CS 68: BIOINFORMATICS

Prof. Sara Mathieson
Swarthmore College
Spring 2018

Outline: Feb 9

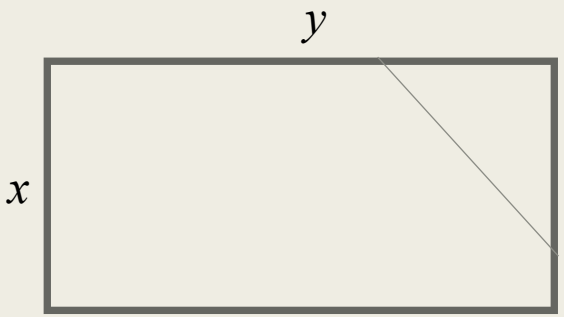
- Finish: local alignment variations
- Begin: Burrows-Wheeler Transform (BWT)
- Application to read mapping

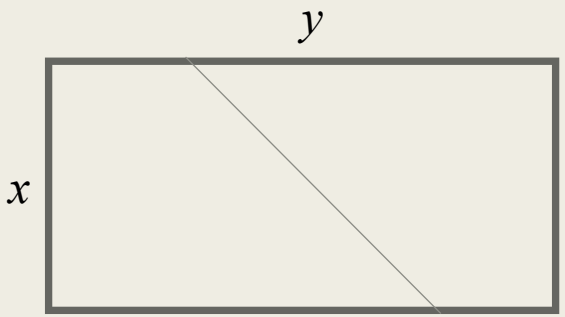
Notes:

- Reading posted
- Feb 15: Lab 4 (BWT and read mapping)
- Feb 22: practice midterm (I am at SIGCSE)
- Mar 1: midterm 1
- Lab 1 returned on Monday

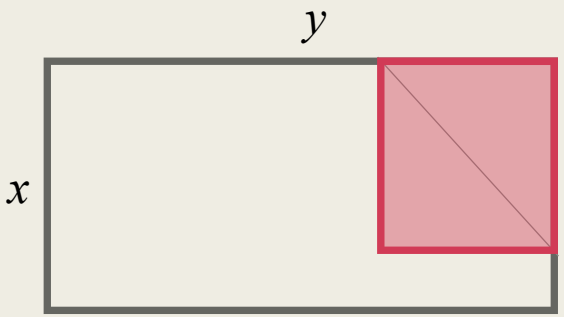
Local alignment variations

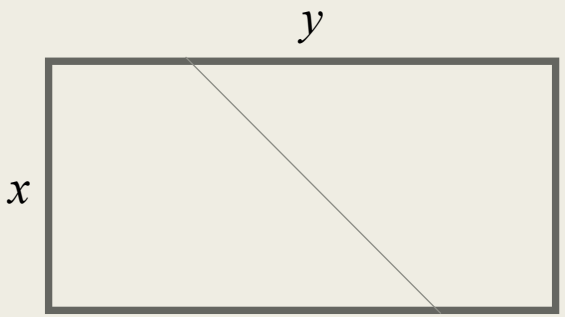
Handout 6: what portion of x aligns to what portion of y?



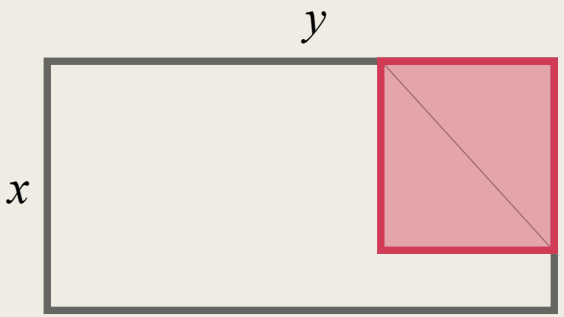


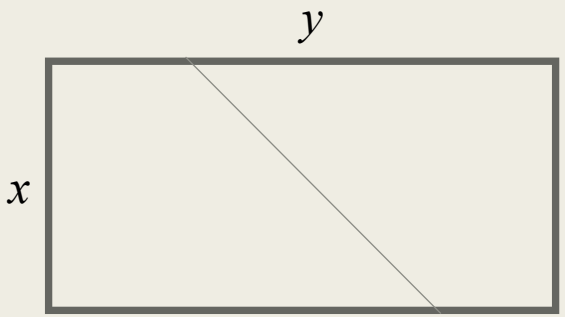
Handout 6: what portion of x aligns to what portion of y?



