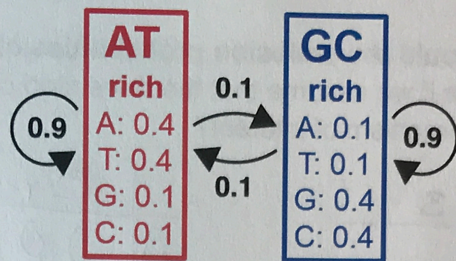
GS 373 Homework 7 - **KEY**

Bioinformatics Questions (80 points)

1. (30 points) **Consider the hidden Markov model that describes DNA sequences in two different states, AT rich and GC rich.** The numbers above the arrows denote transition probabilities, and the numbers within the boxes denote emission probabilities. Assume the initial probabilities of the two states are the same.



   a. (15 points) Based on this model, **calculate the joint probability of observing the sequence ACT with each of the following state paths**:

      i. The state path AT-rich, GC-rich, AT-rich

$= P(\text{init in AT-rich}) * P(\text{emit A} \mid \text{AT-rich}) * P(\text{AT-rich} \to \text{GC-rich}) * P(\text{emit C} \mid \text{GC rich}) * P(\text{GC rich} \to \text{AT rich}) * P(\text{emit T} \mid \text{AT rich})$

$= 0.5(0.4)(0.1)(0.4)(0.1)(0.4) = \boxed{0.00032}$

      ii. The state path AT-rich, AT-rich, AT-rich

$= P(\text{init AT}) * P(\text{emit A} \mid \text{AT}) * P(\text{AT} \to \text{AT}) * P(\text{emit C} \mid \text{AT}) * P(\text{AT} \to \text{AT}) * P(\text{emit T} \mid \text{AT})$

$= 0.5(0.4)(0.9)(0.1)(0.9)(0.4) = \boxed{0.00648}$

      iii. The state path GC-rich, GC-rich, GC-rich

$= P(\text{init GC}) * P(\text{emit A} \mid \text{GC}) * P(\text{GC} \to \text{GC}) * P(\text{emit C} \mid \text{GC}) * P(\text{GC} \to \text{GC}) * P(\text{emit T} \mid \text{GC})$

$= 0.5(0.1)(0.9)(0.4)(0.9)(0.1) = \boxed{0.00162}$

   b. (15 points) **Use the Viterbi algorithm to find the most likely state path for the observed sequence ACAT according to this model**. Show your work and be sure to label the state each base was emitted from.

	A	C	A	T
AT	0.5 (0.4) = 0.2	0.2 * 0.9 * 0.1 = 0.018 0.05 * 0.1 * 0.1 = 0.0005	0.018 * 0.9 * 0.4 = 0.00648 0.018 * 0.1 * 0.4 = 0.00072	0.00648 * 0.9 * 0.4 = 0.0023 0.00162 * 0.1 * 0.4 = 0.00006
GC	0.5 (0.1) = 0.05	0.2 * 0.1 * 0.4 = 0.008 0.05 * 0.9 * 0.4 = 0.018	0.018 * 0.1 * 0.1 = 0.00018 0.018 * 0.9 * 0.1 = 0.00162	0.00648 * 0.1 * 0.1 = 0.000006 0.00162 * 0.9 * 0.1 = 0.00015

Most probable path

A C A T

AT → AT → AT → AT (0.0023 > 0.00015)

2. (30 points) **The hidden Markov model below is a variation of the one proposed in class to identify open reading frames in DNA sequence, where the subscripts denote the position of a nucleotide within a codon.** "s" refers to the start codon, "e" refers to the stop codon, and "ORF" refers to the codons between the start and the end. s_1 refers to the first nucleotide in a start codon, ORF_2 refers to the 2nd nucleotide in a codon between the start and stop codons, etc.

   a. (20 points) **What should the emission probabilities of each of the hidden states e_1, e_2, e_3 and s_1, s_2, s_3 be** if we assume that the three stop codons UAA, UAG, and UGA are equally likely in the organism of interest?

	S_1	S_2	S_3
A	1.0	0.0	0.0
C	0.0	0.0	0.0
G	0.0	1.0	0.0
T	0.0	0.0	1.0

	e_1	e_2	e_3
A	0.0	1/3	2/3
C	0.0	0.0	0.0
G	0.0	2/3	1/3
T	1.0	0.0	0.0

   b. (10 points) Suppose you know that a specific nucleotide in an mRNA sequence is part of a coding region because it was shown experimentally to be directly bound by ribosomes. **Describe how you could use the HMM to infer the most likely reading frame for the mRNA sequence.**

Use forward-back algorithm to calculate the probability the nucleotide is in reading frame 1, 2, or 3.

Reading Frame 1: is the nt in S_1, ORF_1 or e_1?
Reading Frame 2: is the nt in S_2, ORF_2, or e_2?
Reading Frame 3: is the nt in S_3, ORF_3, or e_3?

Choose the one w/ the highest probability

3. (20 points) **Train the parameters of an HMM** with the two states A-rich and T-rich that describe sequences of A's and T's using the training sequence below, whose state sequence is also known.

Sequence	A	T	T	T	T	A	A	A	A	A	A	A	A	A	A	T	T	T	T
State	t	t	t	t	a	a	a	a	t	t	a	a	a	a	a	a	t	t	t

a. (10 points) **Compute the transition probabilities** for the states learned from the above sequence.

- There are 4 transition probabilities: t→t, t→a
  a→a, a→t

- The probabilities most sum to 1.0 coming out to a given state

transition	formula	answer
t→t	(# t→t) / (#t→t + # t→a)	6 / (6+2) = 6/8
t→a	# t→a) / (#t→a + #t→t)	2 / (6+2) = 2/8
a→a	# (a→a) / (#a→a + #a→t)	8 / (8+2) = 8/10
a→t	#(a→t) / (#a→t + #a→a)	2 / (8+2) = 2/10

b. (10 points) **Compute the emission probabilities** for the states learned from the above sequence.

- There are 4 emission probs P(emit T | t)
  P(emit A | t)
  P(emit T | a)
  P(emit A | a)

- The prob most sum to 1.0 for a given state

emission	formula	answer
emit T | t	# emit T | t / (# emit T | t + #emit A | t)	6 / (6+3) = 6/9
emit A | t	#emit A | t / (#emit T | t + #emit A | t)	3 / (6+3) = 3/9
emit T | a	#emit T | a / (# emit T | a + #emit A | a)	2 / (2+8) = 2/10
emit A | a	# emit A | a / (# emit A | a + #emit T | a)	8 / (2+8) = 8/10