Lecture 15: Protein Sequencing

Study Chapter 8.10-8.15
From DNA to Proteins

- DNA sequences
  - “OS” that controls living biological systems
  - Sections of DNA (Genes) encode proteins, like programs
  - Triplets of nucleotides (codons) encode the amino-acid sequences, as well as the stop codes, used to assemble proteins
  - Complications in going from DNA → Protein: introns, RNA editing prior to translation, post-translational modifications
Proteins

• Proteins are the “machinery” or “hardware”
  – Compose the cellular structures
  – Control the biochemical reactions in cells
  – Regulate and trigger the chain reactions (metabolic pathways) that result in the cell’s life cycle
  – Determine which parts of the DNA “code” are activated, executed, and when

• Like DNA, proteins are long molecular chains
  – Sequences of 20 amino acid residues rather than 4 nucleic acids
## Protein Components

- Proteins are made from 20 amino acids
- Peptide bonds join amino acids into long chains
- 100’s to 1000’s of amino acid residues long

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>3-Letter Code</th>
<th>1-Letter Code</th>
<th>Molecular Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>Ala</td>
<td>A</td>
<td>89.09</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys</td>
<td>C</td>
<td>121.16</td>
</tr>
<tr>
<td>Aspartate</td>
<td>Asp</td>
<td>D</td>
<td>133.10</td>
</tr>
<tr>
<td>Glutamate</td>
<td>Glu</td>
<td>E</td>
<td>147.13</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Phe</td>
<td>F</td>
<td>165.19</td>
</tr>
<tr>
<td>Glycine</td>
<td>Gly</td>
<td>G</td>
<td>75.07</td>
</tr>
<tr>
<td>Histidine</td>
<td>His</td>
<td>H</td>
<td>155.16</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile</td>
<td>I</td>
<td>131.18</td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys</td>
<td>K</td>
<td>146.19</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leu</td>
<td>L</td>
<td>131.18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>3-Letter Code</th>
<th>1-Letter Code</th>
<th>Molecular Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methionine</td>
<td>Met</td>
<td>M</td>
<td>149.21</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asn</td>
<td>N</td>
<td>132.12</td>
</tr>
<tr>
<td>Proline</td>
<td>Pro</td>
<td>P</td>
<td>115.13</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Gln</td>
<td>Q</td>
<td>146.15</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg</td>
<td>R</td>
<td>174.20</td>
</tr>
<tr>
<td>Serine</td>
<td>Ser</td>
<td>S</td>
<td>105.09</td>
</tr>
<tr>
<td>Threonine</td>
<td>The</td>
<td>T</td>
<td>119.12</td>
</tr>
<tr>
<td>Valine</td>
<td>Val</td>
<td>V</td>
<td>117.15</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Trp</td>
<td>W</td>
<td>204.23</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Tyr</td>
<td>Y</td>
<td>181.19</td>
</tr>
</tbody>
</table>
Protein Assembly

- Amino acids are joined by peptide bonds into long chains
- These chains “fold” into proteins
- Interact with other proteins and large molecules
Protein Sequencing

• Purify a sample

• Break into pieces
  – Proteases cleave proteins into smaller “peptide” chains

• Read fragments
  – Edman degradation for short peptide sequences
  – Mass spectrometry measures mass/charge
  – The “Hard” part

• Reassemble
  – Relatively easy
Peptide Fragmentation

Collision Induced Dissociation

\[
\text{H...-HN-CH-CO} \quad \ldots \quad \text{NH-CH-CO-NH-CH-CO-…OH}
\]

\( R_{i-1} \quad R_i \quad R_{i+1} \)

Prefix Fragment  Suffix Fragment

- Peptides tend to fragment along the backbone.
- Fragments can also lose neutral chemical groups like \( \text{NH}_3 \) and \( \text{H}_2\text{O} \).
Breaking Peptides into Fragment Ions

• Proteases, e.g. trypsin, break proteins 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

