Assignment 4.2 -- Phylogenomics (33.3 points)

We are going to work on the Regier et al. dataset on arthropod relationships. We have seen this study in class, where we learnt that “Crustacea” are paraphyletic because they include Hexapoda. Your task is to run a concatenation analysis with IQ-TREE 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 distance between the resulting ASTRAL tree and the IQ-TREE 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.

In the IQ-TREE analysis, we will generate a tree from an alignment excluding third codon positions from this matrix.

1. Why would one want to experiment with excluding third codon positions in protein-coding data for phylogenetic inference? (1 sentence suffices, 5 points)

DATA files

We are using a concatenated file called regier.nex.

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

We can tell IQ-TREE to only use first and second codon positions by setting up a partition file like this, which will tell IQ-TREE to only consider the first and second codon positions but not the third:

```
#nexus
begin sets;
charset part12 = 1-41974\3 2-41975\3;
end;
```

Do you see how it skips every third codon position? Save this text into a file (`part12.nex`) and use it as a partition file in IQ-TREE.

We want to run under the following parameters:

- Use the concatenated nucleotide alignment `regier.nex` file.
- Use the partition file `part12.nex` .
- We will run this under the HKY+I+G model `(-m HKY+I+G)`. I have previously tested that this is the appropriate model for the dataset using ModelFinder so you can skip the step to save time.
- Estimate 1000 ultrafast bootstraps.

- Use 4 threads.

2. Paste the IQ-TREE command here (3 points)**:**

3. Paste the resulting newick tree string here (1 point):

After the analysis is done, paste the newick file into PRESTO. Re-root on the node containing all taxa ending in the code xxxONYCH (short for Onychophora, velvet worms). Please answer the following questions regarding your tree.

4. Looking at the clade formed by terminals ending in XxxDIPLUR (short for Diplura or bristletails), which clade is their sister group? (List the terminal names of the sister group. 5 points).

5. Looking at the terminals ending in XxxNEOPT (short for Neoptera, a group of winged insects), does this tree support that Neoptera are a monophyletic group (= a clade)? (Answer with Yes/No. 5 points).

6. Looking at the terminals ending in XxxARACH (short for Arachnida, spiders), does this tree support that spiders are a monophyletic group (= a clade)? (Answer with Yes/No. 5 points).

ASTRAL analysis and Robinson-Foulds distance

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
```

7. Run ASTRAL as we did in class on the `regier.gene.trees` **file. Paste the ASTRAL command here** (3 points)**:**

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

9. Calculate the Robinson-Foulds (RF) distance between the ASTRAL species tree you computed and the IQ-TREE concatenation tree. It does not matter which one of the two trees you use as the reference tree (parameter `-r`**). Use TreeCmp and paste your command below** (3 points).

10. Give the normalized RF distance (RF(0.5)_toUnifAvg) as we did in class (give the value, 2 points).