HW3

Philipp Ross

2017-01-31

Last updated: 2017-02-02

Code version: ca22fdf

Part 1. Phylogenetic analysis

Problem 1

We are interested in sequence AAAGGCCTTT.

(a)

The JC69 assumes equal nucleotide frequences: $\pi_A = \pi_C = \pi_T = \pi_G$. This means the probability of any individual nucleotide is 0.25. With a sequence of 10 nucleotides, the probability of this sequence is going to be $L(D|H_0) = 0.25^{10} = 9.5367432 \times 10^{-7}$.

(b)

Utilizing the F84 model, where $\pi_{AT} + \pi_{GC} = 1$ and π_{AT} stands for the frequency of an A or T, for the above sequence we can calculate that $\pi_{AT} = \frac{6}{10}$ and $\pi_{GC} = \frac{4}{10}$.

For our likelihood, we can calculate $L(D|H_1) = (\frac{1}{6})^6(\frac{1}{4})^4 = 0.0011944$

(c)

The likelihood ratio is just the ratio of both likelihoods. If we compute the likelihood ratio as $\Lambda = \frac{L(D|H_1)}{L(D|H_0)}$ where H_1 is the F84 model while H_0 is the JC69 model, we get $\Lambda = \frac{(\frac{1}{6})^6(\frac{1}{4})^4}{0.25^{10}} = 1252.4124635$. The degrees of freedom are determined by the difference in the number of free parameters between the two models, which in this case is 2 because the JC69 models has one equilibrium frequency parameter, while the F84 model has 2. Thus $\Lambda \sim \chi^2(1)$ giving us a p-value of $2.482043 \times 10^{-274}$.

Problem 2

The command run to generate the output was:

```
phyml -i mito3.phy -q -d nt -m JC69 -f d -c 1 -s SPR --print_site_lnl \
-b -1 --run_id JC69
```

(a)

The log-likelihood of the tree under the JC69 model is -5569.51634 The generalized expression for calculating the likelihood is:

$$L(D|T, B, M = JC69) = \prod_{k=1}^{K-1} \prod_{j=1}^{N} \sum_i P(D|T, B, X_i)\pi_i$$

Where D is the data, T is a particular tree, B are the branch lengths of that tree, M is the model, N is the number of nucleotides in a particular alignment, K is the number of sequences (observed samples), and $X = A, C, T, G$.

The likelihood becomes smaller proportionate to the size of your data set. When you multiple together more and more numbers that are smaller than 1, the likelihood will continue to get smaller and smaller. It doesn't mean this tree is incorrect as it could still be the tree with the **largest** likelihood relative to the others.

(b)

