Fall 2020
Computer Science I
Section 5 – Tuesday and Thursdays – 12:00PM – Fulton 250

Maíra Marques Samary PhD

TA's Office Hours – online

James Monahan - monahajm@bc.edu - https://bccte.zoom.us/j/2822792652
Tuesdays 7:00 PM – 8:00 PM
Wednesdays 4:00 PM – 5:00 PM

Jennifer Joseph - - josephjz@bc.edu - https://bccte.zoom.us/j/5882755193
Wednesdays 11:00 AM – 12:00 PM
Thursdays 3:00 PM - 4:00 PM

Liam Murphy- - murpaue@bc.edu - https://bccte.zoom.us/j/3085424208
Tuesdays 2PM-4PM

Discussion Groups:

CSCI100701 - Tuesday 6:00 PM – 6:50 PM- Fulton Hall 220 (James Monahan)
CSCI100702 - Thursday 5:00 PM – 5:50 PM – Fulton Hall 220 (Jennifer Joseph)
CSCI100703 - Wednesday 4:00 PM – 4:50 PM – Gasson Hall 203 (Liam Murphy)

Created by Hang Yin

Due date – 10/23/20 11:59 PM

General Instructions

Create a folder named ***LASTNAME_FIRSTNAME***. You will populate the folder with **ALL** of the .py files you write for this homework. To submit the homework, verify the folder includes all your .py files, compress (zip) the folder then upload to Canvas. Remember to include the following comments at the **top of each** of your .py files:

```
# author:
# assignment:
# description:
```

What to submit in Canvas?

Make sure all your files are saved in the folder LASTNAME_FIRSTNAME, then compress (zip) the folder and upload to Canvas.

If you encounter any problems in completing the assignment or in the submission process, please don't hesitate to ask for help. The sooner, the better!

Problem

Our DNA (deoxyribonucleic acid) is the molecule that contains the genetic code of organisms. DNA is composed of four nucleotides, A, T, C and G.

We (human-beings) have numerous DNA fragments in our DNA library. In molecular biology, a library is a collection of DNA fragments that is stored and propagated in a population of micro-organisms through the process of molecular cloning.

We want to create a software that automatically distinguish each DNA molecule through attaching a unique DNA barcode at the end of each fragment. DNA barcoding is the method utilized for species identification using a short section of DNA from a specific gene or genes.

However, according to research done in the area, a DNA barcode in needs to be unique (from each other), and must satisfy the following four criteria:

1. Each DNA barcode needs to have a **GC-content** from 40% to 60% (GC-content definition is below)

2. Any barcode that contains more than three consecutive redundancy (repeated) must be voided. In other words, there should not be three identical nucleotides in a row, such as ’AAA’ or ’GGG’

3. Barcodes must exclude restriction sites for AgeI, AscI, BamHI and Sbf1 listed below:

   ’AgeI’ = ’ACCGGT’

   ’AscI’ = ’GGCGCGCC’

   ’BamHI’ = ’GGATCC’

   ’SbfI’ = ’CCTGCAGG’

4. **(BONUS – 1 POINT IN ANY HW OR MIDTERM)** Each DNA barcode needs to have a **hamming distance** greater or equals to 3 from all bar- codes (hamming distance definition below)

Your goal in this homework is to generate a software that will ask the user how many sequences of DNA barcodes (n) she/he wants and what is the length (size) of the DNA barcode he wants. And according to the user parameters you must generate a text file containing a total of **n** sequences of **size-nucleotides** long DNA barcodes. Attached to the instructions is a file that was generated according to the rules presented here with the parameters 50 and 8 (50 sequences of 8-nucleotides long DNA barcodes) and 20 and 5 (20 sequences of 5-nucleotides long DNA barcodes). The output examples consider the bonus question!

If you follow all the rules here, you may encounter a different result then the ones given (probability problem).

Output Format

Your file must generate the file in the **EXACT** format showed below (as in the examples):

barcode (sequence number): DNA barcode

barcode 1: GGTCAATG

For a full grade of the assignment your file output will go through automatic grading.

