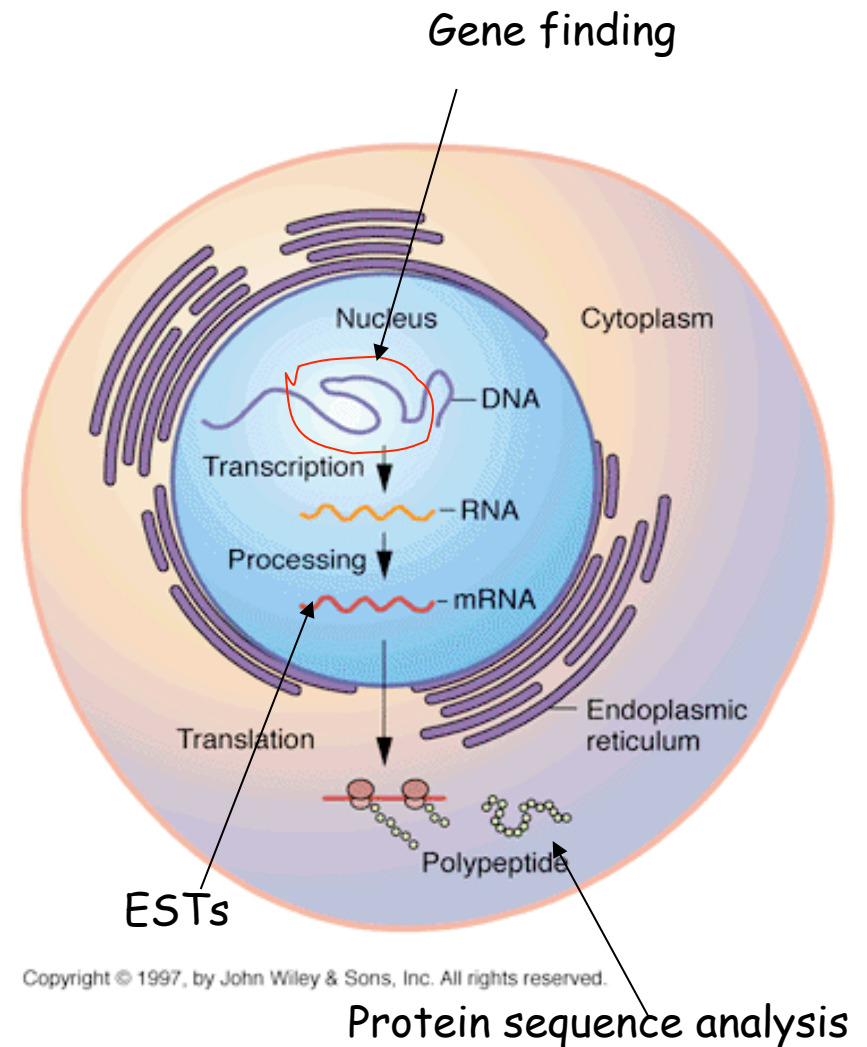


# CSE182-L11

Protein sequencing and Mass Spectrometry

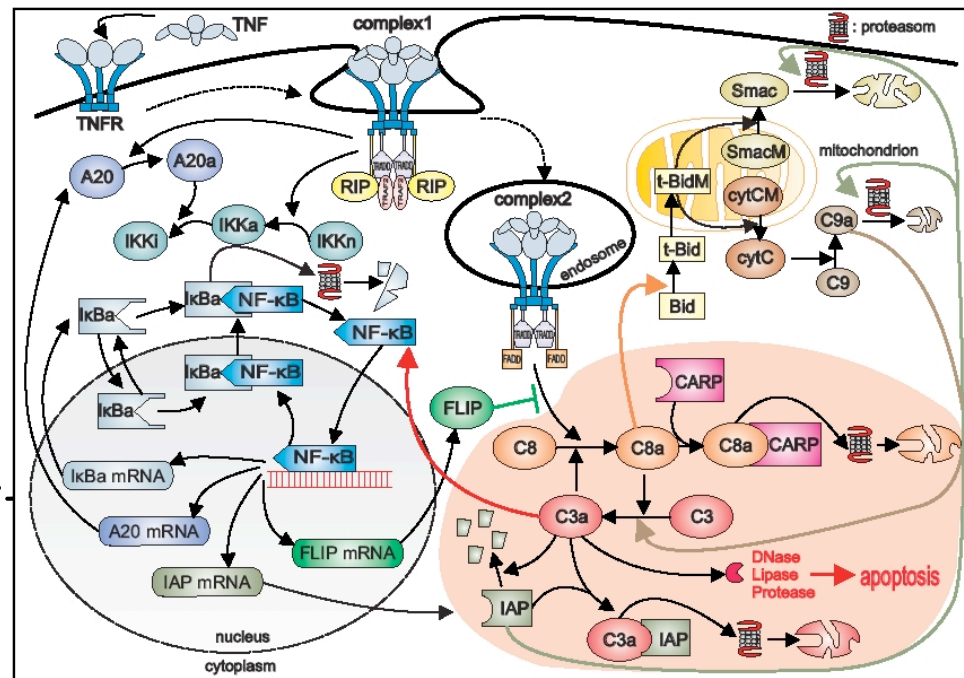
# Course Summary

- Sequence Comparison (BLAST & other tools)
- Protein Motifs:
  - Profiles/Regular Expression/HMMs
- Discovering protein coding genes
  - Gene finding HMMs
  - DNA signals (splice signals)
- ~~• How is the genomic sequence itse obtained?
  - LW statistics
  - Sequencing and assembly~~
- Next topic: the dynamic aspects of the cell



# The Dynamic nature of the cell

- The molecules in the body, RNA, and proteins are constantly turning over.
  - New ones are 'created' through transcription, translation
  - Proteins are modified post-translationally,
  - 'Old' molecules are degraded

