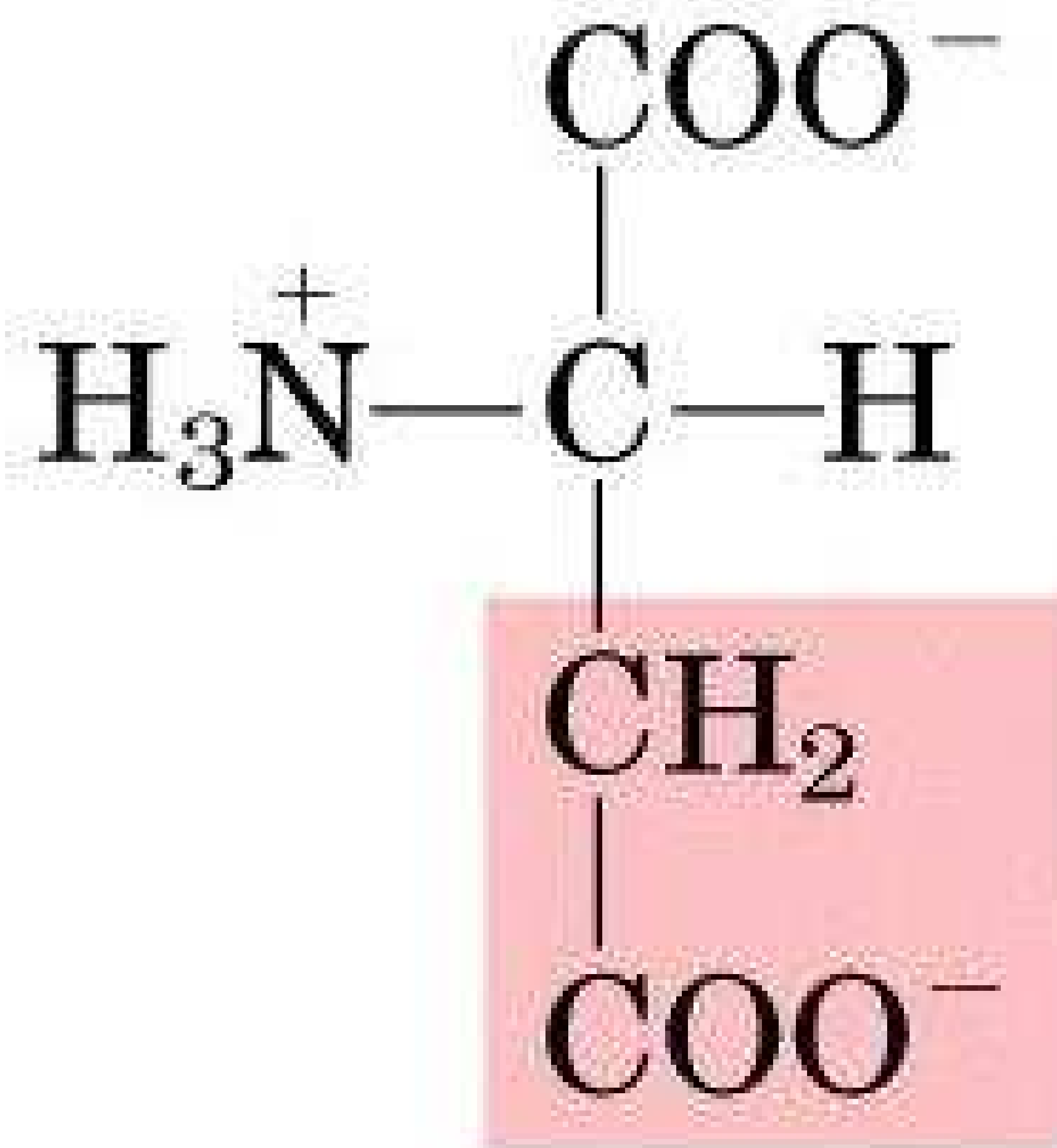


Protein Sequencing and Identification by Mass Spectrometry

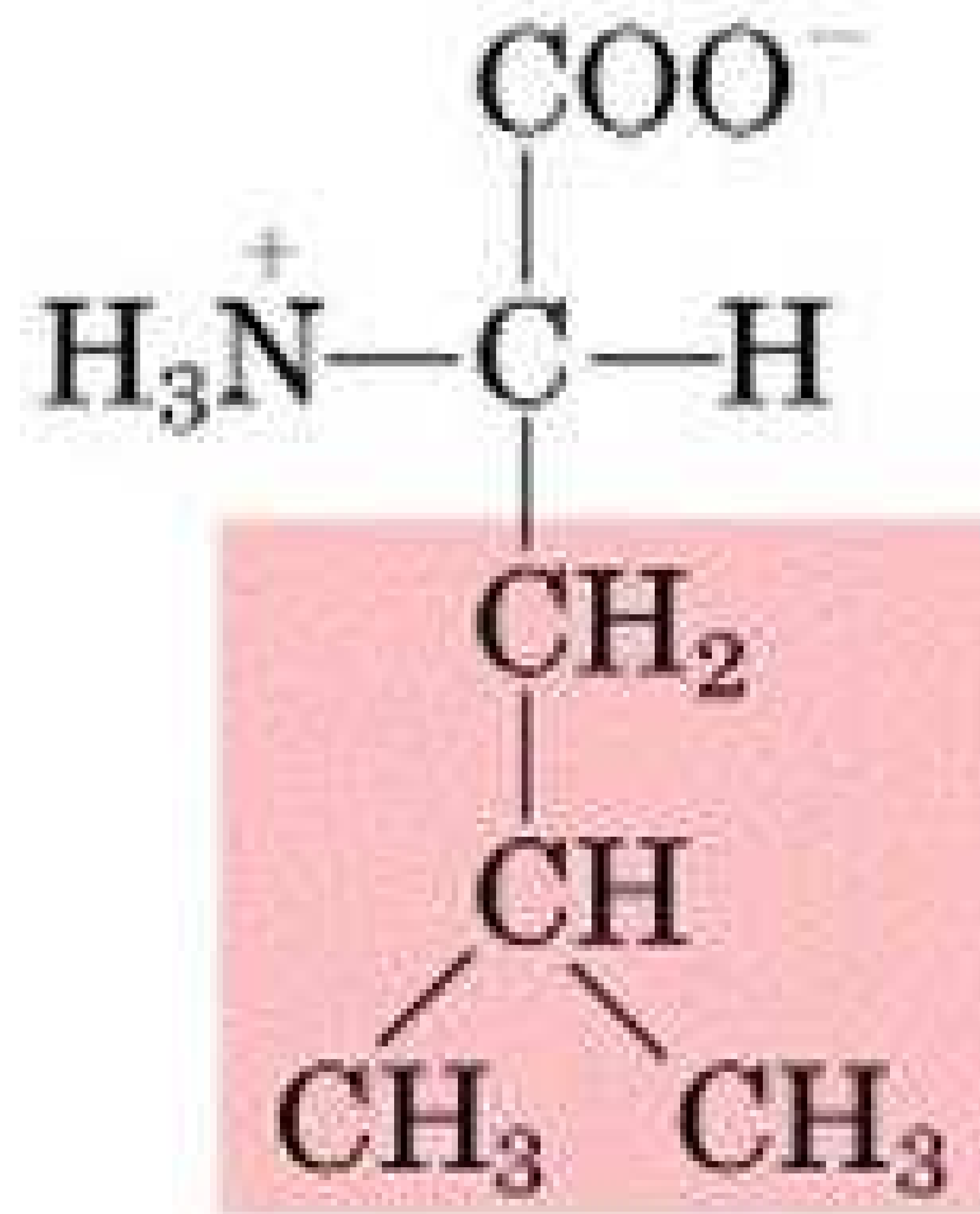
Outline

- Tandem Mass Spectrometry
 - De Novo Peptide Sequencing
 - Spectrum Graph
 - Protein Identification via Database Search
 - Identifying Post Translationally Modified Peptides
 - Spectral Convolution
 - Spectral Alignment
-

Masses of Amino Acid Residues

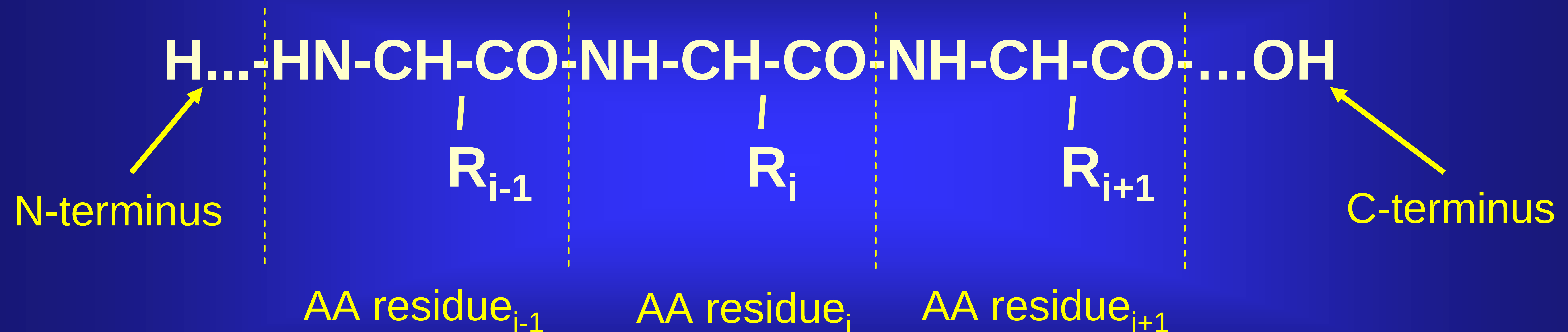


Aspartate

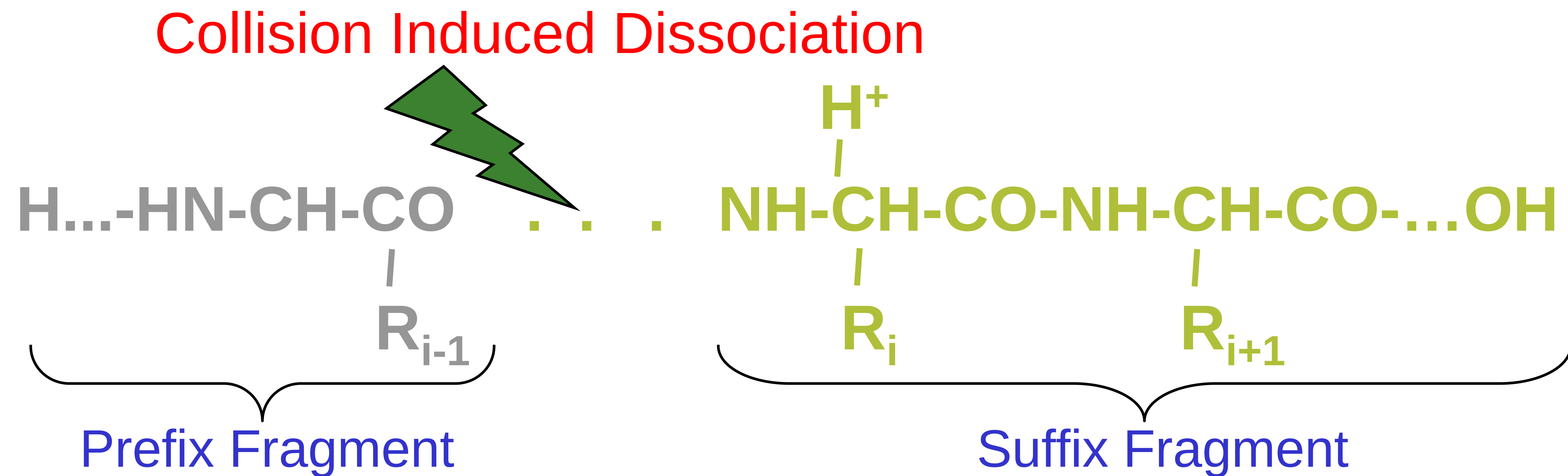


Leucine

Protein Backbone



Peptide Fragmentation

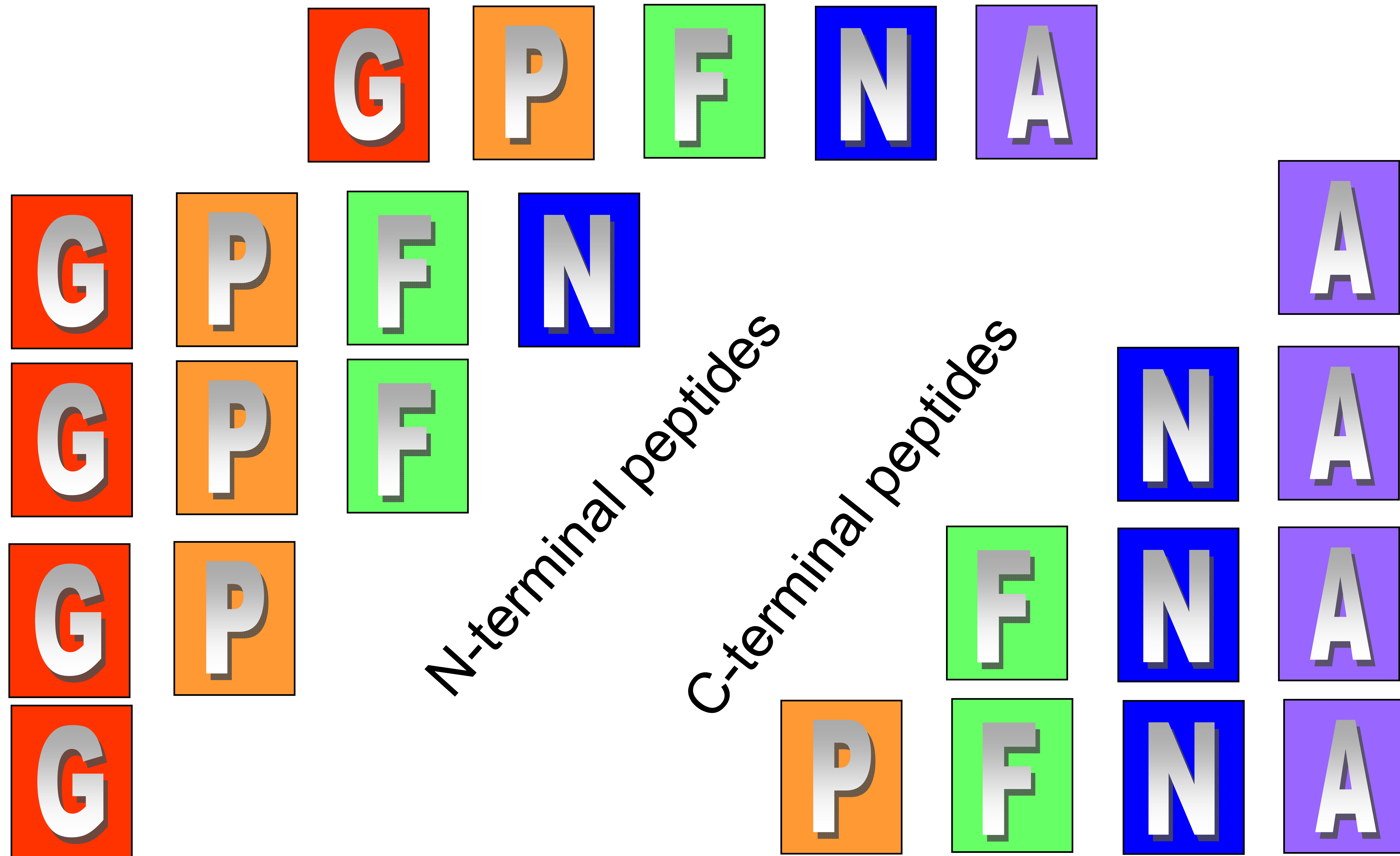


- Peptides tend to fragment along the backbone.
- Fragments can also lose neutral chemical groups like NH_3 and H_2O .

Breaking Protein into Peptides and Peptides into Fragment Ions

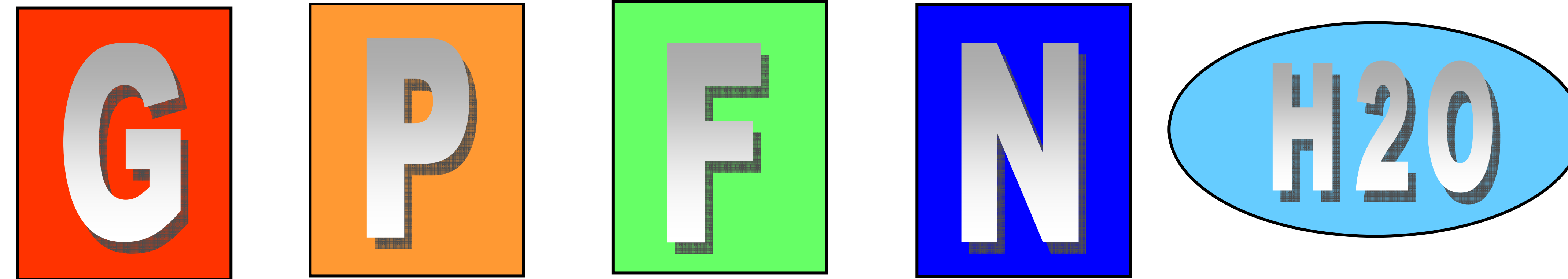
- Proteases, e.g. trypsin, break protein into *peptides*.
- A Tandem Mass Spectrometer further breaks the peptides down into *fragment ions* and measures the mass of each piece.
- Mass Spectrometer accelerates the fragmented ions; heavier ions accelerate slower than lighter ones.
- Mass Spectrometer measure *mass/charge* ratio of an ion.

N- and C-terminal Peptides



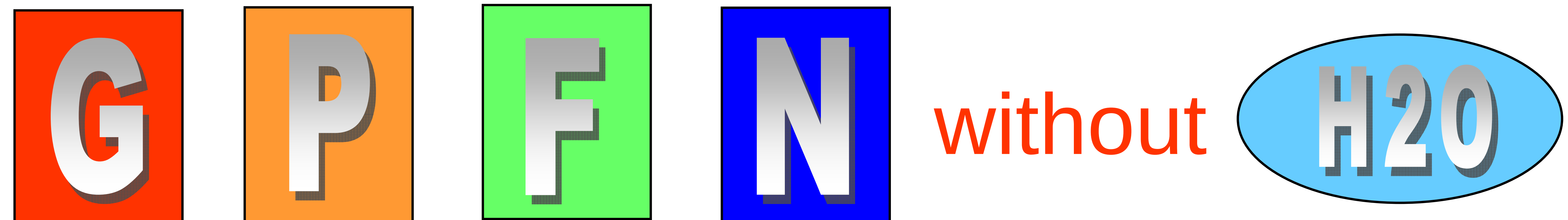
Terminal peptides and ion types

Peptide



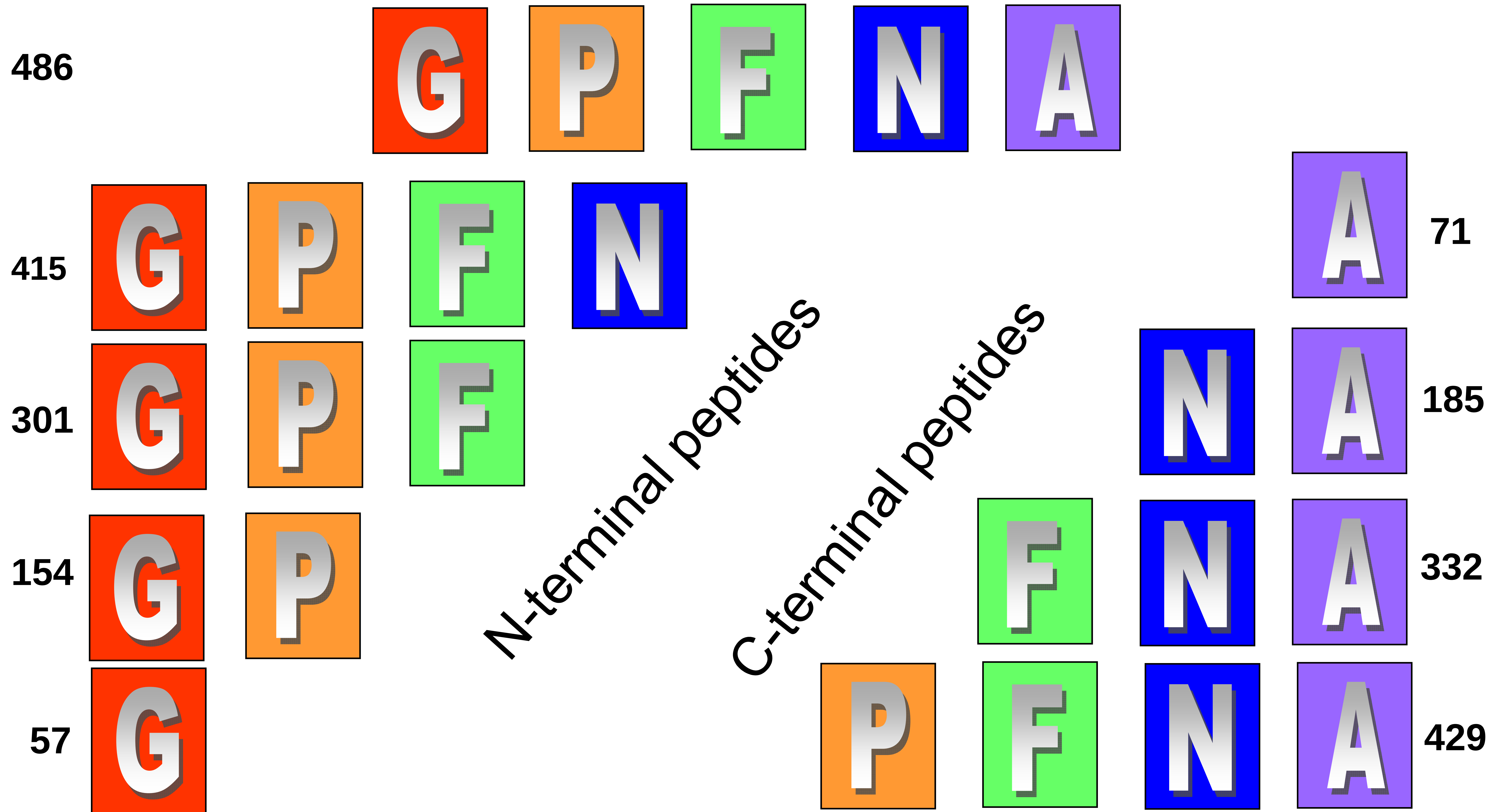
$$\text{Mass (D)} \quad 57 + 97 + 147 + 114 = 415$$