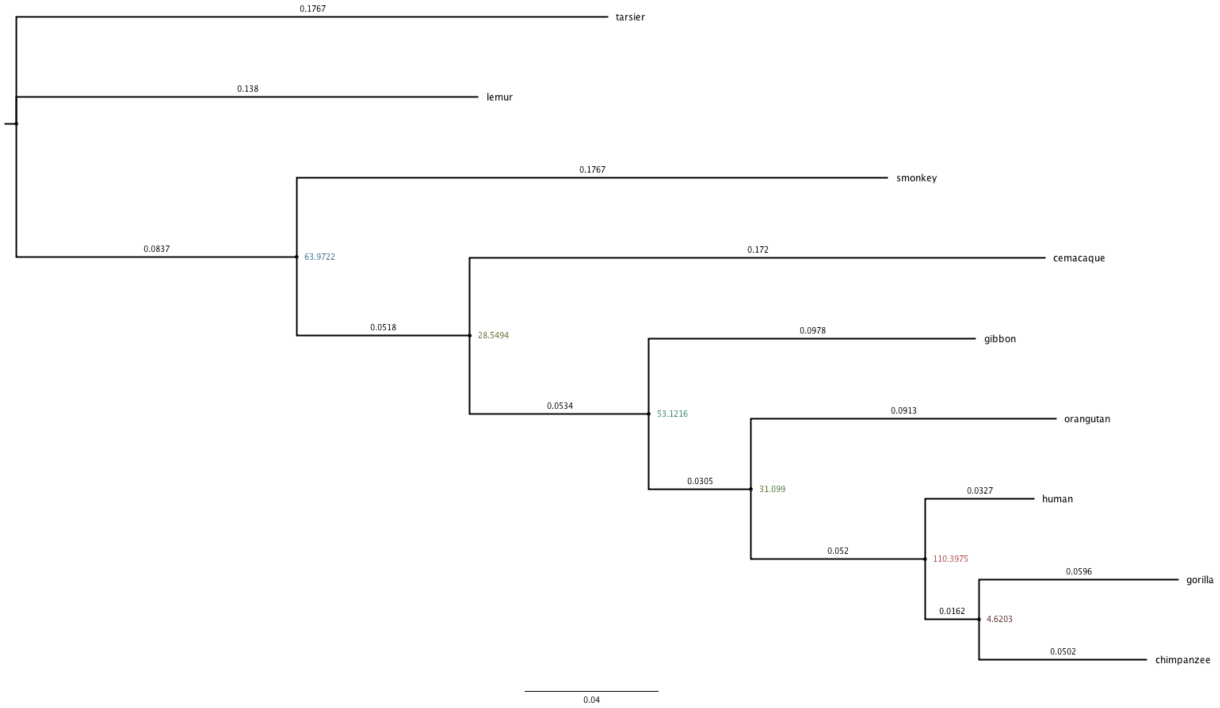


Figure 1: JC69 Tree

The sister group of chimpanzee in this analysis is gorilla with an approximate LRS of 0.9092.

Problem 4

The command run to generate the output was:

```
phyml -i mito3.phy -q -d nt -m HKY85 -f d -c 1 -s SPR --print_site_lnl \
-b -1 --run_id HKY85
```

(a)

Whereas the JC69 model has fixed equilibrium frequencies and rate parameters, the HKY85 model has independent equilibrium frequencies and two rate parameters. Thus we have $6 - 2 = 4$ additional degrees of freedom for HKY85 compared to JC69.

(b)

The log-likelihood of the HKY85 model is -5234.64610. Since the difference of the logs of two numbers is the same as the log of the ratio of those numbers, we can calculate that the $LRS = e^{-5234.64610-(-5569.51634)} = 2.7058105 \times 10^{145}$

(c)

The sister group of chimps in this analysis is human. The approximate LRS is 0.7889.

