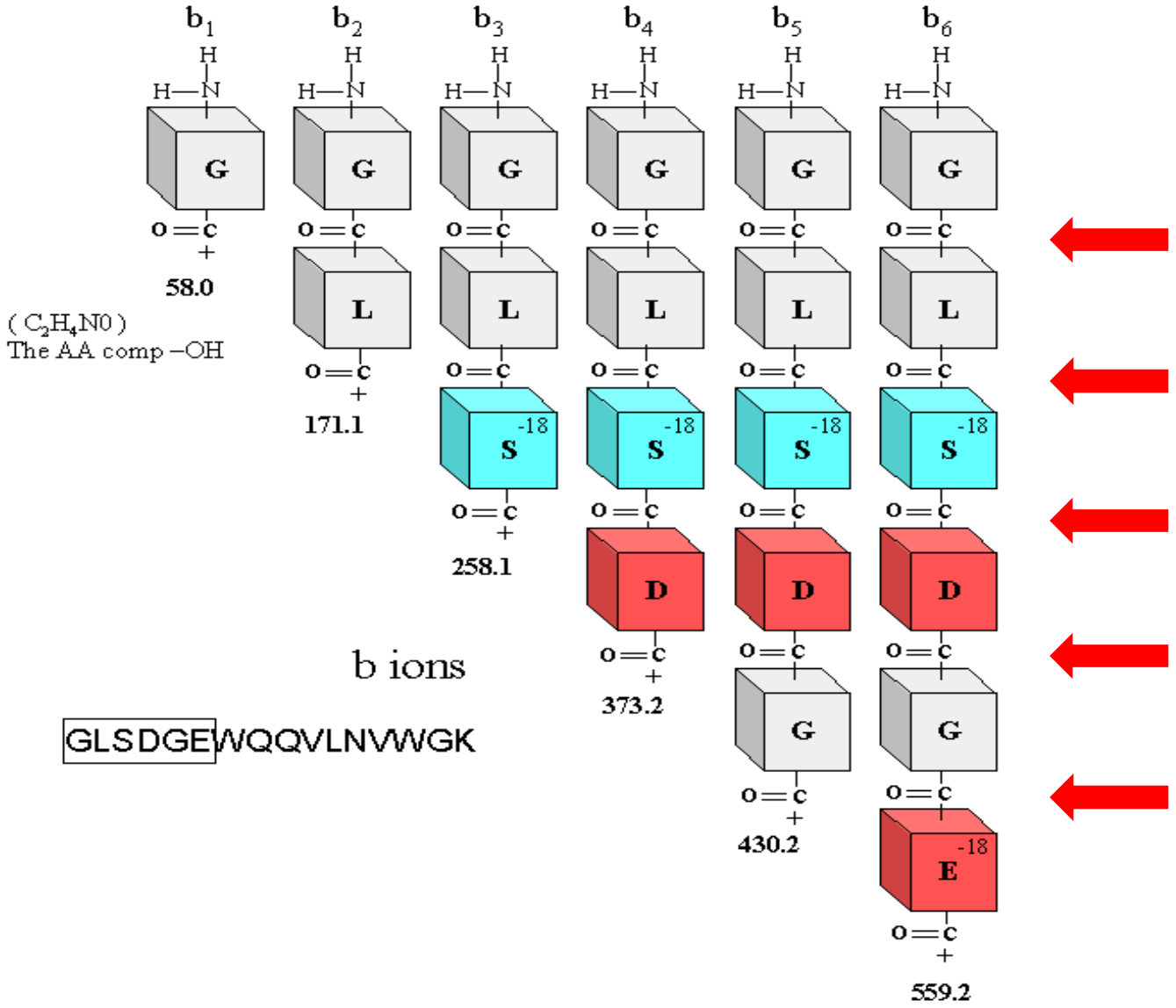
The cleavage generates y- and b-type fragment ions

The fragmentation of the peptides does not occur sequentially: the first fragmentation does not occur at the level of the N-terminal amino acid and then proceeds one residue at a time up to the C-terminal amino acid, but occurs randomly.

Some fragmentations are preferred with respect to others (weaker bonds) and therefore the intensity of the fragment ions is variable.

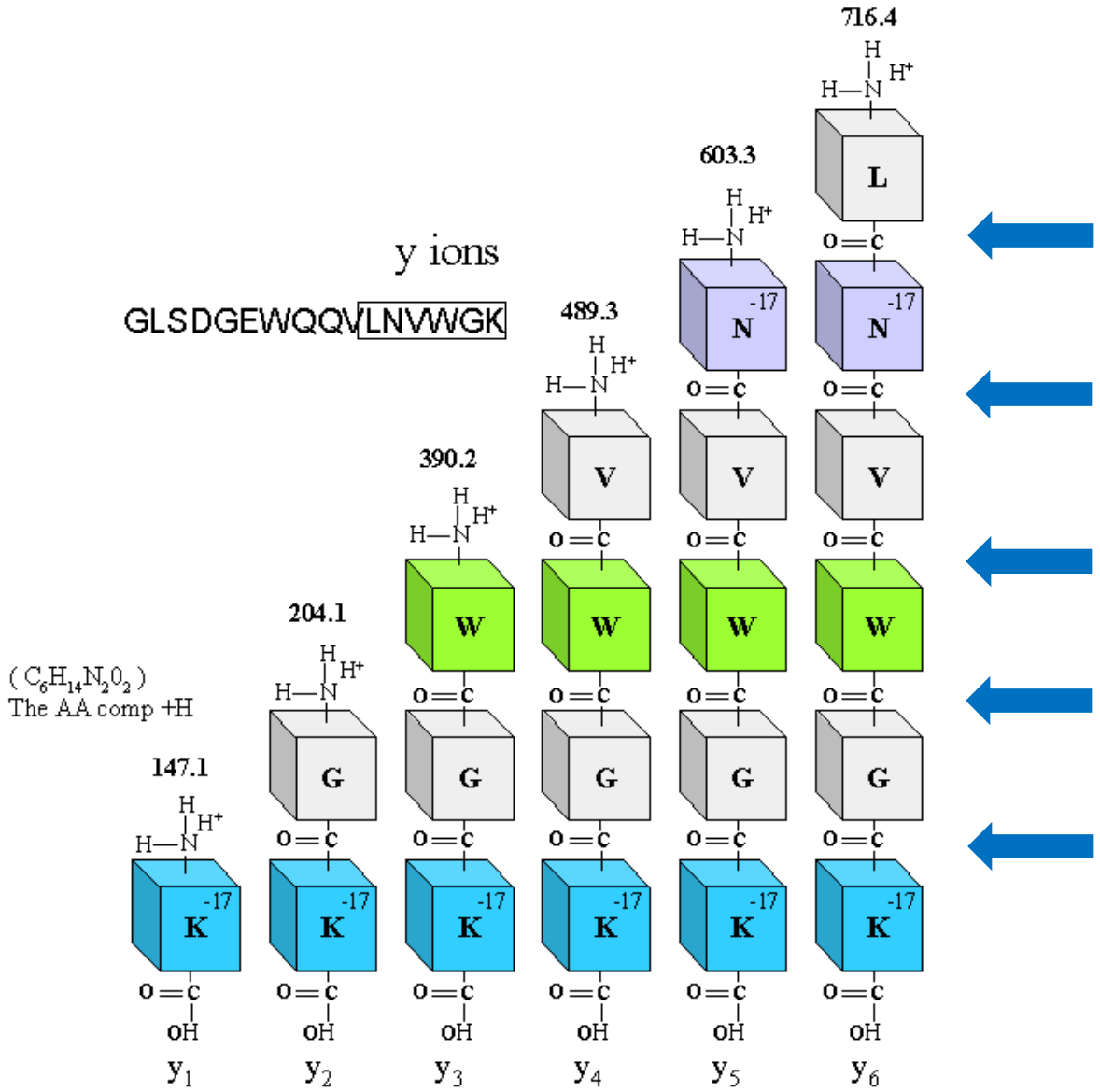
Most of the fragment ions (peaks) differ from the previous and / or subsequent ones for a mass value that roughly corresponds to the mass of an amino acid. Thus the sequence of the peptide can be established by the differences in mass of the fragment ions.

