PHYLOGENY – TME1

Academic Year 2017/2018

Riccardo Vicedomini
riccardo.vicedomini@upmc.fr

5 October 2017

General rules

- Reports must be sent by e-mail, using the subject "`[PHYG] TME1`", including in the body the names of the students who worked on it (two maximum). You have one week to send me your report (points will be deducted for each day of delay)
- Multiple files should be grouped in a compressed archive (`.tar.gz` or `.zip`)
- Your report *must be* in PDF format and named `student1_student2_TME1.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.

Preliminaries

First, install PHYLIP.

- Source code:
  `http://evolution.gs.washington.edu/phylip/download/phylip-3.696.tar.gz`
- Build instructions:
  `http://evolution.gs.washington.edu/phylip/install.html`
- Documentation:
  `http://evolution.genetics.washington.edu/phylip/tuimala3.pdf`

Configure your `.bashrc` file in order to have direct access to PHYLIP's executables: simply add the following line at the end of the file `${HOME}/.bashrc`

```
export PATH=${PATH}:PHYLIP_DIR/phylip-3.696/exe
```

where `PHYLIP_DIR` is the directory where you extracted the file `phylip-3.696.tar.gz`.

Finally, download and extract the archive `PHYG-TME1.tar.gz` from the repository `https://github.com/vice87/PHYG` in order to have the sequences required by the following exercises.

Exercise 1: UPGMA

Briefly explain how the UPGMA algorithm works. Then, take the distance matrix below and construct a tree by using the algorithm: describe each step and, at the end, draw the resulting tree.

	S_1	S_2	S_3	S_4	S_5	S_6
S_1	0.00	13.71	17.45	15.81	16.12	16.76
S_2	13.71	0.00	17.69	16.38	13.01	17.67
S_3	17.45	17.69	0.00	8.92	9.43	11.50
S_4	15.81	16.38	8.92	0.00	9.78	11.44
S_5	16.12	13.01	9.43	9.78	0.00	7.75
S_6	16.76	17.67	11.50	11.44	7.75	0.00

Exercise 2: PAH

Here we consider the phenylalanine-4-hydroxylase enzyme. Its role is to degrade phenylalanine and, in human, a mutation in its gene is responsible for the phenylketonuria disease.

1. Download the **human**, **rat** (*Rattus norvegicus*), **mouse** (*Mus musculus*), **bovine** (*Bos taurus*) and *Caenorhabditis elegans* sequences of phenylalanine-4-hydroxylase from UniProt (search for "phenylalanine-4-hydroxylase" in `www.uniprot.org`). Align them using `Clustal` (you can simply use the web service at `http://www.ebi.ac.uk/Tools/msa/clustalo/`). Save it in PHYLIP format and include the content of the file in your report.

2. Use the command `protdist` (PHYLIP) to compute a distance matrix for the sequences and include the matrix in your report.

3. Use the command `neighbor` (PHYLIP) to compute an UPGMA tree and a Neighbor Joining tree. Include both trees in your report. In order to easily view the tree you can go to the following URL: `http://itol.embl.de/upload.cgi` .

4. Considering the fact that *neighbor joining* returns an unrooted tree, are the two trees different? Justify your answer.

Exercise 3: CFTR

The *cystic fibrosis transmembrane conductance regulator* (CFTR) is a protein that regulates the movement of chloride and sodium ions through epithelial cell membranes. A mutation in this protein is the cause of cystic fibrosis.

1. Consider the CFTR alignment in the file `CFTR_in_mammals.fasta` which contains proteins of several mammals. Compute the distance matrix with `protdist` (as done in the previous exercise) and include it in your report.

2. Compute both UPGMA and NJ trees and *compare* them. Include both of the trees in your report.

3. What do you observe for rat and mouse with UPGMA? And with NJ? Look at these species in the distance matrix, what do you notice?

4. Now look at the pig (*Sus scrofa*) in your trees. Consider its position relative to the bovine (*Bos taurus*), mouse (*Mus musculus*) and horse (*Equus caballus*). The correct tree of placental mammals is provided in Figure 1. Is the pig closer to bovine or horse? What was found by UPGMA and NJ?
(To help, you can write a simplified unrooted tree with these four species.)

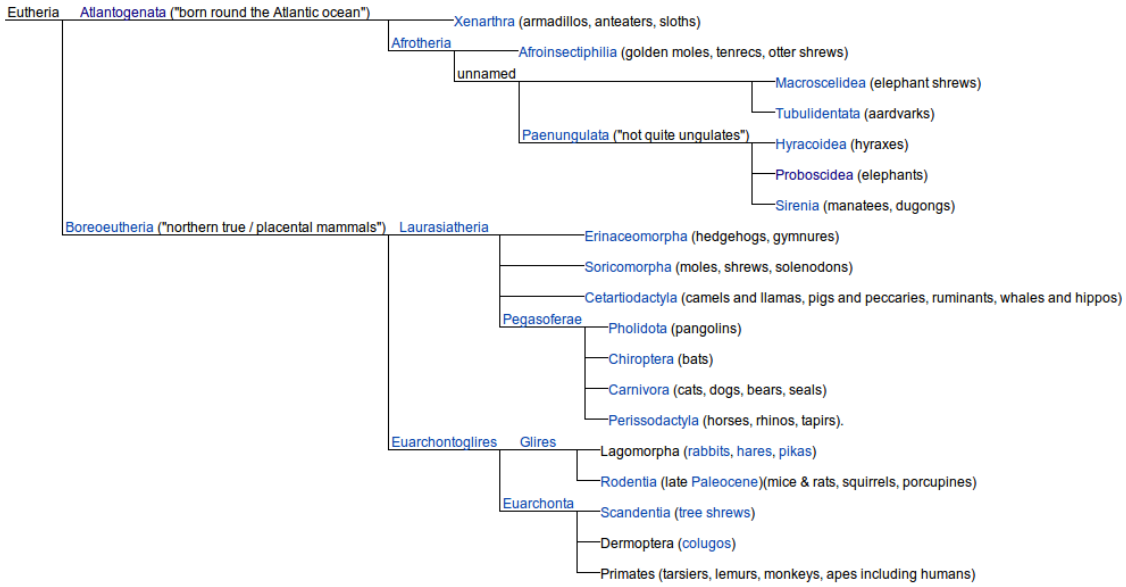

Figure 1: Correct phylogenetic tree of placental mammals

Exercise 4: P53

Tumor suppressor P53 is a protein involved in cell apoptosis and growth regulation. This protein is a protection against cancer as it prevents the cell from forming a tumor. If its gene gets mutated in somatic cells, however, this protein is no longer functional and the cell can generate a tumor. Therefore, when we sequence tumor cells, we often see many mutations in this gene.

1. Consider the P53 alignment (file `p53.fasta`) in order compute the distance matrix and the NJ tree (as in the previous exercises).

2. Now compare the position of the following four species: *Homo sapiens*, *Felis catus* (cat), *Loxodonta africana* (elephant) and *Monodelphis domestica* (opossum), in the CFTR and P53 NJ tree. Which species group together in which tree?

3. Look at the correct tree again (Figure 1). Opossum is a marsupial and therefore is outside this tree. Is the CFTR NJ tree correct? Is the P53 tree correct?