Assignment 4: Phylogenetics

Abdullah Faqih Al Mubarok - vpx267

November 10, 2023

Part 1: Maximum likelihood analysis

1. Why could third codon positions in protein-coding data be problematic in phylogenetic inference, and does that mean it would be desirable to exclude them before analysis? (1-2 sentences, 3 points)

   Answer:

   The third codon has the highest substitution rate than the others. This would lead to wrong phylogenetic which based only from the sequence

2. Paste the IQ-TREE command here (1 points)

   Answer:

   ```
   iqtree2 -s regier.nex -m HKY+G  -spp part12.nex
   -bb 1000 -pre regier_12_ -nt 4
   ```

3. What is the frequency for adenine (A) across the alignment? (2 points):

   Answer:

   From the .iqtree file, the frequence of A is 0.312881

4. What is the value for the shape parameter alpha of the gamma distribution? (2 points):

   Answer:

   The value of alpha of the gamma distribution is 0.310154

5. What was the estimated ratio of transitions to transversions? (2 points):

   Answer:

   The estimated transitions/transversion is 1.96312

6. Paste the inferred newick tree here (1 points):

   Answer:

   (((ThulTARD:0.12039484030000003,MtdTARD:0.1225671452):0.16465804170000004,(((((TorPYCNO:0.054804379299999983,AeliPYCNO:0.03751191840000001):0.011850918300000013,AhiPYCNO:0.03969483629999998):0.010167561500000005,Ele

```
PYCNO:0.06045363670000001):0.01436093590000001,Col2PYCNO:0.06166487250
000002):0.08661913399999999,(((((LpoXIPHOS:0.011316868400000013,Cro2XIPH
OS:0.011874586799999975):0.0787960948,((((((MgaARACH:0.04813309730000001,
StpARACH:0.0624391624):0.015672708099999988,Pma2ARACH:0.0499085011000
0003):0.012043593699999994,AchARACH:0.0822366349):0.01039491330000003,(H
ariARACH:0.022638747899999978,HspARACH:0.02193941319999998):0.06356217
010000004):0.014094677399999977,(LnigARACH:0.10420489779999997,EgigARAC
H:0.06710466530000003):0.014303108499999995):0.005900759400000011,(Crp2AR
ACH:0.07550597270000003,PwhARACH:0.11742563439999998):0.01165233220000
0004):0.007476167899999997):0.0080229,Din2ARACH:0.1721293842):0.0053012899
00000006,(Amb2ARACH:0.09494611659999996,IpumARACH:0.1254456285999999
7):0.017625795200000016):0.026362793399999973,((((((LfoCHILO:0.058088731500
000046,SpoCHILO:0.06702977110000002):0.015565266799999977,ScolCHILO:0.072
04761479999999):0.008828663200000031,Ctas2CHILO:0.07536465990000002):0.025
09840879999997,(((NamDIPLO:0.06867609879999997,AmaDIPLO:0.064175644799
99999):0.014677816399999988,Pge2DIPLO:0.08107998529999999):0.017172144399
999978,Pol2DIPLO:0.1045355765):0.013037350400000036):0.009491998499999987,
((Han2SYMPH:0.07048765729999995,Scu3SYMPH:0.05606145549999997):0.04462
271750000002,EuryPAURO:0.14285372220000003):0.012171392599999997):0.0143
00909799999983,(((OstOST:0.15743881670000004,DtyMYSTACO:0.151946531999
99997):0.015752290999999974,((Arg2BIURA:0.10273635539999998,AarPENTA:0.
17091366619999998):0.04805038660000005,(HapaOST:0.0791003495,SkleOST:0.05
743093389999998):0.07670382090000005):0.013095239899999944):0.0151654068000
00003,(((((MesoCOPE:0.02577204649999998,A369COPE:0.0416537783):0.0725879
3260000002,EafCOPE:0.1396552869):0.06697035709999999,((NheMALA:0.0922348
9629999998,((LemMALA:0.0640345366,Avu3MALA:0.10559239689999997):0.0123
9000139999999,NeoMALA:0.076549125):0.017976417500000008):0.07239133730000
002,(((BbaTHECOS:0.026986401199999982,CfrTHECOS:0.027144243099999993):
0.0101787884,LeanTHECOS:0.04485327569999997):0.028588130200000006,LoxTH
ECOS:0.1217667369):0.081317466):0.011532978199999988):0.012094388899999975
,((TloBRANCH:0.10485274600000002,((DmaBRANCH:0.07636654359999998,Lle2
BRANCH:0.05112762589999997):0.01995543710000003,LynBRANCH:0.089298896
60000001):0.01741944510000004):0.019527008999999984,(Asa3BRANCH:0.063237
16809999999,ufsBRANCH:0.03749004139999998):0.1029818103):0.02395845619999
9978):0.010209260300000023,((StuREMI:0.08755072869999997,HmaCEPHAL:0.11
710509259999996):0.01418615330000006,(((PaqCOLL:0.10719381589999999,(Oim
COLL:0.08099389699999998,Tom2COLL:0.06511380150000001):0.02335750919999
996):0.08117643070000002,(JapDIPLUR:0.08014016359999998,EfrDIPLUR:0.1182
0350149999997):0.017390664200000017):0.008504334499999988,((PsaARCHEO:0.0
1416774110000002,MbaARCHEO:0.012366443800000015):0.07376980849999998,((
CliZYGEN:0.05463287259999999,NmeZYGEN:0.05118357069999996):0.024010856
800000036,((MayEPHEM:0.0411245261,EinEPHEM:0.04641966850000001):0.0615
3045260000001,((IveODONAT:0.0443690283,LlyODONAT:0.036575897800000035
):0.06303897549999998,(((ApaukNEOPT:0.04563680530000003,CpoNEOPT:0.032
7702835):0.029074999999999962,PquNEOPT:0.041544444999999985):0.111022231
20000004,(FauNEOPT:0.11346716480000002,(PamNEOPT:0.041002043200000005
```

