BI/BE/CS 183 Set 1

Chris Pukszta

Problem 1

Single-cell RNA sequencing (single-cell RNA-seq or scRNA-seq) technologies utilize numerous recent technology breakthroughs. While many popular methods utilize similar ideas, such as barcoding transcripts to inform on their cell of origin, there are many ways to achieve this as discussed in Lecture.

a) The Unique molecular identifier (UMI) is a barcode on the reverse transcriptase primer that is unique to that primer set of primers. This allows for the cDNA from individual cells to be separated from one another because all pieces of cDNA from a cell will have the same UMI.

b) the 'library' produces by ss methods refers to the group of complimentary DNA strands (cDNA) that result from the single-cell method. This library can then be manipulated, or most commonly, sequenced to determine the RNA in the molecule.

c) the Indrops technology exhibits sub-Poisson loading of cells per droplet. Sub-Poisson means that the loading of cells into droplets is not a Poisson process (a series of discrete events with an average time/distance between events, but no way to know the exact timing of each event). In this case, the known time/distance is one cell per droplet.

d) 3' capture refers to the sequencing of DNA from the 3' end.

e) Order for 10X sequencing:
1) Cell capture and lysis
2) Addition of sample index/label (sample index PCR)
3) Reverse transcription and amplification
4) cDNA fragmentation and size selection
5) Single- or paired-end sequencing

Problem 2

a) Cell 3 has the highest expression level of gene 2

b) Gene 3 is most highly expressed in cell 2

c) i) The rank of G is 2.

This can easily be seen by examining the columns of the matrix. It is easy to see that:

$3 * \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 9 \\ 3 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix} + \begin{pmatrix} 2 \\ 1 \\ 4 \end{pmatrix} = \begin{pmatrix} 3 \\ 4 \\ 5 \end{pmatrix}$.

Therefore, because two of the vectors can be made as a liner combination of other vectors in the matrix, there are only two linearly independent vectors and therefore the rank of G is 2.

ii) The rank of G suggest that at minimum there are 2 independent drivers of gene expression in this set of cells.

D) i) let $v = \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix}$

$v^T G = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \begin{pmatrix} 1, 2, 3, 3 \\ 3, 1, 9, 4 \\ 1, 4, 3, 5 \end{pmatrix} = \begin{pmatrix} \frac{5}{3} \\ \frac{7}{3} \\ 5 \\ 4 \end{pmatrix}$

ii) $\begin{pmatrix} \frac{5}{3} \\ \frac{7}{3} \\ 5 \\ 4 \end{pmatrix}$

iii) Genes 3 and 4 have the highest mean expression. Therefore, we want to only retain these genes and wipe out everything else.

let $P = \begin{pmatrix} 0, 0 \\ 0, 0 \\ 1, 0 \\ 0, 1 \end{pmatrix}$

Then $GP = \begin{pmatrix} 3, 3 \\ 9, 4 \\ 3, 5 \end{pmatrix}$

E)

i) Computed from a python notebook:

$D_{L1} = \begin{pmatrix} 0, 10, 4 \\ 10, 0, 12 \\ 4, 12, 0 \end{pmatrix}$

$D_{L2} = \begin{pmatrix} 0, 6.48, 2.828 \\ 6.48, 0, 7.07 \\ 2.828, 7.07, 0 \end{pmatrix}$

$D_c = \begin{pmatrix} 1, 0.8869, 0.9635 \\ 0.8869, 1, 0.731 \\ 0.9635, 0.731, 1 \end{pmatrix}$

ii) For $L1$ distance, cells 1 and 3 are the closest

For $L2$ distance, cells 1 and 3 are also the closest

For cosine distance, cells 2 and 3 are the closest

F) Only the cosine similarity will be affected by the contamination. Let the level of contamination be n such that every entry to the gene expression matrix is now:

$x_{ij} + n$

In general for some vector x and y with components x_i and y_i:

$(x_i + n) - (y_i + n) = x_i + y_i$ meaning that the both the L_1 and L_2 distances are not affected as shown below:

$L_1(x + n, y + n) = \sum_{i=1}^{n} |(x_i + n) - (y_i + n)| = \sum_{i=1}^{n} |x_i - y_i| = L_1(x, y)$

$L_2(x + n, y + n) = \sqrt{\sum_{i=1}^{n} ((x_i + n) - (y_i + n))^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} = L_2(x, y)$

For the cosine similarity contamination does affect the result. This can be seen in a counter-example.

let $x = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ and $y = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix}$ with the level of contamination being 2.

Thus contaminated matricies x_c and y_c are:

$x_c = \begin{pmatrix} 3 \\ 4 \\ 5 \end{pmatrix}$ and $y_c = \begin{pmatrix} 6 \\ 7 \\ 8 \end{pmatrix}$

Computing the cosine similarities for these two sets of vectors gives:

$c(x, y) = 0.9746318461970762$

wheres as

$c(x_c, y_c) = 0.9963692477452066$

Because these two values are different from one another, the cosine similarity is affected by contamination.

G) The cosine similarity satisfies the property of scaling because the communicative property of multiplication applies to both the norm and dot product. Therefore:

$$c(ax, by) = \frac{ax * by}{||ax||_2||by||_2} = \frac{ab(x * y)}{ab(||x||_2||y||_2)} = \frac{x * y}{||x||_2||y||_2} = c(x, y)$$

For the L_1 and L_2 distances, however, the scaling property does not hold because $ax_i - by_i = x_i - y_i$ is true only for certain values of x_i, y_i, a, b as shown below

$ax_i - by_i = x_i - y_i$

$x_i(a - 1) = y_i(b - 1)$

$\frac{x_i}{y_i} = \frac{(b-1)}{(a-1)}$

iff the above ratio holds then $ax_i - by_i = x_i - y_i$. However for other values of x_i, y_i, a, b, $ax_i - by_i \neq x_i - y_i$. Therefore, L_1 and L_2 distances are not invariant to scaling.

$L_1(ax, by) = \sum_{i=1}^{n} |ax_i - by_i| \neq \sum_{i=1}^{n} |x_i - y_i| = L_1(x, y)$

$L_2(ax, by) = \sqrt{\sum_{i=1}^{n} (ax_i - by_i)^2} \neq \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} = L_2(x, y)$

3) Clickable link to Colab notebook:
https://colab.research.google.com/github/CPukszta/BI-BE-CS-183-2023/blob/main/HW1/Problem3.ipynb