The tricky point concerns the correct attribution of the observed signals to the b series or the y series. If the attribution is wrong, the determined sequence could be the inverse of the real one.

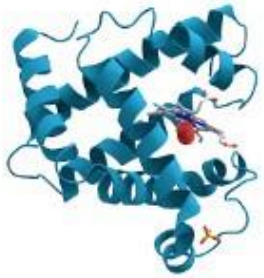


The first six b ions of a peptide. The m/z values of monocharged b ions correspond to the peptide mass without OH, or -17 Da.

The first six y ions of the same peptide. The m/z values of monocharged y ions correspond to the peptide mass.



protein smaple

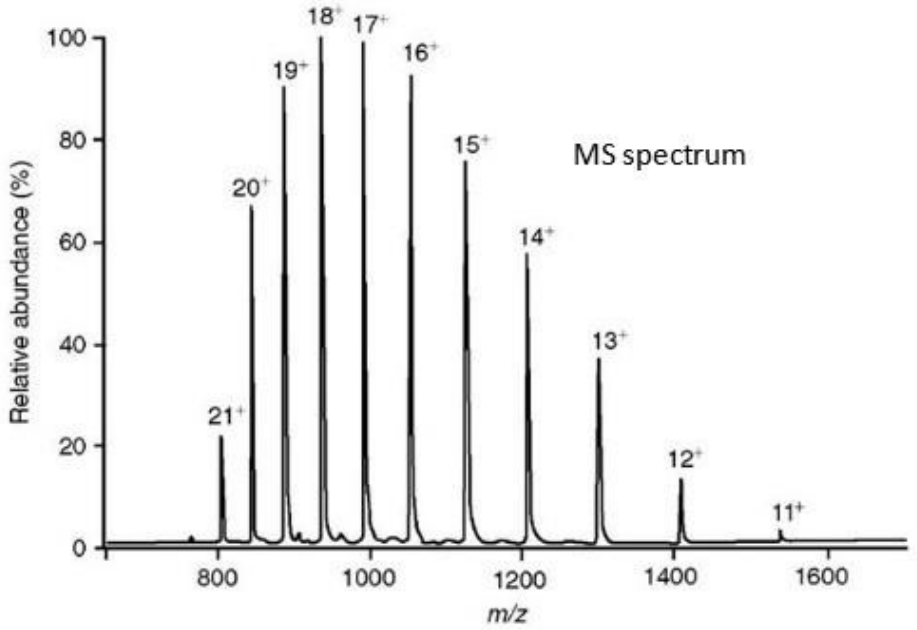


enzyme digestion

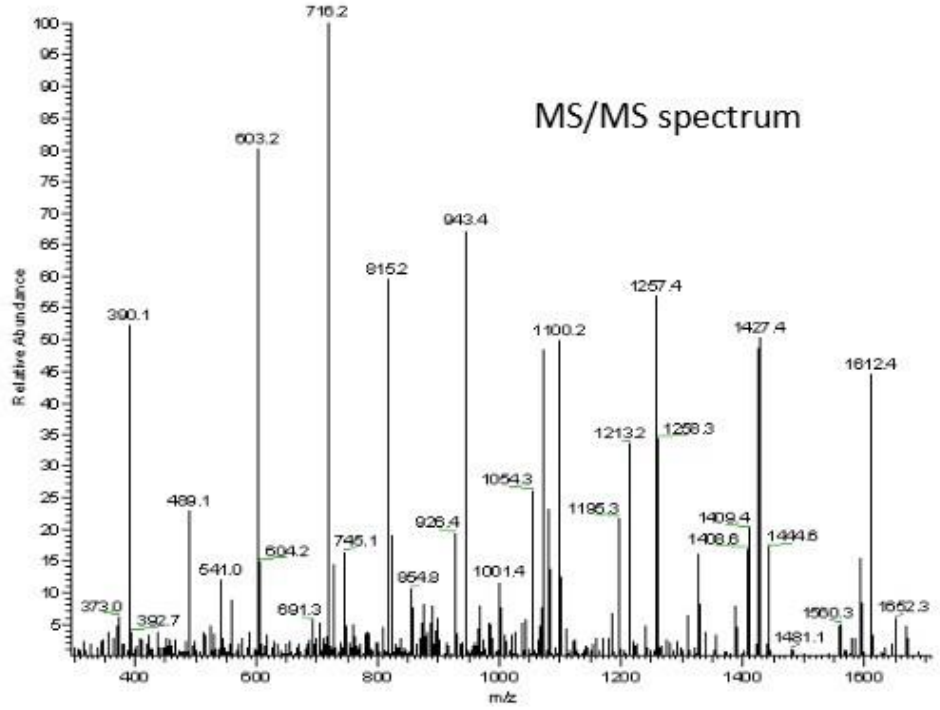
peptide mixture

HPLC

MS analysis



MS/MS analysis



Provides primary sequence

de novo sequencing

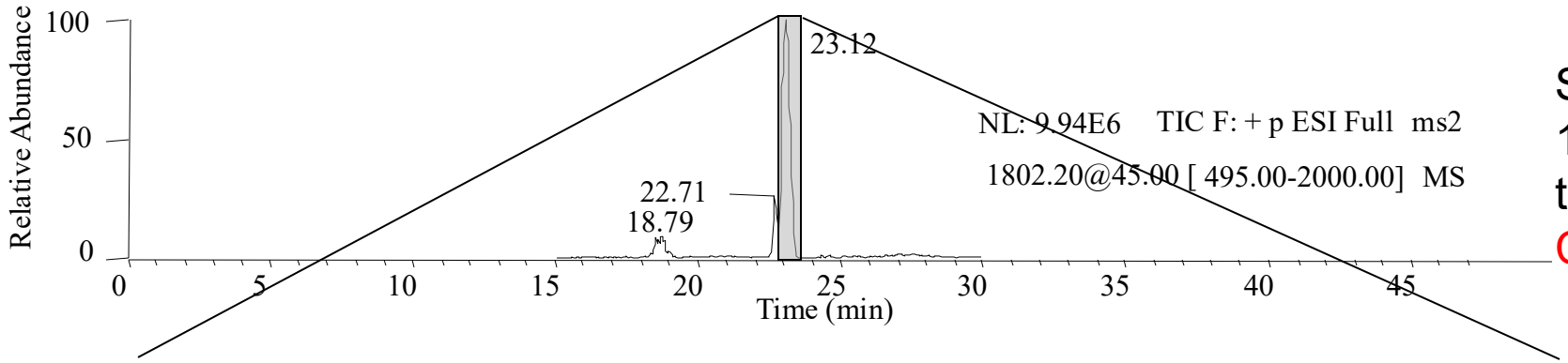
In mass spectrometry, **de novo peptide sequencing** is the method in which a peptide amino acid sequence is determined from tandem mass spectrometry.

MS MS experiments on a tryptic peptide from a salivary protein eluting at 23.12 min (PRP1)

