**BC205: Algorithms for Bioinformatics.**

**Exercise 1. Introduction and Sequence Analysis (Deadline: April 7th, 2023)**

**Identifying non-mers in a bacterial genome**

**Non-mers are k-mers that don't have a single instance in a greater corpus (e.g. a genome), that is they do not exist in a genome. Search the genome of E. coli for any given 10-mer and report the 10-mers that do not exist in the genome:**

The code is provided in the file named Ex1\_BC205.ipynb and it is a Jupiter notebook file.

Firstly, the libraries and packages that will be needed are imported. The package itertools is used, to create all the permutations with repetition that can be created by the 4 bases A,T,C,G. They will be saved in a list called **all\_kmers**. There is the variable k that gives the opportunity to change the length of the permutation. In this problem the function will be called with the variable k = 10.

Continuing, the file containing the E.coli genome will be manipulated. The whole sequence without the header of the file will be saved in a string named ecoli. Then, a list named **kmers** will be created that will contain all the kmers that will be found in the ecoli genome string. Specifically, it will read the whole genome with iteration step 1 by 1 and will save all the substrings till the last possible k-mer of the genome (len(genome)-k+1 = the index that will stop the iteration).

Next, using the python function ‘set’ all the elements of the previous lists will be taken exactly once. Then, using the function non\_kmers = list(set(all\_kmers) - set(kmers)) only the kmers that exists in all\_kmers list (hence in all k-mers that can be created) but not in the kmers found in the genome will be saved in a new list named non\_kmers.

There are some kmers that contain ‘Ns’, that will not be considered. Those kmers are not able to be found in the permutations that were computed at the first steps. That is the reason that if we add the number of unique kmers found in the genome and the kmers that don’t exist in the genome we will find a larger number than all the possible k-mers. But if we remove those substrings (aka kmers that contain Ns) we will have the same expected number and our function we return the Boolean True.

Before the return statement, the function creates a file that is called k(the length of the kmer) \_non\_mers\_in\_Ecoli.txt that will contain all the kmers that do not exist in the E.coli genome. In our case that we called the function with k=10, a file named 10\_non\_mers\_in\_Ecoli.txt will be created and it is attached to the email that was send.