
EEOB563 – Assignment #5

Dennis Psaroudakis

March 12th 2019

03/05

1

a) Assumptions of JC-model:

- all bases equally likely in root sequence & all following sequences
- mutation rate equal for all $x \rightarrow y$, $x, y \in \{A, G, C, T\}$

So first I'll look at base distribution

	A	G	C	T
S ₀	111	105	94	90
S ₁	109	102	97	92
S _{0'}	98	92	107	103
S _{1'}	100	87	110	103

None of them are truly equally distributed.
The expected value for each cell would be 100. According to the model.

I will calculate a summed squared error for each sequence to find out if they are significantly different from each other in terms of compliance:

$$\sum_{\text{cell}}^{\text{row}} (\text{cellValue} - 100)^2$$

$$S_0 = \underbrace{282, S_1 = 158, S_0' = 126, S_1' = 278}_{440}$$

So the second table fits closer with this assumption, but I don't think that's what we're looking for yet, so I'll look at the second assumption

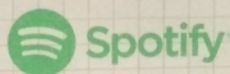
Aaah I don't even need to calculate anything: In the first table the probability for transitions is much higher than for transversions
→ that contradicts the model.

⇒ JC is more appropriate for the second pair of sequences, S_{0'} and S_{1'}

b) An appropriate model for the first pair would be a Kimura model.

$$L(AG \xrightarrow{0.3} AG) = \underbrace{L(A \xrightarrow{0.3} A)}_{\text{part I}} \times \underbrace{L(G \xrightarrow{0.3} G)}_{\text{part II}}$$

At least I assume that you multiply the site likelihoods, anything else doesn't make sense.



$$\text{Part I: } L(A \rightarrow \begin{matrix} \nearrow A \\ \nwarrow A \end{matrix}) = L(A \rightarrow \begin{matrix} \nearrow A \\ \nwarrow A \end{matrix}) + 3 \cdot L(A \rightarrow \begin{matrix} \nearrow \text{not } A \\ \nwarrow A \end{matrix})$$

Since in gc all probabilities for changing a nucleotide are the same, I can summarize G, C, and T to "not A".

$$L(A \rightarrow \begin{matrix} \nearrow A \\ \nwarrow A \end{matrix}) = \frac{1}{4} \cdot P_{AA}(0.3) \cdot P_{AA}(0.1)^2 = \frac{1}{4} \cdot \left(\frac{1}{4} + \frac{3}{4}e^{-0.4/3}\right) \cdot \left(\frac{1}{4} + \frac{3}{4}e^{-0.1/3}\right)^2$$

↑
choose this as root node.

$$= \frac{1}{4} \cdot 0.75274 \cdot 0.90638^2 = \underline{0.154599}$$

$$L(A \rightarrow \begin{matrix} \nearrow \text{not } A \\ \nwarrow A \end{matrix}) = \frac{1}{4} \cdot P_{AX}(0.3) \cdot P_{AX}(0.1)^2 = \frac{1}{4} \cdot \left(\frac{1}{4} - \frac{1}{4}e^{-0.3/3}\right) \cdot \left(\frac{1}{4} - \frac{1}{4}e^{-0.1/3}\right)^2 = 2.0066 \cdot 10^{-5}$$

$$L(\text{Part I}) = 0.1546592$$

$$L(\text{Part II}) = L(G \rightarrow \begin{matrix} \nearrow G \\ \nwarrow G \end{matrix}) + L(G \rightarrow \begin{matrix} \nearrow C \\ \nwarrow G \end{matrix}) + 2 \cdot L(G \rightarrow \begin{matrix} \nearrow [AT] \\ \nwarrow G \end{matrix})$$

$$L(G \rightarrow \begin{matrix} \nearrow G \\ \nwarrow G \end{matrix}) = \frac{1}{4} \cdot P_{GG}(0.3) \cdot \underbrace{P_{GG}(0.1)}_{\substack{\text{same as } P_{GC}(0.1) \\ \text{so I'm saving the value for later}}} \cdot \underbrace{P_{GC}(0.1)}_{\substack{\text{same as } P_{GC}(0.1) \\ \text{so I'm saving the value for later}}} = \frac{1}{4} \cdot 0.75274 \cdot 0.028285 = 5.3228 \cdot 10^{-3}$$

