HW3_Submission

Justin Rigby

11 February 2018

1.

a.

$L_{JC69} = (0.25)^{10} = 9.5 * 10^{-7}$

b.

$\pi_A = 0.3 \quad \pi_T = 0.3 \quad \pi_G = 0.2 \quad \pi_C = 0.2$

$L_{84} = (0.3)^3 * (0.3)^3 * (0.2)^2 * (0.2)^2 = 1.1664 * 10^{-6}$

c.

$\frac{L(M_{69})}{L(M_{84})} = \frac{9.5*10^{-7}}{1.1664*10^{-6}} = 0.8144719$
From this we can see that the HYK85 model is a better fit for the sequence.

$-2 * ln(\frac{9.5*10^{-7}}{1.664*10^{-6}}) = 0.4104307$

Degrees of Freedom = 1

```
p.val <- (1-pchisq(0.4104307, df= 2))
p.val
```

```
## [1] 0.8144719
```

Since the null model was not refuted the most likely model is the simplest – the JC69 model.

Comment: Your likelihood-ratio statistic (LRS) is wrong. Check what mistake you made. -0.5

2. Run the analysis for the JC69 model.

a. What is the ln-likelihood of the tree and JC69 model? Write the generalized expression for the likelihood (don't try to calculate it: it's too small!). Why is it so small? Does this mean the tree is almost certainly incorrect (2pt)?

$L(JC69) = (\pi_{A,T,G,C})^N$

Log-likelihood = -5569.51599

No it does not mean that at all. What it means is that, for this relative seuqence, this is the relative measurement of the log-likelihood for the JC69 model for this sequence.

b. What is the sister-group of chimpanzee in this analysis? What is the value of the approximate likelihood ratio statistic supporting this grouping (1pt)?

The sister group is Gorilla.
Log-likelihood value = 4.6045

4.

a. How many extra degrees of freedom does this model have compared to JC69 (1pt)?

It has three extra degrees of freedom since each base has an independent value.

Comment: Check HKY85 model. -1

b . What is the ln-likelihood of the model? Calculate the LRS relative to JC69 and determine which is the better model (2pt).

HYK85
Ln-likelihood = -5232.71479

JC69
Ln-likelihood = -5569.51599

$LRS = -2 * ln(\frac{L_{Null}}{L_A}) = -2 * (ln(L_N) - ln(L_A))$

$LRS \ = \ -2(ln(L_{JC69}) - ln(M_{HYK85})) = -2(-5569.51599 - (-5232.71671)) = 673.5986$
Degrees of Freedom = 1

```
p_value <- (1-pchisq(673.5986, df = 4))
p_value
```

```
## [1] 0
```

The alternative model (HYK85) is more likely due to the significant p-value which rejects the null model (JC69).

Comment: You are using wrong likelihood for JC69 (check 2.(a)), and also a wrong formula for doing likelihood-ratio test. -1.5

c. What is the sister-group of chimps in this analysis? What is the value of the approximate likelihood ratio statistic supporting this grouping (1pt)?

The sister group to chimpanzees is Humans. The likelihood ratio statistic is 3.2092.

d. Examine the _stat file. What is the maximum likelihood estimate of the frequency of the four nucleotides in the data set (1pt)?

f(A)= 0.31363
f(C)= 0.29735
f(G)= 0.10337
f(T)= 0.28565

5

a. How does this model differ from the model you used in the last problem? How many extra degrees of freedom does it have (2pt)?

It differs due to the use of a gamma distribution in conjunction with a standard HKY85 model. It also gains one additional degree of freedom.

Comment: Check the gamma distribution of rate to find the extra degree of freedom. -0.5

b . Calculate the LRS and evaluate whether which of the three models you have evaluated is the best model (1pt).
$LRS = -2ln(\frac{M_{Null\ Model}}{M_{Alt\ Model}}) = 2ln(\frac{M_{Alt\ Model}}{M_{Null\ Model}}) = 2(ln(M_{Alt\ Model}) - ln(M_{Null\ Model}))$

HYK85
Ln-likelihood = -5232.71479
4 Degrees of Freedom

JC69
Ln-likelihood = -5569.51599
1 Degree of Freedom

HYK85 Gamma
Ln-likelihood = -5042.90027
5 Degree of Freedom

$LRS = -2 * ln(\frac{L_{Null}}{L_A}) = -2 * (ln(L_N) - ln(L_A))$

$LRS = -2(ln(L_{JC69}) - ln(M_{HYK85})) = -2(-5569.51599 - (-5232.71671)) = 673.5986$
Degrees of Freedom = 1

```r
p_value <- (1-pchisq(673.5986, df = 1))
p_value
```

```
## [1] 0
```

$LRS = 2ln(M_{HYK85} - M_{HYK85\ Gamma}) = -2 * (-5232.71479 - (-5042.90027)) = 379.629$
Degrees of Freedom = 4

```r
p_value <- (1-pchisq(379.629, df = 2))
p_value
```

```
## [1] 0
```

From the log likelihood ratio statistic we can see that the best model for this systems is the HYK85-Gamma model due to its lower p-value.

