Algorithms in Bioinformatics: A Practical Introduction

Peptide Sequencing

What is Peptide Sequencing?

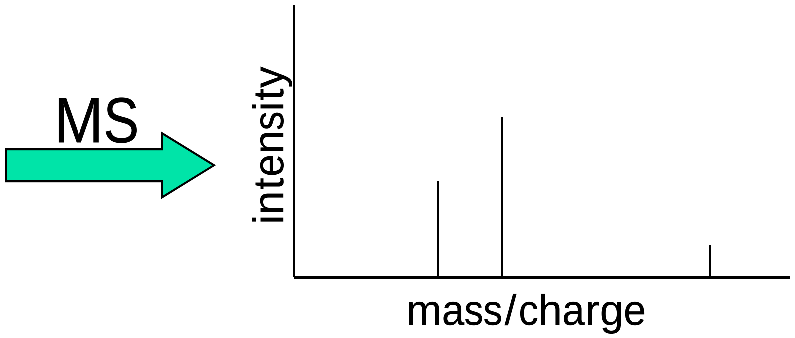
- High-throughput Protein Sequencing is to deduce the amino acid sequence of a protein. It is still very difficult.

- Currently, research focus on Peptide Sequencing, that is, getting the amino acid sequence of a short fragment of a protein (of length ≈ 10).

Enabling technology: Mass Spectrometry

- Idea for deducing the peptide sequence: Mass!
- Mass Spectrometry is a machine which can separate and measure samples with different mass/charge ratio.
- Example:

Sample 1: m/z=100Da, 10mol
Sample 2: m/z=50Da, 50mol
Sample 3: m/z=33Da, 30mol



Dalton(Da) is a mass unit. E.g. H is of mass 1Da

History

- Peptide sequencing is discovered by Pehr Edman (1949) and Frederick Sanger (1955).
- In 1966, Biemann et al successfully sequenced a peptide using a mass spectrometer machine.
- During 1980s, sequencing using mass spectrometry becomes popular.

Agenda

- Biological Background
- De Novo Peptide Sequencing
  - PEAK
  - Spectrum graph
- Protein Database Searching Problem
  - SEQUEST

Amino acid residue mass

A	71.08		M	131.19
C	103.14		N	114.1
D	115.09		P	97.12
E	129.12		Q	128.13
F	147.18		R	156.19
G	57.05		S	87.08
H	137.14		T	101.1
I	113.16		V	99.13
K	128.17		W	186.21
L	113.16		Y	163.18

- Amino acid residue = amino acid losing a water
- I and L have the same mass
- Smallest mass is G (57.05 Da)
- Largest mass is W (186.21 Da)

Mass Spectrometry can separate different peptides

- Previous table shows that most of the amino acids have different masses.
- Hence, with high chance, different peptides have different masses.

- The mass given by a mass spectrometer has a maximum error ±0.5Da. It can separate most of the peptides.

Protein identification process (LC/MS/MS)

Input: a protein sample

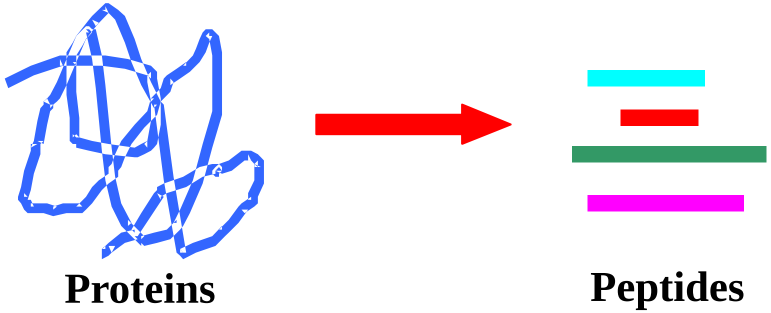
A. Biology part:
   1. Digest the protein into a set of peptides
   2. By HPLC+Mass Spectrometer, separate the peptides.
   3. Select a particular peptide
   4. Fragment the selected peptide
   5. Get the tandem mass (MS/MS) spectrum of the selected peptide

B. Computing part:
   - De Novo Sequencing
   - Protein Database Search

Digest a protein into peptides

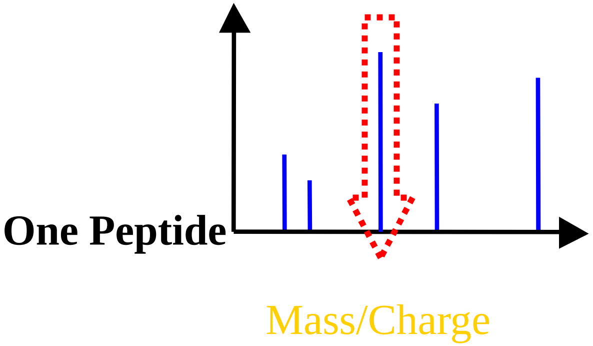
- By an enzyme, digest a protein into short peptides.
- If we digest a protein using trypsin,
  - it digests the protein at K or R that are not followed by P.
  - After digestion, we will get a set of peptides end with K or R!

- E.g. ACCHCKCCVRPPCRCA → ACCHCK, CCVRPPCR



Selecting a particular peptide

- HPLC stands for High Performance Liquid Chromatograph. It can separate a set of peptides in a high pressure liquid chromatography
- After HPLC, the mixture of peptides are analyzed by MS.
  - Then, we get the MS spectrum



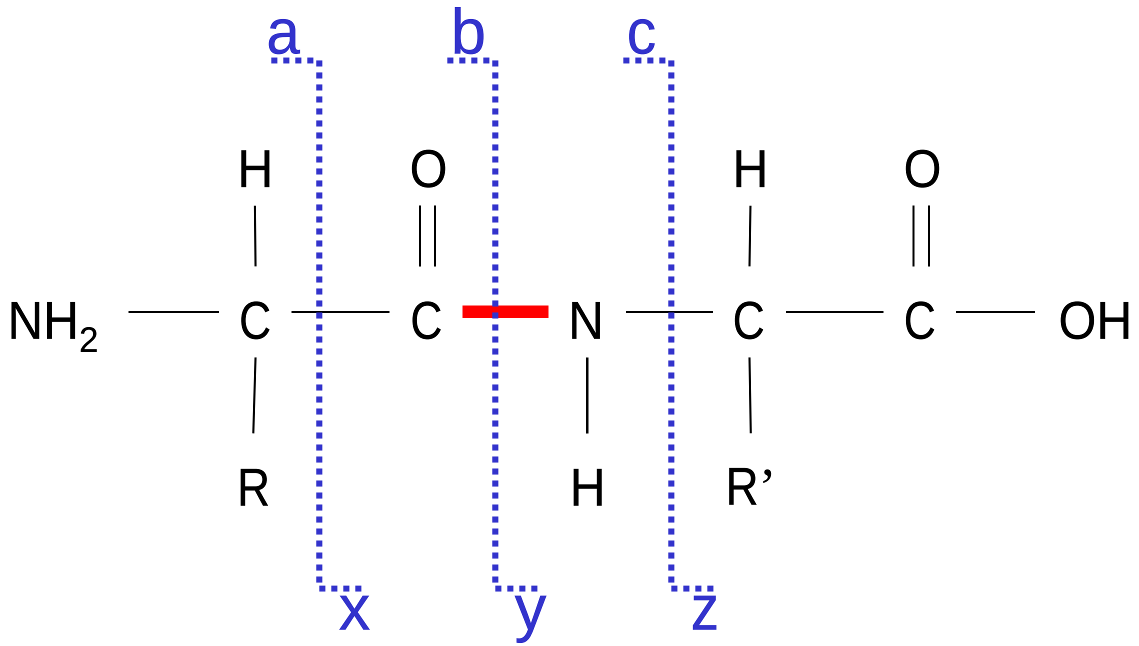
- The peptide of a particular mass is selected.