,AdoNEOPT:0.06716830810000002):0.0126800299):0.011697055999999983):0.0069
2646210000003):0.005600474899999985):0.006253184799999978):0.01052835330000
0015):0.017407814700000024):0.009206904699999963):0.007993899000000027):0.00
8438700000000021):0.0379223595):0.011344494799999993):0.01157097600000001):0
.02133225759999996):0.03547401315,((Pno2ONYCH:0.022241230599999995,ErwO
NYCH:0.008119459399999973):0.03470012289999999,PepONYCH:0.0401822439):0
.03547401315);

7. Looking at the clade formed by terminals ending in XxxDIPLUR (short for Diplura or bristletails), which clade is their sister? (List the terminal names of the sister group; 2 points).

   Answer:

   There are: PaqCOLL, OimCOLL, Tom2COLL

8. Using the complete definition of a monophyletic group, look at the terminals ending in XxxNEOPT (short for Neoptera, a group of winged insects) and explain: does this tree support Neoptera as a monophyletic group? (Yes/No and why; 3 points).

   Answer:

   Yes it is. the XxxNEOPt are a monophyletic group since they share the same MRCA.

9. Looking at all the terminals ending in XxxARACH (short for Arachnida, spiders), does this tree support spiders as a monophyletic group? (Yes/No and why; 3 points).

   Answer:

   No they are not since the MRCA for all XxxARACH include XxxXIPHOS

Part 2: ASTRAL analysis and RF distances

10. Run ASTRAL as we did in class on the regier.gene.trees file. Paste the ASTRAL command here (1 points):

    Answer:

    ```
    java -jar Astral/astral.5.15.5.jar
    -i regier.gene.trees -o regier.gene.species.tre
    -T 4 > regier.gene.astral.log
    ```

11. Paste the resulting newick tree string here (1 points):

    Answer:

    (AarPENTA,(Arg2BIURA,((((((HmaCEPHAL,((DtyMYSTACO,((JapDIPLUR,(
    OstOST,(LoxTHECOS,(LeanTHECOS,(BbaTHECOS,CfrTHECOS)0.9:0.2579462
    326917244)1:0.6131044728864088)1:2.015354133102803)0.7:0.13919500218232522)0
    .94:0.2617485026787862,(EafCOPE,(A369COPE,MesoCOPE)1:2.179525000236819

