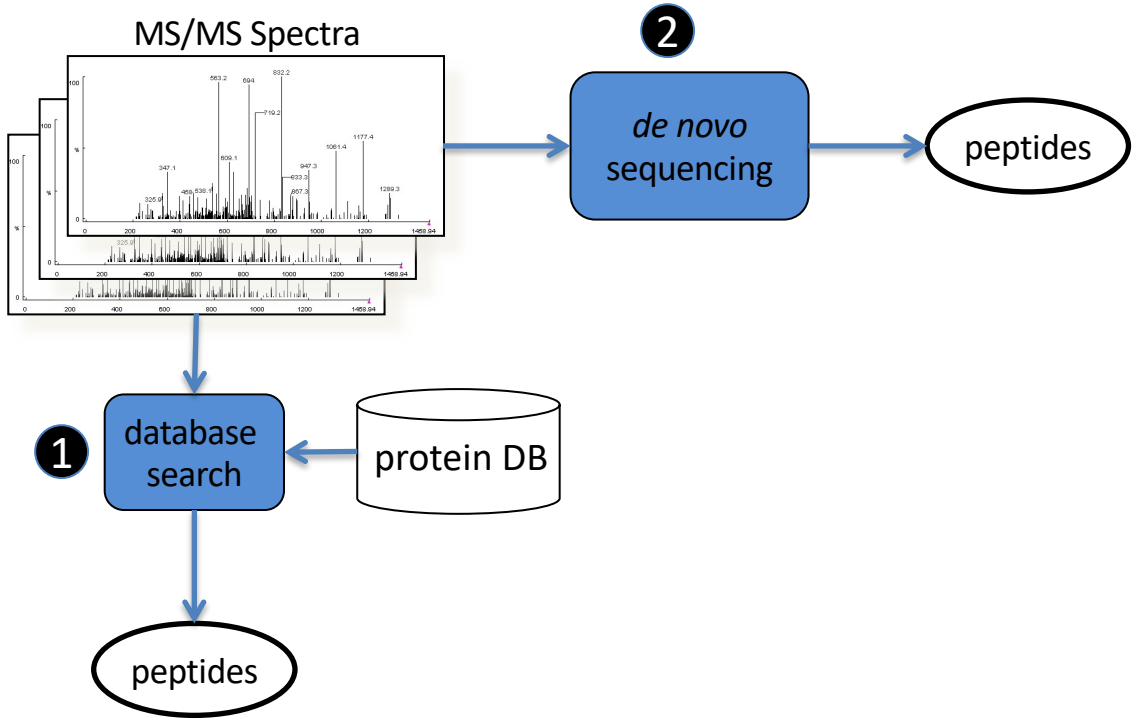
De Novo Peptide Sequencing

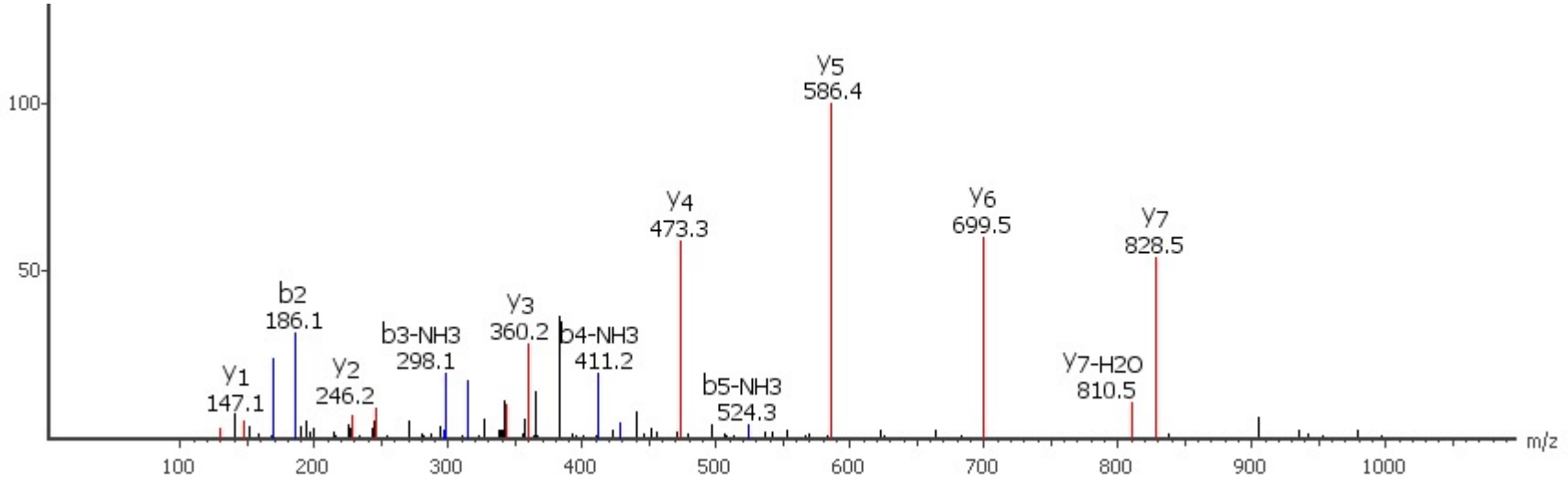
Possible Ways to Interpret MS/MS Data



De Novo Peptide Sequencing

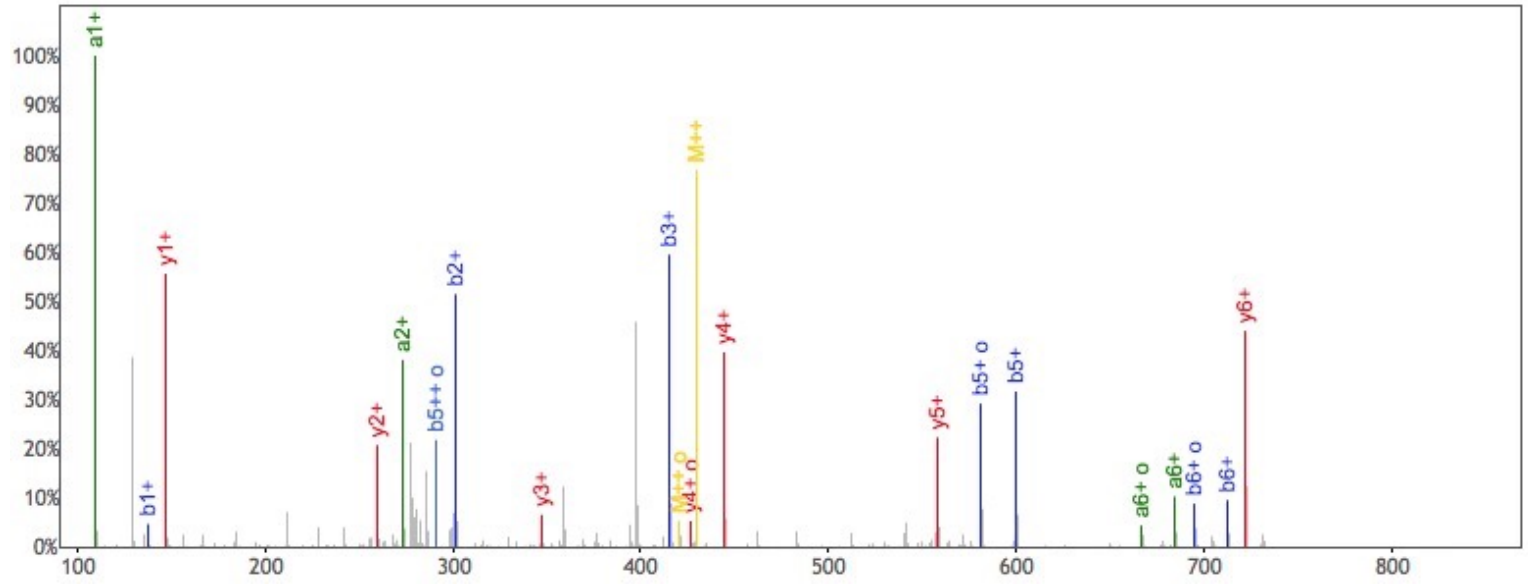
b_1	**A NELLLNVK**	y_8
b_2	**AN ELLLNVK**	y_7
b_3	**ANE LLLNVK**	y_6
b_4	**ANEL LLNVK**	y_5
b_5	**ANELL LNVK**	y_4
b_6	**ANELLL NVK**	y_3
b_7	**ANELLLN VK**	y_2
b_8	**ANELLLNV K**	y_1

