EEOB563 – Assignment #5

Dennis Psaroudakis

March 12th 2019

EEOB563 - Assignment 5

Dennis Psaroudakis

03/05

1/

a) Assumptions of JC-model:

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

So first I'll look at base distribution

	A	G	C	T
S_0	111	105	54	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:

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

$S_0 = 282$, $S_1 = 158$, $S'_0 = 126$, $S'_1 = 278$

$S_0 + S_1 = 440$, $S'_0 + S'_1 = 404$

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.

2/ $L(AG \xrightarrow{0.3} \begin{matrix} \xrightarrow{0.1} AG \\ \xrightarrow{0.1} AG \end{matrix}) = \underbrace{L(A \xrightarrow{.3} \begin{matrix} \xrightarrow{.1} A \\ \xrightarrow{.1} A \end{matrix})}_{\text{Part I}} \times \underbrace{L(G \xrightarrow{.3} \begin{matrix} \xrightarrow{.1} C \\ \xrightarrow{.1} G \end{matrix})}_{\text{Part II}}$

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

Part I: $L(A \xrightarrow{3} \begin{matrix} A \\ A \\ A \end{matrix}) = L(A \xrightarrow{1} A \begin{matrix} A \\ A \\ A \end{matrix}) + 3 \cdot L(A \xrightarrow{3 \text{ not } A} \begin{matrix} A \\ A \end{matrix})$

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

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

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

↑ choose this as root node.

$L(A \xrightarrow{3 \text{ not } A} \begin{matrix} A \\ A \end{matrix}) = \frac{1}{4} \cdot P_{AX}(0.3) \cdot P_{AA}(0.1)^2 = \frac{1}{4} \cdot \left(\frac{1}{4} - \frac{1}{4} e^{-0.3 \cdot 4/3}\right) \cdot \left(\frac{1}{4} - \frac{1}{4} e^{-0.1 \cdot 4/3}\right)^2 = 2.0046 \cdot 10^{-5}$

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

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

$L(G — G \begin{matrix} C \\ G \end{matrix}) = \frac{1}{4} \cdot P_{GG}(0.3) \cdot \underbrace{P(0.1)}_{GG} \cdot \underbrace{P(0.1)}_{GC} = \frac{1}{4} \cdot 0.75274 \cdot 0.028285 = 5.3228 \cdot 10^{-3}$

same as $P(0.1) \cdot P(0.1)$ so I'm saving the value for later. (P_{CC}, P_{CA})

$L(G — C \begin{matrix} C \\ G \end{matrix}) = \frac{1}{4} \cdot P_{GC}(0.3) \cdot \text{saved value} = \frac{1}{4} \cdot 0.08242 \cdot \text{saved value} = 1.09677 \cdot 10^{-4}$

$L(G — [AT] \begin{matrix} C \\ G \end{matrix}) = \frac{1}{4} \cdot \downarrow \text{same value} \cdot P_{[AT][CA]}(0.1)^2 = \frac{1}{4} \cdot 0.08242 \cdot 9.71856 \cdot 10^{-4} = 2.0066 \cdot 10^{-5}$

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

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

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:

```
Raxml:  -54568.308901
FastMe: -54674.432495
```

