Local alignment

Given a string **P** (“pattern”) of length **m** and a string **T** (“text”) of length **n**, find substrings **a** and **b** of **P** and **T** respectively having maximal optimal global alignment score.

Example:

P = xyaxbacsll, **T** = pqraxabcstvq

<table>
<tr><th>Event</th><th>Example</th><th>Penalty</th></tr>
<tr><td>Match</td><td>TTG
|||
TTG</td><td>+15</td></tr>
<tr><td>Mismatch</td><td>GTG
| |
GAG</td><td>-30</td></tr>
<tr><td>Gap in text</td><td>ATC
| |
A-C</td><td>-30</td></tr>
<tr><td>Gap in pattern</td><td>C-G
| |
CCG</td><td>-30</td></tr>
</table>

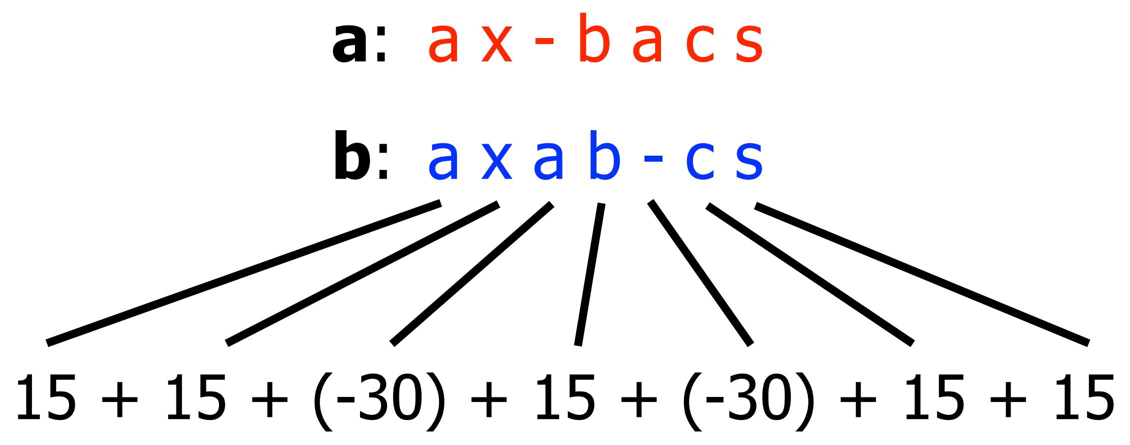
Local alignment

Given a string **P** (“pattern”) of length **m** and a string **T** (“text”) of length **n**, find substrings **a** and **b** of **P** and **T** respectively having maximal optimal global alignment score.

Example:

P = xyaxbacsll, **T** = pqraxabcstvq

Answer:



a: a x - b a c s

b: a x a b - c s

15 + 15 + (-30) + 15 + (-30) + 15 + 15

Score = 15

Event	Example	Penalty
Match	TTG ||| TTG	+15
Mismatch	GTG | | GAG	-30
Gap in text	ATC | | A–C	-30
Gap in pattern	C–G | | CCG	-30

Smith-Waterman

Text

Pattern

	-	c	c	c	t	t	c	c	t	t	a	c	g	c	g	a	c	c	c	a
-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
t	0																			
c	0																			
t	0																			
t	0																			
a	0																			
c	0																			
g	0																			
a	0																			
c	0																			

Event	**Penalty**
Match	+15
Mismatch	-30
Gap in text	-40
Gap in pattern	-40

Smith-Waterman

When filling in a cell...

$$H(i,j) = \max \begin{cases} 0 \\ H(i-1,j-1) + \ w(a_i,b_j) \\ H(i-1,j) + \ w(a_i,-) \\ H(i,j-1) + \ w(-,b_j) \end{cases}$$

Text

Pattern

	-	c	c	c	t	t	c	c	t	t	a	c	g	c	g	a	c	c	c	a
-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
t	0																			
c	0																			
t	0																			
t	0																			
a	0																			
c	0																			
g	0																			
a	0																			
c	0																			

$w(a_i, b_j)$ Match or mismatch

$w(a_i, -)$ Text gap

$w(-, b_j)$ Pattern gap

Event	Penalty
Match	+15
Mismatch	-30
Gap in text	-40
Gap in pattern	-40

Smith-Waterman

When filling in a cell...

$$H(i,j) = \max \begin{cases} 0 \\ \boxed{H(i-1,j-1) + \ w(a_i,b_j)} \\ H(i-1,j) + \ w(a_i,-) \\ H(i,j-1) + \ w(-,b_j) \end{cases}$$

Consider diagonal move from upper-left (match or mismatch)

Text

Pattern

	-	c	c	c	t	t	c	c	t	t	a	c	g	c	g	a	c	c	c	a
-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
t	0																			
c	0																			
t	0																			
t	0																			
a	0																			
c	0																			
g	0																			
a	0																			
c	0																			