NH$_2$- G P F N A -CO$_2$H

G P F N A
G P F N A
G P F N A
G P F N A

N-terminal peptides

G P F N A
G P F N A
G P F N A
G P F N A

C-terminal peptides

P F F N A
P F F N A
P F F N A
P F F N A

11/5/13
Comp 555
Fall 2013
Terminal peptides and ion types

Peptide: GPFN

Mass (D): 57 + 97 + 147 + 114 = 415

Peptide: GPFN (without H₂O)

Mass (D): 57 + 97 + 147 + 114 - 18 = 397
N- and C-terminal Peptides

NH₂-GPFNA-CO₂H

N-terminal peptides

GPFNA

C-terminal peptides

PFFNA

11/5/13
Comp 555
Fall 2013

11/5/13
Comp 555
Fall 2013

11/5/13
Comp 555
Fall 2013

11/5/13
Comp 555
Fall 2013

11/5/13
Comp 555
Fall 2013

11/5/13
Comp 555
Fall 2013

11/5/13
Comp 555
Fall 2013

11/5/13
Comp 555
Fall 2013

11/5/13
Comp 555
Fall 2013

11/5/13
Comp 555
Fall 2013

11/5/13
Comp 555
Fall 2013

N- and C-terminal Peptides

486
NH₂-

415
401
154
57

-CO₂H

N-terminal peptides

71
185
332
429
C-terminal peptides

11/5/13  Comp 555  Fall 2013
N- and C-terminal Peptides
N- and C-terminal Peptides

Reconstruct peptide from the set of masses of fragment ions

(mass-spectrum)
The peaks in the mass spectrum:

- **Prefix** and **Suffix** Fragments.
- Fragments with **neutral losses** (-H$_2$O, -NH$_3$)
- Noise and missing peaks.
Protein Identification with MS/MS

Peptide Identification:

MS/MS

Intensity

mass

G V D L K
De Novo vs. Database Search

Database Search

Database of known peptides

MDERHILNM, KLQWVCSDL, PTYWASDL, ENQIKRSACVM, TLACHGGE, NGALPQWRT, HLLERTKMN, GGPASSDA, GGLITGMQSD, MQPLMNWE, AAKKMRRPQ, AVGELTK, HEWAILF, GHNWAMMNAC, GVFGSLRA, EHLKNAATYIN..

AVGELTK

De Novo

Database of all peptides $20^N$

AAAAAAA AAAAACC, AAAAADD, AAAAEE, AAAAACC, AAAAASS, AAAAHAAAAAA,

AVGELTK, AVGELTK, AVGELTL, AVGELTM,

YYYYYYYYS, YYYYYYYS, YYYYYYYYV, YYYYYYY

Mass, Score
A Paradox

• Database of all peptides is huge $\approx O(20^n)$.
• 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
a
a is an ion type shift in b
noise
Some Mass Differences between Peaks Correspond to Amino Acids
Ion Types

- Some masses correspond to fragment ions, others are just random noise
- Known ion types $\Delta = \{\delta_1, \delta_2, \ldots, \delta_k\}$ allow us to distinguish fragment ions from noise
- We can learn ion types $\delta_i$ and their probabilities $q_i$ by analyzing a large test sample of annotated spectra.
Example of Ion Type

• \( \Delta = \{ \delta_1, \delta_2, \ldots, \delta_k \} \)

• Ion types

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

correspond to

\[ \Delta = \{ 0, 17, 18 \} \]

*Note: In reality the \( \delta \) value of ion type \( b \) is -1 but we will “hide” it for the sake of simplicity*
Matching Spectra

- 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
- $\Delta$: set of possible ion types
- $m$: parent mass

Output:
- $P$: peptide with mass $m$, whose theoretical spectrum best matches the experimental $S$ spectrum
Vertices of Spectrum Graph

