CS313 Exercise 8 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.

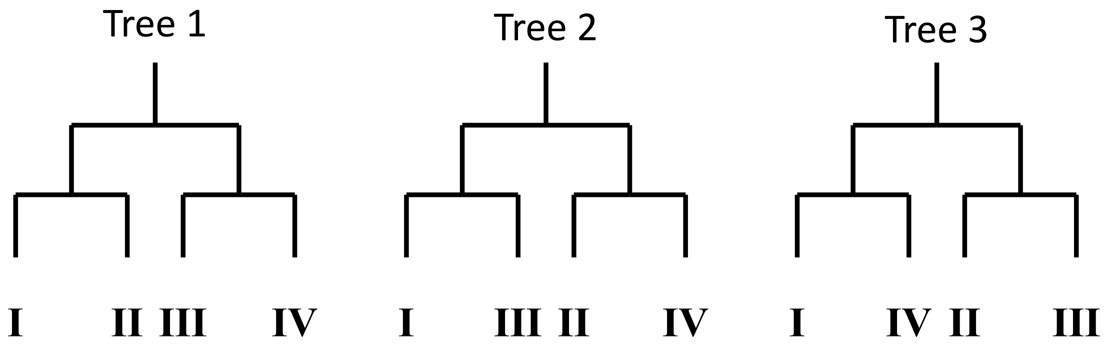
PART	**TIME**	**SCORE**
Exercise		

Task 1: Constructing a Phylogenetic Tree Based on Maximum Parsimony

Consider the following four aligned sequences:

```
I    TCGTTCG
II   TAGGTTA
III  CCGGCAA
IV   AAGTCCG
```

For these four sequences, which of the three trees below is the most parsimonious with respect to evolutionary changes? Show your work.



Task 2: Phylogenetic Tree Concepts

Prior to completing this task please read the perspectives article by Baum *et al.* entitled "The Tree-Thinking Challenge". The article can be found in the Exercise8 folder in the `download` folder on the CS server.

Phylogenetic trees represent hypotheses of the evolutionary relationships between taxa. Molecular data have provided an invaluable supplement to pre-existing trait-based phylogenies, and have greatly improved our phylogenetic knowledge. However, care must be taken when constructing phylogenetic trees based on molecular data.

Phylogenetic trees based on gene sequences do not always correspond exactly to phylogenetic trees showing the evolutionary relationships between species. What are two reasons that this might happen?

When using molecular data to construct phylogenetic trees, protein sequence data are generally more useful than nucleotide sequence data for inferring evolutionary events that occurred in the distant past. Why is this?

What would be the consequence(s) of using a gene in your phylogenetic analysis that is a paralog that is no longer functional (called a pseudogene) instead of an ortholog?

Among the Bacteria and the Archaea, transfer of genetic material from the genome of one species to another is not uncommon. For example, a fragment of DNA might be transferred from the chromosome of *Escherichia coli* to the chromosome of *Enterobacter cloacae*, since these bacteria are often present together in mammalian gastrointestinal tracts. This is a process called **horizontal gene transfer** or **lateral gene transfer** (transfer of genetic material from parent to offspring is called **vertical gene transfer**).

What would be the consequence of using a gene that has been laterally transferred in a phylogenetic analysis that includes both the horizontal gene transfer donor bacterium and the horizontal gene transfer recipient bacterium?

Task 3: Epidemiological Case Study of HIV Virus Phylogeny

Human Immunodeficiency Virus (HIV) is a virus with a single-stranded RNA genome that is 9749 nucleotides long. Because RNA replication is highly error prone when compared to DNA replication, the HIV virus is constantly mutating. Many of these nucleotide changes result in non-functional viruses, but some produce viable viruses with altered cell surface antigens. This represents a significant challenge to producing an effective HIV vaccine.

The HIV genome has 9 ORFs that produce 15 proteins, which is accomplished via the action of a protease encoded by the HIV genome. Because this HIV protease has a mechanism of action that is distinct from human proteases, and because HIV protease activity is crucial for virus replication, a mixture of drugs that specifically block the HIV protease (protease inhibitor) and the HIV polymerase (reverse transcriptase) has been reasonably effective at controlling HIV infections. The outside of the HIV virus is coated with a glycoprotein called gp120. gp120 specifically binds to CD4, a human cell surface protein. gp120-CD4 binding is a critical event for viral binding and subsequent infection of the host cell.

