Space-Optimal Alignments

Wednesday, February 10, 2021 10:46 AM

Previously,

Alignment models: { Global (NW algo) : Local (SW algo) : semi-global (Gotoh's algo) }

×

Gap models: { Linear gap (g) : Affine gap (h+g) }

Complexities:

Runtime $= O(mn)$

→ Space $= O(mn)$

"Reads" ~ 100 bp

S_1: ——— $m \sim 100$

S_2: ——— $n \sim 100$

Space for DP table $\propto 10^4$ cells

$\{S, D, I\} \Rightarrow$ each cell ≈ 12 bytes

Space (DP table) $\propto$ 120 kB

Genes ~ 10^4 bp

Space $\propto (10^4 \times 10^4) \times 12$

$\approx 12 \times 10^8$

$\approx$ 1200 MB

$\approx$ 1.2 GB

Genomes ~

- COVID (viral): 30K
- 10^9 (human): 10^{18} cells → Exabytes

900 M

~ 1 B cells

12 GB

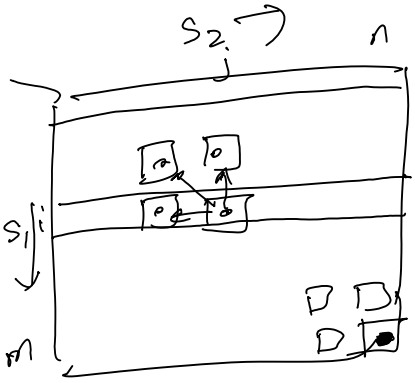
Space-optimal alignments

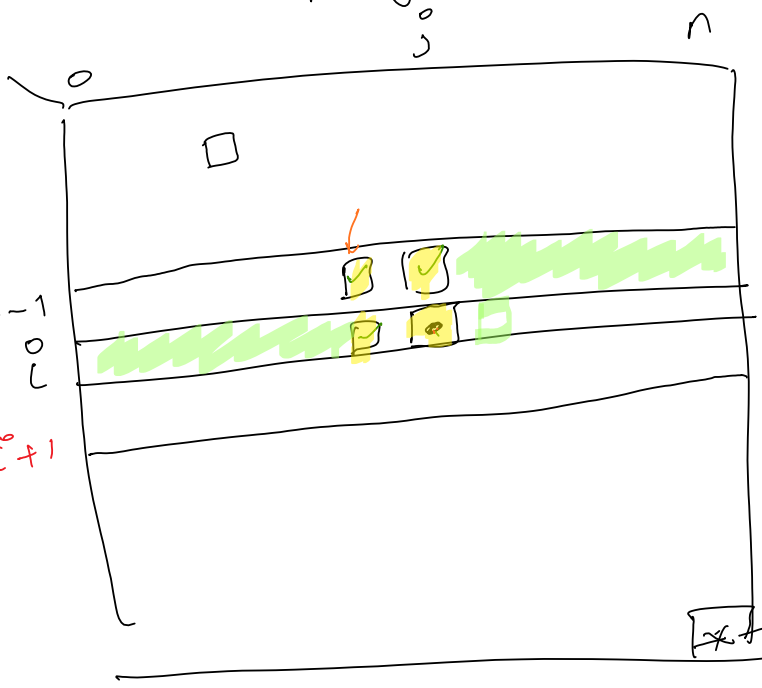
Wednesday, February 10, 2021 10:46 AM

Q) Can we store only part of the DP table?

Approach:

a) Do we need the full table for computing the optimal score?

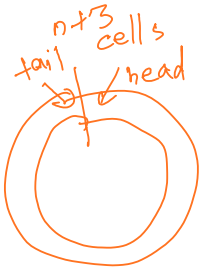




Space cost

$= O(m + n + 2n)$

m: s_1; n: s_2; $2n$: 2 rows (curr, prev)