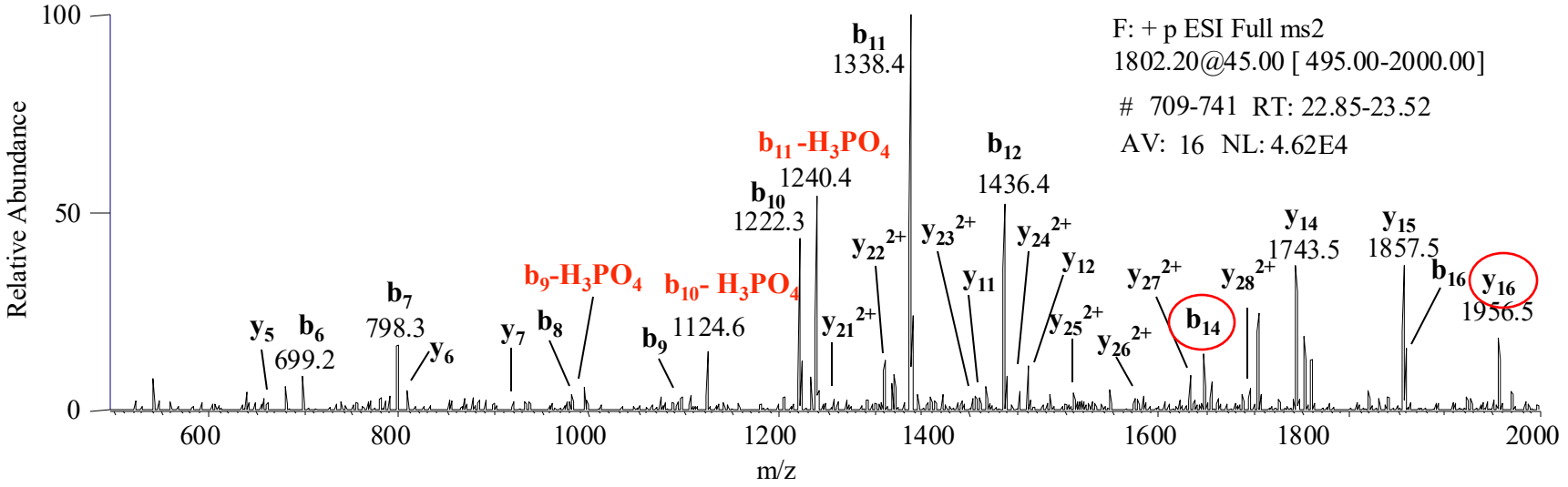
y16

<QDLDEDVSQEDVPLVISDGGDSEQFIDEER aPRPs fr. 1-30

b14



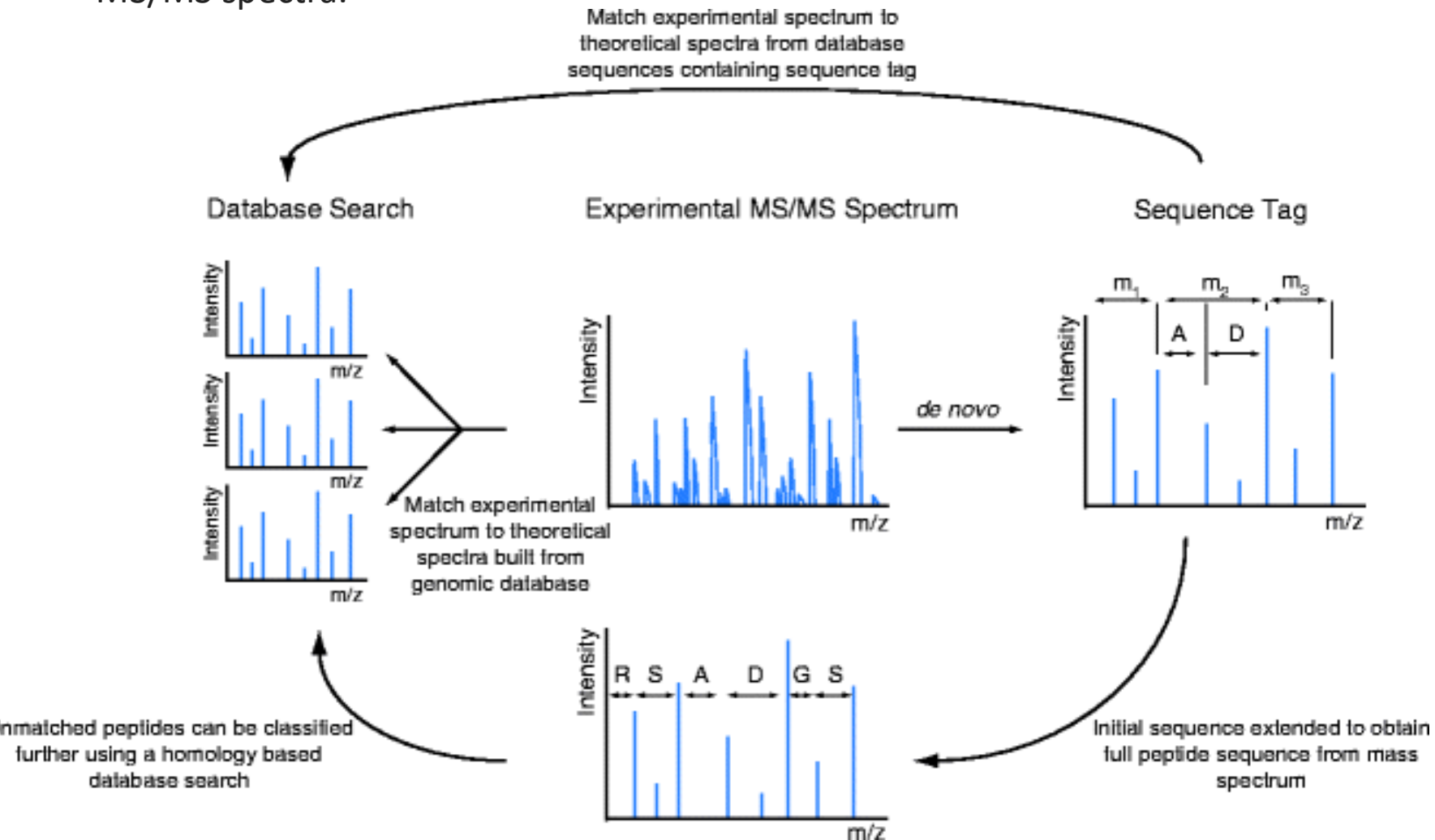
Select Ion at 1802.20 and Send to Collision Cell for CID dissociation



De novo sequencing

Utilize computational approaches to deduce the sequence or partial sequence of peptides directly from the experimental MS/MS spectra.

Manual inspection of the MS/MS spectrum

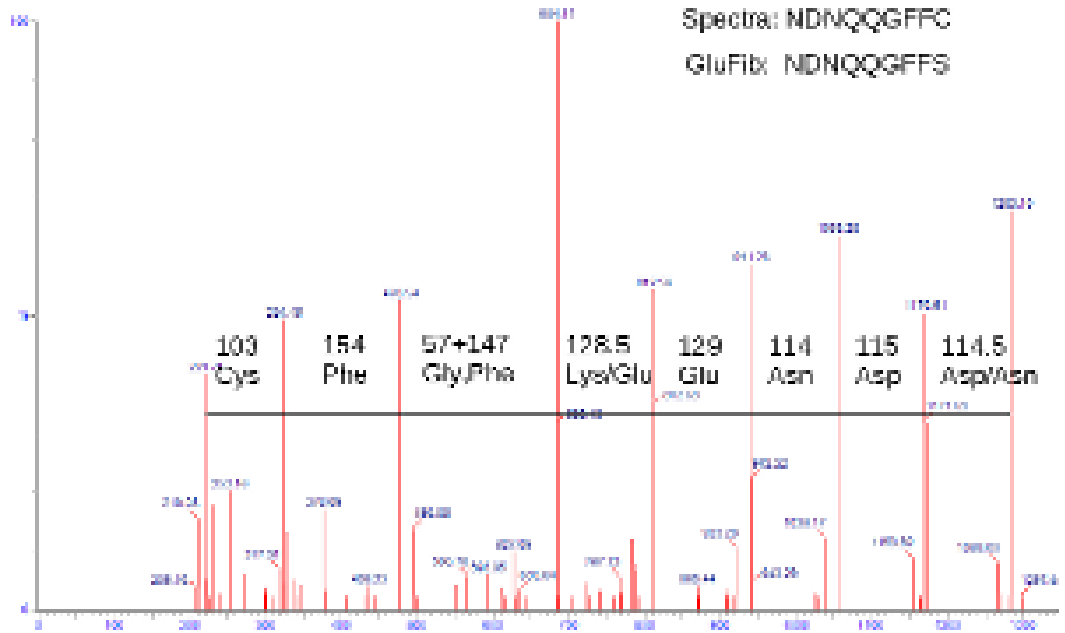


De novo sequencing of tandem (MS/MS) mass spectra represents the only way to determine the sequence of proteins from organisms with unknown genomes.

Residue	3/1 letter code	Composition	Mono. mass	Avg. mass
Alanine	Ala / A	C_3N_5NO	71.03712	71.08
Arginine	Arg / R	$C_6H_{12}N_4O$	156.10112	156.19
Asparagine	Asn / N	$C_4H_6N_2O_2$	114.04293	114.10
Aspartic acid	Asp / D	$C_4H_5NO_3$	115.02695	115.09
Cysteine	Cys / C	C_3H_5NOS	103.00919	103.14