$$L(G \rightarrow \begin{matrix} \nearrow C \\ \nwarrow G \end{matrix}) = \frac{1}{4} \cdot P_{GC}(0.3) \cdot \text{saved value} = \frac{1}{4} \cdot 0.08242 \cdot \text{saved value} = 1.03677 \cdot 10^{-4}$$

$$L(G \rightarrow \begin{matrix} \nearrow [AT] \\ \nwarrow G \end{matrix}) = \frac{1}{4} \cdot \underbrace{P_{G[AT]}(0.3)}_{\text{same value}} \cdot P_{G[AT]}(0.1)^2 = \frac{1}{4} \cdot 0.08242 \cdot 9.73856 \cdot 10^{-4} = 2.0066 \cdot 10^{-5}$$

$$L(\text{Part II}) = 5.4726 \cdot 10^{-4}$$

$$\Rightarrow L(\text{Tree}) = \underline{8.461 \cdot 10^{-4}}$$

3

F84 is not available in raxml, so I'm using F81. The trees are not the same (see following tanglegram), they also have different log-likelihood scores:

- 1 **Raxml**: -54568.308901
- 2 **FastMe**: -54674.432495

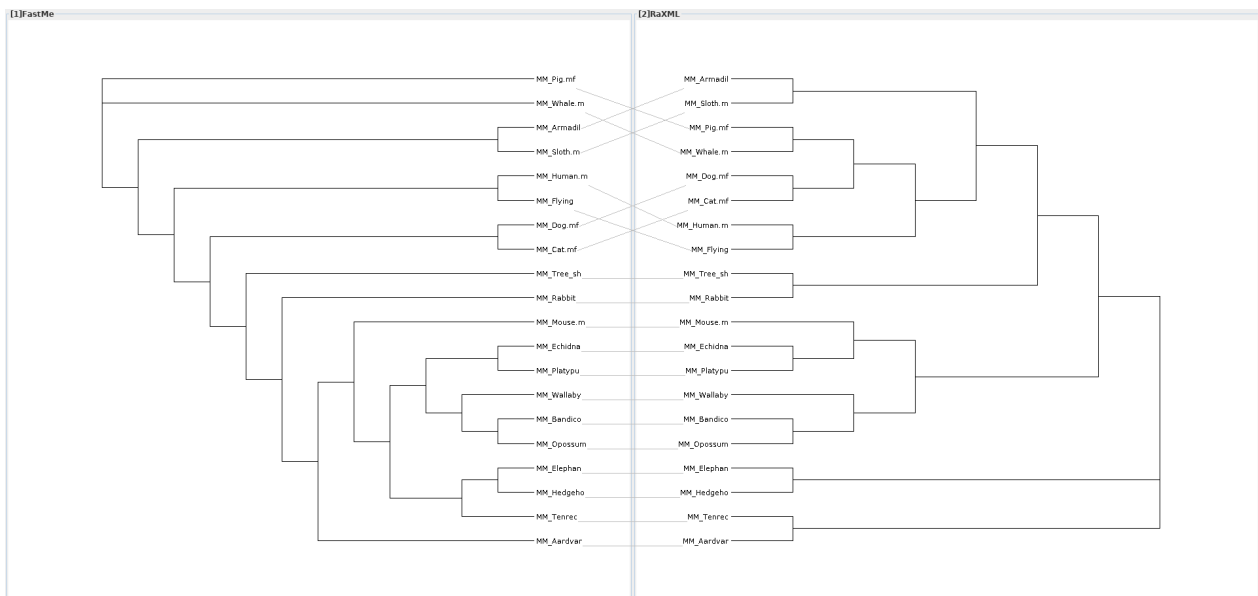


Figure 1: Tanglegram made with Dendroscope (zoom in)

4

These are the two commands I used:

```
1 raxml-ng --search --msa ../alignment.phy --model model_codons --prefix codons --seed 12 --brlen scaled
2 raxml-ng --search --msa ../alignment.phy --model model_genes --prefix genes --seed 12 --brlen scaled
```

The two models looked like this:

```
1 model_codons:
2   GTR+G+FO, COBX=1-4482/3
3
4 model_genes:
5   GTR+G+FO, cob=1-1248
6   GTR+G+FO, cox1=1249-2808
7   GTR+G+FO, cox2=2809-3543
8   GTR+G+FO, cox3=3544-4482
```

These are the log-likelihoods for the best tree:

```
1 Codons: -45373.634679
2 Genes: -45270.430490
```

They are much better than the original tree BUT the evolution models I used are different so I'm not sure if it's directly comparable.