- Masses of potential N-terminal peptides
- Vertices are generated by reverse shifts corresponding to ion types
  \[ \Delta = \{\delta_1, \delta_2, \ldots, \delta_k\} \]
- Every N-terminal peptide can generate up to \( k \) ions
  \[ m - \delta_1, m - \delta_2, \ldots, m - \delta_k \]
- Every mass \( s \) in an MS/MS spectrum generates \( k \) vertices
  \[ V(s) = \{s + \delta_1, s + \delta_2, \ldots, s + \delta_k\} \]
  corresponding to potential N-terminal peptides
- **Vertices of the spectrum graph:**
  \[ \{\text{initial vertex}\} \cup V(s_1) \cup V(s_2) \cup \ldots \cup V(s_m) \cup \{\text{terminal vertex}\} \]
Reverse Shifts

Shift in $\text{H}_2\text{O}$

Shift in $\text{H}_2\text{O}+\text{NH}_3$
Edges of Spectrum Graph

- Two vertices with mass difference corresponding to an amino acid $A$:
  - Connect with an edge labeled by $A$
Paths

- Paths in the labeled graph spell out amino acid sequences

- There are many paths, how to find the correct one?

- We need **scoring function** to evaluate paths
Path Score

- \( p(P, S) = \text{probability that peptide } P \text{ produces spectrum } S = \{s_1, s_2, ... s_q\} \)

- \( p(P, s) = \text{the probability that peptide } P \text{ generates a peak } s \)

- **Scoring** = computing probabilities

- \( p(P, S) = \prod_{s \in S} p(P, s) \)
Peak Score

• For a position \( t \) that represents ion type \( d_j \):

\[
p(P,s_t) = \begin{cases} 
q_j, & \text{if peak is generated at } t \\
1-q_j, & \text{otherwise}
\end{cases}
\]
Peak Score (cont’d)

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

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

• $q_R =$ the probability of a noisy peak that does not correspond to any ion type
For a given MS/MS spectrum $S$, find a peptide $P'$ maximizing $p(P,S)$ over all peptides $P$:

$$p(P',S) = \max_P p(P,S)$$

- Peptides = paths in the spectrum graph
- $P'$ = the optimal path in the spectrum graph
Ions and Probabilities

- Tandem mass spectrometry is characterized by a set of ion types \( \{\delta_1, \delta_2, \ldots, \delta_k\} \) and their probabilities \( \{q_1, \ldots, q_k\} \)

- \( \delta_i \)-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 \textit{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.

• $T_i = \{t_1, t_2, \ldots, t_k\}$ - set of positions that represent ions of partial peptides $P_i$.

• A peak at position $t_{\delta j}$ is generated with probability $q_j$.

• $R = T - (\bigcup T_i)$ - set of positions that are not associated with any partial peptides (noise).
• 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:

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

Database Search

De Novo

Database of known peptides

MDERHILNM, KLQWVCSDL, PTYWASDL, ENQIKRSACVM, TLACHGGEM, NGALPQWRT, HLLERTKMNVV, GGPASSDA, GGLITGMQSD, MQPLMNWE, AKKMNRT, AVGELTK, HEWAILF, GHNLWAMNAC, GVFGSVLRA, EKLNAATYIN..

AVGELTK
Peptide Identification Problem

**Goal:** Find a peptide *from the database* with maximal match between an experimental and theoretical spectrum.

**Input:**
- $S$: experimental spectrum
- *database of peptides*
- $\Delta$: set of possible ion types
- $m$: parent mass

**Output:**
- A peptide of mass $m$ *from the database* whose theoretical spectrum matches the experimental spectrum the best
MS/MS Database Search

Database search in mass-spectrometry has been very successful in identification of already known proteins.

Experimental spectrum can be compared with theoretical spectra of database peptides to find the best fit.

SEQUEST (Yates et al., 1995)

But reliable algorithms for identification of new protein forms via mutation is a much more difficult problem.
Modified Peptides

- Virtual Database Approach

- Yates et al., 1995: an exhaustive search in a virtual database of all modified peptides.

- Exhaustive search leads to a large combinatorial problem, even for a small set of modifications types.

- Problem (Yates et al., 1995). Extend the virtual database approach to a large set of modifications.
Exhaustive Search for modified peptides.

- For each peptide, generate all modifications.
- Score each modification.