Peptide

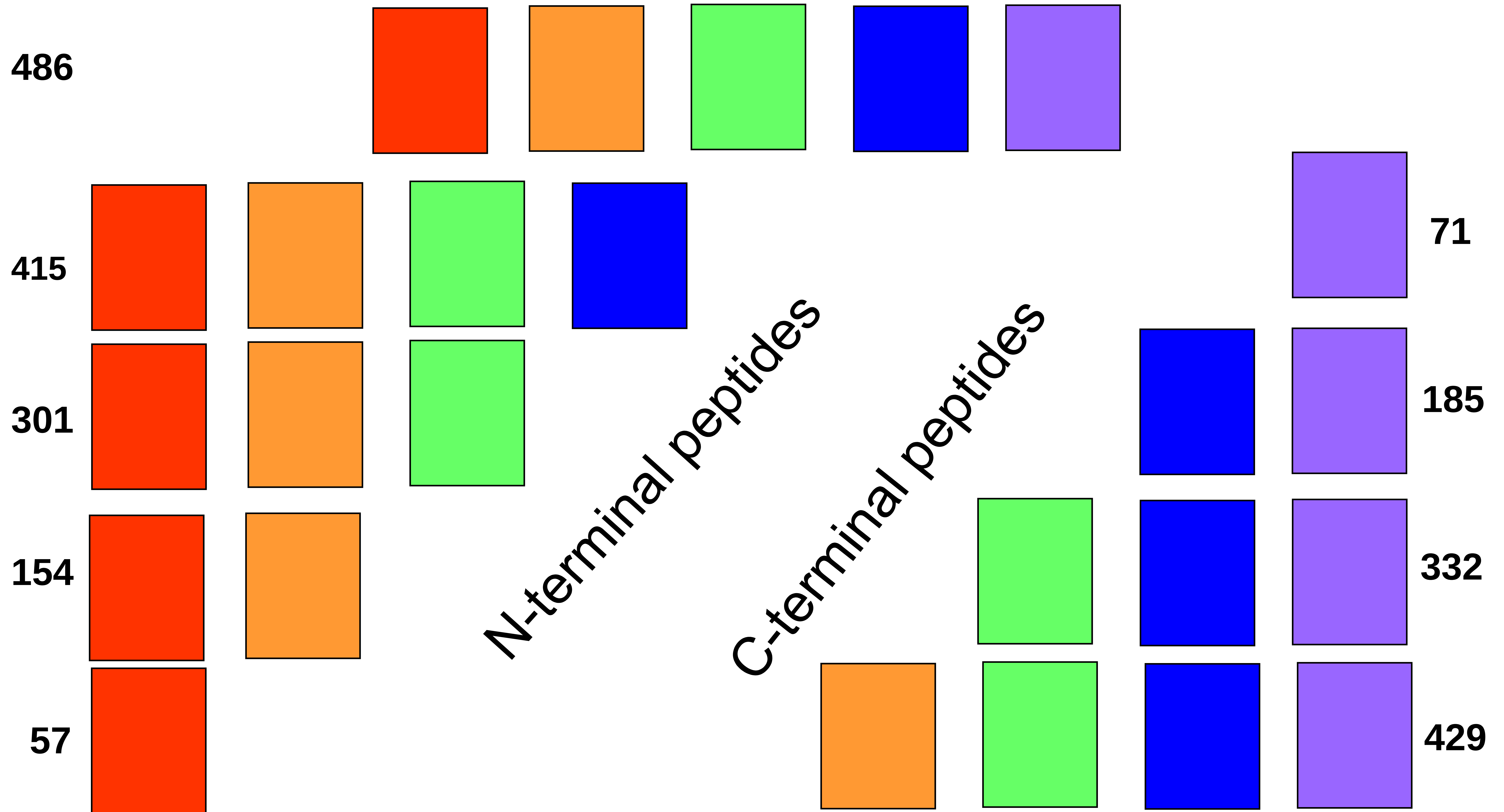


$$\text{Mass (D)} \quad 57 + 97 + 147 + 114 - 18 = 397$$

N- and C-terminal Peptides



N- and C-terminal Peptides



N- and C-terminal Peptides

486

415

301

154

57

71

185

332

429

N- and C-terminal Peptides

486

415

Reconstruct peptide from the set of masses of fragment ions

301

(**mass-spectrum**)

154

57

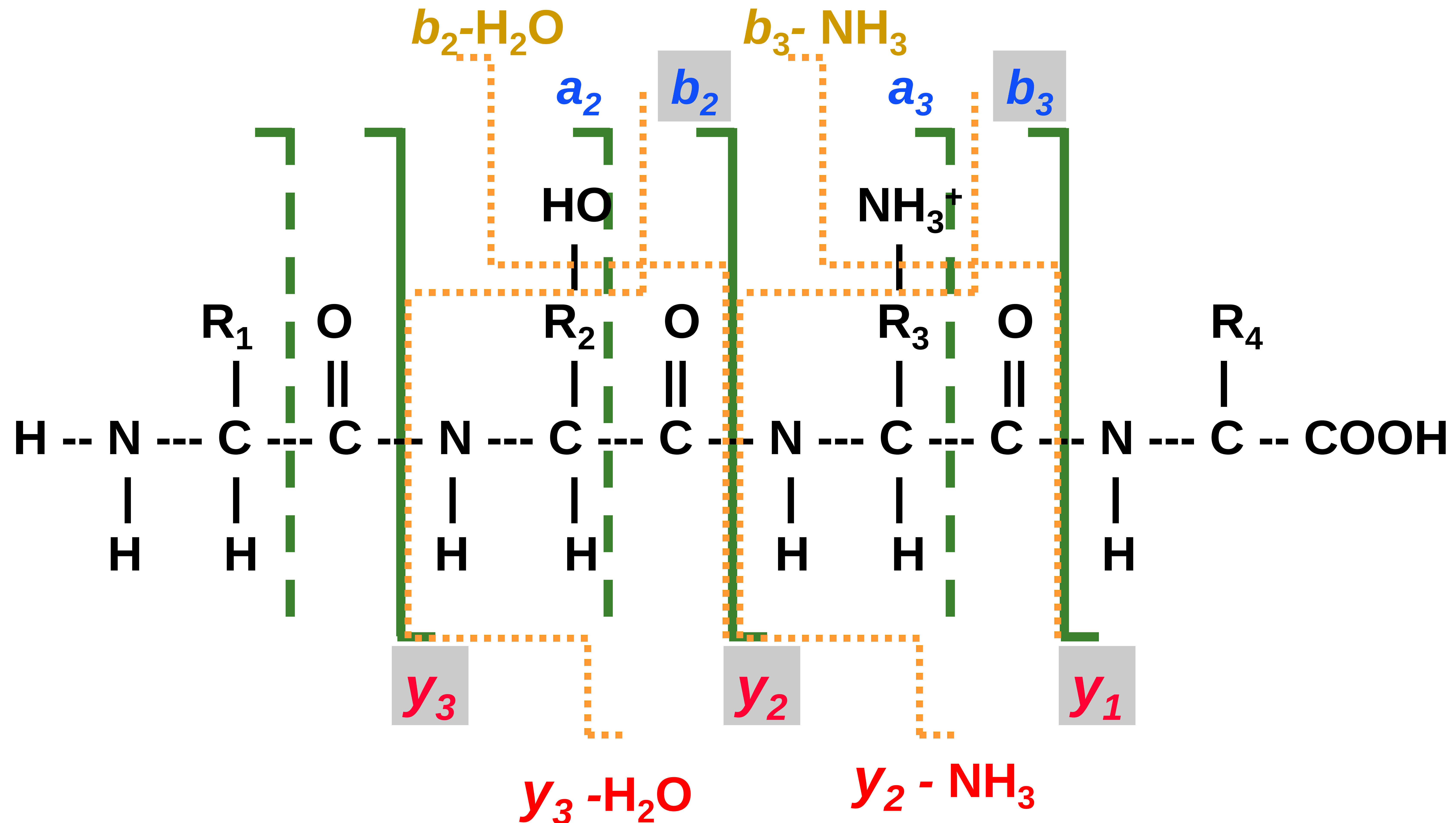
71

185

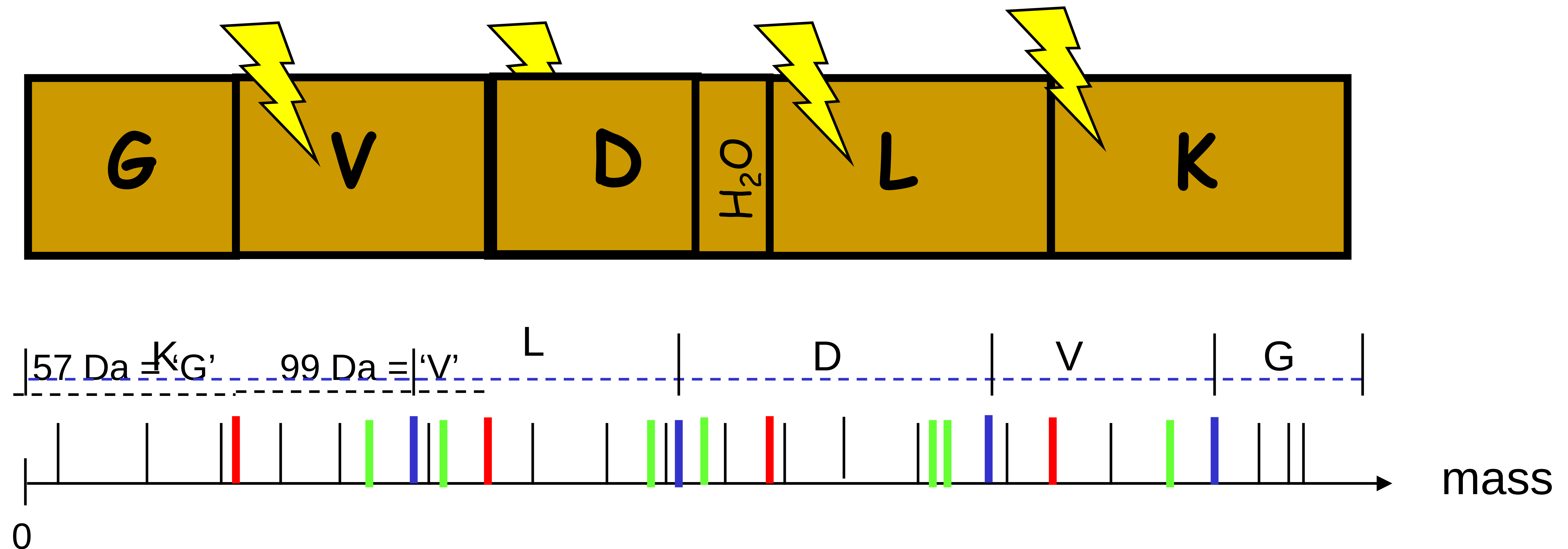
332

429

Peptide Fragmentation

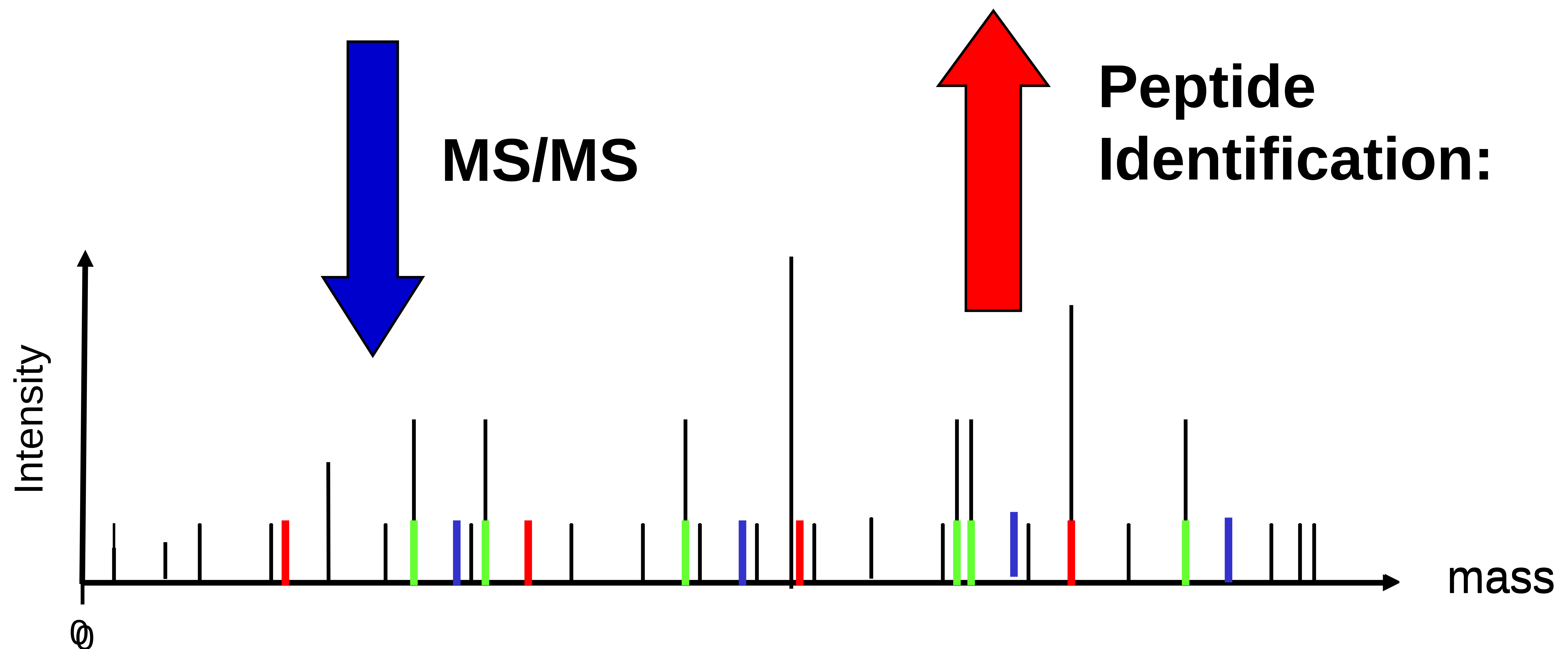
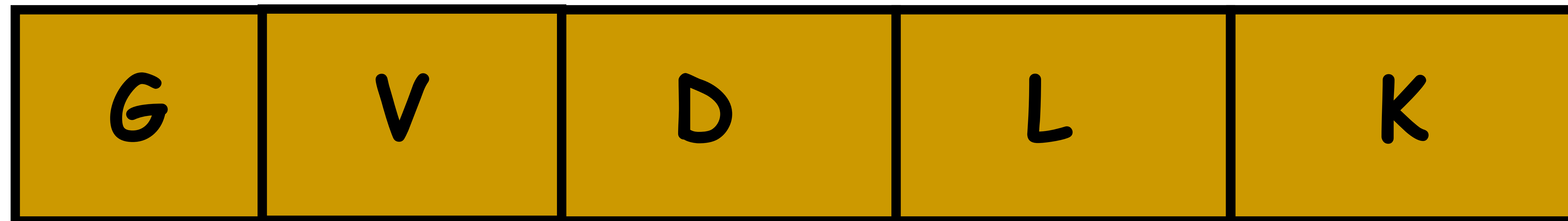


Mass Spectra

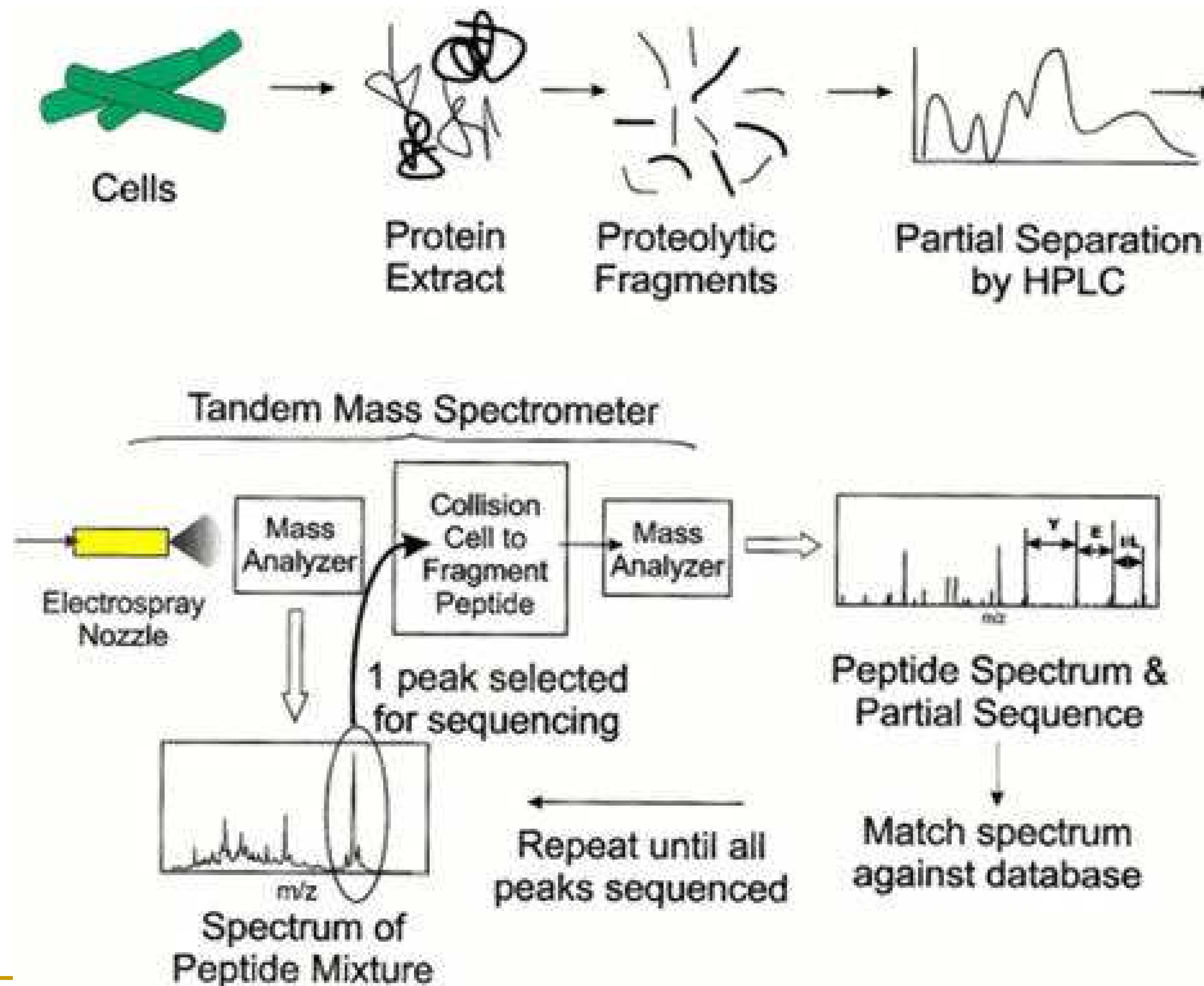


- The peaks in the mass spectrum:
 - **Prefix** and **Suffix** Fragments.
 - Fragments with **neutral losses** ($-\text{H}_2\text{O}$, $-\text{NH}_3$)
 - Noise and missing peaks.

