QBIO 478: Homework 3

Due on March 23rd Midnight [Submit to Blackboard]

March 8, 2021

Total points: 70 points

Question 1 *25 points*

Guide tree for MSA

Given below are five sequences:
S1: GATTCA; S2: GTCTGA; S3: GATATT, S4: GTCAGC, S5: GAGACA.
The pair-wise sequence alignments for S1-S4 are as follows:

S1:	**G**	**A**	**T**	**-**	**T**	**C**	**A**		**S1:**	**G**	**A**	**T**	**-**	**T**	**C**	**A**
S2:	**G**	**-**	**T**	**C**	**T**	**G**	**A**		**S3:**	**G**	**A**	**T**	**A**	**T**	**T**	**-**
S1:	**G**	**A**	**T**	**T**	**C**	**A**	**-**	**-**	**S2:**	**G**	**-**	**T**	**C**	**T**	**G**	**A**
S4:	**G**	**-**	**T**	**-**	**C**	**A**	**G**	**C**	**S3:**	**G**	**A**	**T**	**A**	**T**	**-**	**T**
S2:	**G**	**T**	**C**	**T**	**G**	**A**			**S3:**	**G**	**A**	**T**	**-**	**A**	**T**	**T**
S4:	**G**	**T**	**C**	**A**	**G**	**C**			**S4:**	**G**	**-**	**T**	**C**	**A**	**G**	**C**

(a) [10 points] Find pairwise alignment of S5 with other sequences. Use match score = 1 and -1 for mismatch and gap.

(b) [15 points] Construct a guide tree for MSA for these sequences.

Question 2 *20 points*

MSA of Alu repeats

Go to **UCSC Table Browser**. Find the DNA sequences (in FASTA format) for all repeats that belong to the **Alu** family, and are located within the genomic region chr12:56,735,000-56,750,000 of human GRCh38/hg38 assembly in RepeatMasker track (under repeats group). Select output format to be "sequence". Tutorial about Table Browser can be found here. Answer the following questions based on the output from the Table Browser.

(a) [8 points] How many Alu repeats are there in the result?

(b) [2 points] How many unique repeat names are there in the result?

(c) [10 points] Take any one sequence for the following repeats:
AluJb,AluJo,AluJr,AluSc,AluSg,AluSp,AluSx and AluSz. Use the EBI multiple sequence alignment tool **Clustal Omega** to align these sequences.

1. What is the percentage identity between AluJo and AluSp?
2. Find the phylogenetic tree based on the MSA.

Question 3 *25 points*

De Bruijn graph

(a) [5 points] Construct a De Bruijn graph for r_1 = CGATTCTAAGT for $k = 4$?

(b) [5 points] Now construct a De Bruijn graph for $k = 4$ given the information that there is a single-nucleotide variant (C/G) at position 6 of r_1. Hint: Now there are two possible sequences and you have to construct the graph for
r_1 = CGATTCTAAGT
r_2 = CGATTGTAAGT

(c) [5 points] How many extra nodes are present in the graph (b) compared to the graph in part (a)?

(d) [10 points] A 1Mb diploid genome is sequenced by a technology that produces perfect reads. There are no repeats in the genome longer than 25 bases, but there is a single-nucleotide variant every 1,000 bases. Assuming error-free sequencing by 100-base reads only from the forward strand and the spacing between two SNPs is greater than 50 bases, how many nodes are in the resulting De Bruijn graph built with $k = 27$?