Fragmentation of peptide (I)

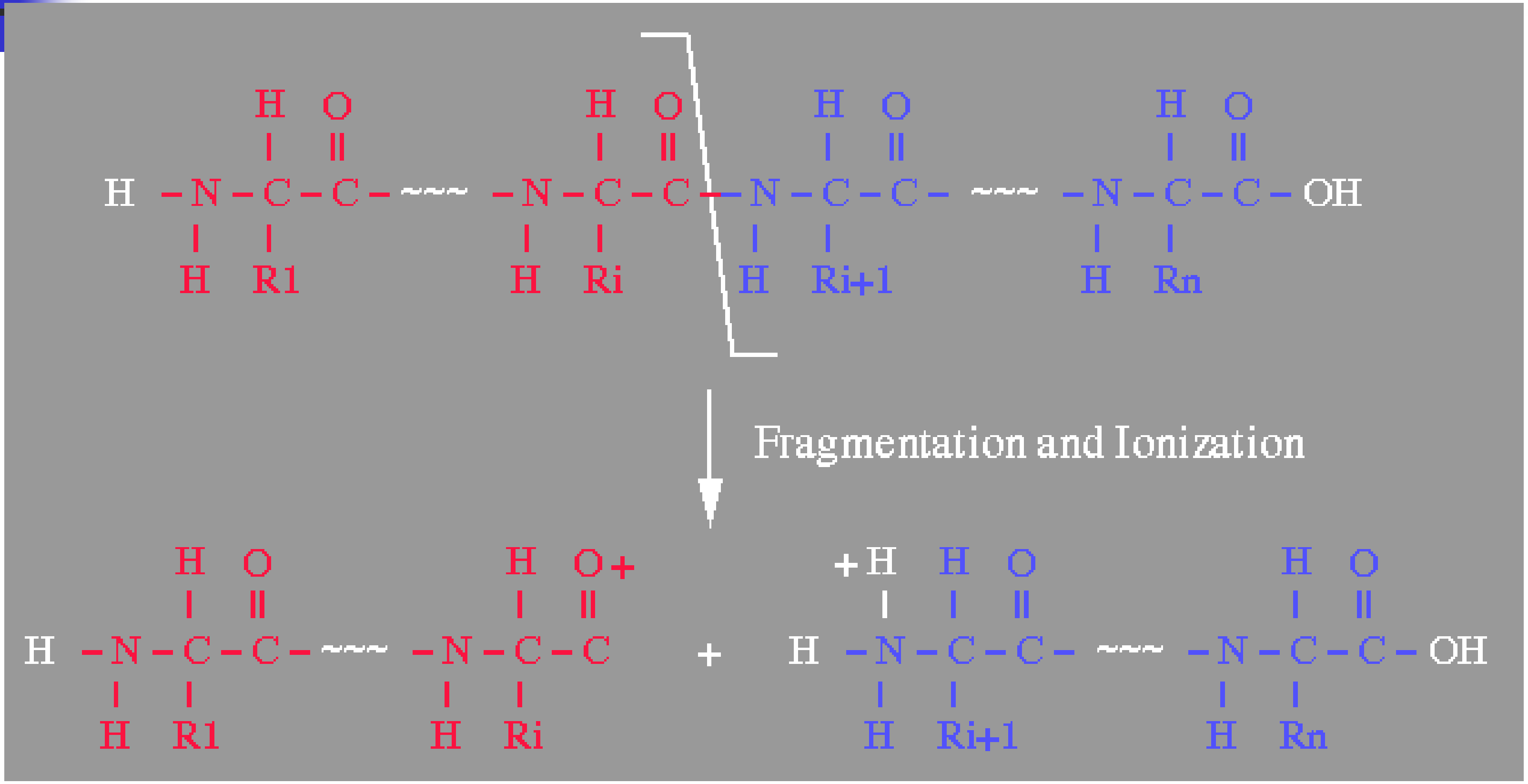
- Fragmentation tries to break the selected peptide at all positions in the peptide backbond.
- Usually, fragmentation is by Collision Induced Dissociation (CID).
  - The peptide is passed into the collision cell (which has been pressurized with argon [inert gas]).
  - Collision between peptide and argon break the peptide.
- Each peptide is usually fragmented into 2 pieces.
  - prefix fragment and suffix fragment (either one fragment will be charged but not both)

Fragmentation of peptide (II)

- Most often, the peptide is broken at C-C, C-N, N-C bonds.
  - Resulting a-ions, b-ions, c-ions, x-ions, y-ions, and z-ions.
  - Based on experiment,
    - The intensity of y-ions > that of b-ions
    - The intensities of other ions are even smaller



Fragmentation of peptide (III)


H O
H -N-C-C-~~~ -N-C-C-N-C-C- ~~~ -N-C-C-OH
H R1
H Ri
H Ri+1
H Rn
Fragmentation and Ionization
H O+
+H
+
R1
Ri
Ri+1
Rn