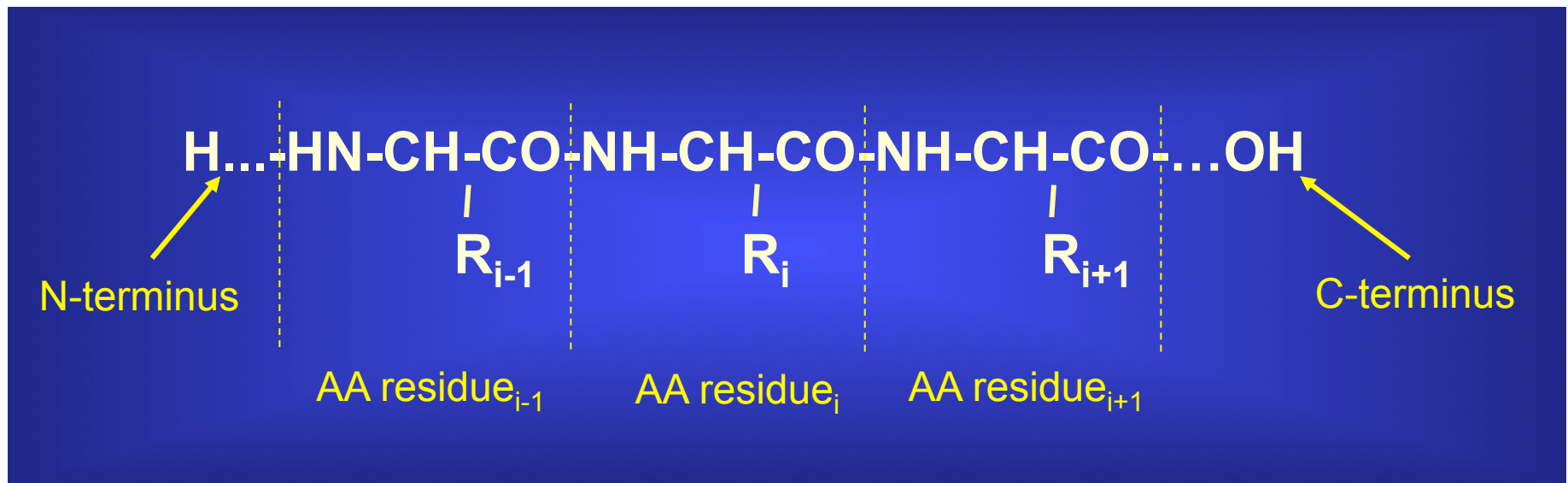


# Dynamic aspects of cellular function

- Expressed transcripts
  - Microarrays to 'count' the number of copies of RNA
- Expressed proteins
  - Mass spectrometry is used to 'count' the number of copies of a protein sequence.
- Protein-protein interactions (protein networks)
- Protein-DNA interactions
- Population studies

# The peptide backbone

The peptide backbone breaks to form fragments with characteristic masses.



# Mass Spectrometry

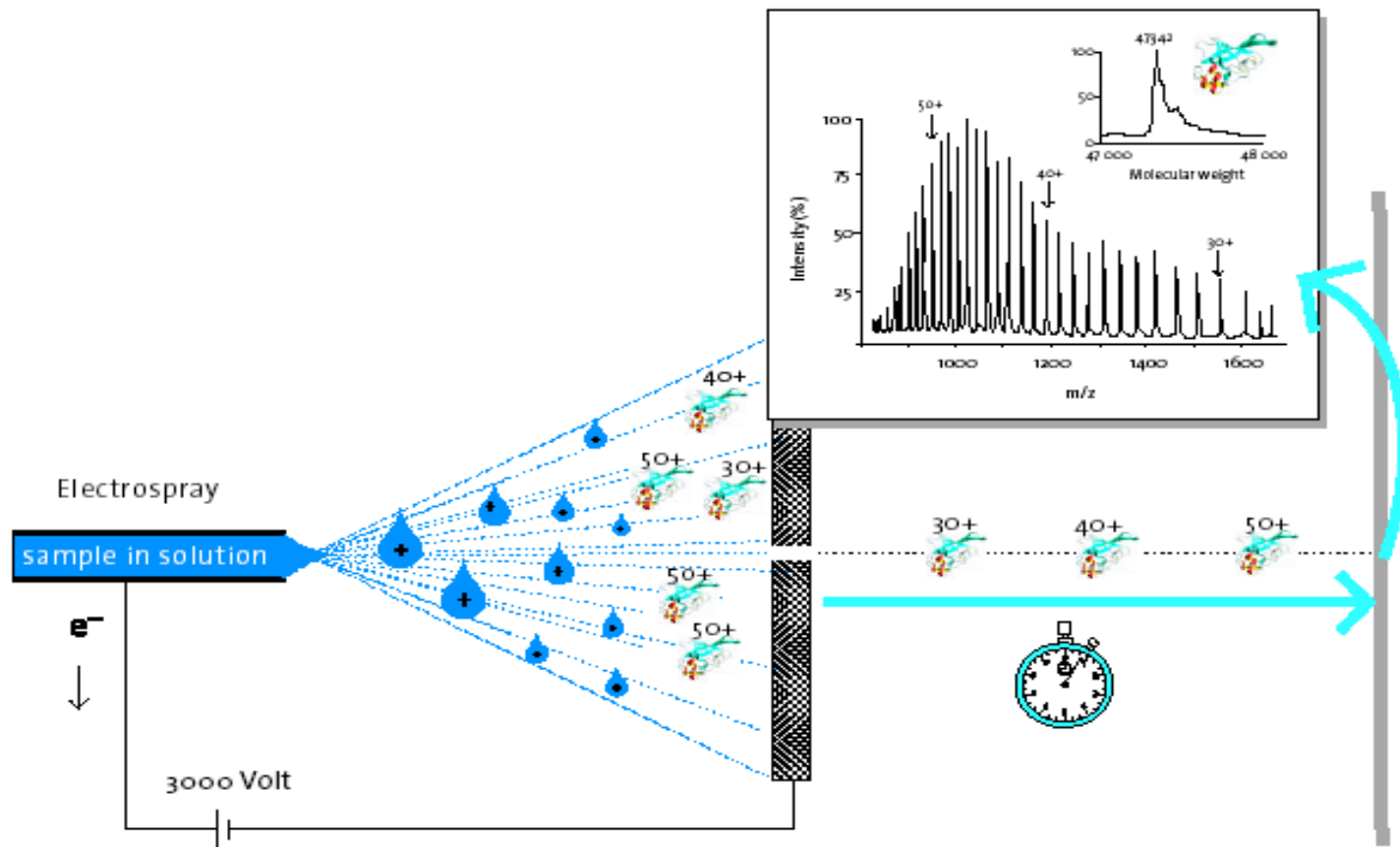


Figure 1. The electrospray process.

# Nobel citation '02

detection has been most important for molecular weight determinations of biological macromolecules. It is today a cornerstone of proteomics.