Cysteine-cm [1]	Cys / C	$C_5H_8N_2O_2S$	160.03065	160.20
Glutamine	Gln / Q	$C_5H_8N_2O_2$	128.05858	128.13
Glutamic acid	Glu / E	$C_5H_7NO_3$	129.04260	129.12
Glycine	Gly / G	C_2H_3NO	57.02147	57.05
Histidine	His / H	$C_6H_7N_3O$	137.05891	137.14
Isoleucine	Ile / I	$C_6H_{11}NO$	113.08407	113.16
Leucine	Leu / L	$C_6H_{11}NO$	113.08407	113.16
Methionine	Met / M	C_5H_9OS	131.04049	131.19
Methionine-ox [2]	Met / M	$C_5H_9O_2S$	147.03540	147.18
Phenylalanine	Phe / F	C_9H_9NO	147.06842	147.18
Proline	Pro / P	C_5H_7NO	97.05277	97.12
Serine	Ser / S	$C_3H_5NO_2$	87.03203	87.08
Threonine	Thr / T	$C_4H_7NO_2$	101.04768	101.10
Tryptophan	Trp / W	$C_{11}H_{10}N_2O$	186.07932	186.21
Tyrosine	Tyr / Y	$C_9H_9NO_2$	163.06333	163.18
Valine	Val / V	C_5H_9NO	99.06842	99.13

The difference observed between one fragment and the next in the two series corresponds to the relative mass of the amino acid minus one molecule of water (18 Da).

