BIO310 Introduction to Bioinformatics
Homework 2
Spring 2019

March 3, 2020

Instructions:

- We expect you to start working on this assignment in the lab, at the end of the lab you will submit how far you came along. Your overall effort will be graded and it will eventually contribute to your lab grade. This grade will be assigned as a number from 0-10; no effort being 0 and full effort being 10. You are not expected to finish the entire assignment during the lab; you will have a chance to submit the final version till the due date and this will be your homework 2 grade, which is out of 100.
- For the homework submission, submit a PDF document for the answers of the write-up questions, the plots should be appropriately labeled, figures should have captions and should be appropriately cited within the main text. Name your submission as `BIO310-HW2-YourName.pdf` where you substitute in your first and last names into the filename in place of 'YourName' and submit online through SUCourse as a single file. Upload your final report on SuCourse by the due date.
- Upload the code online on SuCourse by the due date. The code you submit should be in a format that is ready to run. In submitting the code on SuCourse, compress it as a ZIP file with the name `BIO310-HW2code-YourName.zip` where you substitute in your first and last names into the file name in place of 'YourName' and X with the current homework number.
- An ipynb with code and report together is also acceptable.
- If you are considering to submit the homework late, please see the late submission policy in the syllabus.
- Please follow the submission instructions, not adhering the submission standards will lead to point deduction.

1 Crime Investigation [15 pts.]

You have been called to assist in a crime scene investigation: the body of a tourist was found at the airport. He seems to have suffered from convulsions and internal bleeding. Detectives at the crime scene found a drink carton with some sort of beverage: it still contained some fluid which looks like milk. This may be key evidence.
The fluid was sent to the lab and you receive a list of the components of the beverage. Some small molecules such as sugar were found, but also four unidentified proteins were detected. It is your job to analyze these proteins to see if you can help figuring out how the tourist died.
A list containing the amino acid sequences of the 4 proteins (called suspect1 through 4) is given in a separate document. You now have enough information to start your investigation. For each of the unidentified proteins, answer these four questions :

1. Which protein is it?
2. From which organism does it originate?
3. What is the function of this protein?
4. Is this protein guilty? Could it be responsible for the death of the tourist? Why (not)?

How did the victim die?

2 Local Alignment [60 pts.]

1. Implement the local sequence alignment algorithm with linear gap penalty. The user should be able to specify the input filename, and the mismatch, gap penalty and match scores. The two DNA sequences should be in two separate lines of the input file. The first sequence will form the rows of the scoring matrix, the second sequence will be the columns.
2. You should write the output into an output file. The file should include the alignment of the two sequences and the values for mismatch, gap penalty and match scores used to generate the alignment and the score achieved by the alignment.
3. Test your program with several test cases. Especially test edge cases carefully. For example, how would your algorithm run if two very short strings are input, for example 'A' vs 'T' alignment. We have provided additional test cases separately `test_cases`. Submit the output of these test cases.

3 Hidden Markov Models [25 pts.]

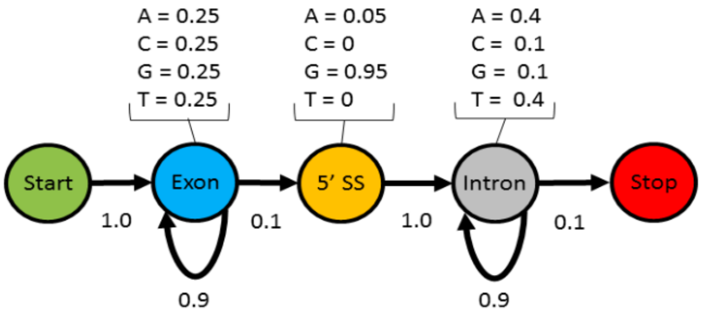

Figure 1: A toy HMM model for 5' splice site recognition.

For most eukaryotic genes and some prokaryotic ones, the precursor messenger RNA must be processed before it becomes a mature messenger RNA (mRNA). One of the steps in this processing, called RNA splicing, involves the removal of introns. The final mRNA consists of exons. You may find some references about RNA splicing in more detail here.
Consider a very simplified version of the recognition of 5' splice site. Assume we are given a DNA sequence that begins in an exon, contains one 5 splice site and ends in an intron. The problem is to identify where the switch from exon to intron occurred that is where the 5 splice site (5SS) is. The HMM model is shown in Figure 1.

1. Using the HMM model shown in Figure 1, calculate the probability of each of the following state paths: Show your work.

		A	G	T	G	A		Probability
Path 1	Start	E	E	E	5	I	End	
Path 2	Start	E	E	5	I	I	End	
Path 3	Start	E	5	I	I	I	End	

2. Specifically, which of the three state paths is most likely to annotate the sequence?
3. Note that each state path has the 5' splice site at a different position in the sequence. At which position the splice site is more likely to be in?