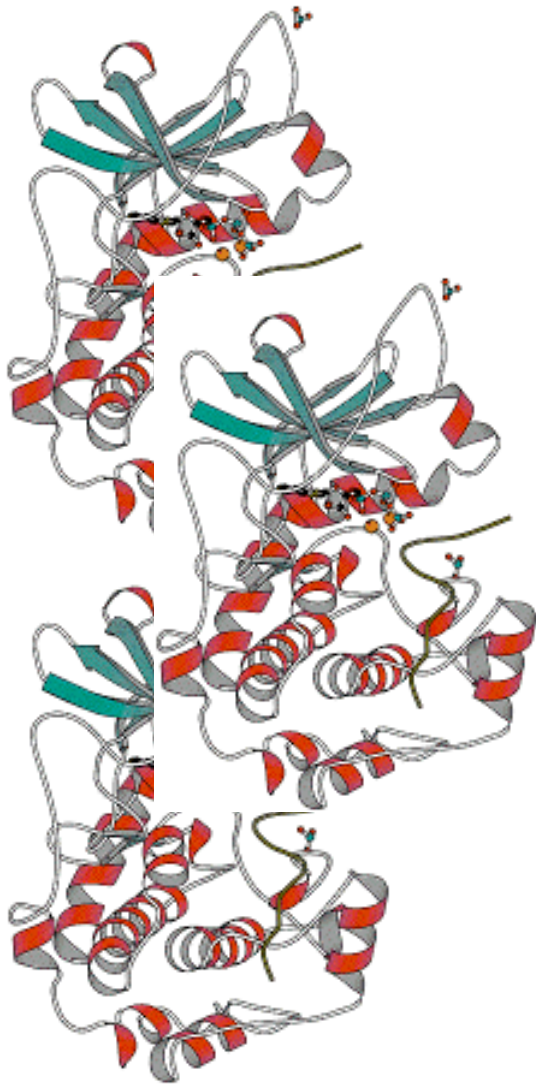
## RECENT DEVELOPMENTS AND APPLICATIONS OF MS

Some five years ago, mass spectrometry definitively crossed the border to biochemistry. The general ways that it provides structural determination, identification and trace level analysis have many applications in the biochemical field. It has become an attractive alternative to Edman sequencing, earlier dominant, and has an unsurpassed ability to identify posttranscriptional modifications and non-covalent interactions in for example antigen-antibody binding studies for identifying ligands to orphan receptors. An important development was the demonstrated usefulness of the MS technique for protein identification after 2-D gel electrophoretic separation and after liquid separation. With the recent interest in microfabricated devices for sample preparation, low-flow (nano-flow) techniques are being developed for optimised utilization of time and sensitivity.

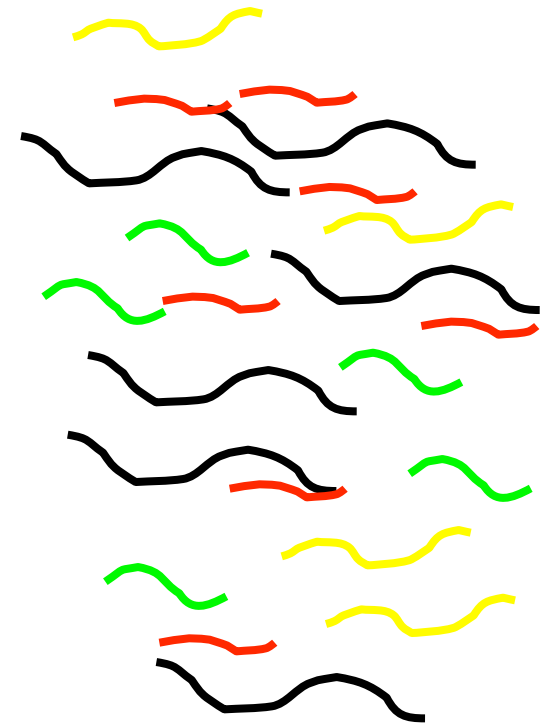
# The promise of mass spectrometry

- Mass spectrometry is coming of age as the tool of choice for proteomics
  - Protein sequencing, networks, quantitation, interactions, structure....
- Computation has a big role to play in the interpretation of MS data.
- We will discuss algorithms for
  - Sequencing, Modifications, Interactions..

