Study Questions

These are optional problems to give you additional experience working with Hidden Markov models. These problems will not be graded. A solution set will be posted tomorrow.

1. Define an HMM $\mathcal{H}$ with three states $\{S_1, S_2, S_3\}$; alphabet $\{a, b, c\}$; initial state probabilities $\pi_1 = 0.25, \pi_2 = 0.75, \pi_3 = 0$; and the following transition and emission probabilities:

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>0.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>$S_2$</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>$S_3$</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.25</td>
<td>0.0</td>
<td>0.75</td>
</tr>
</tbody>
</table>

This problem asks you to calculate the total probability of a sequence and the most probable path, given that sequence. Normally, we calculate these quantities using dynamic programming algorithms. With this very simple HMM, it is possible to calculate these quantities by enumerating all paths with non-zero probability, which is what we ask you to do here.

(a) Draw the state diagram of this HMM and show the transition probabilities.

(b) Write down all of the possible state paths for the sequence $O = a, c, a$. 
(c) What is $P(O)$?

(d) What is the most probable path, $Q^*$? What is $P(O, Q^*)$, the probability of emitting $O$ via the most probable path?

(e) The Forward algorithm calculates $P(O)$, the probability that the model will emit $O$ over all possible paths. One might consider approximating this probability by calculating $P(O, Q^*)$ using the Viterbi algorithm. For this particular HMM, would $P(O, Q^*)$ be a good approximation $P(O)$? Explain your reasoning.
2. Hidden Markov model design:

(a) Some prokaryotes use non-canonical start codons, in addition to the canonical start, AUG. A study in *E. coli* reported that 3,544 genes used the canonical start codon, 612 used *GUG*, and 130 used *UUG*.

Design an HMM to model start codons in *E. coli*. Give the topology of your model and the initial, transition, and emission probabilities. Use a pseudocount of $b = 1$ to account for codon variations not observed in the data. You may assume that insertions and deletions never occur within start codons.
(b) Many genes in E. coli have a Shine-Dalgarno (SD) sequence upstream of the start codon. The SD sequence contributes to translation initiation by binding to a complementary sequence in 16S rRNA. The canonical SD motif is GGAGG, although the variants GGAG and GAGG are observed in 16% and 12% of SD sequences, respectively. The SD motif is separated from the start codon by a variable length spacer of nucleotides (n): GGAGGnn...nnAUG. In E. coli, this spacer ranges from 4 to 6 nucleotides in length.

Design an HMM to model the SD sequence in E. coli. Give the topology of your model and the initial, transition, and emission probabilities. Do not use pseudocounts. Your model should emit the sequences GGAGG, GGAG and GAGG in the appropriate frequencies. It should not emit GAG, AGG, etc. Your model should emit spacers of 4, 5, or 6 nucleotides, with equal probability. The frequencies of the nucleotides in the spacer region should be 0.25 for all four bases. Give your model silent Start and End states.
3. *Multiple sequence alignment using HMM’s:* Your goal is to obtain a global multiple alignment of the following unlabeled sequences

\[ \text{WRCCTGC, WCCGGCC, WCGCC, WRCCCGC, WHCCGC.} \]

You decide to align the sequences by constructing a Profile HMM with seven match states, \( M_1 \) to \( M_7 \). States \( M_0 \) and \( M_8 \) are silent *start* and *end* states, respectively, and do not emit symbols.

(a) How many insertion states will there be in this model, assuming you use the canonical Profile HMM architecture? How many deletion states?

(b) What estimation method will you use to determine the initial, transition and emission probabilities of this model? What algorithms are required for this method?

(c) Once you have instantiated your model, you need to label your sequences. What algorithm(s) could you use to assign a state to each amino acid in each protein sequence?
(d) Suppose the paths through the HMM corresponding to the five sequences are

\[
\begin{align*}
& m_0 \quad m_1 \quad i_1 \quad m_2 \quad m_3 \quad m_4 \quad m_5 \quad d_6 \quad m_7 \quad m_8 \\
& m_0 \quad m_1 \quad m_2 \quad m_3 \quad m_4 \quad m_5 \quad m_6 \quad m_7 \quad m_8 \\
& m_0 \quad m_1 \quad m_2 \quad d_3 \quad d_4 \quad m_5 \quad m_6 \quad m_7 \quad m_8 \\
& m_0 \quad m_1 \quad i_1 \quad m_2 \quad m_3 \quad m_4 \quad m_5 \quad d_6 \quad m_7 \quad m_8 \\
& m_0 \quad m_1 \quad i_1 \quad m_2 \quad m_3 \quad d_4 \quad m_5 \quad d_6 \quad m_7 \quad m_8 
\end{align*}
\]

Give the alignment of the sequences specified by this labeling.

(e) Based on the alignment, would you perform “model surgery” on this profile HMM? Why or why not? If so, which states would you add and/or delete?
(f) If you do decide to perform model surgery, you will need to update your parameter values. What estimation method should you use to determine the new initial, transition, and emission probabilities? Give the name of the method and any equations you might need.