Sequence Alignment Problem

Dynamic Programming

A DNA strand consists of a string of molecules called bases:

Adenine	A
Cytosine	C
Guanine	G
Thymine	T

S_1 = A C C G G T C G

S_2 = C C A G G T G G C

A C C _ G G T C G _

_ C C A G G T G G C

3 gaps

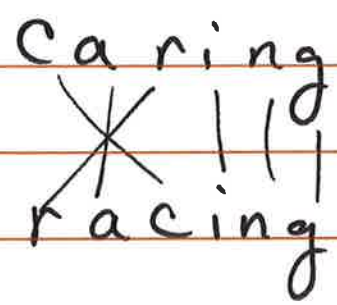
mismatch

Suppose we have 2 strings X & Y

$X = \{x_1 x_2 \ldots x_m\}$

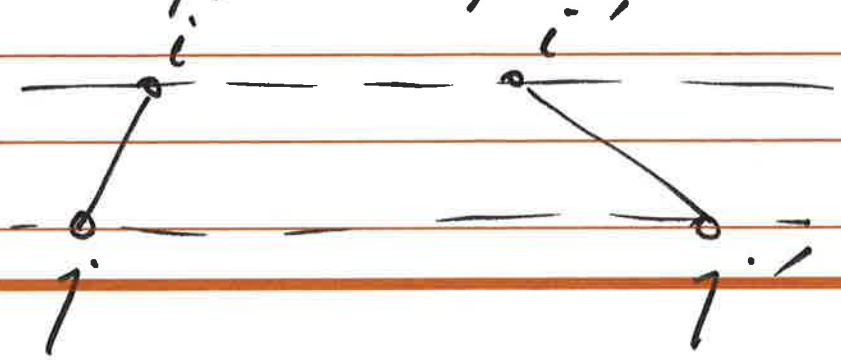
$Y = \{y_1 y_2 \ldots y_n\}$

<u>Def.</u> A matching is a set of ordered pairs w/ property that each item occurs at most once.



<u>Def.</u> A matching is an alignment if there are no crossing pairs

$(i,j), (i',j') \in M$ & $i < i'$ $\Rightarrow j < j'$



For a given alignment M between X & Y

1- We incur a "gap penalty" of δ (cost of δ)

2- For each mismatch (of letters q & p) we incur a mismatch cost α_{pq}

	A	C	G	T
A	0	α	α	α
C		0	α	α
G			0	α
T				0

<u>Similarity</u> between strings X and Y

is the Min. Cost of an alignment between X and Y

$X = \{x_1 \ldots x_m\}$

$Y = \{y_1 \ldots y_n\}$

either $(x_m, y_n) \in M$ or $(x_m, y_n) \notin M$

Define $OPT(i, j)$ as the min. cost of an alignment between $x_1 \ldots x_i$ & $y_1 \ldots y_j$

In an opt. alignment M, at least one of the following is true:

1- $(x_m, y_n) \in M \Rightarrow$

$$OPT(m,n) = OPT(m-1, n-1) + \alpha_{x_m y_n}$$

2- x_m is not matched $\Rightarrow$

$$OPT(m,n) = OPT(m-1, n) + \delta$$

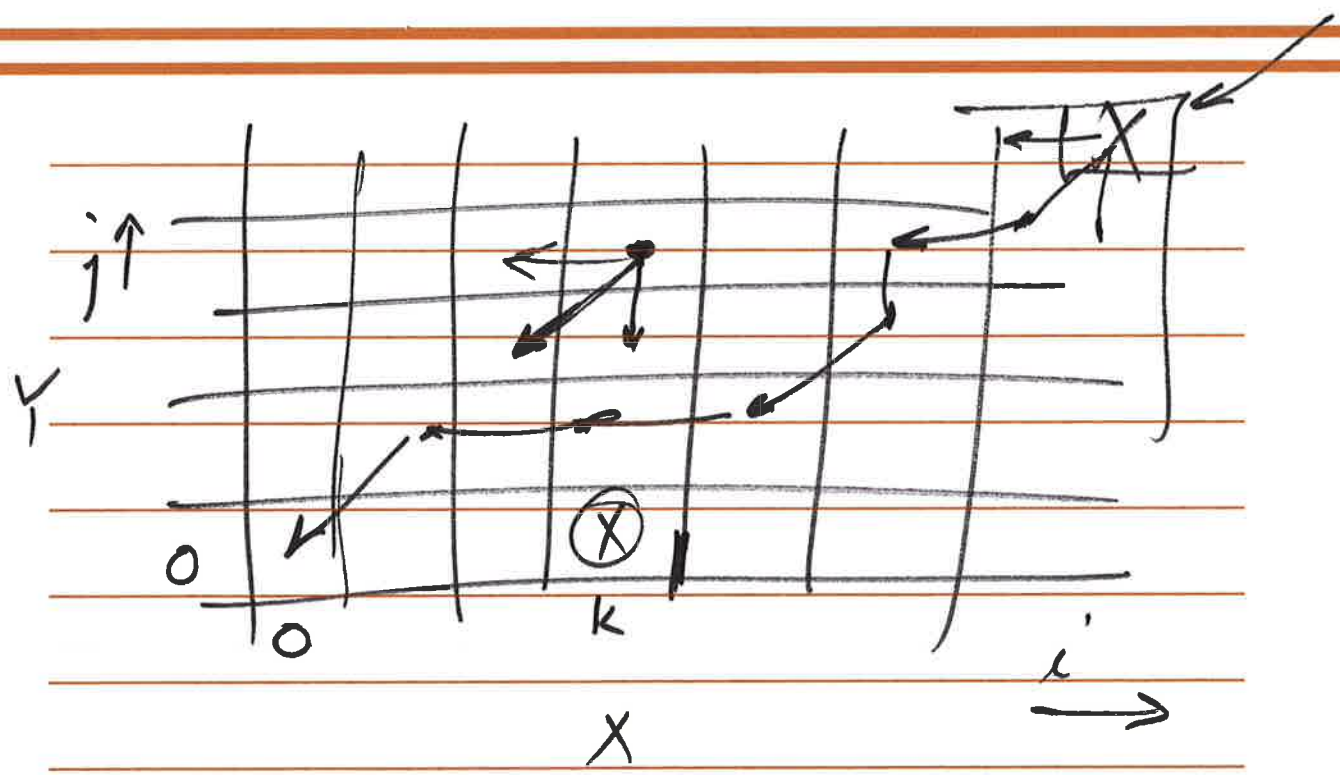
3- y_n is not matched $\Rightarrow$

$$OPT(m,n) = OPT(m, n-1) + \delta$$

$$OPT(i,j) = \min \Big[\alpha_{x_i y_j} + OPT(i-1, j-1),$$

$$\delta + OPT(i-1, j),$$

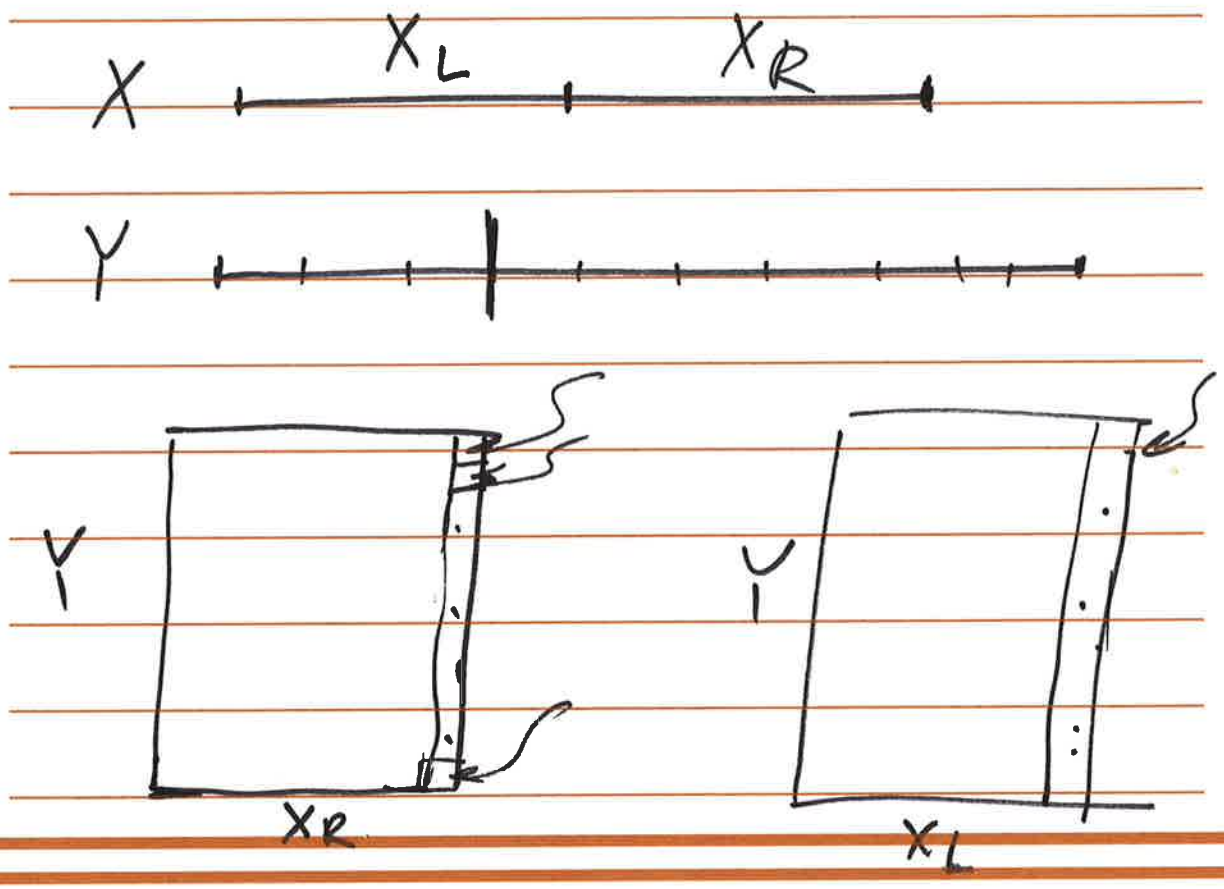
$$\delta + OPT(i, j-1) \Big]$$



$$10^7 \times 10^7 \times 4 = 4 \times 10^{14}$$

```
Alignment(x, y)
    Initialize A[i,0] = iδ for each i
               "       A[0,j] = jδ   "   "  j
for j=1 to n
    for i=1 to m
        A[i,j] = Min (α_{x_i y_j} + A[i-1, j-1],
                      δ + A[i-1, j],
                      δ + A[i, j-1] )
    endfor
endfor
return A[m,n]
```

takes $O(mn)$


X
X_L
X_R
Y
Y
X_R
Y
X_L