Problem: To construct a sequence that matches the spectrum the best.



MS/MS Spectrum of Peptide HYNPSLK

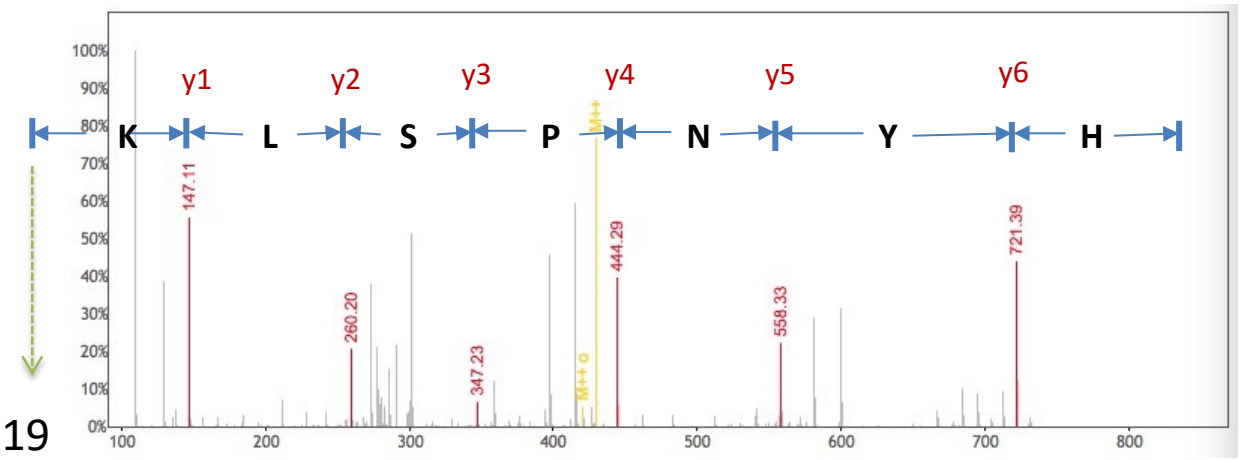
b6	**HYNPSL K**	y1
b5	**HYNPS LK**	y2
b4	**HYNP SLK**	y3
b3	**HYN PSLK**	y4
b2	**HY NPSLK**	y5
b1	**H YNPSLK**	y6



- Recall that y-ion m/z = (total of amino acid residue mass + 18.011 + z * 1.007) / z
- We use nominal mass for simplicity.

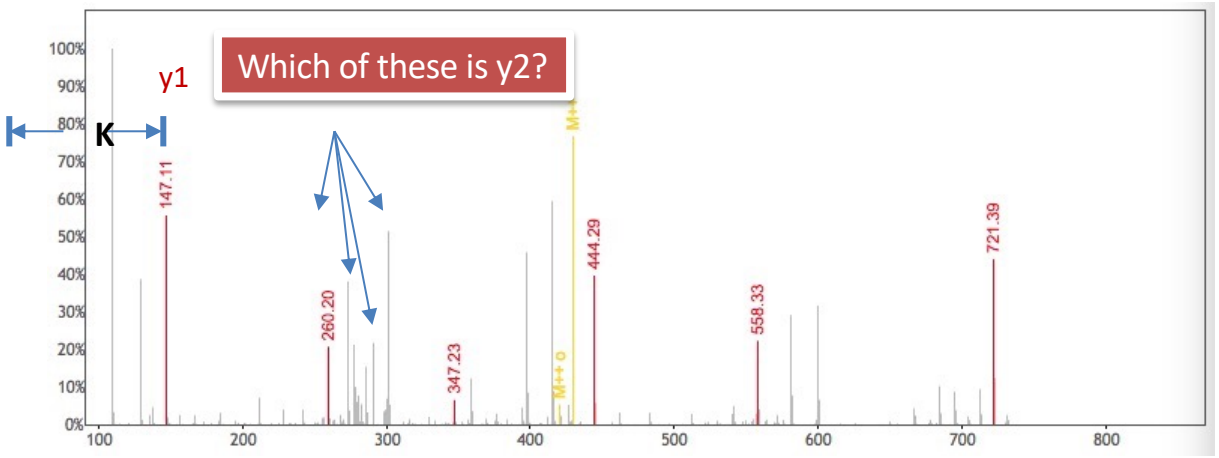
Manual De Novo Sequencing

b6	**HYNPSL K**	y1
b5	**HYNPS LK**	y2
b4	**HYNP SLK**	y3
b3	**HYN PSLK**	y4
b2	**HY NPSLK**	y5
b1	**H YNPSLK**	y6



Challenge

b6	**HYNPSL K**	y1
b5	**HYNPS LK**	y2
b4	**HYNP SLK**	y3
b3	**HYN PSLK**	y4
b2	**HY NPSLK**	y5
b1	**H YNPSLK**	y6