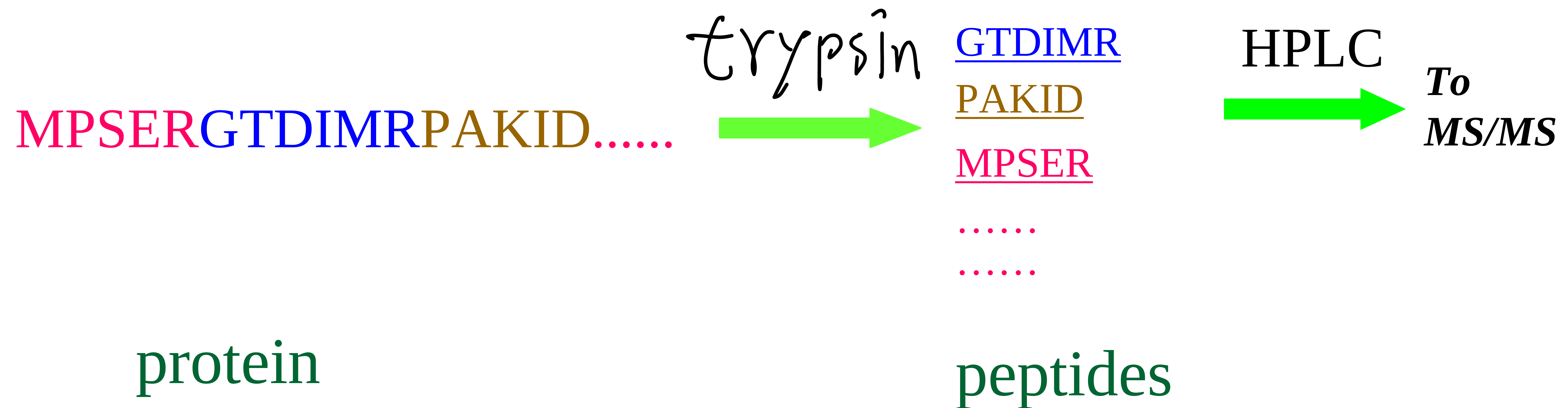
Protein Identification with MS/MS



Tandem Mass-Spectrometry



Breaking Proteins into Peptides



Mass Spectrometry

Matrix-Assisted Laser Desorption/Ionization (MALDI)

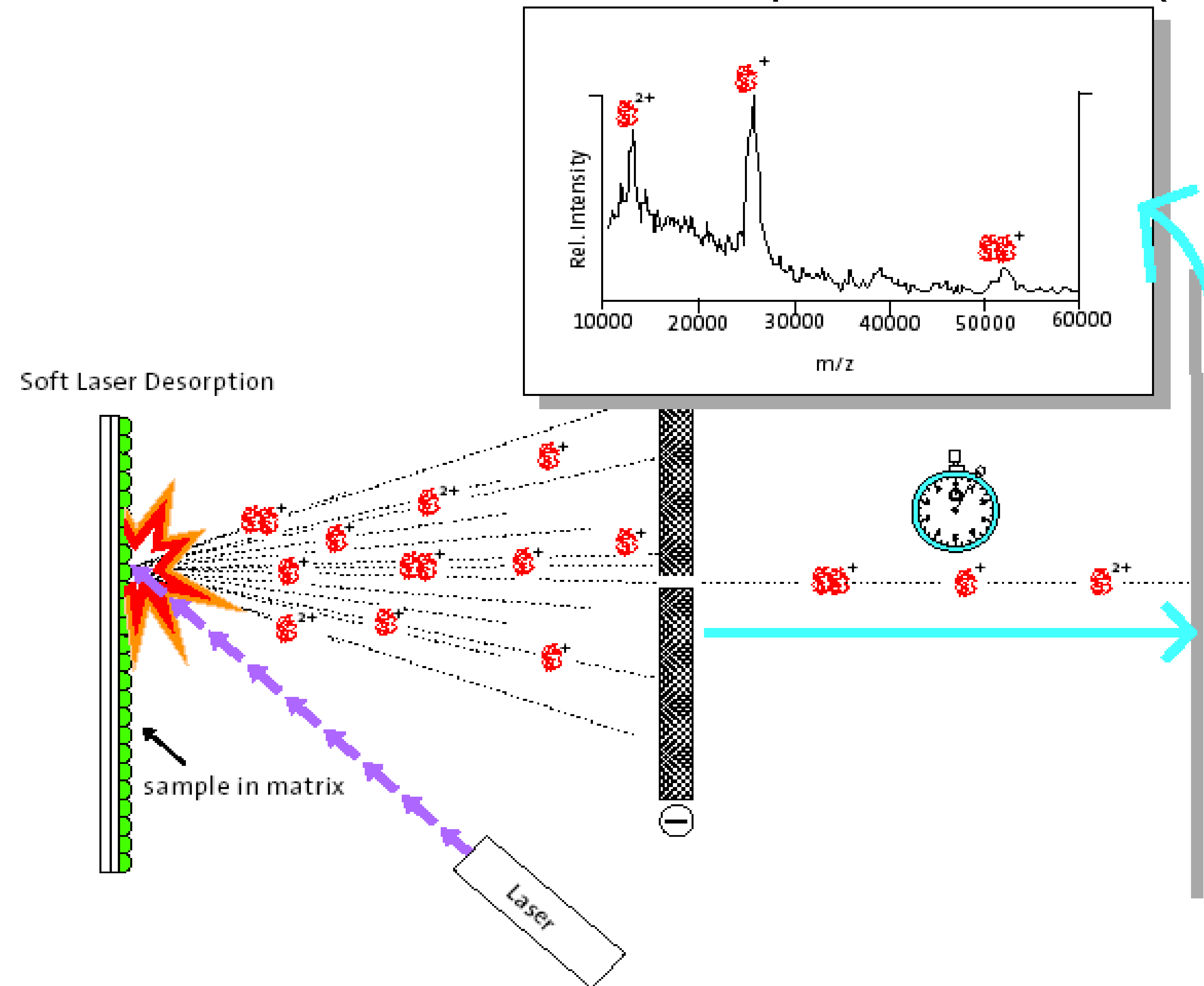
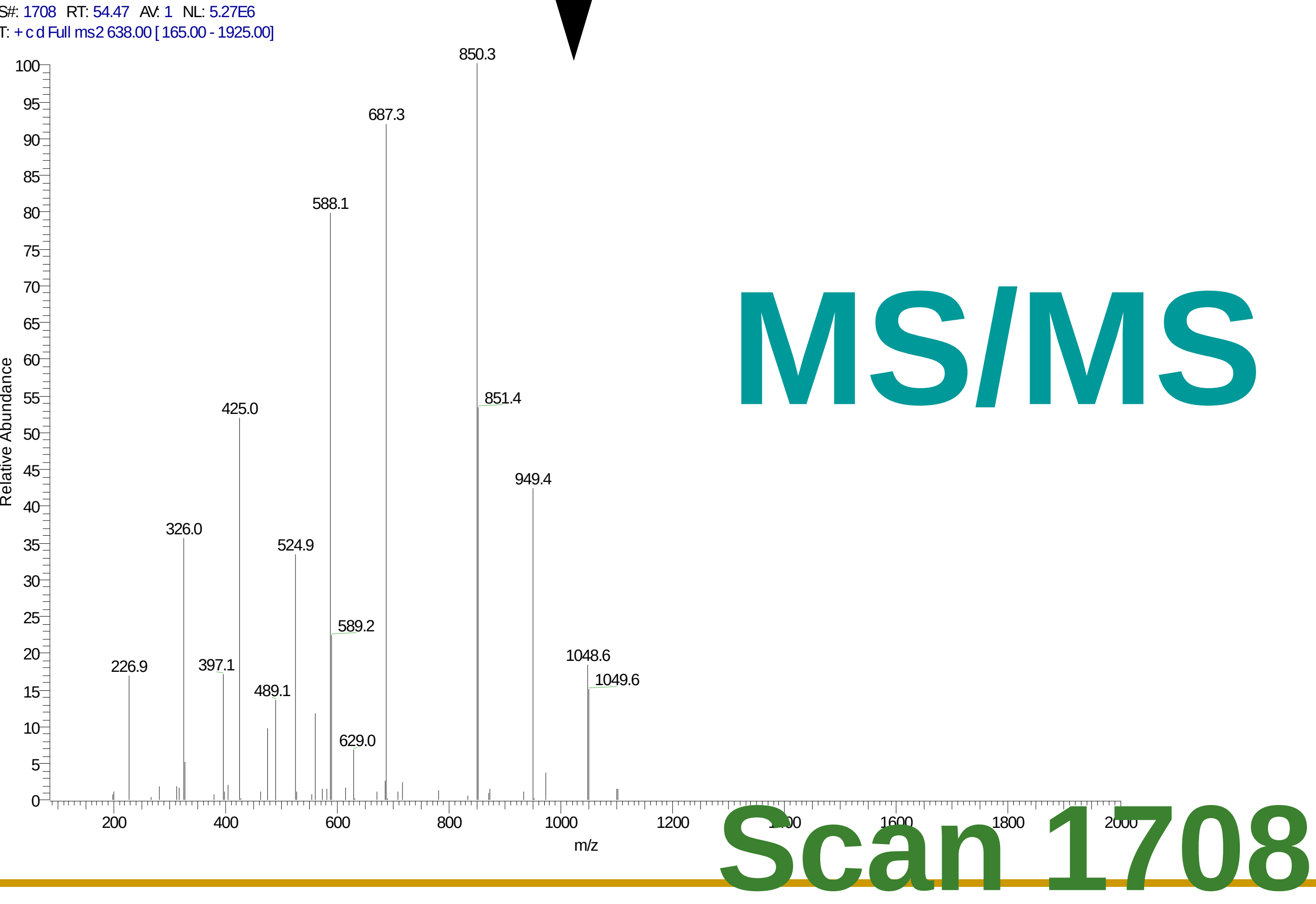
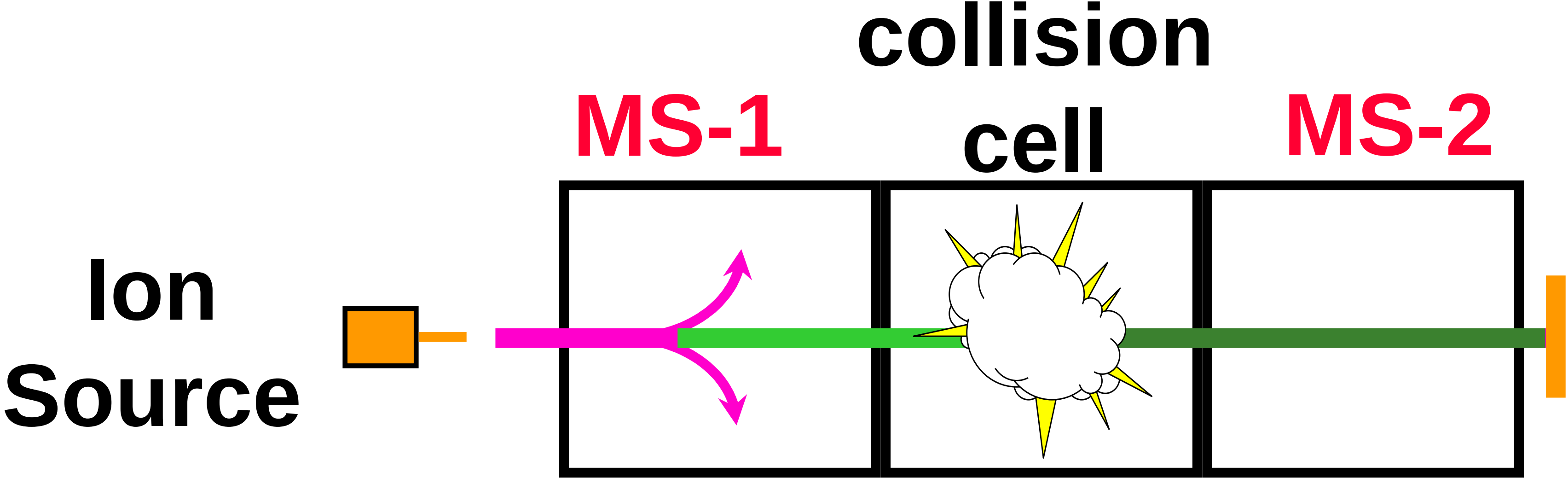
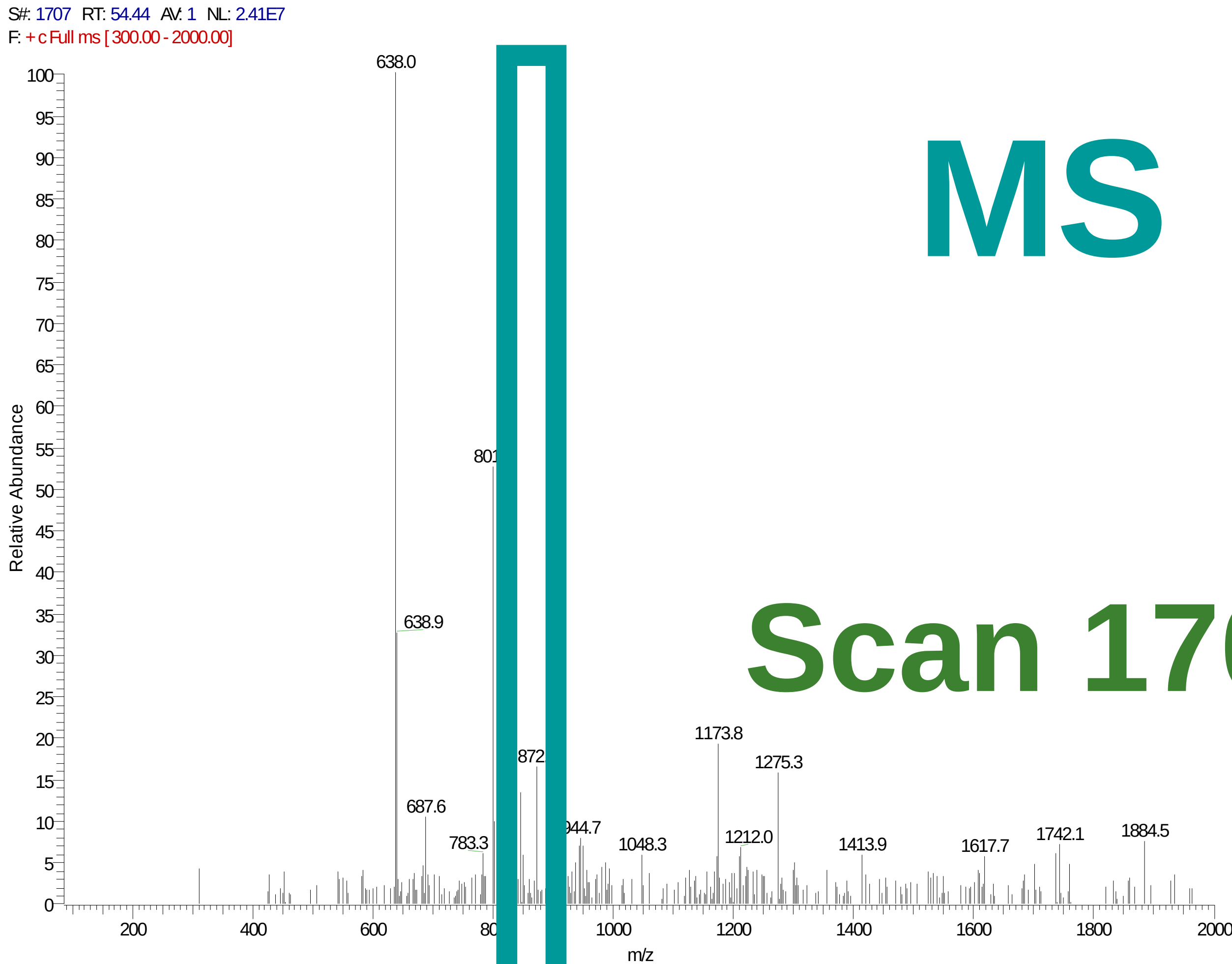
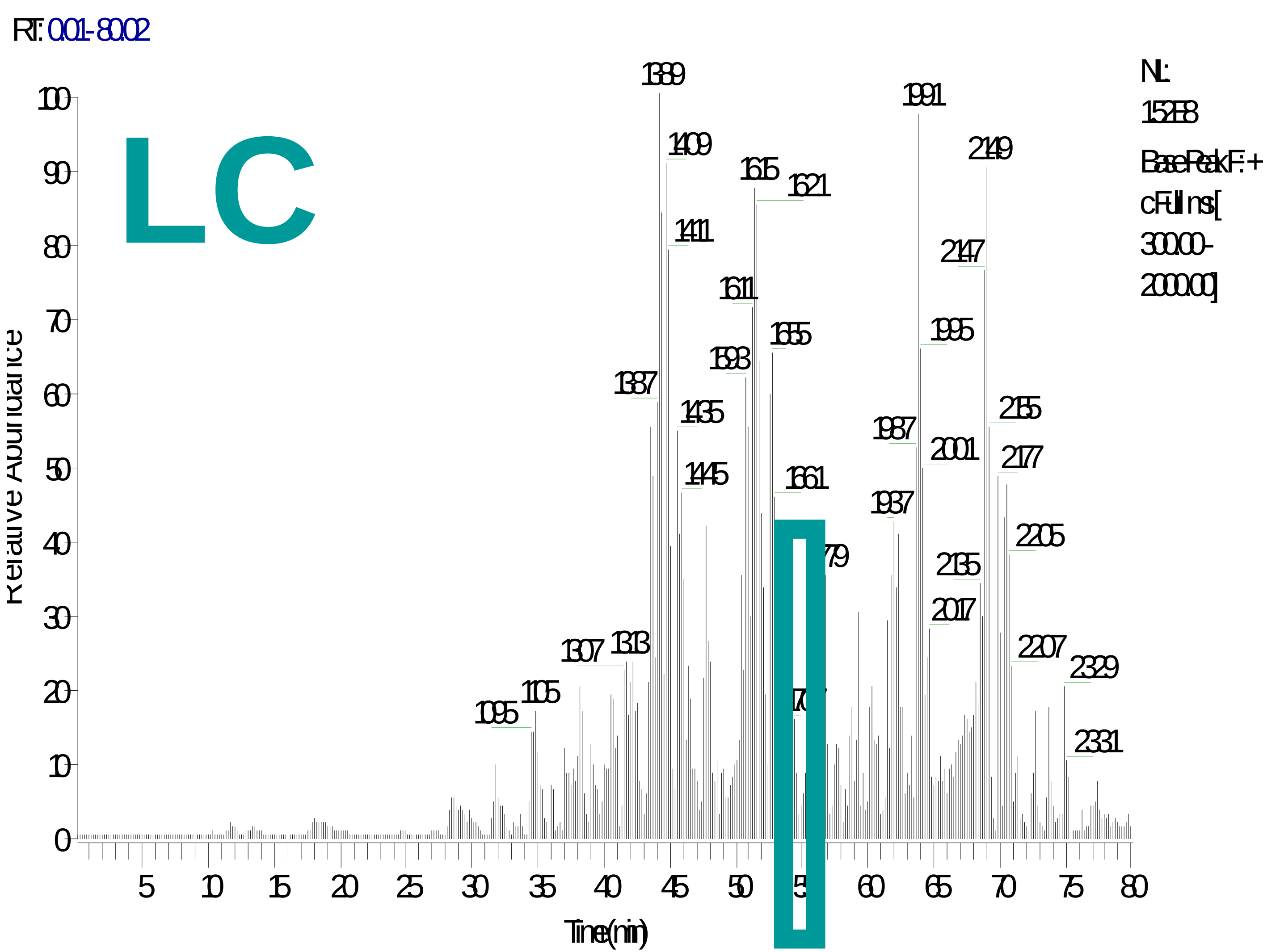


Figure 2. The soft laser desorption process.