For example, the mass of alanine is 89.0 Da but the differences observed in the loss of an alanine are 71.0 Da,

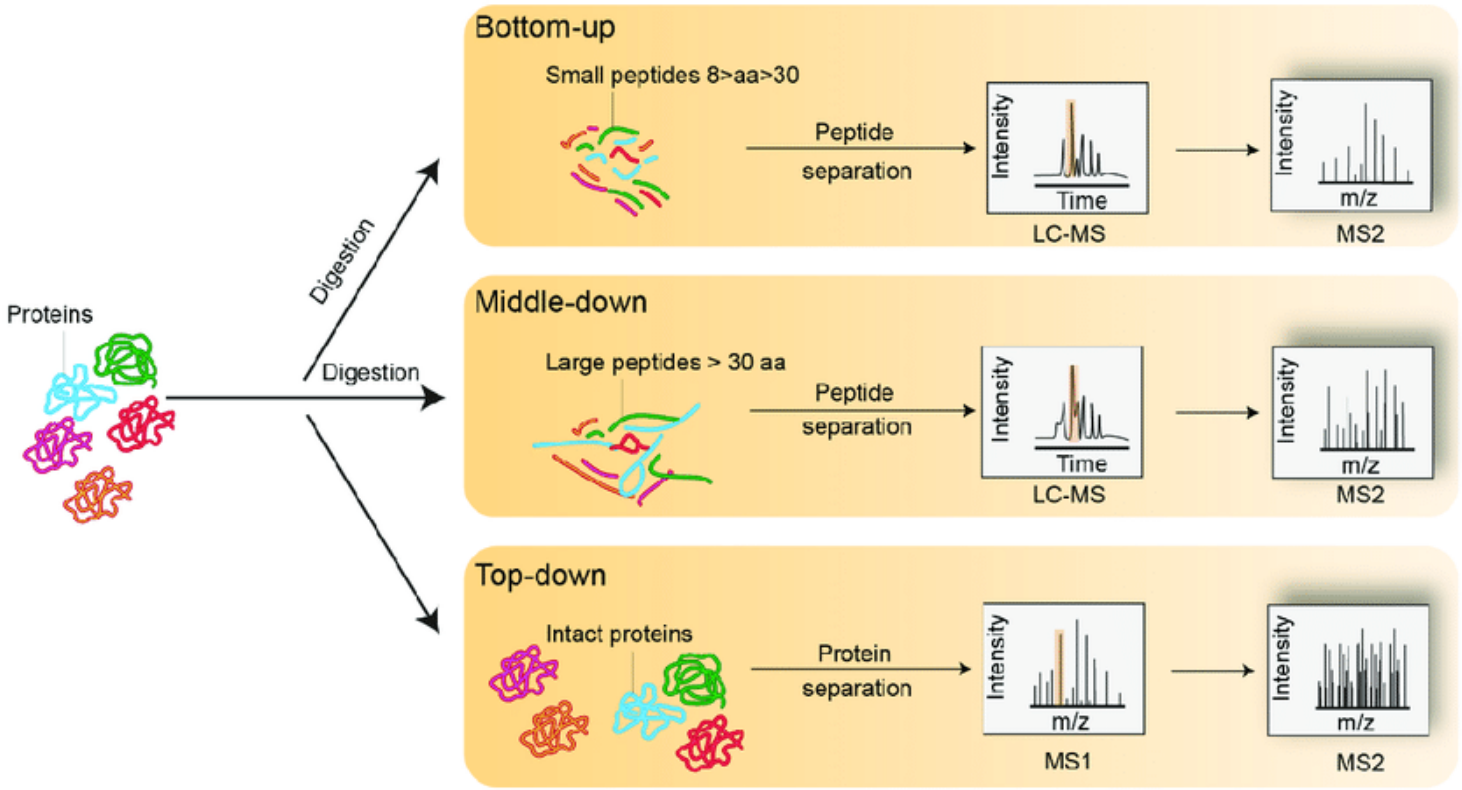


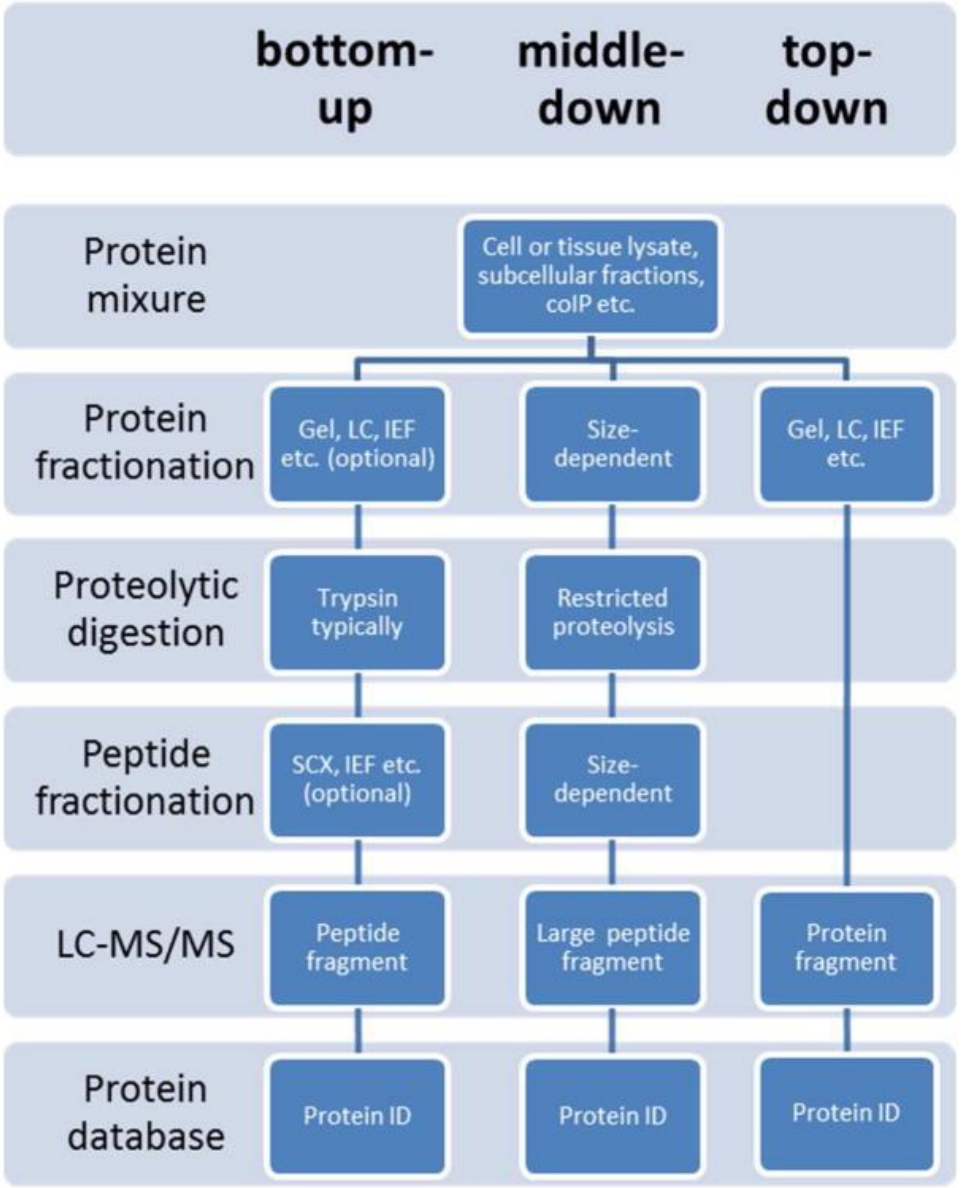
CONCLUSIONS

Why *de novo* sequencing is difficult

1. Leucine and isoleucine have the same mass
2. Glutamine and lysine differ in mass by 0.036Da
3. Phenylalanine and oxidized methionine differ in mass by 0.033Da
4. Cleavages do not occur at every peptide bond (or cannot be observed on the MS-MS
   - Poor quality spectrum (some fragment ions are below noise level)
