CS466: Introduction to Bioinformatics Name: Payal Mantri

Problem Set 4

Handed out: November 9, 2022 *Due: Nov 17, 2022*

Instructions: This homework assignment consists of two questions worth a total of 50 points.
Do not forget to write your name at the top!

1. **Assembly I** [25 points]

   a. Compute for each permutation of the set $S = \{\text{ABB, BAB, ABA}\}$ the corresponding shortest common superstring (SCS) respecting the order prescribed by the permutation. Indicate the permutation(s) with overall shortest common superstring. [10 points]

   Hint: There are $3! = 6$ permutations of S.

   i. $S = \{\text{ABB}, \underline{\text{B}}\text{AB}, \underline{\text{AB}}\text{A}\}$
   $SCS(S) = ABBABA \Rightarrow length = 6$

   ii. $S = \{\text{ABB}, \text{ABA}\ , \underline{\text{BA}}\text{B}\ \}$
   $SCS(S) = ABBABAB \Rightarrow length = 7$

   iii. $S = \{\text{BAB}, \underline{\text{AB}}\text{B}\ , \text{ABA}\ \}$
   $SCS(S) = BABBABA \Rightarrow length = 7$

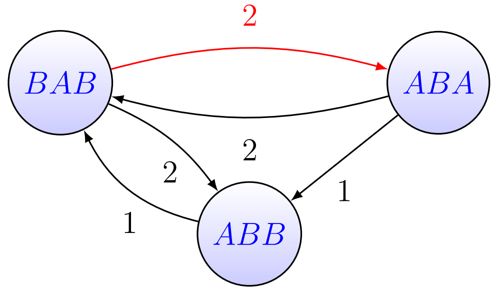
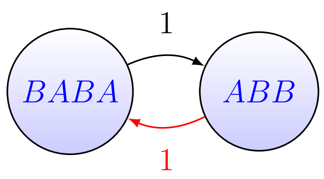
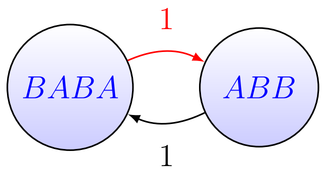
   iv. $S = \{\text{BAB}, \underline{\text{AB}}\text{A}\ , \underline{\text{A}}\text{BB}\ \}$
   $SCS(S) = BABABB \Rightarrow length = 6$

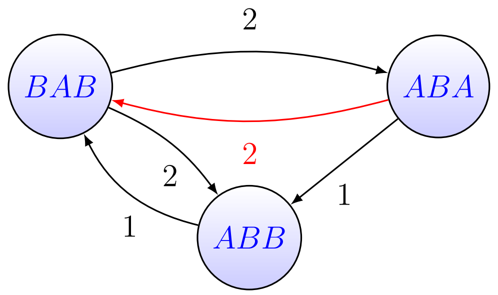
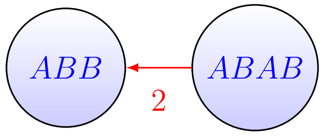
   v. $S = \{\text{ABA}, \underline{\text{A}}\text{BB}\ , \underline{\text{B}}\text{AB}\ \}$
   $SCS(S) = ABABBAB \Rightarrow length = 7$

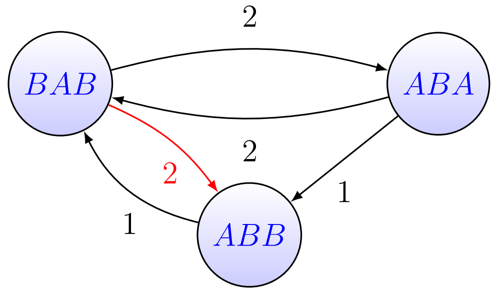
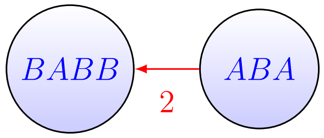
   vi. $S = \{\text{ABA}, \underline{\text{BA}}\text{B}\ , \underline{\text{AB}}\text{B}\ \}$
   $SCS(S) = ABABB \Rightarrow length = 5$

   Overall shortest common substring $SCS(S) = ABABB \Rightarrow length = 5$

b. Consider the same set $S = \{\text{ABB, BAB, ABA}\}$ as in the previous question. Use the greedy heuristic to approximate the SCS problem. Give the edge-weighted directed graph for each step of the greedy heuristic. In the case of multiple edges with the same maximum weight, enumerate all solutions. [10 points]