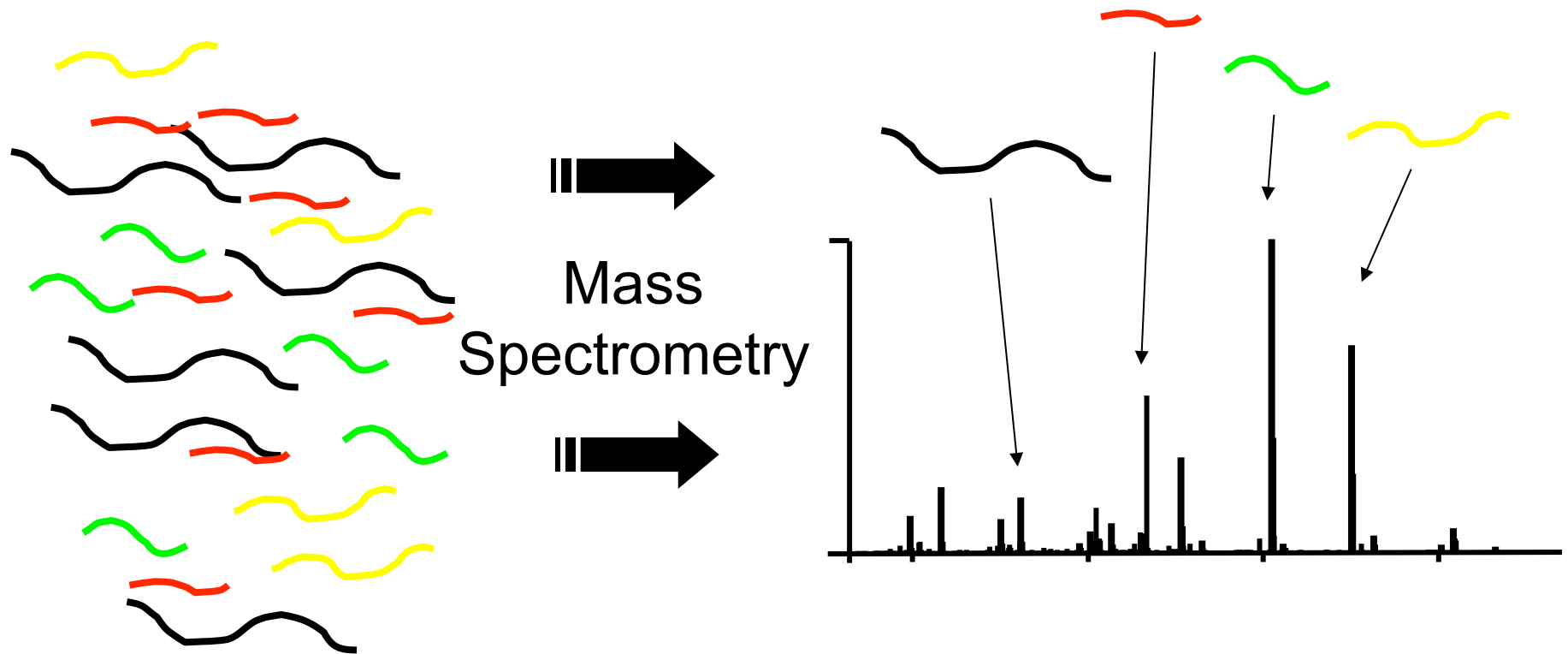
# Sample Preparation



Enzymatic Digestion  
(Trypsin)  
+  
Fractionation

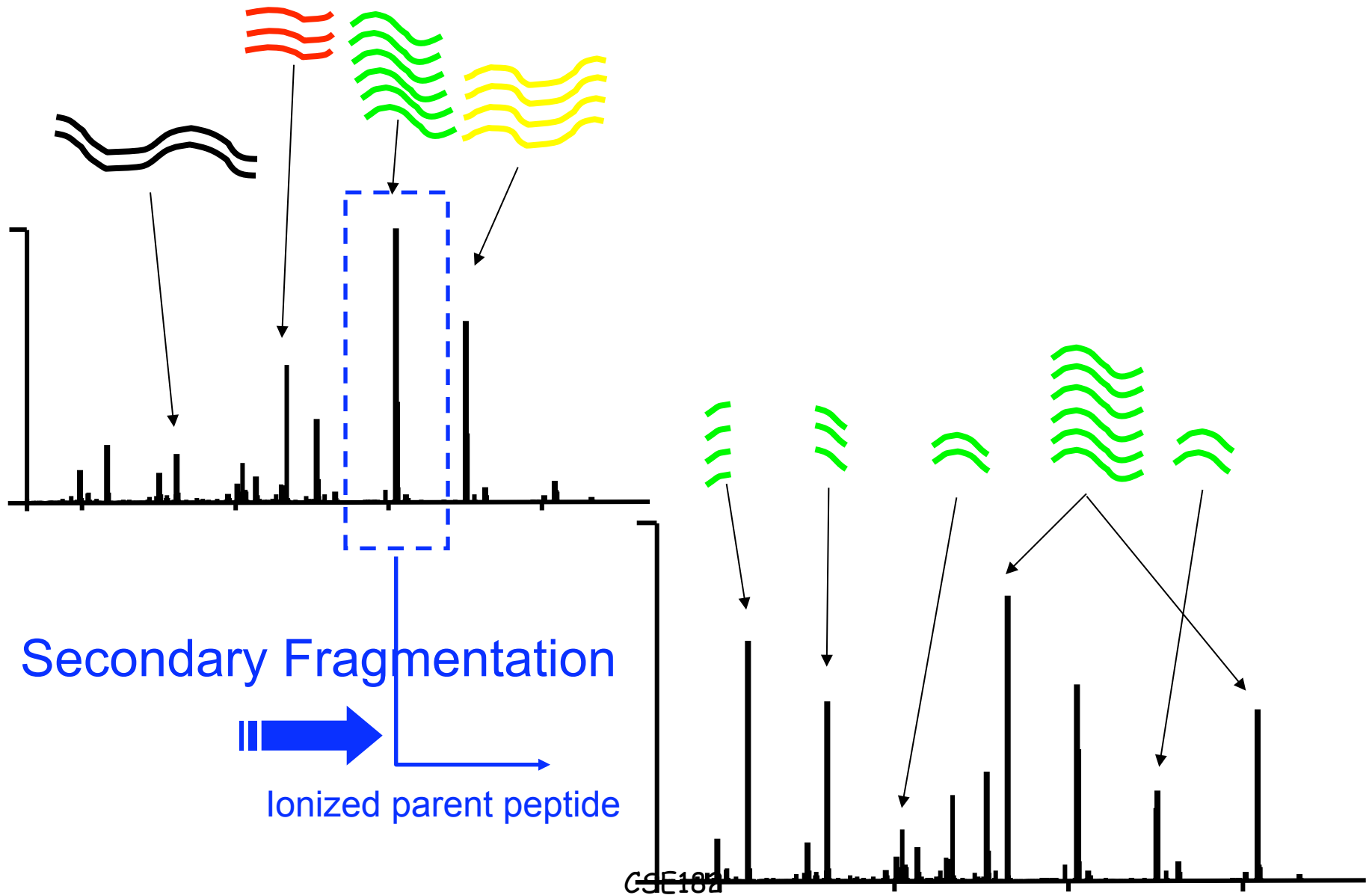


# Single Stage MS



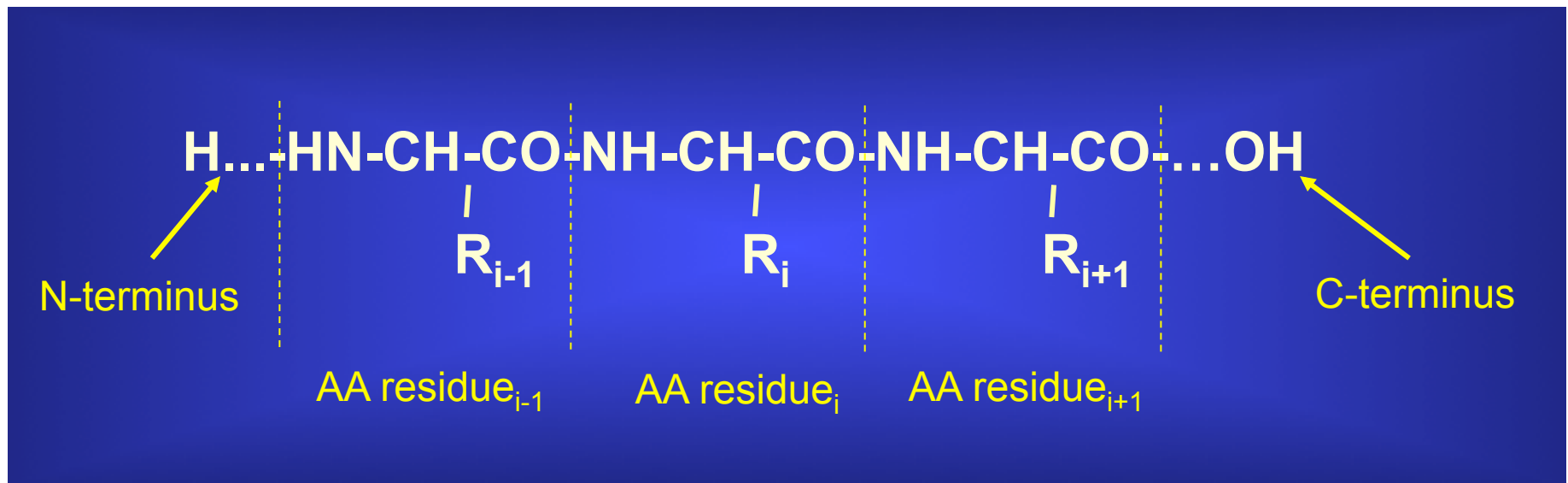
LC-MS: 1 MS spectrum / second

