HMM Open Questions

Start Assignment

- Due 13 Dec by 23:59
- Points 30
- Submitting a file upload
- File types pdf
- Available after 25 Nov at 8:00

This is a quiz of open questions for the third assignment HMM of the course, which will count for **30% of HMM assignment grade (20% for Q1 and 10% for Q2)**. In this assignment, you need to answer 2 questions given below.

Please submit your answer as a pdf file by Week 7 Friday 23:59. In the beginning of the pdf, please write down your name and student number.

>>>

Question 1 (20 points)

Train the HMM using the Baum-Welch algorithm on a set of 200 training sequences located in input/training_sequences.fasta, to estimate the "real" A and E matrices. You will do the training with several prior A and E matrices.

Perform the Baum-Welch training using A1 and E1 as priors. (The same matrices used for your implementation, provided in the input folder).
Manually derive biologically sensible parameter values to define transition matrix A2 and emission matrix E2 . You may assume that hydrophobic residues are more common in domains and less abundant in linker regions, compared to charged and polar amino acids. Assume linker regions exclusively occur between two domains.
Use the same parameters as in A2 and E2 , but swap the emission probabilities between the L and D states to define E3.
Use A2 and an emission matrix E4, for which the emission probabilities are identical for both states.

Note that you have to iterate the training process until it converges to a local maximum or until the maximum number of iterations is reached.

For each of the above, what are your prior matrices and what are the new A and E matrices you

obtain through training? Do the different training runs converge to the same parameters? Explain your results. In your discussion, at least include your convergence threshold, iteration limit, and how fast your training runs converged. Tables for prior and posterior probabilities will not be counted towards the word limit. **(Max. 300 words)**

Question 2 (10 points)

Imagine you were able to find structural information for a small subset of your protein sequences. Considering the results of question 1 and the influence that the priors have on the training performance, how would you modify your workflow in order to improve the HMM training? **(Max. 150 words)**