Comment: Again, you are using a wrong formula for likelihood-ratio test. Also, the likelihood values are slightly different from the answers. -1

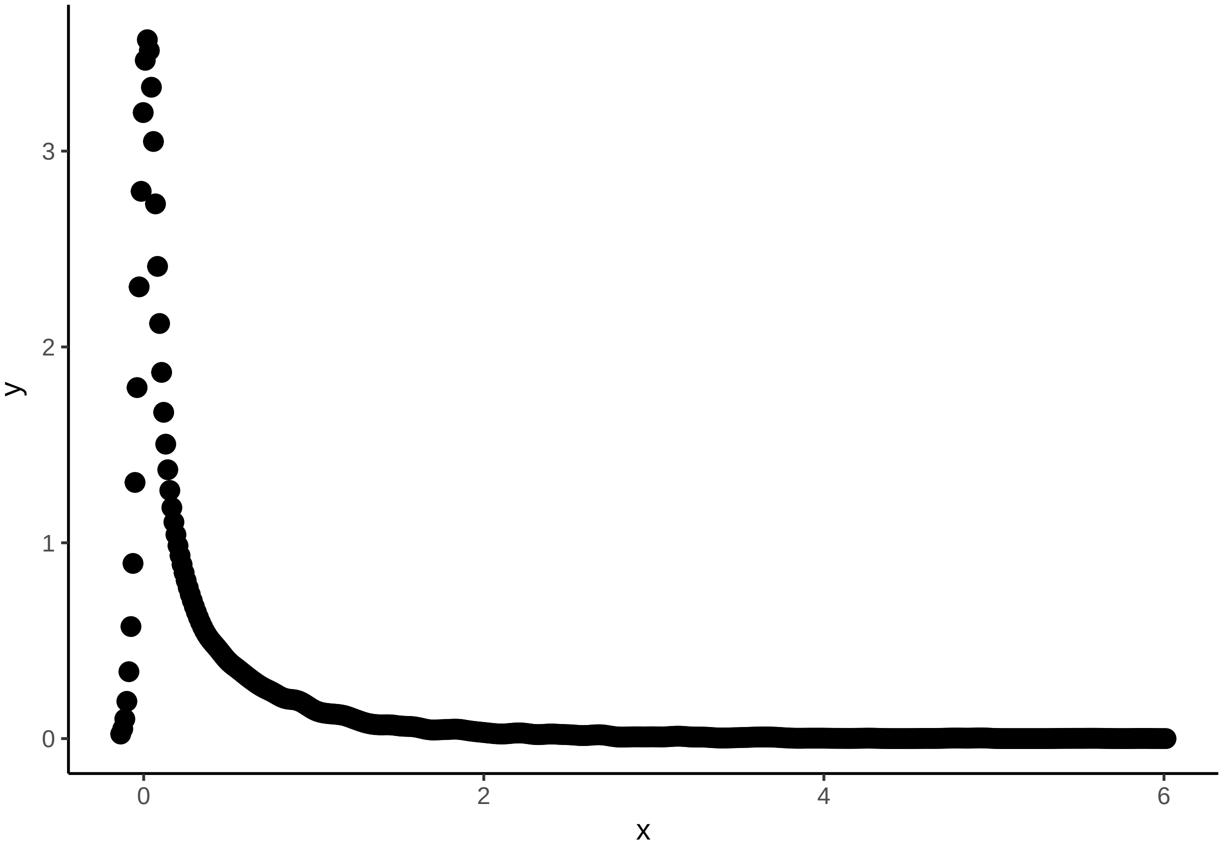
c. What is the sister-group of chimps in this analysis? What is the support for this grouping (1pt)? The sister group for chimps is humans with a ln-likelihood of 12.2718.

d. Examine the _stat file. What is the estimate of alpha, the shape parameter of the gamma distribution? What is the approximate shape of the distribution? (You can use a reference here or plot the distribution given alpha. An approximate shape is sufficient.) (1pt)

$\alpha = 0.411$

```r
library(ggplot2)
  library(MASS)
r <- seq(0, 3, length =10000)
xr <- rgamma(r, shape = 0.364)
den <- density(xr)
dat <- data.frame(x = den$x, y = den$y)

ggplot(data = dat, aes(x = x, y =y)) + geom_point(size = 3) + theme_classic()
```



Comment: This is not the value you should be getting. Combined with the slightly different likelihood, this seems to say that you are using slightly different options for running PhyML. -0.5

6. From all the models you have compared, what do you consider the best model of sequence evolution to be (1pt)?

From the log-likelihood ratios I find that the non-gamma HYK85-Gamma is best suited for this particular data set.

