Krish Patel 605 796 227

Final C~~S~~1121

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

Consider q_{ij}, ~~where~~ ~~$y_i = x_i$~~ : $$\frac{(1+\|y_i - y_j\|^2)^{-1}}{\sum_{k=1}^{N} \sum_{l \neq k} (1+\|y_k - y_l\|^2)^{-1}}$$

if $y_i = x_i$ based on the assumption, $P_{ij} = Q_{ij}$

~~For~~ Thus KL Divergence $= \sum_{i \neq j} P_{ij} \log \underset{\substack{\Downarrow \\ 1}}{\left(\frac{P_{ij}}{q_{ij}}\right)} = \sum_{i \neq j} P_{ij}\,(0)$ $\boxed{=0}$

b) Doing the substitution: $q_{ij} = P_{ij}$

$$KL(P \| Q) = \sum_{k=1}^{N} \sum_{j \neq j} P_{ij} \log\left(\frac{P_{ij}}{Q_{ij}}\right) \qquad z_i = y_i - c$$

$$q_{ij}(y) = \frac{(1+\|y_i - y_j\|^2)^{-1}}{\sum_{k=1}^{N} \sum_{l \neq k} (1+\|y_k - y_l\|^2)^{-1}} \qquad q_{ij}(z) = \frac{(1+\|z_i - z_j\|^2)^{-1}}{\sum_{k=1}^{N} \sum_{l \neq k} (1+\|y_k - y_l\|^2)^{-1}}$$

$$z_i - z_j = (y_i - c) - (y_i - c) = y_i - y_j = z_i - z_j$$

$$z_k - z_l = (y_k - c) - (y_l - c) = y_k - y_l = z_k - z_l.$$

Thus, both distribution are equal. and hence KL Diveryerge $KL(P \| Q(y)) = KL(P \| Q(z))$

~~According~~ This illustrates that shifting a fixed set with lower dimension space Q, doesnt alter the overal relationship and mapping between higher & and lower dimesional space. Thus, relationship depends on the relative positions and distances between points, unaffected by such translations.

c) Using formula. $\|x_1 - x_2\|^2 = \alpha$ $\|x_1 - x_3\|^2 = \beta = 100\alpha$

$$p_{3|1} = \frac{\cancel{e} \; e^{-\beta/2\sigma_1^2}}{e^{-\alpha/2\sigma_1^2} + e^{-\beta/2\sigma_1^2}} = \frac{e^{-50\alpha/\sigma_1^2}}{e^{-\alpha/2\sigma_1^2} + e^{-50\alpha/\sigma_1^2}}$$

Limits $\sigma_1^2 \to \infty$ and $\sigma_1^2 \to 0$.

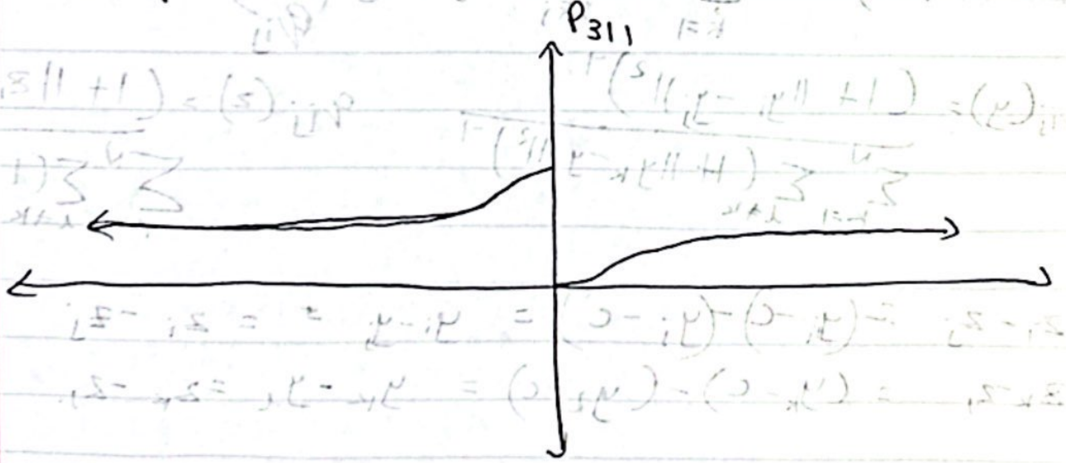
Both terms in denom approach 1 as $e^0 = 1$

