Some things to keep in mind:

- You can use whatever resources online, just please don't post the questions. If you do, I will find out and haunt your nightmares with really obscure probability questions for the rest of your life.
- Please don't work together. This assignment is meant to be completed on your own. If you do, I will find out and haunt your nightmares with really obscure biology questions for the rest of your life.
- I will be available for clarifying questions in person and on Zoom. Please keep an eye out for the schedule on Piazza.
- You must complete all problems for a chance at full credit.
- Finally, I don't expect this to take any more time than a normal final. If you've been keeping up with the assignments, it should be completable within a 3 hour window.

Good luck.

Take care and have a great summer. Congratulations to those of you graduating. Hope to see you around in the future.

Problem 1

t-SNE

[30 points]

(a) In t-SNE we normally use a different high-dimensional distribution (P) than the low-dimensional one (Q). Consider the case where we define p_{ij} using a t-distribution:

$$p_{ij} = \frac{(1 + ||x_i - x_j||^2)^{-1}}{\sum_{k=1}^{N} \sum_{l \neq k} (1 + ||x_k - x_l||^2)^{-1}}.$$

In this case, the distribution of P and Q are technically the same, though the dimensionality of x_i and y_i may be different. Now, assume that $y_i = x_i$ for every i. Show that in this special case that the KL-divergence is exactly 0.

Hint: You can start writing down an arbitrary (i, j) pair and their contribution to the KLD ($p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right)$). Expand that out and some things will cancel. You will then need to argue that the remaining components are equal to each other. Once you show it for an arbitrary pair, the entire KLD follows.

(b) Consider we have found two "solutions" to standard t-SNE: $\mathbf{y} = (y_1, \ldots, y_N)$ and $\mathbf{z} = (z_1, \ldots, z_N)$. In this case we know $z_i = y_i - c$ for any vector c with the appropriate dimension. Show that under this special case, the KL-divergence is exactly the same when Q is built from y or Q is built from z. What's the take away from this result?

Hint: there is no need to expand the terms of p_{ij}. You can simply assume p_{ij} is the same in both cases. This isn't as ugly as it sounds, I promise. Just write out the KL-divergence for some (i, j) pair with the known facts, stare a little, and the result should follow.

(c) Let there be three points in high-dimensional space: x_1, x_2, x_3. Points x_1 and x_2 are "close" to eachother, as $||x_1 - x_2||^2 = \alpha$ and α is "small". Points x_1 and x_3 are "far" from eachother, as $||x_1 - x_3||^2 = \beta = 100\alpha$. Show that changing values of σ_1^2 under the standard definition of $p_{3|1}$ can result in relatively large $p_{3|1}$ versus effectively zero.

What are the consequences of this result in t-SNE? In particular, we are looking for a connection to p_{ij} and the interpretation of p_{ij}. In words is sufficient.

Hint: write out the full definition of $p_{3|1}$. Substitute all the known facts of this problem. You will get values that depend on datapoint x_2. Since there are only 2 possibilities for the PMF when you condition on 1, $p_{2|1}$ must go up when $p_{3|1}$ goes down.

Problem 2

Pseudoalignment

[30 points]
Consider these isoforms:

$$
\begin{aligned}
t_1 &= ACCGGTATC \\
t_2 &= ACCCCTATG \\
t_3 &= \quad GGTAGCCT \\
t_4 &= \quad GGTATCCCG
\end{aligned}
$$

In this problem, reads can be generated from both the forward and reverse strand.

(a) Draw a transcript de Bruijn graph with $k = 3$. You can get full credit simply drawing the forward strand graph. You can pick colors if you'd like, otherwise, annotate the equivalence classes with labels 1, 2, ..., 4.

(b) Pseudoalign (with skipping) the read $TATCCCG$. Number the nodes which you visit.

(c) Pseudoalign (with skipping) the read $CGGGATC$. Number the nodes which you visit.

(d) Add the following isoform, t_5 to your graph: $CGGGATACC$.

(e) Pseudoalign (with skipping) the read $GGTATC$. Number the nodes which you visit.

Problem 3

Generative models

[30 points]

Consider the following experimental protocol for single-cell RNA-seq:

1. Single-cells are run through a machine that can identify the cell cycle state ($G1, S, G2,$ and M). Given the cell state, this machine can label the cell with a specific barcode for the corresponding state.

2. Once the cell state has been labeled and integrated into the cell, these cells go through the standard single-cell protocol we discussed in class. for example, there is cell barcode labeling by droplet creation, and *no perturbation*.

3. Some additional wrinkles:

   - At the end of the process, cells in state M die randomly with probability δ.
   - When a cell dies, we assume the observation of that cell goes to zero.

4. Some additional necessary details you will need:

   - You can assume that for each cell, the protocol only samples the expression of one gene. Thus, your entire set of observations from the entire experiment will be a vector of length C cells.
   - You need not specify the distribution of the gene expression when the cell is alive, but you must specify the distribution when it is dead, and you must specify the conditional dependencies of gene expression.

And now, the problem:

(a) Like we did in class for the functional genomics module, describe in words/steps the generative model for each cell. Make sure your assumptions are clear about which steps are important and which dependencies are important.

(b) Write down the generative model with random variables and the conditional dependencies.

(c) Draw a plate model of the process.