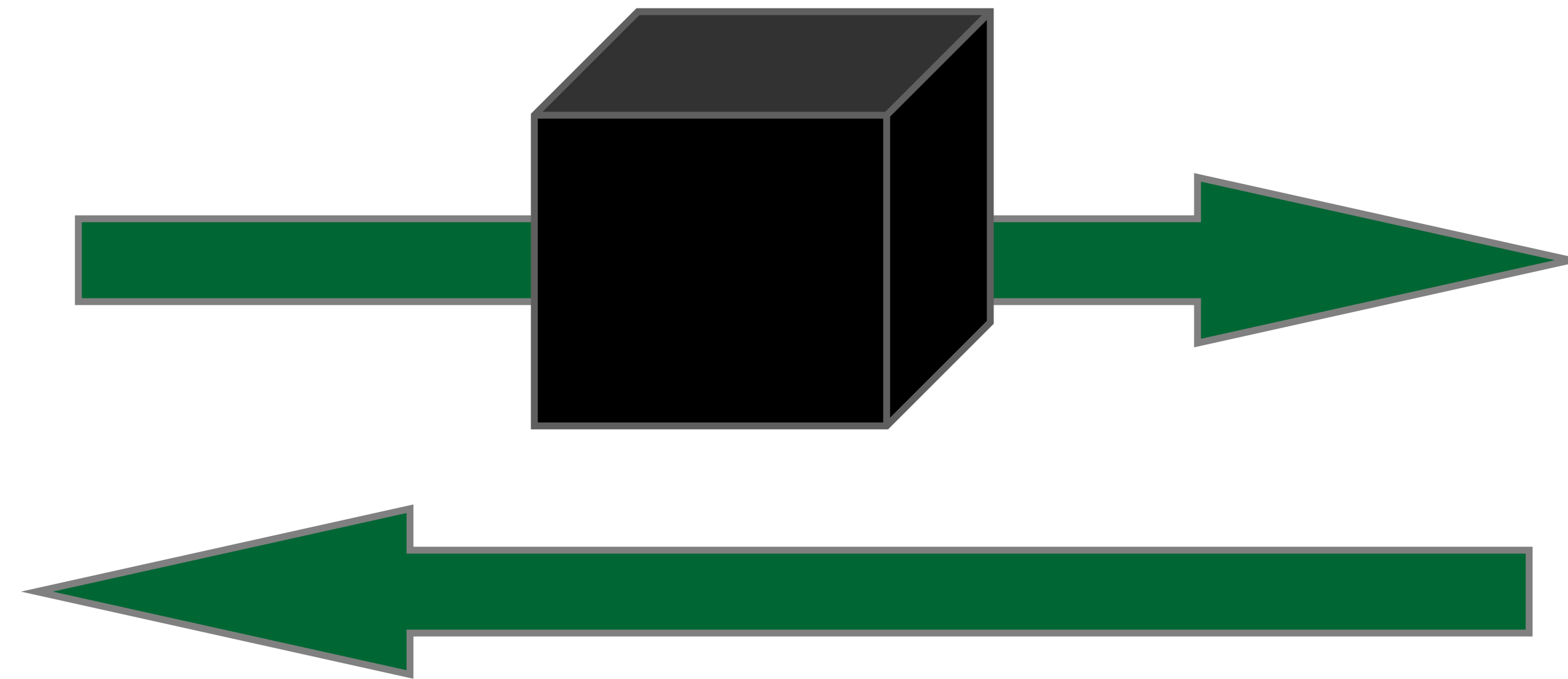
Tandem Mass Spectrometry



Protein Identification by Tandem Mass Spectrometry

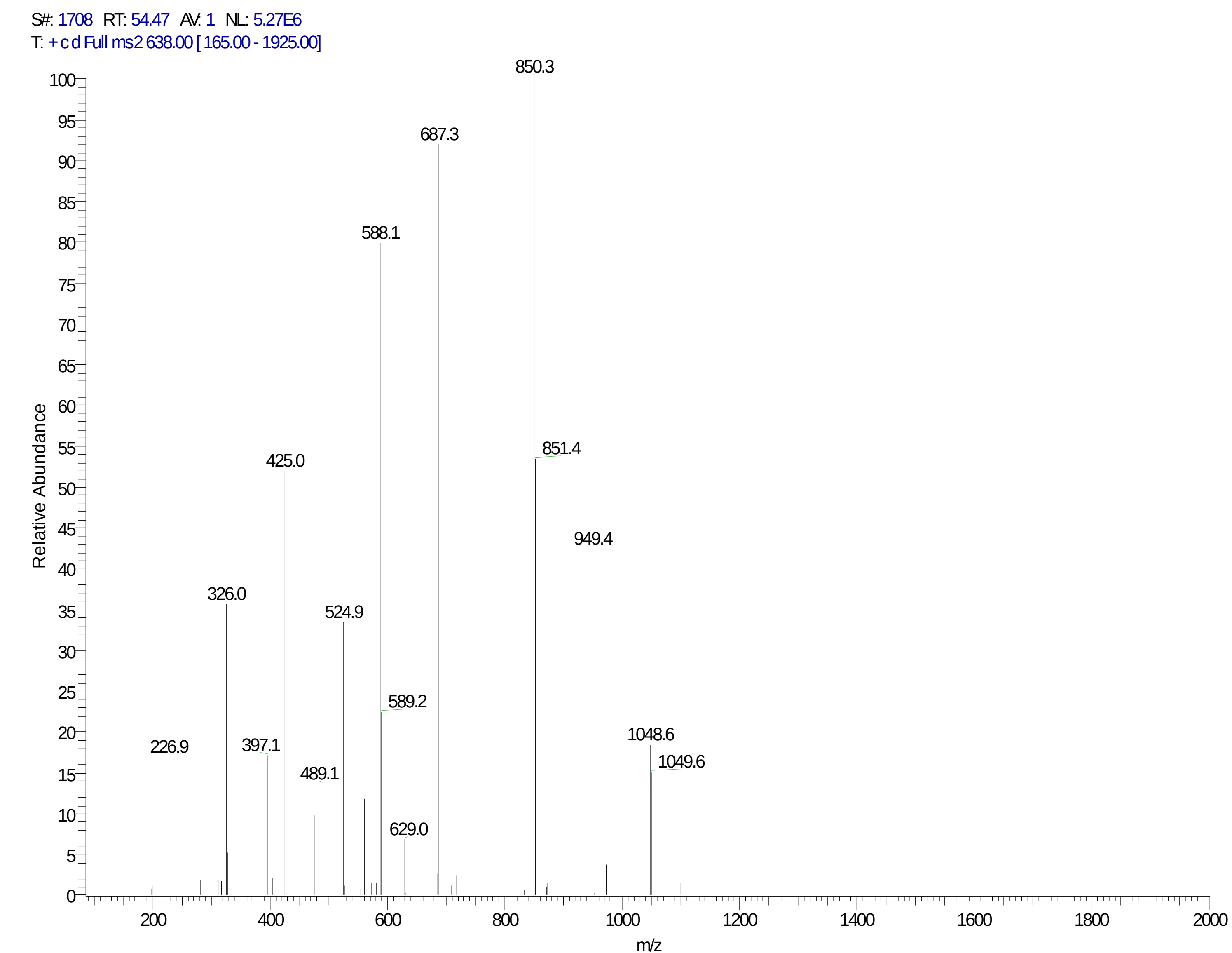
Sequence

MS/MS instrument



Database search

- **Sequest**
- **de Novo** interpretation
- **Sherenga**



Tandem Mass Spectrum

- Tandem Mass Spectrometry (MS/MS): mainly generates partial N- and C-terminal peptides
- Spectrum consists of different ion types because peptides can be broken in several places.
- Chemical noise often complicates the spectrum.
- Represented in 2-D: mass/charge axis vs. intensity axis

Database Search

MDERHILNM, KLQWVCSDL,
PTYWASDL, ENQIKRSACVM,
TLACHGGEM, NGALPQWRT,
HLLERTKMNVV, GGPASSDA,
GGLITGMQSD, MQPLMNWE,
AAKKMMNRRT, **AVGELTK**,
HEWAILF, GHNLWAMNAC,
GVFGSVLRA, EKLNKAATYIN..

Mass, Score

De Novo

Database of all peptides $\approx 20^n$

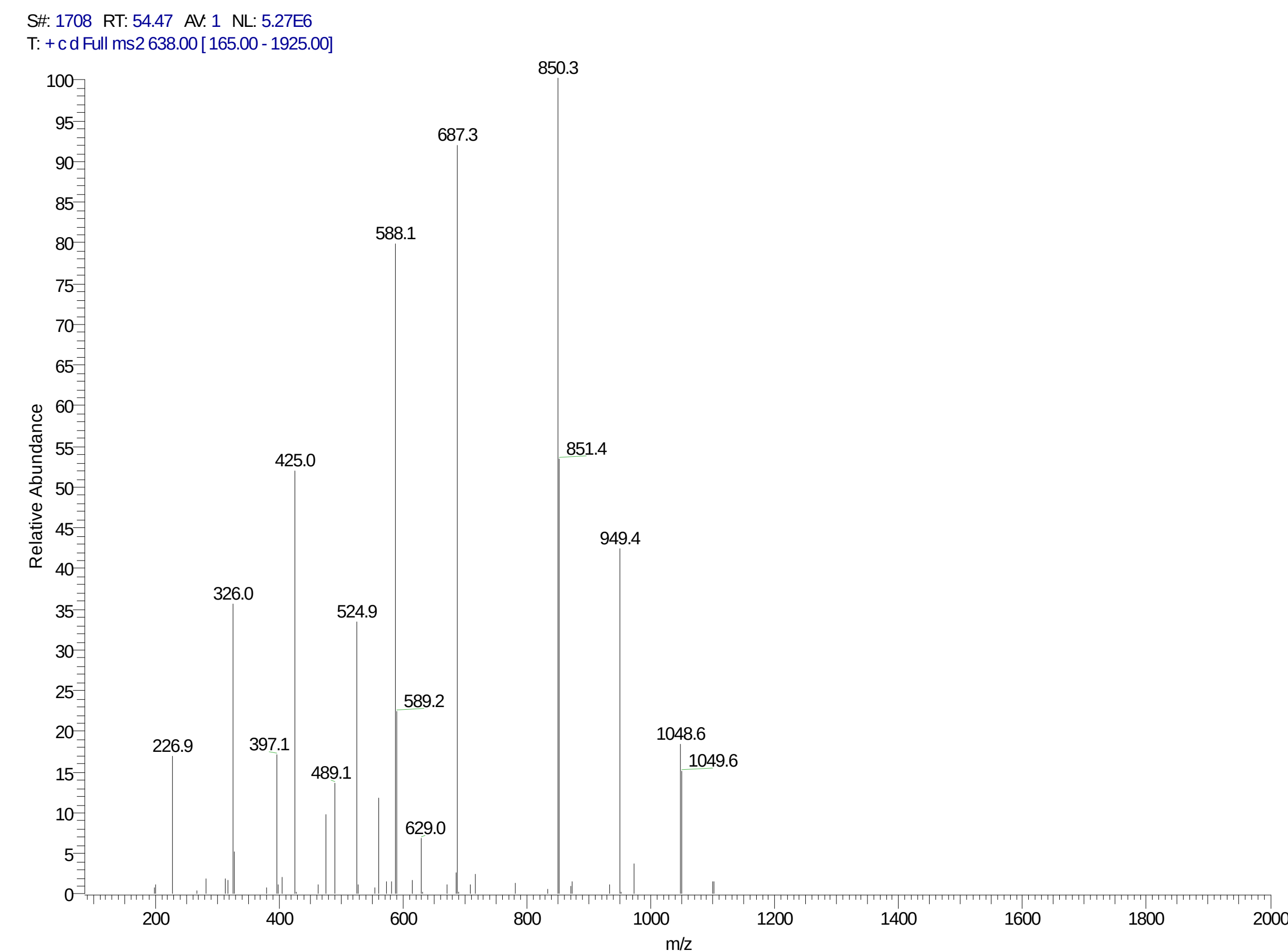
The diagram illustrates a sequence graph. Nodes are represented by blue circles, and edges by black arrows. A specific path is highlighted in red, starting from a yellow node on the left, passing through several blue nodes, and ending at a yellow node on the right. The path is labeled with red letters: A, G, L, T, K. The graph is surrounded by text sequences: 'AAAAAAAAA,AAAAAAC,AAAAAAD,AAAAAAE,AAAAAAG,AAAAAAF,AAAAAAH,AAAAAAA' at the top and 'YYYYYYYYS,YYYYYYYT,YYYYYYYV,YYYYYYY' at the bottom. In the center, the sequence 'AVGELTI, **AVGELTK**, AVGELTL, AVGELTM,' is shown, with 'AVGELTK' in bold blue. Other letters like 'C', 'L', 'P', 'W', 'D', and 'T' are also visible near the nodes.

AVGELTK

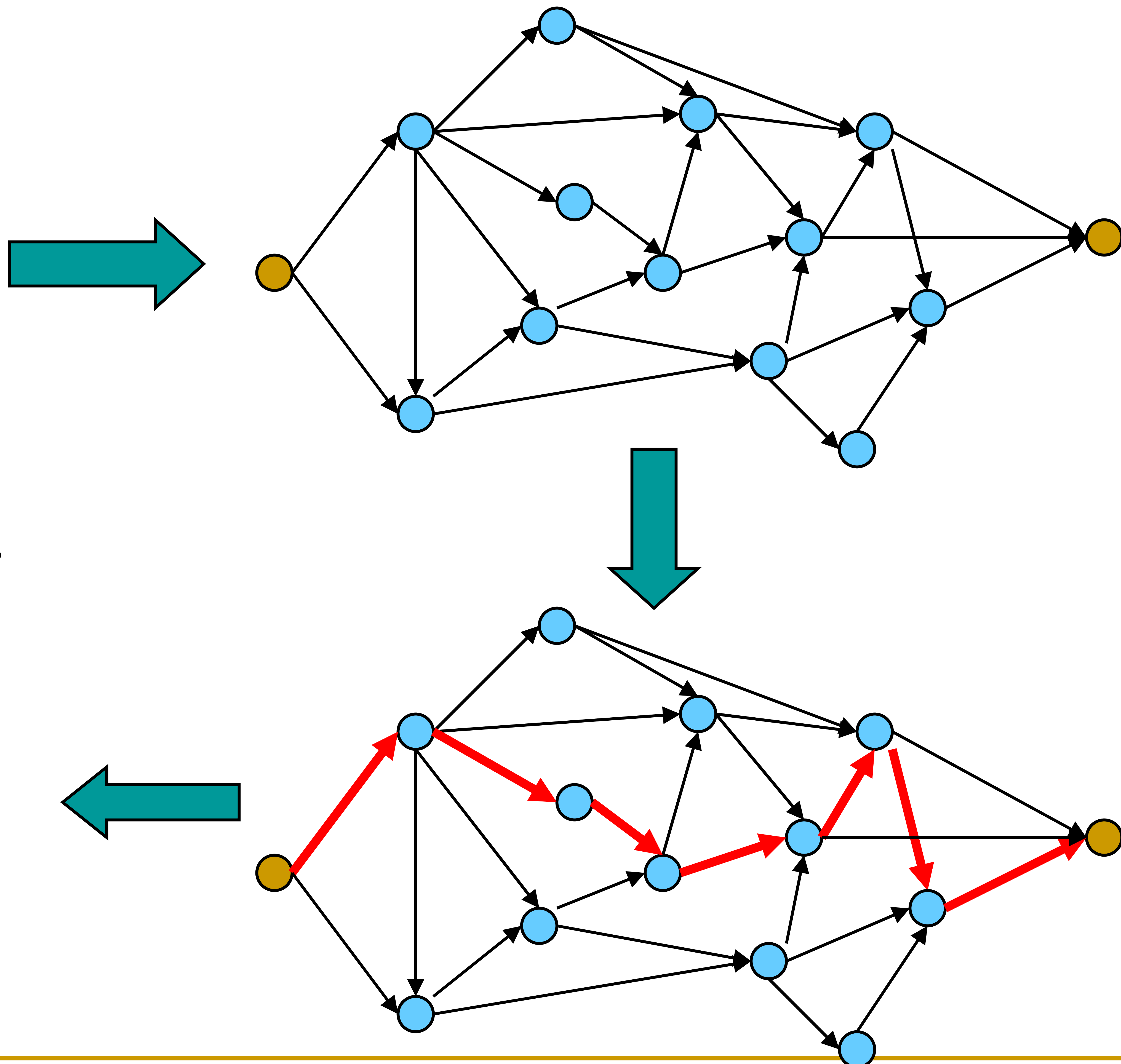
De Novo vs. Database Search: A Paradox

- The database of all peptides is huge $\approx O(20^n)$.
- The database of all known peptides is much smaller $\approx O(10^8)$.
- However, *de novo* algorithms can be much *faster*, even though their search space is much *larger*!
- A database search scans all peptides in the ***database of all known peptides*** search space to find best one.
- De novo eliminates the need to scan ***database of all peptides*** by modeling the problem as a graph search.

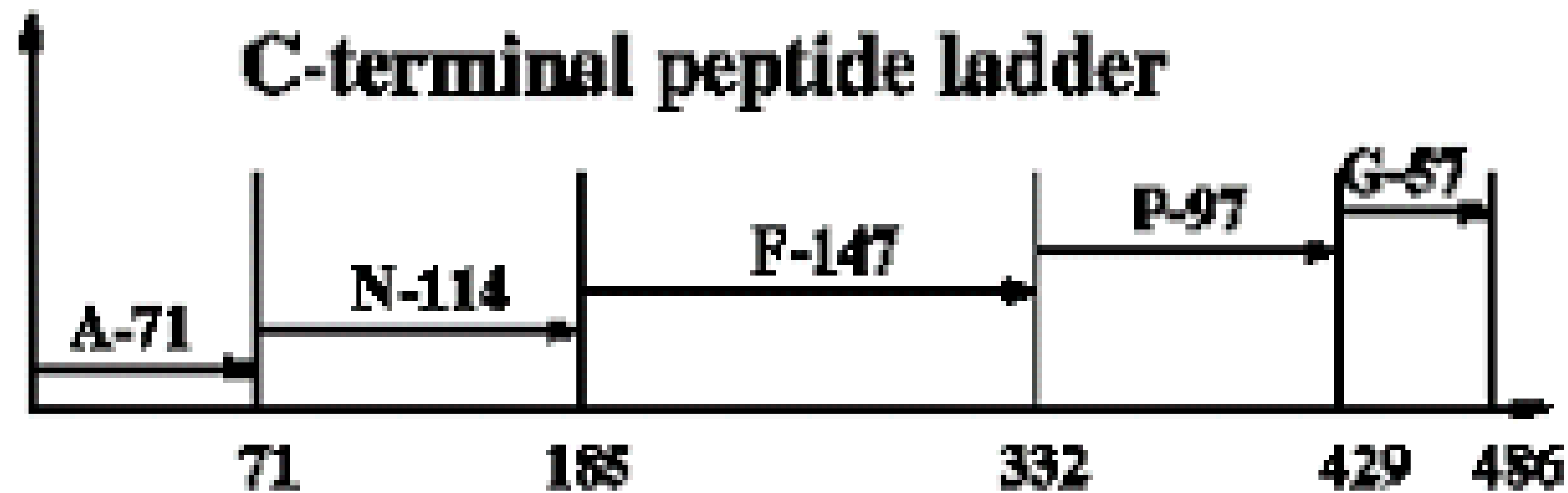
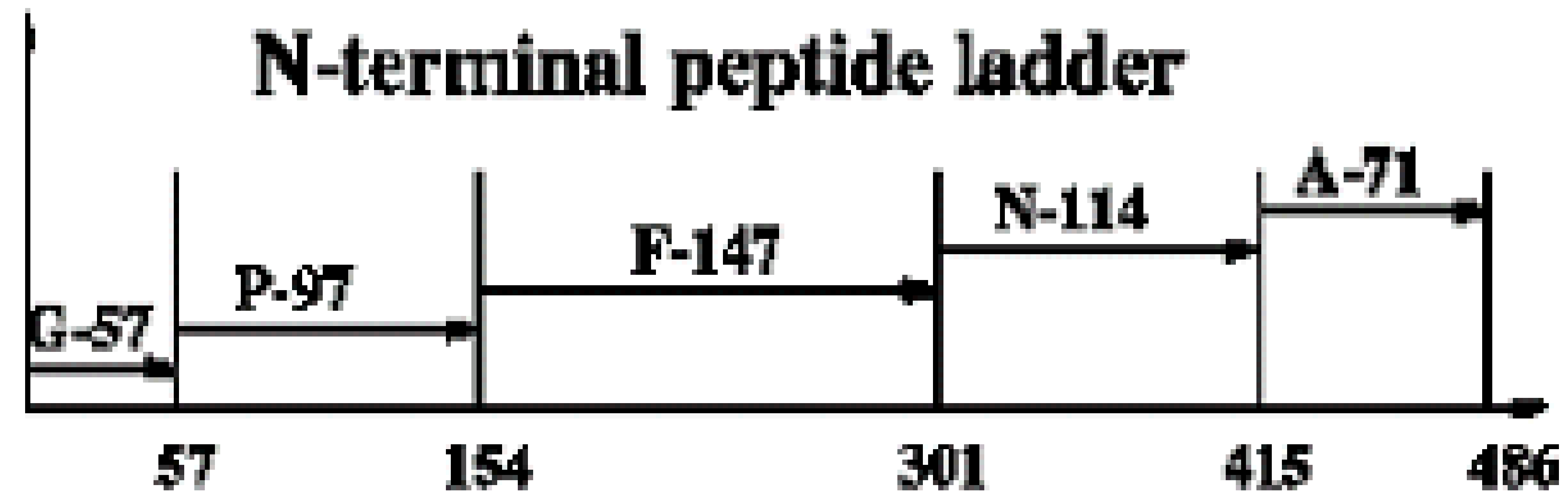
De novo Peptide Sequencing