# Tandem MS



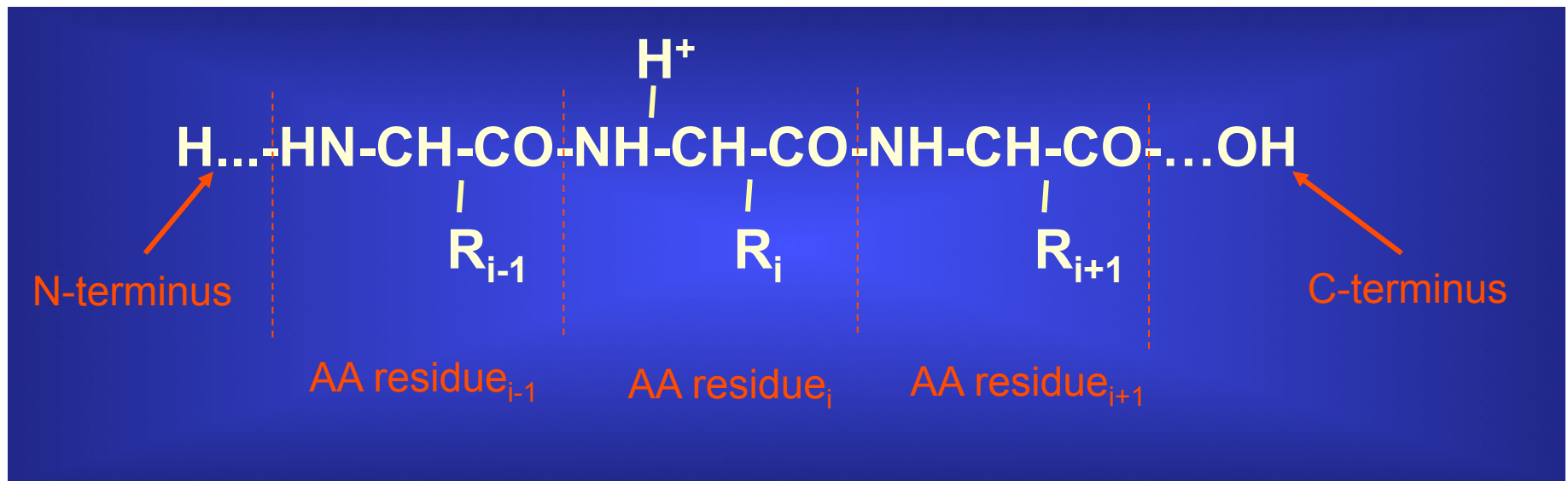
# The peptide backbone

The peptide backbone breaks to form fragments with characteristic masses.



# Ionization

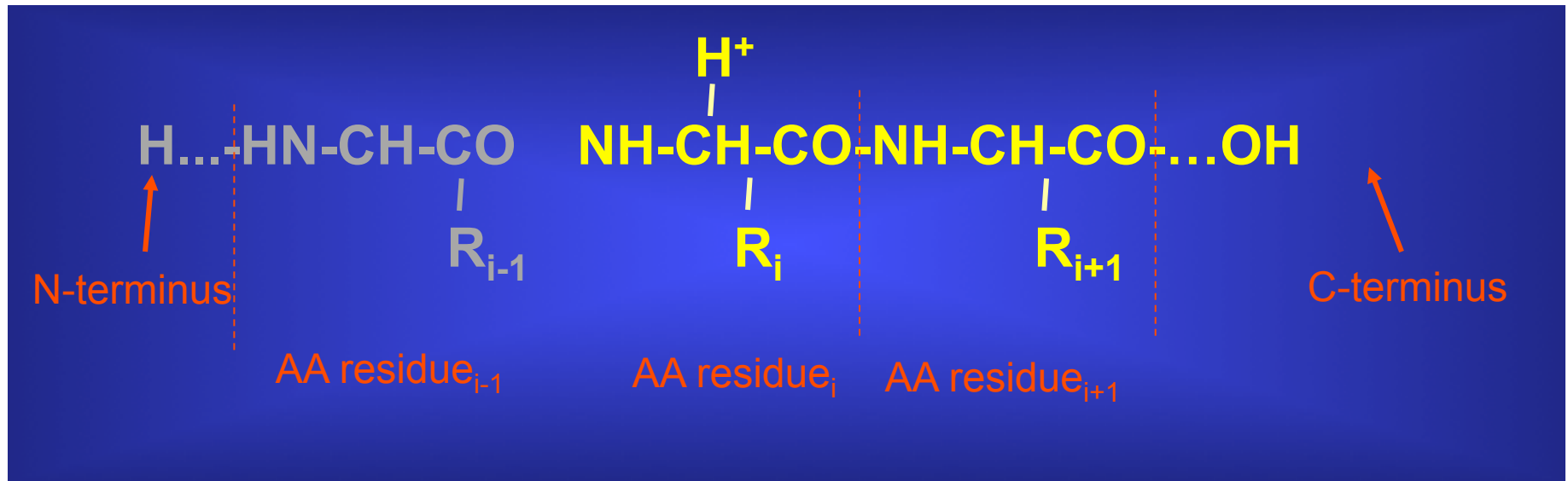
The peptide backbone breaks to form fragments with characteristic masses.



