1st Report (Exercise 1.2)

Vouzina Olympia Dialekti

Identifying non-mers in 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.

To find the 10-mers that do not exist in the *E. coli* genome, the first step was to create a list with all the possible 10-mers, which are combinations of the four nucleotides (A, C, G, T), 4^10 in number (permutations with replacement were used). These 10-mers were saved in a list named “kmers”.

```
from itertools import product

nucleotides = ['A', 'C', 'G', 'T']
k = 10
kmers = [''.join(x) for x in product(nucleotides, repeat=k)]
```

The next step was to turn the fasta file into a string (named “string”) containing the whole genome of *E. coli*. Before this, a variable named “count” was created to count the lines of the file, to stop the following “for” loop.

```
fn = "Ecoli_genome.txt"
count=0
with open(fn) as f:
    for l in f:
        count+=1
ln=[]
with open (fn) as file:
    for y in range(count):
        line = next(file)
        lines = line.strip().split()
        ln.append((lines))

ln=list(chain.from_iterable(ln))
string=(''.join(ln))
```

A list (”kmers_genome”) containing all the 10-mers that exist in the genome was created to be compared with the list of all the possible 10-mers. This was done by scanning the whole genome and saving all the 10-mers.

```
kmers_genome = [string[i:i+10] for i in range(len(string)-9)]
```

The last step was to find the elements that were not common in the two lists (“kmers” and “kmers_genome”), and save them in a list (”missing_kmers”) and finally in a file named “missing_kmers.txt”.

```
missing_kmers = list(set(kmers).difference(kmers_genome))

with open("missing_kmers.txt","w") as file_missing:
    for i in missing_kmers:
        file_missing.write(i+'\n')
```

An example of 10 random 10-mers that do not exist in the genome is shown bellow.

```
import random
random_kmers_missing=random.sample(missing_kmers, k=10)

for i in random_kmers_missing:
    print(i)
```

Output:

GACACTACGT
AGATTCTCGG
CTATCCAAGC
AGGACGTATC
GAAGCGCTAG
ACTCCTGAGG
GAGGCAGTGG
CGCGTTAGGG

GGCCTAGCAG
GTCCTAATGA

Conclusion

The number of all the possible 10-mers is 4^10=1048576, the total number of 10-mers (non-unique) is 3101344, and the number of the 10-mers that do not exist in the genome is 223321. Some of the 10-mers exist multiple times in the genome and some others not even once, this happens because the existence of a 10-mer is not entirely random, as they can be a part of a regulatory region that is common in more than one gene. On the other hand, it is expected that some 10-mers may not exist in a genome, because they may not have a functional role, and so they might have been eliminated by natural selection. So, the results were expected, and nothing suspicious was observed, as *E. coli* has a relatively small genome.