

Bi/BE/CS 183 2021-2022
Instructor: Lior Pachter
TAs: Tara Chari, Meichen Fang, Zitong (Jerry) Wang

Problem Set 8

Submit your solutions as a single PDF file via Canvas by **10am Friday March 4th**.

- If writing up problems by hand, please use a pen and not a pencil, as it is difficult to read scanned submission of pencil work. Typed solutions are preferred.
- For problems that require coding, Colab notebooks will be provided. Please copy and save the shared notebook and edit your own copy, which you should then submit by including a clickable link in your submitted homework. Prior to submission make sure that you code runs from beginning to end without any error reports.

A hidden Markov model (HMM) has n hidden states and m observed states. At any time t , the random variable X_t is the hidden state at time t and the random variable Y_t is the observation at time t . There is a $n \times n$ transition matrix $S = \{S_{ij}\}$, where $S_{ij} = \mathbb{P}(X_t = j | X_{t-1} = i)$ is the probability of transitioning from hidden state i to hidden state j . There is a $n \times m$ matrix $T = \{T_{ij}\}$, where $T_{ij} = \mathbb{P}(Y_t = j | X_t = i)$ is the probability of observing state j while in hidden state i .

Problem 1 (30 points)

Consider the HMM with $n = 2$, $m = 4$ (corresponding to A, C, T, G), and an equal of probability of starting in each of the two hidden states. Suppose that

$$S = \begin{pmatrix} 0.8 & 0.2 \\ 0.05 & 0.95 \end{pmatrix} \tag{1}$$

and

$$T = \begin{pmatrix} 0.2 & 0.5 & 0.1 & 0.2 \\ 0.1 & 0.25 & 0.25 & 0.4 \end{pmatrix}. \tag{2}$$

- (a) (15 points) (decoding) Find the most probable sequence of hidden state to have produced the sequence ACG
- (b) (15 points) (likelihood) Compute p_{ACG} .

Problem 2 (30 points)

Suppose that an HMM with $n = m = 2$ (corresponding to observations 0,1) has $p_{000} = p_{111} = 0$. Show that S and T must satisfy the following conditions,

$$S = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \text{or} \quad \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \quad (3)$$

Problem 3 (40 points)

In this problem you will implement the Needleman-Wunsch algorithm for global alignment of a pair of sequences. Your program will read in a FASTA file containing a pair of DNA sequences, and then run the Needleman-Wunsch algorithm to find their optimal alignment given parameters for matching, mismatching, and unaligned base-pairs. The algorithm constructs an optimal global alignment by dynamic programming.

The Problem notebook is here.

Your edited version of the notebook *must be submitted* for this problem. Reminder to check that your notebook runs all the way through with the the `Runtime` → `Restart` and `Runtime` → `Run All` commands.