GS 373 Homework 6 - KEY

Bioinformatics Questions (80 points)

1. Gene structure (20 points)
   Go to the **UCSC Genome Browser** (https://genome.ucsc.edu/cgi-bin/hgGateway use the **hg38** build) and look up the human gene ***CD3E***.
   a) (5 points) **How many transcript variants does this gene have**, according to GENCODE v24? **How many exons does each one have**?
   2 variants, 9 exons
   b) (5 points) Find the "Human ESTs" option below the plot, set the drop-down menu to full and refresh: **What are ESTs and what does this EST track represent**?
   ESTs (expressed sequence tags) are empirical data, usually from RNA-seq, of a transcript. Each line in the browser is a different piece of evidence.
   c) (5 points) Based on this track, **which of the transcripts from part a) above has more EST evidence**? (can be answered by eye)
   The 9 exon variant
   d) List **5 specific sequence structures** that might be used during *ab initio* gene prediction.
   start codon
   stop codon
   TATA box
   splice opener/closer
   shine-Delgarno
   Terminator
   (others)

2. Markov models (20 points)
   a. (10 points) **Draw a diagram of a Markov model of DNA sequence** that can generate at least one sequence that **contains all four nucleotides**. Include a valid set of **transition probabilities** (each state's outgoing transitions must sum to 1). Then, **calculate the probability of the sequence CATG** given your model, assuming a 25% probability of starting at any particular nucleotide.
   There should be a non-zero path connecting each nucleotide to at least one other nucleotide. The outgoing transition probabilities for each nucleotide must sum to 1.0.
   b. (10 points) Draw a diagram of **another Markov model of DNA sequence in which every possible sequence of the same length is equally probable** (has the same probability according to the model).
   There should be a non-zero path connecting each nucleotide to every other nucleotide. Every transition probability is 0.25.

3. Hidden Markov models (20 points)
   Imagine some researchers studying the spread of flu infection. In one study, these researchers identified and tracked flu-infected individuals, and measured how long they were highly contagious over the course of their infection. They calculated the share of infected individuals that displayed different symptoms - a fever and/or a cough – while they were contagious, and also during the stages at the beginning and end of their infections when they were not.

In a second study, the researchers again collected daily records on whether a different group of flu-infected individuals displayed a fever and/or a cough. However, for this study they were unable to measure whether each participant was contagious on each day of their infection. They decide to use a Hidden Markov Model to infer how long and when each participant in the new study was likely contagious.

a. (10 points) **Draw a diagram of this HMM.** Include all arrows, but no need for probabilities.
b. (5 points) What would the **hidden states** of this HMM be? **What data could you use to assign their transition probabilities**?
   Hidden states include beginning/end of infection and contagious/non-contagious. You could assign transition probabilities based on how long people spent in each state in the first study.
c. (5 points) What would the **emissions** be? **What data could you use to assign the emission probabilities for each hidden state**?
   Emissions include symptoms or no symptoms. You could estimate the emission probabilities from the first study as well.

4. (20 points)
   Read the short paper "What is a hidden Markov model?" by Sean Eddy expanding on ideas from lecture. Answer the following questions about the splice site-finding HMM proposed in the paper:
   a. (9 points) What **prior knowledge** about exons, splice sites, and introns is used?
      nucleotide content of each feature.
   b. (5 points) **State what calculation should be performed with this model to find the most likely state assignment for a particular site in the nucleotide sequence**.
      forward-backward algorithm
   c. (6 points) **What is another example of a genomics problem where HMMs would not be appropriate** (besides the one given in the paper) and **why**?
      Anything where each base isn't independent