2^5 = 32 possibilities, with 2 types of modifications!
Peptide Identification Challenge

Very similar peptides may have very different spectra!

**Goal**: Define a notion of spectral similarity that correlates well with the sequence similarity.

If peptides are a few mutations/modifications apart, the spectral similarity between their spectra should be high.
Deficiency of Shared Peaks Count

Shared peaks count (SPC): intuitive measure of spectral similarity.

Problem: SPC diminishes very quickly as the number of mutations increases.

Only a small portion of correlations between the spectra of mutated peptides is captured by SPC.
S(PRTEIN) = \{98, 133, 246, 254, 355, 375, 476, 484, 597, 632\}
S(PRTEYN) = \{98, 133, 254, 296, 355, 425, 484, 526, 647, 682\}
S(PGTEYN) = \{98, 133, 155, 256, 296, 385, 425, 526, 548, 583\}

no mutations
SPC=10

1 mutation
SPC=5

2 mutations
SPC=2
Spectral Convolution

\[ S_2 \ominus S_1 = \{ s_2 - s_1 : s_1 \in S_1, s_2 \in S_2 \} \]

Number of pairs \( s_1 \in S_1, s_2 \in S_2 \) with \( s_2 - s_1 = x \) :
\[ (S_2 \ominus S_1)(x) \]

The shared peaks count (SPC peak):
\[ (S_2 \ominus S_1)(0) \]
Elements of $S_2 \oplus S_1$ represented as elements of a **difference matrix**. The elements with multiplicity >2 are colored; the elements with multiplicity =2 are circled. The SPC takes into account only the red entries.
Spectral Convolution: An Example

![Graph showing spectral convolution with labeled axes and values.](image-url)
Spectral Comparison: Difficult Case

\[ S = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\} \]

Which of the spectra

\[ S' = \{10, 20, 30, 40, 50, 55, 65, 75, 85, 95\} \]

or

\[ S'' = \{10, 15, 30, 35, 50, 55, 70, 75, 90, 95\} \]

fits the spectrum \( S \) the best?

SPC: both \( S' \) and \( S'' \) have 5 peaks in common with \( S \). Spectral Convolution: reveals the peaks at 0 and 5.
## Spectral Comparison: Difficult Case

### \( S \oplus S' \)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>-10</th>
<th>-20</th>
<th>-30</th>
<th>-40</th>
<th>-50</th>
<th>-60</th>
<th>-70</th>
<th>-80</th>
<th>-90</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0</td>
<td>-10</td>
<td>-20</td>
<td>-30</td>
<td>-40</td>
<td>-50</td>
<td>-60</td>
<td>-70</td>
<td>-80</td>
<td>-90</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>0</td>
<td>-10</td>
<td>-20</td>
<td>-30</td>
<td>-40</td>
<td>-50</td>
<td>-60</td>
<td>-70</td>
<td>-80</td>
</tr>
<tr>
<td>30</td>
<td>20</td>
<td>10</td>
<td>0</td>
<td>-10</td>
<td>-20</td>
<td>-30</td>
<td>-40</td>
<td>-50</td>
<td>-60</td>
<td>-70</td>
</tr>
<tr>
<td>40</td>
<td>30</td>
<td>20</td>
<td>10</td>
<td>0</td>
<td>-10</td>
<td>-20</td>
<td>-30</td>
<td>-40</td>
<td>-50</td>
<td>-60</td>
</tr>
<tr>
<td>55</td>
<td>45</td>
<td>35</td>
<td>25</td>
<td>15</td>
<td>5</td>
<td>-5</td>
<td>-15</td>
<td>-25</td>
<td>-35</td>
<td>-45</td>
</tr>
<tr>
<td>65</td>
<td>55</td>
<td>45</td>
<td>35</td>
<td>25</td>
<td>15</td>
<td>5</td>
<td>-5</td>
<td>-15</td>
<td>-25</td>
<td>-35</td>
</tr>
<tr>
<td>75</td>
<td>65</td>
<td>55</td>
<td>45</td>
<td>35</td>
<td>25</td>
<td>15</td>
<td>5</td>
<td>-5</td>
<td>-15</td>
<td>-25</td>
</tr>
<tr>
<td>85</td>
<td>75</td>
<td>65</td>
<td>55</td>
<td>45</td>
<td>35</td>
<td>25</td>
<td>15</td>
<td>5</td>
<td>-5</td>
<td>-15</td>
</tr>
</tbody>
</table>