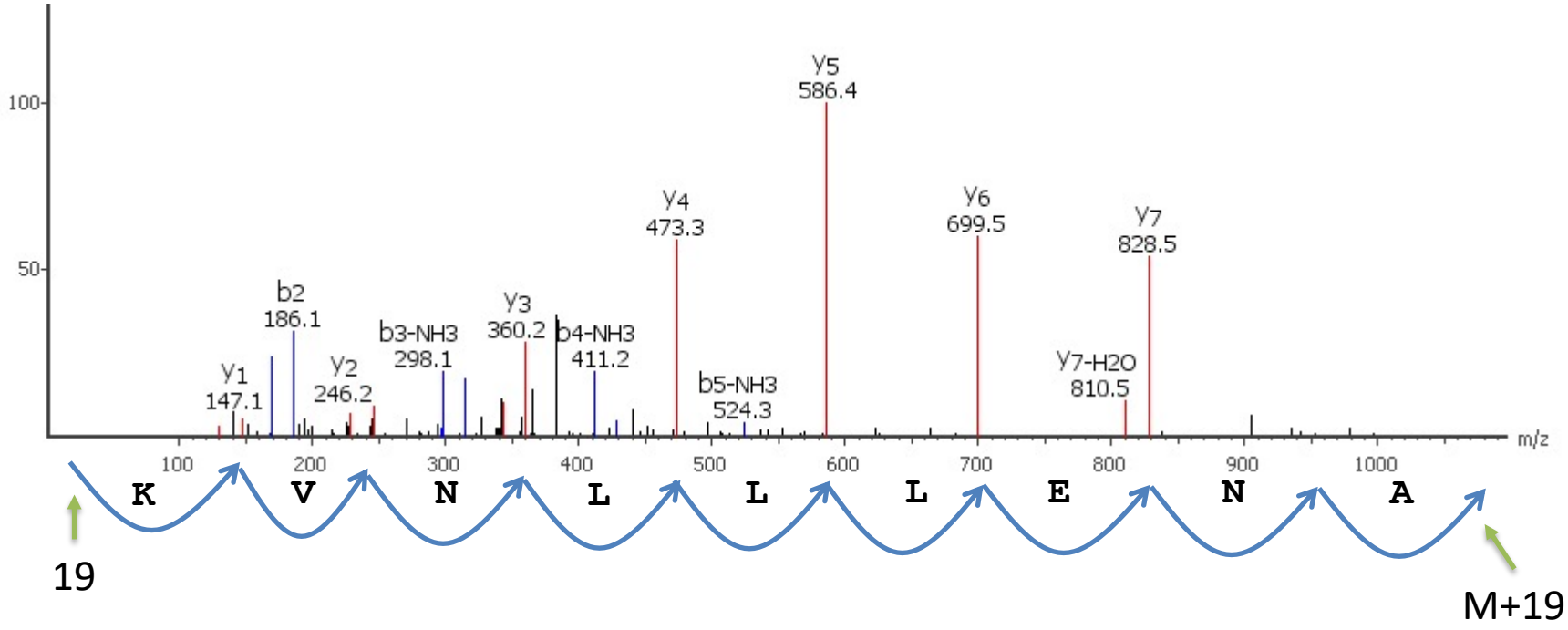
Exhaustive Search

- Exhaustively search for all combinations?
- Length-30 peptides: 20^{30}
- 1 billion peptides per sec ⇒ over 10^{22} years to search.
- Let’s develop an efficient algorithm instead.

De Novo Peptide Sequencing

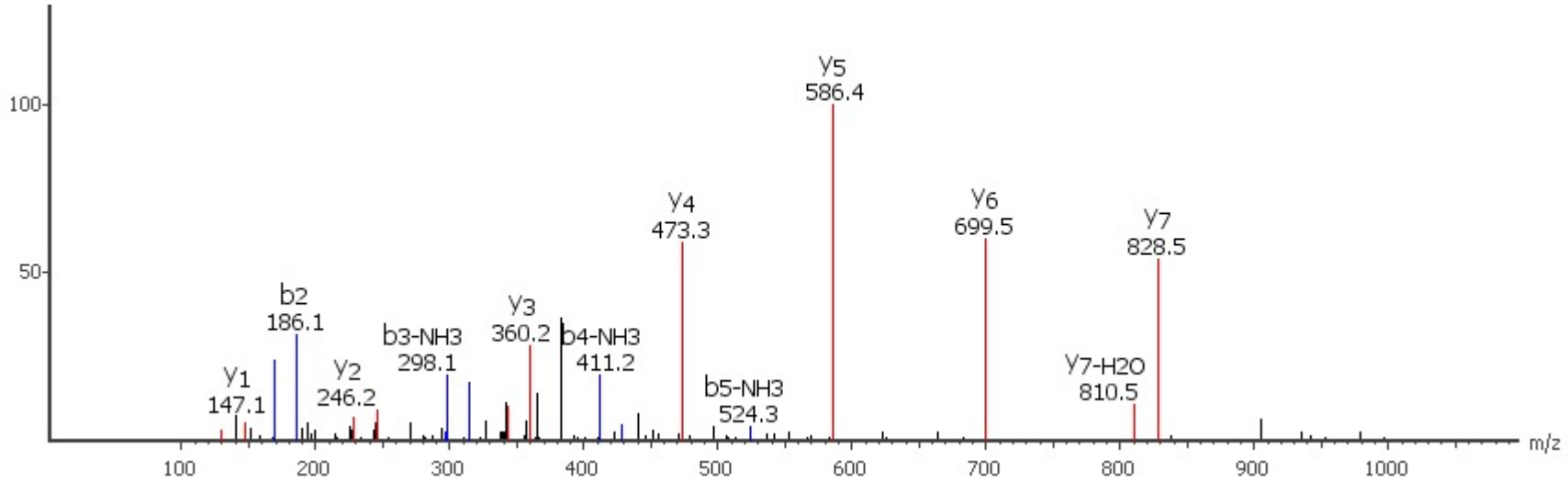
b_1	**A NELLLNVK**	y_8
b_2	**AN ELLLNVK**	y_7
b_3	**ANE LLLNVK**	y_6
b_4	**ANEL LLNVK**	y_5
b_5	**ANELL LNVK**	y_4
b_6	**ANELLL NVK**	y_3
b_7	**ANELLLN VK**	y_2
b_8	**ANELLLNV K**	y_1

- A sequence corresponds to a path that connects the y-ions.
- Note the reversed sequence in the path because of the use of y-ions.



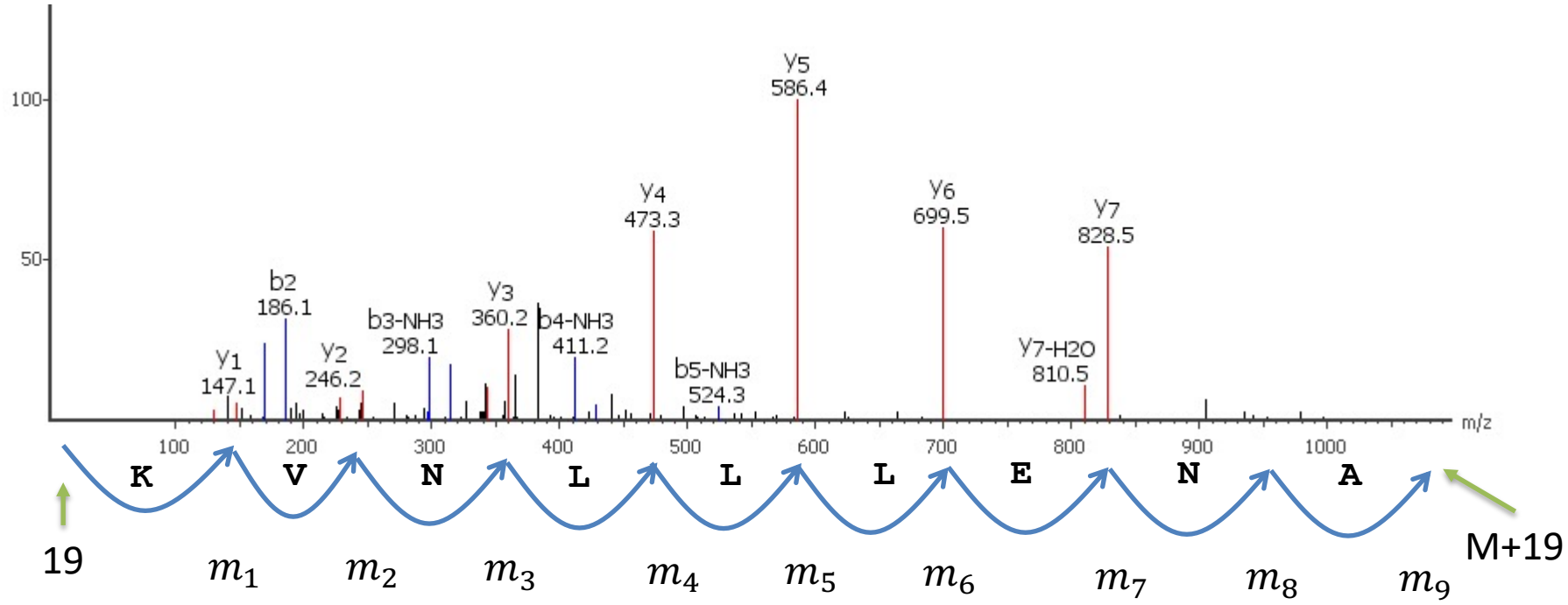
M is total residue mass

Notations



- Let $f(m)$ be the ion matching score at m/z value m.
- For example, if the log relative intensity score is used, then
- $f(m) = \begin{cases} \log_{10} 100x \text{ , if there is a peak nearby m with relative intensity } x > 0.01. \\ 0, \text{otherwise} \end{cases}$
- Note that this score can be precalculated for each m.

