C2079C AY2020
Individual Coursework Submission Cover Page

Personal Details	
Name	
Admin No.	

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. Coursework1_Flowchart.docx or Coursework_Final_Flowchart.jpg)
7. Save your video recording for the program in **MP4** file (e.g. Coursework1.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-AY2020CW1

   e.g. 2001111-John Khoo-C2769C-AY2020CW1

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 **5 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 1

In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Sequence alignment appears to be extremely useful in a number of bioinformatics applications. In this coursework, you are tasked to code your own sequence alignment program.

The program that you are tasked to write is a simplified version that is adapted from a *publication (Gerald Z. et al.) that focuses on a weighted approach. In this approach, a weighted table is constructed based on four DNA sequences (Table 1) with length of six DNAs each.

Columns	1	2	3	4	5	6
Seq1	A	A	T	T	G	A
Seq2	A	G	G	T	C	C
Seq3	A	G	G	A	T	G
Seq4	A	G	G	C	G	T

Table 1: DNA Sequences

The DNA sequences (Table 1) will be used to calculate the weighted table (Table 2). The weighted table is calculated by looking at the distribution of each DNA for each column. Each column of the four sequences have an average weight of 0.25. As Adenine (A) occurs in the first column of all four sequences, the weight of Adenine is 1.0. In the second column, Guanine is seen to occur three times across all four sequence and hence the weight of 0.75. The same is calculated for all other DNAs for all columns.

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 2: Weighted Table

From the weighted table (Table 2), we need to identify the consensus DNA sequence. The consensus DNA of each column is the DNA with the highest weight (in yellow). As a result, the consensus DNA sequence formed is “AGGTGX”. Column 6 does not have any consensus DNA and hence it is represented by “X”, as a neutral (hence the weight for column 6 is ignored for all calculations).

The corresponding weights of the consensus DNA sequence is summed up to produce the base score of 3.5 (summation of all highlighted cells from Table 2). The base score will be used to calculate the matching probability.

We will use the weighted table (Table 2) to calculate all DNA sequences including the original four DNA sequences. The weight of each DNA is extracted from the weighted table based on the column of each DNA in the sequence. An example of calculation for “Seq1” can be seen in Table 3 below.

Seq1	A	A	T	T	G	A	Total
	1	0.25	0.25	0.5	0.5	X	2.5

Table 3: Example of calculation for “Seq1”

The summation of all weights for the “Seq1” (Table 3) is performed to produce a sequence weight of 2.5. The sequence weight is then divided by the base score of 3.5 to produce a match probability of 71.4%. As “Seq1” produces the lowest match probability, you may use 70% as the threshold to determine if new input sequences produce a match or not.

School of Infocomm

C2769C Introduction to Programming

A summary of the coursework requirements.

Part 1:

1. Read the four DNA sequences into the program through user input (Table 1).
2. Calculate the weighted table (Table 2), for each column in the DNA sequence table (Table 2).
3. Identify the consensus DNA sequence and calculate and display the base score.
4. Calculate and display the weighted score for all four sequences (Table 1).
5. Request for a threshold from the user.
6. Request for a new sequence from the user.
7. Calculate and display the match probability of the new sequence, indicating if the new input sequence is a match or not.

*Refer to "Steps.xlsx" for more detailed breakdown of steps.

Part 2:

1. Modify your codes to allow the user to define the number of input sequences (Table 1), to allow more than four sequences to be entered.

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.)

Publication reference:
https://pgfe.umassmed.edu/TFDBS/Documentation/Freq2PWM.pdf