)1:0.6309259731654561)0.5:0.0640038494982401)0.47:0.10439583424976343,(PquNEOPT,(ApaukNEOPT,CpoNEOPT)1:0.9270779974933916)1:2.1711067312023538)0.8:0.20492161245496154)0.77:0.15659725710423306,(NheMALA,(Avu3MALA,(LemMALA,NeoMALA)0.8:0.13855364459033945)0.63:0.09299136469708288)1:1.3969055987237824)0.38:0.04913238675333127,((Asa3BRANCH,ufsBRANCH)1:2.8455295153324665,(TloBRANCH,(LynBRANCH,(Lle2BRANCH,DmaBRANCH)1:0.4582052925409832)0.81:0.16444962380621866)1:0.7381386032175303)1:0.3235588711977975)0.73:0.13491000171501777,(HapaOST,SkleOST)1:1.0118872735596771)0.62:0.08879940999650603,((StuREMI,(((((IveODONAT,LlyODONAT)1:1.788827001369514,(FauNEOPT,(AdoNEOPT,PamNEOPT)0.66:0.0940259144302094)0.99:0.296854074557852)0.84:0.14868880340129356,(EinEPHEM,MayEPHEM)1:1.8853248309082922)0.44:0.07089754999969072,(CliZYGEN,NmeZYGEN)1:0.41791952931726517)0.7:0.1088044146904491,(MbaARCHEO,PsaARCHEO)1:3.4231762883809314)1:0.49591927954738935)0.6:0.06918113355472015,(EfrDIPLUR,(PaqCOLL,(OimCOLL,Tom2COLL)0.84:0.20321688304635005)1:1.3708234817375347)0.74:0.14051735950832323)0.69:0.09063459188189088)0.83:0.14368948752087068,((((((Din2ARACH,((PwhARACH,(((HariARACH,HspARACH)1:3.091042453358317,(AchARACH,(Pma2ARACH,(MgaARACH,StpARACH)1:0.5200044000945292)1:0.4120015092708607)0.57:0.06710876048911354)1:0.3607003712330088,(Cro2XIPHOS,LpoXIPHOS)1:3.3929166902645442)0.88:0.1816951471044377)0.37:0.007842276045730205,(LnigARACH,EgigARACH)0.45:0.03105243345556902)0.95:0.21025065363247483)0.88:0.18680232319521617,(Crp2ARACH,IpumARACH)0.61:0.08559258044288205)1:0.3614172184956892,((Col2PYCNO,(ElePYCNO,(TorPYCNO,(AeliPYCNO,AhiPYCNO)0.71:0.13361626298733612)0.98:0.28776589704785277)0.61:0.08254583789165482)1:2.3506457279215396,(PepONYCH,(ErwONYCH,Pno2ONYCH)1:1.7047480922384253)1:2.716839140927625)0.39:0.03430571175380222)1:0.31642311244111654,(EuryPAURO,((Pol2DIPLO,(Pge2DIPLO,(AmaDIPLO,NamDIPLO)0.84:0.2504116510748734)0.9:0.20841720892919036)0.59:0.07141936560271873,(Ctas2CHILO,(ScolCHILO,(LfoCHILO,SpoCHILO)0.95:0.25820608923625615)0.95:0.22181491556424696)1:0.4297521010375983)0.74:0.11195926010977736)0.2:0.07269074245584187)1:0.3294316951496192,(Han2SYMPH,Scu3SYMPH)1:0.7899852185996011)0.52:0.04905348323418594,(Amb2ARACH,(MtdTARD,ThulTARD)1:2.0734044383934935)0.43:0.04094979932086465)0.91:0.2339324044375095)1:0.8876216504443147):0.0);

12. What units do the branch lengths have? (2 points):

    Answer:

    The branch length from the ASTRAL output is in coalescent time unit ($\frac{t}{Ne}$) where t indicates the time since divergence and Ne is the sample size.

13. What is the branch length leading to the group of terminals ending in XxxCOLL (i.e., the stem length)? (1 value; 2 points):

    Answer:

    The branch length is 1.3708

14. If you were to simulate some alignment data in IQTREE2 using this tree from ASTRAL, what command would you use? You can decide on the model and alignment length. (paste command; 2 points).

    Answer:

    Here, I used the HKY for the base substitution rate that allows different transition/transversion rate. In addition, it would be a good idea to allow different rate among sites we would use discrete gamma model. In addition, I simulated it with 10,000 sequence length. Below is the command to execute it:

    ```
    iqtree2 --alisim sim_regier_alignment
    -m HKY+I+G4 -t regier.gene.species.tre
    --length 10000
    ```

15. Calculate the Robinson-Foulds (RF) distance between the ASTRAL species tree you computed and the IQ-TREE concatenation tree. Use IQ-TREE (version 1.6, not IQTREE2 because it has a bug in RF calculation!). Paste your command below (1 points).

    Answer:

    ```
    iqtree -rf regier.gene.species.tre regier_12_.treefile
    ```

16. Give the RF distance between the ASTRAL tree and the IQ-TREE tree (1 number; 1 points).

    Answer:

    The RF distance is: 62

17. You are being asked to be a reviewer for a scientific article that includes a phylogenetic analysis. The method description is brief: "To obtain a phylogenetic hypothesis, the 2000 genes were concatenated (400,567 amino acids total) and were analyzed in a maximum likelihood framework with IQ-TREE v.1.6." Give two suggestions to the authors of additional steps to improve their phylogenetic analysis (2 sentences; 5 points)?

    Answer:

    The concatenated model, having a uniform evolution model for all genes, does not reflect the evolution process since each gene is expected to have different selection pressure. Thus, there are two additional steps: creating a **partition file** for identifying the location of the genes to allow different model for each gene and allowing IQTREE to search best possible model for each gene using greedy strategy with **-m MFP** command.