   - The C-terminal side of proline is often resistant to cleavage
   - Absence of mobile protons
   - Peptides with free N-termini often lack fragmentation between the first and second amino acids

Exist another way to classify the Proteomic approaches



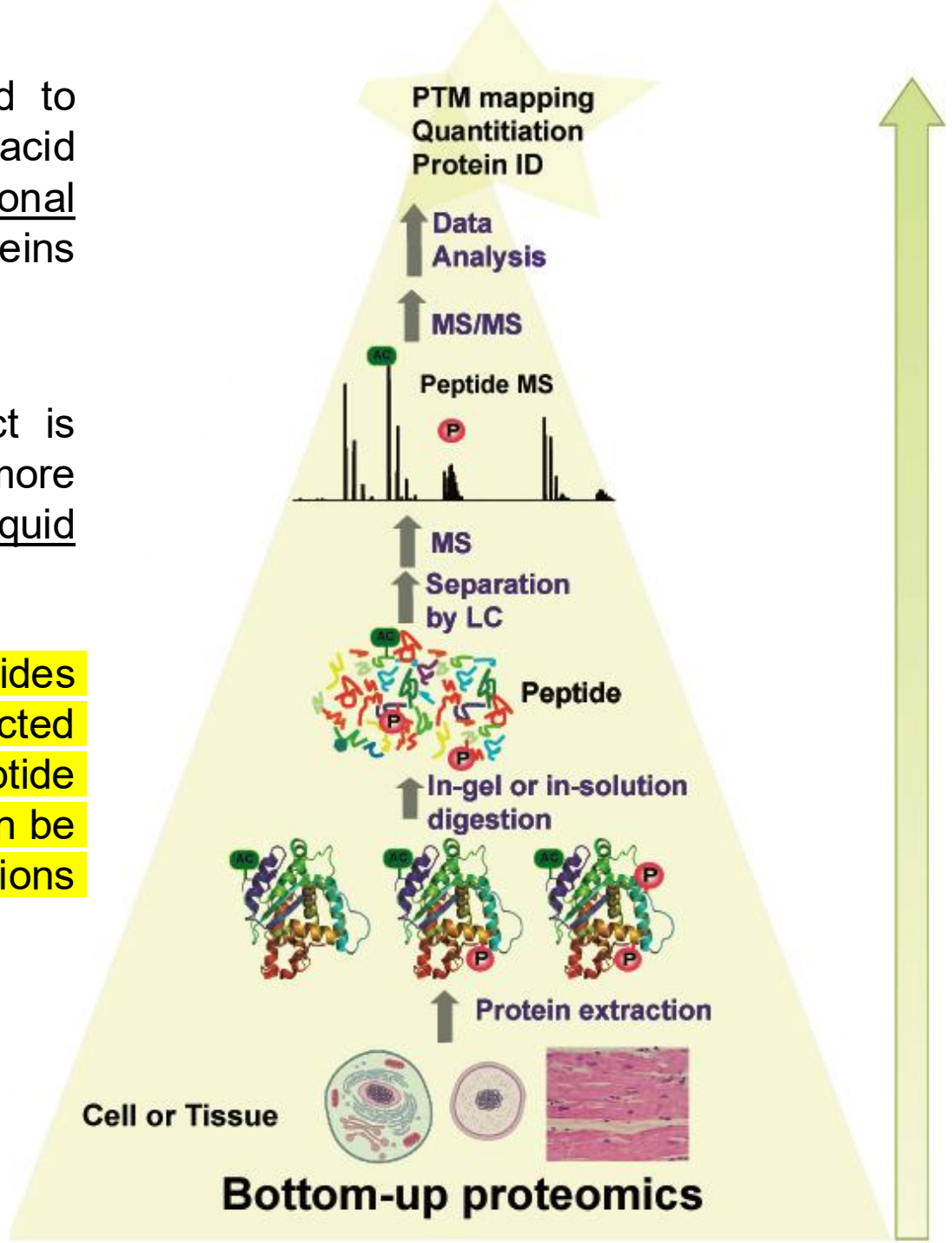


Bottom-up proteomics is a common method to identify proteins and characterize their amino acid sequences and post-translational modifications by proteolytic digestion of proteins prior to analysis by mass spectrometry.

