02-510/710: Computational Genomics, Spring 2025

HW3: Epigenetics

Version: 1
Due: 23:59 EST, Mar 30, 2025 on Gradescope

Topics in this assignment:

1. Haplotype Inference
2. Metagenomics
3. Hidden Markov Models

What to hand in.

- One write-up (in pdf format) addressing each of following questions.
- All source code. If the skeleton is provided, you just need to complete the script and send it back. Your code is tested by autograder, please be careful with your main script name and output format.

Submit the following file which contain the completed code and the pdf file to gradescope separately.

./S2025HW3.pdf

Please note that all the solutions must be your own. We will check for plagiarism after the final submission.

1. **[20 pts] Haplotyping by hand**

   This will be a simple exercise in haplotyping that should be done by hand. Please show your work.

   After mapping reads and performing variant calling, you have filtered down to only heterozygous sites and encoded the variants in 0/1 format to denote wild-type/variant.

   Below each read is listed on a line, where '-' means that the read doesn't cover that variant locus.

```
----------------1101011111----------------------------
----------1010010010100001001111O----------------------
-----001000101101101-----------------------------------
0011000100010------------------------------------------
------101110100100101----------------------------------
-----------------------------001010001010101----------
-----------------------------1101011101010------------
---------------------------000010100010---------------
----------------------00000011110101------------------
-----------------------------------------01010001010---
-------------------------------------------------1110101011
-----------------------------------------010001110100
----------------------------------00010101010---------
1100111010101------------------------------------------
---------------001010000000-------------------------
---------------------1111110010101--------------------
-------------------------------------10101010111010----
----------01011011010---------------------------------
```

   (a) (10pts) Assuming the reads come from a diploid organism, determine the two haplotypes.

   Solution

   (b) (5pts) What is the MEC score of your proposed diploid haplotypes?

   Solution

   (c) (5pts) You should have noticed that in your solution, the two haplotypes are entirely complementary (if there is a 0 in haplotype A, then there is a 1 in haplotype B). Explain why this is the case.

   Solution

2. **[40 pts] Metagenomic read binning**

In this problem, we will be building a simple metagenomic read classifer from scratch.

(a) (10pts) Inside the data folder is 4 genome fasta files, each of which are 1,000,000 bp long. Using $k = 4$, report the 10 most common kmers for each genome.

Solution

(b) (10pts) The reads.fa file contains 1000 reads from the synthetic genomes. Using the most common kmers from each genome, design a classifier to determine the source genome of the reads. Describe your method below, and implement it with code. State the relative frequency of each genome in the reads file.

Solution

(c) (10pts) An alternative method for binning metagenomic reads is to match them to discriminative kmers. A kmer will be discriminative if it is present in only 1 of the reference genomes. Using k = 10, report the 5 most common discriminative kmers for each genome.

Solution

(d) (10pts) Similar to part b), design a new classifier, this time using discriminative kmers from each genome. Describe your method below, and then state the relative frequency of each genome in the reads file. (Hint: for this method, would recommend using the entire set of discriminative kmers for a given genome)

Solution

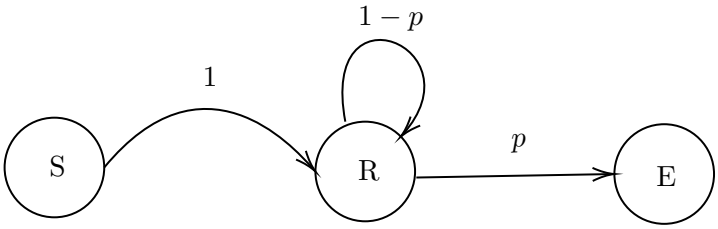

Figure 1: Random Genome HMM

3. **[40 points] Hidden Markov Models**

Warm-Up

(a) Consider the state transition diagram for the very simple HMM shown in Figure 3. The state S is a silent start state, and the state E is a silent end state. The state R, which stands for *random*, emits nucleotides with the following probabilities:

nucleotide	emission probability
A	0.15
C	0.35
G	0.35
T	0.15

The human genome is approximately 3.2 billion base pairs long. Suppose we generate a genome using the HMM in Figure 3. Find the value of p such that the expected length of this random genome is equal to the size of the human genome.

Solution

(b) Find the expected GC content of the genome generated in the previous part (write the answer as a percentage).

Solution

Supervised Learning with HMMs

(c) Consider the problem of supervised learning using HMMs. Specifically, we are given a set of n observation sequences $O^{(i)} = o_1^{(i)}, \ldots, o_{T_i}^{(i)}$ drawn from an alphabet set O, along with state annotation data $Q^{(i)} = q_1^{(i)}, \ldots, q_{T_i}^{(i)}, (i = 1, \ldots, n)$, from a set of possible states Q. Here T_i is the length of the i-th observation. Show that the maximum likelihood estimates for the HMM are

$$\hat{a}_{st} = \frac{\sum_{i=1}^{n} \sum_{j=2}^{T_i} \mathbb{I}\left\{q_{j-1}^{(i)} = s, q_j^{(i)} = t\right\}}{\sum_{i=1}^{n} \sum_{j=2}^{T_i} \sum_{t' \in Q} \mathbb{I}\left\{q_{j-1}^{(i)} = s, q_j^{(i)} = t'\right\}},$$

and

$$\hat{e}_s(b) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{T_i} \mathbb{I}\left\{q_j^{(i)} = s, o_j^{(i)} = b\right\}}{\sum_{i=1}^{n} \sum_{j=1}^{T_i} \sum_{b' \in O} \mathbb{I}\left\{q_j^{(i)} = s, o_j^{(i)} = b'\right\}}.$$

Here, $\mathbb{I}$ is the indicator function, which takes value 1 when its argument is true and 0 otherwise.

Solution

HMMs Application

CpG islands are DNA regions rich in cytosine (C) and guanine (G) nucleotides, commonly associated with gene promoter regions and regulation. Suppose you want to build a Hidden Markov Model (HMM) to recognize CpG islands within genomic DNA sequences.

(d) Design a simple Hidden Markov Model to model genomic sequences for CpG island detection. Using a graphical illustration of your HMM design and clearly define the hidden states and the possible observations.

Solution

(e) After training your HMM, you receive an unlabeled genomic DNA sequence. Describe two methods you could use to identify and locate CpG islands within this sequence. Briefly explain how each method works.

Solution

(f) Consider now that you have two separate biological conditions, e.g. healthy vs. diseased tissues, each potentially having different distributions of CpG islands. You wish to construct two HMMs, one per condition, from unlabeled sequences. What algorithm should you use to estimate the transition and emission probabilities for each model and why?

Solution

(g) Suppose you have trained the above 2 HMMs. You wish to determine if a new given unlabeled genomic sequence is more likely to originate from healthy or diseased tissue. What method would you use to do this?

Solution