Assignment #2

Part I. Drawing phylogenetic trees (using paper and pencil).

1) **Draw the 15 unrooted binary topological trees that could describe the relationship between 5 taxa a, b, c, d, e. Group them naturally according to the relationship they display for the first 4 taxa a, b, c, d.**

2) An unrooted tree has vertices v1, v2, . . . , v9 with edges {v1, v2}, {v1, v6}, {v1, v9}, {v2, v7}, {v2, v8}, {v3, v9}, {v4, v9}, {v5, v9}.
a. Without drawing the tree, determine the degree of each vertex.
b. Use your answer to part (a) to determine the leaves of the tree.
c. Use your answer to part (a) to determine whether the tree is binary.
d. Draw the tree to check your work.
e. Write this tree in the Newick format.

Part 2. Git & GitHub

3) You'll need some knowledge of Git for this part. If you lack it, play with this interactive tutorial or read my non-interactive one or do both. Now create a ~/EEOB563/labs directory on your computer and initialize it as a Git repository. Add a README file to the directory and commit it to the repository (there are a few steps inbetween!). Create an empty GitHub repository and link your local repository to it. Push the changes you made. **Include the link to your GitHub repository in your assignment!**

Part 3. Genbank and BLAST

4) Michael Crichton's fantasy about cloning dinosaurs, Jurassic Park, contains a putative dinosaur DNA sequence that you can retrieve from the NCBI ftp site. Use NCBI's nucleotide BLAST tool to find identical/similar sequences in the nucleotide (nr/nt) database. **a) What sequence did Michael Crichton use?**

Mark Boguski, who was at the NCBI at the time, noticed this obvious contaminant and supplied Crichton with a better sequence for the sequel, The Lost World. You can also retrieve this sequence from the NCBI ftp site. Identify the most likely source of Mark's sequence using translating BLAST (blastx).
b) Is Mark's sequence identical to the source sequence you found in the database? What are the differences?

5) Download 20 sequences from NCBI most similar to Mark's "dinosaur" sequence. **Are these sequences**

homologous/orthologous/paralogous? Explain!

Part 4. Mafft (extra point).

Install mafft on your computer or use it on the HPC-class (issue `module load mafft` first!). There are onle servers that can run MAFFT, but do not use them or use them only after you try the other options. One of our goals in this course is to become comfortable with the UNIX/command line interface.

Use mafft to create a multiple sequence alignment for amino-acid sequences you downloaded from GenBank. **What search strategy did you/mafft use for the alignment?**

In addition to the alignment itself mafft program can output the guide tree it uses to build it (see website for details). You can view this tree with FigTree. **What does this tree represent? Does it correspond to the species phylogeny you expected? Explain!**

Include your tree (no alignment, please!) and your answers to the questions above in a single document and either send it by a direct message to me in Slack or print it out and bring to class by the due date.

GOOD LUCK!