PHYLOGENY – TME1

Academic Year 2021/2022

Yasser Mohseni
yasser.mohseni_behbahani@sorbonne-universite.fr

7 October 2021

General rules

- Reports must be sent by e-mail, *mandatorily* using the subject "`[PHYG] TME1`", including in the body the names of the persons who worked on it (maximum two students per group).
- 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.
- All required materials can be found in the repository `https://github.com/yassermb/PHYG2021.git`
- A discord server is created so we can exchange our questions, answers and comments `https://discord.gg/sPYg3kaegB`.

Preliminaries

First, install PHYLIP.

- Source code:
  Linux: `http://evolution.gs.washington.edu/phylip/download/phylip-3.697.tar.gz`
  Windows: `http://evolution.gs.washington.edu/phylip/download/phylip-3.698.zip`
- 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-3.697/exe
```

where `PHYLIP_DIR` is the directory where you extracted the file `phylip-3.697.tar.gz`.

Finally, you can clone the repository `https://github.com/yassermb/PHYG2021.git` in order to have the sequences and materials 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.

S1	0.00	13.71	17.45	15.81	16.12	16.76
S2	13.71	0.00	17.69	16.38	13.01	17.67
S3	17.45	17.69	0.00	8.92	9.43	11.50
S4	15.81	16.38	8.92	0.00	9.78	11.44
S5	16.12	13.01	9.43	9.78	0.00	7.75
S6	16.76	17.67	11.50	11.44	7.75	0.00

Exercise 2: Neighbor-Joining (NJ)

Take the distance matrix below and construct/draw trees by using both NJ and UPGMA algorithms. Which one of these trees is more reliable? why? Is this matrix ultrametric? Is it additive?

S1	0	6	7	5
S2	6	0	11	9
S3	7	11	0	6
S4	5	9	6	0

Exercise 3: 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.

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. For better visualization, consider using `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 4: 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 (*mucoviscidose* in French).

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. Display and compare the tree. Include both 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 UGPMA and NJ?
(To help, you can write a simplified unrooted tree with these four species.)

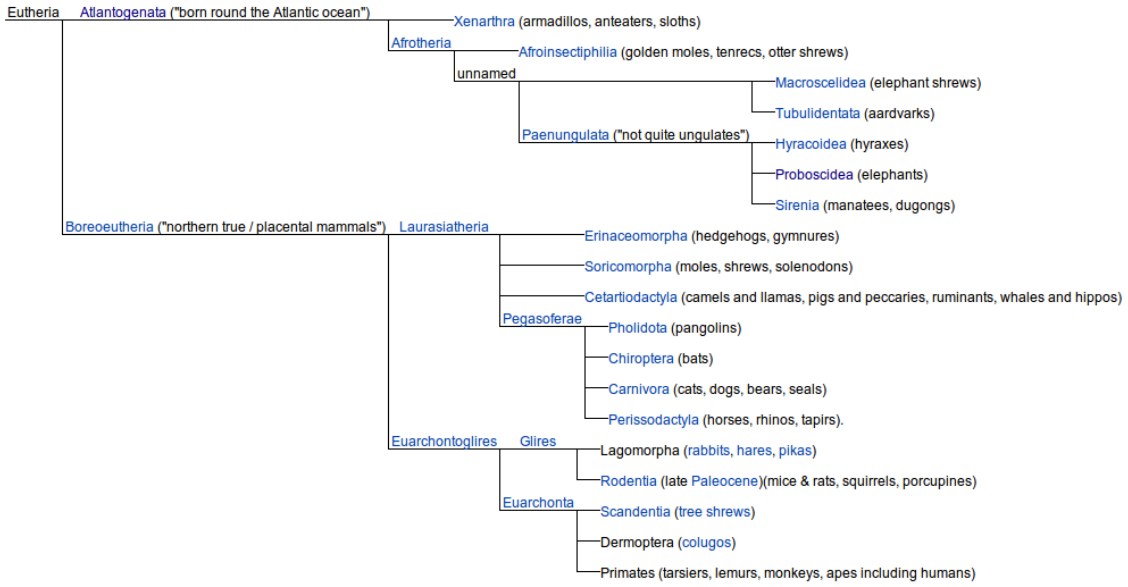

Figure 1: Correct phylogenetic tree of placental mammals

Exercise 5: 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 to 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?

Exercise 6: Coronaviridae

Coronaviridae is a family of enveloped viruses that rely on membrane fusion with the host cell as a primary step to enter the cell and begin to use its molecular machinery to translate the viral RNA. They are called Coronaviridae because the fusion protein is spike protein that decorates the barbed virion surface as crowns (corona). This family is composed of four lineages: α, β, γ, and δ. (Of course these lineages are not associated with the variants of the SARS-CoV-2 that you hear in the news). Human-infecting viruses from this family (including one common-cold) belong to $\alpha-$, $\beta-$Coronaviridae. The Severe Acute Respiratory Syndrome (SARS) viruses (such as SARS-Cov and SARS-CoV-2) and Middle East Respiratory Syndrome (MERS) viruses belong to the $\beta-$Coronaviridae. We will review the articles in the reference and answer to the following questions to better understand the structure and the mechanism of SARS-CoV-2 and how it functions.

- What are the four structural proteins composing the SARS-CoV-2 virion?
- What are the two major parts of viral spike (S) protein?
- Briefly explain the membrane fusion process and what is the role of spike protein?
- What are the two biggest domains of the S1 fragment of spike protein? Please explain their roles.
- Consider the alignment for S protein (file `S_protein.fasta`). Reconstruct and visualize the tree for the spike protein.
- From which animal the infectious virus SARS-CoV-1 (Human-SARS) is transmitted to humans? Highlight the related branches.
- From which animal the infectious virus SARS-CoV-2 is transmitted to humans? Highlight the related branches.

References

[1] Hoffmann, Markus, et al. "SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor." cell 181.2 (2020): 271-280.

[2] Hoffmann, Markus, Hannah Kleine-Weber, and Stefan Pöhlmann. "A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells." Molecular cell 78.4 (2020): 779-784.

[3] Casalino, Lorenzo, et al. "Beyond shielding: the roles of glycans in the SARS-CoV-2 spike protein." ACS Central Science 6.10 (2020): 1722-1734.

[4] Cai, Yongfei, et al. "Distinct conformational states of SARS-CoV-2 spike protein." Science 369.6511 (2020): 1586-1592.

[5] https://www.nature.com/articles/d41586-021-02039-y