Ionized parent peptide

# Fragment ion generation

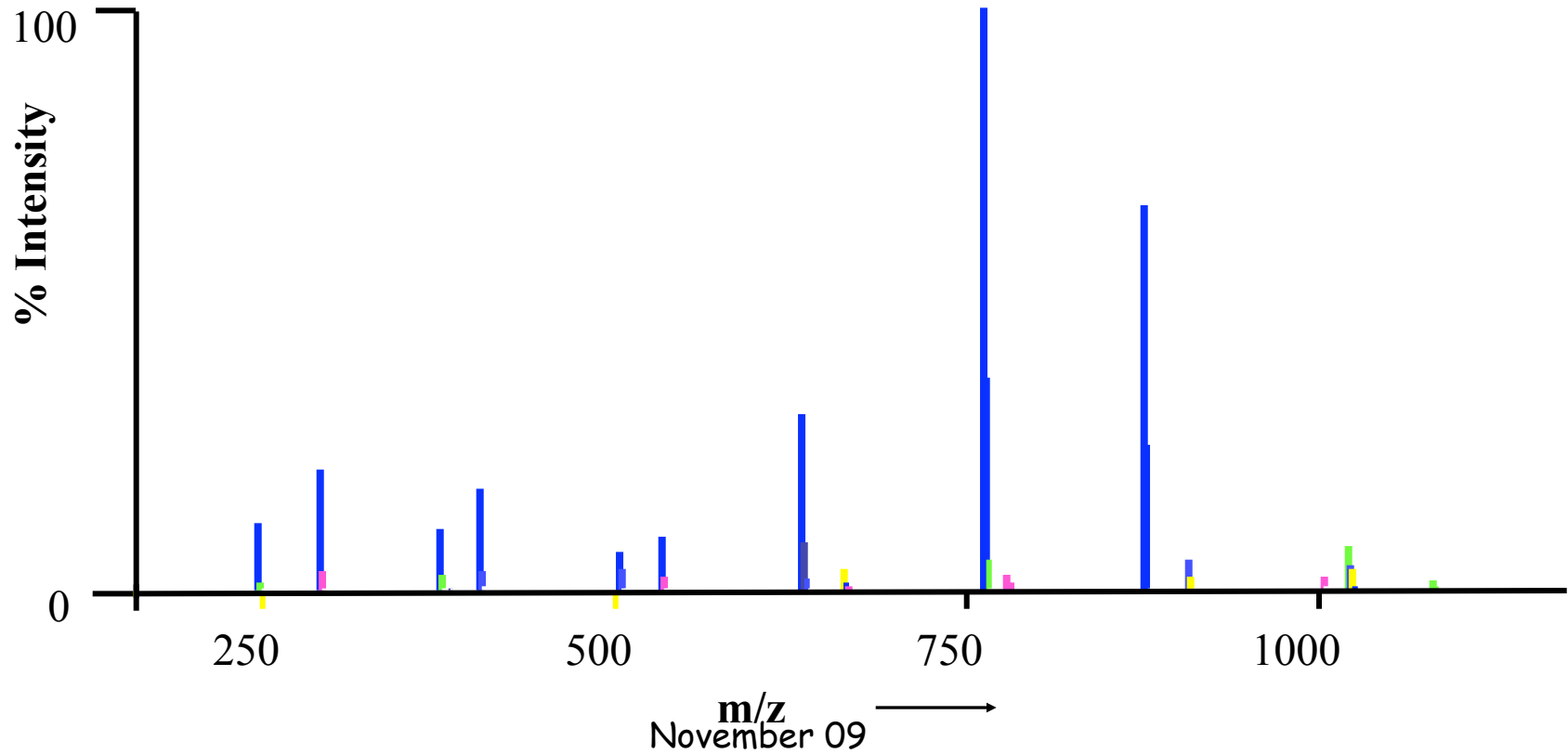
The peptide backbone breaks to form fragments with characteristic masses.



