CS 175

Problem Set 1

Due: 01 October 2018 11:59pm

General Instructions

- The answer sheet for this problem set should be submitted as a PDF file. You may use any word processing software to create the answer sheet. The name of the PDF file to be submitted should follow the following format: [CS 175] < *Last Name*, *First Name* > − Problem Set 1.pdf. For example: [CS 175] De la Cruz, Juan - Problem Set 1.pdf. Include the answer sheet in the .zip file detailed next.
- If you have consulted references (books, journal articles, online materials, other people), cite them as footnotes to the specific item where you used the resource/s as reference.
- This problem set will require you to submit a .zip file containing several files (including the answer sheet) as detailed below. The name of the .zip file to be submitted should follow the following format: [CS 175] < *Last Name*, *First Name* > − Problem Set 1.zip. For example: [CS 175] De la Cruz, Juan - Problem Set 1.zip
- Submission of the problem set answers should be done via e-mail. Attach the .zip file, and write as the subject header of the e-mail: [CS 175] < *Last Name*, *First Name* > − Problem Set 1. For example: [CS 175] De la Cruz, Juan - Problem Set 1. Send your answers to jcyap@dcs.upd.edu.ph.
- **You should receive a confirmation e-mail from me stating receipt of your deliverable within 24 hours upon your submission of the problem set**. If you have not received any, forward your previous submission using the same subject header once more.
- If you have any questions regarding an item (EXCEPT the answer and solution) in the problem set, do not hesitate to e-mail me to ask them. However, **questions regarding this problem set forwarded/received on or after 12:01am of 28 September 2018 will NOT be entertained**.

Questions

Consider the FASTA formatted file named *20180910 Problem Set 1 Dataset.txt* containing 12 sequences. Perform a multiple sequence alignment using mafft making sure that the output is in Clustal format.

1. How many base pairs did each of the sequences contain?
2. How many fully conserved positions were there in the resulting multiple alignment?
3. What is the longest *substring* that is fully conserved in all of the sequences in the resulting alignment?
4. Given the alignment, what *pair of sequences* have the most number of matched regions?

Perform a sequence database search using blastn in the NCBI website on default settings on each of the sequences stored in the file.

5. From what organism did the nucleotide sequences (most likely) come from?
6. What is the over-all theme of the data set? (e.g. The data set contains nucleotide sequences sampled from the buttocks of primates indigenous to South America.)