$p_{3|1} = \frac{1}{2}$

The term $e^{-50\alpha/\sigma_1^2}$ approaches 0 much faster than $e^{-\alpha/2\sigma_1^2}$

Therefore, $p_{3|1} \approx 0$

Thus, graph looks like



Implications:

When σ_1^2 is large, model considers both nearby and distant points equally, leading to less distinction between x_2 and x_3 relative to x_1. When σ_1^2 is small, distinction becomes sharp, effectively ignoring distant points. x_3 relative to x_1 and x_1 preserving more local structure. Changing σ_1^2 impacts relative influence of distant points in t-SNE. greater σ_1^2 leads to a more global perspective, while smaller σ_1^2 values emphasizes local relationsh

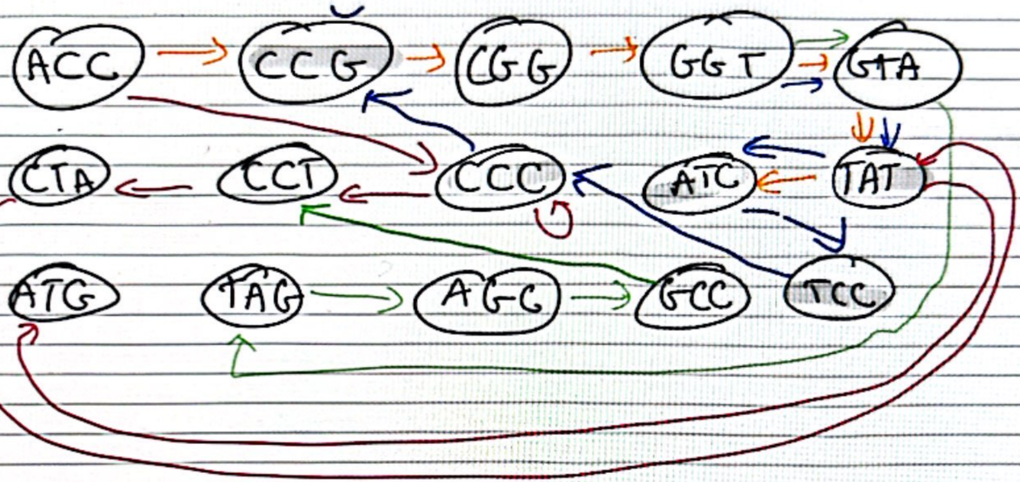
Q2)

t1 ACC CCG CGG GGT GTA TAT ATC

t2 ACC CCC CCC CCT CTA TAT ATG

t3 GGT GTA TAG AGC GCC CCT

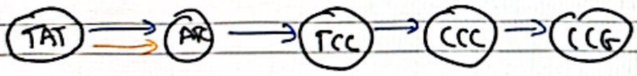
t4 GGT GTA TAT ATC TCC CCC CCG



b) Pseudoalign with skipping

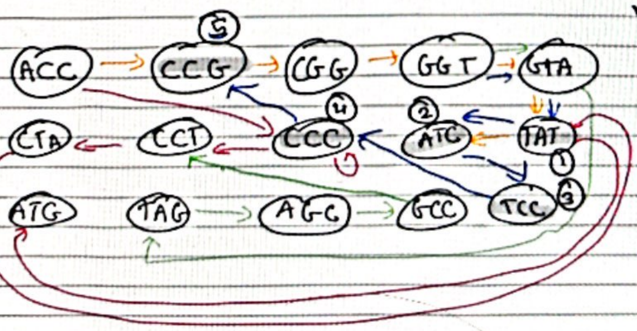
TATCCCG → TA

TAT → ATC → TCC → CCC → CCG



forward complement.