Ionized peptide fragment

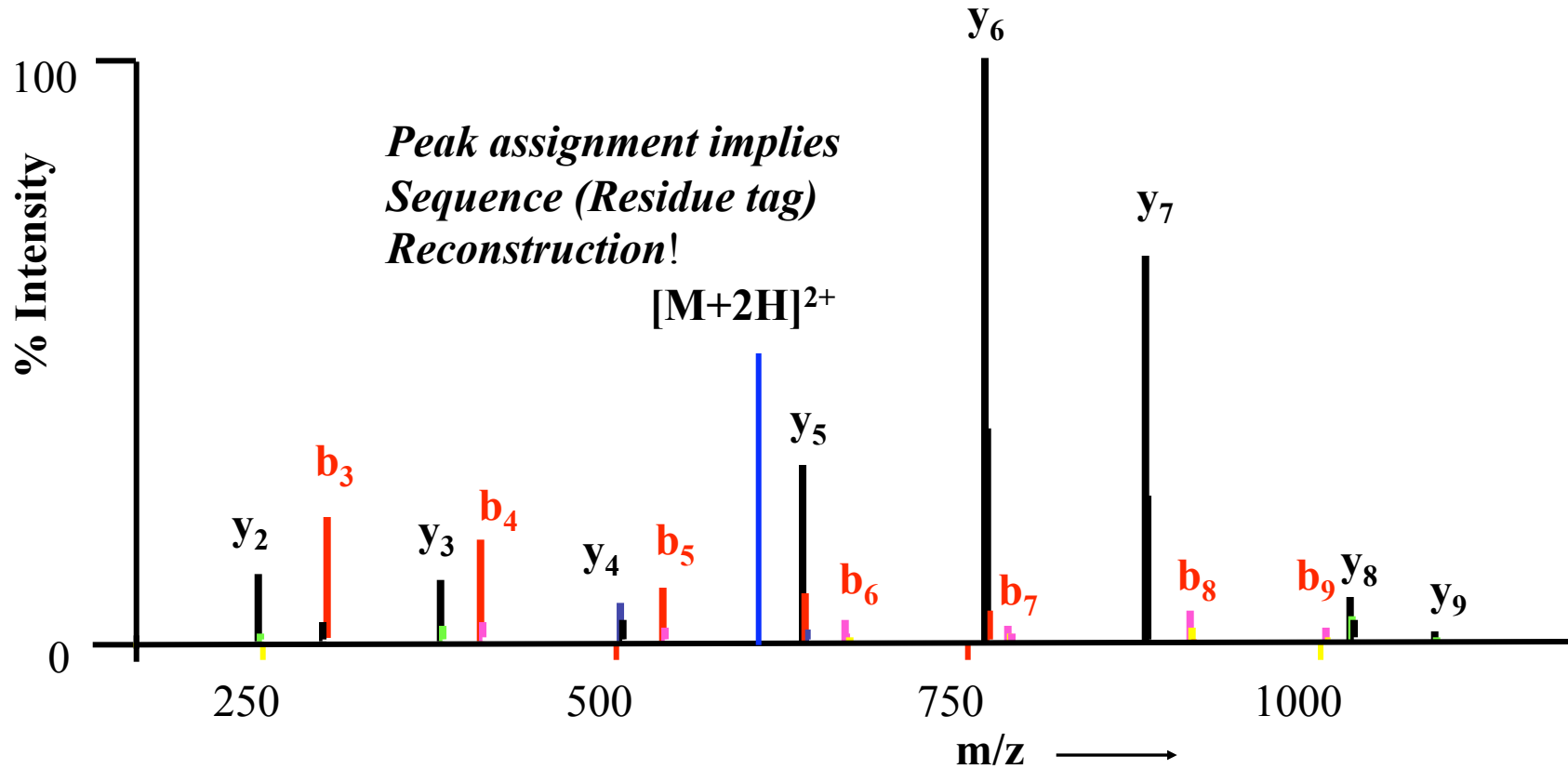
# Tandem MS for Peptide ID

<u>88</u>	<u>145</u>	<u>292</u>	<u>405</u>	<u>534</u>	<u>663</u>	<u>778</u>	<u>907</u>	<u>1020</u>	<u>1166</u>	b ions
<b>S</b>	<b>G</b>	<b>F</b>	<b>L</b>	<b>E</b>	<b>E</b>	<b>D</b>	<b>E</b>	<b>L</b>	<b>K</b>	
<u>1166</u>	<u>1080</u>	<u>1022</u>	<u>875</u>	<u>762</u>	<u>633</u>	<u>504</u>	<u>389</u>	<u>260</u>	<u>147</u>	y ions



# Peak Assignment

<u>88</u>	<u>145</u>	<u>292</u>	<u>405</u>	<u>534</u>	<u>663</u>	<u>778</u>	<u>907</u>	<u>1020</u>	<u>1166</u>	<b>b ions</b>
<b>S</b>	<b>G</b>	<b>F</b>	<b>L</b>	<b>E</b>	<b>E</b>	<b>D</b>	<b>E</b>	<b>L</b>	<b>K</b>	
<u>1166</u>	<u>1080</u>	<u>1022</u>	<u>875</u>	<u>762</u>	<u>633</u>	<u>504</u>	<u>389</u>	<u>260</u>	<u>147</u>	<b>y ions</b>



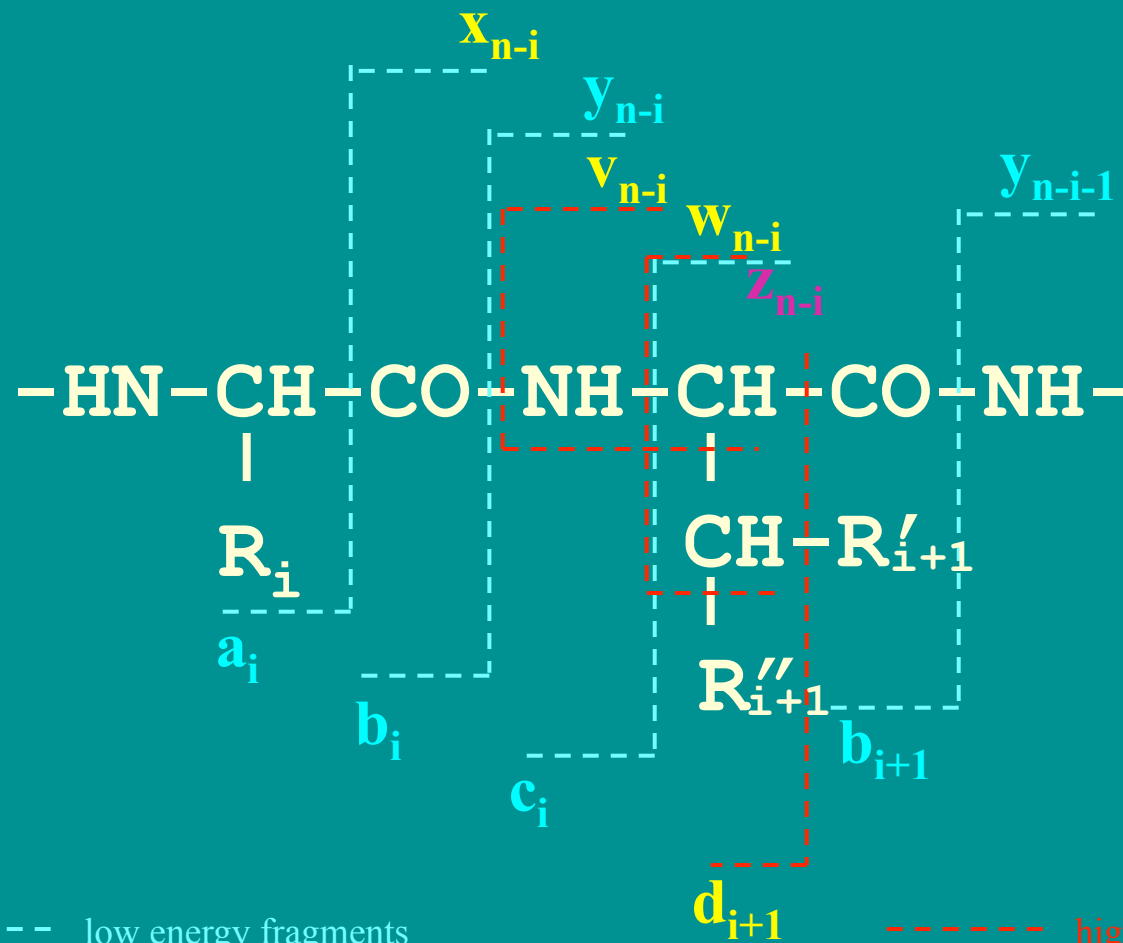
# Database Searching for peptide ID

- For every peptide from a database
  - Generate a hypothetical spectrum
  - Compute a correlation between observed and experimental spectra
  - Choose the best
- Database searching is very powerful and is the *de facto* standard for MS.
  - Sequest, Mascot, and many others

# Spectra: the real story

- Noise Peaks
- Ions, not prefixes & suffixes
- Mass to charge ratio, and not mass
  - Multiply charged ions
- Isotope patterns, not single peaks

