02510/02710 - Midterm 2 - Spring 2024

Name: _____ Andrew ID: _____

Problem 1 [20 pts]. Haplotype Phasing with Hidden Markov Models
Assume we have 3 ancestral haplotypes with 8 variant blocks each.

Haplotype A: 1, 0, 0, 1, 0, 1, 1, 0
Haplotype B: 0, 1, 1, 1, 0, 0, 1, 1
Haplotype C: 0, 1, 0, 0, 1, 1, 0, 1

Assume that the transition probability for switching between haplotypes is 50% at each location, with an equal chance of swapping to any of the other haplotypes. Now you are given a mixed Sequence: 1, 0, 1, 1, 0, 1, 1, 0, and you need to answer the following questions. Assume the initial probability $\pi = [1, 0, 0]$.

(a) Calculate the probability that the specified sequence originated from the following haplotype assignment sequence: A, A, B, A, B, C, B, A.

(b) Determine the most likely source haplotype for each block in the given sequence (list all possibilities) and calculate the likelihood of this sequence(s).

(a) $P = \frac{1}{2} * (\frac{1}{4})^6 = \frac{1}{8192}$. (8pt)

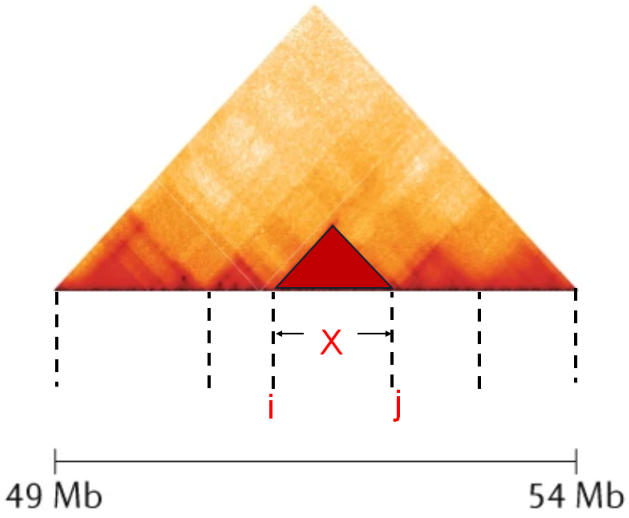
(b) For the transition probability:

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

For the emission probability:
For a given position, if the haplotype has 1 and the mixed sequence also has 1, the emission probability is 1, otherwise 0.
Perform Viterbi Algorithm and perform trace back, the most probable paths are listed below:
$AABAAAAA$ (3pt)
$AABBAAAA$ (3pt)
$AABBBAAA$ (3pt)
$P = \frac{1}{512}$ (3pt)

Problem 2 [20 pts]. 3D Genome

The spatial organization of the genome within the nucleus plays a critical role in gene regulation. Hi-C is a high-throughout technique to map chromatin contacts genome-wide at a length scale of hundreds of kilobases to a few megabases. Within the provided Hi-C contact map, you have identified specific patterns, referred to as X (a squared box or triangle in the Hi-C contact map), at a certain resolution on the Hi-C contact map.



(a) Based on what you have learned in the lectures on the 3D genome, can you identify what X represents?

(b) Suppose the Hi-C contact map has $n \times n$ entries, where each entry denotes the interacting frequency between the genomic loci i and j as $M(i, j)$. Design a dynamic programming strategy to find all non-overlapping square boxes like X that cover the diagonal of the contact map that maximizes the average contact frequency within those boxes. Write the dynamic programming recurrence with brief explanation for correctness.

(c) If you want to highlight key structural features within the Hi-C data by minimizing the number of detected X-like patterns, how would you adjust your dynamic programming approach? Hint: add a penalty.

- TAD (5pt)
- Let $f(i, j)$ represents the average contact frequency within a squared box that both row and column start at i and end at j. Suppose there are k square boxes represented by $(T_1, T_2, ..., T_k)$, and T_i has a starting index s_i and ending index e_i, then we want to maximize $\sum_{n=1}^{k} f(s_k, e_k)$. The objective function is $OPT(i) = max\{f(1, i), max_{1 \le k \le i}\{OPT(k) + f(k + 1, i)\}\}$. The base case is $OPT(1) = M[1, 1]$. (10pt)
- We can add a penalty score α that represents the penalty to introduce a new TAD. And the modified recurrence would be $OPT(i) = max\{f(1, i) - \alpha, max_{1 \le k \le i}\{OPT(k) + f(k + 1, i) - \alpha\}\}$ (5pt)

Problem 3 [20 pts]. Heritability

(a) A study measures the height of 100 pairs of MZ twins raised apart. The correlation coefficient for their heights is 0.8. Estimate the heritability of height based on the MZ twins' data and explain.

(b) A separate study measures the height of 100 pairs of DZ twins and 100 pairs of non-twin siblings. The correlation coefficient for the DZ twins' heights is 0.4, and for the siblings' heights, it is 0.3. Estimate the heritability of height using the DZ twins' data and explain.

(c) Estimate the heritability of height using the siblings' data and explain.

(d) Explain why heritability estimates might differ between DZ twins and non-twin siblings. Which one is more accurate for genetic heritability only?

5pts each

(a) Heritability: 0.8. Any differences in the heights of MZ twins would be due to environmental factors, as their genetic factors are identical.

