QBIO 478: Homework 2

Due on February 23^{rd} Midnight [Submit to Blackboard]

February 21, 2021

Total points: 90 points

Question 1 *20 points*

(a) [5 points] List the suffixes of the string T = ATATAAT\$.

(b) [5 points] Draw the suffix trie for T. Include edge labels. Enumerate your nodes $1, 2, \ldots |V|$, where $|V|$ is the total number of nodes in the trie.

(c) [5 points] Repeated sequences in T are shared prefixes of any suffix of T. Using the trie, list the repeated substrings of T.

(d) [5 points] The depth of a node in a tree is the number of edges traversed on the path from the root to the node. A branching node is a parent node with more than one child. Can the branching node with the maximum depth among all branching nodes in a trie have a depth different from the longest repeat in T? Explain your answer.

Question 2 *25 points*

COVID-19 PCR primer

RT-qPCR is commonly used for clinical diagnosis of Coronavirus disease 2019 (COVID-19). The genome sequence of COVID-19 (MN908947.3) is available in the public database. Now you are given a set of short sequences:

seq1: CCCTGTGGGTTTTACACTTAA
seq2: TCATAGGCTGCGGTATCGGC
seq3: TTATAAACTACGGTATCGGC
seq4: ACGATTGTGCATCAGCTGA

Your task is to determine the potential primers for COVID-19 diagnosis among the four sequences above using the basic local alignment search tool (BLAST). A BLAST search enables a researcher to compare a subject protein or nucleotide sequence (called a query) with a library or database of sequences, and identify database sequences that resemble the query sequence above a certain threshold.

(a) [10 points] Perform BLAST search for the above sequences using COVID-19 genome as the database. Use Nucleotide BLAST and restrict the query to the COVID-19 genome by mentioning genome ID (MN908947.3) in the Entrez Query field. Make sure you enter the sequences in FASTA format. Report the query coverage (Query Cover) and percentage identity (Per. Ident) available on the result page for each of the sequences (seq 1-4).

(b) [10 points] Based on the results from the BLAST search, identify the two primer sequences.

(c) [5 points] Report the forward and reverse primer. [Hint: Use Graphics visualization for individual results under the Alignment tab to show the alignment]

Question 3 *20 points*

Create a substitution scoring matrix

The goal of this question is to learn how to create a scoring matrix for nucleotides (A,C,G,T) based on high confidence mutliple sequence alignment (MSA) block. We will use the same strategy followed by the **BLOSUM** scoring matrixs creation. These are the steps to be followed:

1. Calculate the probablity of occurence (p_i) for each nucleotide i, which is given by:

$$p_i = \frac{\text{number of occurences of nucleotide } i \text{ in the block}}{\text{total number of nucleotides in the block}}, \text{where } i = \text{A,C,G,T}.$$

   At the end of this step you will find: $\{p_A, p_C, p_G, p_T\}$.

2. Count pair frequencies $c_{ij}^{(k)}$ for each pair of nucleotide i and j, for each column k of the block. Note: if n sequences are aligned, there are $\binom{n}{2} = n(n-1)/2$ pairs for each column. Then sum the scores across all columns

$$c_{ij} = \sum_k c_{ij}^{(k)}.$$

   Enter the pair frequencies to fill up a lower triangle matrix. After this step your matrix will look like this (The entries shown here are not the actual values):

	A	**C**	**G**	**T**
A	15			
C	7	10		
G	2	3	8	
T	8	7	3	9

3. Normalize the pair frequencies so that they will sum to 1. If q_{ij} is the normalized pair frequency, then

$$q_{ij} = \frac{c_{ij}}{T},$$

   where $T = \sum_{i \geq j} c_{ij} = \frac{w \times n(n-1)}{2}$, where $w =$ number of columns, $n =$ number of sequences.

   For the above shown matrix, $T = 72$ and $q_{AC} = \frac{7}{72}$.

4. Calculate the expected probability for each (i, j) pair (assuming independence):

$$e_{ii} = p_i^2,$$
$$e_{ij} = p_i \times p_j + p_j \times p_i = 2 \times p_i \times p_j \ (i \neq j).$$

   p_i is the probablity calculated in step 1. For example, $e_{AC} = 2p_A \times p_C$.

5. Create log-odds score matrix. For each entry (i, j), the log odds score can be calculated by:

$$s_{ij} = \log_2(\frac{q_{ij}}{eij}).$$

6. Substitution matrix: The final score stored in the BLOSUM matrix is obtained by rounding to the nearest integer after multiplying by 2.

$$matrixScore = \text{(rounded) } 2 \times s_{ij}.$$

Yay finally we created the scoring matrix!

Now following steps 1-6, **create a substitution scoring matrix** for the given MSA block:

ACTGCC
AGTGAC
TCTTGT

Inorder to help you with the steps, the answers for some of the steps are partially completed below.

- $p_A = 0.167, p_C = 0.278, p_G =?, p_T =?$

- Frequency matrix:

	A	**C**	**G**	**T**
A	1			
C	1	2		
G	1	3	?	
T	?	?	?	?

- Normalized frequency matrix

	A	**C**	**G**	**T**
A	0.056			
C	0.056	0.111		
G	0.056	0.167	?	
T	?	?	?	?

log odds score matrix:

	A	**C**	**G**	**T**
A	1			
C	?	0.53		
G	-0.42	0.43	?	
T	?	?	?	?

Question 4 *15 points*

Global sequence alignment

Given two sequences:

GCATAC and **GATACA**

Using the substitution matrix from Question 3 and gap penalty value = -1,

(a) [10 points] Find the optimal global alignment(s). Partially filled dynamic program table is shown below:

	-	G	C	A	T	A	C
-							
G		0	0			-3	-4
A		-1	-1			0	-1
T		-2	-2	1	3	2	1
A				0			
C				-1			6
A		-4	-3	0			

(b) [5 points] Is there anything unusual in the optimal alignment? Comment.

Question 5 *10 points*

Local sequence alignment: Given two sequences:

CATCGC and **TATCGT**

Using the substitution matrix given below and gap penalty value = -3, compute the optimal local alignment(s) using Smith-Waterman algorithm.

	A	G	C	T
A	2	-1	-2	-2
G	-1	2	-2	-2
C	-2	-2	2	-1
T	-2	-2	-1	2

Partially filled dynamic program table is shown below:

	-	C	A	T	C	G	C
-							
T		0	0	2	0	0	0
A		0		0			
T		0		4	1		
C		2			6	3	2
G		0					
T		0		3			