Homework#3

Aksahy Sanjeev

May 2025

The data consists of the sequence of 40 genes from species, organized into 40 `.fasta` (One for each gene). Each file had 18 DNA sequences corresponding to each of the species(specified by `>` character).

1 Multiple Sequence Alignment

The first task was to do a multiple sequence alignment and to output the aligned data in `mult-fasta` format. I used a python script to do that.

2 Cleaning the data

Once the alginment is over I cleaned up the data in the following manner:

- `fasta_dict`: It reads a `fasta` file and returns a dictionary mapping sequence ids to their corresponding sequences.
- `remove_gap_columns`: This removes alignment columns that contain gaps (-) in any sequence from a dictionary of aligned sequences.
- `write_fasta`: Finally this function writes sequences from a dictionary to a `fasta` file.

Then I combined all the sequences of different genes, that belongs to a particular species to generate 18 sequences (each 40211bp long)- file named `species_aligned.fasta`

3 Phylogenetic tree

Then I converted `species_aligned.fasta` to `species_aligned.phy`, for easier Phylogenetic tree construction. I used the Montpellier Bioinforamtics platform to use phyML. You can find the results here.

I also computed the distance matrix and used it to build a tree using the UPGMA method (using the code you shared).

Both of these trees are given below:

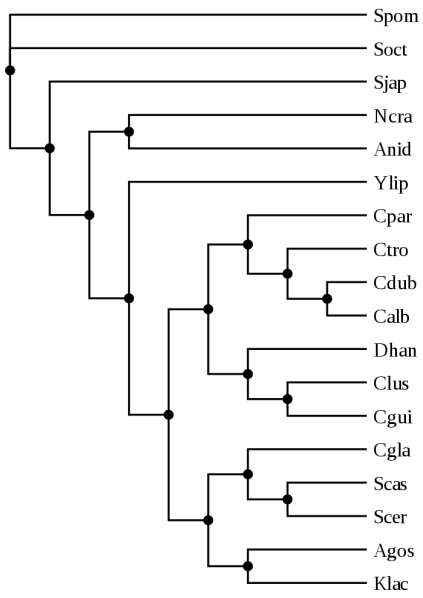

Figure 1: Tree from phyML

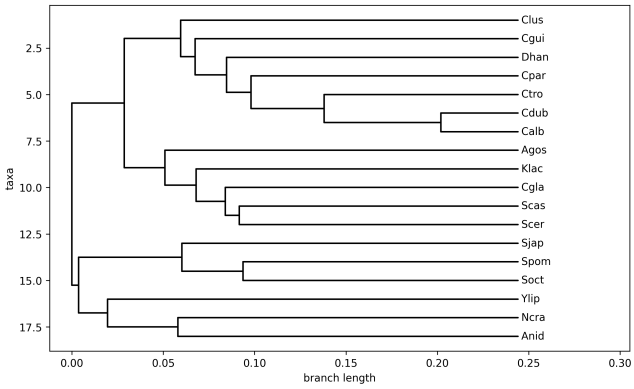

Figure 2: Tree from UPGMA