Path Score

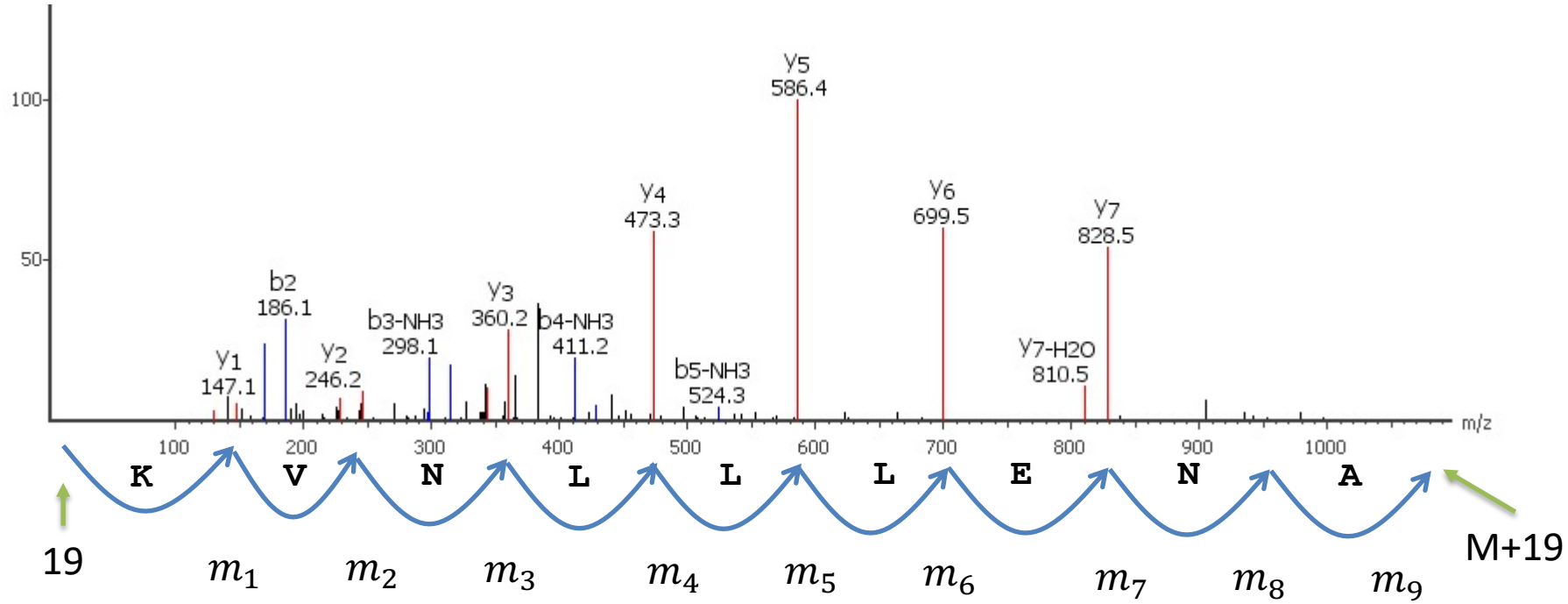


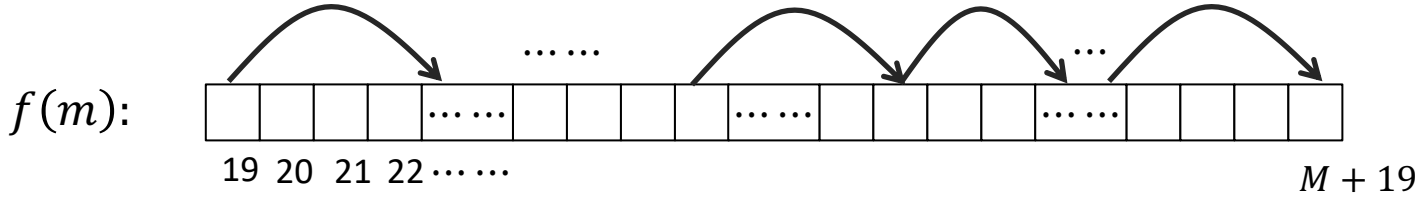
- Path score is the total of $f(m)$ for all y-ions ions
- E.g. score of path in $=f(m_1)+f(m_2)+f(m_3)+\cdots$
- **De novo sequencing:** To find a path with maximum score.

Input

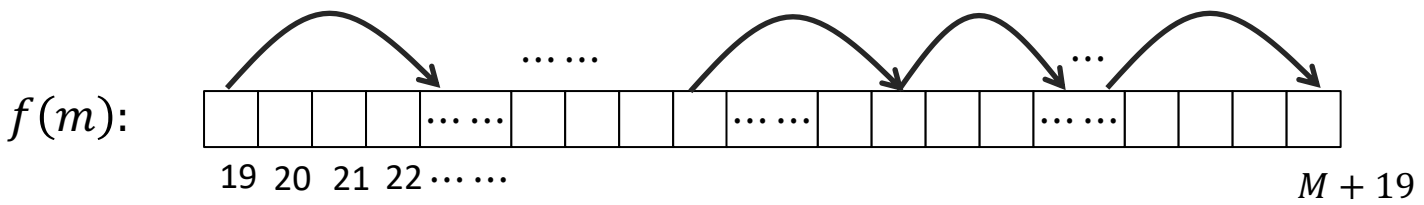
- Given spectrum, $f(m)$ can be precomputed for each m/z value m, without knowing the peptide sequence.
- Also, the total residue mass M can be computed from the precursor m/z and charge state.
- Thus, we assume we have $f(m)$ and M in our input.

