

Workshop#1

Bioinformatics\_exercise

Alejandro Gómez Moyano 20222020073

Facultad de ingeniería, Universidad Distrital Francisco José de Caldas

Group 020-84, System\_Analysis

Carlos Andrés Sierra Virguez

13 de marzo de 2024

### Technical concern

The first concern that i have when i started with the workshop is the diferrents way of code in Python , it can be a Little bit confused understand some lines of code in the process of créate database , get the motifs and compare .and the second is learn and grasp the context of the biological genomic.

### Decision

The best Decision was learn about information theory of shanon in videos or documents , with this knowledge its posible transform or use the math formula of entropy to calculate and filter the occurences to make a best Analysis of the genomic .

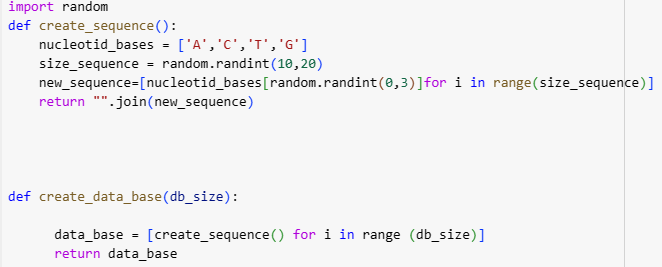
Difficulty

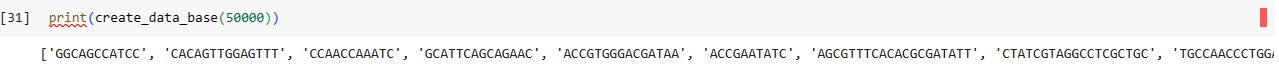
At first the principal dificulty was understand the Shanon´s formula for can apply in the context of the genomic combinations , also another dificulty was i downt have a knowledge in python but was easy to learn and apply knowledge of another subject like basic programming

# Workshop

First step: make a function to create a sequence using the nucleotid bases (A ,C ,T,G) , create a sequence with diferents size with a random between 10 and 20 .

After that create the database using the previous function to add all the elements to a list.



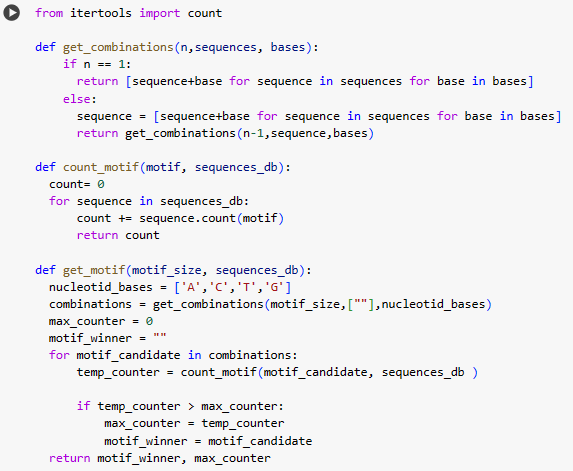


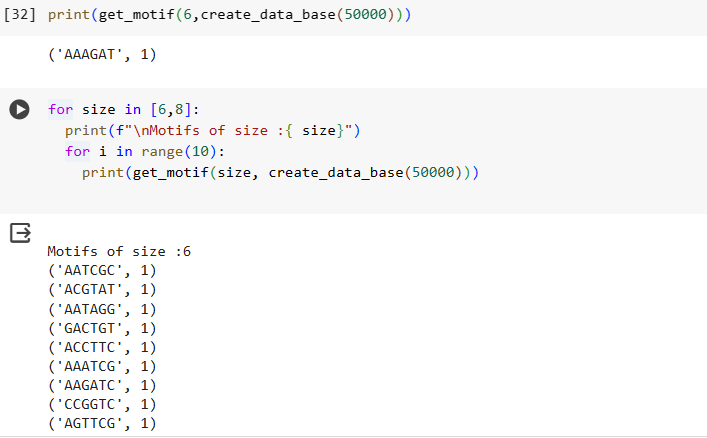
For the Analysis we use a databae of 50000

Second step: create a function call :” get\_combinations “ , and this is a recursive function because the return call to the same function, ,use the number of size like the count of times that do the iteration .

In count motif use a variable count the time that the motif repeat in the sequence

In get motif foound the motif that repeat mosto f times in the sequence





Output of  print(get\_motif(size, create\_data\_base(50000)))

Motifs of size :6

('AATCGC', 1)

('ACGTAT', 1)

('AATAGG', 1)

('GACTGT', 1)

('ACCTTC', 1)

('AAATCG', 1)

('AAGATC', 1)

('CCGGTC', 1)

('AGTTCG', 1)

('ACACGA', 1)

Motifs of size :8

('ACGGATTC', 1)

('AACAACAG', 1)

('AAACTACA', 1)

('GAATAACT', 1)

('AGAGCCGT', 1)

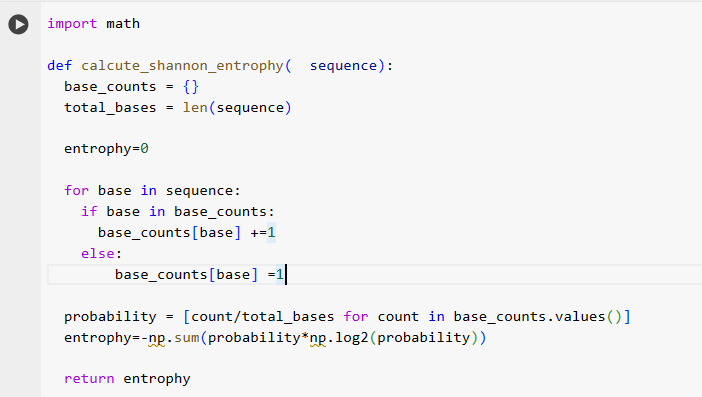
('ATCCCCTC', 1)

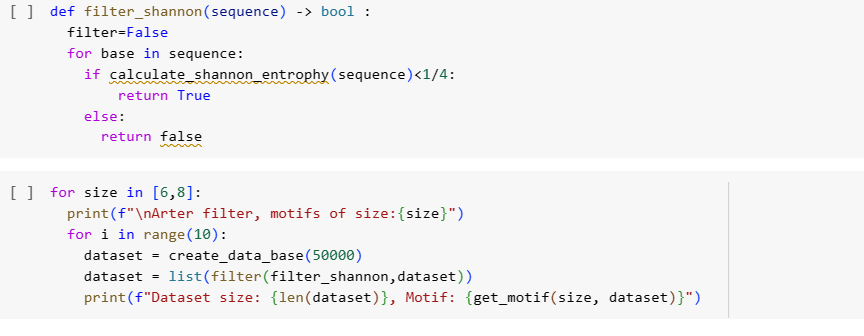
('CGTGGGTT', 1)

('ACGATCAC', 1)

('GGGGCAGA', 1)

('AAAACACA', 1)



to calculate the shanon´s theory we have to know that the way to know the base of genomic is ¼ because there are 3 questions to know , if the base is not g can be :t or c or t , and thats the parametter , we rule out the sequence that have an entrophy more of that.