BMEG3105 Data Analytics for Personalized Genomics and Precision Medicine

Lecture 18: Data Visualization and Protein-RNA/DNA
Lecturer: Yu LI (李煜)
Scriber: Kenny TANJAYA (1155198328)

18.0.1 - Review:

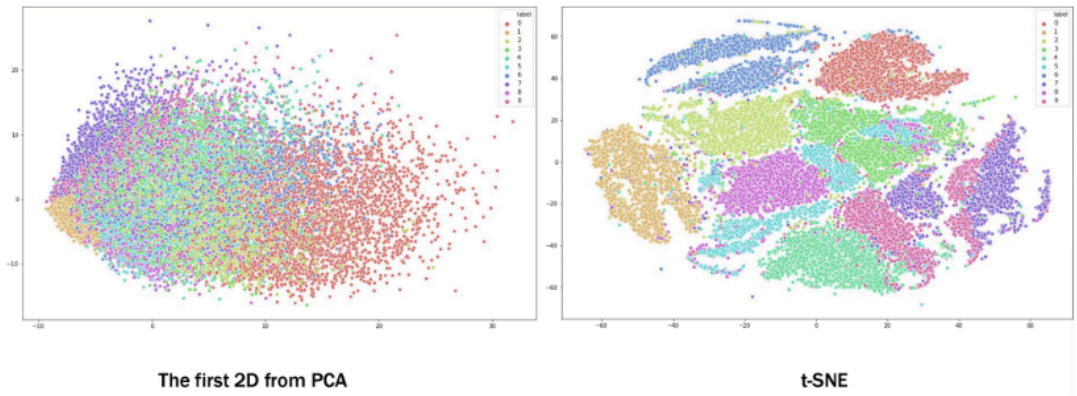
- Single-cell analysis examines the sequence information of individual cells separately, leading to a very sparse gene expression matrix. We can normalize the output data to counts per million to account for this sparsity. There are also some problems that arise from the down-stream analysis of single cells, such as noise, doublet, dropout and the batch effect.
- As long as you understand the basic idea, you should be good.

18.0.2 - Learning outcomes:

- Visualization -> t-SNE [**18.1**]
- Protein-RNA/DNA Interaction -> Motif Analysis [**18.2**]

18.1 - Visualization

- We want to **map gene expression into a 2D space**, whilst **preserving the clustering** purely for visualization, without any regards to distance perseverance or dimensional reduction.
- By projecting the data to the direction with the highest variance using PCA, we can preserve as much data as possible. However, this method is not perfect. The original clusters are not preserved, which tends to be more problematic in higher spaces.
- **T-Distributed Stochastic Neighbour Encoding (t-SNE)** is a non-linear dimensionality reduction technique used for embedding high-dimensional data for visualization in a low-dimensional space of 2 or 3 dimensions. It does this by modelling similar objects with nearby points and dissimilar objects with distant points with high probability iteratively.
- **Process:** Random initiation, Slightly updating the position for each point, Repeating until there are no more updates to be made.
- The updating step is done by comparing the cluster to the original cluster. The points from the same cluster attract each other while those from different clusters are pushed apart from each other.
- Unlike PCA, t-SNE is irreversible. However, the resulting map also vastly differs.



The first 2D from PCA t-SNE

- Some **disadvantages** of t-SNE include having a long running time for larger datasets since it is an iterative process, being non-deterministic as the initialization seed is randomized, having noisy patterns, and not precisely preserving the original distance.
- **UMAP** can also be an alternative to t-SNE.

18.2 - Protein-RNA/DNA Interaction

- Proteins have different preferences for binding sites, having binding motifs and patterns that we can look for by biotechnological experimentation followed by computer analysis.
- By experimentation, we can find these protein binding motifs. We start by using an Epitope-tagged RBP mixed with a random RNA pool of 30-41 nucleotides, going through a GST pulldown assay and microarray hybridization to get data to be analyzed.
- A **motif** is a kind of sequence pattern that appears many times.
- **Process:** After getting our sequence, we need to first align the sequence as the motif will be smaller than the sequence itself. Then, we count how many times each nucleotide appears for each position, giving us a position count matrix which then gives us the position probability matrix, directly giving us the motif.
- **What if our sequences are not aligned?** There would be a 0.25 probability for each nucleotide for each position. So, the sequence alignment step is very important.

Position	1	2	3	4	5	6
A	1.00	0.67	0.00	0.83	0.83	0.66
C	0.00	0.00	0.33	0.00	0.00	0.00
G	0.00	0.00	0.50	0.00	0.00	0.00
T	0.00	0.33	0.17	0.17	0.17	0.33

18.3 - Take-Home Message

- We will learn about Deep Learning, Deep Neural Networks, Medical Imaging and Convolutional Neural Networks in our next lecture.