HMM assignment part 2

In this exercise we will train an HMM to recognise domain (D) and linker (L) regions in protein sequences. To simplify the exercise, we will have a three letter amino acid alphabet: H (hydrophobic), P (polar) and C (charged) amino acids. You may assume that hydrophobic residues are more common in domains and while be less abundant in linker regions compared to charged and polar amino acids. Furthermore, the linker region links two domains by definition.

- States Q = {B, D, L, E}
- Alphabet Σ = {H, P, C}

Let X_1= CCHHPCCPHHCH,

Transition probabilities between the states, A_1 =

	B	L	D	E
B	0	0.5	0.5	0
L	0	0.7	0.2	0.1
D	0	0.2	0.7	0.1
E	0	0	0	0

Emission probabilities, E_1 =

	H	P	C
L	0.5	0	0.5
D	0	0.5	0.5

These matrices are also provided as .txt files in the exercise folder.

1) Run the Viterbi algorithm on the sequence X_1= CCHHPCCPHHCH with the HMM defined by A_1 *and* E_1. Please give the Viterbi matrix, the most likely state path, and its probability. (Modify MainProgram.py and look at Viterbi.py)

2) Implement the forward and backward algorithm (see *Biological sequence analysis*, p. 59-60). What probability do you get for P(X_1 | *HMM*) in each case? Please also give the forward and backward matrices. [Modify BaumWelch.py, look at Viterbi.py. Note that you can look for the phrase "###start coding here" in the provided template to see where you will need to insert code.]
3) Implement the Baum-Welch algorithm (see *Biological sequence analysis*, p. 65) by completing the scripts that are provided for you – or by implementing your own version. Run one iteration of the Baum-Welch algorithm on the sequence X_1 and A_1

and E_1 matrices given above. Please give the new estimates of the A and E matrices and the new probability P(X_1 | *HMM*). [Modify BaumWelch.py]

4) Now we will start training the HMM on a set of 200 training sequences, to estimate the "real" A and E matrices. You will do the training with different prior A and E matrices.
   a) Do the Baum-Welch training using the different prior A_1 and E_1 matrices as given above.
   b) Use a sensible setting based on the description of the biological problem above to define to transition matrix A_2 and Emission matrix E_2. (Note that the parameters provided by A_1 and E_1 are not biologically sensible!)
   c) Use the same parameters as in A_2 and E_2, but swap the emission probabilities between the L and D states to define E_3.
   d) Use an emission matrix E_4, for which the emission probabilities are the same for both states.

   For each of the above, what are the new A and E matrices you obtain through training? Explain your results. How (fast) did your search converge? Explain your conclusions in ~ 500-1000 words (1/2-1 A4).
5) Given your findings at 4) and after reading Chapter 3 of *Biological sequence analysis*: what HMM training strategy would you propose if you were given a small set of training sequences *with* state annotation, and a large set of training sequences *without* such annotation? Note that you may assume here that the sequences from both sets are generated by an identical (biological) process.

Optional assignments:

A) Using the Viterbi training algorithm (see *Biological sequence analysis*, p. 66)., repeat question 4).

B) Write a sequence generator which will generate a set of sequences for a given given HMM (predefined emission and transition matrices). Perform the training as proposed in 5).

Marking scheme:

- Part 1: 20 %
- Part 2: 70 % (including your code)
- optional assignments: 10 %