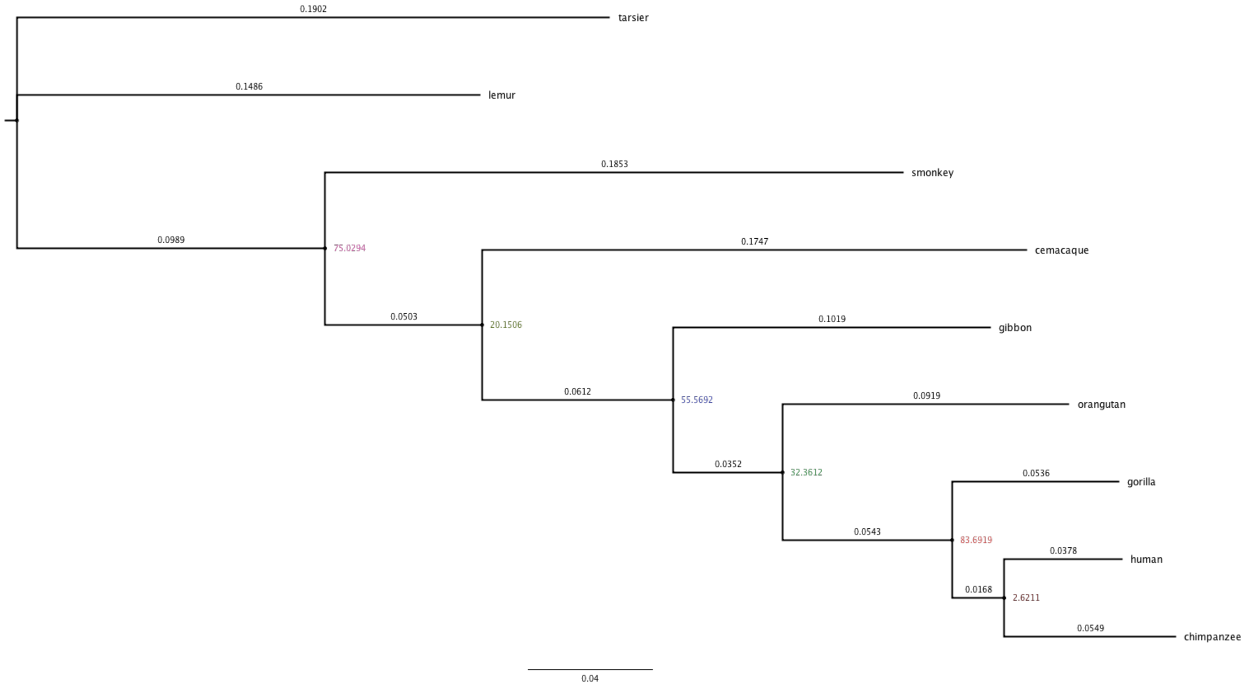


Figure 2: HKY85 Tree

(d)

The estimate of the nucleotide frequences was:

```
- f(A)= 0.32195
- f(C)= 0.30443
- f(G)= 0.10761
- f(T)= 0.26602
```

Problem 5

The command run to generate the output was:

```
phyml -i mito3.phy -q -d nt -m HKY85 -f d -c 4 -s SPR --print_site_lnl \
-b -1 -a e --run_id HKY85_cat4
```

(a)

This model differs from the last in that we're now estimating the most likely value for alpha within the gamma distributed among-site variation model for independent branch lengths per site. Previously we assumed that the probability of substitution along any branch is the same for every site in our sequences. Now we're estimating alpha and discretzing that distribution to properly weight slow and fast evolving sites accordingly in our branch length calculations. This has one extra degree of freedom due to the estimation of alpha.

(b)

The log-likelihood for including the four-category discrete gamma distribution model is -5042.89454 making it the best model. The LRS compared to each previous model is:

Assuming H_1 = HKY85_Gamma_4 and H_0 can be either JC69 or HKY85:

$LRS = 1.8907948 \times 10^{83}$ when $H_0 = HKY85$ and $LRS = 5.1161326 \times 10^{228}$ when $H_0 =$ JC69.

(c)