Anyway, here is the difference between the Gene partitioned tree and the original RaXML tree, they are very similar except for the rabbit:

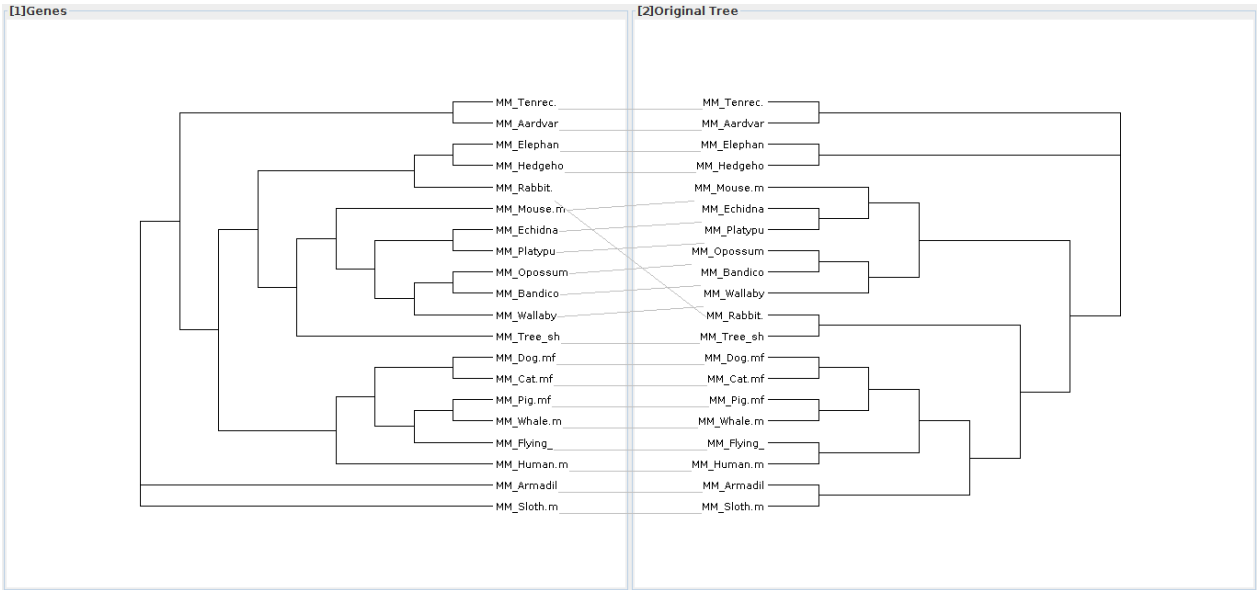


Figure 2: Tanglegram made with Dendroscope (zoom in)

5

I had problems with slurm so I couldn't start it until Tuesday evening, so it's still running. I will resubmit this assignment tomorrow with part 5 when it is done. I am expecting a runtime of ≥ 10 hours from what the others have told me, I don't remember the runtimes from the lab anymore. This is my slurm script:

```
1  #!/bin/bash
2
3  #SBATCH --time=10:30:00    # walltime limit (HH:MM:SS)
4  #SBATCH --nodes=1         # number of nodes
5  #SBATCH --ntasks-per-node=16  # 16 processor core(s) per node
6  #SBATCH --job-name="5.5_assignment"
7  #SBATCH --mail-user=dpsaroud@iastate.edu    # email address
8  #SBATCH --mail-type=BEGIN
9  #SBATCH --mail-type=END
10 #SBATCH --mail-type=FAIL
11
12 # LOAD MODULES, INSERT CODE, AND RUN YOUR PROGRAMS HERE#!/bin/bash
13 cd /home/dpsaroud/EEOB563/assignments/5/5
14
15 /home/dpsaroud/bin/raxml-ng --all --msa ../alignment.phy --model GTR+G --seed 12 --
    threads 16 --bs-metric fbp --bs-trees 1000
```