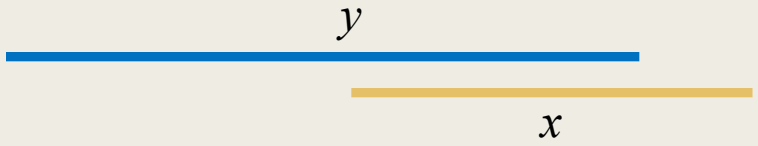


Handout 6: what portion of x aligns to what portion of y?

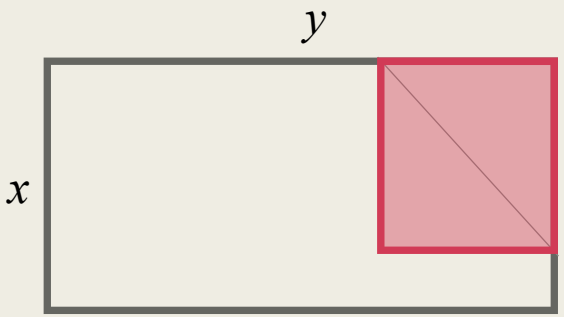


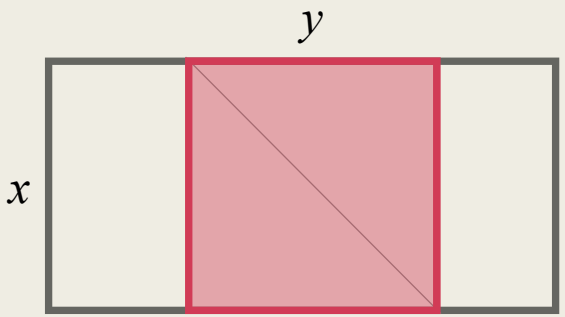


Beginning of *x* with end of *y*

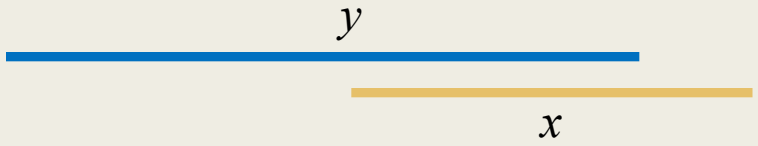


Handout 6: what portion of x aligns to what portion of y?

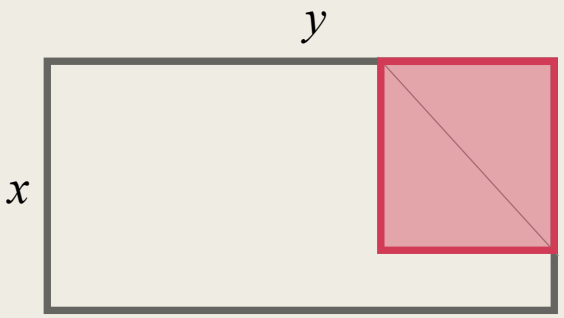




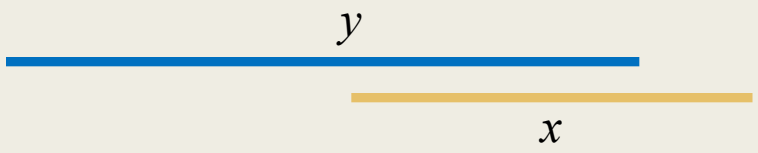
Beginning of *x* with end of *y*

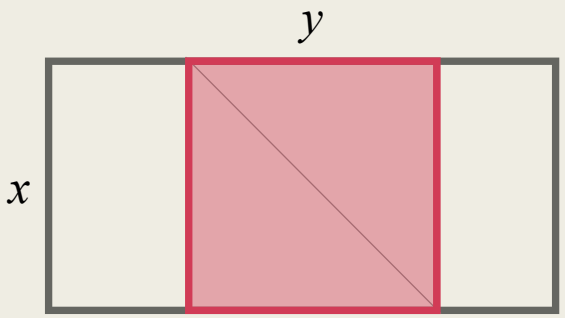


Handout 6: what portion of x aligns to what portion of y?