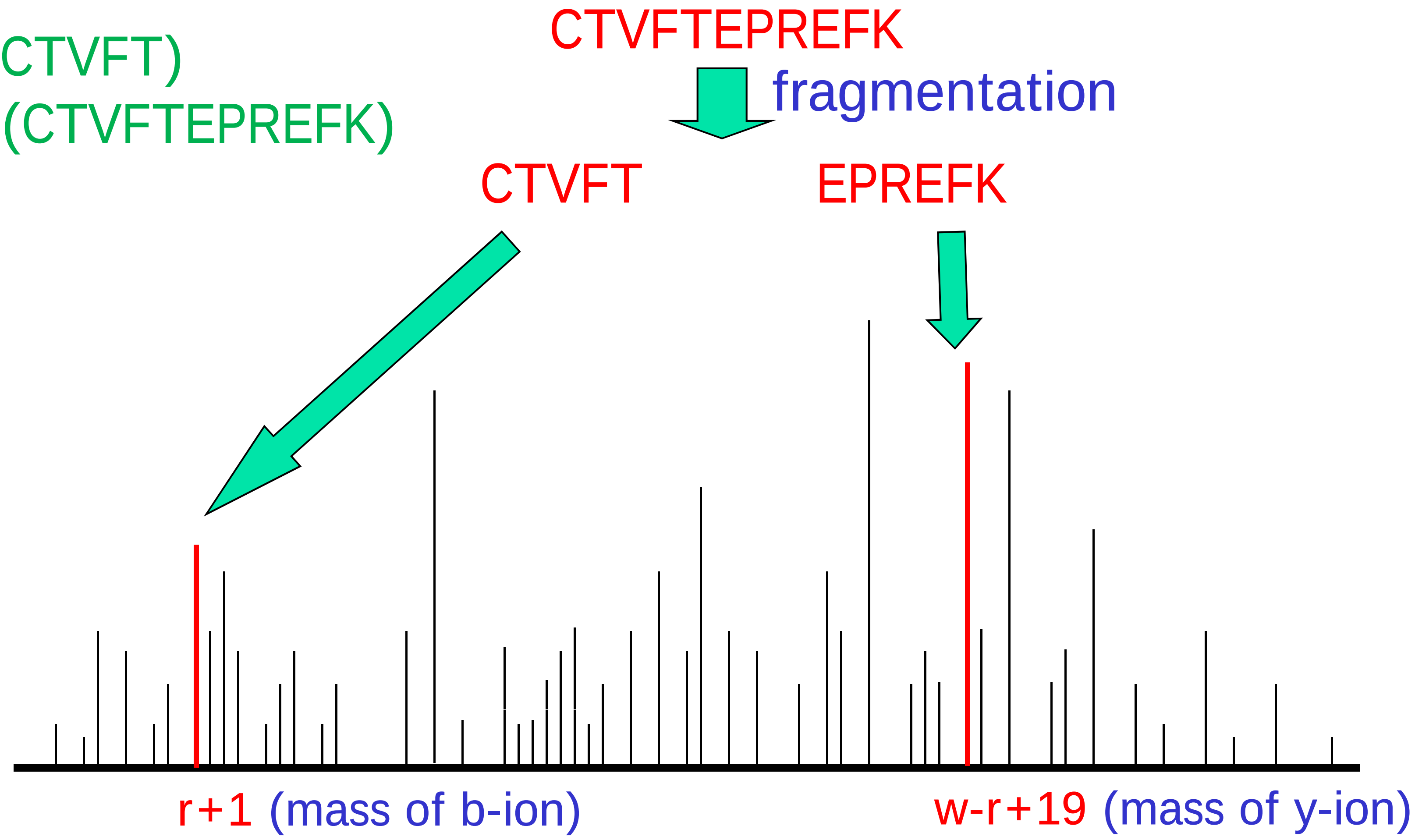
B-ion Y-ion

Complementary: Mass(B-ion)+Mass(Y-ion) = Mass(peptide)+4H+O

Fragmentation of peptide (IV)

r = w(CTVFT)

w = w(CTVFTEPREFK)



CTVFTEPREFK

fragmentation

CTVFT

EPREFK

r+1 (mass of b-ion)

w-r+19 (mass of y-ion)

Mass of the ions (I)

- Let A be the set of amino acid. For every $a \in A$, $w(a)$ = mass of its residue
- Let $P=a_1a_2...a_k$ be a peptide.
  - $w(P) = \Sigma_{1 \le j \le k} w(a_j)$.
- Actual mass of the peptide with sequence P is
  - $w(P)+18$ (since it has an extra H_2O)
- Mass of b-ion of the first i amino acids is
  - $b_i = 1 + w(a_1a_2...a_i)$
- Mass of y-ion of the last i amino acids is
  - $y_i = 19 + w(a_i...a_k)$
- Note: $b_i + y_{i+1} = 20 + w(P)$

Mass of the ions (II)

- E.g. P=SAG
  - w(P) = w(S)+w(A)+w(G) = 215.21
  - Actual mass of P = w(P)+18 = 233.21
  - y_1 = w(SAG)+19 = 234.21
  - y_2 = w(AG)+19 = 147.13
  - y_3 = w(G)+19 = 76.05
  - b_1 = w(S)+1 = 88.08
  - b_2 = w(SA)
  - b_3 = w(SAG)+1 = 216.21

