Computational Genomics Midterm
02-510/02-710
Spring 2023

This exam has 7 questions, for a total of 100 points.

Name: ______________________________

Instructions:

- Write clearly. Unless stated otherwise, make sure to write down your steps for the calculations/derivations, do NOT just write a number.
- This exam is open book, open notes.
- You have 1 hour and 35 minutes to finish this exam.
- **DO NOT DISCUSS** the materials on this midterm (on Piazza or otherwise) until after the answer key for the midterm has been posted.
- **Note:** The questions vary in effort and difficulty. We recommend if you get stuck to move on and return to harder problems with remaining time.
- Good luck!

No.	Topic	Max. Score	Your Score
1	Sequence Alignment	20	
2	Suffix Tries	18	
3	Markov Models	23	
4	Differential Expression	15	
5	Population Structure	12	
6	Motif Discovery	12	
7	Extra Credit (Last Page)	10	

1. **[20 points] Sequence Alignment (Shane)**

   (a) (4 points) Define sequence alignment, discuss the drawbacks associated with its runtime, and explain a method to speed up sequence alignment in practice. One sentence for each part of the question should suffice.

   Solution

   Example answer: Sequence alignment establishes a distance metric between strings by employing a set of fundamental operations known as insertion, deletion, and substitution. Strings S_1 and S_2 are aligned by identifying the least-cost sequence of steps that transform S_1 into S_2, while adding gaps where a deletion or insertion has taken place to construct the alignment. Although this approach is theoretically valid, it encounters a practical limitation in terms of runtime, which is $\Theta(mn)$, where $m = |S_1|$ and $n = |S_2|$. Even when incorporating space-saving solutions (e.g., Hirschberg's algorithm), this runtime cost is prohibitive for genome-scale strings. As a result, numerous heuristics have been developed to tackle this challenge, one of which involves considering only a linear-sized subset of the dynamic programming cells in the matrix, such as a banded diagonal.

   (b) (4 points) Describe the differences between global and local alignment and provide a biological example of when each would be appropriate.

   Solution

   Example answer: Global alignment is designed to align sequences in their entirety, from start to end. It is particularly suitable for comparing sequences of similar length with relatively conserved regions throughout. On the other hand, local alignment focuses on identifying regions of high similarity within the sequences without considering the entire length of the sequences, which is appropriate when comparing sequences with only short regions of similarity or sequences of significantly different lengths.

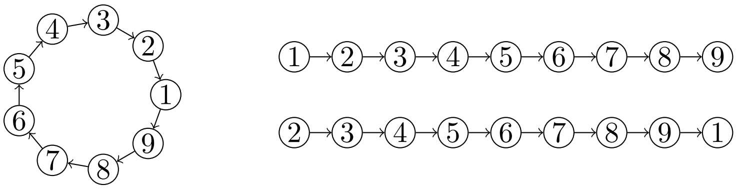
   Biological examples:

   - Global alignment would be appropriate for comparing orthologous genes or proteins from different species, which are expected to have a high degree of similarity across their entire length.
   - Local alignment is suitable when comparing two proteins or genes that share a conserved domain or motif but may have different functions and structures outside of that shared region.

(c) (7 points) Fill out the matrix below by running the Needleman-Wunsch algorithm on the strings ACTG and ATTGA with alignment costs MATCH = 1, MISMATCH = -1, and GAP = -2.

	–	A	T	T	G	A
–						
A						
C						
T						
G						

(d) (5 points) Suppose you need to align two circular bacterial genomes. You will soon face a challenge: circular strings do not have a distinct start or end, which means conventional alignment algorithms cannot be applied without adjustments. One straightforward approach is to convert a circular string into a linear string by designating a character as the starting point and the character before it as the endpoint. The illustration below demonstrates a circular string on the left and two potential linear strings on the right that can be created using this technique.



Considering this technique, describe and analyze an algorithm to compute the edit distance between two circular strings. Your algorithm should focus solely on determining the alignment cost and **does not** need to return the actual alignment.