# Peptide fragmentation possibilities (ion types)



# Ion types, and offsets

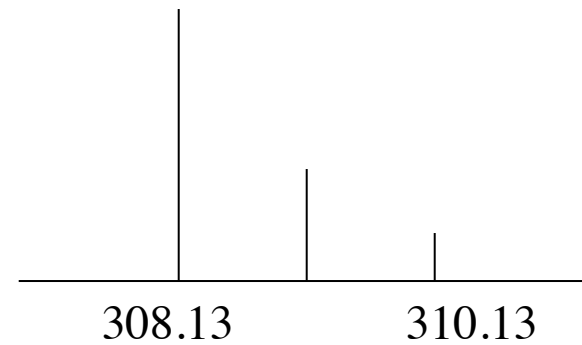
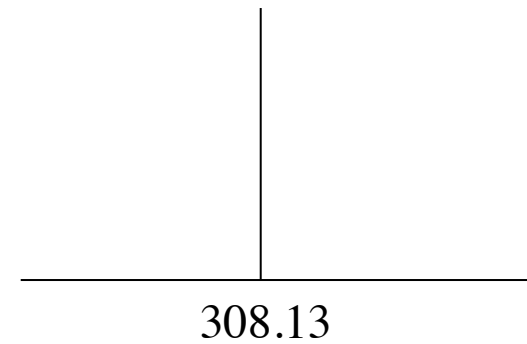
- $P$  = prefix residue mass
- $S$  = Suffix residue mass
- b-ions =  $P+1$
- $\gamma$ -ions =  $S+19$
- a-ions =  $P-27$

# Mass-Charge ratio

- The X-axis is not mass, but  $(M+Z)/Z$ 
  - $Z=1$  implies that peak is at  $M+1$
  - $Z=2$  implies that peak is at  $(M+2)/2$ 
    - $M=1000, Z=2$ , peak position is at 501
- Quiz: Suppose you see a peak at 501. Is the mass 500, or is it 1000?

# Isotopic peaks

- Ex: Consider peptide SAM
- Mass = 308.12802
- You should see:
- Instead, you see



# Isotopes

- C-12 is the most common. Suppose C-13 occurs with probability 1%
- EX: SAM
  - Composition: C11 H22 N3 O5 S1
- What is the probability that you will see a single C-13?

$$\binom{11}{1} \cdot 0.01 \cdot (0.99)^{10}$$

- Note that C,S,O,N all have isotopes. Can you compute the isotopic distribution?

# All atoms have isotopes

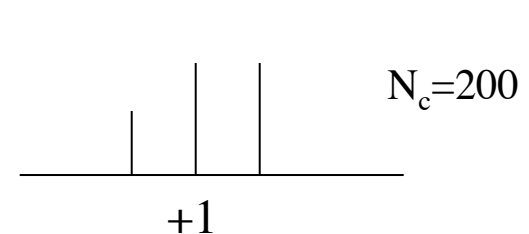
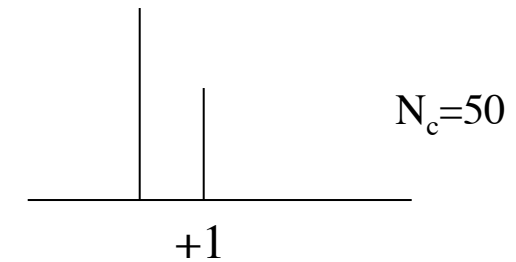
- Isotopes of atoms
  - O<sup>16,18</sup>, C-<sup>12,13</sup>, S<sup>32,34</sup>....
  - Each isotope has a frequency of occurrence
- If a molecule (peptide) has a single copy of C-<sup>13</sup>, that will shift its peak by 1 Da
- With multiple copies of a peptide, we have a distribution of intensities over a range of masses (Isotopic profile).
- How can you compute the isotopic profile of a peak?

# Isotope Calculation

- Denote:
  - $N_c$  : number of carbon atoms in the peptide
  - $P_c$  : probability of occurrence of C-13 (~1%)
  - Then

$$\Pr[\text{Peak at } M] = \binom{N_c}{0} p_c^0 (1 - p_c)^{N_c}$$

$$\Pr[\text{Peak at } M + 1] = \binom{N_c}{1} p_c^1 (1 - p_c)^{N_c - 1}$$



# Isotope Calculation Example

- Suppose we consider Nitrogen, and Carbon
- $N_N$ : number of Nitrogen atoms
- $P_N$ : probability of occurrence of N-15
- Pr(peak at M)
- Pr(peak at M+1)?
- Pr(peak at M+2)?

$$\text{Pr}[\text{Peak at } M] = \binom{N_C}{0} p_c^0 (1-p_c)^{N_C} \binom{N_N}{0} p_N^0 (1-p_N)^{N_N}$$

$$\begin{aligned} \text{Pr}[\text{Peak at } M+1] &= \binom{N_C}{1} p_c^1 (1-p_c)^{N_C-1} \binom{N_N}{0} p_N^0 (1-p_N)^{N_N} \\ &\quad + \binom{N_C}{0} p_c^0 (1-p_c)^{N_C} \binom{N_N}{1} p_N^1 (1-p_N)^{N_N-1} \end{aligned}$$

How do we generalize? How can we handle Oxygen (O-16,18)?

# General isotope computation

- Definition:
  - Let  $p_{i,a}$  be the abundance of the isotope with mass  $i$  Da above the least mass
  - Ex:  $P_{0,C}$ : abundance of C-12,  $P_{2,O}$ : O-18 etc.
- Characteristic polynomial

$$\phi(x) = \prod_a \left( p_{0,a} + p_{1,a}x + p_{2,a}x^2 + \dots \right)^{N_a}$$

- Prob{M+i}: coefficient of  $x^i$  in  $\phi(x)$  (a binomial convolution)

# Isotopic Profile Application

- In DxMS, hydrogen atoms are exchanged with deuterium
- The rate of exchange indicates how buried the peptide is (in folded state)
- Consider the observed characteristic polynomial of the isotope profile  $\phi_{t_1}, \phi_{t_2}$ , at various time points. Then

$$\phi_{t_2}(x) = \phi_{t_1}(x)(p_{0,H} + p_{1,H})^{N_H}$$

- The estimates of  $p_{1,H}$  can be obtained by a deconvolution
- Such estimates at various time points should give the rate of incorporation of Deuterium, and therefore, the accessibility.

# Quiz

- How can you determine the charge on a peptide?
- Difference between the first and second isotope peak is  $1/Z$
- Proposal:
  - Given a mass, predict a composition, and the isotopic profile
  - Do a 'goodness of fit' test to isolate the peaks corresponding to the isotope
  - Compute the difference

# Tandem MS summary

- The basics of peptide ID using tandem MS is simple.
  - Correlate experimental with theoretical spectra
- In practice, there might be many confounding problems.
  - Isotope peaks, noise peaks, varying charges, post-translational modifications, no database.
- Recall that we discussed how peptides could be identified by scanning a database.
- What if the database did not contain the peptide of interest?

# De novo analysis basics

- Suppose all ions were prefix ions? Could you tell what the peptide was?
- Can post-translational modifications help?