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 \( \approx 10 \)).
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

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>71.08</td>
<td>M</td>
<td>131.19</td>
</tr>
<tr>
<td>C</td>
<td>103.14</td>
<td>N</td>
<td>114.1</td>
</tr>
<tr>
<td>D</td>
<td>115.09</td>
<td>P</td>
<td>97.12</td>
</tr>
<tr>
<td>E</td>
<td>129.12</td>
<td>Q</td>
<td>128.13</td>
</tr>
<tr>
<td>F</td>
<td>147.18</td>
<td>R</td>
<td>156.19</td>
</tr>
<tr>
<td>G</td>
<td>57.05</td>
<td>S</td>
<td>87.08</td>
</tr>
<tr>
<td>H</td>
<td>137.14</td>
<td>T</td>
<td>101.1</td>
</tr>
<tr>
<td>I</td>
<td>113.16</td>
<td>V</td>
<td>99.13</td>
</tr>
<tr>
<td>K</td>
<td>128.17</td>
<td>W</td>
<td>186.21</td>
</tr>
<tr>
<td>L</td>
<td>113.16</td>
<td>Y</td>
<td>163.18</td>
</tr>
</tbody>
</table>

- 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 $\pm 0.5$Da. 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 backbone.
- 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

![Diagram of peptide fragmentation](image)
Fragmentation of peptide (III)

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

\[ r = w(CTVFT) \]
\[ w = w(CTVFTEPREFK) \]

**Diagram:**
- Fragmentation of peptide: CTVFTEPREFK
- Mass of b-ion: \( w-r+19 \)
- Mass of y-ion: \( r+1 \)

Legend:
- CTVFT
- EPREFK
- Fragmentation arrow
Mass of the ions (I)

- Let $A$ be the set of amino acid. For every $a \in A$, $w(a) = \text{mass of its residue}$
- Let $P = a_1 a_2 \ldots a_k$ be a peptide.
  - $w(P) = \sum_{1 \leq j \leq k} w(a_j)$.
- 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
  - $b_i = 1 + w(a_1 a_2 \ldots a_i)$
- Mass of $y$-ion of the last $i$ amino acids is
  - $y_i = 19 + w(a_i \ldots a_k)$
- 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$_2$O, y-NH$_3$, y-H$_2$O-H$_2$O, y-H$_2$O-NH$_3$
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 \( \delta \) (default \( \delta = 0.5 \))

- **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 $P=a_1a_2\ldots a_k$
  - 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 $i=1,2,\ldots,k$

- So, we define the score function $\text{score}(M,P) = \sum \{h| (x,h) \in M, \ |x-y_i|\leq \delta \text{ for } i=1,2,\ldots,k\}$
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 $\delta$

- **Output:**
  - A peptide $P$ such that $\text{wt} - \delta \leq \text{wt}(P) \leq \text{wt} + \delta$
  - which maximizes $\text{score}(M, P)$. 
Brute-force solution

- For every possible peptide \( P \) such that \(|w(P) - wt| \leq \delta\), compute \( \text{score}(M, P) \).
- Report the peptide \( P \) such that \(|w(P) - wt| \leq \delta\) which maximizes \( \text{score}(M, P) \).

- 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 $f_M(r) = \sum \{ h \mid (x,h) \in M \text{ and } |x-r| \leq \delta \}$.
  - $f_M(r)$ is the sum of all peaks in $M$ whose mass is close to $r$.
- Observation:
  - $\text{score}(M, a_1a_2..a_k) = \text{score}(M, a_1a_2..a_{k-1}) + f_M(w(a_1a_2..a_k) + 19)$
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 \( \max_{|r-wt| \leq \delta} V(r) \). 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.

$$V(147.13) = \max \left\{ 
\begin{align*}
V(76.05) + 450 & \quad (due \ to \ A) \\
V(43.99) + 450 & \quad (due \ to \ C) \\
\ldots & 
\end{align*}
\right. $$
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 $score_Y(M, P)$.

1. Set $V(r) = 0$ for $r < 0$
2. for $r = 0$ to $W + \delta$ do
3. \quad $V(r) = \max_{a \in A} \{V(r - w(a)) + f_M(r + 19)\}$
4. end for
5. $r' = \arg\max_{W - \delta \leq r \leq W + \delta} V(r)$
6. Through back-tracing, we find the peptide $P$ of mass $r'$ which maximizes $score_Y(M, P)$
Example

- Given the spectrum M and \( \text{wt}=215.21 \).
  - \( V(76.05) = V(0) + 210 = 210 \) (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!
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 $\text{score}(M, P)$ as
  $$\sum \{ h | (x, h) \in M, |x-y_i| \leq \delta \text{ or } |x-b_i| \leq \delta \text{ for } i=1,2,...,k \}$$
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
Observations

1. Suppose $P = a_1a_2\ldots a_k$.
   - $b_i$ 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 $i = 1, 2, \ldots, k$, form a set of nested regions.
   - For the adjacent nested intervals, the mass difference is at most $\max_{a \in A} w(a) = 186.21$.

Consider $P = a_1a_2\ldots a_7$.

$m = \frac{(w(P) + 20)}{2}$
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 $i + (k - j + 1) = l$.

- Let $\text{score}'(M, a_1 \ldots a_i, a_j \ldots a_k)$ 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

\[
\text{score}'(M, a_1 \ldots a_i, a_j \ldots a_k) = \begin{cases} 
\text{score}'(M, a_1 \ldots a_{i-1}, a_j \ldots a_k) + f_M(b_i, y_j) & \text{if } b_i \geq y_j \\
\text{score}'(M, a_1 \ldots a_i, a_{j+1} \ldots a_k) + f_M(y_j, b_i) & \text{otherwise}
\end{cases}
\]
Solution (a more complicated dynamic programming)

- Let $\hat{a}$ be $\max_{a \in A} w(a) = 186.21$.
- For every $|r-s| \leq \hat{a}$, 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 $a \in A$.
  - 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 $\text{score}(M, P)$ is maximized and $|w(P) - W| \leq \delta$

1: Let $\hat{a} = \max_{a \in A} w(a)$
2: Initialize all $V(r, s) = -\infty$; Let $V(0, 0) = 0$
3: for $r = 1$ to $(W/2 + \hat{a})$ do
4:   for $s = r - \hat{a}$ to $\min\{r + \hat{a}, W - r\}$ do
5:     for $a \in A$ such that $r + s + w(a) < W$ do
6:       if $r < s$ then
7:         $V(r, s) = \max\{V(r, s), V(r - w(a), s) + f_M(r + 1, s + 19)\}$
8:       else
9:         $V(r, s) = \max\{V(r, s), V(r, s - w(a)) + f_M(s + 19, r + 1)\}$
10:    end if
11:   end for
12: end for
13: end for
14: 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 $P = P' a P''$ where $w(P') = r$ and $w(P'') = s$. 
Time complexity

- We need to fill-in $V(r,s)$ for all $|r-s| \leq \hat{\alpha}$.
- So, we need to fill-in $w_t \cdot \hat{\alpha}$ entries.
- Each can be filled-in using $O(|A|)$ time.
- The time complexity is $O(w_t \cdot \hat{\alpha} \cdot |A|)$ time.
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 $score_b(v)$ and $score_y(v)$, 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 $\pm 1$ 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