Find TAT. Find ATC we then find connections between two kmer nodes we do this and note down all



Completely aligns with true t4

c) Pseudoalign with skipping

CGGGATC

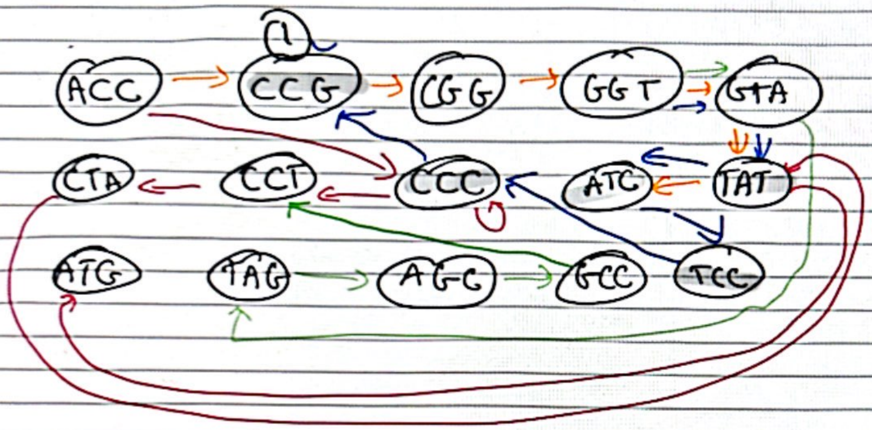
CGG → GGG → GGA → GAT → ATC

c) Pseudoalign with skipping

CGGGATC

CGG → GGG → GGA → GAT → ATC

GGG Not found. GGA Not found. and GAT not found. Thus, we can't pseudoalign the sequence because no match is found from the first kmer to the last kmer (CGG → ATC).

