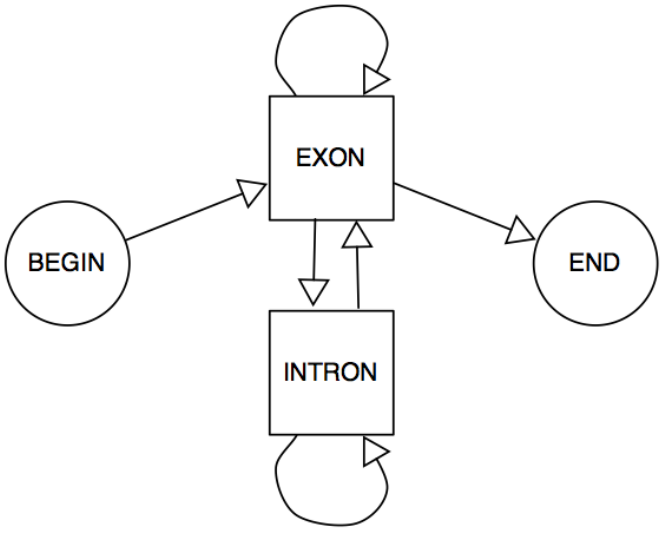
ECE 551 Homework 1 (Spring 2016)

Due date is March 25th, 2016 beginning of class

1. Programming (40 points) Write a program that takes as input two sets of mRNA sequences, trains an exon-intron predicting HMM based on the first set (i.e., training set), and predicts the exon intron locations of the second set (i.e., test sequences). The HMM that you need to train is shown below. Your program should train the transition and emission probabilities using Maximum Likelihood Estimate technique with Laplace smoothing (i.e., adding one pseudocount to each count). It should then run the Viterbi algorithm on each test sequence to predict which positions belong to exons and which positions belong to introns. In order to avoid numerical issues, you should use the logarithms of probabilities in the Viterbi algorithm instead of the probabilities themselves. During the traceback step, if there is a tie, prefer to trace back to the intron state.

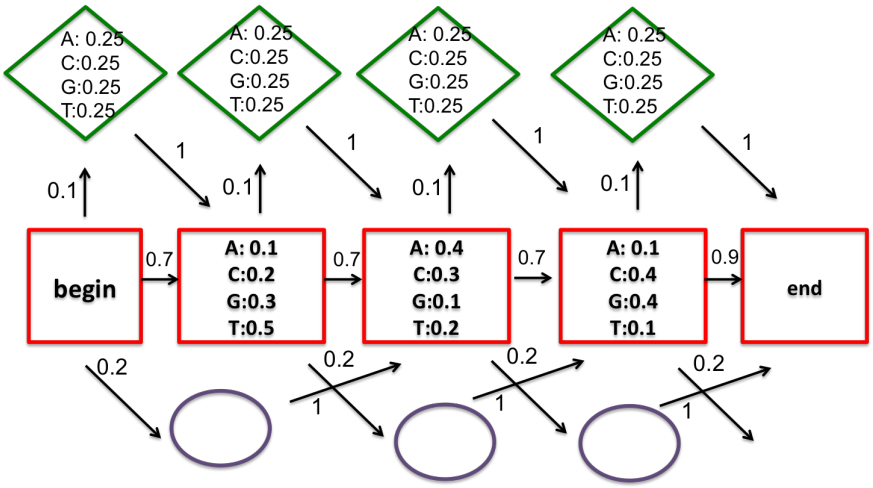


Your program should take as input two filenames, the first one contains the training sequences and the second one contains the test sequences. Each line of these files will correspond to a single sequence. The training sequence file will have exon positions in uppercase and intron positions in lowercase. The test file will have all the characters in lowercase (i.e., labels are unknown). Your program should output its predictions to a file where each test sequence is printed on a separate single line with predicted exon positions uppercase and predicted intron positions in lowercase. Sample input and output files are provided. You can start with the small example first to debug

your code. Write another program that takes as input two files, true annotations and predicted annotations, and outputs the accuracy, recall and precision of the predictions. These measures are defined as follows:

$$
\begin{aligned}
accuracy &= \frac{\text{\# of correctly predicted positions}}{\text{total \# of positions}} \\
recall &= \frac{\text{\# of correctly predicted exonic positions}}{\text{total \# of true exonic positions}} \\
precision &= \frac{\text{\# of correctly predicted exonic positions}}{\text{total \# of predicted exonic positions}}
\end{aligned}
$$

2. Written (15 points) Find five different ways in which the sequence ATG could be generated based on the profile HMM below (red: match, green:insertion, purple:deletion). Calculate the probabilities of these paths. You can label the match states with M_1, M_2 and M_3; insertions states with I_1, I_2 and I_3 and deletion states with D_1, D_2 and D_3



3. Written (20 points) Many genes in E.coli have a specific motif upstream of the start codon (the first codon of a gene). This motif could be GGAGG, GGAG or GAGG with frequencies 72%, 16% and 12%, respectively. The distance between this motif and the start codon ranges from 4 to 6 nucleotides in length. Design an HMM to model this motif until the beginning of the start codon (the end state of the HMM should correspond to the beginning of start codon). Draw the HMM and also give transition and emission probabilities.

4. Written (25 points) Estimate the parameters of a profile HMM built from the following multiple sequence alignment.

G	C	A	G
G	-	-	G
G	-	A	G
G	C	T	G
A	-	A	C
G	-	A	C
G	-	G	G
A	-	A	C

Use Laplace smoothing (pseudocounts of 1) when calculating the emission and transition probabilities. You should design a column as a match state if the number of gap symbols is less the number of letter symbols.

a) Draw the HMM. Show the emission and transition probabilities.

b) Use Viterbi algorithm to find the most likely path for GCCAG.