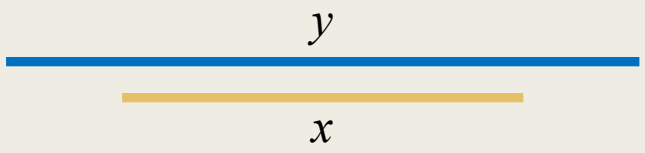


Beginning of *x* with end of *y*

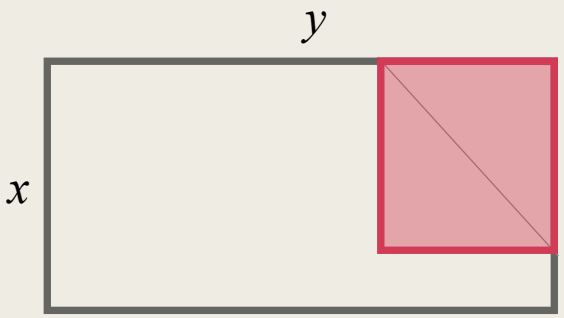




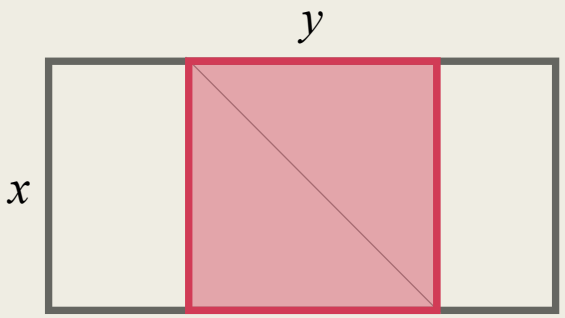
x aligns with the middle of *y*



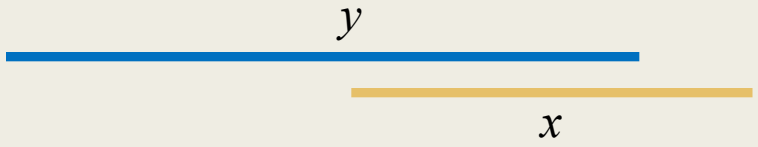
Handout 6: what portion of x aligns to what portion of y?

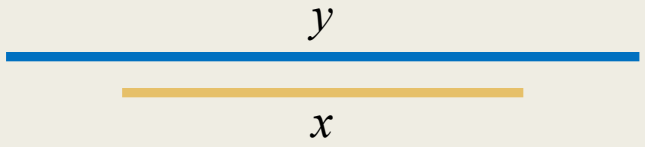


Beginning of *x* with end of *y*



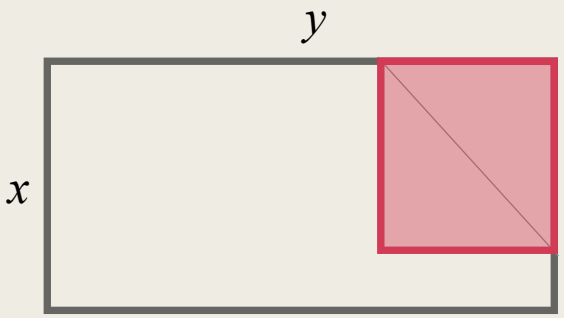
x aligns with the middle of *y*



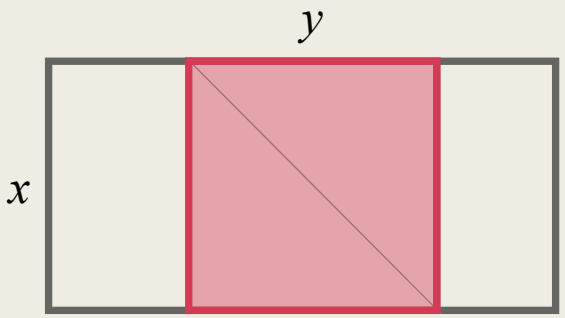


<u>DP algorithm modifications</u>:

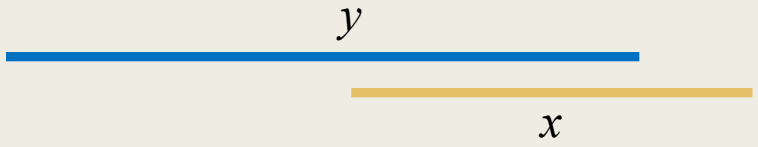
Handout 6: what portion of x aligns to what portion of y?

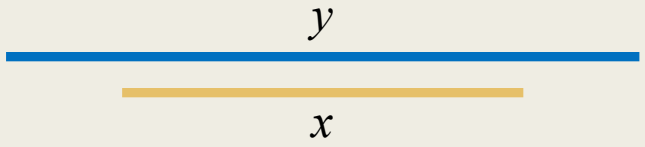


Beginning of *x* with end of *y*



x aligns with the middle of *y*

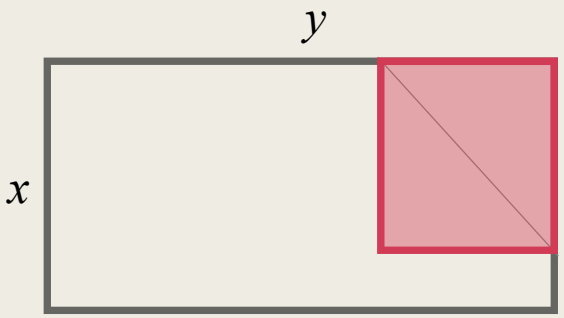




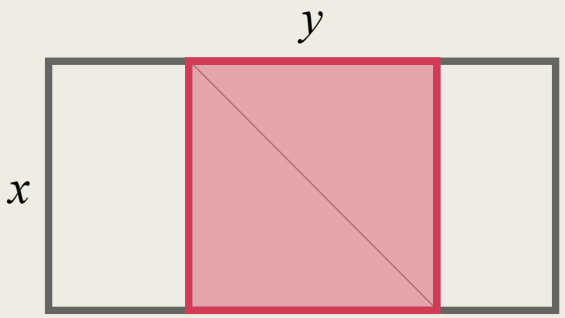
<u>DP algorithm modifications</u>:

1) Initialization: *0th row and 0th column with 0's to not penalize leading/trailing gaps*

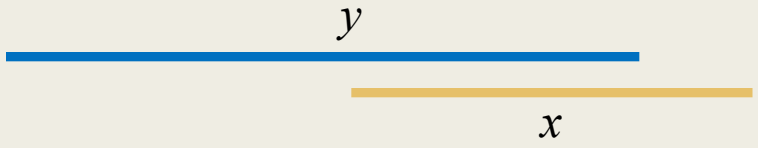
Handout 6: what portion of x aligns to what portion of y?