Linear

Space-optimal alignments

Wednesday, February 10, 2021 10:46 AM

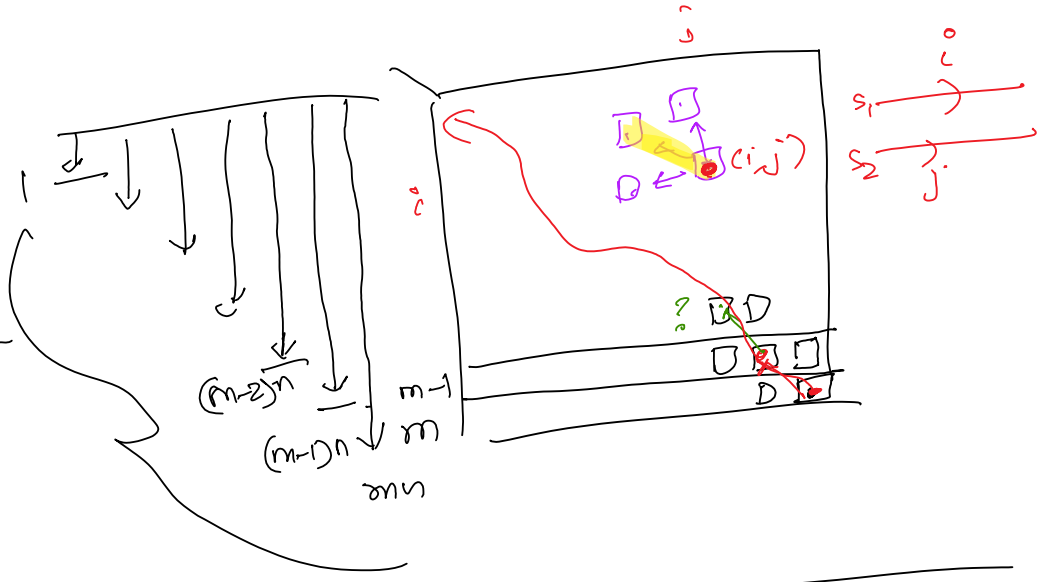
Q) what if we also need to output/reconstruct the optimal path?

