BIO310 Introduction to Bioinformatics
Lab 6 and Homework 3 Spring 2019

April 4, 2019

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 1-5; no effort being 1 and full effort being 5. 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 3 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-HW3-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-HW1code-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.
- 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.

Question 1 [20 pts.]

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.

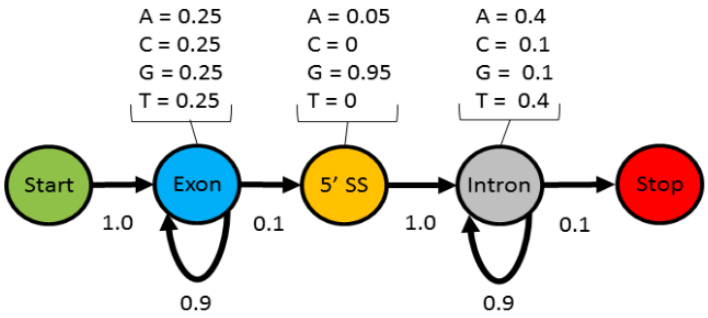

Figure 1: A toy HMM model for 5' splice site recognition.

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?

Question 2 [60 pts.]

You are provided with an incomplete code for the implementation of the Viterbi algorithm. There are 3 parts missing: (1) the recursion step, (2) the traceback step and (3) calling the Viterbi function. You will fill out the missing lines in the code, `viterbi_implementation_hm3.py`. All variables, transition and emission probabilities etc. are already defined.

Using the code you implemented identify the best state path for the following observation sequence: CTTCATGTGAAAGCAGACGTAAGTCA using the model with the same parameters as provided in Question 1.

Question 3 [20 pts.]

The following papers make use of Hidden Markov Models to solve different problems in biological sequences. Choose one of the papers and desribe in a few paragraphs describing what problem it aims to solve, and how HMMs are used (describe the states, observations, probabilities, etc.):

1. Chromatin-state discovery and genome annotation with ChromHMM `doi:10.1038/nprot.2017.124`
2. Profile hidden Markov models https://www.ncbi.nlm.nih.gov/pubmed/9918945
3. Protein fold recognition using HMM–HMM alignment and dynamic programming `doi:10.1016/j.jtbi.2015.12.018`
4. Hidden Markov models for detecting remote protein homologies `doi:10.1093/bioinformatics/14.10.846`
5. Microbial gene identification using interpolated Markov models `doi:10.1093/nar/26.2.544`