Solution

Example algorithm: Denote the two circular strings by S and T and fix a reference node in each called S_1 and T_1. One of many ways to solve this problem is to notice that we only need to consider the starting point of one of the strings, not both. To see this, suppose the optimal alignment starts at nodes S_i and T_j with respect to the fixed reference node. We recover the same alignment by starting at S_1 and $T_{(j-i+1) \bmod n}$ because all moves in the optimal alignment may be performed. Letting $\mathcal{E}$ denote the edit distance, the preceding discussion guarantees that

$$\min_{k=1}^{n} \mathcal{E}(S, T[k:] + T[:k])$$

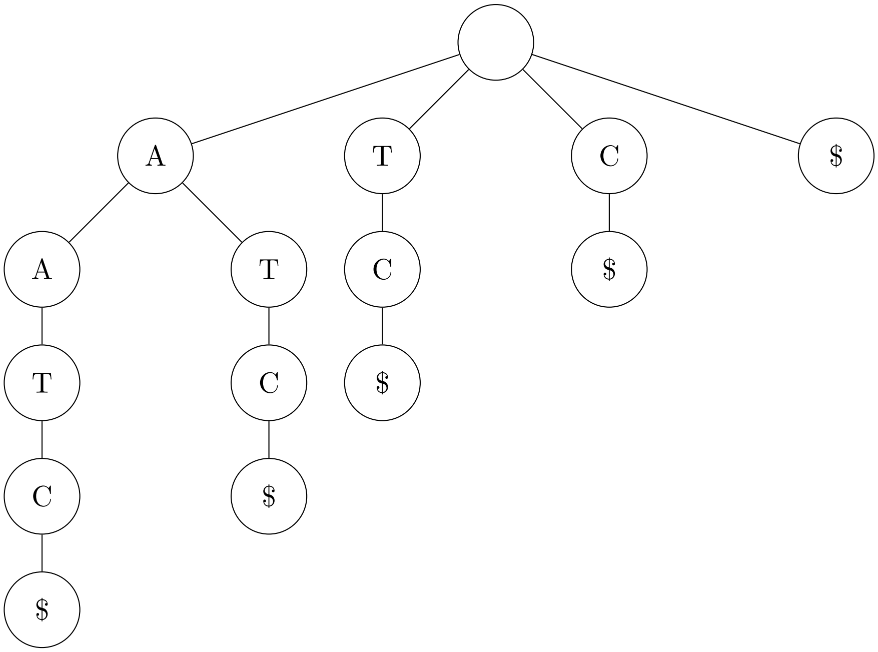
recovers the alignment cost. This requires calculating the edit distance between two strings of length m and n exactly n times, giving a runtime of $\Theta(mn^2)$.

2. **[18 points] Suffix Tries. (Shane)**

   (a) (7 points) Draw a suffix trie for the string AATC with the symbol $ appended to the end.

Solution

To Do: transition labels to edges instead of nodes



   (b) (6 points) Circle true or false to the following questions:

      i. True or False: A suffix trie is a data structure that stores all the suffixes of a given string, allowing for efficient pattern matching and string operations.

      ii. True or False: The construction of a suffix trie takes O(n) time complexity, where n is the length of the input string.

      iii. True or False: A suffix trie is a space-efficient data structure, requiring only O(n) space, where n is the length of the input string.

      iv. True or False: Compressed suffix tries, also known as suffix trees, combine identical branches and reduce the overall size of the trie by representing edges with substrings instead of single characters.

      v. True or False: Suffix tries can be used to solve various string processing tasks, such as finding the longest common substring, counting the number of distinct substrings, and searching for a pattern within a given string.

      vi. True or False: In a suffix trie, searching for a pattern of length m in a string of length n has a time complexity of O(m).

(c) (5 points) You have built a suffix trie (or tree) T from a string S on the alphabet $\Sigma = \{A, T, C, G\}$. Suppose you need to delete a character in the original string S at index k. Assuming you have access to the original string S, describe an algorithm that modifies T to account for the deleted character. You **do not** have to analyze runtime.