brute force

Time cost

$\propto n(1+2+\cdots+m)$

$\propto n(m^2)$

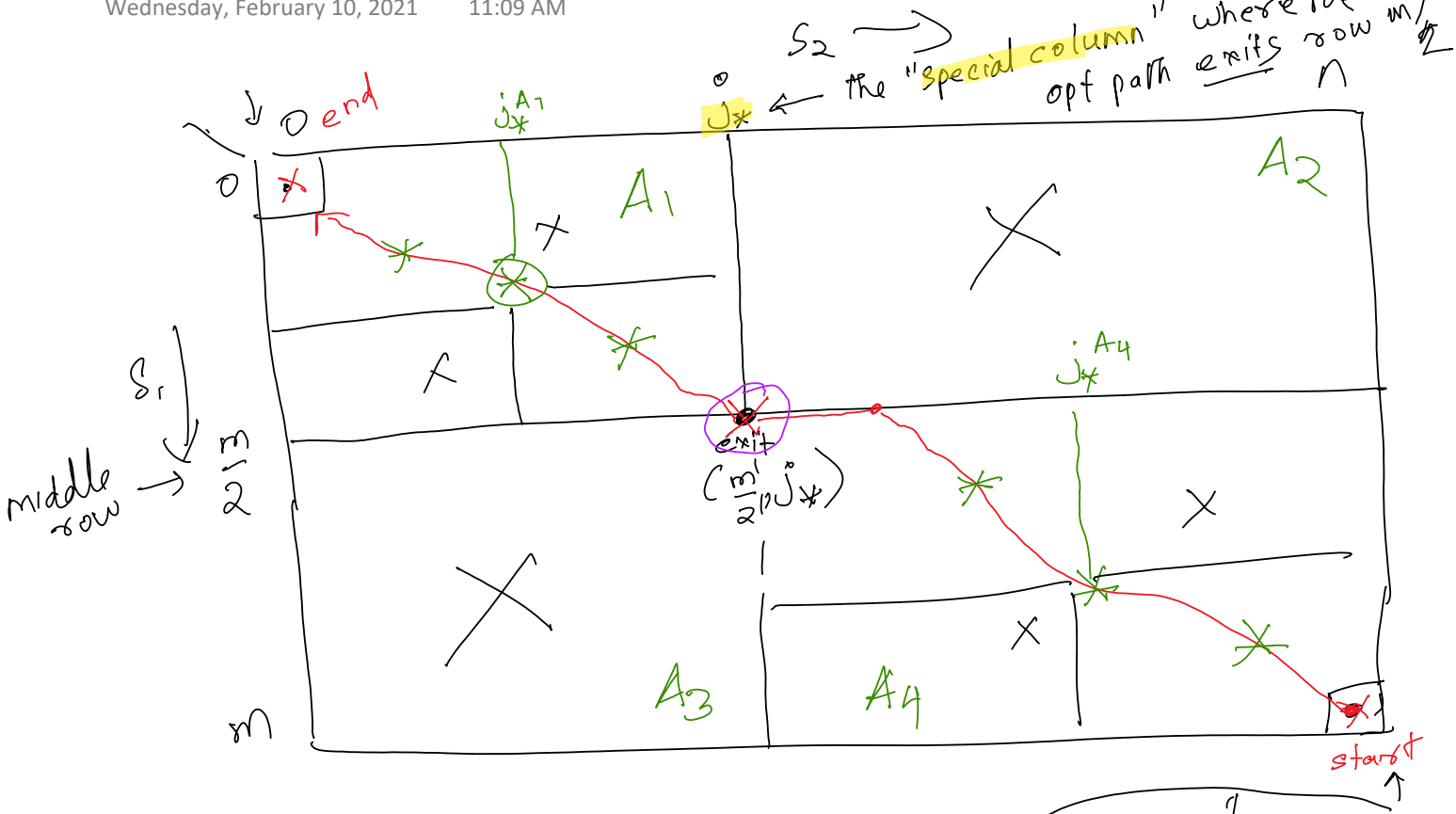


Goal: To Compute the opt path in $O(m+n)$ space and in $O(mn)$ time

(space-optimal)

Space-optimal alignments: The Hirschberg Technique

Wednesday, February 10, 2021 11:09 AM

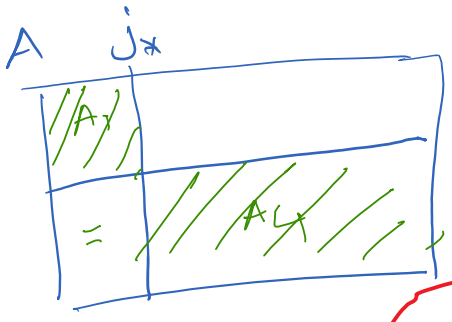


$A_1 + A_2 + A_3 + A_4 = (m+1)(n+1)$

"oracle ⇒ j^*"

Recursive step	# cells computed
1	mn
2	$+ \frac{mn}{2}$
3	$+ \frac{mn}{4}$
4	$mn/8$
⋮	⋮

$= A_1 + A_4$



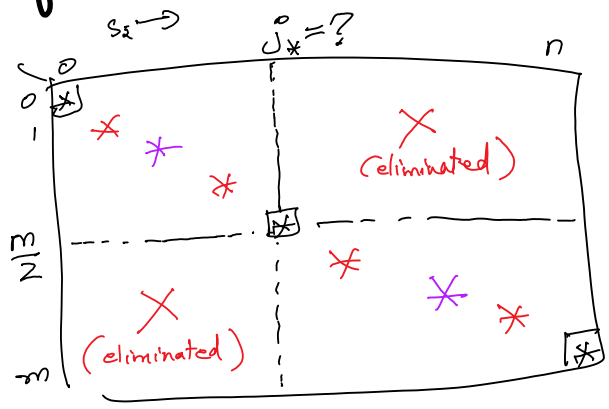
$mn(1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots) \approx 2mn = O(mn)$ time

If we could find j^* ⇒ then we could do the optimal retrace in $O(mn)$ time.

Determining the oracle (for the special column)

Friday, February 12, 2021 10:20 AM

Q) How to determine j^* (i.e. design of the oracle)?



Key Observation:

If you split an optimal path $\mathcal{P}$ into disjoint subpaths $\{P_1, P_2, \ldots P_k\}$ then each subpath P_i has to be optimal for that sub problem.

opt path $\mathcal{P} := P_1 \cdot P_2 \ldots P_i \ldots P_k$

An example:

Let the (unknown) optimal alignment be:

$a_1\ a_2\ a_3\ a_4\ a_5$ | $a_6\ a_7\ a_8\ a_9$

s_1: a a c g c | − − t a c g

s_2: a t c g c | a a t a − g

$b_1\ b_2\ b_3\ b_4\ b_5$ | $b_6\ b_7\ b_8\ b_9\ b_{10}$

j^*

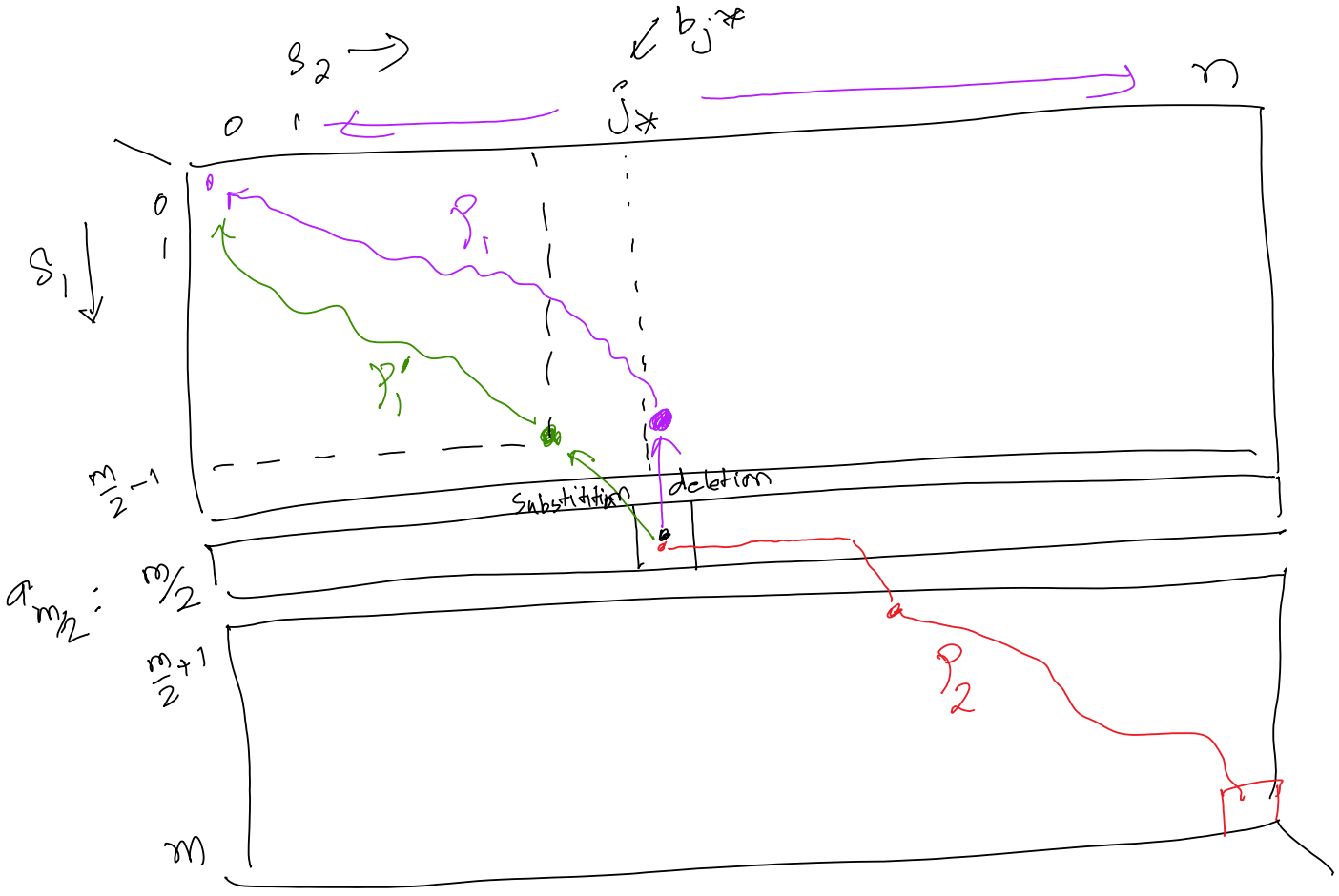
P : P_1 P_2

$(s_1[a_1 \ldots a_5], s_2[b_1 \ldots b_5])$

$(s_1[a_6 \ldots a_9], s_2[b_6 \ldots b_{10}])$

Oracle for j*

Friday, February 12, 2021 10:49 AM



Our optimal path

$$p : \begin{cases} p_1 \circ \begin{pmatrix} a_{m/2} \\ - \end{pmatrix} \circ p_2 \\ p_1' \circ \begin{pmatrix} a_{m/2} \\ b_{j^*} \end{pmatrix} \circ p_2 \end{cases} \text{(or)}$$

$$\text{Score}(p) = \max \begin{cases} \text{score}(p_1) + g + \text{score}(p_2) \\ \text{score}(p_1') + \delta(a_{m/2}, b_{j^*}) + \text{score}(p_2) \end{cases}$$

Oracle for j*

Friday, February 12, 2021 11:32 AM

Hirschberg technique (1975)

At row $m/2$:

Evaluate all possible j and figure out j^*

---

for $j = 0$ to n {

$Score_1 = Score(S_1[1..\frac{m}{2}-1], S_2[1..j]) + s(a_{\frac{m}{2}}, b_j) + Score(S_1[\frac{m}{2}+1..m], S_2[j+1..n])$

$Score_2 = Score(S_1[1..\frac{m}{2}-1], S_2[1..j]) + g + Score(S_1[\frac{m}{2}+1..m], S_2[j+1..n])$

$Score_j = \max\{Score_1, Score_2\}$

}

$O(mn)$

$j^* = Arg\,Max_j(Score_j) \quad 0 \le j \le n$