Equivalent Problem



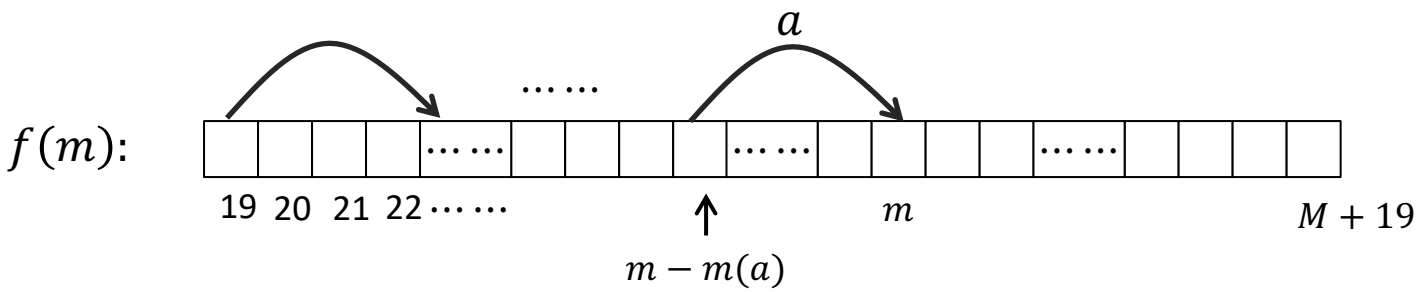


Equivalent Problem



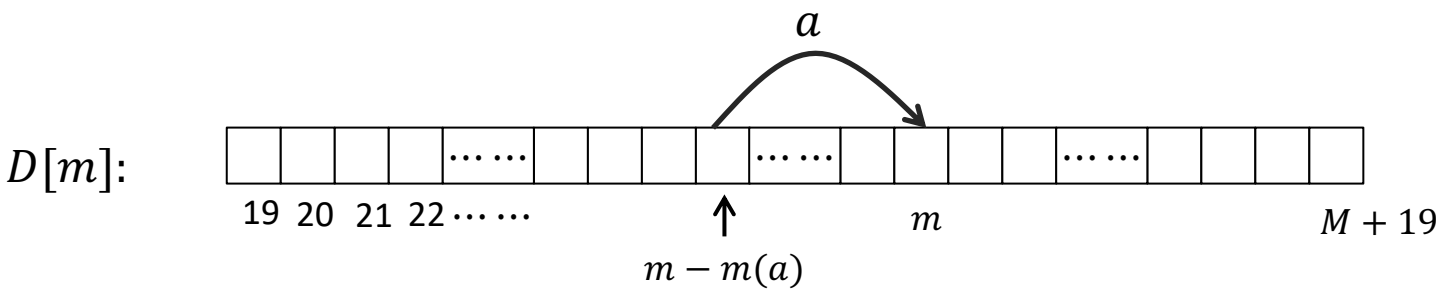
- Given an array $f(m)$ and M, to find a path from 19 to $M + 19$, such that
  - Each step length is an amino acid residue mass.
  - The total of the scores in the visited cells is maximized.

Dynamic Programming



- Let $D[m]$ be the maximum score a path from 19 to m can achieve.
- If the path is not empty, assume a is the last amino acid, then $D[m] = D[m - m(a)] + f(m)$.
- Thus, $D[m] = f(m) + \max_a D[m - m(a)]$.

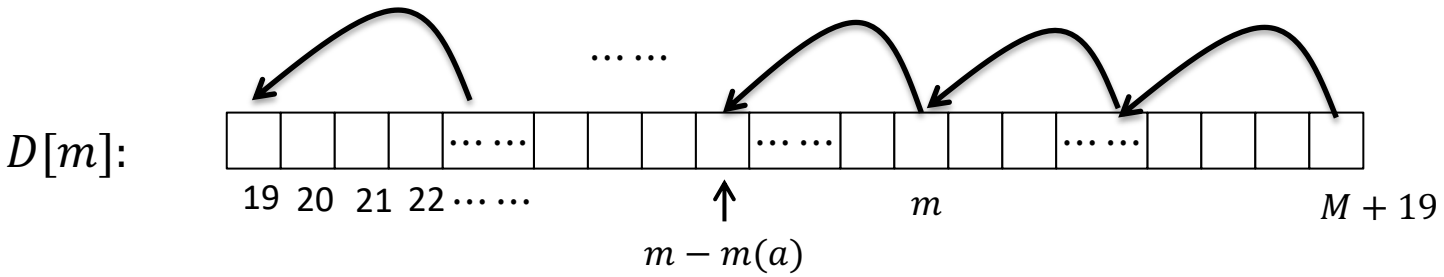
Dynamic Programming



- Initializes $D[19] = 0$ and all other cells to be $-\infty$.
- For m from 20 to M+19
- $$D[m] = f(m) + \max_{a} D[m - m(a)]$$

Time complexity: $O(M)$

Backtracking



- The best sequence can be retrieved by a backtracking process by repetitively computing the last amino acid a that maximizes the recurrence relation.

$$D[m] = f(m) + \max_{a} D[m - m(a)]$$

- Time complexity: O(length of peptide).

Practical Concerns

- As usual, the basic algorithm looks simple. But the reality is more difficult.

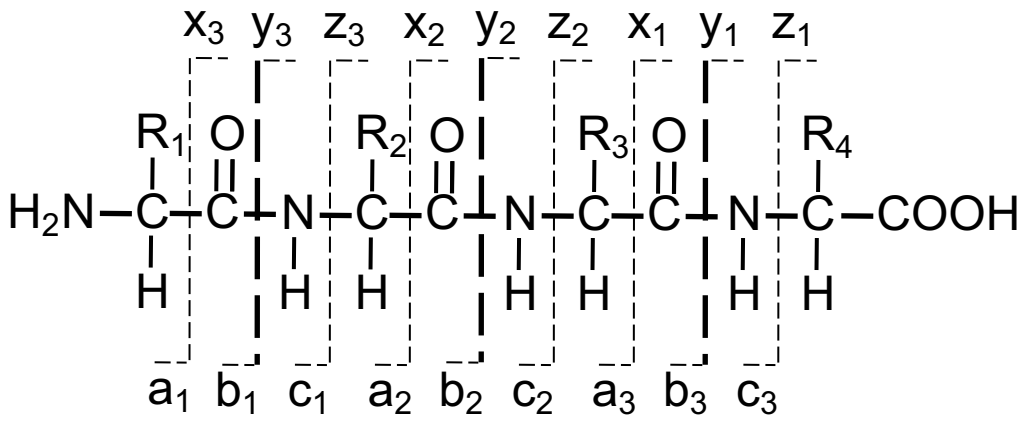
Dealing with High Resolution Data

- In high res data, nominal mass is not good enough. Error up to $\pm$0.5Da.
- We can multiple each mass (including both amino acid mass and peak m/z) with 1000 and round to integer.
- E.g. 123.4567 => 123458
- The rounding error is then limited to $\pm$ 0.0005Da.