### \( S \oplus S'' \)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>-10</th>
<th>-20</th>
<th>-30</th>
<th>-40</th>
<th>-50</th>
<th>-60</th>
<th>-70</th>
<th>-80</th>
<th>-90</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>-5</td>
<td>-15</td>
<td>-25</td>
<td>-35</td>
<td>-45</td>
<td>-55</td>
<td>-65</td>
<td>-75</td>
<td>-85</td>
<td>-95</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>0</td>
<td>-10</td>
<td>-20</td>
<td>-30</td>
<td>-40</td>
<td>-50</td>
<td>-60</td>
<td>-70</td>
<td>-80</td>
</tr>
<tr>
<td>25</td>
<td>15</td>
<td>5</td>
<td>-5</td>
<td>-15</td>
<td>-25</td>
<td>-35</td>
<td>-45</td>
<td>-55</td>
<td>-65</td>
<td>-75</td>
</tr>
<tr>
<td>40</td>
<td>30</td>
<td>20</td>
<td>10</td>
<td>0</td>
<td>-10</td>
<td>-20</td>
<td>-30</td>
<td>-40</td>
<td>-50</td>
<td>-60</td>
</tr>
<tr>
<td>60</td>
<td>50</td>
<td>40</td>
<td>30</td>
<td>20</td>
<td>10</td>
<td>0</td>
<td>-10</td>
<td>-20</td>
<td>-30</td>
<td>-40</td>
</tr>
<tr>
<td>65</td>
<td>55</td>
<td>45</td>
<td>35</td>
<td>25</td>
<td>15</td>
<td>5</td>
<td>-5</td>
<td>-15</td>
<td>-25</td>
<td>-35</td>
</tr>
<tr>
<td>80</td>
<td>70</td>
<td>60</td>
<td>50</td>
<td>40</td>
<td>30</td>
<td>20</td>
<td>10</td>
<td>0</td>
<td>-10</td>
<td>-20</td>
</tr>
<tr>
<td>85</td>
<td>75</td>
<td>65</td>
<td>55</td>
<td>45</td>
<td>35</td>
<td>25</td>
<td>15</td>
<td>5</td>
<td>-5</td>
<td>-15</td>
</tr>
</tbody>
</table>
Limitations

Spectral convolution does not reveal that spectra $S$ and $S'$ are similar, while spectra $S$ and $S''$ are not.

Clumps of shared peaks: the matching positions in $S'$ come in clumps while the matching positions in $S''$ don't.

This important property was not captured by spectral convolution.
Shifts

\[ A = \{a_1 < ... < a_n\} : \text{an ordered set of natural numbers.} \]

A *shift* \((i, \Delta)\) is characterized by two parameters, the starting position \((i)\) and the shift distance \((\Delta)\). The shift \((i, \Delta)\) transforms

\[ \{a_1, ..., a_n\} \]

into

\[ \{a_1, ..., a_{i-1}, a_i + \Delta, ..., a_n + \Delta\} \]
The shift \((i, \Delta)\) transforms \(\{a_1, \ldots, a_n\}\) into \(\{a_1, \ldots, a_{i-1}, a_i + \Delta, \ldots, a_n + \Delta\}\)

\[\begin{array}{ccccccccccc}
10 & 20 & 30 & 40 & 50 & 60 & 70 & 80 & 90 \\
& & & & & & & & & \\
shift (4,-5) & & & & & & & & & \\
10 & 20 & 30 & 35 & 45 & 55 & 65 & 75 & 85 \\
& & & & & & & & & \\
shift (7,-3) & & & & & & & & & \\
10 & 20 & 30 & 35 & 45 & 55 & 62 & 72 & 82 \\
\end{array}\]
Spectral Alignment Problem

• Find a series of $k$ shifts that make the sets $A=\{a_1, \ldots, a_n\}$ and $B=\{b_1, \ldots, b_n\}$ as similar as possible.