		SCS(S)= ABBABA length = 6
		SCS(S)= BABABB length = 6

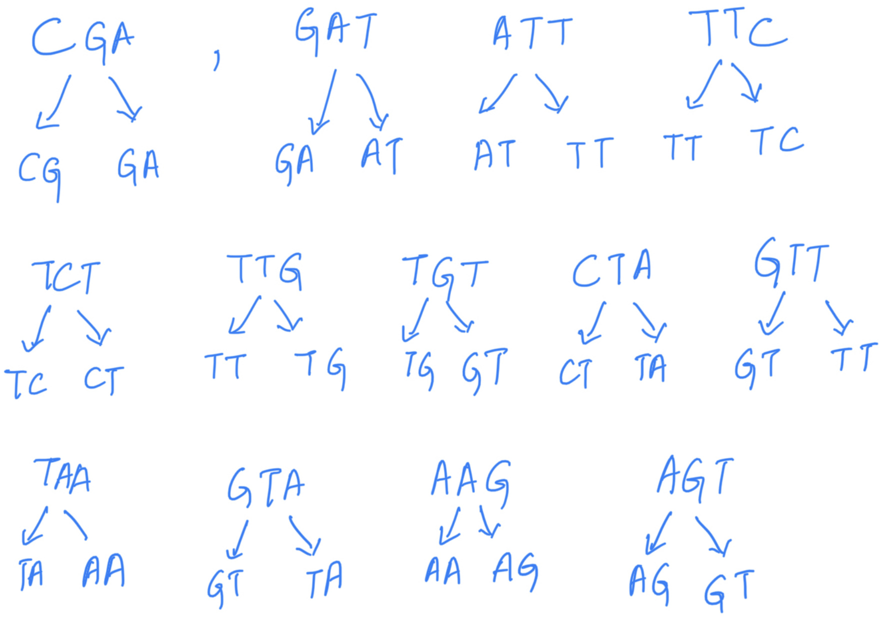
		SCS(S)= ABABB length = 5

		SCS(S)= ABABB length = 5

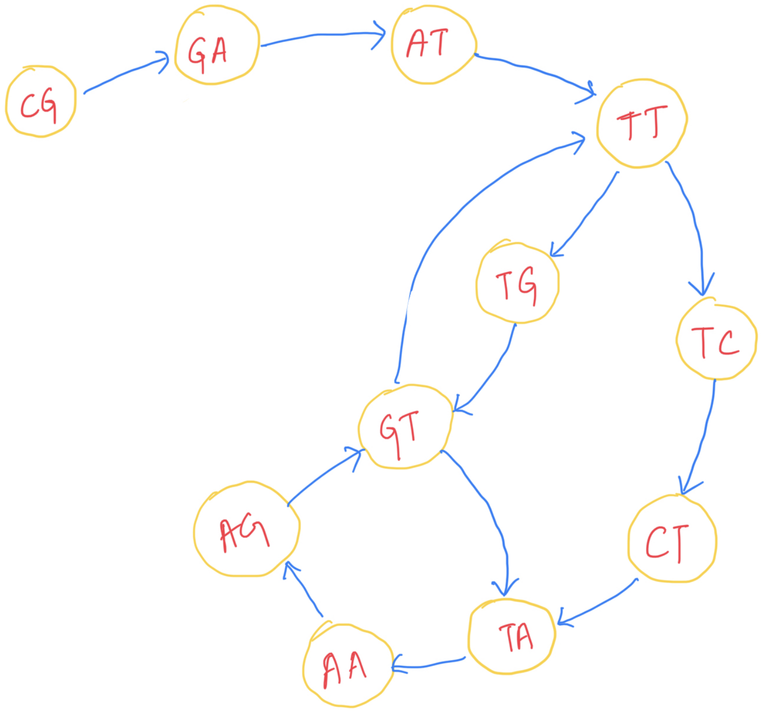
Overall shortest common substring $SCS(S) = ABABB \Rightarrow length = 5$

c. Consider the following 3-mer set $S = \{$CGA, GAT, ATT, TTC, TCT, TTG, TGT, CTA, GTT, TAA, GTA, AAG, AGT$\}$. Construct the De Bruijn graph, identify an Eulerian walk and reconstruct the corresponding assembly. [5 points]