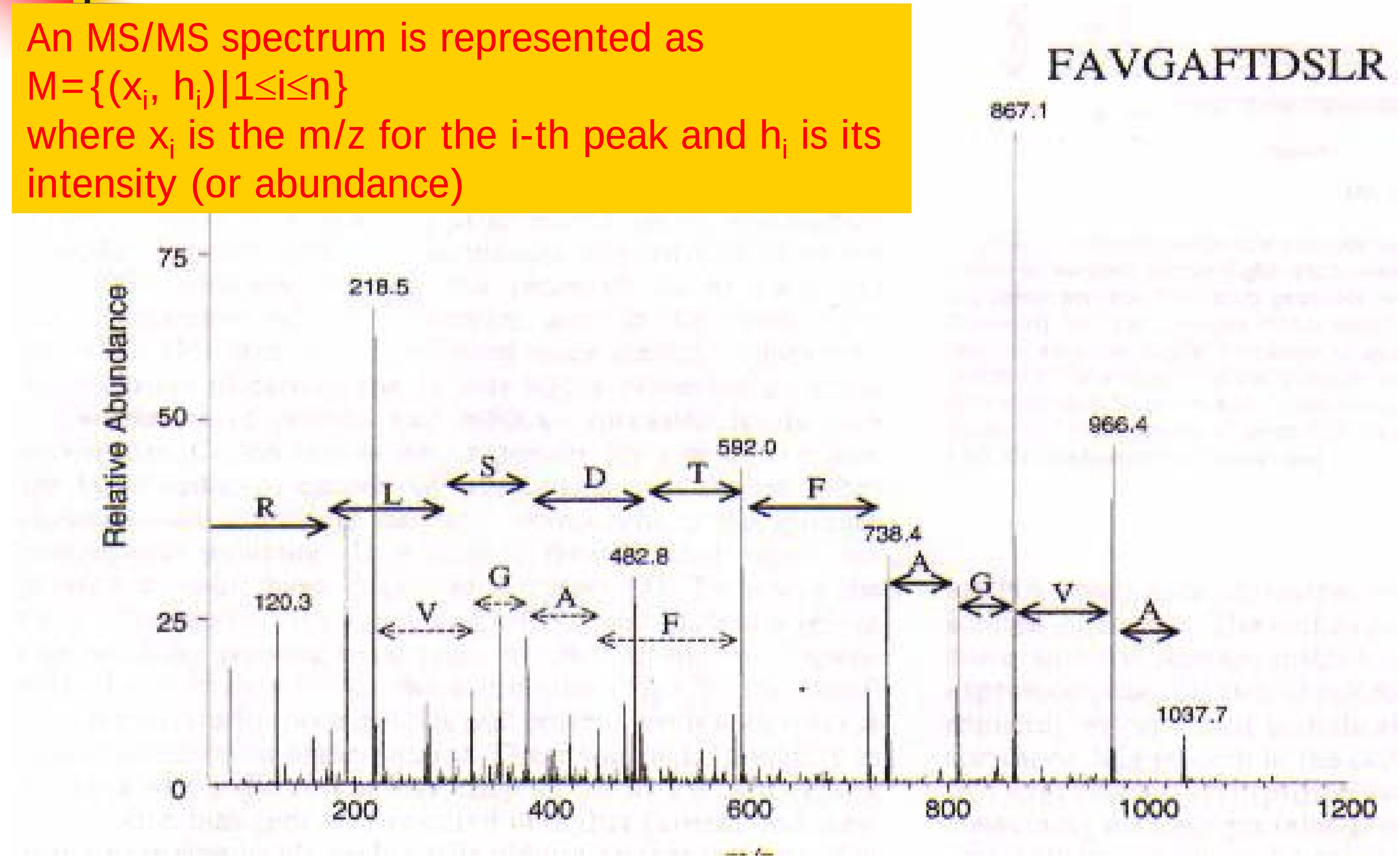
Other ion types

- Apart from a-ion, b-ion, c-ion, x-ion, y-ion, and z-ion, we also have variations with additional loss of
  - a water molecule
  - an ammonia molecule
  - a water and an ammonia molecule
  - Two water molecules

- E.g. y-H_2O, y-NH_3, y-H_2O-H_2O, y-H_2O-NH_3

Tandem Mass Spectrum (MS/MS Spectrum)

An MS/MS spectrum is represented as
$M=\{(x_i, h_i) | 1 \leq i \leq n\}$
where x_i is the m/z for the i-th peak and h_i is its intensity (or abundance)



Computational problems

- There are three computational problems:
  1. De novo peptide sequencing
  2. Peptide Identification
  3. Identification of PTM (Post-translational modification)

- We will discuss problems 1 and 2.

De Novo Peptide Sequencing Problem

- Input:
  - A MS/MS spectrum M; and
  - the total mass wt of the peptide
  - An error bound δ (default δ=0.5)
- Output:
  - The peptide sequence

Assumption of the spectrum

- We assume all the ions are singly charged.

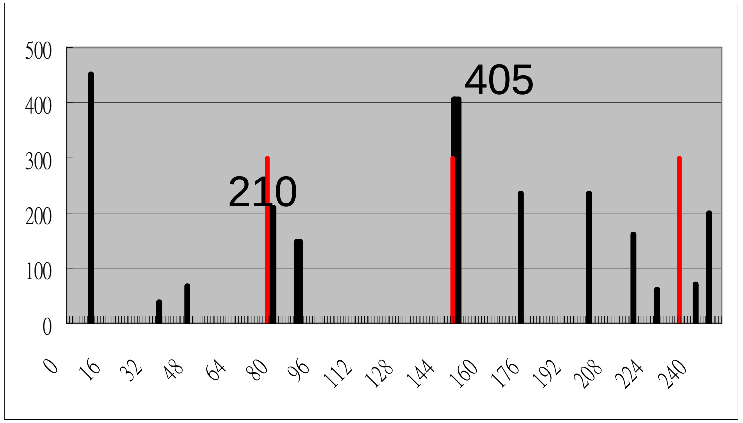
- In fact, in a MS/MS experiment,
  - an ion can be charged with different charges.
- Fortunately,
  - if a spectrum has peaks corresponding to multiply charged ions, there exists standard method to convert those peaks to their singly charged equivalents.

Simple scoring scheme

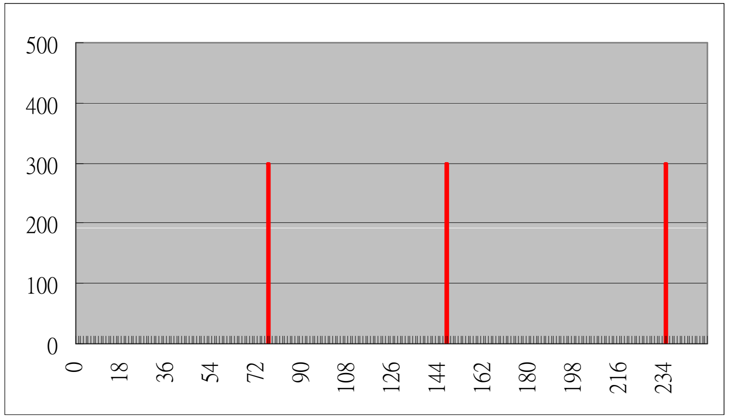
- Consider a peptide $P=a_1a_2...a_k$
  - Recall that y-ions are expected to have the highest intensities.
  - If M is a spectrum for P, we can find peaks for m/z = y_i for i=1,2,...,k
- So, we define the score function score(M,P) = $\Sigma\{h|(x,h)\in M, |x-y_i|\leq\delta$ for $i=1,2,...,k\}$

Simple scoring scheme example

- E.g. P=SAG
  - $y_1 = 57.05+71.08+87.08+19 = 234.21$
  - $y_2 = 57.05+71.08+19 = 147.13$
  - $y_3 = 57.05+19 = 76.05$
- Score(M,P) = 210+405 = 615



Black peaks: real peaks



Red peaks: artificial y-ions

Refined problem

- Input:
  - A MS/MS spectrum M
  - The total mass wt of the peptide
  - An error bound δ
- Output:
  - A peptide P such that $wt-\delta \le w(P) \le wt+\delta$ which maximizes score(M,P).

Brute-force solution

- For every possible peptide P such that $|w(P)-wt| \leq \delta$,
  - Compute score(M,P)
- Report the peptide P such that $|w(P)-wt| \leq \delta$ which maximizes score(M,P)!

- Exponential time! Very slow!

- Can we solve the problem faster?
  - Yes! By dynamic programming.

Idea of the dynamic programming

- Try to identify the residues one by one from right to left.
- Let $f_M(r) = \Sigma\{ h \mid (x,h) \in M \text{ and } |x-r| \leq \delta \}$.
  - $f_M(r)$ is the sum of all peaks in M whose mass is close to r.
- Observation:
  - $\text{score}(M,a_1a_2\ldots a_k) = \text{score}(M,a_1a_2\ldots a_{k-1}) + f_M(w(a_1a_2\ldots a_k)+19)$

Simple dynamic programming solution

- Let V(r) be the maximum score(M,P) among all possible P such that w(P)=r.
- Our aim is to find $\max_{|r\text{-}wt|\leq\delta} V(r)$. Then, by back-tracking, we can recover the peptide.

- We have
  - $V(0)=0$.
  - $V(r) = \max_{a\in A} \{ V(r-w(a)) + f_M(r+19) \}$.

