Specialist Diploma in Biomedical Informatics and Analytics
Coursework Final / C2769C

C2079C AY2020
Individual Coursework Submission Cover Page

Personal Details	
Name	Julia Loh Jie Min
Admin No.	20053577

Instruction to Students

1. This coursework assignment is to be completed and submitted by each candidate
2. Student must ensure that it is their own work and must be responsible for the safeguarding of the assignment
3. The maximum score achievable for this coursework assignment is **100 marks**.

Submission Procedures

4. You are to upload the assignment in LEO 2.0.
5. Save your answer for the program in **python** file (e.g. Coursework1.py or Coursework_Final.py)
6. For flowchart, save your diagram in **MS Word** document or **JPEG** file (e.g. Coursework_Final_Flowchart.docx or Coursework_Final_Flowchart.jpg)
7. Save your video recording for the program in **MP4** file (e.g. Coursework_Final.mp4)
8. Endorse this document (cover page), copy all answer files into an empty folder, the folder must be named with the following file naming convention:

   <Student ID>-<Name>-C2769C-AY2020CWF

   e.g. 2001111-John Khoo-C2769C-AY2020CWF
9. Compress your folder in **zip** file format
10. Candidate must submit his/her zip file to the Programme Co-ordinator no later than **2359h** on **21 January 2021** **in LEO 2.0**. Late submissions of assignment-based coursework component without leave of absence (LOA) for the module will be subjected to the following late penalty:

Time after submission deadline	Between 0 and < 24 hours	Between 24 and <48 hours	Between 48 and <72 hours	After 72 hours
Percentage of total marks deducted by	5%	10%	15%	100%

School of Infocomm

C2769C Introduction to Programming

Coursework Final

Extending the work from Coursework 1 (CW1), there is a need to apply the consensus DNA sequence to the gene sequences from organisms. The purpose of this is to find the DNA location with the highest affinity to the consensus sequence. Your task is to apply the consensus sequence of "AGGTG" (from CW1) to the gene sequences of a photosynthetic bacteria, "Synechococcus elongatus". You are to work with the gene sequences found in the given file "gbbct1_truncated.seq.txt"*.

(*The file is a truncated version of the original file "gbbct1.seq.gz" that can be obtained from genebank https://ftp.ncbi.nih.gov/genbank/.)

You are to parse the file "gbbct1_truncated.seq.txt"* and extract only the first gene sequence ("ctgcagccgc … tggctcgcca tc") of 2992 DNAs. Once the first sequence of 2992 DNAs are extracted from the file, you are to use the weighted table from CW1 (Table 1), to calculate the weights at all locations (from locations 1 through 2988), in blocks of 5 DNAs. For example, at location 1, calculate the sequence weight for "CTGCA" and at location 2, calculate the sequence weight for "TGCAG". Repeat for all locations until you reach the location 2988, where you calculate the sequence weight for "CCATC". Once the sequence weight for all locations are calculated, you are to list down the sequence weight and the corresponding location(s), for all sequence weights that fall within the threshold of 70%.

Columns	1	2	3	4	5	6
A	1.00	0.25	0.00	0.25	0.00	0.25
C	0.00	0.00	0.00	0.25	0.25	0.25
G	0.00	0.75	0.75	0.00	0.50	0.25
T	0.00	0.00	0.25	0.50	0.25	0.25

Table 1: Weighted Table

School of Infocomm

C2769C Introduction to Programming

A summary of the coursework requirements.

Part 1:

1. Parse the file “gbbct1_truncated.seq.txt” into the program and extract the first DNA sequence (all 2992 DNAs).
   (Make sure to remove all unnecessary characters/spaces and leave only the characters denoting DNAs – “A”, “T”, “C”, “G”.)
2. Calculate the sequence weight (in blocks of 5 DNA length) at each location of the extracted DNA sequence from (Part 1, step 1).
3. List down both the sequence weight and the corresponding location(s) for all sequence weights that fall within the threshold of 70%.

Part 2:

1. Modify your codes to extract all the three sequences and process them according to Part 1, steps 1-3.
2. Modify your codes to consolidate all overlapping regions. (Do this by selecting the sequence with the higher sequence weight between two or more overlapping sequences.)

You will be graded based on the following **THREE** submissions:

1. Flowchart of program design.
2. Codes written for the program.
3. A video recording of you explaining the codes that you have written (maximum of 10 minutes).

*Your video recording is to include your live capture of your own face for verification purpose. You are to appear at the bottom right hand corner of the screen where it is not obstructing your codes. The software that you can use if Flash Back express. (Installation file can be found in the coursework folder.)