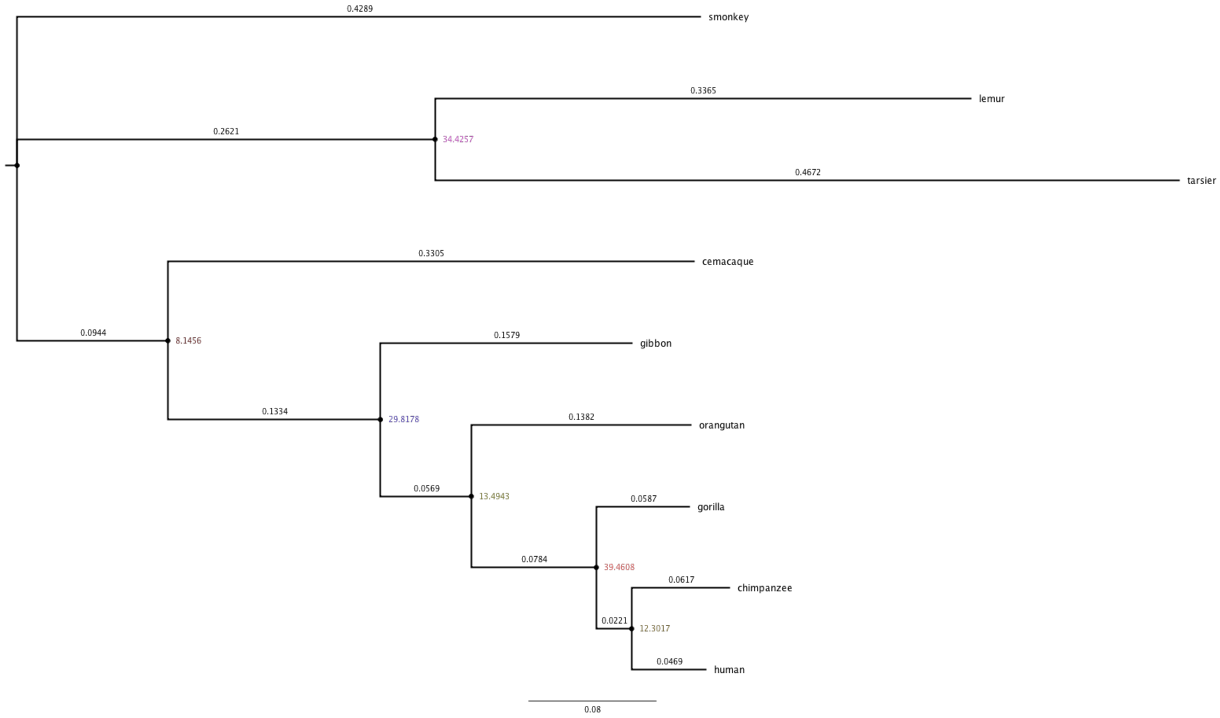


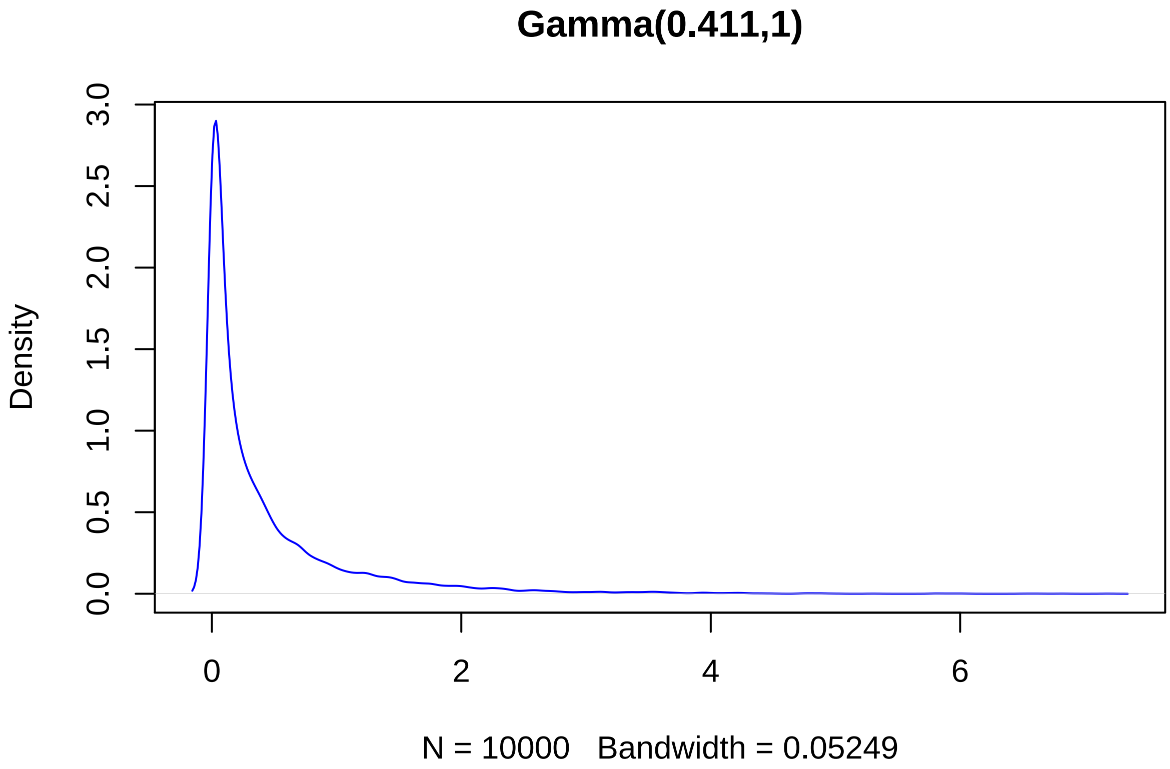
Figure 3: HKY85 with Discrete Gamma 4 Tree