So, you must follow the specified format.

YOU ARE FREE TO CREATE HOW MANY FILES AND FUNCTIONS YOU WANT

BUT ALL YOUR CODE MUST BE INSIDE FUNCTIONS

PUT COMMENTS EXPLAINING WHAT YOU ARE DOING IN EACH FUNCTION

Definitions

GC-content

GC-content (or guanine-cytosine content) is the percentage of nitrogenous bases in a DNA or RNA molecule that are either guanine (G) or cytosine (C); The equation is given by:

$$\frac{G + C}{A + T + G + C}\ x\ 100\%$$

Restriction Enzyme

A restriction enzyme is an enzyme (endonuclease) that cleaves DNA into fragments at specific recognition sites known as restriction sites.

Restriction means the enzyme operates only at specific sites. Enzyme is a protein that carries out specific chemical reactions.

Restriction Site

A specific nucleotide sequence, typically ranging from four to eight nucleotides long, that a given restriction enzyme cuts. For the purpose of this homework, certain restriction site needs to be avoided in the DNA barcode sequences.

Hamming Distance

The Hamming distance between two strings of DNA of equal length is the number of positions at which the corresponding symbols are different. In other words, it measures the minimum number of substitutions required to change one string into the other, or the minimum number of errors that could have transformed one string into the other.

```
ATCGGCA
||.|...
ATGGCGT
```

Some ideas on how to divide all you need to do (guidelines)

1. According to the number of sequences (number) and the size of the sequence (size), create a function that generates a list of z with all the possible combinations of genes to have a list of the specified size but an amount as number*10 (assuming that a bunch on the sequences will be erased because they will fail one of the rules. The function must the return the list generated

2. Create a function that will receive as a parameter the list of the previous function, and validate the GC-content according to the rules stated. If a sequence breaks the GC-content rule you take out the sequence of the list. And you must return a smaller list where no sequence breaks the GC-content.

3. Create a function that receives the list from the previous function and that excludes from the list any sequence that has 3 identical nucleotides in a row. Again, you must return a smaller list where there is no sequence that has 3 repeated nucleotides connected.

4. Create a function that receives the list from the previous function and that excludes the listed restrictions.

5. (BONUS) Create a function that will receive as a parameter the list of the previous function and that checks that the hamming distance of each sequence is greater or equal to 3.

OBS: Your code must not assume a specific DNA size or a specific number of sequences. You must ask the user for that information.

RUBRIC

	Excellent (100% of points))	**Average (60% of points)**	**Needs Improving (40% of points)**	**Possible Points**
Function module	• Functions are implemented properly • Function headers are complete and accurate	• Functions are generally implemented properly, but exhibit minor errors. • Function headers are generally complete and accurate, but some minor details are missing.	• Functions are implemented improperly. • Function headers are sketchy or missing.	**3**
User interface	• UI formatting is appropriate. • Prompts are complete and concise. • Information is presented in a meaningful form.	• UI formatting exhibits minor flaws. • Prompts are not completely clear and concise. • Information presentation is slightly confusing.	• UI formatting is sketchy or haphazard. • Prompts are confusing or missing completely. • Information presentation is completely confusing.	**1**
Output (file with DNA barcode) Parameters smaller than 30 (sequences) and 5 (size)	• File formatting is appropriate. • Information is presented in a meaningful form. • Contains the information as requested	• File formatting is not the most appropriate. • It does not contain all information requested • Few of the DNA barcodes does not follow the criteria	• File formatting is not appropriate. • It does not contain all information requested • Many of the DNA barcodes does not follow the criteria	**3**
Output (file with DNA barcode) Parameters bigger than 30 (sequences) and 5 (size)	• File formatting is appropriate. • Information is presented in a meaningful form. • Contains the information as requested	• File formatting is not the most appropriate. • It does not contain all information requested • Few of the DNA barcodes does not follow the criteria	• File formatting is not appropriate. • It does not contain all information requested • Many of the DNA barcodes does not follow the criteria	**3**
FINAL SCORE				**10**