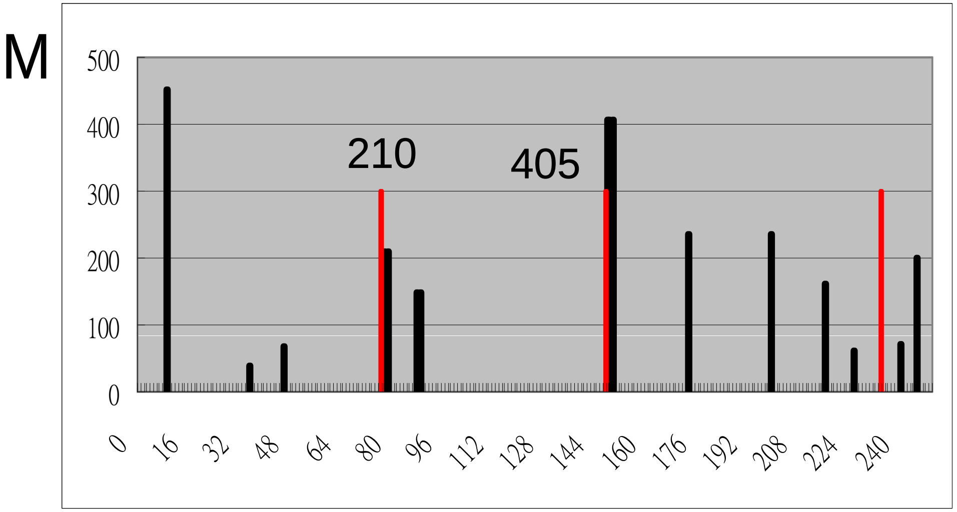
Example

- Recall $V(0)=0.$

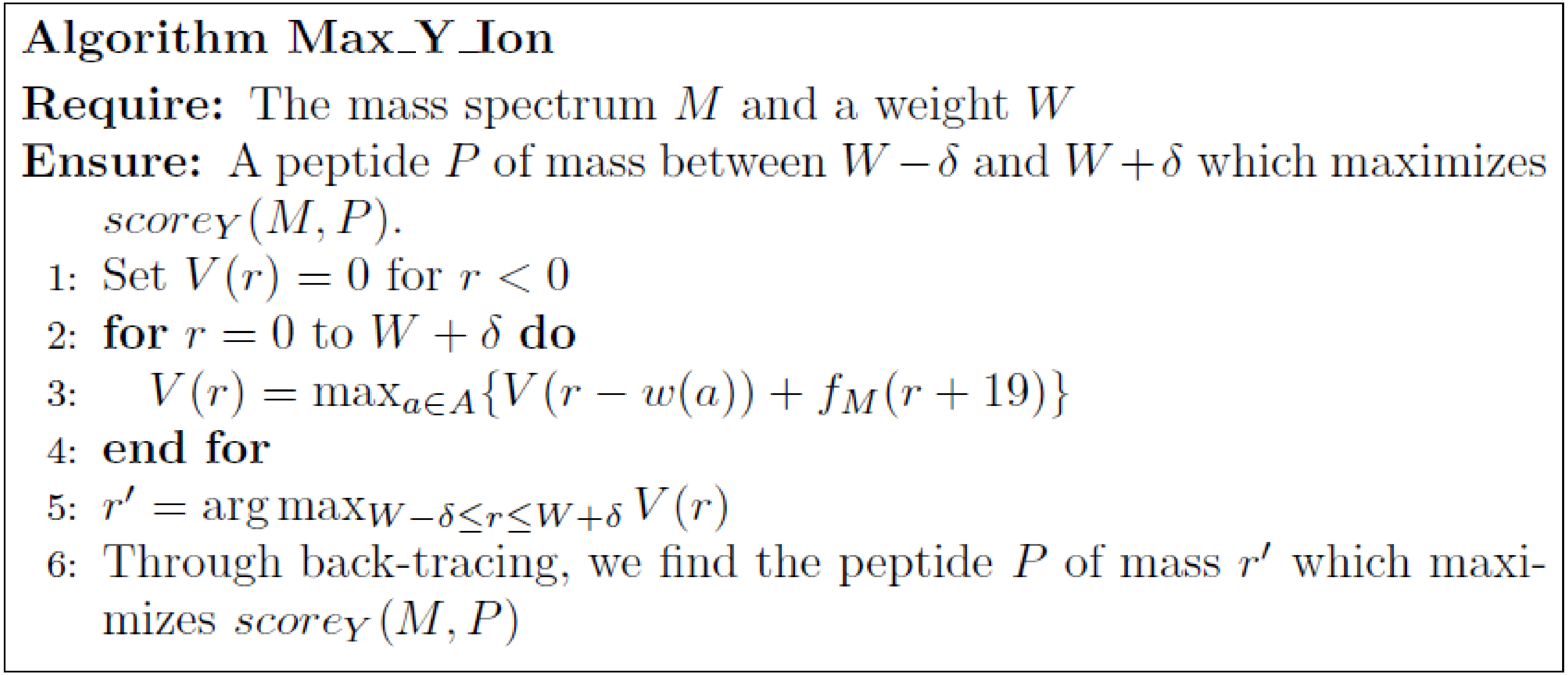
  $V(r) = \max_{a\in A} \{ V(r-w(a)) + f_M(r+19) \}.$

- E.g.

$$V(147.13) = \max\begin{cases} V(76.05)+450\ (due\,to\,A) \\ V(43.99)+450\ (due\,to\,C) \\ \ldots \end{cases}$$



Algorithm



Algorithm Max_Y_Ion

Require: The mass spectrum M and a weight W

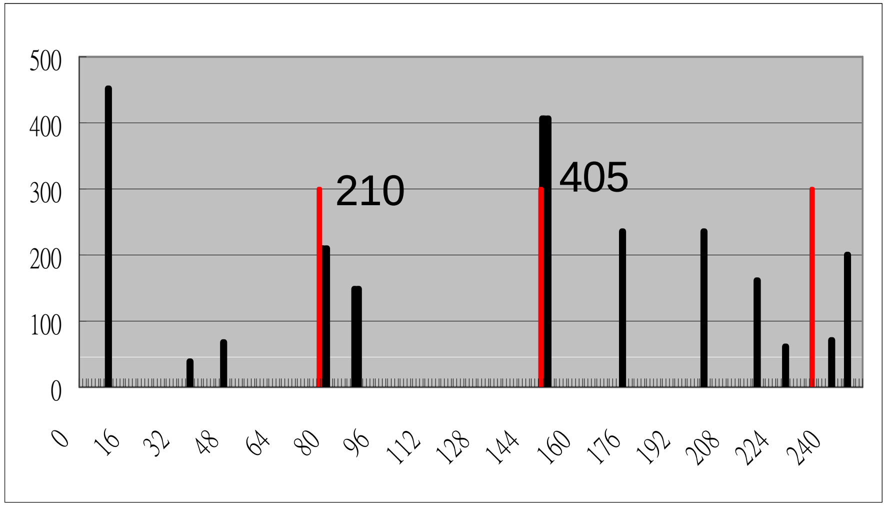
Ensure: A peptide P of mass between $W-\delta$ and $W+\delta$ which maximizes $score_Y(M, P)$.

```
Set $V(r) = 0$ for $r < 0$
for $r = 0$ to $W + \delta$ do
    $V(r) = \max_{a \in A}\{V(r - w(a)) + f_M(r + 19)\}$
end for
$r' = \arg\max_{W-\delta \le r \le W+\delta} V(r)$
Through back-tracing, we find the peptide $P$ of mass $r'$ which maximizes $score_Y(M, P)$
```