PTM and De Novo Sequencing

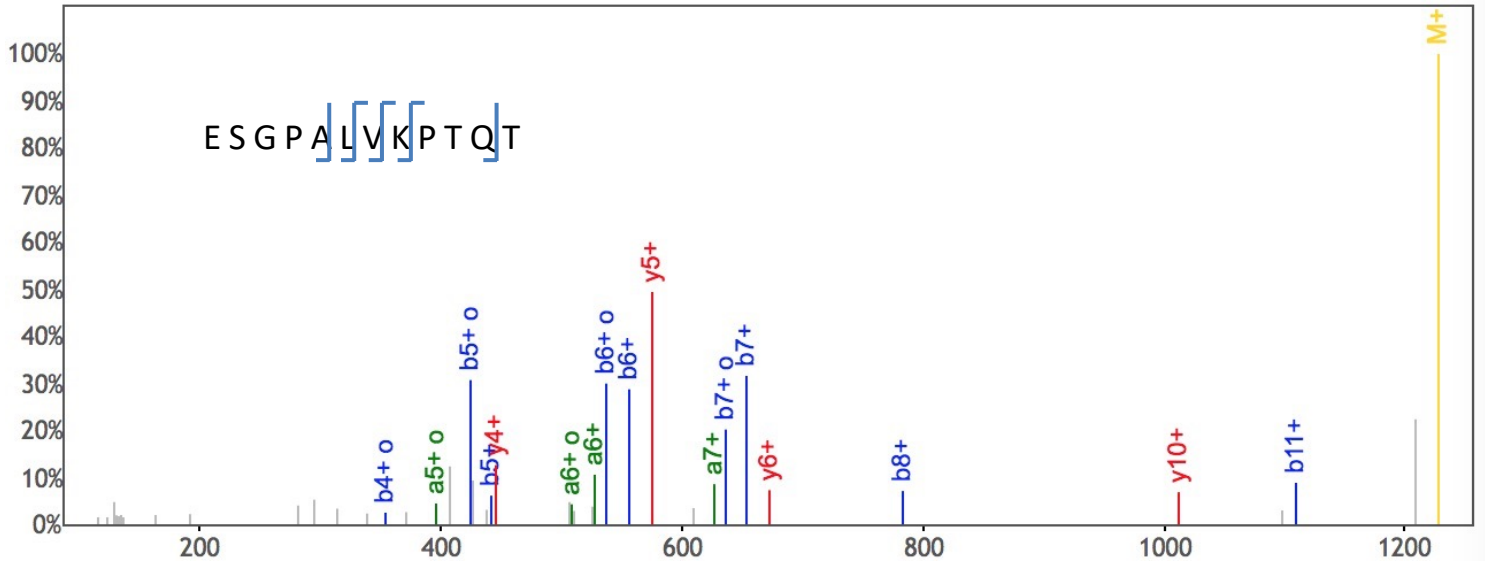
- Variable PTM does not cause major speed slow down for *de novo* sequencing algorithms.
  - Instead of trying 20 regular amino acids in the maximization, the algorithm simply tries all modified amino acids too.
  - The time complexity is increased by a constant factor. (Compare to the exponential growth in database search approach).
- However, since the solution space is larger when many variable PTMs are allowed, the accuracy of the algorithm is reduced.

Other Fragment Ions



- Between two adjacent residues, there are 3 fragmentation possibilities, causing 6 fragment ion types.
- Each ion type has a mass offset
  - a: -27, b: +1, c: +18, x: +45, y: +19, z: +2
- b and y ions are complementary.
  - Charge one b + y = total residue mass +20.
- y ion usually the most abundant.
- Also neutral loss ions such as y-H_2O and b-NH3

De Novo Peptide Sequencing Accuracy

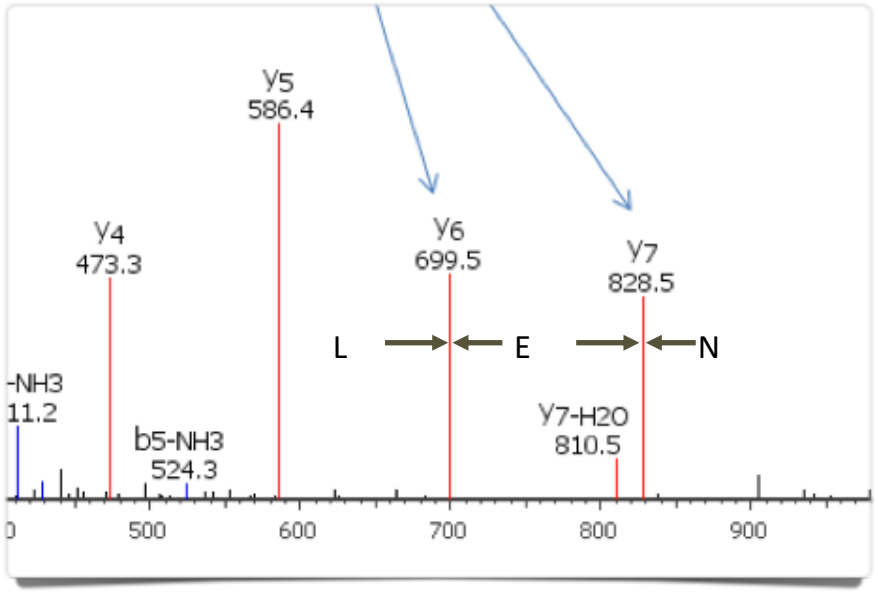


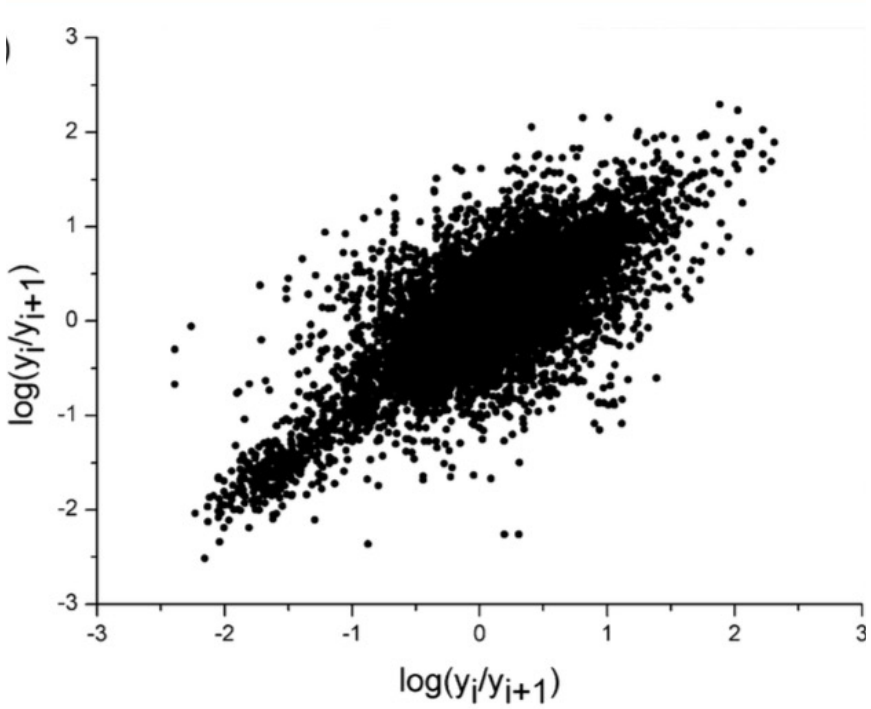
- For below average spectrum, as much as 50% error rate!
- Mostly mass gap error.
- Error source:
  - spectrum quality
  - Inaccurate scoring function

Solutions

- Make use of partially correct de novo sequence tags.
- Improve the scoring function.
- Next let us examine one such effort, the Novor software.