• Provides a notion of “$k$-similarity” between sets

• $D(k)$ - the maximum number of elements in common between sets after $k$ shifts (Like SPC).
Representing Spectra in 0-1 Alphabet

- Quantize (bin) the mass dimension
- Convert spectrum to a 0-1 string with 1s corresponding to the positions of the peaks.
Comparing Spectra=Comparing 0-1 Strings

- A modification with positive offset corresponds to inserting a block of 0s
- A modification with negative offset corresponds to deleting a block of 0s
- Comparison of theoretical and experimental spectra (represented as 0-1 strings) corresponds to a (somewhat unusual) edit distance/alignment problem where elementary edit operations are insertions/deletions of blocks of 0s
- Use sequence alignment algorithms!
Spectral Alignment vs. Sequence Alignment

- Manhattan-like graph with different alphabet and scoring.
- Movement can be diagonal (matching masses) or horizontal/vertical (insertions/deletions corresponding to PTMs).
- At most $k$ horizontal/vertical moves.
A = \{a_1, \ldots, a_n\} and B = \{b_1, \ldots, b_n\}

Spectral product $A \otimes B$: two-dimensional matrix with $nm$ 1s corresponding to all pairs of indices $(a_i, b_j)$ and remaining elements being 0s.

SPC: the number of 1s at the main diagonal.

$\delta$-shifted SPC: the number of 1s on the diagonal $(i, i + \delta)$
Spectral Alignment: $k$-similarity

$k$-similarity between spectra: the maximum number of 1s on a path through this graph that uses at most $k+1$ diagonals.

$k$-optimal spectral alignment = a path.

The spectral alignment allows one to detect more and more subtle similarities between spectra by increasing $k$. 
SPC reveals only $D(0) = 3$ matching peaks.

Spectral Alignment reveals more hidden similarities between spectra: $D(1) = 5$ and $D(2) = 8$ and detects corresponding mutations.
Black line represents the path for $k=0$

Red lines represent the path for $k=1$

Blue lines (right) represent the path for $k=2$
Spectral Convolution’s Limitation

The spectral convolution considers diagonals separately without combining them into feasible mutation scenarios.
$D_{ij}(k)$: the maximum number of 1s on a path to $(a_i, b_j)$ that uses at most $k+1$ diagonals.

$$D_{ij}(k) = \max_{(i', j') < (i, j)} \begin{cases} D_{i', j'}(k) + 1, & \text{if } (i', j') \sim (i, j) \\ D_{i', j'}(k - 1) + 1, & \text{otherwise} \end{cases}$$

$$D(k) = \max_{ij} D_{ij}(k)$$

Running time: $O(n^4 k)$
Edit Graph for Fast Spectral Alignment

$\text{diag}(i,j)$ – the position of previous 1 on the same diagonal as $(i,j)$
Fast Spectral Alignment Algorithm

\[ M_{ij}(k) = \max_{(i', j') < (i, j)} D_{i'j'}(k) \]

\[ D_{ij}(k) = \max \left\{ D_{\text{diag}(i, j)}(k) + 1, M_{i-1, j-1}(k-1) + 1 \right\} \]

\[ M_{ij}(k) = \max \left\{ D_{ij}(k), M_{i-1, j}(k), M_{i, j-1}(k) \right\} \]

Running time: \( O(n^2 k) \)
Spectral Alignment: Complications

Spectra are combinations of an increasing (N-terminal ions) and a decreasing (C-terminal ions) number series.

These series form two diagonals in the spectral product, the main diagonal and the perpendicular diagonal.

The described algorithm deals with the main diagonal only.
Spectral Alignment: Complications

- Simultaneous analysis of N- and C-terminal ions
- Taking into account the intensities and charges
- Analysis of minor ions