Example

- Given the spectrum M and wt=215.21.
  - V(76.05) = V(0)+210 = 210 (due to G)
  - V(147.13) = V(76.05)+450 = 615 (due to A)
  - V(234.21) = V(147.13)+0 = 615 (due to S)
- By backtracking, we recover SAG!

M



Time analysis

- We need to fill-in the V table with wt entries.
- Each entry can be computed in O(|A|) time.
- So, total time complexity is O(|A|wt) time.

Can we use more information other than y-ions?

- Yes. We can also use information from b-ions.

Better scoring scheme

- Consider a peptide $P=a_1a_2...a_k$
  - If M is a spectrum for P, we can find peaks for $m/z = y_i$ or $m/z = b_i$ for $i=1,2,...,k$
- So, we redefine the score function score(M,P) as $\Sigma\{h|(x,h)\in M, |x-y_i|\leq\delta \text{ or } |x-b_i|\leq\delta \text{ for } i=1,2,...,k\}$

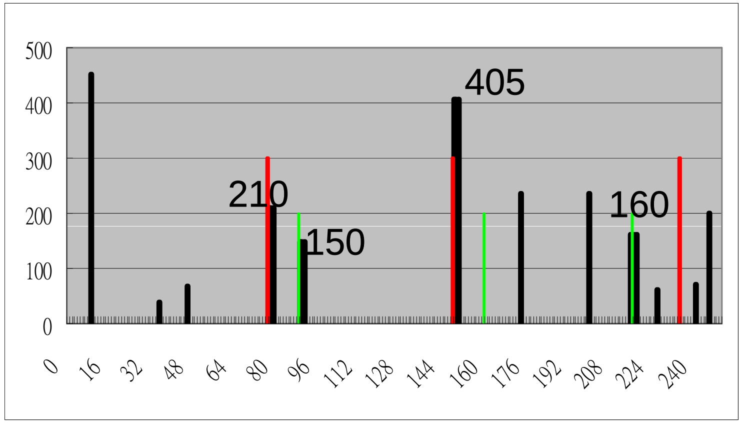
Better scoring scheme example

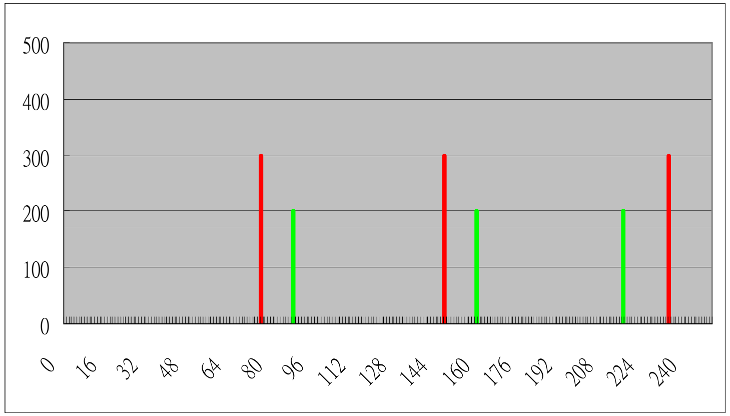
- E.g. P=SAG
  - y_1 = 57.05+71.08+87.08+19 = 234.21
  - y_2 = 57.05+71.08+19 = 147.13
  - y_3 = 57.05+19 = 76.05
  - b_1 = 87.08+1 = 88.08
  - b_2 = 87.08+71.08+1 = 159.16
  - b_3 = 87.08+71.08+57.05+1 = 216.21

Score(M,P)

= 210+405+150+160

= 925





Black peaks: real peaks

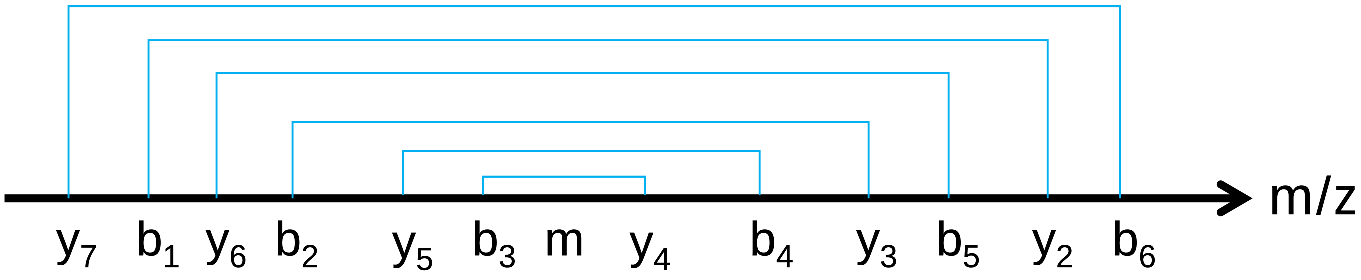
Green peaks: artificial b-ions

Red peaks: artificial y-ions

Observations

- Suppose $P=a_1a_2...a_k$.

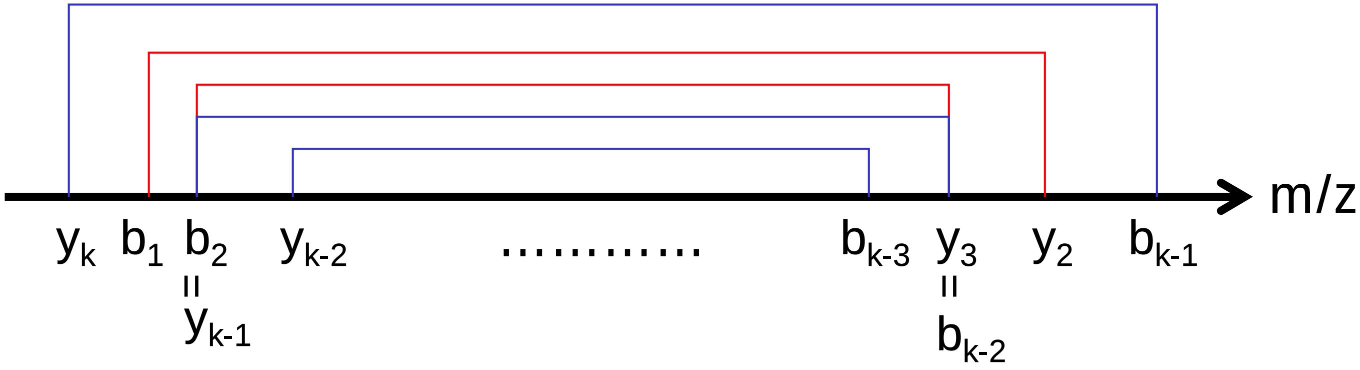
1. b_i is strictly increasing while y_j is strictly decreasing.
   - Proof: For any peptide Q and amino acid a, w(Qa), w(aQ) > w(Q).
   - Hence, $b_{i+1}-b_i$, $y_j-y_{j+1} \geq \min_{a\in A} w(a) = 57.05 > 0$
2. Note that $b_i+y_{i+1} = w(P)+20$.
   - Hence, we have (b_i, y_{i+1}), for all i=1,2,...,k, form a set of nested regions.
   - For the adjacent nested intervals, the mass different is at most $\max_{a\in A} w(a) = 186.21$.



