PHYLOGENY – TME4

2023-2024

Marina Abakarova
marina.abakarova@sorbonne-universite.fr

26 October 2023

General rules

- Reports must be sent by e-mail, using the subject "`[PHYG] TME4`", including in the body the names of the persons who worked on it (maximum two students per group). The deadline is 9th of November.
- Multiple files should be grouped in a compressed archive (`.tar.gz` or `.zip`).
- Your report *must be* in PDF format and named `student1_student2_TME4.pdf`. It should be simple, clear and well organized. Answers should be given in an exhaustive manner. Consider adding at the beginning a summary indicating the page of each answer.
- Source code must be well explained, commented and, most importantly, it should work without errors. Provide all needed information (*e.g.,* compiler/interpreter version) in a `README` file.
- All required materials can be found in the repository `https://github.com/abakarovaMarina/PHYG2023`
- A discord server is created so we can exchange our questions, answers and comments `https://discord.gg/w7JUvJV4`.

Use `http://itol.embl.de/upload.cgi` to visualize and export trees using the circular display mode. Labels have to be colored according to the clade a sequence belongs to (see file `clades.list`).
You can install Clustal Omega from the archive `clustalo-1.2.4.tar.gz` (check `http://www.clustal.org/omega/`) and use it with the option `--outfmt=phy` in order to get sequence alignments in phylip's format (`./clustalo --help` for more details).

Exercise 1

1. What are the problems related to the construction of phylogenetic trees with thousands species? What are the strategies for reconstructing phylogenetic trees on a large scale? How is it possible to handle data fragmentation?

2. Describe the *Maximum Agreement Subtree* and the *Maximum Compatible Tree* algorithms.

Exercise 2 – Phylogenetic tree from a single domain family

1. Get the *Ribosomal_S27* (`PF01599`) proteins fasta file from the archive `TME4_sequences.tar.gz` and select only those belonging to the species listed in the file `species.list`. If there are several sequences for a specific species, just take the first one (or any of your choice). Then, align the sequences and build two phylogenetic trees using the commands `neighbor` and `proml` of the phylip package.

2. Compare the trees you obtained and include them in your report. Were the clades grouped together? In order to facilitate the comparison, put a different color for each clade (see file `clades.list`).

Exercise 3 – Phylogenetic tree from multiple domain families

1. Consider all Pfam families in the archive `TME4_sequences.tar.gz` and, again, for each family, select only those sequences who belong to the species listed in the file `species.list` (as done in Exercise 2). Moreover, if a family *does not* contain *all* such species *do not* consider it.

2. Align the sequences of each selected family, concatenate the alignments (write a script to perform this task) and build a tree using the commands `neighbor` and `proml` of the phylip package. Compare the trees you obtained and include them in your report. Were the clades grouped together?

Exercise 4 – Super Trees

1. Consider the file `TME4_clades.tar.gz`, where the sequences of each family are partitioned according to their clades. For each pair of clades, align sequences of each family, concatenate the alignments, and build a phylogenetic tree. Write a script to perform these tasks and automatize the reconstruction of trees.

   For simplicity, please consider the `clades.list` file. In this file, we have just 7 clades. Remove the species that do not exist in the `species.list` file. We will have 40 species left. Now from `TME4_clades.tar.gz` we only consider the families containing all the clades and all the species of each clade. For each pair of clades we align, concatenate, and reconstruct a tree. It's like repeating Exercise 3 several times, each time we consider sequences from a pair of clades.

2. Use the program `treePack` to combine the generated trees. The input must be a single file where every line corresponds to a different tree created in the previous step (in newick format). See `treePack.readme` to generate a super tree. Include it in your report. Were the clades grouped together?

Exercise 5 – Tree comparison

Compare and discuss the best trees generated in Exercise 2, 3 and 4. What are your conclusions?