# PHYLOGENY – TME4

**2025-2026**

VINH-SON PHO
VINH-SON.PHO@SORBONNE-UNIVERSITE.FR

11 December 2025

**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 8th of January.
- 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/20sxn/PHYG2025`.

## 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?

## 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 – Bootstrap

The goal of this exercise is to compute the bootstrap values for each node of the `proml` phylogenetic tree.

1. Build up to 10 bootstrap MSAs by sampling columns from the original MSA (with replacement) until the bootstrap MSAs have the same amount of columns as the original MSA.

2. For each bootstrap MSA, build a bootstrap tree using `proml`. (Can be done automatically using the provided code)

3. Compute boostrap values for each node of the original tree, annotate and save the tree in newick format.

## Exercise 4 – 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 protein 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?