Fragment Assembly.

(21) DNA sequencing & assembly:

Shotgun sequencing starts with a large sample of genomic DNA. The sample is sonicated, a process which randomly partitions each piece of DNA in the sample into inserts; the inserts that are smaller than 500 nucleotides are removed from further consideration. Before the inserts can be read, each one must be multiplied billions of times, so that it is possible to read them through gel electrophoresis.

(*) Shortest superstring problem:

Given a set of strings find the shortest string that contains all of them.

→ ~~Graph~~ Set of strings $\{s_1, s_2, \ldots, s_n\}$

→ Graphical representation:

$s_i \rightarrow s_j$ if s_i overlaps s_j
(prefix of s_i = prefix of s_j).

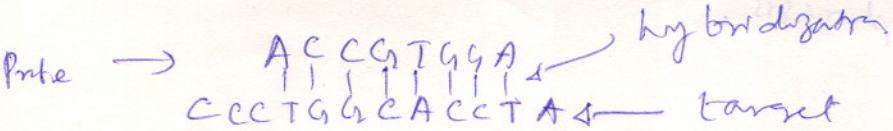
→ equivalent to Traveling Salesman Problem (or shortest Hamiltonian Path finding problem) in a directed graph with n vertices (correspond to s_i's) & edges of length (overlap (s_i, s_j)).

(NP complete)

→ greedy strategy: repeatedly merge a pair of strings with max. overlap until only one string remains.

⊛ Sequencing by hybridization (SBH)

SBH involves in building a miniature DNA array (known as also as DNA chip) that contains thousands of short DNA fragments called probes. Given a short probe (a 8 - to 30 nucleotide single-stranded synthetic DNA fragment), the target will hybridize with the probe if the probe is substring of the target's ~~WATSON~~ complement. ~~When the~~



SBH relies on the hybridization of the target DNA fragment against a very large array of short probes. In this manner, probes can be used to test the unknown target DNA to determine its l-mer composition. The universal DNA array ~~consists~~ contains all 4^l probes of length l.

~~add to Opl~~

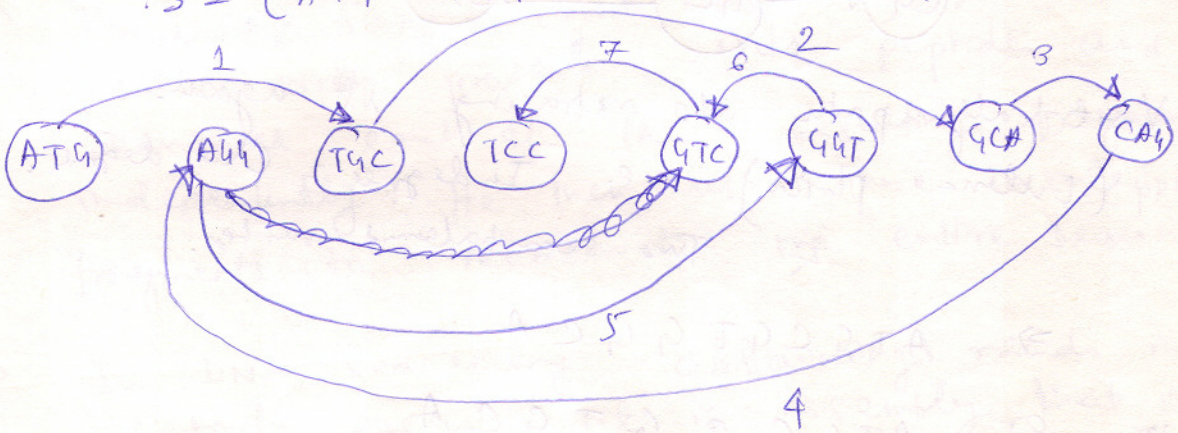
- Spectrum (S, l): For a string S of length n, the l-mer composition or spectrum of S, is the multiset of $n-l+1$ l-mers in S,

  Eg. Spectrum (TATGGTGC, 3)

  $= \{TAT, ATG, TGG, GGT, GTG, TGC\}$

- SBH Problem:
  Reconstruct a string from its l-mer composition.
  Input: Spectrum $\equiv (s_1, s_2, \ldots, s_n)$; O/p = S.
- Hamiltonian approach:
  $V = \{s_1, s_2, \ldots, s_n\}$
  $E = \{s_i \rightarrow s_j\}$, if '$l-1$' mer suffix of s_i = '$l-1$' mer prefix of s_j.

eg. $S = (ATG\ \ AGG\ \ TGC\ \ TCC\ \ GTC\ \ GGT\ \ GCA\ \ CAG)$

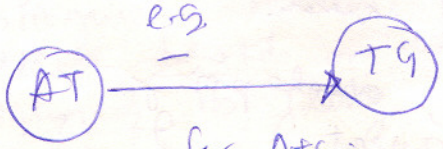


(nos. denotes sequence of edge to be traversed).