The sister-group of chimps in this analysis is human. The support for this grouping is a LRS of 12.3017.

(d)

The estimate for alpha is 0.411. A gamma distribution with alpha = 0.411 and beta = 1 looks like the following:

```r
plot(density(rgamma(10000, shape = 0.411, rate = 1)), col = "blue", main = "Gamma(0.411,1)")
```



Problem 6

The best model of sequence evolution is the HKY85_Gamma_4 model based on the LRS compared to the other two models tried.

Problem 7

Based on these analyses, I would conclude that humans are the sister group of chimps because in both the HKY85 and HKY85_Gamma_4 models, human was chosen as the sister group to chimp. Both models proved to be better options than the JC69 model according to the LRS.

Part 2. Ancestral reconstruction

Here is an image of the tree:

Problem 1

The mean posterior probability for AncSR1 is 0.99913 and for AncSR2 is 0.98882.

Problem 2

The mean posterior probability of the entire sequence for AncSR1 is 0.93578 and for AncSR2 is 0.36853. Compared to question 1, the per site and entire sequence posterior probabilities are nearly the same, but for AncSR2, the entire sequence probability is much lower.

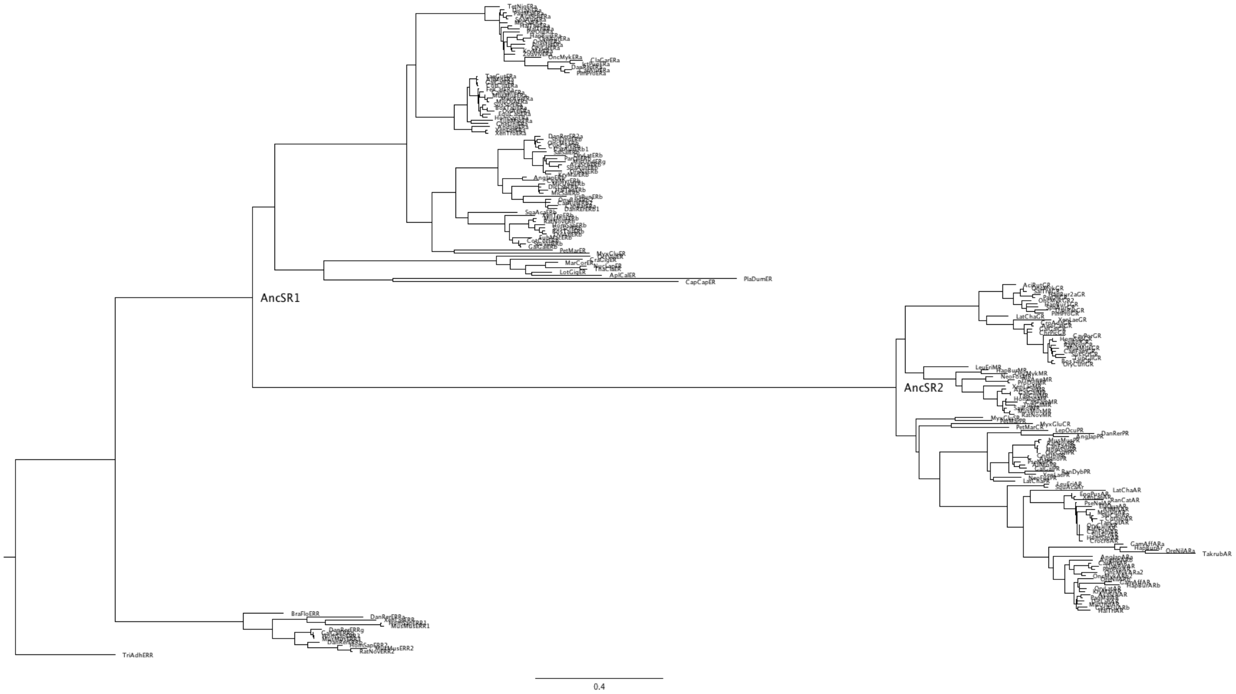


Figure 4: SR Tree

Problem 3

Position 62 is an ambiguous site in AncSR2 with amino acids F (Phenylalanine) and Y (Tyrosine) have posterior probabilities of 0.582 and 0.308, respectively.

Problem 4

34 replacements occured in the DBD between AncSR1 and AncSR2:

```
Branch 203:  220..308

     3 V 0.992 -> I 0.847
     7 N 0.352 -> E 1.000
    11 F 0.927 -> C 1.000
    16 W 0.543 -> L 0.994
    17 S 1.000 -> T 1.000
