**BIOMI 609 Computational Genomics and Bioinformatics**

**Spring 2022**

**San Diego State University**

**Dr. Arun Sethuraman**

**Assignment 2**

**Due on Monday, 3/7/2022 at 8 AM before class on Canvas**

For this week’s assignment, you’ll have to install MEGA on your machine first (<https://www.megasoftware.net/>). Installation instructions are on the website.

1. Here’s a distance matrix generated from a multiple sequence alignment of the FOXP2 gene in humans, gorillas, macaques, and rabbits. Build a UPGMA tree from this distance matrix. Now go to Genbank, look for the accession [NM\_014491.4](https://www.ncbi.nlm.nih.gov/nucleotide/NM_014491.4?report=genbank&log$=nucltop&blast_rank=1&RID=YN9PFZYG014), BLAST it, obtain hits for the same 4 species, download it as a multiple sequence alignment FASTA, upload it into MEGA, build a UPGMA tree in MEGA, and compare that with your results, paste your final tree in your online part (10 points). Also construct a neighbor joining tree using the F&M algorithm discussed in class, compare it to the NJ tree constructed using MEGA (5 points).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Genetic Distances | *Homo sapiens* | *Gorilla gorilla* | *Macaca mulatta* | *Pan paniscus* |
| *Homo sapiens* | 0 |  |  |  |
| *Gorilla gorilla* | 1.07 | 0 |  |  |
| *Macaca mulatta* | 1.79 | 1.84 | 0 |  |
| *Pan paniscus* | 1.88 | 1.8 | 1.94 | 0 |

Graphical user interface, text, application, email

Description automatically generated

From MEGA (UPGMA)

From MEGA (NJ):

Graphical user interface, text, application, email

Description automatically generated

2. Compute several likelihood trees based on different models (Jukes Cantor, Kimura 2 parameter, Generalized Time Reversible), then compare the likelihoods using a log-likelihood ratio test. JC model has 1 parameter, GTR has 10 parameters, K2 model has 2 parameters) - 5 points.

Under GTR: LogL = -31520.85 (10 parameters)

Under Jukes Cantor: LogL = -31950.90 (1 parameter)

Under Kimura 2 parameter model: LogL = -31950.89 (2 parameters)

So comparing GTR vs Jukes Cantor, we have 9 df, LLR = 2 \* (-31520.85+31950.90) = 860.1, p-value << 0.005, which provides support for GTR over the Jukes Cantor Model

Comparing Jukes Cantor vs Kimura 2 parameter, we have 1 df, LLR = 2 \* (-31950.90 +31950.89) , p-value = 1.0, which doesn’t provide support for Jukes Cantor Model.

Similarly, comparing GTR vs Kimura 2 parameter, we have 8 df, LLR = 2 \* (-31520.85+31950.89) = 860.08, p-value << 0.005, which provides support for GTR over the Kimura 2 parameter model.

Overall, you can see that the GTR has the highest likelihood as well, therefore a better model.