CS313 Exercise 5 Cover Page

Name(s): _____

In the *TIME* column, please estimate the time you spent on the parts of this exercise. Please try to be as accurate as possible; this information will help us to design future exercises.

PART	TIME	SCORE
Exercise		

Task 1: Construction of Guide Tree for Multiple Sequence Alignment

Suppose that you have generated the following table of pairwise similarity scores for six orthologous sequences from the following organisms: fruit fly (*Drosophila*), dog (*Canis*), mosquito (*Anopheles*), puffer fish (*Fugu*), humans (*Homo*), and zebrafish (*Danio*)

	Drosophila	***Anopheles***	***Canis***	***Fugu***	***Homo***	***Danio***
Drosophila		46	39	41	49	43
Anopheles			34	30	33	36
Canis				57	58	60
Fugu					53	54
Homo						50
Danio						

Using the unweighted pair group method with arithmetic mean (UPGMA), draw the guide tree for the six sequences.

Task 2: Generating Pairwise Similarity and Pairwise Distance Scores

Consider the following 5 genomic sequences:

```
seq1 = CGATAGTGCTATATCTAGCGCCGTCTAGATGCATTATACGATATCG
seq2 = AACGACATGGCTCGTGCTATTACGCGCGAATATCC
seq3 = ATAGTGCTATACTCGTGCTATTCTAGATGCCGCGATATAT
seq4 = GGATAGGCTATATCTAGCGCGTCTAGATGCATTTACGATATC
seq5 = TACGACATGCGCTCGTGCATATTAGCGCGCGATATATCG
```

In this task, you will calculate pairwise alignment scores for each of the $\binom{5}{2} = 10$ possible pairwise combinations of sequences involving `seq1`, `seq2`, `seq3`, `seq4`, and `seq5`.

To begin, download the `/home/cs313/download/GuideTree_for_MSA` directory from the `CS` server. This directory contains two class files, `SequenceOps.class` and `Alignment.class`, that you are encouraged to use. You can view the contract for the `SequenceOps` class at

http://cs.wellesley.edu/~cs313/projects/project1/project1.html

and the contract for the `Alignment` class at

http://cs.wellesley.edu/~cs313/projects/project3/doc/Alignment.html

You must write a Java class that contains the following method:

```
public static int getSimilarityScore(String s1, String s2);
```

Returns the optimal pairwise global alignment score for sequence `s1` and `s2`. The global alignment should be performed using a match score of +5, a mismatch score of -4, and a linear gap score of -6.

Using your `getSimilarityScore` method described above, you should compute the optimal pairwise global alignment score for each of the $\binom{5}{2} = 10$ possible pairwise combinations of sequences involving `seq1`, `seq2`, `seq3`, `seq4`, and `seq5`. Enter your scores in the table below.

	seq1	**seq2**	**seq3**	**seq4**	**seq5**
seq1					
seq2					
seq3					
seq4					
seq5					

Each entry in the table above corresponds to a **similarity score** (specifically, the optimal global alignment score) of two sequences. Higher similarity scores represent more similar sequences, and lower similarity scores represent less similar sequences.

Now, you will calculate a **distance score** for each pair of sequences. A distance score is a dissimilarity score, i.e., higher distance scores represent less similar (more distant) sequences, and lower distance scores represent more similar (less distant) sequences.

A distance score, `D`, for two sequences `s1` and `s2` can be calculated as follows:

$$D = 100.0 * (-\ln(S_{norm}))$$

where `ln` refers to the natural logarithm. In the above equation, S_{norm} is defined as

$$S_{norm} = (S_{global} - S_{rand}) / (S_{iden} - S_{rand})$$

S_{global} is the optimal global alignment score of `s1` and `s2`.

S_{iden} is the average of the optimal global alignment score of `s1` aligned with `s1` and the optimal global alignment score of `s2` aligned with `s2`.

S_{rand} is the average of 1000 optimal global alignment scores. Each of the 1000 global alignment scores is calculated by aligning two randomly generated sequences, the first being of the same length and same expected nucleotide composition as `s1`, and the second being of the same length and same expected nucleotide composition as `s2`.