    19 E 1.000 -> G 1.000
    20 G 0.994 -> S 1.000
    23 A 1.000 -> V 1.000
    28 S 0.992 -> A 1.000
    29 I 0.994 -> V 0.999
    30 Q 0.999 -> E 1.000
    32 H 0.502 -> Q 0.999
    33 V 0.721 -> H 0.999
    34 D 0.834 -> N 1.000
    36 M 0.390 -> L 1.000
    38 P 1.000 -> A 1.000
    39 A 0.984 -> G 1.000
```

```
40 T 0.972 -> R 1.000
42 Q 0.506 -> D 1.000
44 T 0.904 -> I 1.000
48 H 0.606 -> I 0.999
52 S 0.998 -> N 1.000
54 Q 1.000 -> P 1.000
62 Y 0.685 -> F 0.582
63 E 0.875 -> Q 1.000
64 V 0.997 -> A 1.000
67 M 0.661 -> T 0.803
68 K 0.953 -> L 1.000
70 G 0.951 -> A 0.999
71 V 0.593 -> R 0.997
72 R 0.997 -> K 0.994
73 K 0.645 -> S 0.959
74 D 0.620 -> K 1.000
75 R 0.970 -> K 0.999
```

Problem 5

Sites 62, 63, 70, and 72 are conserved in most decendents of AncSR2.

Problem 6

The side chains I identified are colored in cyan below.

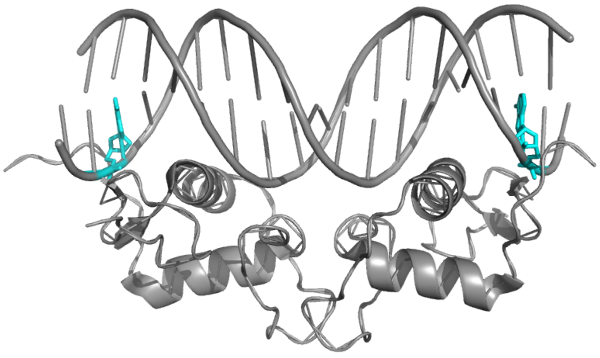


Figure 5: SR Tree

Problem 7

Based on the residues that have remained highly conserved in decendents of AncSR2, based on the structure provided, these residues directly contact DNA on the far sides of the DBD. One hypothesis might be that all, some, or one of these sites was mutated during the evolution of AncSR1 to AncSR2 that caused a change in DNA binding specificity.

Problem 8

To test that experiment, we could synthesize the ancestral protein and transfect it into living decendents followed by biochemical characterization. One would have to make several synthetic versions corresponding to the different combination of potential substitutions in AncSR2 relative to AncSR1 that could have caused the change in DNA specificity.