Amino Acid Combination Affects the Peak Intensity

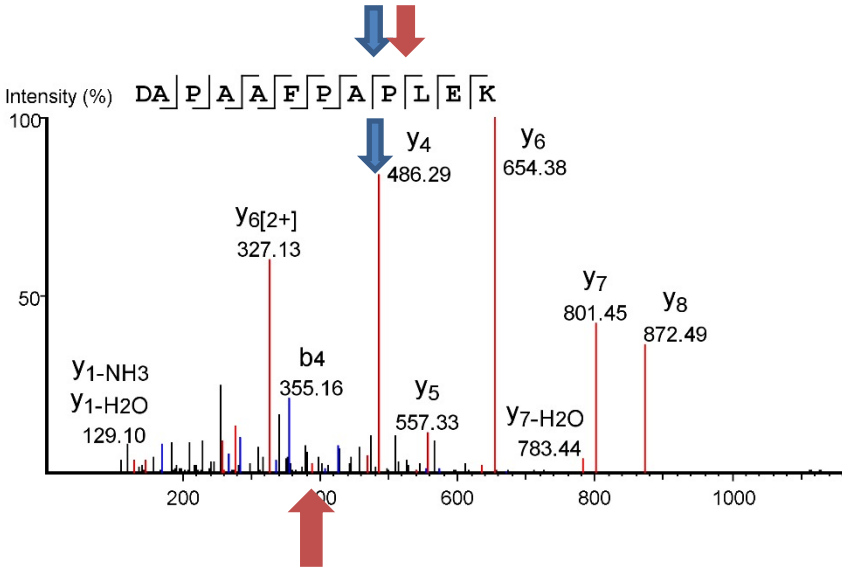




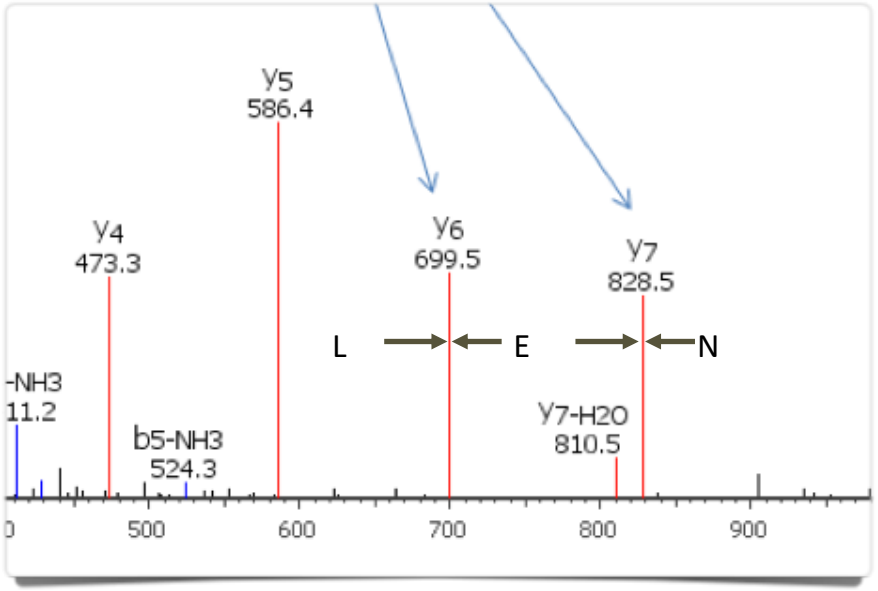
- The neighbouring 3 amino acids approximately determine the peak intensity.

An Example: Proline (P)

- Most software's scoring function prefers more abundant peaks.
- Pro enhances fragmentation at left, and reduces at right.