∴ String: ATGCAGGTCC

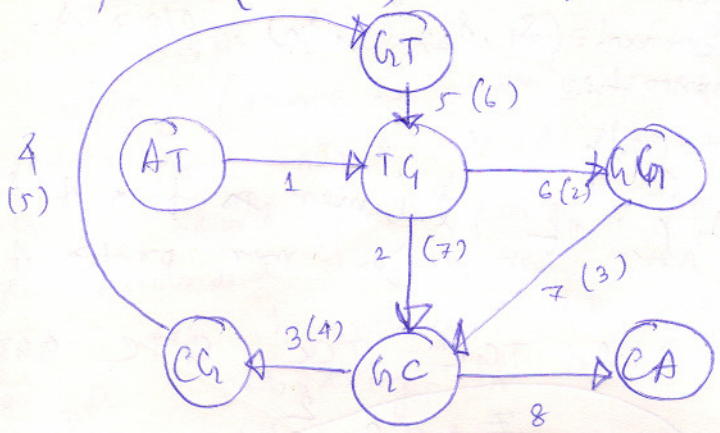
However the problem belongs to NP complete set.

- Eulerian Path approach:
  Build a graph with set of all $l-1$ - mers. an vertex & an edge is connected between two vertex, if a string has its $(l-1)$ prefix & suffix as those vertex.



for ATG:

2. (2) S2 { ATG, TGG, TGC, GTG, GGC, GCA, GCG, CGT }



Visit all path w/o retracing it again.
(Eulerian path) i.e. Indegree = Out degree
& Two semi balanced vertex.

⇒ ATGCGTGGCA

or, or, ATGGCGTGCA

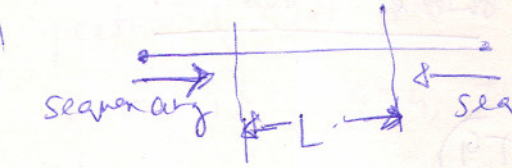
→ linear time traversal & reconstruction

(*) Problems in Fragment Assembly in DNA sequencing:

→ 1% to 3% error rate in reading inserts.
→ Double stranded DNA → which strand?
→ Repeats in DNA ⇒ major problem.

Weber & Myers suggested reading from two ends of a long insert (with a fixed gap).

mate-pair read



L > repeat length (mostly)

(This is unlikely, overlapping will be ambiguous due to repeat). It is unlikely both end will contain repeats

① Protein Sequencing and identification:

→ Edman degradation reaction to chop of an amino acid at a time from the end of a protein and read it. However, this only works for a few terminal amino acids before ~~this~~ the results become impossible to interpret.