Solution

This is trivial when the string is of length 1. In that case, there is only one character to delete, so we are left with a single node for the character \$.

When the string is of length $n > 1$, we should perform a depth-first search on T and must carefully choose when to delete paths. When we encounter an edge with the deleted symbol, we recursively check if it belongs to a path of length $n - k$ starting from the current node and ending on the \$ character. If such a path exists, edges and nodes on the path are deleted while recursion terminates until a node with more than one child is encountered. Once such a node is encountered, we know that the path has been effectively eliminated from the trie.

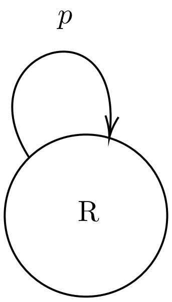


Figure 1: Random Genome HMM

3. **[23 points] Markov Models (Aditya)**

HMM Basics

Consider the state transition diagram for the very simple HMM shown in Figure 1. The state R, which stands for *random*, emits nucleotides with the following probabilities:

nucleotide	emission probability
A	0.20
C	0.30
G	0.30
T	0.20

(a) (1 points) What must be the value of p in Figure 1?

Solution

We must have $p = 1$.

(b) (8 points) Let $S = s_1, s_2, s_3, \ldots$ be the sequence of a genome generated by this HMM. Find the expected length of an open reading frame in the genome generated by this HMM.
Hint: The first three bases of the open-reading frame will be that of the start codon, ATG. The last three bases will form a stop codon, which can be any one of the following: {TGA, TAG, TAA}.

Solution

An open reading frame can be thought of as a sequence of codons beginning with a start codon, ATG, followed by any number of codons until we terminate with one of the stop codons; for example for an open-reading frame with 5 codons (including the stop codon):

ATG CCC GAC GAA TAA

Once we are in an open-reading frame (that is, once we have seen an ATG), each of the triplets following this initial start codon can be thought of as a success or failure: success if it is a stop codon (terminating the open-reading frame), and failure otherwise. The probability of success is the probability of observing a stop codon:

$$\theta = p(TAA)+p(TAG)+p(TGA) = (0.2)(0.2)(0.2)+(0.2)(0.2)(0.3)+(0.2)(0.3)(0.2) = 0.032.$$

So the probability of generating X codons following the initial start codon (and including the stop codon) is

$$P(X) = (1-\theta)^{X-1}\theta,$$

and so the expected number of codons generated after the initial start codon is

$$E(X) = \frac{1}{\theta}.$$

Since we need to account for the initial start codon as well, the expected number of codons in the open-reading frame is thus

$$E(X) + 1 = \frac{1+\theta}{\theta}.$$

Since each codon consists of 3 nucleotides, the expected length of the open-reading frame is

$$3 \cdot [E(X) + 1] = \frac{3+3\theta}{\theta} = \frac{3+0.096}{0.032} = \frac{3096}{32} \approx 96.$$

So on average we expect to see around 96 bases in an open-reading frame.
Grading note: It is ok if they do not count the bases in the stop codon as part of the length. Grading notes:

- +2 for recognizing we should split sequence into triplets
- +2 for calculating probability of stop codon
- +2 for correct calculation of probability of length of open reading frame (in terms of codons)
- +2 for correct calculation of expectation

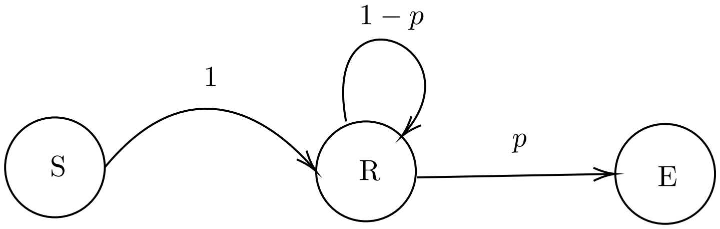


Figure 2: Random Genome HMM with End

Consider the familiar HMM in Figure 2. As before, the states S and E are silent start and end states. This time the emission probabilities of the nucleotides in state R are unknown:

nucleotide	emission probability
A	p_A
C	p_C
G	p_G
T	p_T

Suppose you observe the following dataset for $N = 3$ observations from this HMM:

i	$O^{(i)}$
1	ACAC
2	ACC
3	TTAGG

(c) (8 points) Find the maximum likelihood estimate of the parameters of this HMM. You should provide 5 estimates, as there are 5 parameters for this model.

