**PRESENTATION**

**BIOLOGICAL PROBLEM**

Population studies have used genomic techniques and confirmed that the main bacterial genera isolated in the oral cavity are: Streptococcus, Actinomyces, Veillonella, Fusobacterium, Porphyromonas, Prevotella, Treponema, Neisseria, Haemophilus, Eubacteria, Lactobacterium, Capnophycum, Pseudomonas and Propionibacterium. [1]

Among them we can find different pathogenic species, within the genus Neisseria we have two of great importance, Neisseria meningitidis and Neisseria gonorrhoeae.

In the MALDI-TOF Mass Spectrometry technique, which generates unique protein mass spectrum for each species based on their ribosomal proteins, there are problems for the differentiation of these two species from the rest of the non-pathogenic species of the Neisseria spp [2]. Due to the great similarity of their spectrum, and as a consequence to the great similarity of the species to each other.

For *Neisseria gonorrhoeae* and *Neisseria meningitidis* the most significant m/z peak, that we can see in the protein spectrum, is 5052, this peak is reproducible intra-laboratory and inter-laboratory for *N. gonorrhoeae* and *N. meningitidis* isolates, and correspond to ribosomal protein RL34.

The aim of this script is to be able to differentiate between the species of the genus Neisseria, starting from unknown sequences of the rplF- 50S ribosomal protein L6 [3] and rpmH- 50S ribosomal protein L34 [4] genes, obtaining the molecular weight of the proteins.

**PROGRAME STRUCTURE**

1. We start from a fasta file with different known DNA sequences of the rplF and rpmH genes, and we obtain the sequence of RNA, proteins and finally the molecular weight of said protein.
2. We create a function to obtain the molecular weight of the proteins.
3. We create a function to identify at the gender level.
4. We create a function to identify at the species level.
5. We make a dendrograme to see the relationships between the sequences.

The molecular weight of the 50S ribosomal protein L34 in the different species of the Neisseria genus is the same (5051.96 Da / ~ 5052 Da).

The molecular weight of the 50S ribosomal protein L6 is different in the species of the genus Neisseria.

**USED MODULES**

**Bio.SeqIO:** the standard Sequence Input/Output interface for BioPython . Bio.SeqIO provides a simple uniform interface to input and output assorted sequence file formats (including multiple sequence alignments), but will *only* deal with sequences as [SeqRecord](https://biopython.org/wiki/SeqRecord" \o "wikilink) objects.

**Pyplot from matplotlib:** matplotlib.pyplot is a state-based interface it provides a MATLAB-like way of plotting. Pyplot:interactive plots and simple cases of programmatic plot generation.

[**Scipy.cluster.hierarchy**](https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html#module-scipy.cluster.hierarchy)**:** The [hierarchy](https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html#module-scipy.cluster.hierarchy) module provides functions for hierarchical and agglomerative clustering. Its features include generating hierarchical clusters from distance matrices, calculating statistics on clusters, cutting linkages to generate flat clusters, and visualizing clusters with dendrograms.

**DIFFICULTIES**

When obtaining the molecular weight of the protein, in the function mol\_weigth(), using only the sum of the amino acids was not the expected weight.

* The correct weight was obtained by adding the weight of a water molecule to the sum.

When applying the mol\_weight() function to the L34 and L6 protein lists, it only returned the last value.

* We create an empty list inside the function called listrw to store the weights of all proteins in the fasta file.

To apply the neisseria\_genus() and neisseria\_species() functions we create an object where we first apply the mol\_weigth() function to the list of proteins.

To make the dendrograms we needed to create a list, so we directly used the lists of the molecular weights of the proteins (L6\_MW and L34\_MW).

We can apply this script to files with unknown sequences, DNA, RNA or proteins of both genes, and also search for other genes to differentiate more species based on this code.

1. *Serrano-Coll HA, Sánchez-Jiménez M, Cardona-Castro N. Conocimiento de la microbiota de la cavidad oral a través de la metagenómica. Rev. CES Odont 2015; 28(2): 112-118*
2. *Morel, F., Jacquier, H., Desroches, M., Fihman, V., Kumanski, S., Cambau, E., Decousser, J. y Berçot, B. 2018, "Use of Andromas and Bruker MALDI-TOF/MS MS in the identification of Neisseria", European Journal of Clinical Microbiology & Infectious Diseases, vol. 37, no. 12, pp. 2273-2277.*
3. *Bennett JS, Watkins ER, Jolley KA, Harrison OB, Maiden MCJ. Identifying Neisseria species by use of the 50S ribosomal protein L6 (rplF) gene. J Clin Microbiol. 2014;52(5):1375–81.*
4. *Malakhova, M.M., Maier, T., Kubanova, A.A., Govorun, V.M., Svistunova, T.S., Gazarian, A.O., Borovskaya, A.D., Kostrzewa, M., Ilina, E.N., Vereshchagin, V.A. y Kruglov, A.N. 2009, "Direct Bacterial Profiling by Matrix-Assisted Laser Desorption−Ionization Time-ofFlight Mass Spectrometry for Identification of Pathogenic Neisseria", Journal of Molecular Diagnostics, The, vol. 11, no. 1, pp. 75-86*

This is the link for the recorded presentation: <https://www.canva.com/design/DAEXPH8xqo0/bjKctb_ASkxXJdg2rJp3_g/view?utm_content=DAEXPH8xqo0&utm_campaign=designshare&utm_medium=link&utm_source=recording_view>

With this task I have really learned to handle python, despite going quite a loss at the beginning. The qualification that I consider that I deserve is an 8. I don´t think that the problem is of a very high difficulty. I would have liked to do the step from DNA to RNA and proteins in two different functions. But with the FASTA format type and the use of Seqrecord I haven´t known how to do it, so it would be easier to apply the script to other files. The generation of the different FASTA files throughout the script is very slow, so the code may not be very efficient.

But I think it is an interesting way to combine data from mass spectrometry (MALDI-TOF MS), the molecular weights and spectra of ribosomal proteins, with the sequences that we can obtain from microorganisms from the microbiota of the oral cavity, and their identification. Also, I fulfil all the points marked (the use of modules, the use of two or more functions ...) And finally, my presentation is original and clear, meeting the time limit and slides.