Part 3: Trait evolution

18. Plot the distribution of homeRange onto the phylogeny. (Give the command and resulting figure; 3 points)

    Answer:

```
library(phytools)
library(nlme)
data("mammal.data", "mammal.tree", package="phytools"
   )

mammal.data$lgBodyMass <- log(mammal.data$bodyMass)
mammal.data$lgHomeRange <- log(mammal.data$homeRange)


## Visualize log(homeRange) on tree
contMap(mammal.tree,
        setNames(mammal.data$lgHomeRange, rownames(
           mammal.data)),
        lwd=2, fsize=0.5)
```

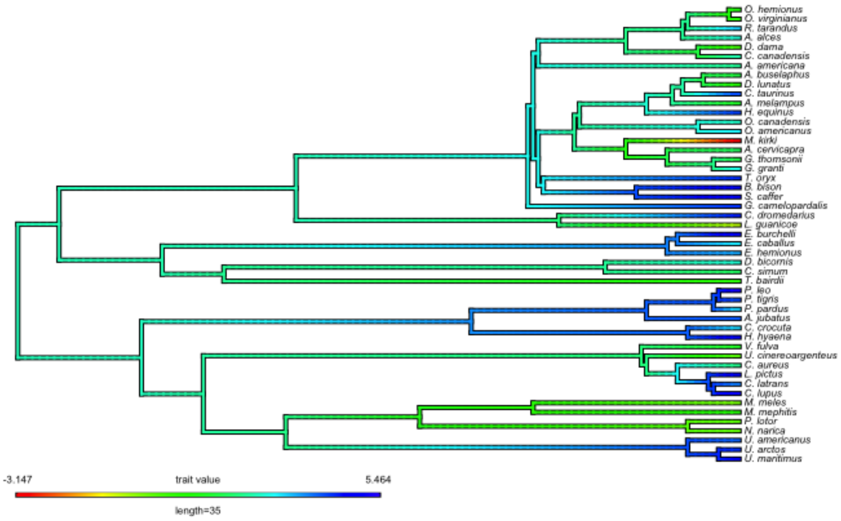

Figure 1: Log Home Range Distribution on Phylogeny

19. Test whether the two traits are correlated using PGLS. (Give the command and the output here; 3 points)

Answer:

Estimating linear model below with GLS and corBrownian (predicted covariance under a random brownian evolution model) :

$$log(homeRange) = log(bodyMass) + v$$

```
## Calculating the Brownian Correlation
### The form needs to be a var
species <- rownames(mammal.data)
corBM <- corBrownian(phy=mammal.tree,form=~species)
```

```
## Run GLS
pgls_fit <- gls(lgHomeRange ~ lgBodyMass,
                     data=mammal.data, correlation=
                       corBM)
summary(pgls_fit)
```

Based on the GLS results, The log(bodyMass) is statistically significant(p-value <0.001). Thus we could confidently reject the null (which stated that the $\beta_{log(bodyMass)} = 0$). Furthermore, the estimated effect ($\hat{\beta}_{log(bodyMass)}$) is 1.262. Which means every 1% increase in bodyMass, is expected to increase 1.262% of homeRange.

20. Test three models of continuous trait evolution (Brownian motion, Early Burst and Ornstein-Uhlenbeck) on the evolution of the mammal home range (log). (Give the command and the output; 3 points)

    Answer:

    ```
    dat <- setNames(mammal.data$lgHomeRange, rownames(
      mammal.data))

    fitBM <- fitContinuous(mammal.tree, dat, model="BM")
    fitEB <- fitContinuous(mammal.tree, dat, model="EB")
    fitOU <- fitContinuous(mammal.tree, dat, model="OU",
                           # Set the new max and min
                             value for the alpha
                           bounds=list(alpha = c(min=0,
                             max=exp(5))))

    aic <- setNames(c(AIC(fitBM), AIC(fitEB), AIC(fitOU))
       ,
                    c("BM","EB","OU"))

    print(aic.w(aic))
    print(aic)
    ```

BM	**EB**	**OU**
208.96 (0.7%)	210.96 (0.3%)	199.11 (99%)

    Table 1: AIC and its AIC.w (in parenthesis)

21. How would you choose the best model? (1 sentence; 1 points)

    Answer:

    To choose the best model we could select the model that has the lowest AIC together with its weighted AIC (probability of a model is really the best among others given data)

22. Which is the best model in this case? (1 word; 1 points)

Answer:

Ornstein-Uhlenbeck (OU) model

23. What was the initial value for the home range at the root of the tree? (give value; 2 points)

    Answer:

    The initial value of log(homRange) on OU model is 2.529030 which equals to homRange = 12.54134

24. Discuss your findings: What is the result in relation to the original prediction/hypothesis about traits (i.e., was the prediction met)? Give a suggestion for a possible next step to gain more confidence in this result. (2-3 sentences; 3 points)

    Answer:

    From the previous results, we could conclude that our hypothesis seems true (homeRange is correlated with bodySize). However, this simple model might over/under estimated the effect since there is another factor that can be added such as the habitat of the animals.