Solution

Since there is only one state which emits any symbols, we know all observed emissions must come from state R. This makes the estimation of p_A, p_C, p_G, p_T easy:

$$\hat{p}_j = \frac{n_j}{\sum_k n_j},$$

for $j \in \{A, C, G, T\}$. Thus $\hat{p}_A = 1/3$, $\hat{p}_C = 1/3$, and $\hat{p}_G = \hat{p}_T = 1/6$. To estimate p, note that the lengths of the sequence have a geometric distribution, as shown in homework 3. Letting l_i be the length of sequence i, the likelihood is

$$\mathcal{L}_n(p|l_1, \ldots, l_N) = \prod_{i=1}^{N}(1-p)^{l_i - 1}p = p^N(1-p)^{\sum_{i=1}^{N} l_i - N}.$$

The mle is easy to solve for, and is

$$\hat{p} = \frac{N}{\sum_{i=1}^{N} l_i}.$$

Thus for this dataset,

$$\hat{p} = \frac{3}{12} = \frac{1}{4}.$$

Grading notes:

- +1.5 for correct calculation of $\hat{p}_i$ for each nucleotide
- +2 for correct calculation of $\hat{p}$

HMMs for Genome Annotation

Consider again the problem of crossing a werewolf with a vampire, resulting in our hybrid creature from homework 3, the “werepyre”. As in the homework, we want to use the werepyre genomic sequence to determine which regions of the werepyre genome originate from the werewolf genome, and which ones from the vampire genome, using HMMs.

Suppose we know that vampire genomes have a GC-content of 30%, and werewolf genomes have a GC-content of 60%. Additionally, the cytosines in vampire genomes are 5 times more likely to be methylated than in werewolf genomes. Therefore, this time we have both methylated and unmethylated C’s in our genomic sequence.

(d) (4 points) Design a Hidden Markov Model for the Werepyre genome and provide its state transition diagram. You can assume that the nucleotides (A, T) and (G, C) have equal frequency, and we have determined the methylation state of each C in the genome. (*Hint*: This is very similar to the problem in HW 3).

Solution

The state transition diagram is exactly the same as the one in HW 3, problem 1(e).
Emission probabilities: $\{e_s(b)\}_{s\in\{H,L\},b\in\{A,C^-,C^+,G,T\}}$, where C^- and C^+ are for unmethylated and methylated cytosines, respectively.
Grading Notes:

- -2 If missing state transition diagram
- -2 If missing emission probabilities
- -2 If not incorporating methylation information
- +1 extra-credit for (correct) numerical calculations

(e) (2 points) Which of the following experimental methods would we use to determine which cytosines are methylated?

- RNA-seq
- ChIP-seq
- Bisulfite sequencing
- Hi-C
- Expectation Minimization

Solution

Bisulfite Sequencing

4. **[15 points] Differential expression (Mike)**

Suppose you are studying the effects of eating poptarts on gene expression. Cases are given poptarts, and controls are given placebo. You measure gene expression changes between these groups for 10,000 genes. You also perform a Bonferroni correction in order to control for Type I error. Your uncorrected p-value is 1/100,000. Answer the following questions.

(a) (2 point) What is the corrected p-value?

Solution

0.1
1 point: showed work that was reasonable attempt but wrong answer. 2 points: correct answer

(b) (1 point) Suppose you found 20 genes whose expression levels are significantly different between the case and control groups using the procedure described above. Your colleague claims that the probability of at least one of these genes being a false positive is 0.2. Is this true or false?