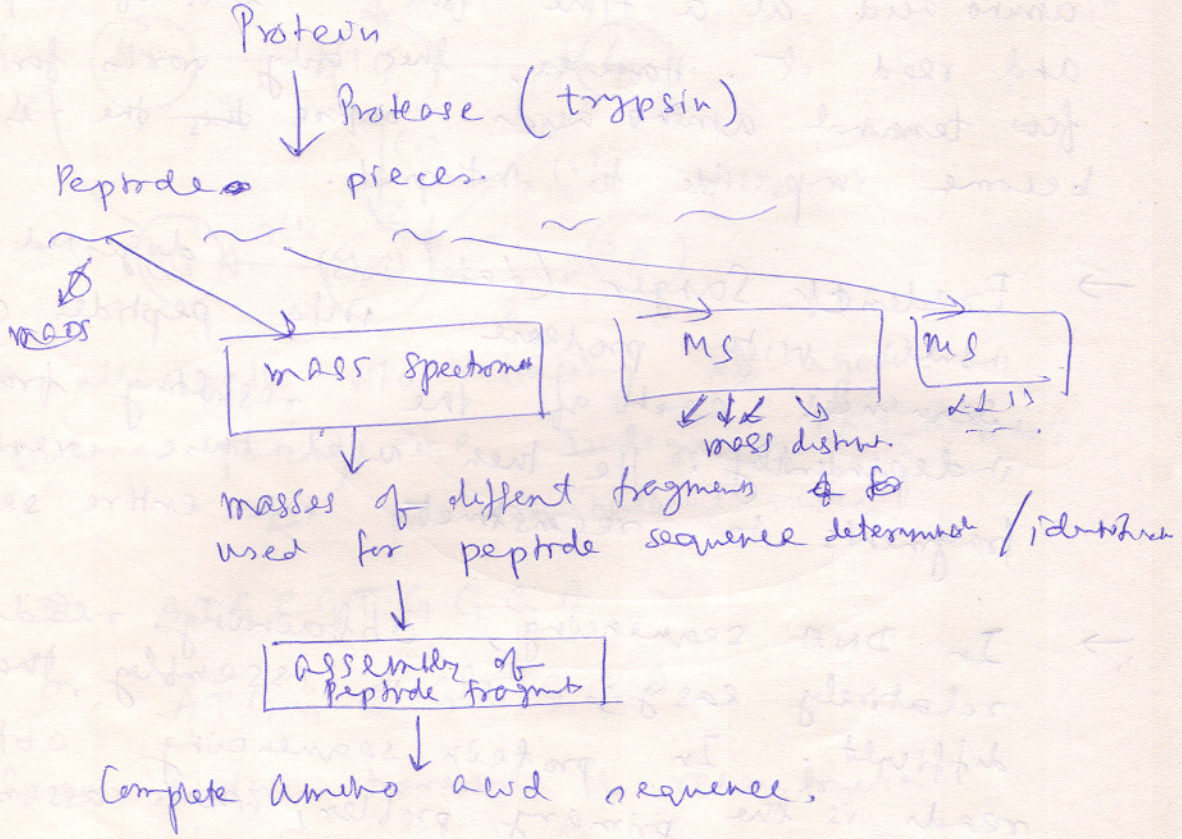
→ Frederick Sanger (Late' 40's) digested insulin with protease into peptides and sequenced each of the resulting fragments independently. He then used these overlapping fragments to reconstruct the entire sequence.

→ In DNA sequencing, obtaining reads is relatively easy; it is assembly that is difficult. In protein sequencing obtaining reads is the primary problem, while assembly is easy.

* Protein analysis by mass spectro~~scopy~~metry:

A mass spectrometer works like a charged sieve. A large molecule (peptide) gets broken into smaller fragments that have an electrical charge. These fragments are then spun around and accelerated in a magnetic field until they hit a detector. Because large fragments are harder to spin than small ones, one can distinguish between fragments with different masses based on the amount of energy required to fling the different fragments around. It happens that most molecules can be broken in several places, generating several diffe

ion types, the problem is to reconstruct the amino acid sequences of the peptide from the masses of these broken pieces.



- Peptide Sequencing problem:

$A = \{a_1, a_2, \ldots, a_{20}\}$ ← Set of amino acids.

$P = p_1 \ldots p_n$ ← a peptide (seq of amino acids), $p_i \in A$.

$m(a_i)$ ← mass of amino acid a_i.

$$\therefore m(P) = \sum_{i=1}^{n} m(p_i)$$

N-terminal peptide : $p_1 \ldots p_i \Rightarrow P_1 = m_i = \sum_{j=1}^{a_i} m(p_j)$

C- " " : $p_i \ldots p_n \Rightarrow P = m(P) - m_i$

$1 \leq i \leq n$

Mass spectra obtained by tandem mass spectrometry (MS/MS) consists predominantly of partial N-terminal peptides & C-terminal peptide.

Partial peptide may also loose a water H_2O ($_{\text{daltons}}^{18}$) or Amonia (NH_3) (17) or both (18 + 17).

Let $\Delta = \{\delta_1, \delta_2, \ldots, \delta_k\}$ be the set of such numbers, which may be lost from the molecular weight of a peptide fragment P_i with mass m_i is so that it converted to any one of the elements of $\{m_i - \delta_1, m_i - \delta_2, \ldots, m_i - \delta_k\}$

δ-ion of N terminal partial peptide P_i modifies its mass m_i to $m_i - \delta$ & the ion is called b-ions. Similarly δ-ions of C-terminal Partial peptide P_i^-, are called y-ions.

e.g. $b - H_2O$, $y - NH_3$ etc., $b - H_2O - NH_3$ etc.