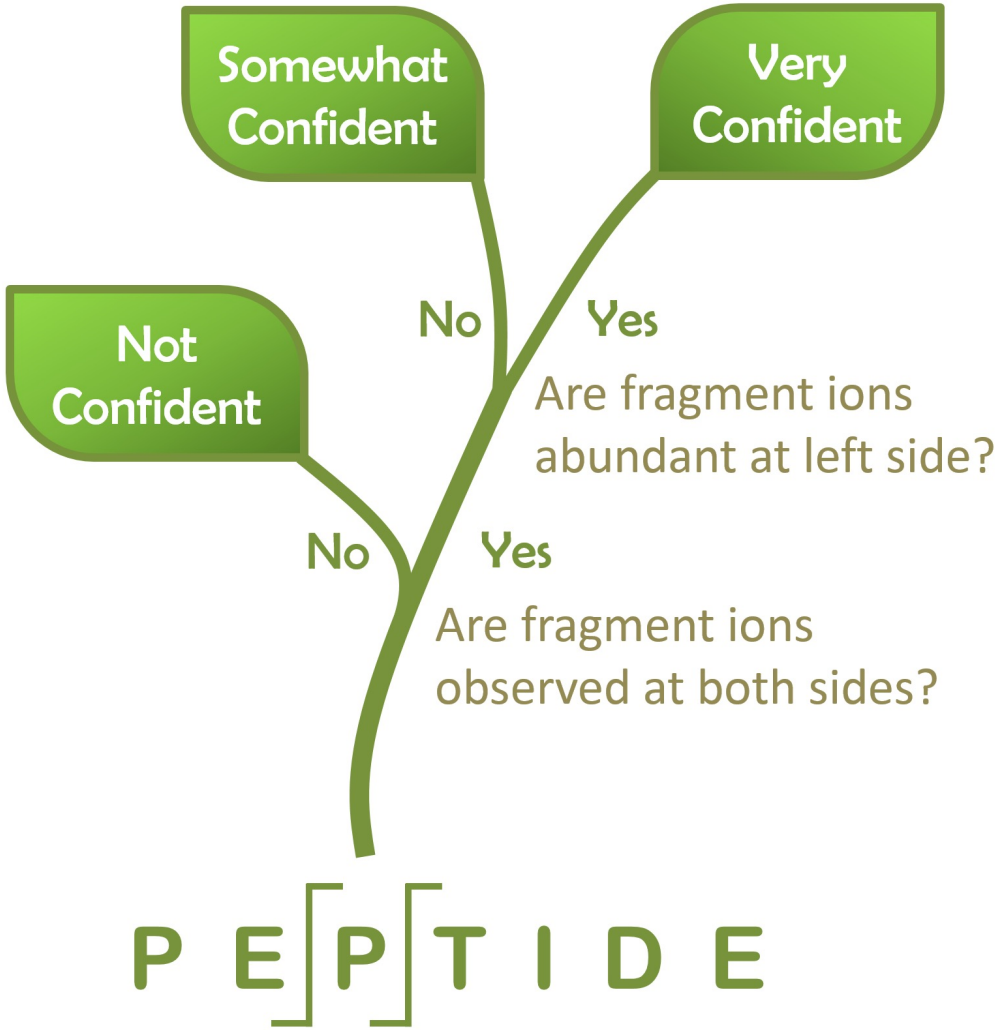


Scoring Features



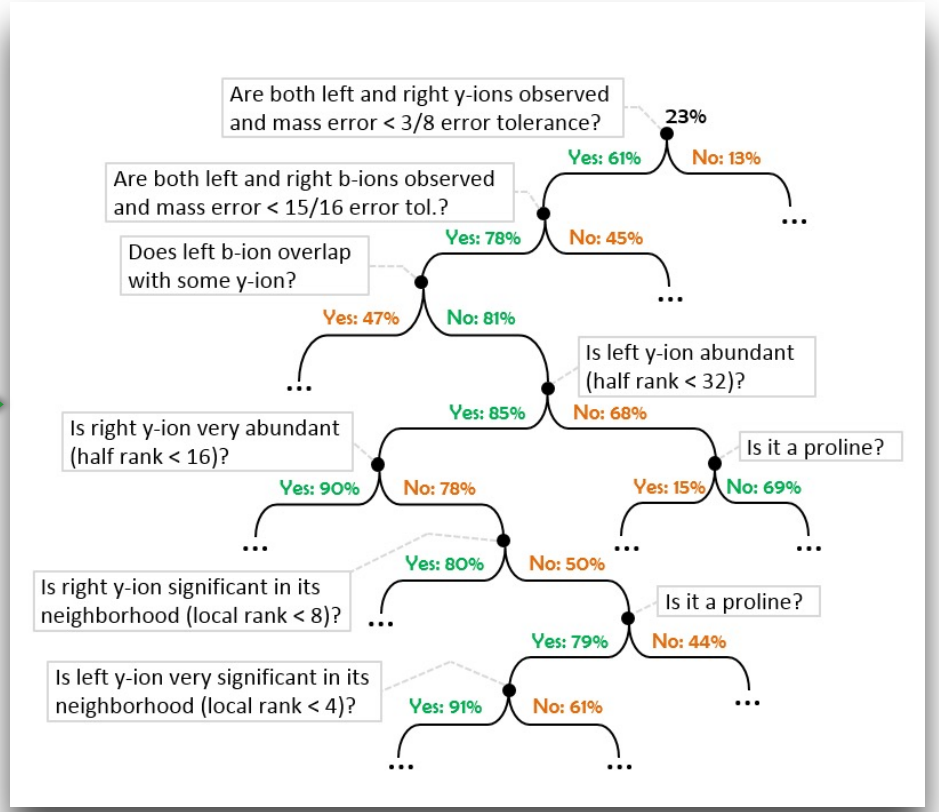
- Pr(E is correct) predicted by features such as
- mass error
- intensity of y6 and y7
- intensity ratio y6/y7
- L, E, and N
- and many others

Decision Tree



Decision Tree Learning

NIST Spectrum Library
340,000 spectra



169 features
14,000 internal nodes
average depth 18.4

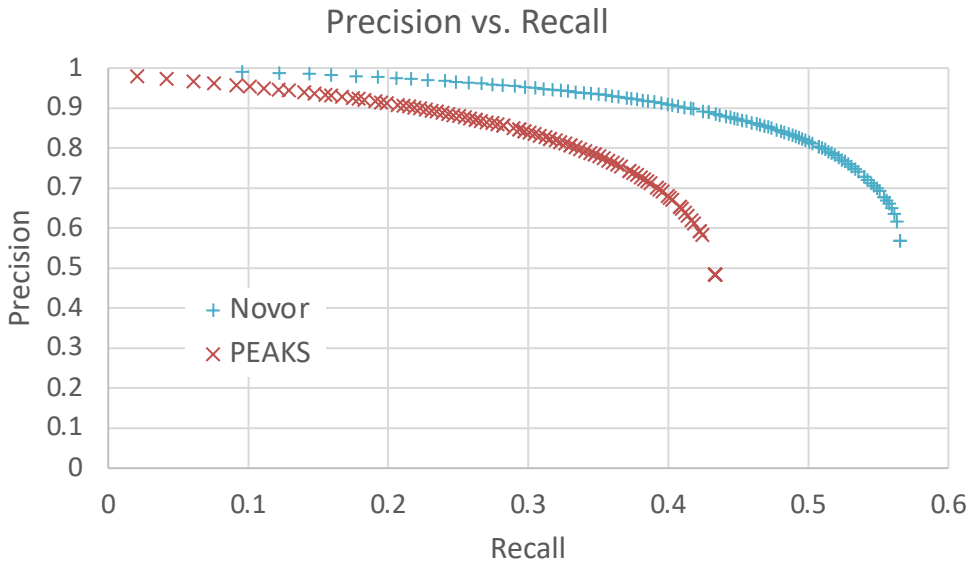
Benefits of Decision Tree

- Allows to use of a large number of scoring features
  - Mass error, sequence pattern, all ion types, intensity, etc.
  - 169 features
- Learn a large number of rules
  - 14,000 branching nodes
- Each evaluation is fast.
  - Path from root to a leaf average length = 18
  - Only most important features are examined according to situation.

Algorithm

- A peptide’s score is the sum of amino acid confidence score.
- Algorithm computes a peptide to maximize this score.

Novor vs. PEAKS (Accuracy)

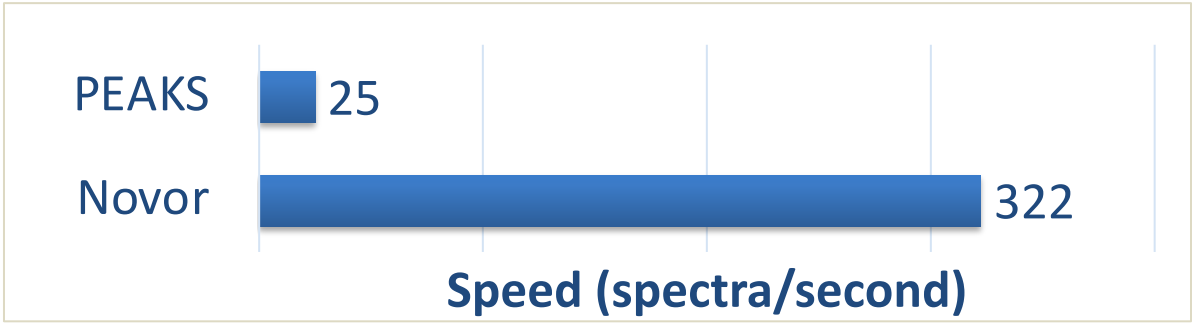


$$\text{Precision} = \frac{\#\text{ Correct AA above threshold}}{\#\text{ AA above threshold}}$$

$$\text{Recall} = \frac{\#\text{ Correct AA above threshold}}{\text{total }\#\text{ AA}}$$

AA: Amino Acid

Novor vs. PEAKS (speed)

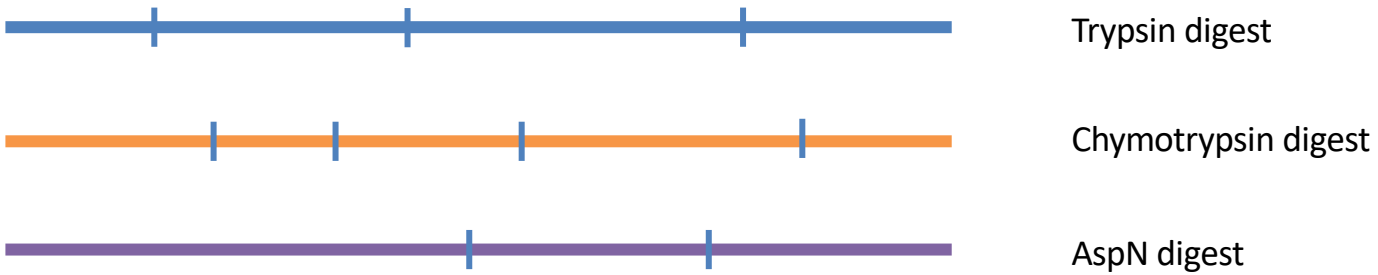


Speed on a laptop PC

- Speed is an order of magnitude faster.
- First and only real-time de novo sequencing software.

Ma, B. (2015). Novor: Real-Time Peptide de Novo Sequencing Software. J. ASMS, 26, 1885–1894.

De Novo Protein Sequencing Basic Idea



1. Digest protein with different enzymes.
2. De novo sequence each peptide.
3. Assemble overlapping peptides to derive the protein sequence.

Automated De Novo Sequencing

- Many de novo sequencing programs
  - Sherenga (1999)
  - Lutefisk (2001)
  - PEAKS (2003)
  - PepNovo (2005)
  - Novor (2015)
  - DeepNovo (2017)