COMP 571: Homework #2
Spring 2017

Assigned on February 9, 2017.

Due at the beginning of class on February 21, 2017.

The Honor Code applies to all homework sets. Sign the pledge on your solutions.

1 HMMs and the Viterbi Algorithm [20 pts]

The Viterbi algorithm for finding the most probable path in an HMM seeks the path Π^* such that

$$\Pi^* = \text{argmax}_{\Pi} P(X, \Pi).$$

1. Show that this is equivalent to

$$\Pi^* = \text{argmax}_{\Pi} P(\Pi|X).$$

2. One might ask why we are interested in this Π^* and not

$$\Pi^* = \text{argmax}_{\Pi} P(X|\Pi)$$

instead. The latter state sequence Π^*, after all, is the one that, if given, has the highest probability of producing X. Discuss why this is the wrong approach (using an example HMM).

2 HMMs and Length Distribution [25 pts]

Consider the HMM of Fig. 1. States X and Y emit symbols x_i and y_j with probabilities q_{x_i} and q_{y_j}, respectively.

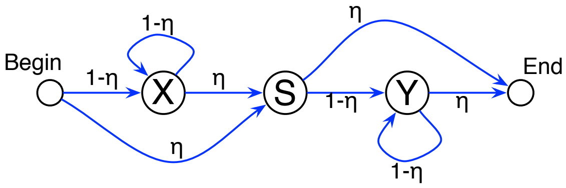

Figure 1: HMM for Problem 2.

1. This HMM can generate pairs of sequences (x, y). What is the probability that sequence x has length t under the condition that $L_x \leq n$ and $L_y \leq m$ for some integers n, m, where L_x and L_y denote the sequence lengths.

2. What is the expected length of sequences from the given HMM? Based on your derivation, what should η be set to if the characteristic length of sequences x and y is L^*? For this part, assume that L_x and L_y are not bounded; that is, $n \to \infty$ and $m \to \infty$. (for the expected length, make use of the definition of expected value of a discrete random variable). You can make use of the formula

$$\sum_{i=0}^{n} i(1-a)^i = \frac{(1-a)(1-(n+1)(1-a)^n + n(1-a)^{n+1})}{a^2}.$$

3 Finding Gene Structure using HMMs [30 pts]

In 1987, Chris Burge invented the first modern gene finder, a program that takes a sequence of DNA as input and outputs a labeling for the sequence indicating the locations of different genes. His tool, Genscan, was the first gene finder to completely represent the known structure of a gene using a Hidden Markov Model. In this exercise you will design, implement, and test a simple HMM gene finder in Matlab.

1. **HMM Design** Construct and draw the HMM that models the following gene structure:

   A gene always consists of three regions in the following order: the *Promoter*, the *Coding Region*, and the *PolyA*. Genes are separated by *Intergenic* regions that do not code for protein. These different regions have differing average nucleotide frequencies shown in Table 1. The average length of each region is as follows: Promoter - 5 base pairs (bp); Coding Region - 10 bp; PolyA - 6 bp; Intergenic Region - 10 bp. There is equal probability of starting in any of the different regions.

Table 1: Nucleotide frequencies in each of the different regions.

Nucleotide	*Promoter*	*Coding Region*	*PolyA*	*Intergenic Region*
A	0.1	0.1	0.7	0.25
T	0.1	0.2	0.1	0.25
C	0.4	0.3	0.1	0.25
G	0.4	0.4	0.1	0.25

2. **HMM Implementation** Implement your HMM model using the HMM package for Matlab (`http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html`) or any package of your choice (but specify clearly what package and programming language you used). Test your HMM model by generating a DNA sequence (length 200 nucleotides) and labeling using your HMM. Run your HMM on the generated DNA sequence and compare the labeling computed to the generated labeling. In your solution, provide the matrices for your HMM, the test DNA sequence, generated labeling, computed labeling, and comments on the errors you see.

3. **Augmenting the Gene Finder** The HMM you designed in Part 1 will only find genes when they are oriented so that the Promoter is read first (this gene is said to be on the *positive strand*). Design and draw an HMM that can find genes on both the positive and negative strands (on the negative strand, the gene is "read" from the PolyA region towards the Promoter). Assume that genes do not overlap.

4 Profile HMMs [25 pts]

1. Estimate the parameters of a profile HMM for the following multiple alignment of DNA sequences:

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's rule to augment the pseudo-counts for the emission and transition probabilities. Further, consider a column with fewer than 50% gap symbols to be a match state.

2. Draw the state diagram.

3. Use the log-odds score version of the Viterbi algorithm to align sequence GCCAG to the profile HMM you built. To define the log-odds scores, assume that the background model is an independence model with $P(A) = P(T) = 0.3$ and $P(C) = P(G) = 0.2$.