1. Given is a hidden Markov model *HMM* with the following definitions:
   - States Q = {B (Begin), Q1, Q2, Q3, E (End)}
   - 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

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

   - Emission probabilities E =

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

   a. Draw the state diagram of this HMM
   b. Give all possible state paths (π) for the sequences below. For each state sequence also give the probability P(X|*HMM*) for the observed sequence X, under the given HMM, using all possible state paths.
      - $X1$ = CGT
      - $X2$ = CTC

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 frequency of G and C than of A and T. 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 (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

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

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

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? Use first the Viterbi algorithm to fill the following matrix, then determine the most probable sequence of states.

	-	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?