(b) Heritability: $0.4 \times 2 = 0.8$. DZ twins share 50% of their segregating genetic variants. The expected correlation for traits fully determined by additive genetic variation would be 0.5. Therefore, to estimate the heritability from DZ twins, we can double the correlation coefficient

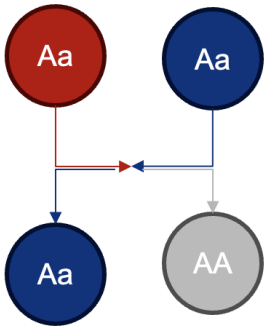
(c) Heritability: $0.3 \times 2 = 0.6$. Siblings share 50% of their segregating genetic variants. Similar to DZ twins.

(d) Accepted answers: Sampling Variability; Additionally, the DZ twins share more of their environmental effects (womb effects, etc.), and so the higher heritability might just be due to shared environment. Thus, assuming all else equal, the sibling studies are probably more accurate than the dizygotic twin studies for genetic heritability.

Problem 4 [20 pts]. Pedigree MLE single-locus linkage analysis
Given:

- A single biallelic locus AA with $p_A = 0.9$, $p_a = 0.1$.
- A trait that we assume is subject to autosomal dominant Mendelian inheritance with incomplete penetrance at some locus $K = 0.8$.
- Trait-state is positive (red), negative (blue) and unknown (grey).

You don't have to compute the precise final answer; instead, express it as the product of several probabilities.



(a) Calculate the probability of the pedigree under the hypothesis that this locus controls the trait.

$$p = 0.2 \cdot 0.18 \cdot 0.8 \cdot 0.18 \cdot 0.5 \cdot 0.2 \cdot 0.25 \cdot (0.2 + 0.8)$$

Probability of each node 2pt + 2pt if all probabilities is correct.

(b) Recompute assuming that trait independent of the locus.
Let $p(trait \mid AA) = p(trait \mid Aa) = p(trait \mid aa) = 0.5$.

$$p = 0.5 \cdot 0.18 \cdot 0.5 \cdot 0.18 \cdot 0.5 \cdot 0.5 \cdot 0.25 \cdot (0.5 + 0.5)$$

Probability of each node 2pt + 2pt if all probabilities is correct.

Problem 5 [20 pts]. Dimensionality Reduction

(a) Consider a scenario where you have a dataset containing gene expression levels of thousands of genes across multiple samples. Discuss how you would choose between t-SNE and PCA for dimensionality reduction and visualization. What factors would you consider, and how would the choice impact your ability to interpret the biological significance of the results?

[Total 8pts]

t-SNE is better for capturing local structure and revealing clusters, useful for exploring complex relationships. However, it can be computationally expensive and may not provide a clear global view. PCA, while simpler and more interpretable, assumes linear relationships and may overlook non-linear patterns. If you're interested in local relationships and hidden clusters, t-SNE may be more suitable. If you need a global overview and dimensionality reduction, PCA might be a better choice.

2pt each for the 8 possible things below, up to a total of 8pts (i.e. student must give at least 4 of the following traits to get full marks). If student gets other traits that are correct, each is worth 2pt.

t-SNE = local structure, clusters, complex relationships, visualization only

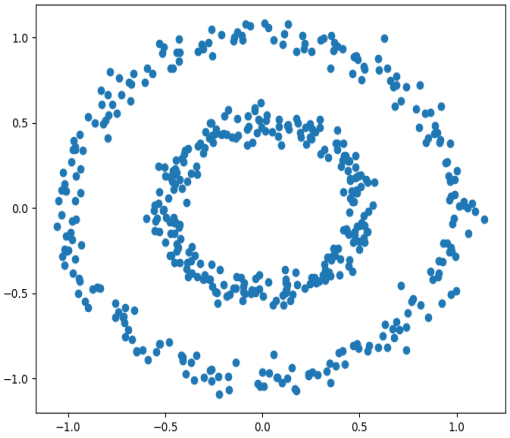
PCA = linear relationships, global overview, interpretable axes, reduced features

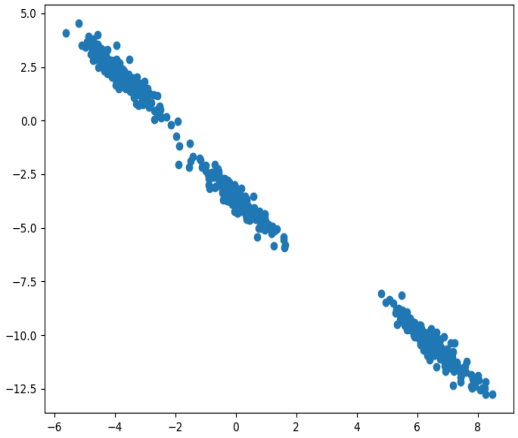
(b) Among PCA, t-SNE, and MDS, which technique is deterministic? Explain how this deterministic nature affects the stability of the results.

PCA is the deterministic technique among PCA, t-SNE, and MDS. Its results are stable and reproducible, making it easier to interpret and compare across analyses. In contrast, t-SNE and MDS are non-deterministic, leading to slightly different results each time they are run on the same dataset.

3pt: PCA is deterministic 3pt: determinism makes it easier to compare across runs

(c) Here are two 2D datasets. Draw the first and second principal components on each plot, and label which vector represents the first principal component and which represents the second.





Each PC is worth 1.5pt.