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

DP algorithm modifications:
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 $y$**

DP algorithm modifications:

1) **Initialization:** $0^{th}$ row and $0^{th}$ column with 0’s to not penalize leading/trailing gaps
Handout 6: what portion of $x$ aligns to what portion of $y$?

DP algorithm modifications:

1) **Initialization**: 0\textsuperscript{th} row and 0\textsuperscript{th} 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 $y$?

DP algorithm modifications:

1) **Initialization**: $0^{th}$ row and $0^{th}$ 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 \( y = \text{entire genome} \) and \( x = \text{single read} \)

- Reading mapping: Given an already assembled reference genome and reads from a newly sequenced 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

- \( n = \# \text{ reads} \)
- \( L = \text{length of read} \)
- \( G \) reference length = \( G \)
- \( L \) reads
- \( O(G \cdot L) \)

BWT

- \( S = \text{banana} \)
- special char alphabetically first
- runtime \( \approx \) billions
- \( 50-100 \) billions
- \( CCCCAAAAAAABBB \)
- \( S = ABCABC \)
- \( 6A + 12B + G \)
<table>
<thead>
<tr>
<th>$T(I(S))$</th>
<th>rank</th>
<th>$T(\text{solved}(S))$</th>
<th>$B(M(S))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>banana $</td>
<td>$ 5</td>
<td>$\text{banana}$</td>
<td>a</td>
</tr>
<tr>
<td>anana $</td>
<td>$ 4</td>
<td>$a$ $\text{banan}$</td>
<td>n</td>
</tr>
<tr>
<td>nana $</td>
<td>$ 4</td>
<td>$a$ $\text{banan}$</td>
<td>n</td>
</tr>
<tr>
<td>ana $</td>
<td>$ 3</td>
<td>$a$ $\text{banan}$</td>
<td>b</td>
</tr>
<tr>
<td>na $</td>
<td>$ 6</td>
<td>$\text{banana}$</td>
<td>$\text{banana}$</td>
</tr>
<tr>
<td>a $</td>
<td>$ 2</td>
<td>$\text{banana}$</td>
<td>$\text{banana}$</td>
</tr>
<tr>
<td>$\text{banana}$</td>
<td>1</td>
<td>$\text{banana}$</td>
<td>a</td>
</tr>
</tbody>
</table>

All cyclic permutations of $S$: $\text{banan} a$
<table>
<thead>
<tr>
<th>$F$</th>
<th>$L$</th>
<th>backtrace</th>
<th>reconstruct S</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rightarrow a_1$</td>
<td>$\rightarrow a_1$</td>
<td>$a$</td>
<td></td>
</tr>
<tr>
<td>$a_1 \rightarrow n_1$</td>
<td>$a_1 \rightarrow n_1$</td>
<td>na$ $</td>
<td></td>
</tr>
<tr>
<td>$a_2 \rightarrow n_2$</td>
<td>$a_2 \rightarrow n_2$</td>
<td>ana$ $</td>
<td></td>
</tr>
<tr>
<td>$a_3 \rightarrow b_1$</td>
<td>$a_3 \rightarrow b_1$</td>
<td>anana$ $</td>
<td></td>
</tr>
<tr>
<td>$b_1 \rightarrow $</td>
<td>$b_1 \rightarrow $</td>
<td>banana$</td>
<td></td>
</tr>
</tbody>
</table>

STOP
Claim: If $F$ has $k$ copies of char $C : C_1, C_2, \ldots, C_k$, the their order is preserved in $L$. 

$T_1 < T_2 < \ldots < T_k \iff T_1 < T_2 < \ldots < T_k$