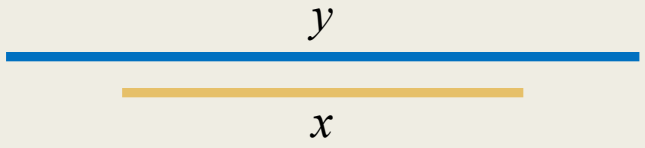


Beginning of *x* with end of *y*



x aligns with the middle of *y*

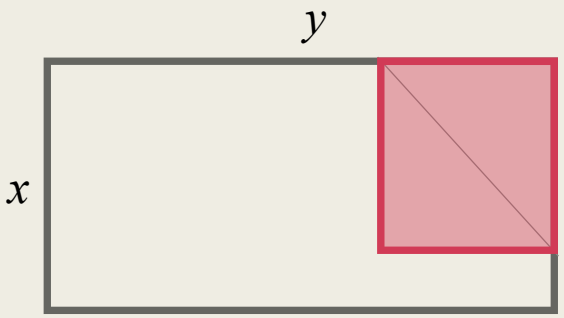




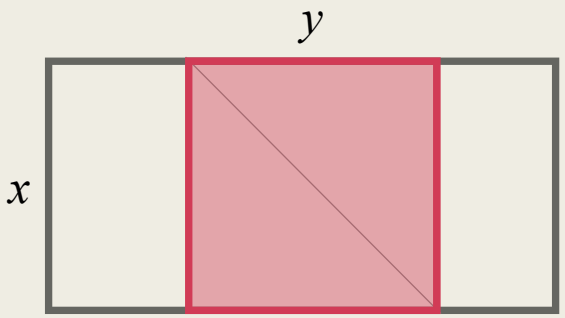
<u>DP algorithm modifications</u>:

1) Initialization: *0th row and 0th column with 0's to not penalize leading/trailing gaps*
2) Recursion: *to fill in the rest of the table, use global alignment (i.e. don't restart at 0)*

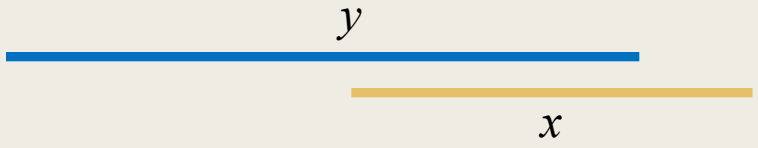
Handout 6: what portion of x aligns to what portion of y?

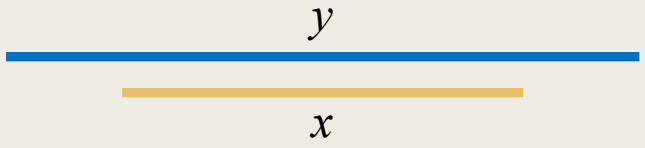


Beginning of *x* with end of *y*



x aligns with the middle of *y*





<u>DP algorithm modifications</u>:

1) Initialization: *0th row and 0th column with 0's to not penalize leading/trailing gaps*
2) Recursion: *to fill in the rest of the table, use global alignment (i.e. don't restart at 0)*
3) Traceback: *start at the maximum value along the last row or last column*

Next topic: read mapping with BWT

What if y is very long and x is very short?

- Exactly the case if **y = entire genome** and **x = single read**

- Reading mapping: Given an already assembled reference genome and reads from a newly sequence individual of the same species, what is the position of each read?

- Three options:
  - *1) Read aligns perfectly with reference*
  - *2) Read aligns with a few differences (representing population-level variation)*
  - *3) Read does not align at all (insertion in the newly sequenced individual)*

Read Mapping

n = # reads

L = length of read

reads

reference → length = G

insertion

	G
L	$O(G \cdot L)$

runtime

$\boxed{O(L \cdot G \cdot n)}$

L: 50–100

G: ≈ billions

n: billions

BWT

S = banana$

- special char
- alphabetically first

BWT(S) = CCCCAAAAAABB

S = ABCABC

L

6·A + 12·B + 6

$\pi(S)$	rank	$\pi^{sorted}(S)$	$BW(S)$
banana$	5	$banana	a
anana$b	4	a$banan	n
nana$ba	7	ana$ban	n
ana$ban	3	anana$b	b
na$bana	6	banana$	$
a$banan	2	na$bana	a
$banana	1	nana$ba	a

all cyclic permutations of S

F (first)