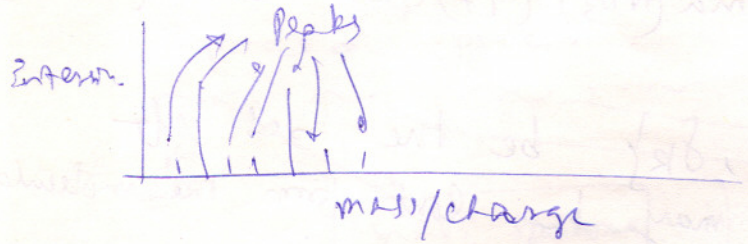
Theoretical spectrum = $T(P)$ = {masses of Partial peptides and masses of all possible derivatives due to subtraction of δ-ions}

$$= \left\{ \begin{matrix} 0, m_1, m_2, \ldots, m_n \\ m(P), m(P) - m_1, m(P) - m_2, \ldots \end{matrix} \middle| \; \forall i \right\}$$

Experimental spectrum = $S = \{s_1, \ldots, s_q\}$ = experimentally observed masses from tandem mass spectrometry.

Shared peak counts = $|S \cap T(p)|$

Mass spectrometry provides the intensities of different mass distribution. Hence, masses are denoted by discrete peaks in the interval.



e.g. $P = GPFNA$ $\delta_2=18$ $\delta_3=17$ $\delta_4=35$

	Sequence	mass	less H_2O	less NH_3	less H_2O+NH_3
	GPFNA	498	480	481	463
b_1	G	58	40	41	23
y_1	PFNA	442	424	425	405
b_2	GP	149	131	132	114
y_2	FNA	351	133	334	316
:	and so on (b_3, y_3), (b_4, y_4), ...				

Prob. Statement:

Find a peptide whose theoretical spectrum has a maximum match to a measured experimental spectrum.

Approach → Exhaustive Search (Generate 20^l combinations) ← Pruning (Probe pruning)

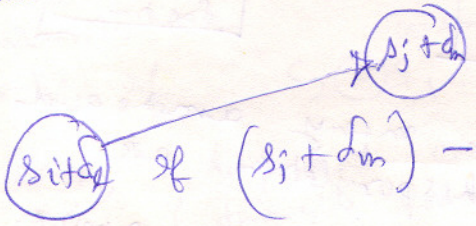
Approach → Analysis through Spectrum Graph

- **Spectrum Graph:** Let us assume Spectrum contains masses of b-ions only (N-terminal partial peptides)

  i.e. $S = \{s_1, s_2, \ldots, s_q\}$

  Let $\Delta = \{\delta_1, \delta_2, \ldots, \delta_k\}$

  For each s_i there may be any of $s_i + \delta_j$, $1 \le j \le k$ masses $\therefore$ there are $(k+1)$ vertices for s_i (including itself), $\delta_i = 0$. We define edge between two vertices if it differs by a single amino acids mol. wt. (out of 20).



If $(s_j + \delta_m) - (s_i + \delta_l) = m(a_t)$

$a_t \in \{a_1, a_2, \ldots, a_{20}\}$

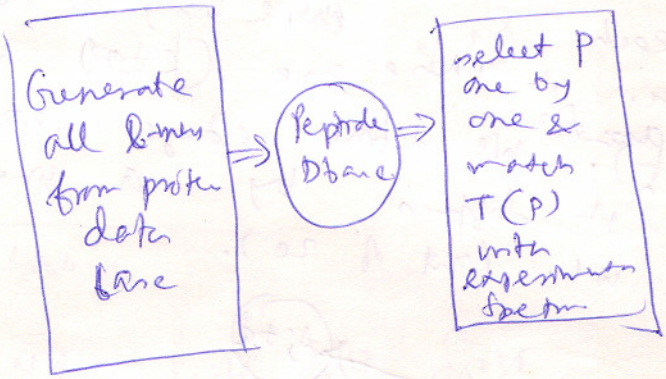
Add vertex 0 & m. Then peptide sequencing problem can be cast as finding a path from 0 to m in the resulting DAG.

No. of vertices: $qk+2$.

Complete Spectrum graph: If every N-type partial peptide is present. For a complete graph, $\exists$ a $(n+1)$ length path from 0 (source) to m-vx (Δ final). Often it is the longest path (max. edges) of the DAG. There may be ambiguous paths. The graph may be also incomplete.