In **bottom-up proteomics**, the protein extract is enzymatically digested, followed by one or more dimensions of separation of the peptides by liquid chromatography coupled to mass spectrometry.

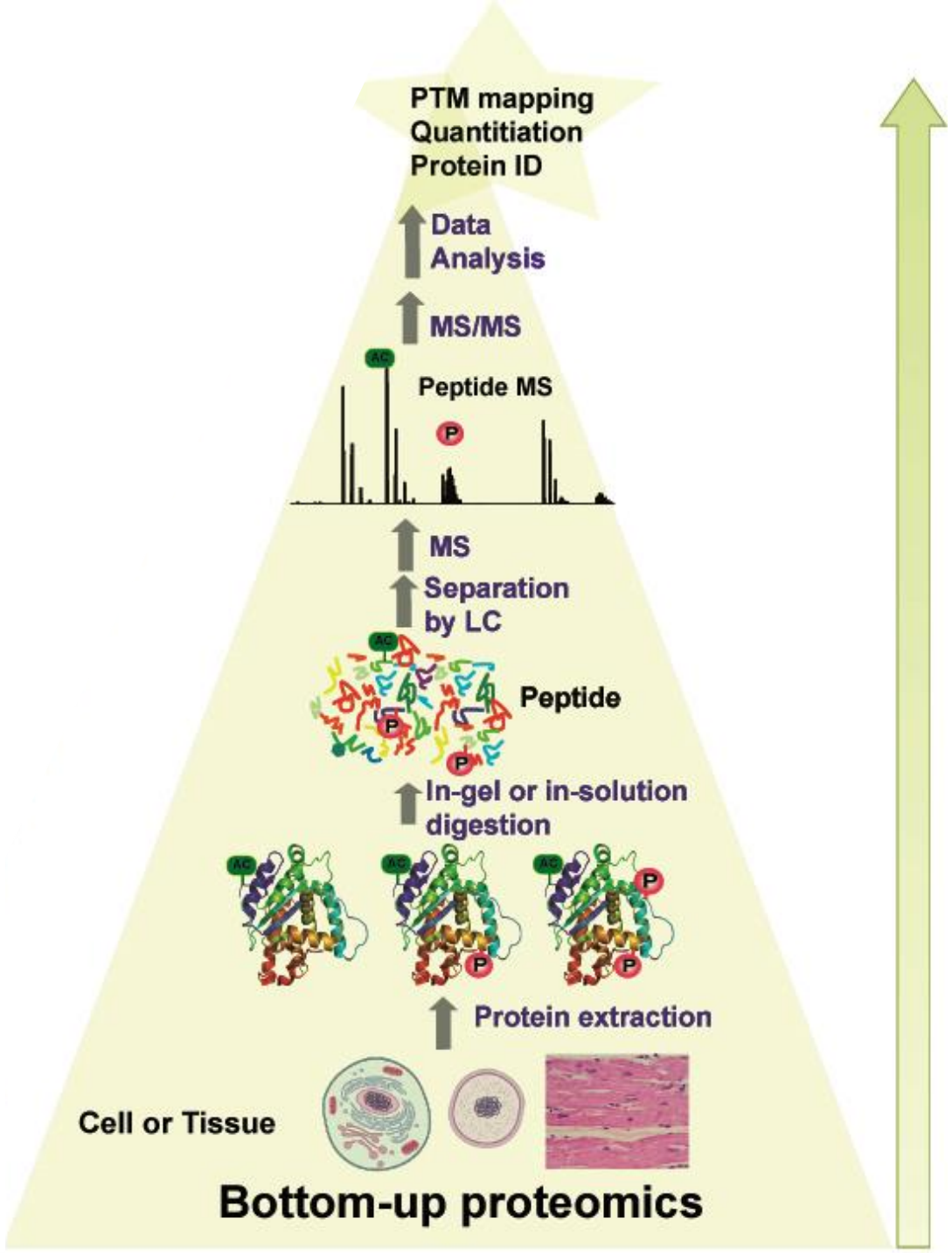
By comparing the masses of the proteolytic peptides or their tandem mass spectra with those predicted from a sequence database or annotated peptide spectral in a peptide spectral library, peptides can be identified and multiple peptide identifications assembled into a protein identification.

Two different approaches are used to identify proteins in bottom-up proteomics - peptide mass fingerprinting and tandem MS (MS–MS).



Usually Bottom-up proteomics is performed on a sigle purified protein or on a very simple protein mixture.

When bottom-up is performed on a more complex mixture of proteins it is called shotgun proteomics, a name coined by the Yates lab because of its analogy to shotgun genomic sequencing.

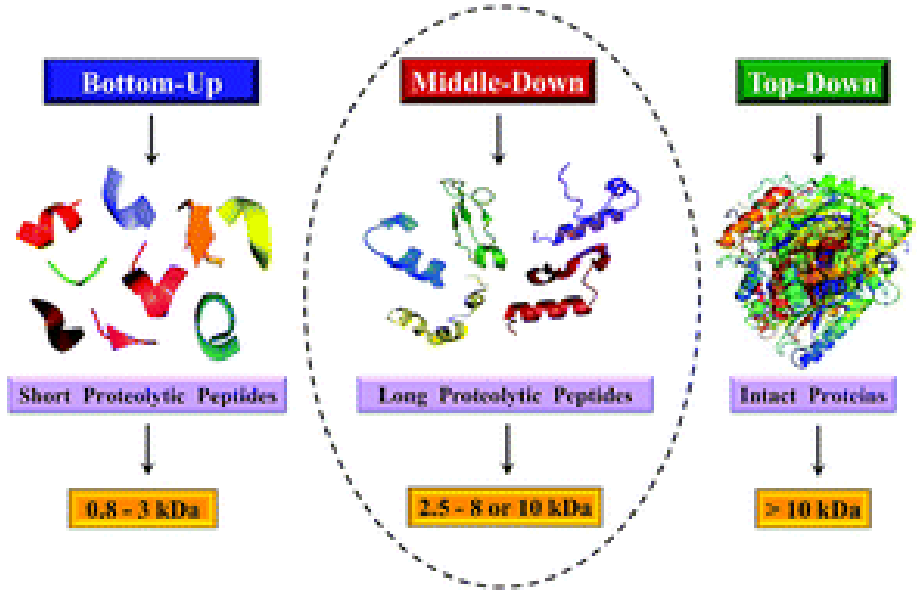


Middle-down proteomics analyzes larger peptide fragments than bottom-up proteomics, minimizing peptide redundancy between proteins.

