Hidden Markov Models - Part 1

Question 1.

Given a hidden Markov model (HMM) with the following elements:

- States, *Q = {B, Q1, Q2, Q3, E}*
- Alphabet, *Σ = {C, G, T}*
- Transition probabilities between the states, *A =*

	B	**Q1**	**Q2**	**Q3**	**E**
B	0	1	0	0	0
Q1	0	0	0.4	0.4	0.2
Q2	0	0.8	0	0	0.2
Q3	0	0	1	0	0
E	0	0	0	0	0

- Emission probabilities, *E =*

	C	**G**	**T**
Q1	0.5	0.5	0
Q2	0.5	0	0.5
Q3	0	0.5	0.5

Note: B and E are special begin and end states*, respectively. They do not emit symbols. When generating the output, the HMM is initiated in state B and terminated in state E, i.e. each sequence of states starts in B and ends in E.*

a) Draw the state diagram of this HMM.

b) Given two observed sequences, X_1 and X_2, find all possible **state paths (π)** for the sequences. Besides, calculate the **probability $P(X_i|HMM)$** for each observed sequence X_i, under the given HMM, using all possible state paths.

From *Durbin et. al*: $P(x,\pi) = a_{0\pi_1} \prod_{i=1}^{L} e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}}$

- X_1 = CGT
- X_2 = CTC

Question 2. Gene finding using a Hidden Markov Model.

Suppose you model a stretch of DNA containing two types of regions. Region 1 contains bases A,T,C, and G with equal frequency. Region 2 has a higher GC frequency. We also have some knowledge about the length of these regions and the overall length of the modeled sequence. Consider the following simple 4-state HMM:

- States, *Q = {B, Q1, Q2, E}*
  (Q = {B (Begin), Q1 (Region 1), Q2 (Region 2), E (End)})
- Alphabet, Σ = *{A, T, G, C}*
- Transition probabilities between the states, *A* =

	B	**Q1**	**Q2**	**E**
B	0	0.5	0.5	0
Q1	0	0.7	0.1	0.2
Q2	0	0.5	0.3	0.2
E	0	0	0	0

- Emission probabilities, *E* =

	A	**T**	**G**	**C**
Q1	0.25	0.25	0.25	0.25
Q2	0.1	0.1	0.5	0.3

a) Probability for a known path. Compute P(X=ATG, π =BQ1Q2Q1E | HMM), which is the probability of the following state path π =BQ1Q2Q1E, and emitting the sequence ATG. (Please, indicate the terms used in the calculation.)

From *Durbin et. al*: $P(x,\pi) = a_{0\pi_1} \prod_{i=1}^{L} e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}}$

b) Most probable path: Viterbi algorithm (see *Biological sequence analysis*, p. 57). Which state sequence was most likely to have generated the observation sequence ATG? First, use the Viterbi algorithm to fill the following matrix, then determine the most probable sequence of states.

Modified from *Durbin et. al* (E for end state):

Initialisation $(i = 0)$: $v_0(0) = 1,\ v_k(0) = 0$ for $k > 0$

Recursion $(i = 1 ... L)$: $v_l(i) = e_l(x_i)max_k(v_k(i-1)a_{kl})$ for emitting states

Termination: $v_E(L+1) = P(x, \pi^*) = max_k(v_k(L)a_{kE})$

	-	**A**	**T**	**G**	**-**
	0	**1**	**2**	**3**	**4**
B					
Q1					
Q2					
E					

c) Would you expect higher, lower, or the same probability from the forward algorithm (see *Biological sequence analysis*, p. 59) compared to the result obtained in b)? Why?