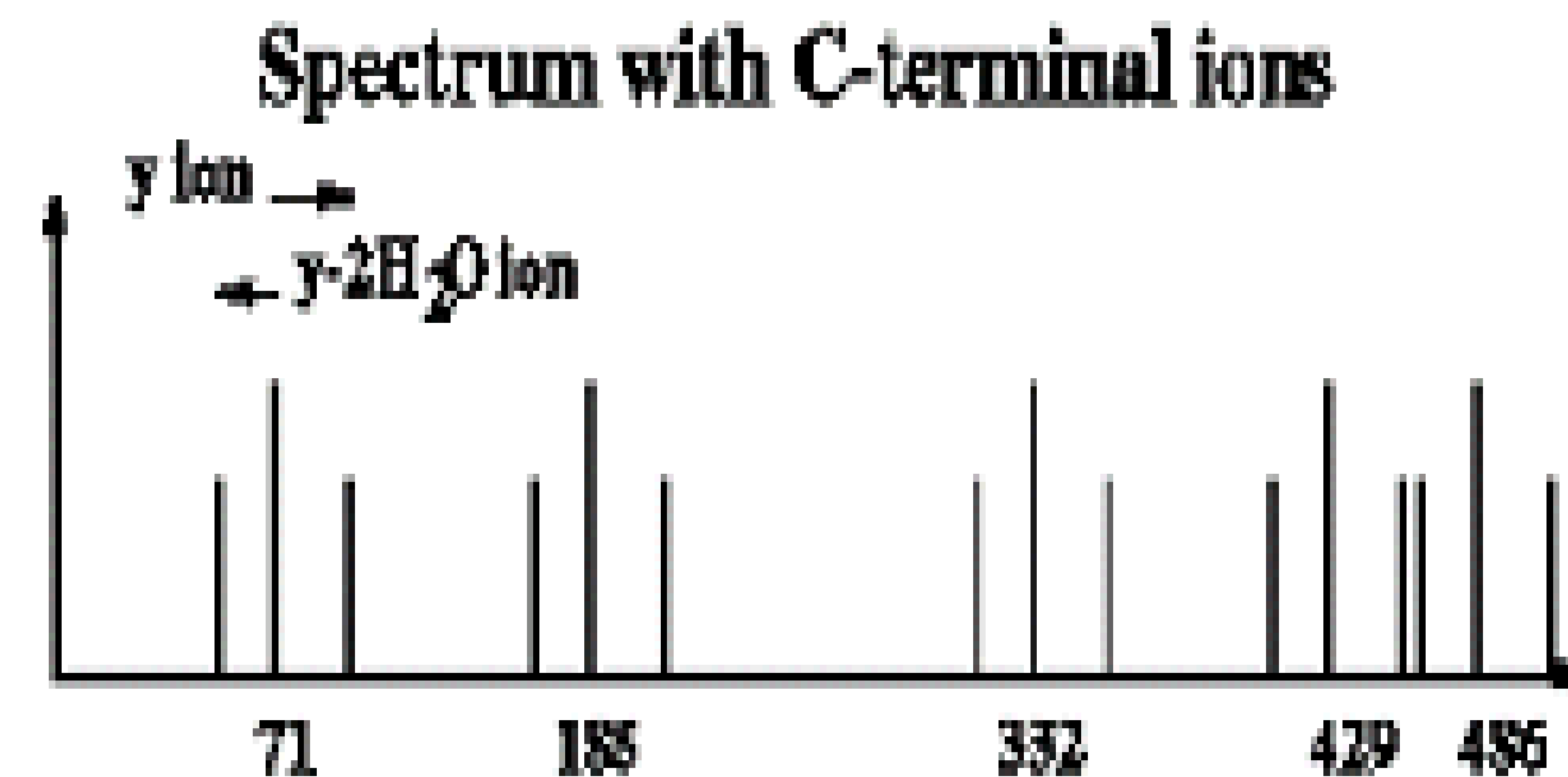
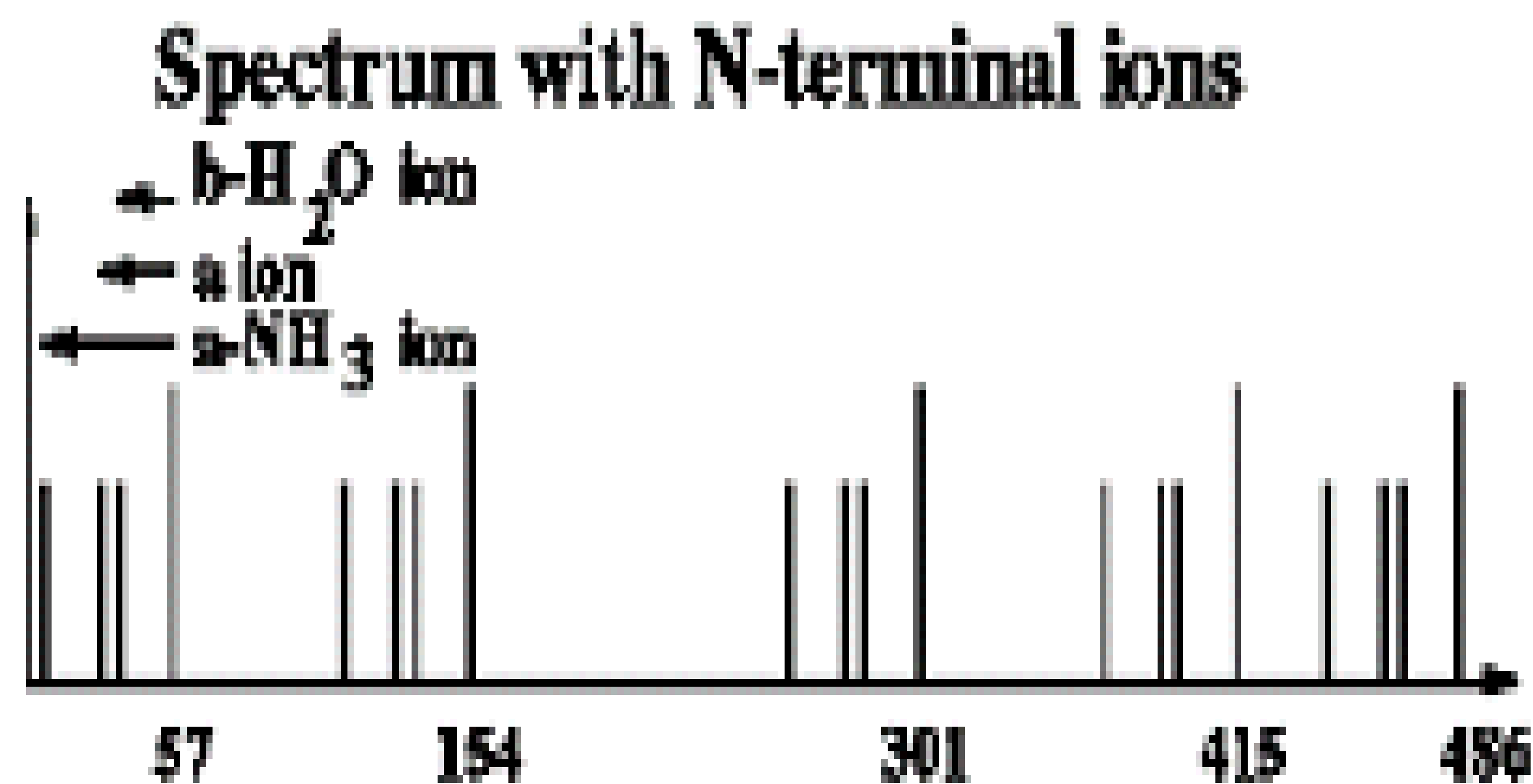
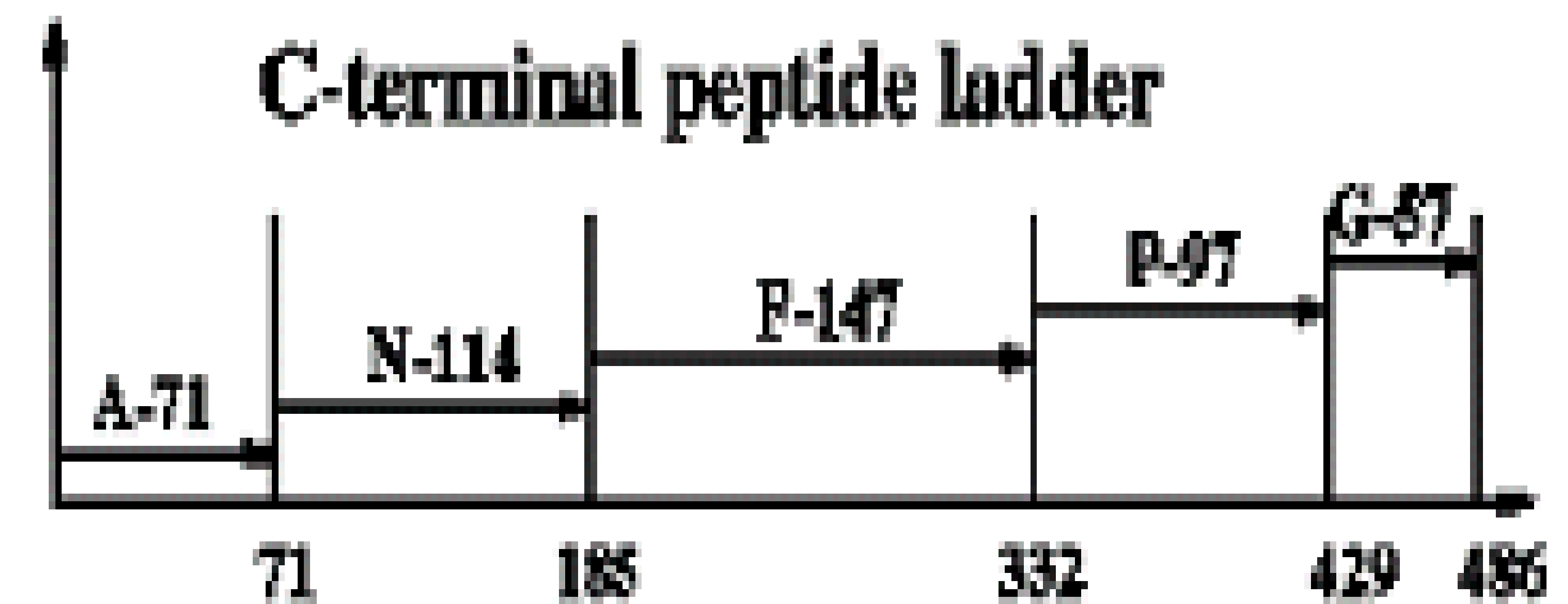
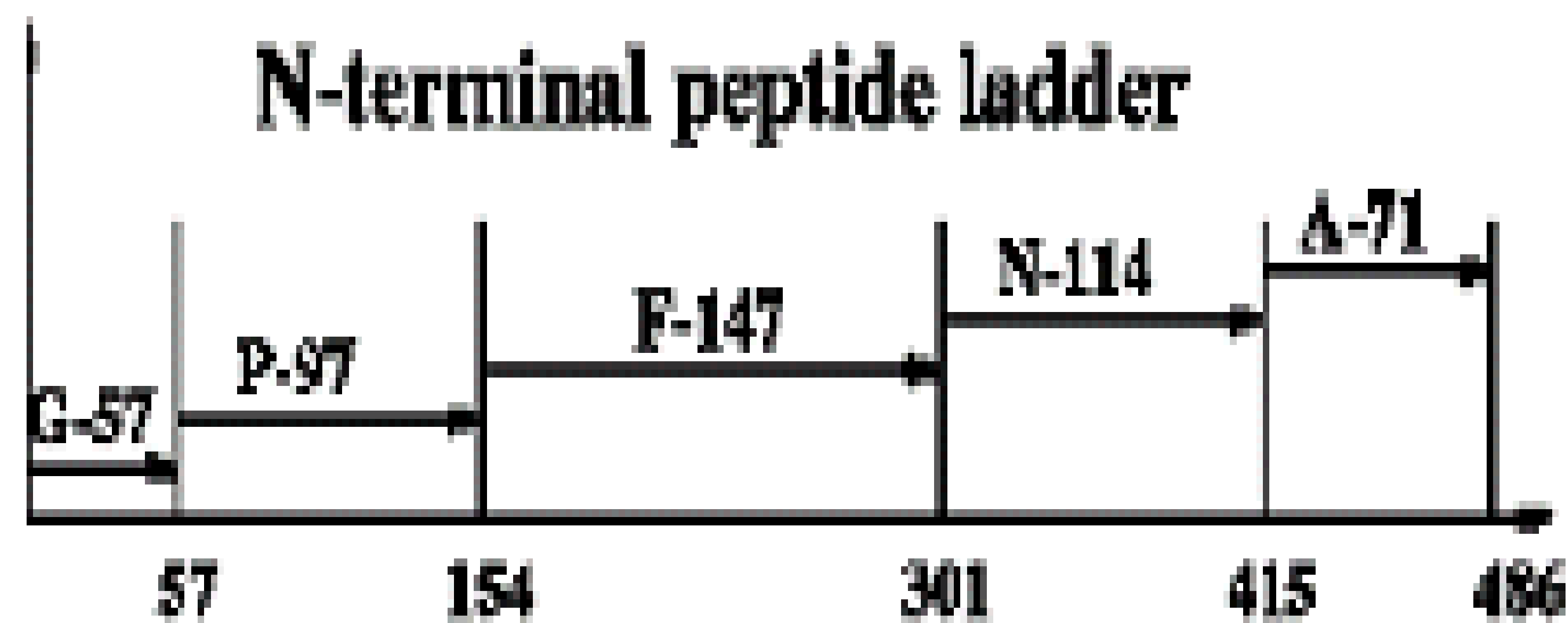
Sequence



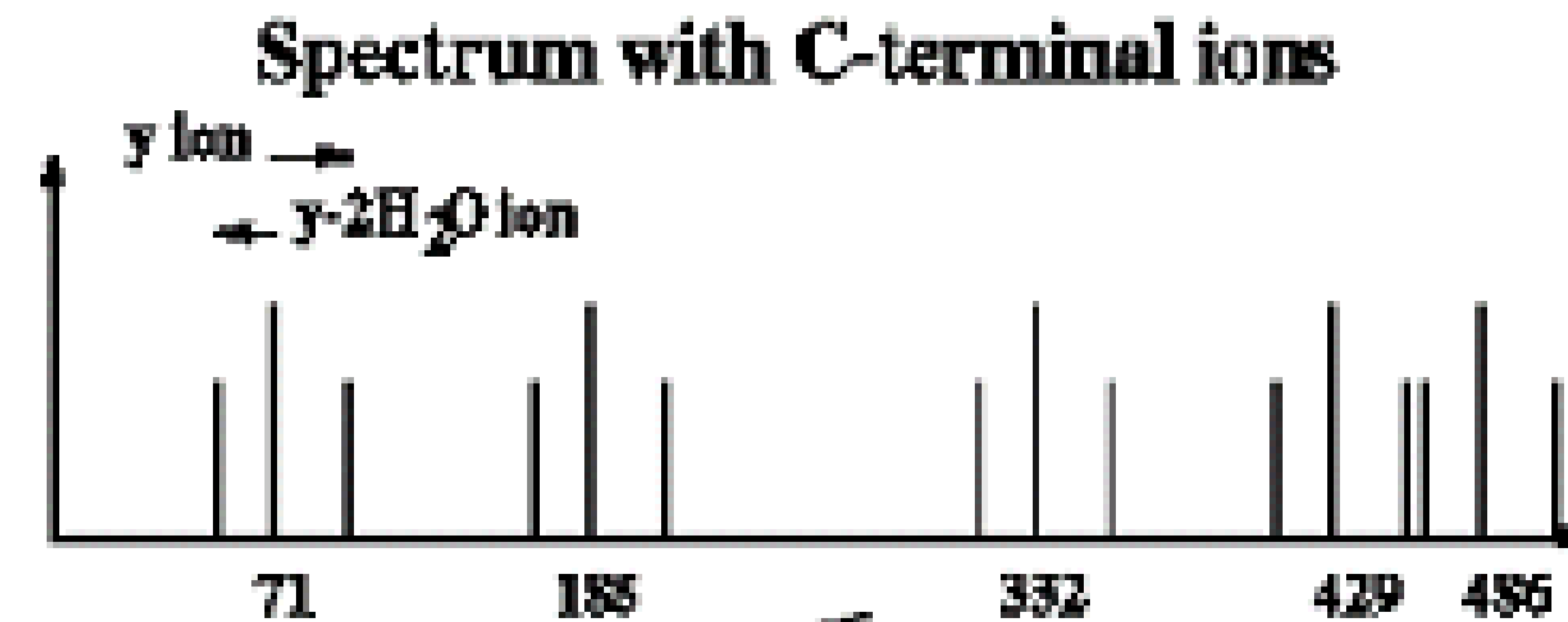
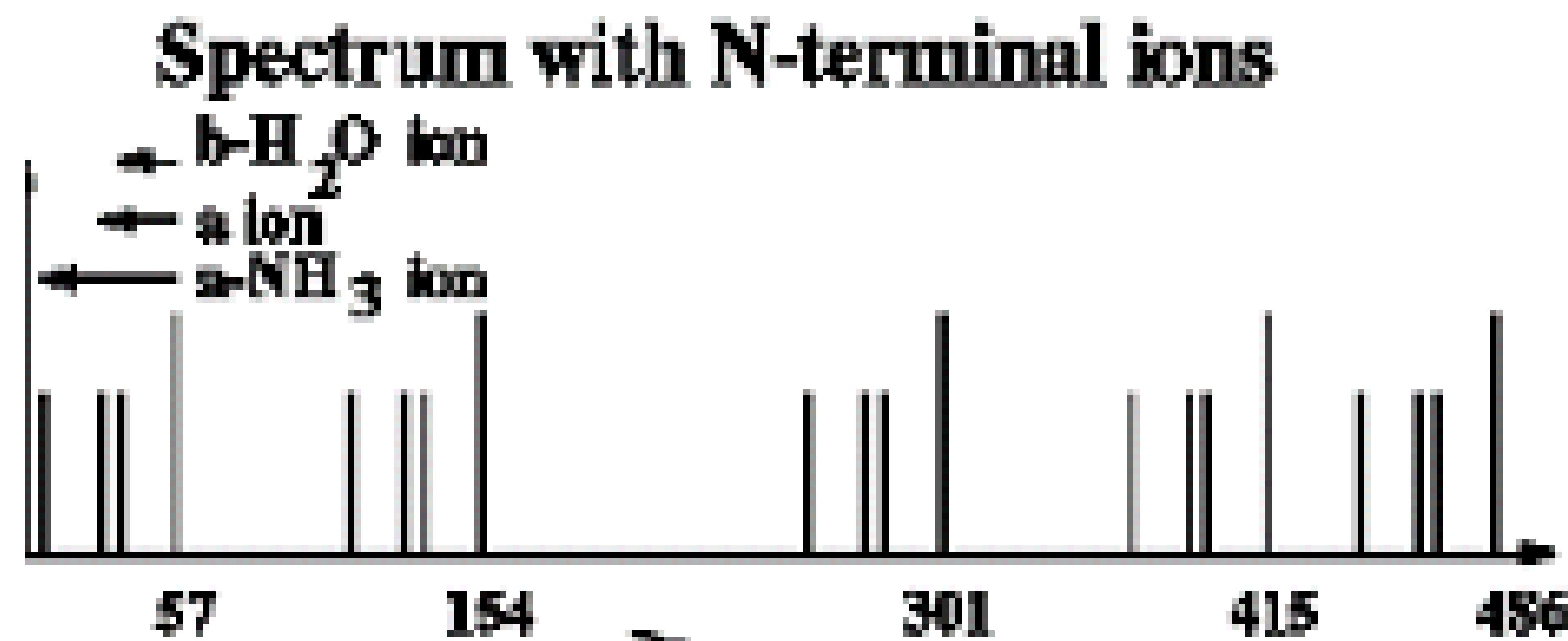
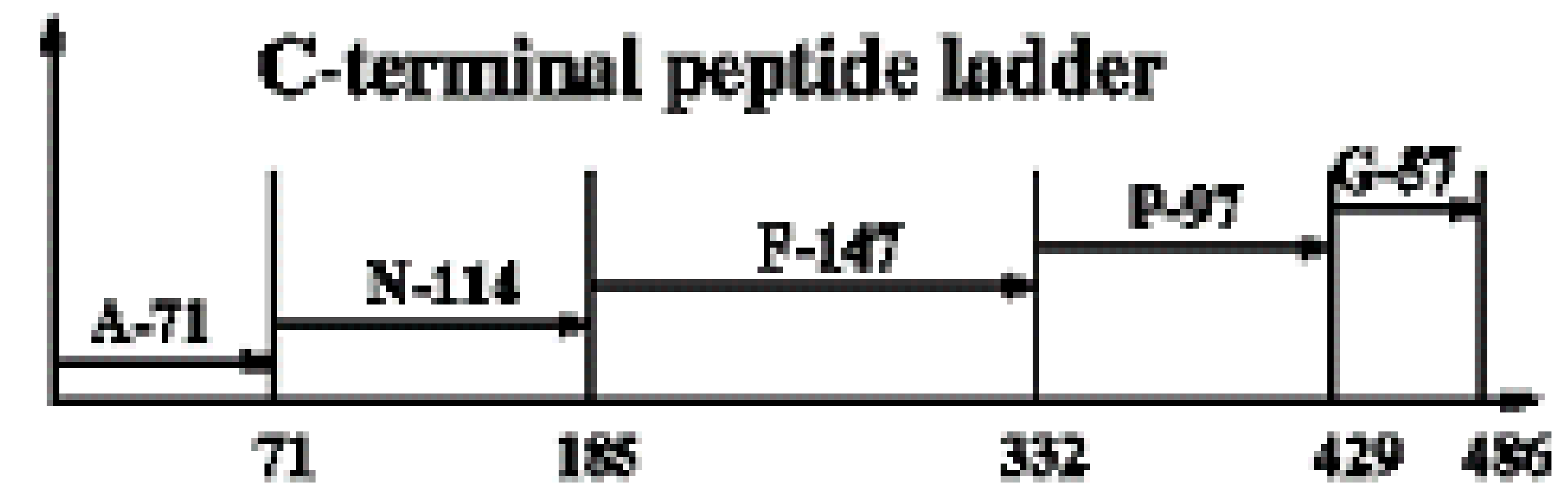
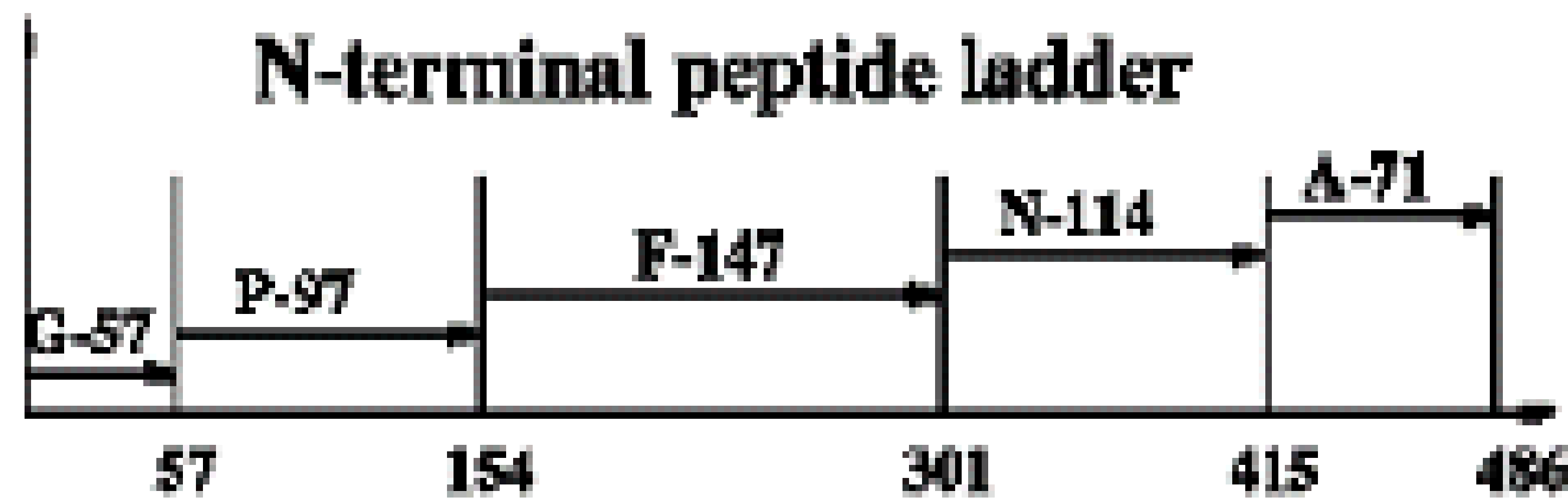
Theoretical Spectrum