Protein Identification:

Prob. Statement: Find a protein from a database that best matches the experimental spectrum.



Problem needs modification as many amino acids get modified due to phosphorylation etc.

i.e. A chemical modification at i^{th} amino acid of a peptide sequence $P_1 P_2 \cdots P_i \cdots P_n$ results in increased mass of the N-terminal peptides $P_i, P_{i+1}, \cdots, P_n$ & increased C-Terminal peptides $P_1, P_2, \cdots, P_{i-1}$.

Modified prob. statement: Find a peptide from the database that best matches the experimental spectrum with up to k modifications.

- **Spectral convolution:**

Multiset $S_2 \ominus S_1 = \{ s_2 - s_1 : s_1 \in S_1, s_2 \in S_2 \}$

Let $S_2 \ominus S_1 (x)$ = multiplicity of x

If S_2 & S_1 match (for, & w/o any modifications) $S_2 \ominus S_1 (0)$ should be high (maximum). It gives the count of shared peak.

For k-modifications : $S_2 \ominus S_1(x)$ will have peaks at different values of x

e.g. for $k=1$, if the mutation takes place at tth amino acid,

$$S_2 \ominus S_1(0) = t - \text{No. of N-terminal peptide}$$

$$S_2 \ominus S_1\left(m(P_{1t}) - m(P_{2t})\right) = n - t.$$

It is therefore reasonable to define spectral similarity as overall height of k highest peaks of $S_2 \ominus S_1$.

However, different combinations may result same $S_2 \ominus S_1$ (though some of the combinations are not valid set of partial peptide masses), as they may violate monotonicity of shifted masses).

- Spectral Alignment:

Let $S = \{s_1, s_2, \ldots, s_n\}$ be an ordered set of integers s.t. $s_1 < s_2 < \cdots < s_n$.

A shift Δ_i transforms S into $\{s_1 + \Delta_i, s_2, \ldots, s_{i-1}, s_i + \Delta_i, s_{i+1} + \Delta_i, \ldots, s_n + \Delta_i\}$. (Ordering property remains intact).

k similarity between A and B = $D(k)$ = max. no. of elements in common between these sets after k-shifts.

e.g.

$A = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$

$B = \{10, 20, 30, 40, 50, 55, 65, 75, 85, 95\}$ ← −5

$\therefore D(1)_{A \& B} = 10$, due to −5 shift at 6th element

$C = \{10, 15, 30, 35, 50, 55, 70, 75, 90, 95\}$

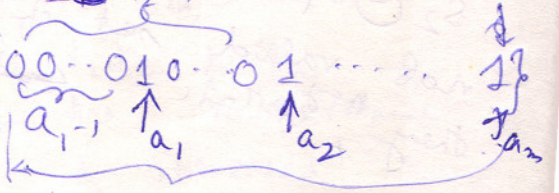
5 element match

$D(1)_{A \& C} = 6.$

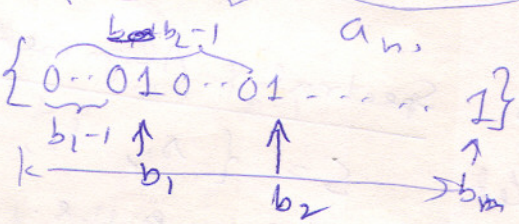
Spectral Alignment Problem:

Find the k-similarity between two sets.

$A = \{a_1, \ldots, a_n\} \equiv$ $a_n - n$ zeros



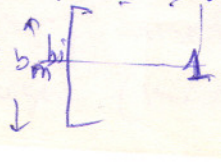
$B = \{b_1, b_2, \ldots, b_m\} =$ $b_m - m$ zeros



