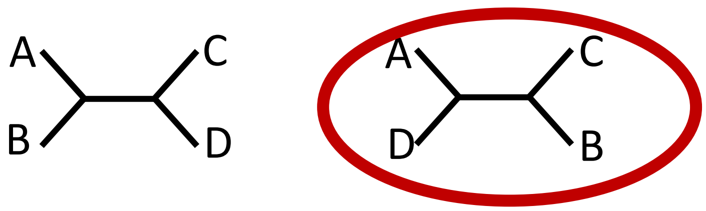
GS 373 Homework 3

Due April 26th before 1:30 PM on Canvas

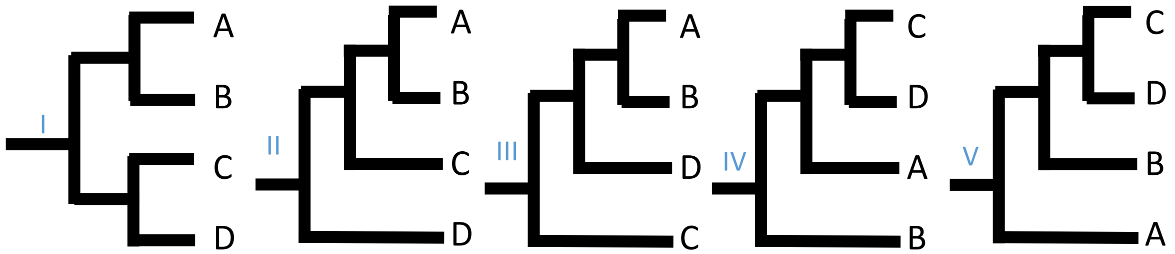
- (100 points): 4 bioinformatics questions (80 points), 1 programming assignment (20 points).
- Submit answers to the bioinformatics questions in a Microsoft Word document or PDF via Canvas. Your answers do not need to contain the text of the questions, but they need to be clearly labeled (e.g., 1a., 3b., etc.)
- Tree topologies can be depicted by inserting shapes and text boxes into a Word Document, by inserting a photo of a hand-drawn tree, or by any method you prefer.
- Submit the programming assignment as a separate .py file onto Canvas. The script should be able to be directly run by Python.

Bioinformatics Questions (80 points)

1. (20 points, 3 parts) Tree topologies
   a. (6 points) **Draw and label three distinct *unrooted* topologies for trees with 4 leaves** labeled A, B, C, and D. **Circle the tree that can be rooted such that A and D share a more recent common ancestor** than either of them share with B or C.



   b. (6 points) Choose one of the trees you drew and **draw three topologically distinct rooted variations of that tree**, with all branches parallel to each other.



   c. (8 points) **How many distinct *rooted* tree topologies with 5 species exist?**

Same number as distinct unrooted tree topologies with 6 species: (# of unrooted topologies)*(# of branches to which the root can be added) = (15)*(2N-3), where N=5, (15*7) = 105

2. (25 points, 2 parts) Distance-based tree methods
   a. (5 points) **What is the upper bound for the maximum raw sequence divergence (fraction of unmatched bases) between two sequences, assuming equal nucleotide frequencies?** Explain why. Suppose the assumption of equal frequencies does not hold and you are comparing two sequences that both have very high G-C content – **how would this upper bound change?**

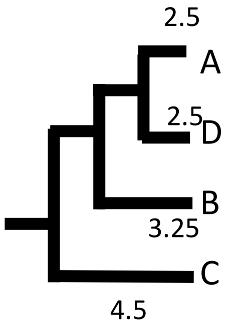
With 4 possible nucleotides, the probability that any site matches even if the sequences are unrelated is ¼, or 0.25. Therefore the probability that they don’t match is 0.75. Therefore the expected distance between two unrelated sequences is 0.75. Upper bound would be lower for higher G-C content because the expected distance between two unrelated sequences is higher (chance that any site matches is greater than 0.25)

b. (20 points) **Use the UPGMA algorithm to construct a tree based on the following distance matrix.** Label the length of each branch on your final tree.

Step1				
	A	B	C	D
A	0	7	10	**5**
B		0	8	6
C			0	9
D				0

Step 2			
	A, D	B	C
A, D	0	**6.5**	9.5
B		0	8
C			0

Step 3	
	C
A, D, B	**9**



3. (20 points, 2 parts) Molecular Evolution
    a. (15 points) For situation below, indicate whether it would be **more appropriate to use a dN/dS method to detect positive selection or a population method and why**.
        i. You have sequencing data from many individuals from two recently diverged populations.
        population (recently diverged, many individuals)
        ii. You are interested in a promoter sequence of a human protein.
        population (dN/dS only works for proteins)
        iii. You are interested in identifying sites on a viral protein which repeatedly changed identity due to selection from the immune system.
        dN/dS (repeated changes, protein, specific sites)

4. (15 points) Take a look at the **2016 paper by Hug *et al.*** describing a new reconstruction of the evolutionary tree of life. **Focus on Figures 1 and 2 and the first three paragraphs of the Methods**.
    a. (8 points) What **method** did the authors use to construct this tree? For the trees in Figures 1 and 2, what **type of sequences** were used from each organism?
    maximum likelihood, ribosomal protein sequences