Solution

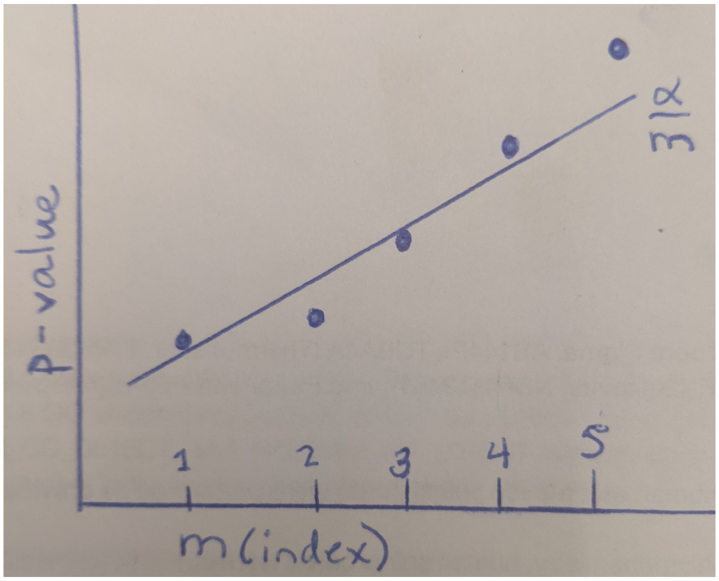
False

(c) (1 point) Several of the genes included in this study are known to have highly correlated expression across a range of conditions. Your colleague suggests using a false discovery rate method instead of Bonferroni correction. Based on this info, should you switch? Answer true or false.

Solution

True

(d) (3 points) Your colleague decided to control Type I error with a false discovery rate. They implemented the Benjamini-Hochberg procedure in order to find genes that are differentially expressed with statistical significance. The figure below shows p-values for the first five genes after sorting all genes using their p-values from lowest to highest (α is the false discovery rate, and m is the index of the p-values of your genes in the sorted list).



Indicate which of the p-value(s) in the figure are considered significant after correction with this procedure. Write the index or indices of the p-value(s) in the solution area below. If none are significant, write 'None'.

Solution

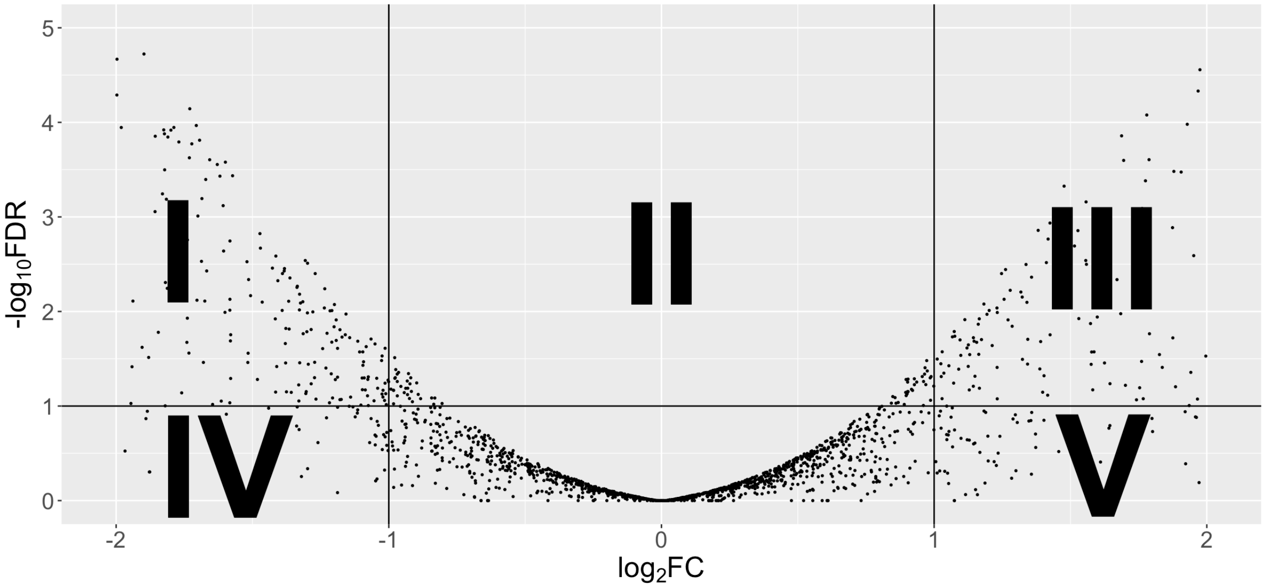
1,2,3
points: correct minus incorrect, unless they miss the first one, which would result in just not getting that point (+0).

