Bioinformatics
Exercise sheet #3

Lena Morrill / Petar Veličković and Sebastian Müller

Lent 2018

Please submit your work by noon on the day before the supervision, either by email at lm687@cam.ac.uk, or leaving it in my pigeonhole in Clare College (Old Court).

Hidden Markov Models

1. State Bayes' theorem.
2. What is the *Markov property* of a Markov chain?
3. Draw a diagram of a Hidden Markov Model and explain its components: transition matrix, emission matrix, etc.
4. Simulate a chain of hidden states ($S = \{\texttt{CpG}, \texttt{non-CpG}\}$) of length 50 given the transition matrix $\begin{pmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \end{pmatrix}$ and initial probabilities $(0.5, 0.5)$. Show your code and the output.
5. From this Markov chain generate emitted states, which correspond to the nucleotides $\{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}$. The emission matrix is

$$\begin{array}{c} \\ \texttt{CpG} \\ \texttt{Non-CpG} \end{array} \begin{array}{c} \begin{array}{cccc} \texttt{A} & \texttt{C} & \texttt{G} & \texttt{T} \end{array} \\ \begin{pmatrix} 0.2 & 0.3 & 0.3 & 0.2 \\ 0.3 & 0.2 & 0.2 & 0.3 \end{pmatrix} \end{array}$$

   Show your code and your hidden and emmitted sequence.
6. Outline the inputs, outputs and time complexities of the following HMM algorithms:
   - Viterbi
   - Forward algorithm
   - Baum-Welch

7. Using the Forward Algorithm, find the likelihood of the sequence `GGCACTGAA` under this model.

8. Run the Viterbi algorithm on the sequence above to find the sequence of hidden states that most likely produced it.

9. Explain the Baum-Welch learning. What are we estimating?

   Optional: implement it and show your results.

10. Explain how you would use HMMs (and which algorithms) in the following scenario:

    Analysis of a transmembrane (located around the cellular membrane) protein secondary structure – namely, for each amino acid of the protein, determining whether it's located *inside the cell, inside the membrane*, or *outside the cell*. You are provided with a training set containing transmembrane protein sequences, along with a labelled sequence of the same length, determining the location of each amino acid. You are also aware that any region of a protein within the membrane will consist of at least 5 and at most 25 amino acids.

11. Describe a profile HMM, and its input and output.