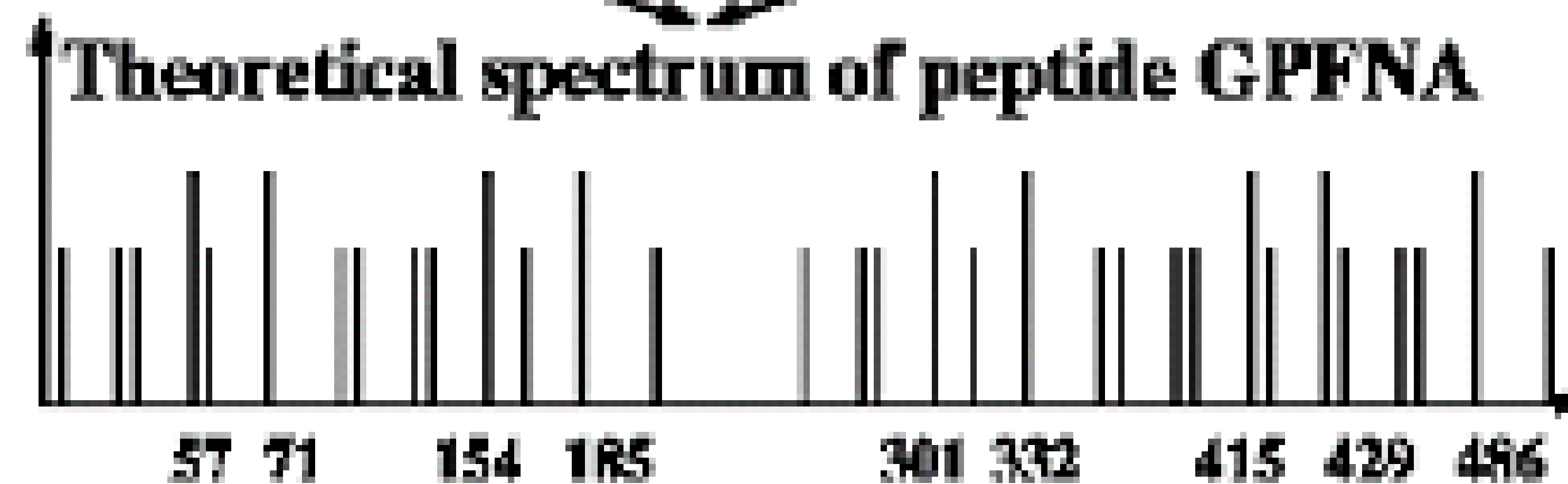
Theoretical Spectrum (cont'd)



Theoretical Spectrum (cont'd)

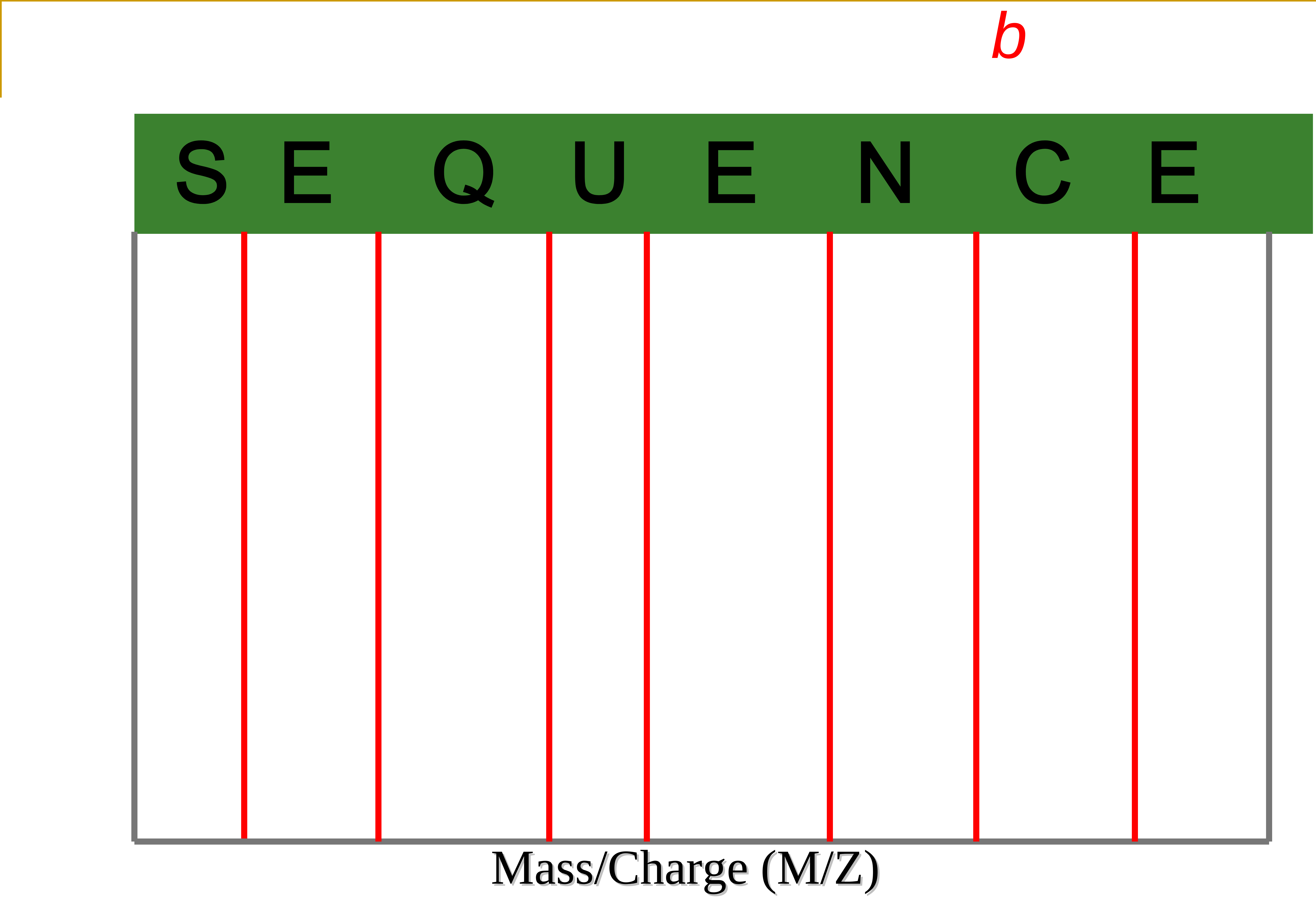


superposition

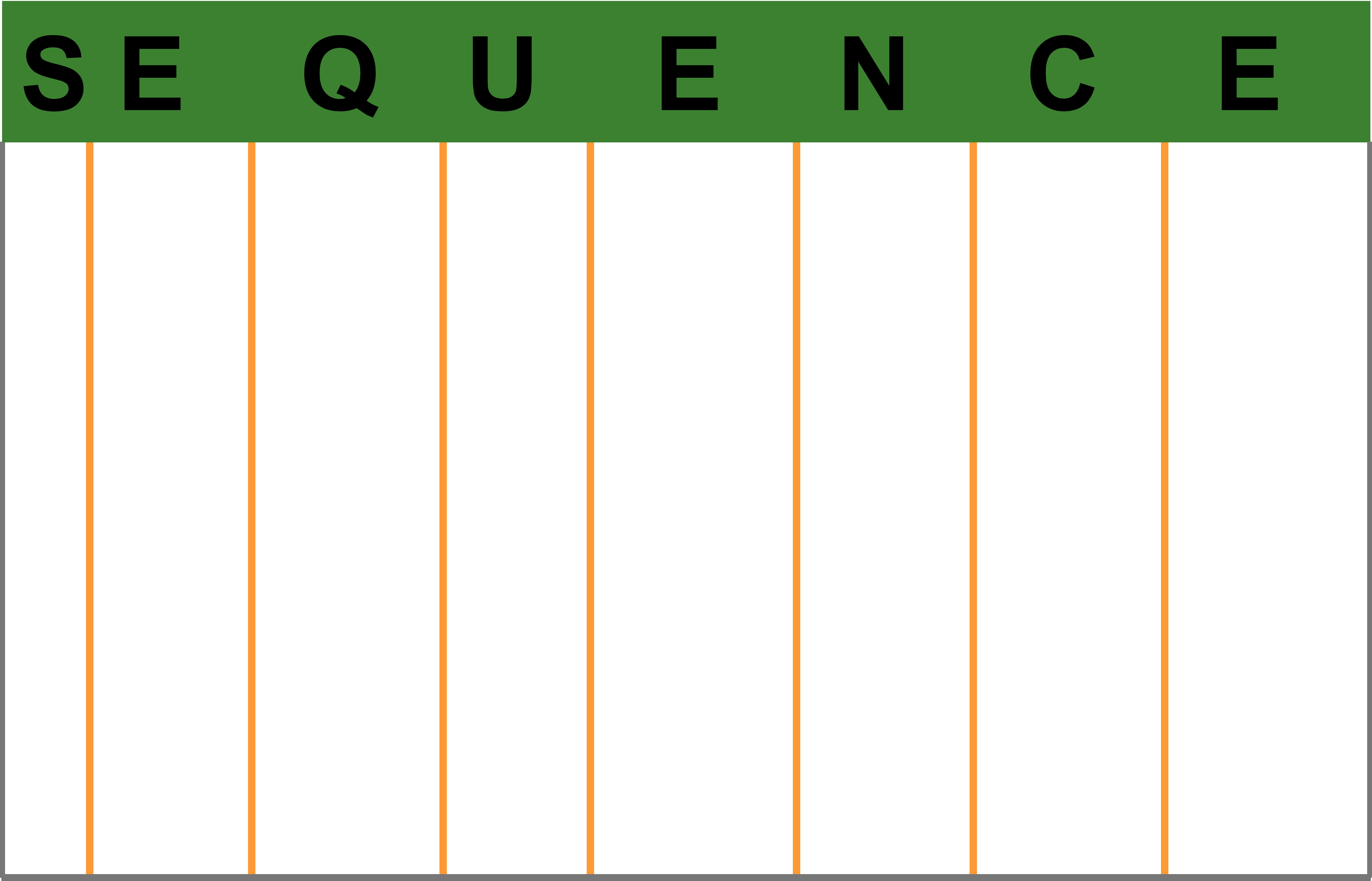


Building Spectrum Graph

- How to create vertices (from masses)
- How to create edges (from mass differences)
- How to score paths
- How to find best path

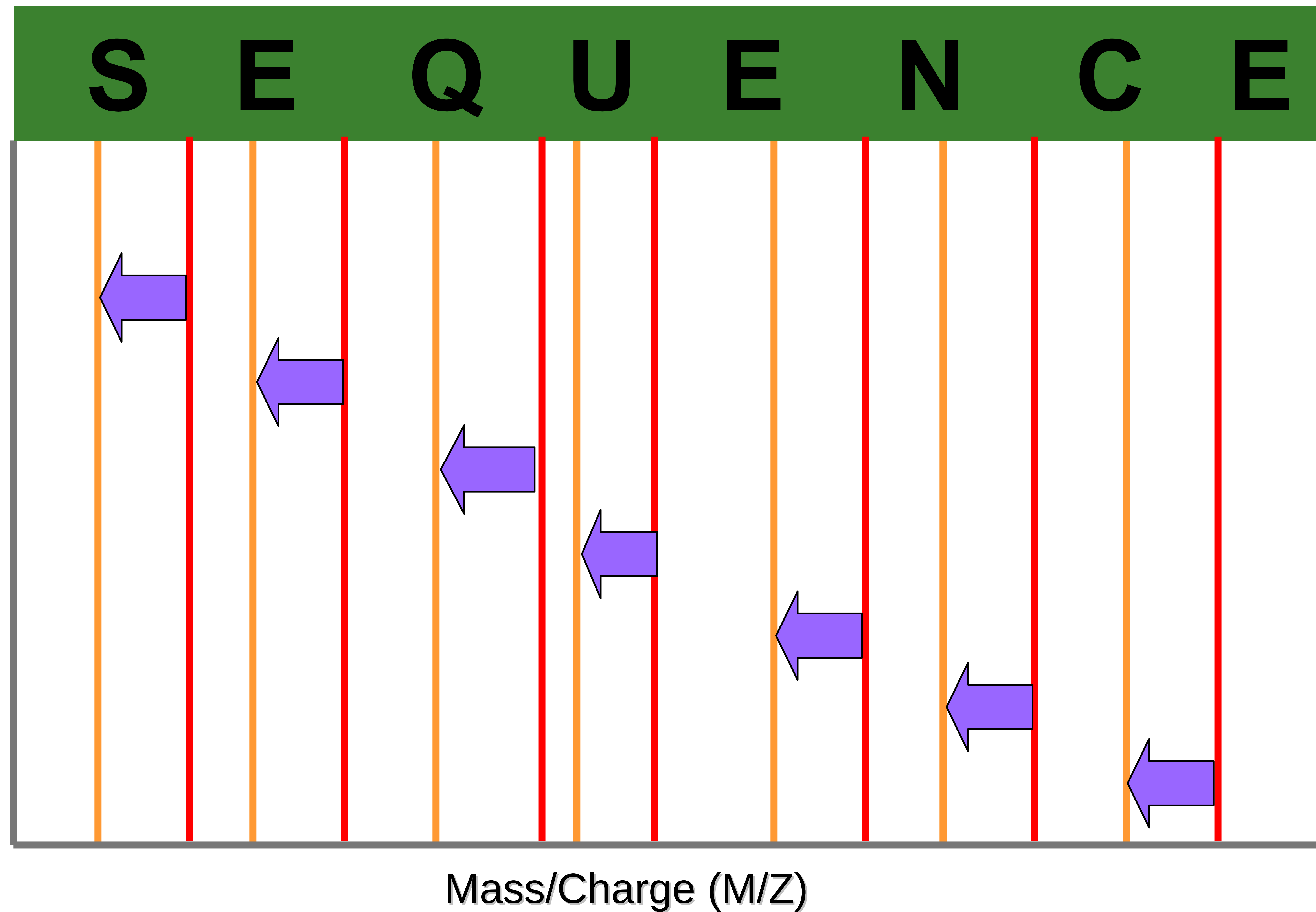


a



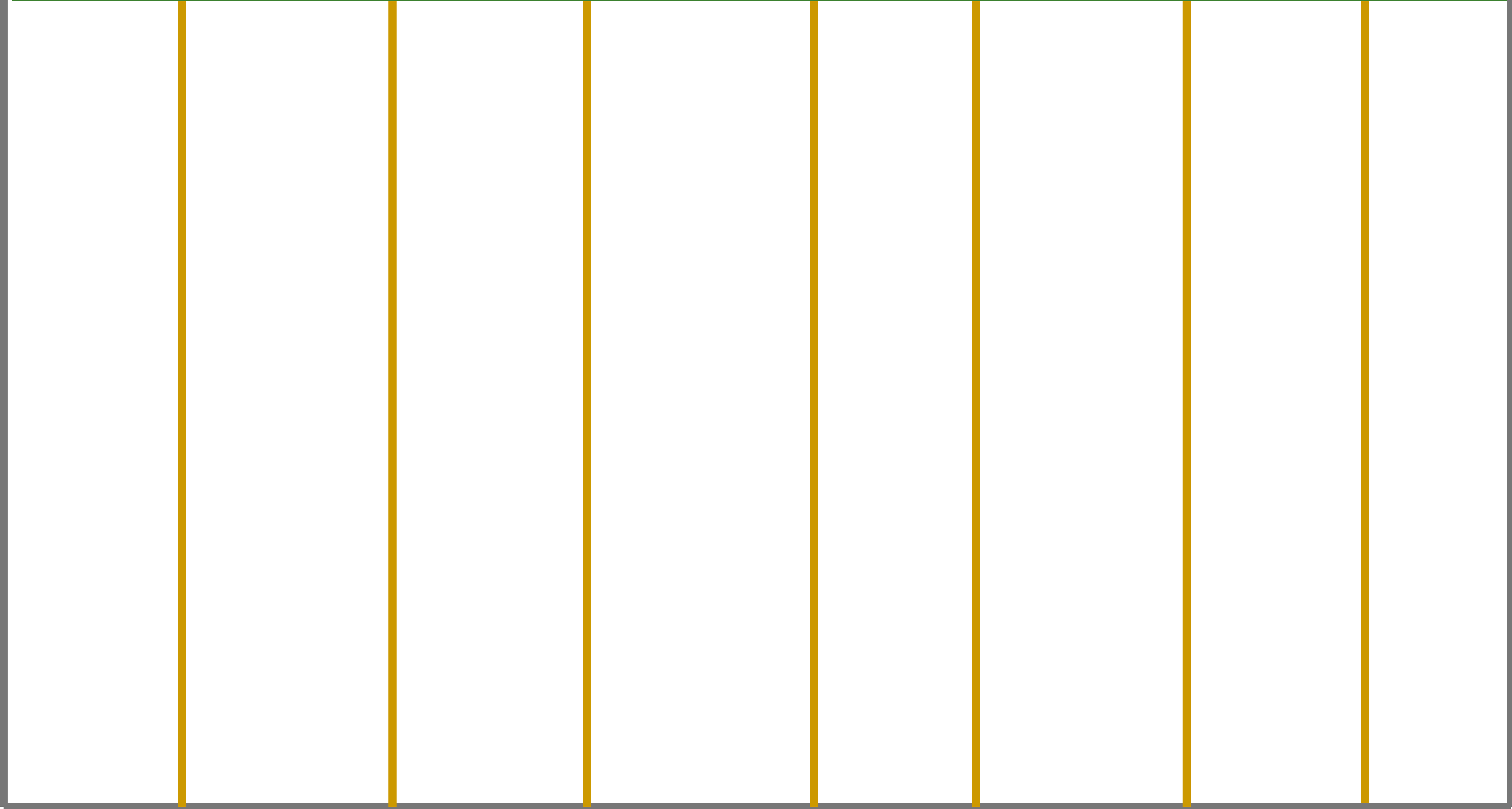
Mass/Charge (M/Z)