Solution:
Get the 2-mers



De Bruin Graph



Eulerian Walks : There are 3 possible solutions

i. $CG \rightarrow GA \rightarrow AT \rightarrow TT \rightarrow TC \rightarrow CT \rightarrow TA \rightarrow AA \rightarrow AG \rightarrow GT \rightarrow TT \rightarrow TG \rightarrow GT \rightarrow TA$

Final String → CGATTCTAAGTTGTA

ii. $CG \rightarrow GA \rightarrow AT \rightarrow TT \rightarrow TG \rightarrow GT \rightarrow TT \rightarrow TC \rightarrow CT \rightarrow TA \rightarrow AA \rightarrow AG \rightarrow GT \rightarrow TA$

Final String → CGATTGTTCTAAGTA

iii. $CG \rightarrow GA \rightarrow AT \rightarrow TT \rightarrow TG \rightarrow GT \rightarrow TA \rightarrow AA \rightarrow AG \rightarrow GT \rightarrow TT \rightarrow TC \rightarrow CT \rightarrow TA$

Final String → CGATTGTAAGTTCTA

2. **Assembly II** [25 points]

Suppose you have the following set of sequenced reads $R = \{$`GTACTG, ACTTGT`$\}$.

a. Draw the De Bruijn graph for this set of reads with $k = 4$. [5 points]

Solution:

Get the 4-mers and split into left and right 3-mers

GTACTG , ACTTGT

4-mer

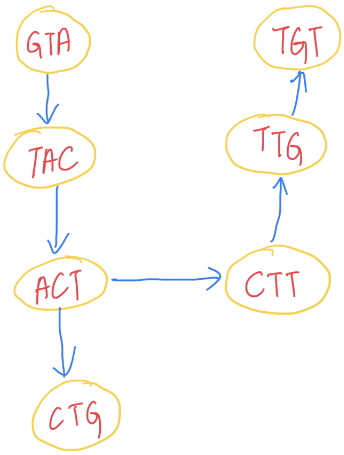
GTACTG ⇒ GTAC , TACT , ACTG

3-mer GTA TAC TAC ACT ACT CTG

ACTTGT ⇒ ACTT , CTTG , TTGT

3-mer ACT CTT CTT TTG TTG TGT

Build the De Bruijn Graph



b. Does a Eulerian walk exist in the graph your drew? If so, give the complete sequence it indicates for the underlying genome (or one if multiple Eulerian paths exist). If not, explain why not. [5 points]

A Node in graph is semi-balanced if indegree differs from outdegree by 1

We know that a directed, connected graph is Eulerian if and only if it has at most 2 semi-balanced nodes and all other nodes are balanced

In 2a, the De Bruijn Graph has 3 semi-balanced nodes GTA, ACT, TGT. Hence it doesn't have a Eulerian walk .

c. Draw the De Bruijn graph for this set of reads with $k = 3$. [10 points]
Solution:

Get the 3-mers and split into left and right 2-mers

GTACTG , ACTTGT

3 mers

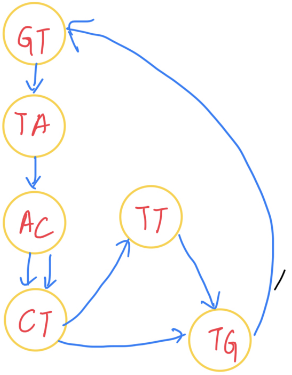
GTACTG ⇒ GTA , TAC, ACT, CTG

GT TA TA AC AC CT CT TG

ACTTGT ⇒ ACT , CTT , TTG , TGT

AC CT CT TT TT TG TG GT

Build the De Bruijn Graph



d. Does a Eulerian walk exist in the graph your drew? If so, give the complete sequence it indicates for the underlying genome (or one such sequence if multiple Eulerian paths exist). If not, explain why not. [5 points]

Since the De Bruin Graph in 2c has exactly two semibalanced nodes - AC and TG, there exists a Eulerian Walk in the graph.
We can have two following sequences

A. $AC \rightarrow CT \rightarrow TT \rightarrow TG \rightarrow GT \rightarrow TA \rightarrow AC \rightarrow CT \rightarrow TG$

Final String → ACTTGTACTG

B. $AC \rightarrow CT \rightarrow TG \rightarrow GT \rightarrow TA \rightarrow AC \rightarrow CT \rightarrow TT \rightarrow TG$

Final String → ACTGTACTTG