Consider $P=a_1a_2...a_7$.
$m = (w(P)+20)/2$

Can we solve the problem using previous DP?

- No!
  - The reason is that, for some masses y_i and b_j, their masses may be very close and correspond to the same peak (x, h)$\in$M.
  - In this case, the previous DP will sum the same peaks two times.



Observation (II)

- Note that the outermost ℓ intervals are formed by breaking the prefix $a_1 \ldots a_i$ and the suffix $a_j \ldots a_k$, where $i+(k-j+1)=\ell$.
- Let score'(M,$a_1 \ldots a_i$, $a_j \ldots a_k$) be
  - the sum of the intensities of all b-ion and y-ion peaks formed by breaking the peptide P between a_x and a_{x+1} for $x \in \{1,\ldots,i\} \cup \{j-1\ldots,k-1\}$.
- Let $f_M(r,s)$ be the sum of all peaks in M which are close to r and wt+20-r but not close to s and wt+20-s. [used to avoid double counting!]
- We have

$$score'(M, a_1 \ldots a_i, a_j \ldots a_k)$$
$$= \begin{cases} score'(M, a_1 \ldots a_{i-1}, a_j \ldots a_k) + f_M(b_i, y_j) \text{ if } b_i \geq y_j \\ score'(M, a_1 \ldots a_i, a_{j+1} \ldots a_k) + f_M(y_j, b_i) \text{ otherwise} \end{cases}$$

Solution (a more complicated dynamic programming)

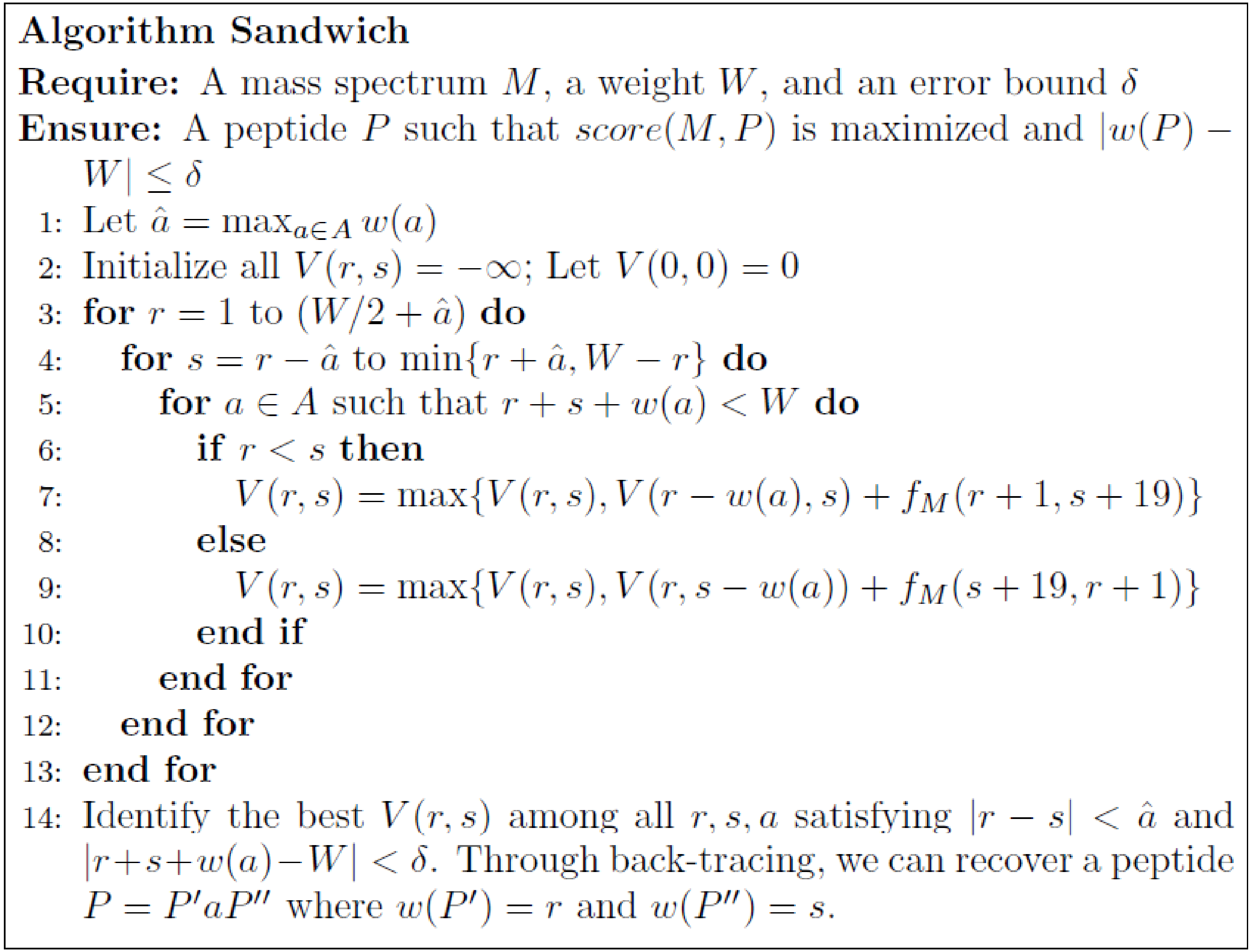
- Let â be $\max_{a \in A} w(a)$ = 186.21.
- For every |r-s|≤â, let V(r, s) be the maximum score'(M,P_1,P_2) among all possible P_1 and P_2 where w(P_1)=r and w(P_2)=s.

$$V(r,s) = \max \begin{cases} max_{a\in A}\{V(r-w(a),s)+f_M(r+1,s+19)\},\ r \geq s \\ max_{a\in A}\{V(r,s-w(a))+f_M(s+19,r+1)\},\ r < s \end{cases}$$

$$with\ base\ case\ V(r,s) = 0(r \leq 0, s \leq 0).$$

Solution (a more complicated dynamic programming)

