PURDUE UNIVERSITY FORT WAYNE

Department of Computer Science
COLLEGE OF ENGINEERING, TECHNOLOGY,
AND COMPUTER SCIENCE

CS 590 Introduction to Bioinformatics
Homework Assignment (8)

Objectives

- Query and read list of DNA and protein sequence data.
- Produce and analyze multiple sequence alignments.

Description

Part A:

NCBI is actively updating its GenBank database with Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) sequence data. *Write R code to perform each of the following tasks :*

1. **(10 points)** Retrieve the protein sequences with the following accession numbers: `YP_009725305, YP_009724395, YP_009725297, QIS60515, QIS60539, QIS60709.`
2. **(10 points)** Use CLUSTAL program to align the sequences. Adjust the pair-wise alignment parameters to use BLOSUM65 and a gap-opening penalty of 20. *Provide a screen-shot of the result.*
3. **(10 points)** Read the alignment into R and compute the score of the alignment.
4. **(10 points)** Print the first 40 characters of the alignment.
5. **(10 points)** Write a function to find the longest stretch of the complete conserved positions in the alignment. Test your function on the alignment of question A-3. Hint: “complete” means 100% identical letters.

Part B:

Consider the DNA sequences given in the attached file (*data.txt*). *Write R code to perform each of the following tasks:*

1. **(5 points)** Read the sequence data and store them in a FASTA file.
2. **(10 points)** Use CLUSTAL program to align the sequences. *Use a penalty for a gap opening equals 15 and a gap extension to be 3.*

3. **(10 points)** Read the alignment in R and print the first 20 positions of the alignment for only the first 3 sequences.
4. **(10 points)** Write a function to filter the alignment and keep only poorly conserved regions. Your function should accept two parameters: the alignment object, and the minimum percent of letters in an alignment column that must be gap characters for the column to be kept.
5. **(5 points)** Test the function in part (B-4) using the alignment in part (B-2) considering at least 75% of gap in a position. *Comment on the output.*
6. **(10 points)** Find the genetic distance of the alignment.