Bi/BE/CS183 2020
Profs. Matt Thomson and Lior Pachter
Problem Set 1
Due on Moodle by Thursday January 16th at 12:00pm

Homework policy: You may collaborate with anyone on the homework and consult/use any resources. Please write down on the homework who you worked with and explain what, if any, help you received. Also list all sources that were used for the homework and explain how they helped.

The notation used is as follows: an HMM has n hidden random variables $X_1, \ldots, X_n$ and n observed random variables $Y_1, \ldots, Y_n$. There is a $k \times k$ transition matrix $S = (s_{ij})$ for each horizontal transition $X_r \to X_{r+1}$ and a $k \times l$ transition matrix $T = (t_{ij})$ for each vertical transition $X_r \to Y_r$.

Problem 1 (10 points)

An upper bound of the number of comma free codes for $|\Sigma| = 4, k = 3$.
(Exercise 1.6 from the Lesson 1 notes)
(1) Using letters "ATCG", which length-3 strings must not appear in a comma free code?
(2) For the remaining length-3 strings, some groups of strings cannot appear in a comma free code at the same time. Specify such relations.
(3) Based on (1)(2), estimate an upper bound of the number of comma free codes for $|\Sigma| = 4, k = 3$.

Problem 2 (15 points)

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

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

and

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

Compute p_{ACGT}.

Problem 3 (15 points)

Suppose that an HMM with $k = l = 2$ and $n = 3$ has $p_{011} = p_{110}$ and $p_{100} = p_{001}$. Also suppose the HMM is stochastic, that is, the transition matrix S or T has entries strictly between 0 and 1. Show that p_{000} and p_{111} cannot both be 0.

Problem 4 (40 points)

Implement the Viterbi algorithm for a binary hidden Markov model ($k = l = 2$). Assume the initial state probabilities are both $\frac{1}{2}$. Demonstrate the algorithm by running it with a fixed set of nonzero parameters on a sequence of length 100 of your choice. How can you be sure the algorithm is working correctly?

We have uploaded two articles on Moodle that you may find helpful:

An Introduction to Hidden Markov Models (1986), by L. R. Rabiner and B. H. Juang.

A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition (1989), by L.R. Rabiner.

Problem 5 (15 points)

On Moodle we uploaded a book chapter called "Probabilistic Models for Error Correction of Nonuniform Sequencing Data" by Marcel H. Schulz and Ziv Bar-Joseph (From the book Algorithms for Next-Generation Sequencing Data, pp 131–145). One of the sections describes an algorithm called *SEECER*, for "De Novo Error Correction Using Profile Hidden Markov Models". Explain in one paragraph what "PHMM learning" means and why it "takes $O(n^t)$". You may find the following articles (also uploaded to moodle) helpful:

Profile Hidden Markov Models (1998) by Sean R Eddy

An Introduction to Hidden Markov Models for Biological Sequences (1998) by Anders Krogh

Problem 6 (Optional - no points given)

Let $S \in \{(,)\}^n$. For example $S = (()(()))(()$. Define a *gene parse* to be a sequence $1 \leq i_1, \ldots, i_{2k} \leq n$ where $S(i_{2r-1}) = ($ and $S(i_{2r}) =)$ for $1 \leq r \leq k$. In the example above $(2, 6, 8, 10)$ is a gene parse. If S contains n parentheses, show that the maximum number of gene parse S can contain is given by the Fibonacci number F_{n+1}. Note that in this enumeration the empty parse is a valid gene parse.