$w(a_i, b_j)$	Match or mismatch
$w(a_i, -)$	Text gap
$w(-, b_j)$	Pattern gap

Event	Penalty
Match	+15
Mismatch	-30
Gap in text	-40
Gap in pattern	-40

Smith-Waterman

When filling in a cell...

$$H(i,j) = \max \begin{cases} 0 \\ H(i-1,j-1) + w(a_i,b_j) \\ H(i-1,j) + w(a_i,-) \\ H(i,j-1) + w(-,b_j) \end{cases}$$

Consider diagonal move from upper-left (match or mismatch)

Consider vertical move from above (gap in text)

Text

Pattern

	-	c	c	c	t	t	c	c	t	t	a	c	g	c	g	a	c	c	c	a
-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
t	0																			
c	0																			
t	0																			
t	0																			
a	0																			
c	0																			
g	0																			
a	0																			
c	0																			

$w(a_i, b_j)$ Match or mismatch

$w(a_i, -)$ Text gap

$w(-, b_j)$ Pattern gap

Event	Penalty
Match	+15
Mismatch	-30
Gap in text	-40
Gap in pattern	-40

Smith-Waterman

When filling in a cell...

$$H(i,j) = \max \begin{cases} 0 \\ H(i-1,j-1) + w(a_i,b_j) \\ H(i-1,j) + w(a_i,-) \\ H(i,j-1) + w(-,b_j) \end{cases}$$

Consider diagonal move from upper-left (match or mismatch)

Consider vertical move from above (gap in text)

Consider horizontal move from left (gap in pattern)

Text

Pattern

	-	c	c	c	t	t	c	c	t	t	a	c	g	c	g	a	c	c	c	a
-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
t	0																			
c	0																			
t	0																			
t	0																			
a	0																			
c	0																			
g	0																			
a	0																			
c	0																			

$w(a_i, b_j)$ Match or mismatch

$w(a_i, -)$ Text gap

$w(-, b_j)$ Pattern gap

Event	**Penalty**
Match	+15
Mismatch	-30
Gap in text	-40
Gap in pattern	-40

Smith-Waterman

When filling in a cell...

$$H(i,j) = \max \begin{cases} 0 \\ H(i-1,j-1) + w(a_i,b_j) \\ H(i-1,j) + w(a_i,-) \\ H(i,j-1) + w(-,b_j) \end{cases}$$

Consider diagonal move from upper-left (match or mismatch)

Consider vertical move from above (gap in text)

Consider horizontal move from left (gap in pattern)

Text

Pattern

	-	c	c	c	t	t	c	c	t	t	a	c	g	c	g	a	c	c	c	a
-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
t	0	0	0	0	15	15	0	0	15	15	0	0	0	0	0	0	0	0	0	0
c	0	15	15	15	0	0	30	15	0	0	0	15	0	15	0	0	15	15	15	0
t	0	0	0	0	30	15	0	0	30	15	0	0	0	0	0	0	0	0	0	0
t	0	0	0	0	15	45	5	0	15	45	5	0	0	0	0	0	0	0	0	0
a	0	0	0	0	0	5	15	0	0	5	60	20	0	0	0	15	0	0	0	15
c	0	15	15	15	0	0	20	30	0	0	20	75	35	15	0	0	30	15	15	0
g	0	0	0	0	0	0	0	0	0	0	0	35	90	50	30	0	0	0	0	0
a	0	0	0	0	0	0	0	0	0	0	15	0	50	60	20	45	5	0	0	15
c	0	15	15	15	0	0	15	15	0	0	0	30	10	65	30	5	60	20	15	0

$w(a_i,b_j)$	Match or mismatch
$w(a_i,-)$	Text gap
$w(-,b_j)$	Pattern gap

Event	Penalty
Match	+15
Mismatch	-30
Gap in text	-40
Gap in pattern	-40

Smith-Waterman

```
         c t t a c g
Answer:  | | | | | |        Score: 90
         c t t a c g
```