Table 1. Distrubution of peptide fragment length from 20,639 proteins

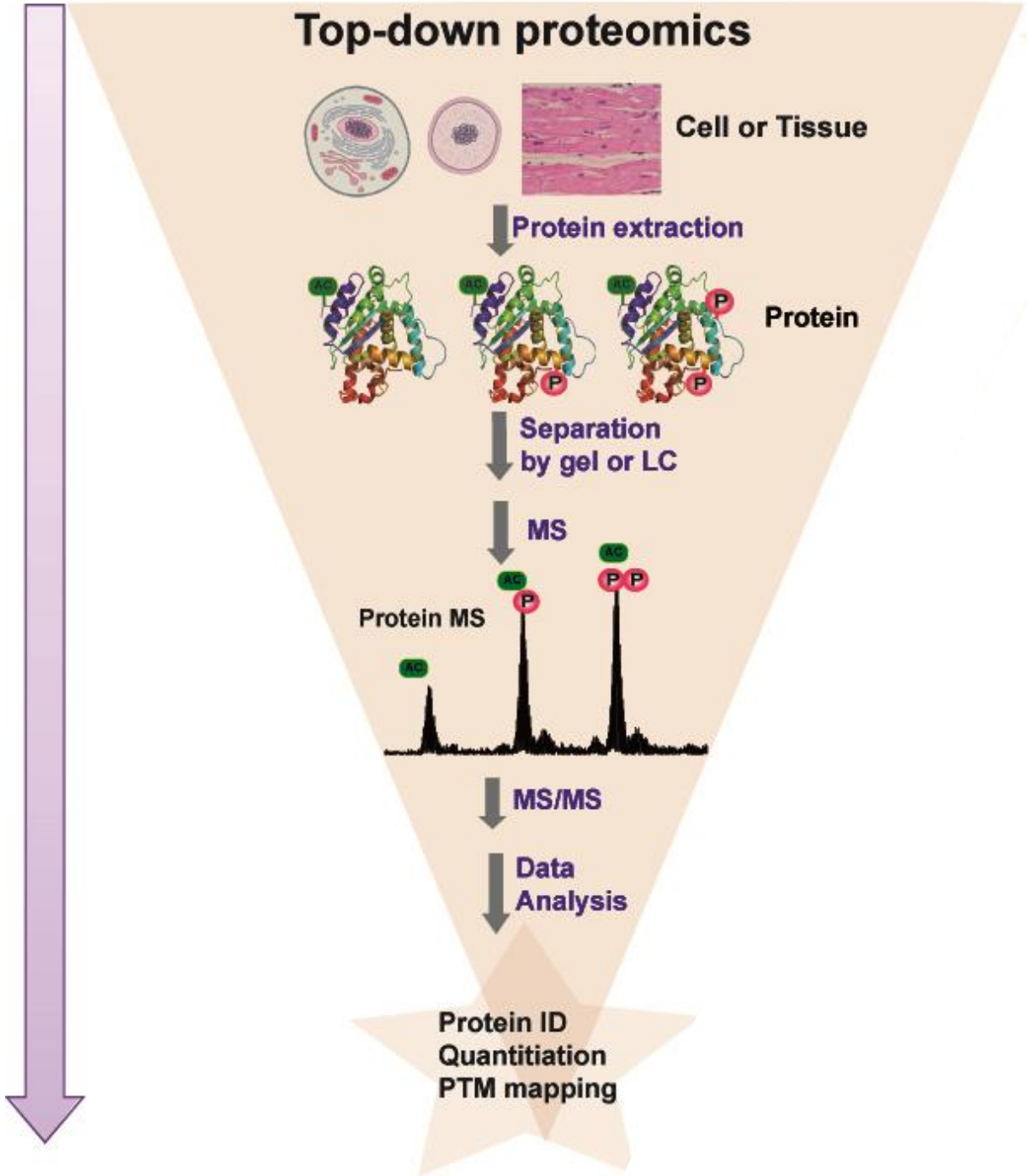
Enzyme/reagent	Residues cleaved	Total fragments	Avg. fragment length
Trypsin	K/R	662,981	8
Lys-C	K	359,140	16
Asp-N	D	321,655	18
Cyanogen bromide	M	150,605	38
Hydroxylamine	N-G	36,643	152
Dilute acid	D-P	35,574	166

- Limited proteolysis with trypsin (1 h) to obtain fragments 25-30 AA in length.
- By using other enzymes/reagents which produce larger fragments



Advantages in using trypsin in proteomics

- It is chip and easy to isolate
- It has a high specificity which is easy to predict
- It is a robust enzyme
- It works over a range of concentrations and conditions
- It works very well on denaturated proteins
- It generates peptides of a good length for BOTTOM-UP mass spec analyses
- All tryptic peptides contain at least an Arg or Lys with their positive charge



Proteins are not enzymatically digested into peptides. Still requires separation of the intact proteins from complex biological samples, and this can be achieved using conventional techniques such as liquid chromatography.

Protein sequence information is obtained by fragmentation of the intact proteins via dissociation methods such as HCD, electron-capture dissociation (ECD) and electron-transfer dissociation (ETD) and/or combinations.

Top-down proteomics is capable of identifying and quantitating unique proteoforms through the analysis of intact proteins.

Comparison between top-down and bottom-up proteomics: advantages and disadvantages

	Bottom-up MS
Protein identification	+++ (More robust and high throughput strategy for protein identification with better bioinformatics tools currently available)
Protein modification	++ (Amenable for large-scale PTM study, but with limited sequence coverage and loss of connectivity between PTMs resulting from peptide loss during sample preparation and sample digestion)
Protein quantification	+++ (Very well-developed methods available for relative and absolute quantitation of protein expression level but with limitation in quantification of peptides with PTMs)

	Top-down MS
Protein identification	+ (More reliable protein identification particularly for proteins with high sequence similarity such as alternatively spliced isoforms, but is of relatively lower throughput)
Protein modification	+++ (Reliable and comprehensive analysis of all types of PTMs simultaneously without a priori knowledge with full sequence coverage, but proteins need to be purified or separated prior to detailed MS study)
Protein quantification	++ (Accurate relative quantification of multiple proteoforms from the same gene product due to genetic variations, alternatively spliced RNA transcripts and PTMs, but the quantification of protein expression level remains underdeveloped.