

CS 490/590 Project 4 and Lab Activity for Chapter 6

Description: For this week's project, you are to implement both the Jukes-Cantor distance measurement of the Guided Programming Project **and** complete most of the Web Exploration Project (exacts parts outlined below). Your programming component is almost trivial and requires that a multiple sequence alignment already be completed. You will use the Web component to both extract that multiple alignment and to see what the output should be.

Web Exploration Component: While you should read the chapter and project, the following are guidelines of specific steps:

1. Find the beta-casein gene sequences for rat, camel, dog, whale, and porpoise from GenBank. Use the scientific species name (see pg. 113) along with "beta casein" or "CSN2". Paste all of your sequences in FASTA format into a single file, e.g. **allseqs.txt**.
2. Go to the [phylogeny FR "one click" site](#). Either paste all of your sequences into the textbox, or upload your allseqs.txt file. Choose option to get results via email, and allow attachments.
3. It will take a while, but you will get multiple emails with different attachments. Two important attachments are the **multiple sequence alignment** (one of the early messages) and the **phylogenetic tree** (in a later message). A usefully formatted file containing the multiple sequence alignment is **alignment.fasta**. The phylogenetic tree represented as a picture is **phylotree.png**. Does the phylogenetic tree look like Fig. 6.4 of text?
4. From file **alignment.fasta** you need to extract the individual aligned sequences (including gaps!) and save into separate files with appropriate file names indicating species. E.g. "rat.txt" is acceptable, though "ratCSN2.txt" is more specific and also acceptable.
5. **You will run your Jukes-Cantor distance implementation on the aligned sequence files.**

Programming component: All you are calculating is the Jukes-Cantor distance of two **already aligned** sequences. This distance is exactly the K value of page 110.

What to turn in: Upload a zip of the following items to Moodle:

1. **Your source code, README, and any other project files needed to compile and run your program (consistent with open source and Java/C/C++ rules of former projects).**
2. **The multiple alignment file alignment.fasta from the phylogeny site.**
3. **The phylogenetic tree picture phylotree.png from the phylogeny site.**
4. **All the individual already aligned sequences (see part 4) used as input into the program.**

This project/activity is due by MIDNIGHT of Thursday, October 17. It is worth 4% of your total grade. (Late policy: -33% per day off.)