PHYG – TME6

2023-2024

Marina Abakarova
marina.abakarova@sorbonne-universite.fr
Content created by Alessandra Carbone

21 December 2023

General rules

- Reports must be sent by e-mail, using the subject "`[PHYG] TME6`", including in the body the names of the persons who worked on it (maximum two students per group). The deadline is 11th of January 2024.
- Multiple files should be grouped in a compressed archive (`.tar.gz` or `.zip`).
- Your report *must be* in PDF format and named `student1_student2_TME6.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`.

Exercise 1 – Codon Bias

1. Select three bacterial genomes from the NCBI database: one cyanobacteria, one proteobacteria and one methanopyrales.

2. Implement the Self-Consistent Codon Index (`SCCI`) calculation algorithm that is searching for a set of genes `S` from all genes `G` of the organism that optimises the following inequality $SCCI(G/S) < SCCI(S)$. SCCI values on genes in S are maximal.
   We recall you the equation seen in class

$$SCCI(g) = (\prod_{k=1}^{L} w_k)^{1/L} \tag{1}$$

   where:

   - g - a gene
   - L - number of codons in g looking for
   - w_k

$$w_k = \frac{frequency\,of\,the\,k^{th}\,codon\,of\,g\,in\,S}{frequency\,of\,the\,dominant\,synonymous\,codon\,in\,S} \tag{2}$$

3. Calculate the codon bias of the genes for the three selected bacteria.

4. Select the 1% of the most biased genes for each of the three bacteria and compare them. Which genes are shared? Which genes are different?

5. Confront the weights obtained in the last iteration and identify the preferred codons of the three bacteria. Do they differ?

You can choose more than three bacteria (in chlamydiales, proteobacteria, firmicutes, thermococcales, spirochaetales....) and extend your study to a larger set. In this case, look for bacteria that share the same environment and try to correlate the weights obtained to the environment. Can you draw any conclusions?