Comment: If you get the right answers and do correct test, this will not be your answer. -1

7. Based on these analyses, would you conclude that gorillas or humans are the sister group of chimps? Justify your answer (1pt)

The null model (JC69) was rejected with each comparison of the models, so I am more inclined to say that the sister group to chimps is humans rather than gorillas.

1. At sites in the DNA-binding domain, what is the mean posterior probability across sites for AncSR1? And for AncSR2? (Hint: this information is labeled as "Overall accuracy of the 204 ancestral sequences . . . for a site" near the bottom of the rst file) (1pt).

AncSR1
Node Label = 220
Site Location = 14

AncSR2
Node Label = 308
Site Location = 102

AncSR1 = 0.90754
AncSR2 = 0.98881

2. What is the posterior probability of the entire sequence for AncSR1 and for AncSR2? That is, given the model, data, and tree, what is the probability that these are the correct sequences? How does this number relate numerically to the mean PP across sites reported in question 1? (Hint: this information is labeled as “Overall accuracy of the 204 ancestral sequences … for the sequence” near the bottom of the rst file) (1pt).

AncSR1 = 0.00012
AncSR2 = 0.36840

It is lower than the posterior probability for the site with the AncSR1 being very small.

3. AncSR1 has 7 ambiguously reconstructed sites and AncSR2 has 1 ambiguous site, if we define ambiguous as having a second-best reconstruction with posterior probability >0.2. Which position in AncSR2 is ambiguously reconstructed, and what are the two ancestral states with reasonable statistical support? (Hint: the “Prob distribution at node xxx, by site”block lists the PP of each of the 20 amino acids at each position in the reconstructed sequence) (1pt).

Ancr2
Site 62 is the site at which there is more than P>0.2 for two residues. The residues are;
Phenylalanine (F): 0.582
Tyrosine (Y): 0.308

4. How many replacements occurred in the DBD between AncSR1 and AncSR2? (Hint: this information is nicely summarized in the section of the rst file designated by Branch xxx:yyy..zzz where yyy represents the node label of AncSR1, and zzz represents the node label of AncSR2) (1pt).

There was a total of 34 amino acid substitutions between the two sequences.

5. How many of these are conserved in one state in most or all of the descendants of AncSR2? (Hint: this can most easily be diagnosed by opening the SR220.phy file in an alignment editor such as Sea-view, Aliview, Jalview or Mesquite, or even by writing a simple script given the Python tools you have used previously) (2pt)

Sites conserved in most of AncSR2 Sequences; 3, 7, 11, 16, 20, 28, 29, 30, 32, 33, 34, 36, 39, 40, 42, 44, 48, 52, 62, 63, 64, 67, 68, 70, 71, 72, 73, 74, 75

Sites conserved in All of AncSR2 Sequences; 17, 19, 23, 38, 54

Comment: This is not an exhaustive list, but I will pass this.

6. Load 2C7A.pdb into pymol. This is the structure of the progesterone receptor DNA-binding domain bound to DNA (as a dimer, on a palindromic response element). Show the structure as a cartoon. Make an object consisting of the residues that changed between AncSR1 and AncSR2; show these side chains and give them a unique color (2pt).

Residues;
3+7+11+16+17+19+20+23+28+29+30+32+33+34+36+38+39+40+42+44+48+52+54+62+63+64+67+68+70+71+72+73-

hide everything
bg_color white
show cartoon, 2C7A
color gray70, 2C7A
select resi 3+7+11+16+17+19+20+23+28+29+30+32+33+34+36+38+39+40+42+44+48+52+54+62+63+64+67+68+70+7
show spheres, sele; color green, sele

7. Formulate a hypothesis for which replacements caused the shift in specificity. Justify your answer (3pt).

It is possible that these residues actually conferred a more stable confirmation for the double helix portion of the protein which reduced the amount of time that it spent in the other confirmations. This lack of staying in the stable conformation could prevent the protein to change its specificity for its substrate.
Most of the residues lie at the beginning, middle, and end of the double helix. The middle portion of residues

lie at opposite ends of the helix, and may be connected to one via salt bridges or through electrostatic interactions (Hydrogen Bonding and Van De Waals). The change of these residues could prevent these interactions from occurring and its time spent in the confirmation.

8. Propose an experiment (or experiments) to test that hypothesis (3pt).

First, the sequences should be transformed into a plasmid alongside an over expressing promoter. The next step would be to change those residues by exposing the plasmids to a mutagenizing agent in order to cause single base-pair changes in the sequences. After-which, using PCR and Sanger sequencing, identify the proteins which had mutations occur in the same region as the residues (may require multiple mutagenizing events) and focus on getting mutations in the regions related to the double helix residues in the middle of the protein where there are residues on both sides of the double helix.
Once that is done and you have the proper mutate sequences the next step is to purify the protein and analyze its crystal structure. From there you can decipher if the mutations are causing a confirmation change from the crystal structures recorded – or you could just run a folding algorithm on the sequences on a cluster. Once the folding confirmations are confirmed you can add the different experimental proteins and exposed them to their respective substrate and record the activity of them.