a is an ion type shift in *b*

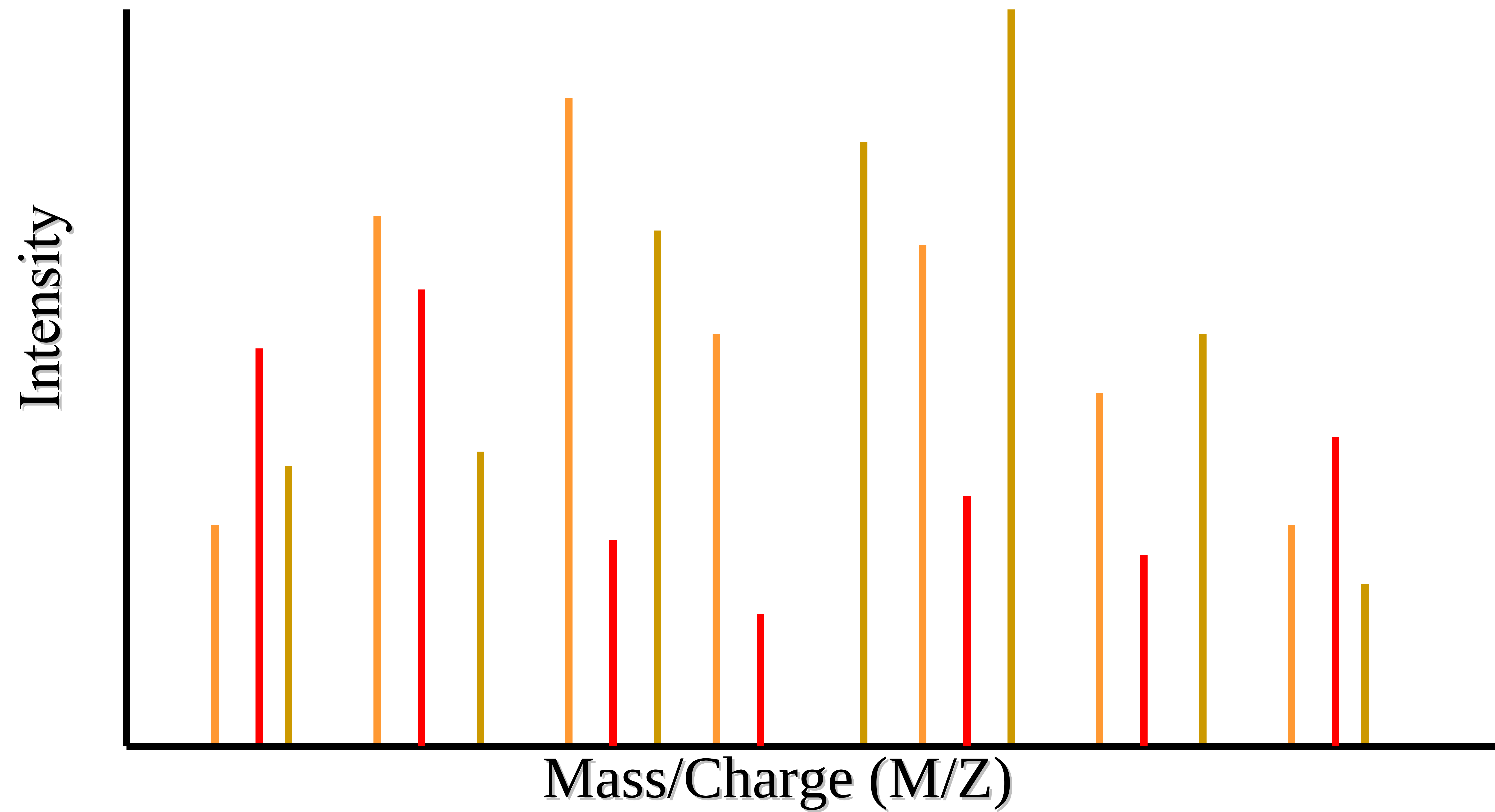


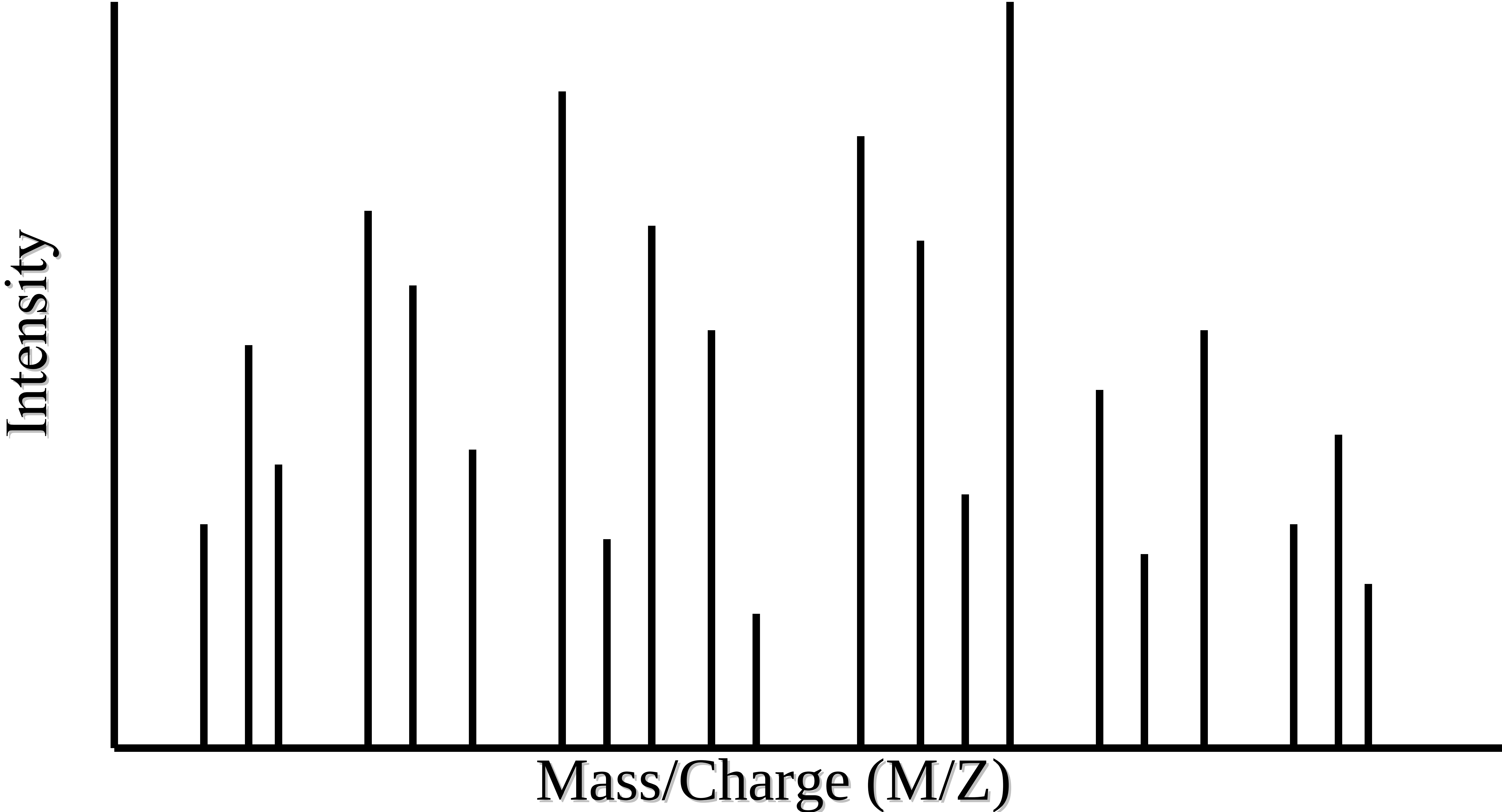
y

E C N E U Q E S

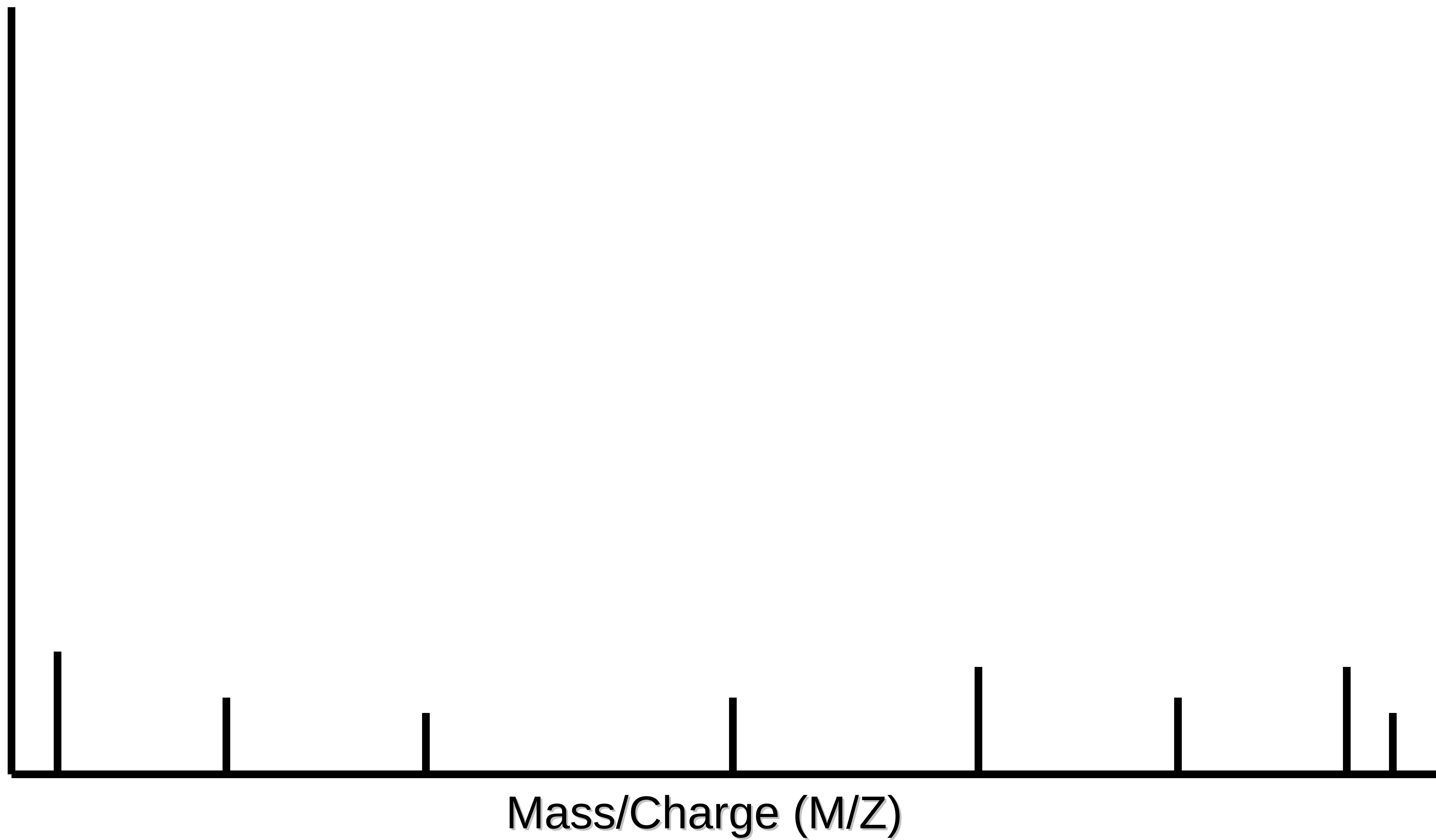


Mass/Charge (M/Z)

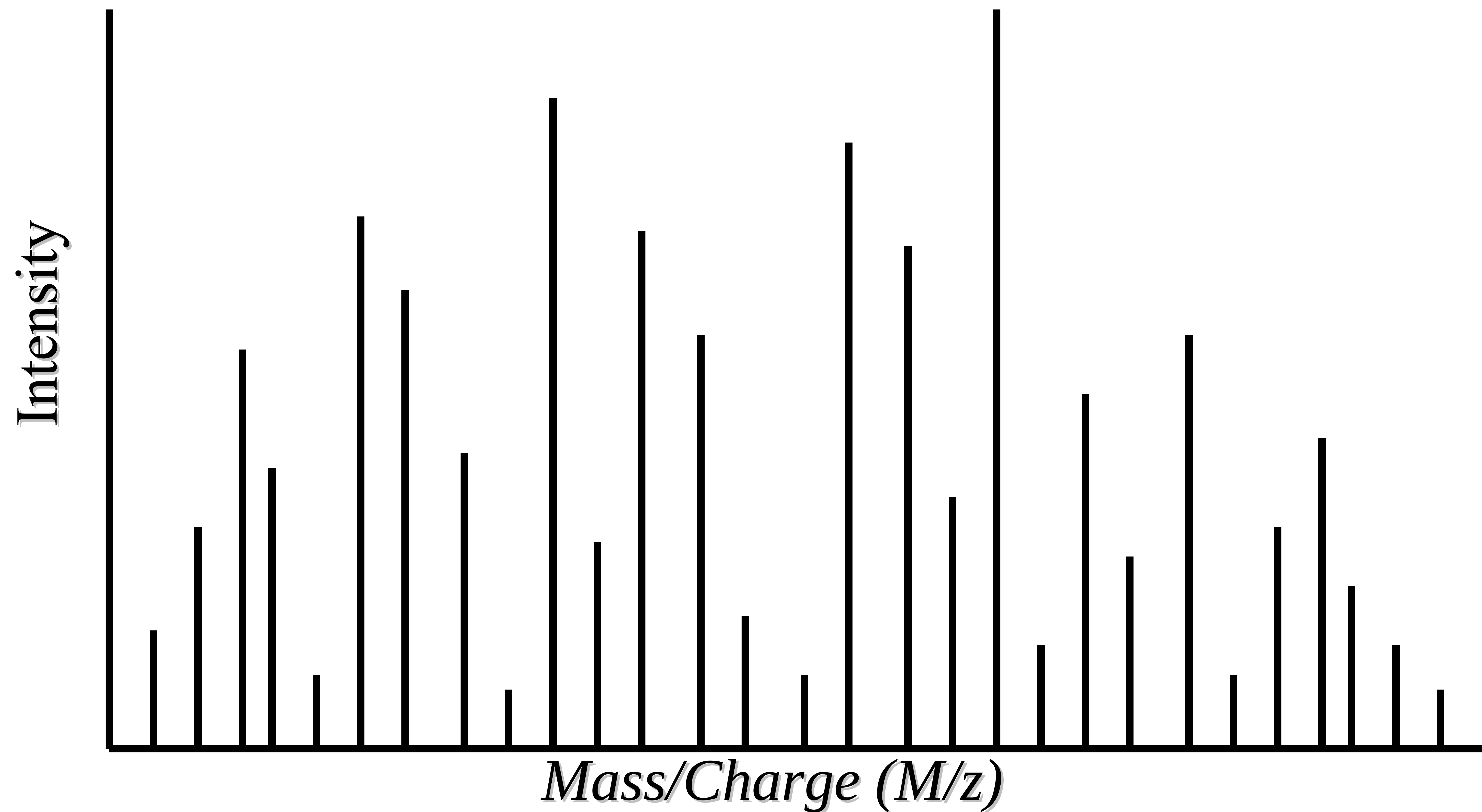




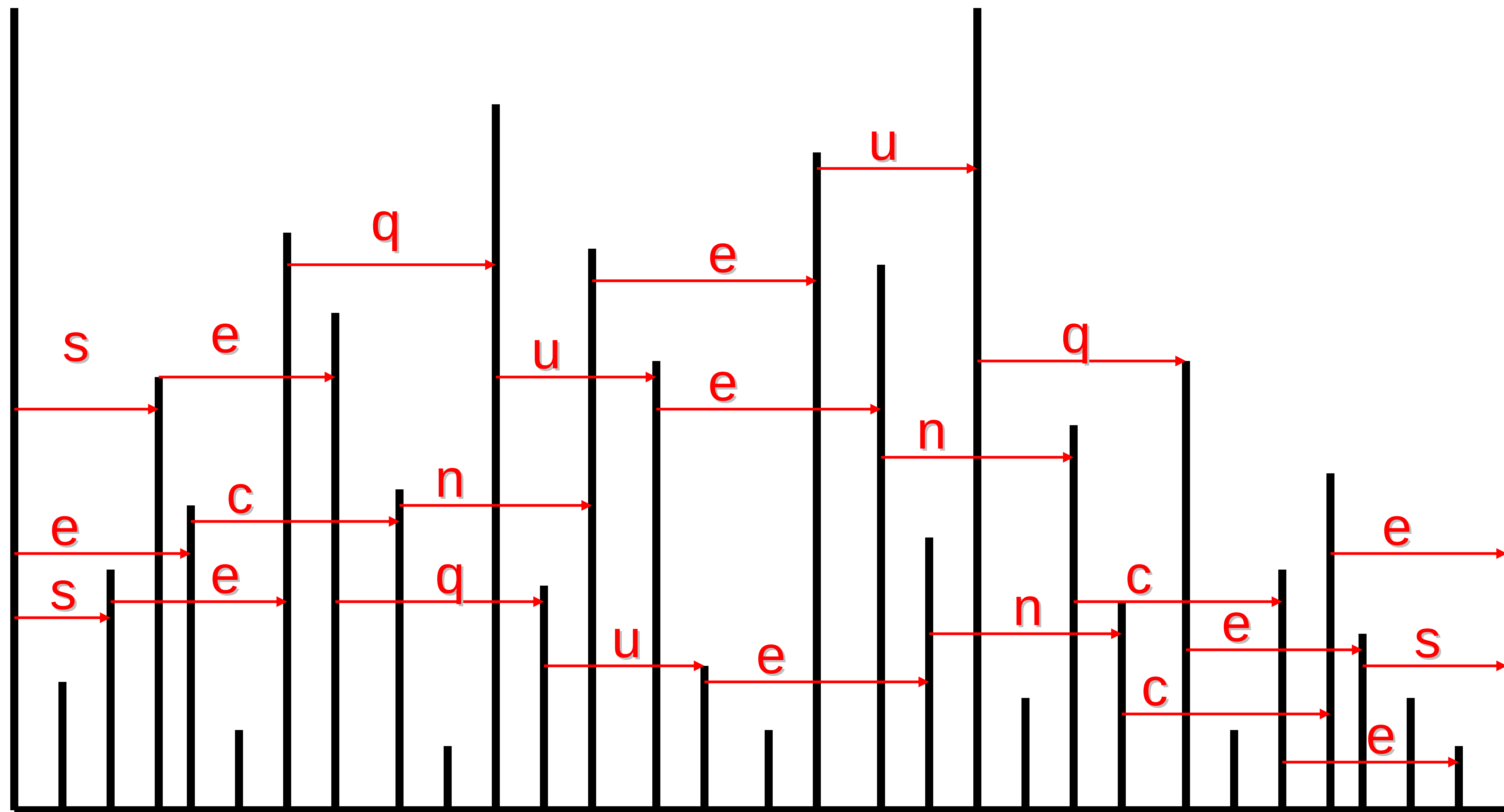
noise



MS/MS Spectrum



Some Mass Differences between Peaks Correspond to Amino Acids



Ion Types

- Some masses correspond to fragment ions, others are just random noise
- Knowing **ion types** $\Delta = \{\delta_1, \delta_2, \dots, \delta_k\}$ lets us distinguish fragment ions from noise
- We can **learn** ion types δ_i and their probabilities q_i by analyzing a large test sample of annotated spectra.