- Aim: Find the best V(r,s) such that wt+20=r+s+w(a) for some a∈A.
  - Then, by back-tracking, we can recover the peptide.



```
Algorithm Sandwich
Require: A mass spectrum $M$, a weight $W$, and an error bound $\delta$
Ensure: A peptide $P$ such that $score(M, P)$ is maximized and $|w(P) - W| \leq \delta$
Let $\hat{a} = \max_{a \in A} w(a)$
Initialize all $V(r, s) = -\infty$; Let $V(0, 0) = 0$
for $r = 1$ to $(W/2 + \hat{a})$ do
  for $s = r - \hat{a}$ to $\min\{r + \hat{a}, W - r\}$ do
    for $a \in A$ such that $r + s + w(a) < W$ do
      if $r < s$ then
        $V(r, s) = \max\{V(r, s), V(r - w(a), s) + f_M(r + 1, s + 19)\}$
      else
        $V(r, s) = \max\{V(r, s), V(r, s - w(a)) + f_M(s + 19, r + 1)\}$
      end if
    end for
  end for
end for
Identify the best $V(r, s)$ among all $r, s, a$ satisfying $|r - s| < \hat{a}$ and
  $|r+s+w(a)-W| < \delta$. Through back-tracing, we can recover a peptide
  $P = P'aP''$ where $w(P') = r$ and $w(P'') = s$.
```

Time complexity

- We need to fill-in V(r,s) for all |r-s|≤â.
- So, we need to fill-in wt·â entries.
- Each can be filled-in using O(|A|) time.
- The time complexity is O(wt·â·|A|) time.

Spectrum Graph approach

- Another method to recover the peptide is based on spectrum graph, which is defined as follows.

Generating vertices in the spectrum graph

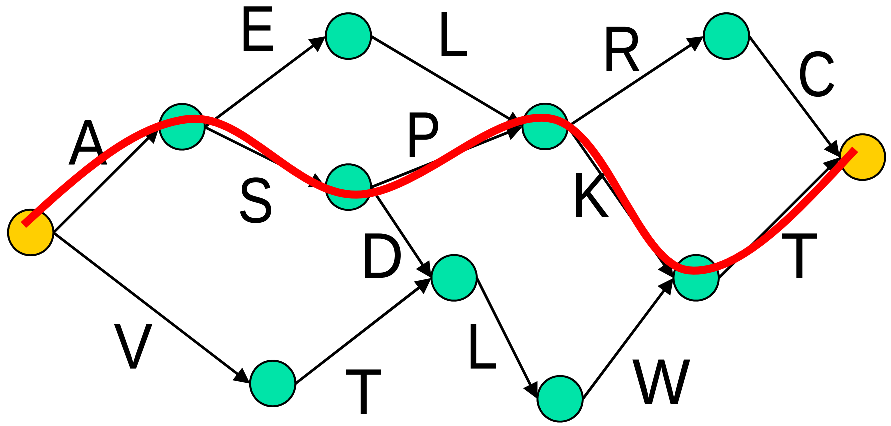
- For each mass r in the spectrum M,
  - We generate two vertices of masses r and wt-r.
- We also include 2 additional vertices:
  - starting vertex with mass = 0 and
  - ending vertex with mass = wt.

Generating edges in the spectrum graph

- For every pair of mass r and s,
  - If r-s equals the mass of an amino acid A,
    - we connect x and y with an edge of label A.
- Since there may be some missing peaks in the spectrum,
  - If r-s equals the total mass of two amino acids A_1A_2,
    - we connect x and y with an edge of label A_1A_2.
  - If r-s equals the total mass of three amino acids $A_1A_2A_3$,
    - we connect x and y with an edge of label $A_1A_2A_3$.

Meaning of a path in the graph

- Every path from start to end corresponds to a possible peptide in the spectrum
- However, there are many possible paths?



Weight of the edges

- Observe that a vertex has higher probability to be real if all ion types are available.
- Hence, we can assign a score depending on whether some ion types are missing.
- Then, this is a problem of finding the heaviest path, which can be solved in polynomial time.

Weighting function for Sherenga

- Assume noise is produced uniformly and randomly with probability q_R.
- Assume q_b is the probability that the b-ion peak exists in M given the b-ion appears in the theoretical spectrum.
- Similarly, assume q_y is the probability that the y-ion peak exists in M given the y-ion appears in the theoretical spectrum.
- The weight of every vertex with mass v is defined as the sum of $\text{score}_b(v)$ and $\text{score}_y(v)$, where

$$score_b(v) = \begin{cases} \log \frac{q_b}{q_R} & \text{if } v+1 \text{ exists in } M \\ \log \frac{1-q_b}{1-q_R} & \text{otherwise} \end{cases}$$

$$score_y(v) = \begin{cases} \log \frac{q_y}{q_R} & \text{if } W-v+19 \text{ exists in } M \\ \log \frac{1-q_y}{1-q_R} & \text{otherwise} \end{cases}$$

Protein Database searching Problem

- Input:
  - a database of proteins (DB)
  - a raw MS/MS spectrum (M)
  - The mass wt of the peptide corresponding to M
- Output:
  - A protein whose peptide is expected to have mass wt and a MS/MS spectrum similar to M.

- This lecture presents a solution called SEQUEST (Eng et al, 1994)

SEQUEST

- Step 1: Reduction of the tandem mass spectrometry data
  - To avoid noise, only 200 most abundant signals of the raw spectrum are used.
  - Also, the total signals of the 200 signals are renormalized to 100.
- Step 2: Search the protein database DB to find all peptides such that each peptide P has mass within (wt±1)Da

SEQUEST

- Step 3: Rank the top 500 fit sequences by a specific scoring function.

SEQUEST

- Step 4: Compare the spectral similarity. Use cross-correlation analysis to generate the final score and rank the sequences.
- The abundance of ions in the hypothetic spectrum: 50 (b-ion, y-ion), 25 (mass/charge within ±1 from b or y), or 10 (a-ion)

Conclusion

- This lecture presents two De Novo Peptide Sequencing algorithms.
- We also present the protein database searching algorithm SEQUEST.
- There are many other problems in this area. For example,
  - Identifying peptide modifications

References

- J. K. Eng, A. L. McCormack, J. R. Yates. “An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database”. J. Am. Soc. Mass Spectrom, 5:976-989, 1994.
- B. Ma, K. Zhang, C. Liang. “An Effective Algorithm for the Peptide De Novo Sequencing from MS/MS Spectrum”. CPM, 266-277, 2003.
- V. Dancik, T. A. Addona, K. R. Clauser, J. E. Vath, P. A. Pevzner. De novo peptide sequencing via tandem mass spectrometry. Journal of Computational Biology, 6:327-342, 1999.