(e) (3 points) Suppose that you performed an experiment where the false discovery rate is $\alpha = 0.1$. You found 20 genes that are significant out of 500 genes tested. What is the expected number of true positives out of those 20 genes?

Solution

18, since 0.1*20 would be false discoveries.
1 point: attempted with major conceptual mistake 2 points: minor mistake leading to wrong answer. 3 points: correct answer

(f) (3 points) Below is a volcano plot measuring the log2 fold change (FC) of genes for cases relative to controls, with a false discovery rate of 0.1. You plan to study significant genes with a minimum magnitude of fold change of 2.



Indicate which numerals correspond to which categories. Possible answers include multiple numerals or 'none'.

Solution

upregulated: III downregulated: I unchanged: II not significant: IV, V
points:
3 points for all correct, none incorrect. 2 points for one conceptual mistake 1 points if attempted with at least one in right spot, multiple mistakes.

5. **[12 points] Population Genetics (Mike)**

(a) (3 points) A given set of alleles {A,a} has a dominant allele A and a recessive allele a. Assume Hardy-Weinberg equilibrium in the population. The probability of allele A is 0.6, the probability of a is 0.4. What is the proportion (a number from 0 to 1) of individuals with at least one A allele? What about the proportion with one of each (A and a)? Two answers expected, show your work.

Solution

At least one A: $p^2 + 2pq = 0.6^2 + 2 * 0.4 * 0.6 = 0.84$
Aa or aA: $2pq = 2 * 0.4 * 0.6 = 0.48$
1 point: got both wrong, somewhat reasonable attempt. 2 points: work shown correctly, got one wrong. 3 points: correct answer

(b) (9 points) Consider genome sequences from two populations, one for werewolves and the other for vampires. Design a mixture of hidden Markov models to learn both haplotypes and population structures from genotype data $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^N]$, where N is the number of individuals and $\mathbf{x}^i = [x_1^i, \ldots, x_P^i]^T$ is the genotype data of the ith individual at P loci, each x_i^n taking values from $\{0, 1, 2\}$. Assume Q populations and M ancestral haplotypes. Hint: introduce latent random variables for unknown haplotypes and unknown population labels and give the model equation for $P(\mathbf{X})$.

Solution

We use the following hierarchical model:

$$\begin{aligned} Z^i &\sim \text{Multinoulli}(\theta_1, \ldots, \theta_Q) \\ \mathbf{x}^i | Z^i = z &\sim \text{HMM}(z). \end{aligned}$$

Here Z is a latent variable corresponding to the population of origin of the i-th observation, and HMM(z) is the haplotype HMM that generates $\mathbf{x}^i$ for the z-th population ($z = 1, \ldots, Q$). The hidden state of the HMM is the ancestral haplotype; the observation is the SNP allele. The parameters of HMM(z) are initial probabilities $\{p_s^z\}_{s \in \{1,\ldots,M\}}$, transition probabilities $\{a_{st}^z\}_{s,t \in \{1,\ldots,M\}}$, and emission probabilities $\{e_{s,j}^z(b)\}_{s \in \{1,\ldots,M\}, j \in \{1,\ldots,P\}, b \in \{0,1,2\}}$.
Given this model, we can write the model likelihood as

$$P(\mathbf{X}) = \prod_{i=1}^{N} \left(\sum_{z=1}^{Q} \theta_z p(\mathbf{x}^i | z) \right),$$

where

$$p(\mathbf{x}^i|z) = \sum_{\mathbf{h} \in \{1,\ldots,M\}^P} p(\mathbf{x}^i, \mathbf{h}^i | z) = \sum_{\mathbf{h} \in \{1,\ldots,M\}^P} p_{h_1^i}^z e_{h_1^i,1}^z(x_1^i) \prod_{j=2}^{P} a_{h_{j-1}^i, h_j^i}^z e_{h_j^i,j}^z(x_j^i)$$

is the probability of observing SNP sequence $\mathbf{x}^i$ assuming the i-th observation belongs to population z.