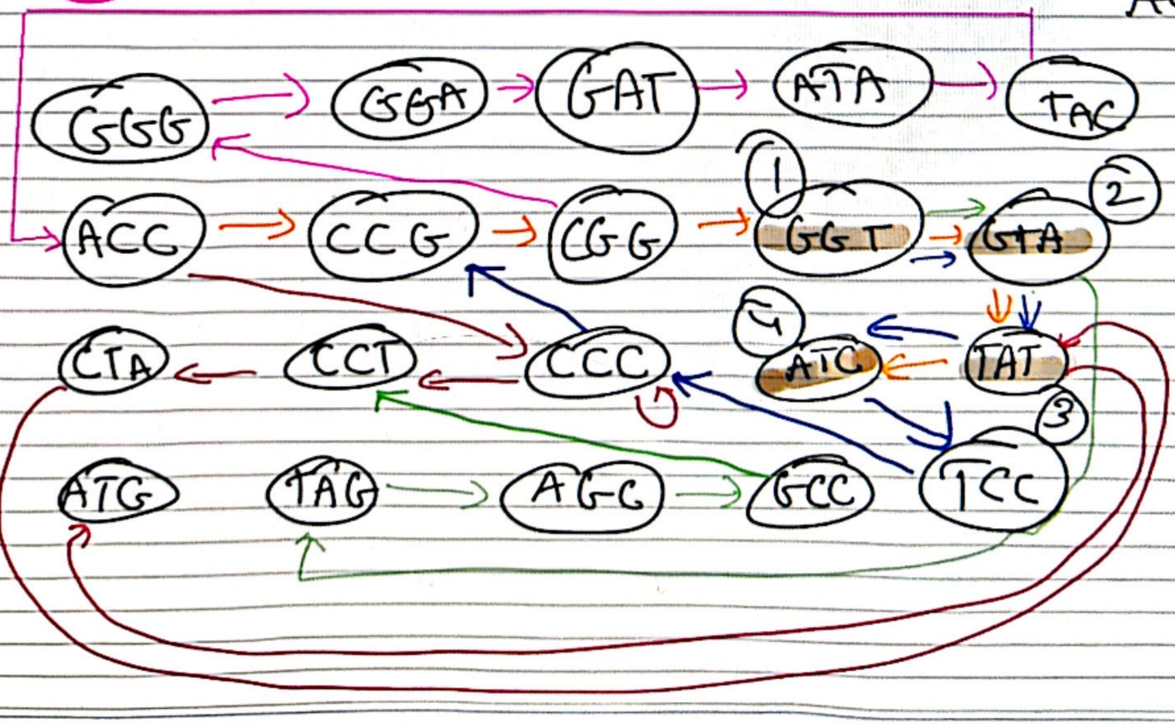


first we find CCG. We find GTA which doesn't exist in the graph. ~~we~~ After finding nullset, we terminate ~~[illegible]~~ alignment.

Reverse complement → GATCCCG → GAT Not found, thus terminate.

Union of forward and reverse is thus the null set

d) t5 CGG GGG GGA GAT ATA ~~TAC~~ ACC



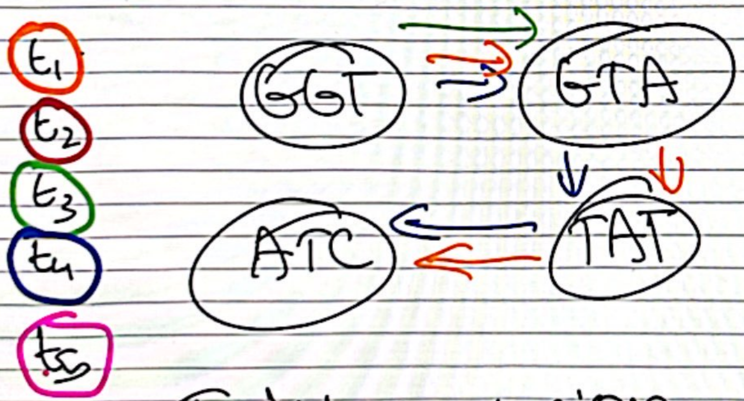
e) Pseudoalign with skipping:

GGT → GTA → TAT → ATC

from the graph above, we first visit GGT. ~~[illegible]~~, then, we visit GTA which aligns with t_3, t_1 and t_4. we then visit TAT, which gives a match with t_1 and t_4. Same with ATC.

Nodes visited

$1 \to 2 \to 3 \to 4$



Taking union

We get ~~read~~ = {🟦 🟧} for the entire length of the read

Q3a) → First , we choose the cell state. (G_1, S, G_2, m)

→ After this step. the machine labels the cell with the specific barcode corresponding to its shape.

→ labelled cells are then integrated in the barcode via droplet creation. No perturbations means ~~that~~ the cell's natural state is preserved throughout.

→ At the end of the process, cells in M state die with δ probability. Observed gene expression of that ~~cell~~ droplet goes to zero.

→ For each cell, ~~the~~ only the expression of 1 gene is sampled. Complete dataset consists of a vector of length C, where each entry represents the expression level of a single gene in 1 cell.

3b) Random Voriables to consider.

→ Cell state → $\overset{S_c}{\cancel{E_S}} \sim$ Categorical (G_1, S, G_2, M)

→ Cell label → $B_c \sim$ (Categorical (P')
↓
distribution of Barcode Lables.

→ ~~As cell~~ Cell death:
cell only dies in state m. ↗ ~~Stan~~ for all cells death rate is same

$P(\text{dead cell} \mid M) = \delta$

(~~cells don't~~ die $P(\text{dead cell} \mid (G_1, S, G_2) = 0$

$(i = 0 \to c)$

→ ~~Observation~~. $P(D_i) = \begin{cases} 1 - \delta^{\varepsilon} \times (A) & D_c = D(\text{Alive}) \\ \delta(A) & D_c = 1(\text{dead}) \end{cases}$ where A is # cells in state $\underline{m}$

Observation:

→ ~~Gene expression~~ → $D_i = \begin{cases} 0 & \text{if cell } i \text{ is alive} \\ 1 & \text{if cell } i \text{ is dead} \end{cases}$

$O_{GE} \sim$ ~~[illegible]~~ 2δ

~~For~~ Gene expression

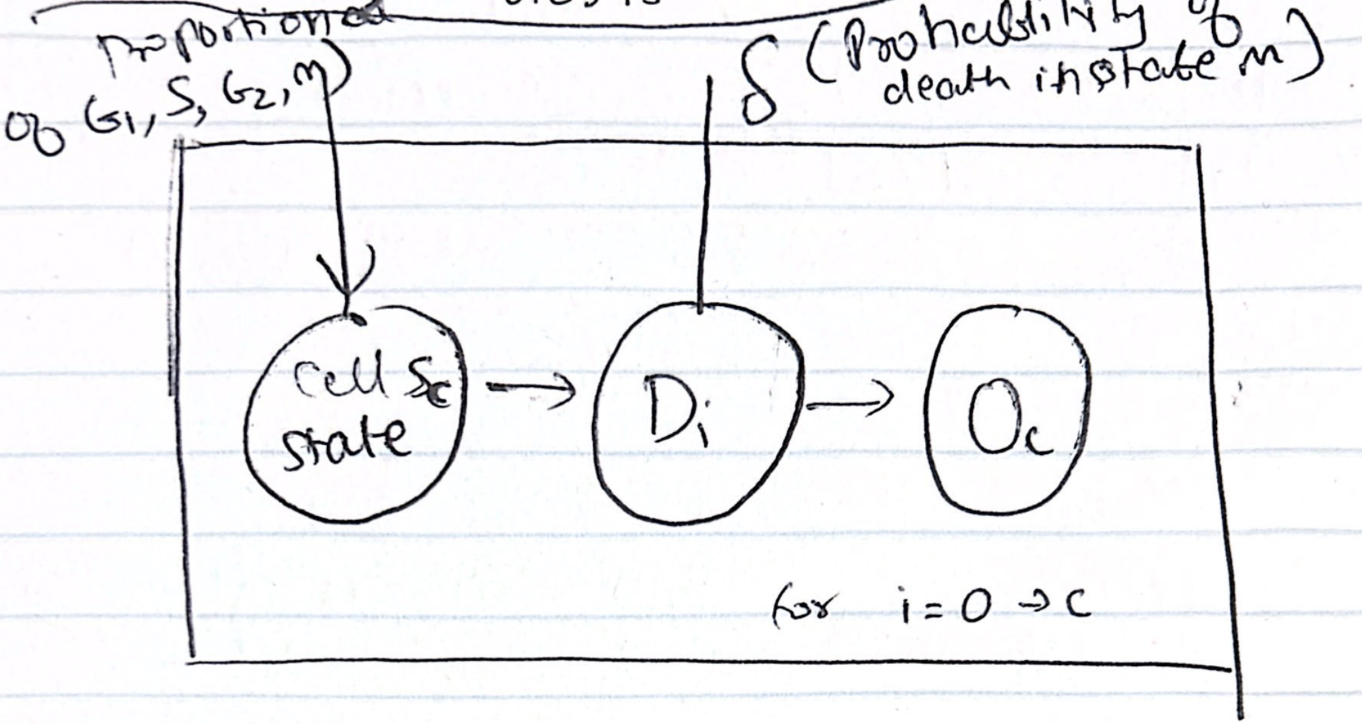
For $D_i = 1$ → Gene expression is zero

Else, $D_i = 0$ → ~~gene expression~~

$O_c =$ Distribution of $\left(\overset{\text{Expected}}{\text{Expression}} \mid S_c\right) \cdot \boxed{F}$ ↗ cells that are alive

$O_1 = 0$ if cell is dead.

$$= \frac{0.8(0.01)}{0.0575} \simeq 0.1391$$


proportioned
of G1, S, G2, M)
δ (Probability of death in state m)
cell S_c state
D_i
O_c
for i = 0 → c