PHYLOGENY – TME2

Academic Year 2022/2023

Marina Abakarova
marina.abakarova@etu.sorbonne-universite.fr

6 October 2022

General rules

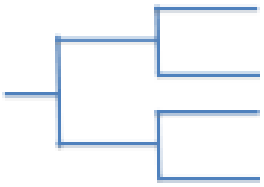
- Reports must be sent by e-mail, using the subject "`[PHYG] TME2`", including in the body the names of the persons who worked on it (maximum two students per group). The deadline is 13th of October.
- Multiple files should be grouped in a compressed archive (`.tar.gz` or `.zip`).
- Your report *must be* in PDF format and named `student1_student2_TME2.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/PHYG2022.git`
- A discord server is created so we can exchange our questions, answers and comments `https://discord.gg/vDmUHtDW`.

Exercise 1: Parsimony

1. What is the main idea of parsimony methods? Give an example.
2. What are the small and large parsimony problems? Which one is harder? Why?
3. What is the number of possible unrooted trees for n species? Justify your answer.

4. Given the following sequences, topology and cost matrices, apply the Fitch and Sankoff's algorithms to calculate the scores.

$A =$ `ATCCTG` $B =$ `ATCCGG` $C =$ `ACGGCC` $D =$ `AGGGCA`



Fitch algorithm

	A	T	G	C
A	0	1	2	3
T	1	0	2	4
G	2	2	0	1
C	3	4	1	0

Sankoff algorithm

	A	T	G	C
A	0	3	4	9
T	3	0	2	4
G	4	2	0	4
C	9	4	4	0

5. What is the main idea of the nearest neighbor interchange algorithm? Why is it considered a heuristic method?

Exercise 2: Reconstruction using reversal distances

1. Go to the web page `http://cinteny.cchmc.org`, choose *human* and *mouse* and click start. Then, select *whole genome analysis* (using human genome as reference). For human, genes are colored by chromosome, while for mouse by chromosome of human's homologous genes. Include both figures in your report.

2. Start again with human and mouse but select *chromosome versus chromosome* for chromosome 1 in human and 4 in mouse. What is the reversal distance? Why was a big part of each chromosome left in white? Include the figure in your report.

3. Now start again with human, mouse, cow, and chimpanzee. Choose a whole genome analysis, write the matrix of reversal distances. Include this matrix in your report.

4. Use PHYLIP command `neighbor` to compute NJ and UPGMA tree from this matrix (use advanced options to select UPGMA). Are these trees correct? Figure1

5. Now, we want to do the same with all mammals. The distance matrix is already available as file `mammalsMatrix.txt`. Compute both NJ and UPGMA trees:

   - use `drawtree` command to draw an *unrooted* tree from the output of NJ;
   - use `drawgram` command to draw a *rooted* tree from the output of UPGMA.

   Alternatively, you can use the page `http://itol.embl.de/upload.cgi` to generate the aforementioned trees. Include both of them in your report.

6. Are these trees correct? What is the limitation of the used approach?

	Has placenta	Lives in water	Lays eggs	Single pair of incisors	Opposable thumb	Enlarged malleolus
Zebrafish	No	Yes	Yes	No	No	No
Opossum	No	No	No	No	Yes	No
Whale	Yes	Yes	No	No	No	Yes
Mouse	Yes	No	No	Yes	No	Yes
Rat	Yes	No	No	Yes	No	Yes
Chimp	Yes	No	No	No	Yes	Yes
Human	Yes	No	No	No	Yes	Yes

Exercise 3: Reconstruction using characters

1. What are the differences between convergent and divergent evolution?
2. Use `pars` command (PHYLIP) with the matrix above in order to compute a tree based on characters. Attach the tree to your report. Is it correct?
3. Run the analysis again without considering the last column (*i.e.*, enlarged malleolus). What happens? Which character is responsible for this incorrect tree?
4. The presence of a character in two species can be explained either by a common ancestor or by convergent evolution. Find all cases of convergent evolution in the table.

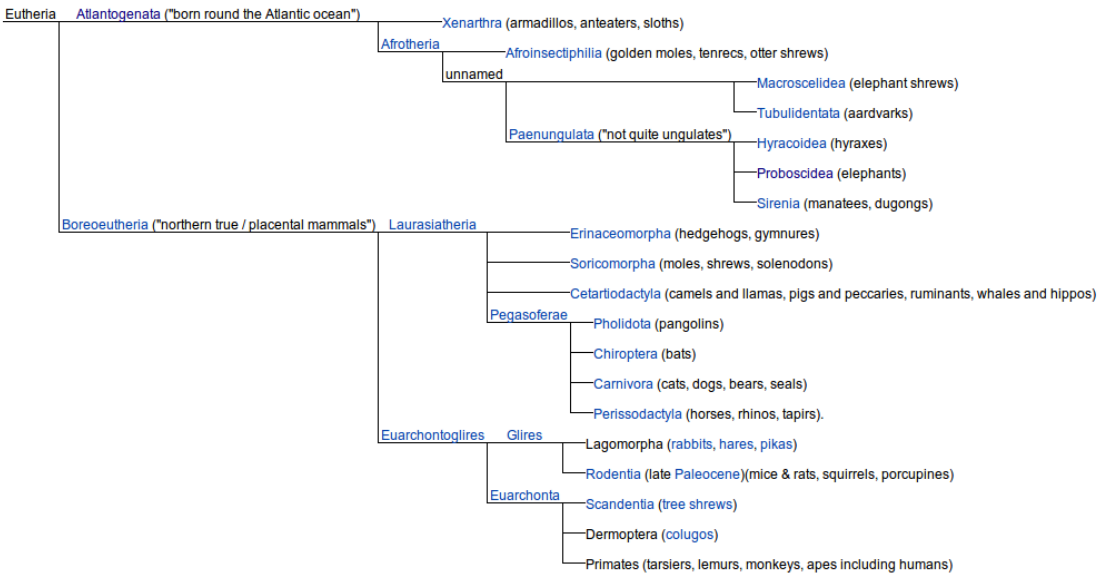


Figure 1: The correct phylogenetic tree of placental mammals (opossum does not appear because it is not a placental mammal).

Exercise 4: Coronaviridae

Following the exercise 5 of TME1, please review the articles in the references and answer the following questions:

- What is the role of Nucleocapsid (N) protein?
- What is the role of Membrane (M) protein?
- What is the role of Envelope (E) protein?
- What is the role of Envelope (E) protein?
- Consider the multiple sequence alignments of N, M, and E proteins (files `N_protein.fasta`, `M_protein.fasta`, and `E_protein.fasta`). Reconstruct and visualize the trees for these proteins. Include the trees in your report.
- Is there a consensus between these trees?

References

[1] Cubuk, Jasmine, et al. "The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA." Nature communications 12.1 (2021): 1-17.

[2] Thomas, Sunil. "The structure of the membrane protein of sars-cov-2 resembles the sugar transporter semisweet." Pathogens and Immunity 5.1 (2020): 342.

[3] Chai, Jin, et al. "Structural basis for SARS-CoV-2 envelope protein recognition of human cell junction protein PALS1." Nature Communications 12.1 (2021): 1-6.