COMP 462 / 561 - Homework #3
Due on November 23rd 2016, in class

Question 1 (40 points).
Protein sequences are generally made of an alternation of hydrophobic regions (that usually end up buried in the core of the folded protein) and hydrophilic regions (that often are that the surface of the folded protein, exposed to water). Different amino acids have different hydrophobicity levels (e.g. Valine is very hydrophobic, while Arginine is very hydrophilic; see https://en.wikipedia.org/wiki/Amino_acid). In this question, you will implement and use the Viterbi algorithm on a small 3-state HMM that will attempt to segment a given protein sequence in three types of regions: Hydrophobic, Hydrophilic, and Mixed. Assume that the expected properties of each type of regions are the following:

- Hydrophobic regions are made of 60% of hydrophobic amino acids (A, V, I, L, M, F, Y, or W, each with equal probability), and 40% of the other amino acids (each with equal probability). They are on average 5 amino acids long. Hydrophobic regions are followed by Mixed regions 80% of the time, and hydrophilic regions 20% of the time.
- Hydrophilic regions are made of 80% of non-hydrophobic amino acids (R, H, K, D, E, S, T, N, Q, S, C, G, P, each with equal probability), and 20% of the hydrophobic amino acids (each with equal probability). They are on average 8 amino acids long. Hydrophilic regions are followed by Mixed regions 70% of the time, and hydrophobic regions 30% of the time.
- Mixed regions are made of any amino acid, each with uniform probability 1/20. They are on average 7 amino acids long. Mixed regions are followed by hydrophobic regions 50% of the time, and hydrophilic regions 50% of the time.
- The first amino acid in a protein belongs to any of the three types of regions with equal probability.

a) (5 points) Draw the HMM that best captures the situation described above.

b) (10 points) Implement the Viterbi algorithm from scratch, in the language of your choice but without the use of any external resources or libraries. Submit your program on MyCourses. Ensure that your program compiles and provides useful output, requiring only the FASTA file to provide a prediction.

c) (20 points) Use your algorithm to identify the best possible annotation of the protein sequences available at:
http://www.cs.mcgill.ca/~blanchem/561/hw3_proteins.fa

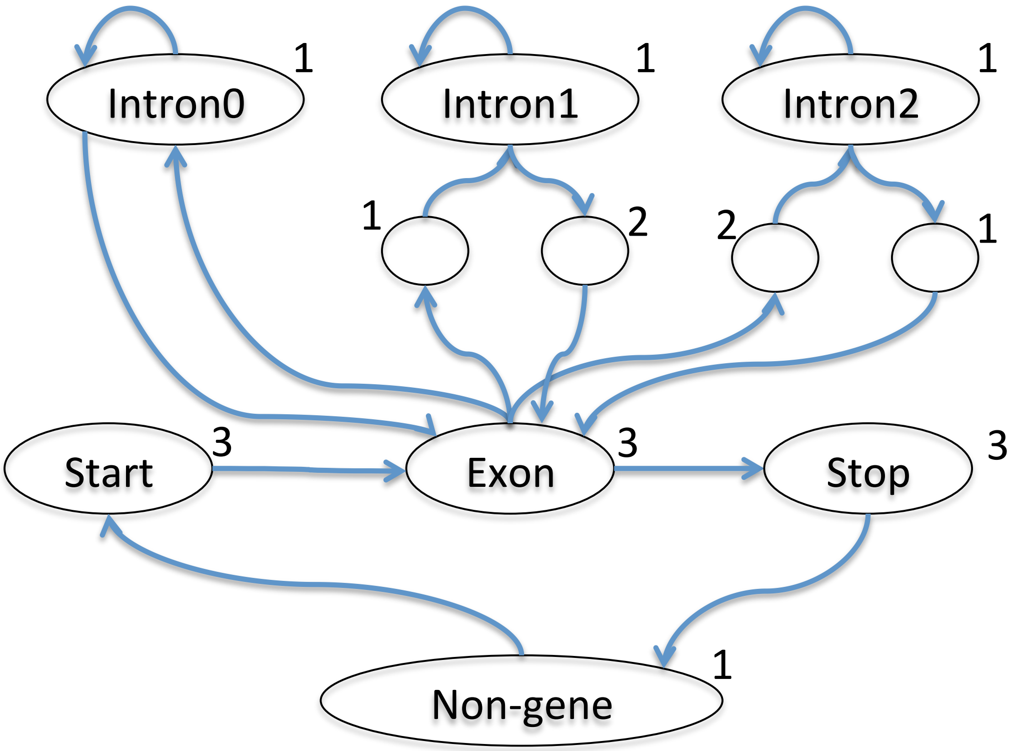
You do not need to submit your complete set of predictions. Instead, use the output of your program to answer the following questions:

c.1) (5 points) Which protein contains the longest hydrophobic region?