In the late 1980's, eight patients of an HIV-positive dentist in Florida were diagnosed as being HIV-positive. Though many of the patients had had invasive dental procedures performed (root canals, tooth extractions), an investigation by the Centers for Disease Control did not uncover systematic hygienic lapses that might account for infection of the patients. Additionally, there were no obvious ways in which the dentist might have deliberately infected his patients.

In an attempt to determine whether the eight HIV-positive patients were infected by the dentist, researchers isolated viral RNA from blood samples from the dentist, the infected patients, and HIV-positive individuals in the area who had had no contact with the dentist. The investigators then amplified DNA copies of the genomic RNA sequences via the polymerase chain reaction (PCR) and determined the nucleotide sequence of pieces of the HIV gp120 gene. This data was then used to determine how closely related the dentist's HIV virus strain was to that of his patients, to that of the individual who was a sexual contact of the dentist, and to that of the HIV-positive individuals who had not had contact with the dentist.

In this task you will use some of the gp120 protein sequences described above to recapitulate the phylogenetic analysis performed in the case of the Florida dentist. To begin, download the file "HIV_data_set.txt" from the Exercise8 folder in the `download` folder on the CS server.

This file contains gp120 protein sequence data from the dentist, his infected patients, one of the dentist's HIV-positive sexual contacts, and local controls (other HIV-positive individuals in the area who had no contact with the dentist).

Using the Clustal Omega program, perform a multiple sequence alignment using the HIV gp120 sequences in the file provided. (Tip: uploading the HIV_data_set.txt file to the Clustal Omega web server may result in faster execution than cutting and pasting the sequences.)

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

Based on the resulting phylogram, do you think it is likely that the Florida dentist infected any of the eight patients? Are there any patients who were unlikely to have been infected by the dentist? Why?

Why was it important for researchers to collect HIV samples from people who had no contact with the HIV-positive dentist as well as an individual with whom the dentist had had sexual contact?

One criticism of the phylogenetic analysis is that HIV virus can mutate within an individual, thereby producing many distinct HIV viruses with different gene sequences, which could obscure the phylogenetic analysis. How might a researcher try to compensate for this possibility in their experimental approach?

Task 4: Analysis of 16S Ribosomal RNAs

Ribosomes are cellular components that translate mRNA into proteins. Ribosomes are composed of RNAs and proteins. One of the RNAs in bacterial and archaeal ribosomes is called 16S rRNA. In this task, you will create a Java program to investigate 16S rRNA sequences from bacteria and archaea. As a starting point, download the `RibosomalAnalysis` folder from the `Exercise8` folder in the `download` folder on the CS server. The folder contains a sub-folder named `data` that has two files, one containing 1,325 16S rRNA sequences from bacteria and one containing 111 16S rRNA sequences from archaea. The folder also contains a class named `PhylogeneticOps`. You can view the contract for the `PhylogeneticOps` class here:

http://cs.wellesley.edu/~cs313/exercises/Exercise8/doc/

Using the methods in the `PhylogeneticOps` class, your goal is to write a Java program that does the following:

- Create a group of 5 randomly chosen 16S rRNA sequences from bacteria and a group of 5 randomly chosen 16S rRNA sequences from archaea.
- Compute the average distance (based on global sequence alignment) between 16S sequences in the same group (i.e., compute the average distance over all pairs of sequences `A` and `B` where `A` and `B` are in the same group), and compute the average distance between 16S sequences in different groups (i.e., compute the average distance over all pairs of sequences `A` and `B` where `A` and `B` are in different groups).
- Create a 10-by-10 matrix of distances between each pair of the 10 16S sequences (5 from bacteria and 5 from archaea) and cluster the 10 sequences into 2 clusters via hierarchical clustering.

What is the average within group distance between 16S rRNA sequences?

What is the average between group distance between 16S rRNA sequences?

Do the 16S sequences cluster as you would expect based on their phylogenies?