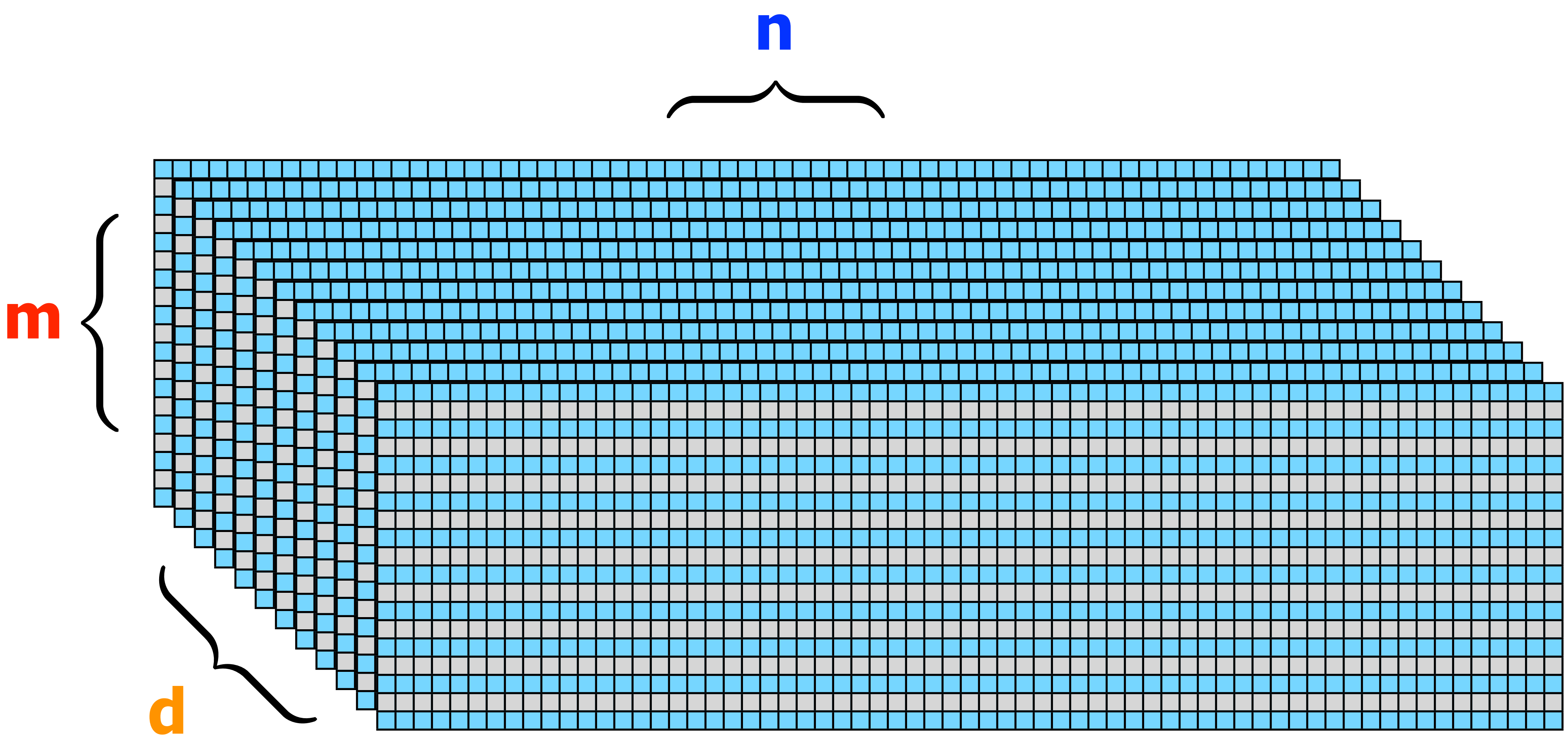
Text

Pattern

	-	c	c	c	t	t	c	c	t	t	a	c	g	c	g	a	c	c	c	a
-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
t	0	0	0	0	15	15	0	0	15	15	0	0	0	0	0	0	0	0	0	0
c	0	15	15	15	0	0	30	**15**	0	0	0	15	0	15	0	0	15	15	15	0
t	0	0	0	0	30	15	0	0	**30**	15	0	0	0	0	0	0	0	0	0	0
t	0	0	0	0	15	45	5	0	15	**45**	5	0	0	0	0	0	0	0	0	0
a	0	0	0	0	0	5	15	0	0	5	**60**	20	0	0	0	15	0	0	0	15
c	0	15	15	15	0	0	20	30	0	0	20	**75**	35	15	0	0	30	15	15	0
g	0	0	0	0	0	0	0	0	0	0	0	35	**90**	50	30	0	0	0	0	0
a	0	0	0	0	0	0	0	0	0	0	15	0	50	60	20	45	5	0	0	15
c	0	15	15	15	0	0	15	15	0	0	0	30	10	65	30	5	60	20	15	0

Smith-Waterman

Alignining **d** reads of length **m** to reference of length **n** is O(**dmn**)



Dynamic programming alignment

$$\left.\begin{array}{l} \mathbf{d} = 6 \text{ x } 10^9 \text{ reads} \\ \mathbf{m} = 100 \text{ nt} \end{array}\right\} \approx 1 \text{ week-long run of}$$

$\mathbf{n}$ = 3 x 10^9 nt ≈ human

Illumina HiSeq 2000

Total of $\mathbf{d}$ x $\mathbf{m}$ x $\mathbf{n}$ = 2 x 10^{21} cell updates

100 processors, each capable of 100 billion cell updates per second, would take >6 years