Add the following method to your Java class:

`public static double getDistanceScore(String s1, String s2);`
Returns the distance score for sequence `s1` and `s2`. The distance score is defined as 100.0 * (-ln(S_{norm})), where S_{norm} is (S_{global} - S_{rand}) / (S_{iden} - S_{rand}). In the aforementioned expression, S_{global} is the optimal global alignment score of `s1` and `s2`, S_{iden} is the average of the global alignment score of `s1` aligned with `s1` and the global alignment score of `s2` aligned with `s2`, and S_{rand} is the average of 1000 global alignment scores based on pairs of sequences generated from random sampling of `s1` and `s2`. The method assumes that global alignments are performed using a match score of +5, a mismatch score of -4, and a linear gap score of -6.

Using your `getDistanceScore` method described above, you should compute the distance score for each of the $\binom{5}{2} = 10$ possible pairwise combinations of sequences involving `seq1`, `seq2`, `seq3`, `seq4`, and `seq5`. Enter your distance scores in the table on the next page.

	seq1	seq2	seq3	seq4	seq5
seq1					
seq2					
seq3					
seq4					
seq5					

Looking at the two tables (one of similarity scores and one of distance scores) in this Task, is it the case that one pair of sequences with a higher similarity score than a second pair of sequences always has a lower distance score than the second pair of sequences? What is a reason why similarity scores are not perfectly anti-correlated with distance scores, i.e., why might one pair of sequences have a higher similarity score than a second pair yet not have a lower distance score?

Phylograms are phylogenetic trees whose branch lengths are representative of the distance between sequences. Distance scores (as opposed to similarity scores) are useful because they can suggest the branch lengths in phylograms. Longer branch lengths are used for more distant sequences and shorter branch lengths are used for less distant sequences.

Based on the second table (of distance scores) in this Task, draw a guide tree for the five sequences as a phylogram so that branch lengths in the tree are indicative of the distance between sequences or clades. The branch lengths in the phylogram need not be drawn perfectly to scale, but it should be the case throughout the tree that two sequences or clades with a larger distance score have longer branch lengths than two sequences or clades with a smaller distance score.

Task 3: Creation of Annotated Multiple Sequence Alignment (MSA)

In this Task, you will perform a multiple sequence alignment (MSA) of putative orthologs. In Exercise #4, you used BLAST to identify sequences similar to that of the yeast gene *rpn12* that codes for a component of the 26S proteasome lid. Most of the sequences reported by BLAST came from the genomes of fungi, but some did not. The following file contains the protein sequences for 10 putative orthologs: one is the original *rpn12* sequence from yeast and nine are similar sequences reported by BLAST from non-fungi organisms.

http://cs.wellesley.edu/~cs313/exercises/Exercise5/rpn12_Seqs.txt

You should perform a multiple sequence alignment of these 10 protein sequences using the Clustal Omega program available from the EBI (European Bioinformatics Institute):

https://www.ebi.ac.uk/Tools/msa/clustalo/

When Clustal Omega finishes aligning your sequences, inspect the results both in the "Alignments" tab and the "Result Summary" tab, specifically the Percent Identity Matrix.

- Click on the "Phylogenetic Tree" tab and scroll down the results page to find the phylogeny. Click the buttons “Cladogram” and “Real” to toggle between cladogram and phylogram versions of the phylogeny.
- One of the advantages of a multiple sequence alignment is that it can provide insight into various properties of a family of proteins. When inspecting multiple sequence alignments, if scientists find portions of their sequences that do not align well, they often remove these non-homologous regions, particularly if they are interfering with optimal alignment within the protein family. Similarly, if particularly divergent sequences from a set of orthologs appear to interfere with optimal alignment, the divergent sequences may be omitted from the alignment.

Which pair of sequences (out of all pairs of sequences, not necessarily including the yeast gene) is the most similar? Which pair of sequences is the most dissimilar?

Based on the 10 sequences, how well does the yeast gene seem to fit its protein family?

For this one protein family, is the phylogenetic tree consistent with your expectations about the phylogenetic relationships between the 10 species?

Are there particular regions within the protein family that appear to be more highly conserved? If so, why might some regions be more highly conserved? If not, why might there be uniform conservation throughout a protein family?