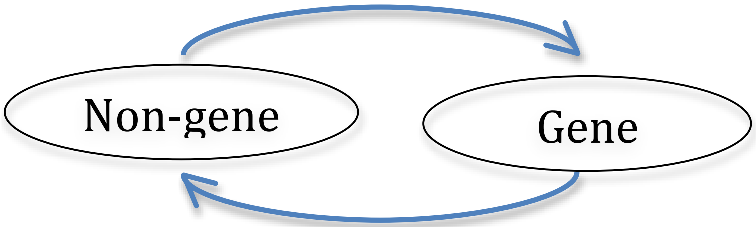
c.2) (5 points) Which protein contains the largest fraction of amino acids annotated as belonging to a Mixed region?
c.3) (5 points) What are the observed overall length distributions of Hydrophobic, Hydrophilic, and Mixed regions? Show the three distributions on a graph. Comment on whether these distributions match the properties listed above.
c.4) (5 points) What are the observed amino acid frequencies in Hydrophobic, Hydrophilic, and Mixed regions? Comment on whether these frequencies match the properties listed above.

d) (5 points) Given the observations made above, how would you revise the your HMM in order to better capture the statistical properties of each type of regions? Only give a high-level description of the changes you would make (5 lines max).

Question 2. (10 points)
The last gene-finding HMM seen in class (depicted below) actually has a small probability of predicting (i.e. emitting) a gene with an in-frame STOP codon.
a) (5 points) How could this happen?
b) (5 points) How could you fix it?



Question 3. (15 points)
If we think of an HMM as a machine to generate a random sequence of observations, the number of consecutive steps the path will remain at a given state follows a geometric distribution (https://en.wikipedia.org/wiki/Geometric_distribution). This means, for example, that, in our gene-finding HMM, the length distribution of exons, introns, and intergenic regions will be assumed to be geometric. However, in reality, these regions have length distributions that can be far from geometric. Consider the very simple two-state gene finding HMM shown below. Assume that we have a desired gene length distribution, provided in the form of a discrete probability distribution for lengths ranging from 1 to 1000: Pr[length = k] = p_k.
Describe (in at most half a page) how you could modify the HMM below to produce a second HMM, where the Gene state will probably need to be subdivided in several Gene sub-states, where the distribution over the duration of stay in the set of gene states is exactly the target length distribution. Describe not only the states of your HMM but also the transition probabilities.



Question 4 (35 points).
A ChIP-Seq experiment was performed to identify genomic regions bound by the GATA2 transcription factor. A subset of the regions found can be downloaded at
http://www.cs.mcgill.ca/~blanchem/561/hw3/GATA2_chr1.fa.gz

A set of genomic regulatory regions that are not bound by GATA2 can be found at:
http://www.cs.mcgill.ca/~blanchem/561/hw3/not_GATA2_chr1.fa.gz

Your goal is to identify what 6-nucleotide consensus sequence (made of the extended alphabet A, C, G, T, A|G, C|T, A|C|G|T) is the most likely to represent the binding site of the GATA2 transcription factor. NOTE: IF YOUR PROGRAM TAKES TOO MUCH TIME TO RUN TO ENUMERATE ALL 7^6 POSSIBLE MOTIFS OF LENGTH 6, REDUCE THE MOTIF LENGTH TO 5. THIS SHOULD MAKE THE ENUMERATION MUCH FASTER.

a) (10 points) Describe the scoring method you will use to identify the best consensus sequence

b) (15 points) Implement an algorithm to identify the highest scoring motif in the set of given sequences. Submit your program on MyCourses. Ensure that your program compiles and provides useful output, requiring only the two FASTA files as input to provide a prediction.

c) (10 points) What consensus sequence obtains the highest score?

Good luck!