Seven11 Assignment 4

Due 11:59pm, Friday, November 19th

Your name:

This assignment is required for students taking the 12-unit course only.

We have been exploring Markov and Hidden Markov models for problems in sequence analysis. This assignment explores properties of sequences that are not well suited to Markov models. Some of these models can be modeled by context sensitive Hidden Markov models (csHMMs), which are described in this paper by Yoon:

- Yoon, Hidden Markov Models and their Applications in Biological Sequence Analysis. *Current Genomics* (2009) 10: p402-415.

  `http://www.cs.cmu.edu/~durand/03-711/2021/Homework/Yoon09.pdf`

NOTE: This text uses different notation from that used during lecture or in the course textbooks. In particular, the article uses O to represent the alphabet, not an observed sequence of symbols. The first half of the paper provides a good review of HMMs. I recommend you read it to familiarize yourself with the notation used throughout the rest of the paper.

Homework must be submitted by 11:59pm electronically to Canvas.

Collaboration is allowed on this homework. You must hand in homework assignments individually. List the names of the people you worked with, and cite all additional references you used to complete this assignment:

1. Palindromes: Restriction enzyme recognition sites are conserved motifs that are 4, 6 or 8 nucleotides long and are commonly DNA palindromes. A DNA palindrome is a sequence that is the same when read 5' to 3' on either strand; that is, the first half of the palindrome is identical to the reverse complement of the second half of the palindrome, read backwards. For example, `C-T-A-G` is a DNA palindrome of length 4; `C-T-T-C` is not.

   (a) Draw a state diagram for a ***Markov chain*** model that assigns a non-zero probability to any nucleotide palindrome of length 4 that starts with an A. All other sequences should have zero probability. Each state in your model should correspond to a single nucleotide. It is not necessary to label the edges with transition probabilities.

   (b) How many states are required for a ***Markov chain*** that models all nucleotide palindromes of length 4, starting with *any* nucleotide?

(c) Draw a state diagram for a Markov chain model that assigns a non-zero probability to any palindromic nucleotide sequence of length 6 starting with an A.

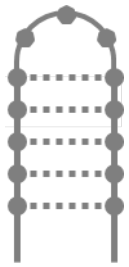
(d) How many states are required for a Markov chain that assigns a probability to any nucleotide palindrome of length 6, with no restrictions on the first nucleotide?

(e) Are Markov chains a suitable model for palindromes? Why or why not?

2. In the context sensitive HMM shown in Figure 4 in Yoon, 2009, the states P_1 and C_1 share memory implemented as a stack (First-In-Last-Out). This model emits sequences of the form $x_1x_2 \ldots x_Nx_N \ldots x_2x_1$ and $x_1x_2 \ldots x_Nx_{N+1}x_N \ldots x_2x_1$.

   Suppose you replaced the stack with a queue (First-In-First-out) data structure. What would the sequences emitted by this modified csHMM look like?

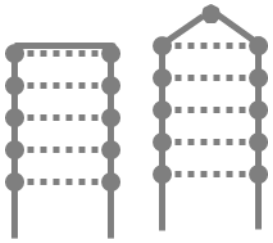
3. In addition to restriction sites, near palindromes with complementary base pairing arise frequently in RNA molecules because they correspond to stem secondary structures, like the hairpin shown here:



The csHMM in Figure 4 in Yoon, 2009 can be easily modified to emit DNA palindromes with complementary base pairing. For example, if the symbol emitted by P_1 at time i is an A, then the emission probabilities will be

$$e(x_j|x_i = A, y_i = P_1, y_j = C) = \begin{cases} 1 & x_j = T, \\ 0 & x_j \in \{A, C, G\}. \end{cases}$$

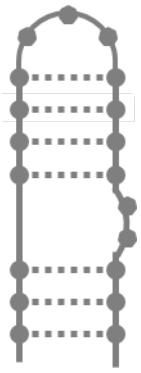
The model in Fig. 4 only emits exact palindromes. If these formed RNA secondary structures, they would look like this:



However, in RNA molecules, the two halves of the palindrome are typically separated by several residues, as shown in the hairpin figure at the top of the page.

Draw a schematic of a modified version of the csHMM in Fig. 4 that emits sequences of a hairpin with *at least* 3 and *at most* 5 residues between the two halves of the stem. The two halves of the stem may be of any length greater than zero.

4. Sometimes RNA stems are interrupted by a bulge, like this one:



Draw a schematic of a modified variant of the csHMM in Fig. 4 that emits an RNA stem secondary structure with a bulge on one side. The two halves of the stem may be of any length greater than zero. The hairpin should be of exactly length 3, and the bulge should be of length 2, exactly.

5. Draw a schematic of a modified variant of the csHMM in Fig. 4 that emits this secondary structure, called a pseudoknot. The two hairpin sequences should each be of length *at least* one.