6. **[12 points] Motif Discovery (Aditya)**

Position Weight Matrices

A motif M can be defined with a Position Weight Matrix where each element $M(i, j)$ represents the probability of having a nucleotide i at position j. For example, in the PWM shown in Table 1, $M(A, 3) = 0.3$ indicates that the probability of "A" at position 3 is 0.3. b_i represents the background probability for nucleotide i, where $i \in \{A, C, T, G\}$. Now, given this PWM, answer the following questions:

	1	2	3	4
A	1/2	3/16	1/8	1/8
C	1/16	1/4	1/4	1/4
T	3/16	5/16	3/8	1/2
G	1/4	1/4	1/4	1/8

Table 1: PWM for motif M

Given the PWM in Table 1 and assume a background probability of p for A, T and a background distribution of $0.5 - p$ for G, C at all positions.

(a) (6 points) For which values of p is A more likely than B? Give your answer as a range.

A. CGCC

B. ACCT

Solution

For $\frac{2}{5} \leq p \leq \frac{2}{3}$. We require

$$\frac{\left(\frac{1}{16}\right)\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)}{\left(\frac{1}{2}-p\right)^4} \geq \frac{\left(\frac{1}{2}\right)\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(\frac{1}{2}\right)}{p^2\left(\frac{1}{2}-p\right)^2},$$

which happens when

$$\frac{1}{16}p^2 \geq \left(\frac{1}{2}-p\right)^2,$$

or equivalently

$$\frac{15}{16}p^2 - p + \frac{1}{4} \leq 0.$$

The roots of this quadratic equation are $p = 2/5$ and $p = 2/3$, and so the inequality is satisfied when $2/5 \leq p \leq 2/3$.

Grading notes:

- +4 for setting up inequality (first equation)
- +2 for each correct number in range

(b) (4 points) List two simplifying biological assumptions of the position-weight matrix model.

Solution

- No gaps (insertions/deletions); i.e., all motifs are of the same length
- Each position within the motif is independent of any other position

(c) (1 point) Recall from HW3 that M^{bg} is the PWM representing background probabilities of nucleotides. Suppose $M^{bg}_{A,1} = 0.25$, $M^{bg}_{C,1} = 0.3$, and $M^{bg}_{G,1} = 0.3$. What is $M^{bg}_{T,1}$?

Solution

0.15

(d) (1 point) What is $M^{bg}_{A,3}$?

Solution

0.25

Extra Credit!!!

In lecture, we saw an EM-based approach (MEME) to learn position-weight matrices when we do not know the locations of the motifs in a collection of sequences. As a reminder, in MEME, we input a set of sequences $Y = \{Y_1, \ldots, Y_N\}$ and an integer W. Each sequence $Y_i \in Y$ is broken up into its set of overlapping W-mers, and these are combined into a single dataset $X = \{X_1, \ldots, X_n\}$. Each $X_i \in X$, is a W-mer originating from sequence in Y. MEME uses the following mixture model for generating the dataset X:

$$X_i \sim \lambda_1\theta_1 + \lambda_2\theta_2,$$

which has the following parameters:

- θ_1: The PWM for the motif M
- θ_2: The PWM for the background model
- λ_1: The mixing proportion for θ_1
- λ_2: The mixing proportion for θ_2.

However, the most important quality of MEME is the following:

(e) (Optional - 10 points) What does "MEME" stand for? Wrong answers only.

Solution

Grading notes: -99 for "Multiple EM for Motif Elucidation".