L (last)

repeat

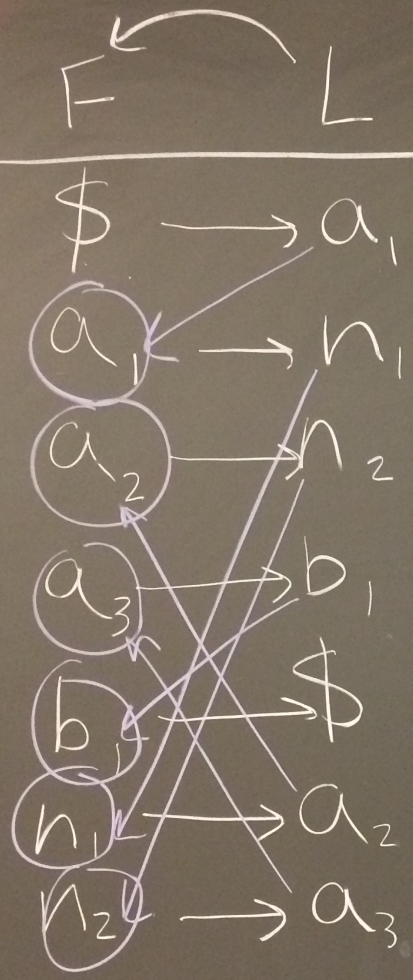
anana

repeat

repeat

B

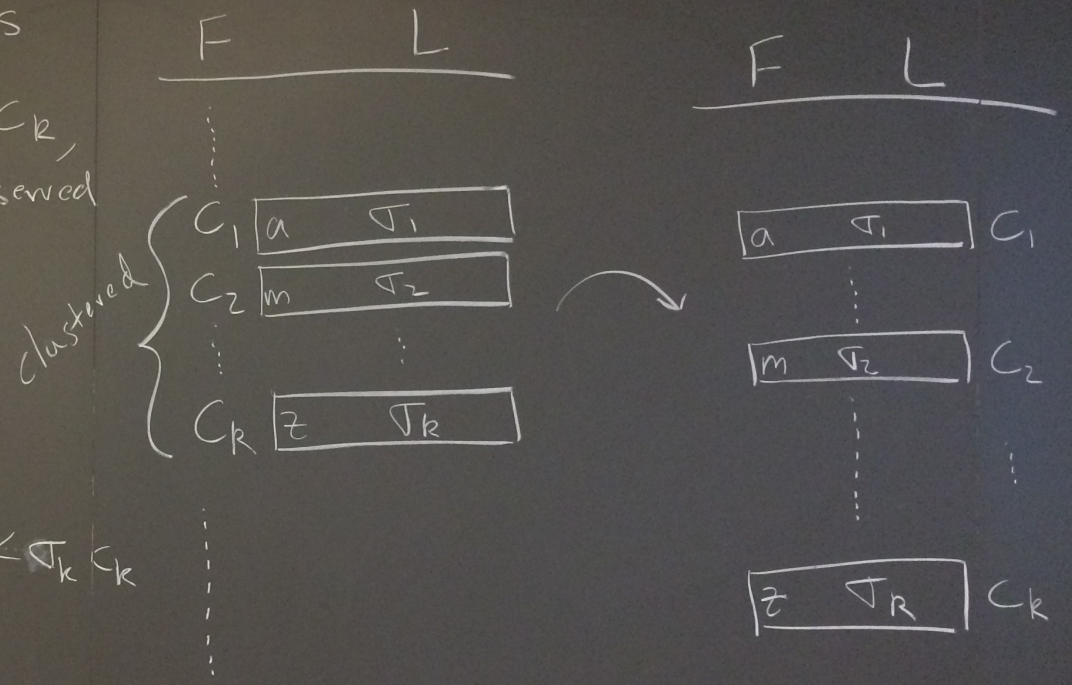
G. A



backtrace	reconstructs
$\$ \to a_1$	a$
$a_1 \to n_1$	na$
$n_1 \to a_2$	ana$
$a_2 \to n_2$	nana$
$n_2 \to a_3$	anana$
$a_3 \to b_1$	banana$
$b_1 \to \$$	

STOP

Claim If F has k copies of char c : $c_1, c_2, \cdots c_k$, the their order is preserved in L



$$\sigma_1 < \sigma_2 < \cdots \sigma_k \iff \sigma_1 c_1 < \sigma_2 c_2 < \sigma_k c_k$$