# 5550-B1-1E22; Advanced Bioinformatics for Next-Generation Sequencing

# Assignment 3: Part 2 Phylogenetics

**Total 50 points**.

**Part 1: Maximum likelihood analysis**

You are going to work on a dataset on arthropod relationships by Regier et al. Your task is to run a concatenation analysis with IQ-TREE2 and answer a few questions about the resulting phylogenetic tree. You will also run an ASTRAL analysis on a small set of gene trees and calculate the Robinson-Foulds distance between the resulting ASTRAL tree and the IQ-TREE2 file. Your answers can be brief and I am giving pointers to the length and format of the desired answers in parentheses at the end of the questions.

We are using a concatenated file called regier.nex.

wget -O regier.nex <https://sid.erda.dk/share_redirect/GkKgpLARjL>

Let’s investigate what happens when we exclude third codon positions from this matrix.

**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)

We tell IQ-TREE2 to only use first and second codon positions by setting up a partition file like this where it only considers every third position from position 1 and 2:

#nexus

begin sets;

charset part12 = 1-41974\3 2-41975\3;

end;

Verify that you follow how it skips every third codon position. Save that nexus text into a file (part12.nex) and use it as a partition file in IQ-TREE. We want to run this under the following parameters:

* Use the concatenated nucleotide alignment regier.nex file.
* Use the partition file part12.nex
* Use a HKY+G substitution model
* Estimate branch supports using 1000 ultrafast bootstrap replicates.
* Use 4 threads.

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

After the analysis is done, paste the newick file into [PRESTO](http://www.atgc-montpellier.fr/presto/). Re-root on the node containing all taxa ending in the code xxxONYCH (short for Onychophora, velvet worms). Answer the following questions regarding your tree.

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

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

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

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

**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).

**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).

**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).

**Part 2: ASTRAL analysis and RF distances**

Download a set of 68 gene trees for this dataset and generate a species tree with ASTRAL. The file contains 68 lines with each a newick tree file that will serve as the ASTRAL input file.

wget -O regier.gene.trees https://sid.erda.dk/share\_redirect/BuBQaOekUc

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

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

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

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

**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).

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!).

**15. Paste your command below** (1 points).

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

**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)?

**Part 3: Trait evolution**

You will work on a simple data set that compares the body size and home range size of various mammals. Why do some mammals have large home ranges while others occupy little space? Is this maybe simply a function of body size -- large animals need more space to find food? You will examine this question.

Make sure you **document all R code here**, so I can reproduce your analyses.

## Get traits & tree for primates

data("mammal.data", "mammal.tree", package="phytools")

Look at the data, note that the species names are rownames in mammal.data. Since these are sizes with orders of magnitude between them, it makes sense to log() transform them in all analyses.

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

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

**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)

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

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

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

**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)