Example of Ion Type

- $\Delta = \{\delta_1, \delta_2, \dots, \delta_k\}$
- Ion types

$$\{b, b\text{-NH}_3, b\text{-H}_2\text{O}\}$$

correspond to

$$\Delta = \{0, 17, 18\}$$

*Note: In reality the δ value of ion type b is -1 but we will “hide” it for the sake of simplicity

Match between Spectra and the Shared Peak Count

- The match between two spectra is the number of masses (peaks) they share (**Shared Peak Count or SPC**)
- In practice mass-spectrometrists use the weighted SPC that reflects intensities of the peaks
- Match between experimental and theoretical spectra is defined similarly

Peptide Sequencing Problem

Goal: Find a peptide with maximal match between an experimental and theoretical spectrum.

Input:

- S : experimental spectrum
- Δ : set of possible ion types
- m : parent mass

Output:

- P : peptide with mass m , whose theoretical spectrum matches the experimental S spectrum the best

Vertices of Spectrum Graph

- Masses of potential N-terminal peptides
- Vertices are generated by **reverse shifts** corresponding to ion types

$$\Delta = \{\delta_1, \delta_2, \dots, \delta_k\}$$

- Every N-terminal peptide can generate up to k ions

$$m - \delta_1, m - \delta_2, \dots, m - \delta_k$$

- Every mass s in an MS/MS spectrum generates k vertices

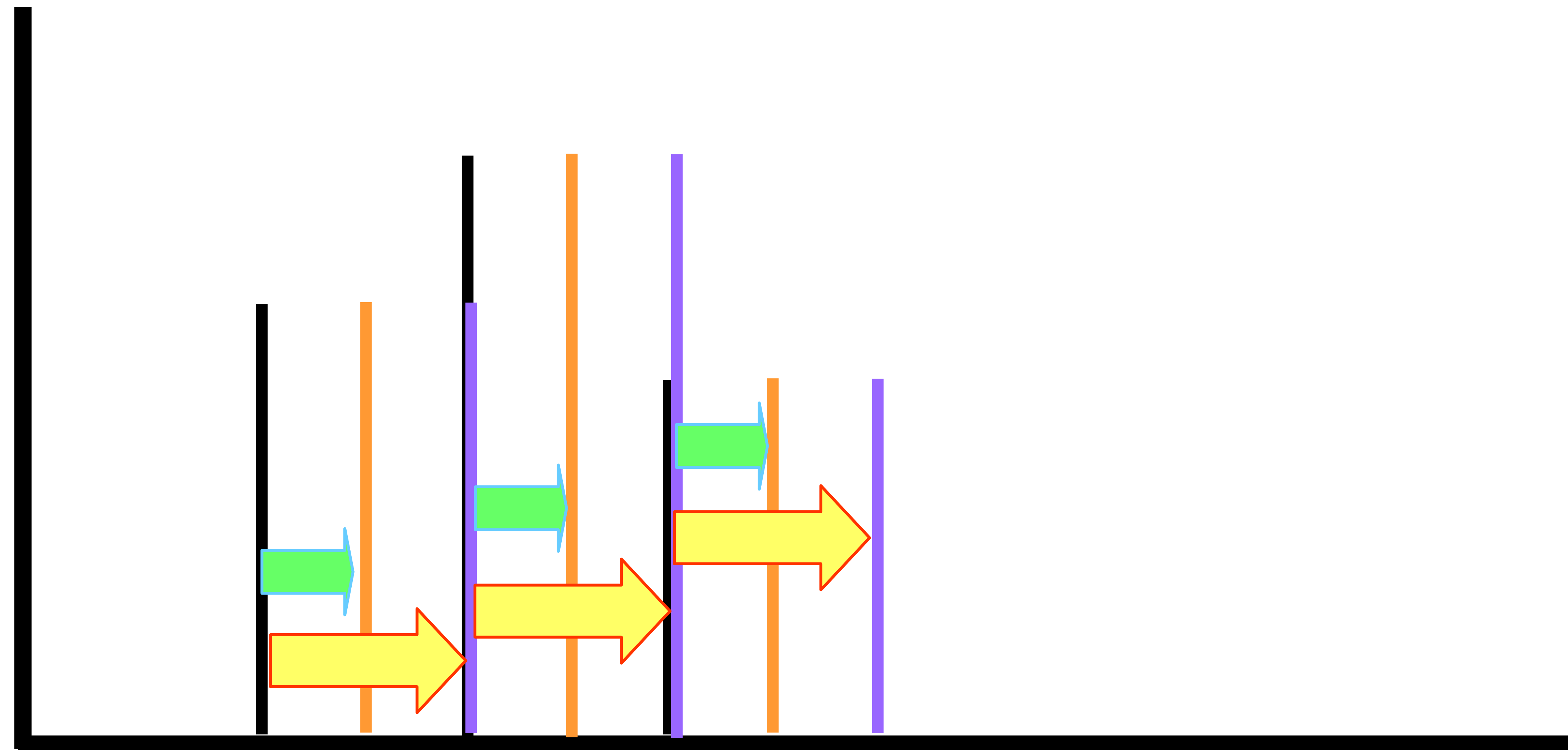
$$V(s) = \{s + \delta_1, s + \delta_2, \dots, s + \delta_k\}$$

corresponding to potential N-terminal peptides

- **Vertices of the spectrum graph:**

$$\{\text{initial vertex}\} \cup V(s_1) \cup V(s_2) \cup \dots \cup V(s_m) \cup \{\text{terminal vertex}\}$$

Reverse Shifts



→ Shift in H_2O

→ Shift in $H_2O + NH_3$

Edges of Spectrum Graph

- Two vertices with mass difference corresponding to an amino acid A :
 - Connect with an edge labeled by A
- Gap edges for di- and tri-peptides

Paths

- Path in the labeled graph spell out amino acid sequences
 - There are many paths, how to find the correct one?
 - We need **scoring** to evaluate paths
-

Path Score

- $p(P, S)$ = probability that peptide P produces spectrum $S = \{s_1, s_2, \dots, s_q\}$
- $p(P, s)$ = the probability that peptide P generates a peak s
- Scoring = computing probabilities
- $p(P, S) = \prod_{s \in S} p(P, s)$

Peak Score

- For a position t that represents ion type d_j :

$$p(\mathbf{P}, s_t) = \begin{cases} q_j, & \text{if peak is generated at } t \\ 1 - q_j, & \text{otherwise} \end{cases}$$

Peak Score (cont'd)

- For a position t that is not associated with an ion type:

$$p_R(\mathbf{P}, s_t) = \begin{cases} q_R, & \text{if peak is generated at } t \\ 1 - q_R, & \text{otherwise} \end{cases}$$

- q_R = the probability of a noisy peak that does not correspond to any ion type

Finding Optimal Paths in the Spectrum Graph

- For a given MS/MS spectrum $\mathbf{S}$, find a peptide $\mathbf{P}'$ maximizing $p(\mathbf{P}, \mathbf{S})$ over all possible peptides $\mathbf{P}$:

$$p(\mathbf{P}', \mathbf{S}) = \max_P p(\mathbf{P}, \mathbf{S})$$

- Peptides = paths in the spectrum graph
- $\mathbf{P}'$ = the optimal path in the spectrum graph

Ions and Probabilities

- Tandem mass spectrometry is characterized by a set of ion types $\{\delta_1, \delta_2, \dots, \delta_k\}$ and their probabilities $\{q_1, \dots, q_k\}$
- δ_i -ions of a partial peptide are produced ***independently*** with probabilities q_i

Ions and Probabilities

- A peptide has all k peaks with probability $\prod_{i=1}^k q_i$
- and no peaks with probability $\prod_{i=1}^k (1 - q_i)$
- A peptide also produces a ``random noise" with ***uniform*** probability q_R in any position.

Ratio Test Scoring for Partial Peptides

- Incorporates **premiums** for observed ions and **penalties** for missing ions.
- Example: for $k=4$, assume that for a partial peptide P' we only see ions $\delta_1, \delta_2, \delta_4$.

The score is calculated as:
$$\frac{q_1}{q_R} \cdot \frac{q_2}{q_R} \cdot \frac{(1-q_3)}{(1-q_R)} \cdot \frac{q_4}{q_R}$$

Scoring Peptides

- T - set of all positions.
- $T_i = \{t_{\delta_1}, t_{\delta_2}, \dots, t_{\delta_k}\}$ - set of positions that represent ions of partial peptides P_i .
- A peak at position t_{δ_j} is generated with probability q_j .
- $R = T - \bigcup T_i$ - set of positions that are not associated with any partial peptides (noise).

Probabilistic Model

- For a position $t_{\delta_j} \in T_i$ the probability $p(t, P, S)$ that peptide P produces a peak at position t .

$$P(t, P, S) = \begin{cases} q_j & \text{if a peak is generated at position } t_{\delta_j} \\ 1 - q_j & \text{otherwise} \end{cases}$$

- Similarly, for $t \in R$, the probability that P produces a random noise peak at t is:

$$P_R(t) = \begin{cases} q_R & \text{if a peak is generated at position } t \\ 1 - q_R & \text{otherwise} \end{cases}$$

Probabilistic Score

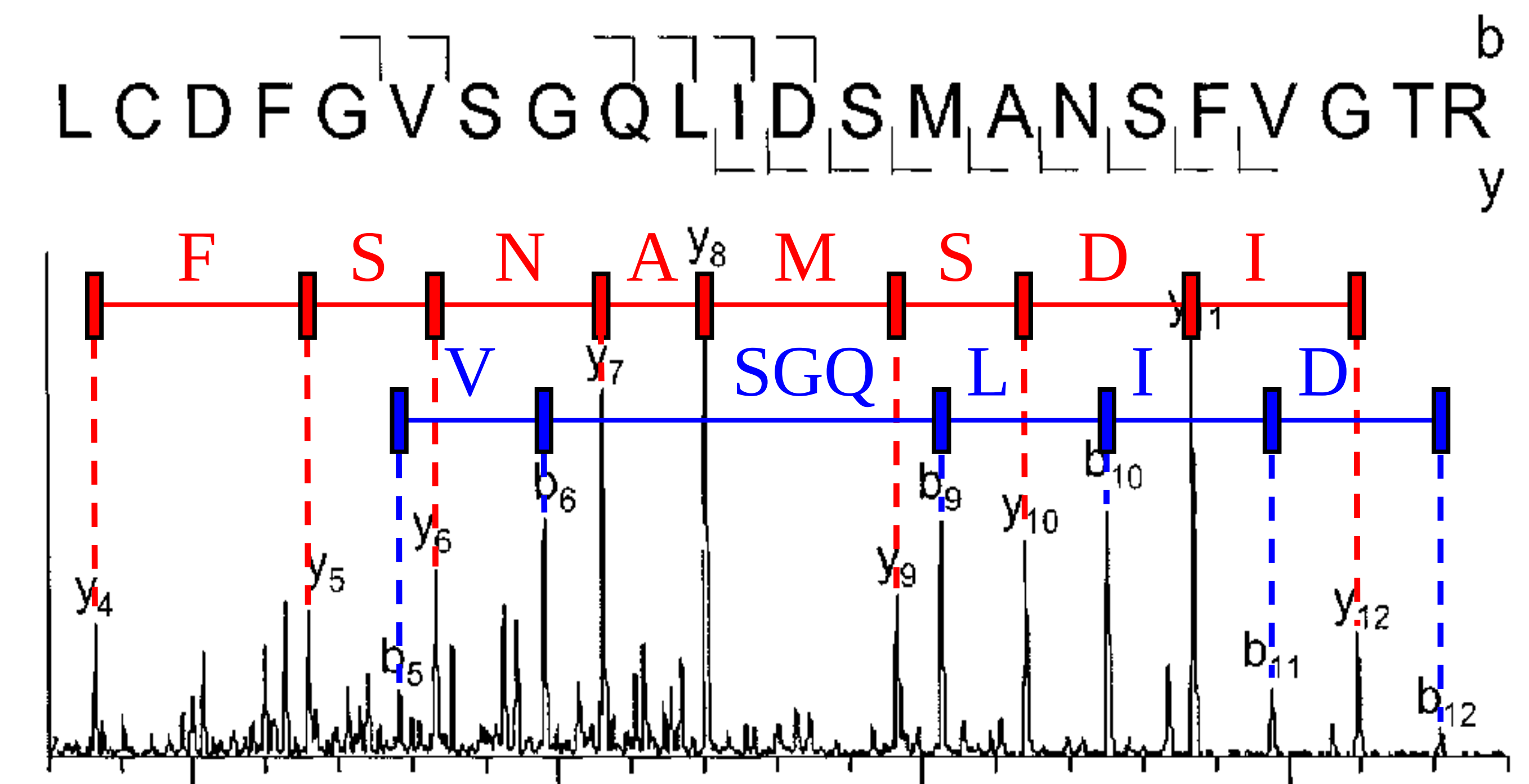
- For a peptide P with n amino acids, the score for the whole peptides is expressed by the following ratio test:

$$\frac{p(P, S)}{p_R(S)} = \prod_{i=1}^n \prod_{j=1}^k \frac{p(t_{i\delta_j}, P, S)}{p_R(t_{i\delta_j})}$$

De-novo sequencing: how?

How can we find the amino acid sequence that best explains the spectrum?

- a) Explains the largest number of peaks
- b) Explains the most intensity



- Exhaustive enumeration?
 1. Generate every possible sequence with the same peptide mass
 2. Match each sequence to the spectrum
 3. Choose the sequence that explains the most intensity in the spectrum

Takes too long!

Initial attempts

Computer programs for *de-novo* interpretation of MS/MS spectra date as far back as 1966 when Biemman et al. proposed a prefix extension algorithm:

- Tries every possible prefix extension
- Eliminates solutions with missing peaks
- Outputs every peptide with a matching parent mass
- ALL prefix peaks must be present in the spectrum

In an attempt to include interpretations with missing peaks, Sakurai et al. (1984) proposed:

- Exhaustive search over the space of all possible permutations of amino acid multisets where the total mass equals the parent mass of the MS/MS spectrum
- Score peptide/spectrum matches by counting prefix and suffix mass matches
- Naturally more sensitive but also *very* slow

Prefix extension revisited

The second half of the eighties saw a few better designed approaches to this problem, based on the same type of algorithm:

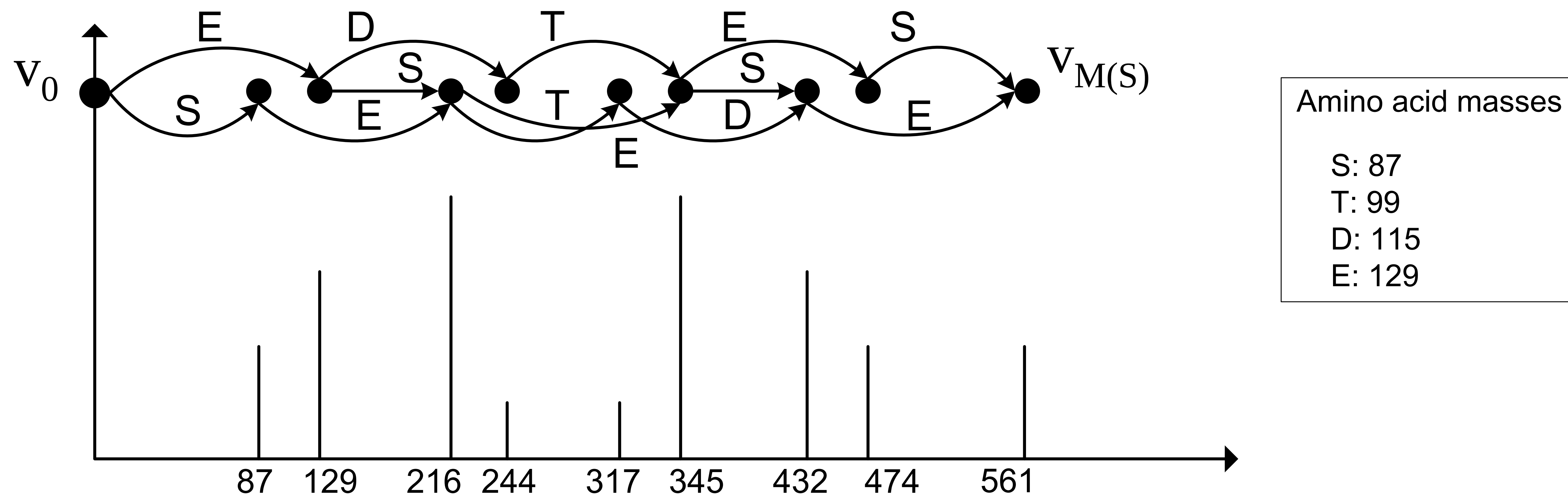
- Prefix extension, one amino acid at a time.
- Tolerate missing peaks.
- Include both prefix and suffix peaks in the score.
- User-specified maximum number of candidates in memory at any point of the execution. (sub-optimal)

Ishikawa and Niwa'86, Siegel and Baumann'88, Johnson and Biemann'89, Zidarov et al.'90

Spectrum graphs

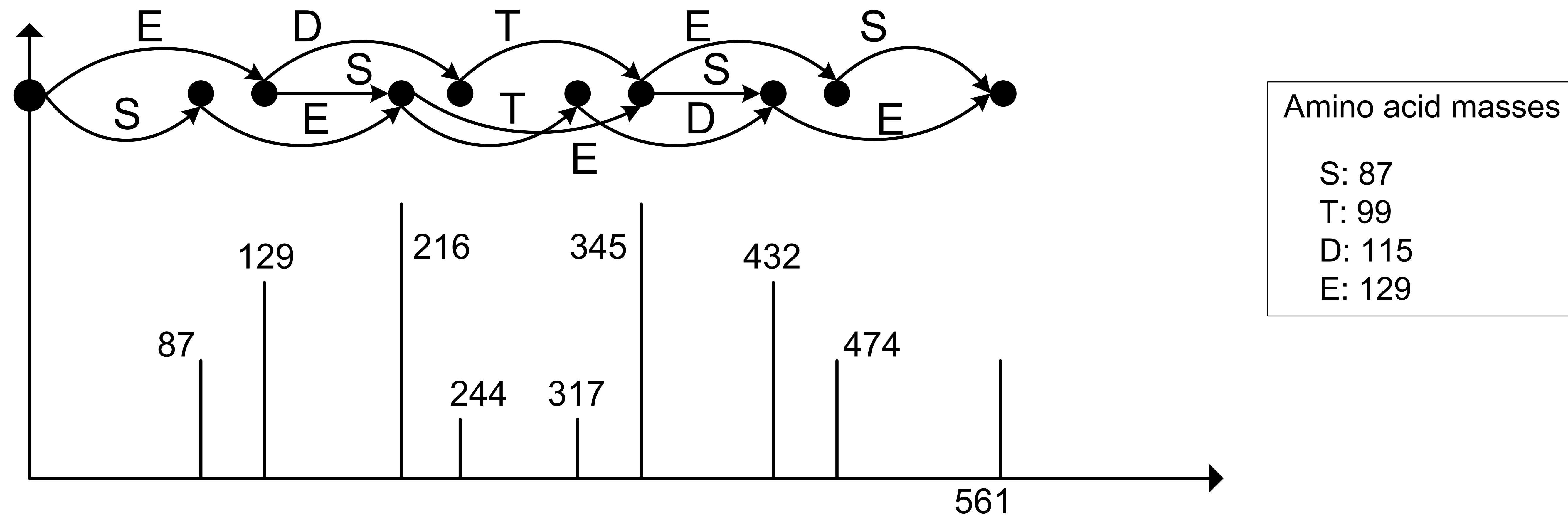
In 1990 Bartels introduced a graph representation of an MS/MS spectrum

- Every peak in the spectrum defines a vertex
- Vertices connected by an edge if peak mass difference is an amino acid mass



- Best peptide is defined as the best path between the two endpoint vertices: v_0 to $v_{M(S)}$
- No detailed algorithm was given for finding the best peptide; interactive exploration tool was made available.

DP de-novo sequencing



What is the DP recursion?

Score(i) = intensity(i) + max(Score(j)),
for all j with mass(i)-mass(j) ∈ Amino acid masses

Recovered de-novo sequence? ESESE