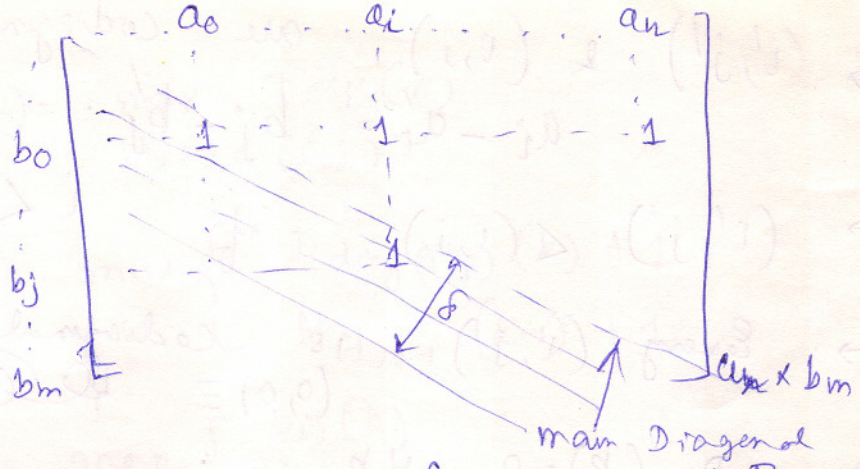
Any shift by δ in A means insertion or deletion of δ zeros. Dynamic programming for the no. of matches with k-shift. ⇒ longest path problem.

Spectral product: $A \otimes B = \{(a_i, b_j) \mid a_i \in A, b_j \in B\}$

2-D matrix representation:

$A \otimes B =$ 

$A \otimes B$ =



$A \otimes B(k, l) = 1$, if $k \in A$ & $l \in B$

$= 0$ otherwise.

$D(k) \Rightarrow$ k similarity measure

$D(0)$ = No. of 1's in main Diagonal.

$D(1) = D(0)$ + ~~Another~~ diagonal containing max. no. of 1 (out of the set of diagonals).

$\therefore D(k) = D(0) +$ max. no. of 1's in k diagonals ~~containing no of 1's~~ ∉.

$D(k) =$ k similarity between spectra

= max. no. of 1's on a path through the spectral matrix that uses at most $(k+1)$ diagonals

k-spectral alignment ⇒ the path that uses $(k+1)$ diagonals leading to $D(k)$.

Algorithm: Let $A_i = \{a_1, a_2, \cdots, a_i\}$

$B_j = \{b_1, b_2, \cdots, b_j\}$

→ Compute $D_{ij}(k)$ for A_i & B_j.

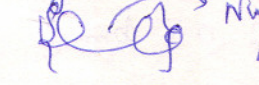
→ Using Dynamic Prog. & solve for $D(k) = D_{mn}(k)$.

→ (i', j') & (i, j) are codiagonal if

$$a_i - a'_i = b_j - b'_j$$

→ $(i', j') < (i, j)$ if $i' < i$ & $j' < j$

→ Every (i, j) is codiagonal with $(0, 0)$.

$(0,0) \equiv$  Null set $A_0, B_0\}$ $A_0 = \phi$ $B_0 = \phi$

→ $D_{00}(k) = 0, \quad \forall k.$

$$\therefore \quad D_{ij}(k) = \max_{(i',j') < (i,j)} \begin{cases} D_{i'j'}(k) + 1, & \text{if } (i',j') \text{ \& } (i,j) \text{ are codiagonal.} \\ D_{i'j'}(k-1) + 1, & \text{else.} \end{cases}$$

Time Complexity:

For computing (i, j) th. state by scanning the preformly visited states (w.r.t. $i' < i$ & $j' < j$) at (i', j') th. state

No. of such states: $i \cdot j$.

$$\therefore \text{ Total cost: } k \cdot \sum_{j=1}^{m} \sum_{i=1}^{m} i \cdot j = k \sum_j \sum_i$$

$$= \frac{m(m+1)\, n(n+1)}{4} \cdot k$$

$$\simeq O(n^4 \cdot k).$$

Compute D_{ij} upto kth level.

$k = 0, 1, \cdots$

* An effluent Algo.:

For reducing the time complexity ~~maintain~~ in compute the similarity from the closest diagonal state (ie maximal diagonal pair (i', j') w.r.t. $i' < i$ & $j' < j$). or $(0, 0)$ if such a pair does not exist)

Define also:

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

$$\therefore \quad D_{ij}(k) = \max \begin{cases} D_{diag(i,j)}(k) + 1 \\ M_{i-1,j-1}(k-1) + 1 \end{cases}$$

$$M_{ij}(k) = \max \begin{cases} D_{ij}(k) \\ M_{i-1,j}(k) \\ M_{i,j-1}(k). \end{cases}$$

⇒ ~~Actu~~ $\Rightarrow O(n^2 k)$

* Complexities:

~~Presence of~~

→ Both N-terminal & C-terminal ions complicate the problem.