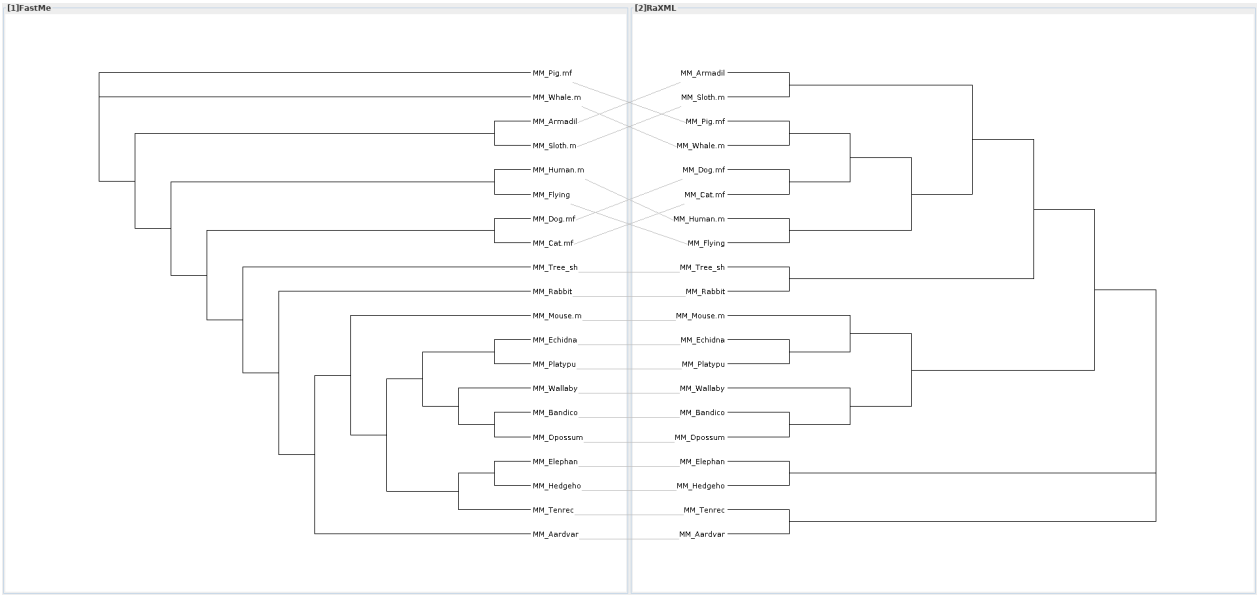

Figure 1: Tanglegram made with Dendroscope (zoom in)

4

These are the two commands I used:

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

The two models looked like this:

```
model_codons:
  GTR+G+FO, COBX=1-4482/3

model_genes:
  GTR+G+FO, cob=1-1248
  GTR+G+FO, cox1=1249-2808
  GTR+G+FO, cox2=2809-3543
  GTR+G+FO, cox3=3544-4482
```

These are the log-likelihoods for the best tree:

```
Codons: -45373.634679
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:

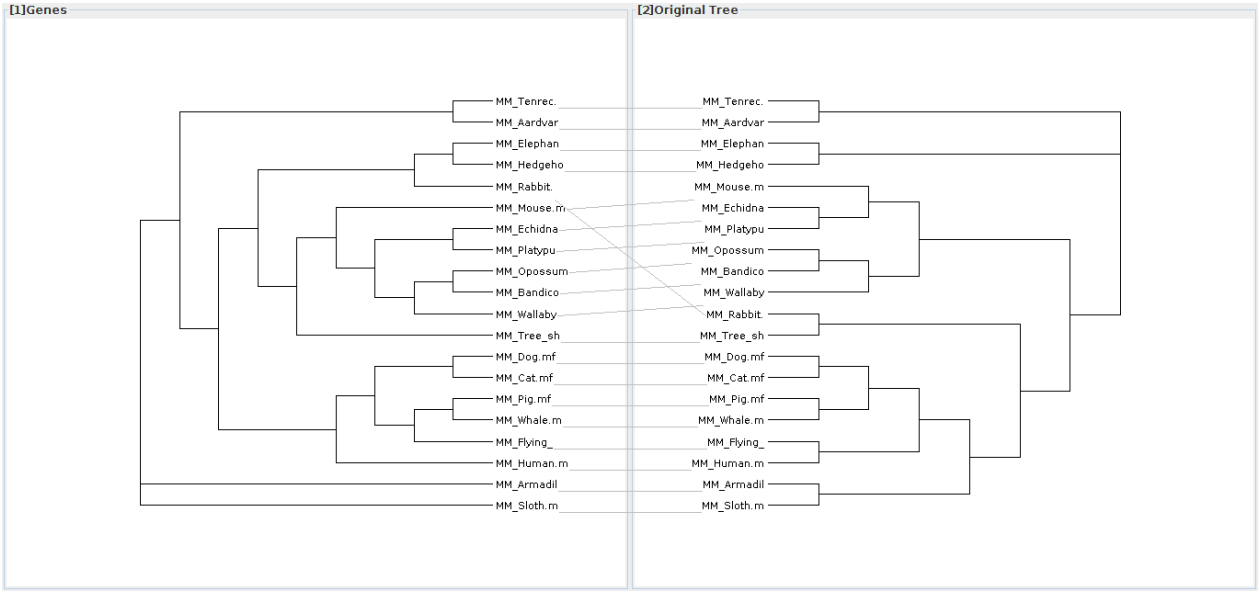

Figure 2: Tanglegram made with Dendroscope (zoom in)

5

I had expected a runtime of $\geq$ 10 hours from what the others have told me, I don't remember the runtimes from the lab anymore. In reality, it only took about 6:30 hours though, so I was lucky. This is my slurm script:

```
#!/bin/bash

#SBATCH --time=10:30:00   # walltime limit (HH:MM:SS)
#SBATCH --nodes=1   # number of nodes
#SBATCH --ntasks-per-node=16   # 16 processor core(s) per node
#SBATCH --job-name="5.5_assignment"
#SBATCH --mail-user=dpsaroud@iastate.edu   # email address
#SBATCH --mail-type=BEGIN
#SBATCH --mail-type=END
#SBATCH --mail-type=FAIL

# LOAD MODULES, INSERT CODE, AND RUN YOUR PROGRAMS HERE#!/bin/bash
cd /home/dpsaroud/EEOB563/assignments/5/5

/home/dpsaroud/bin/raxml-ng --all --msa ../alignment.phy --model GTR+G --seed 12 --
    threads 16 --bs-metric fbp --bs-trees 1000
```

Here is the tree, finally the tree is further away from the animals:

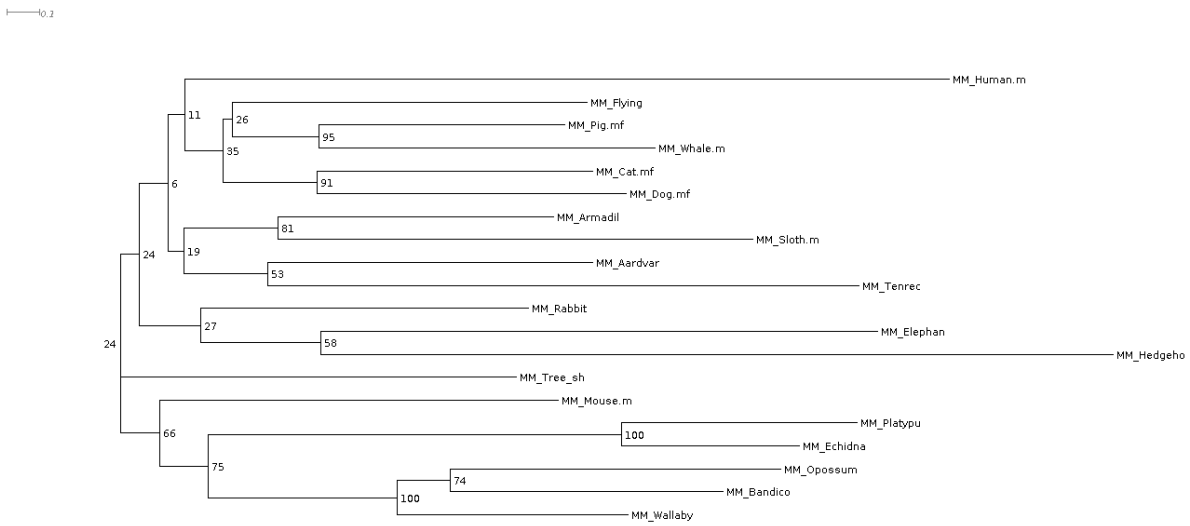

Figure 3: ML tree with bootstrap support values