Lecture 18: Single cell RNA sequencing (short)

Outline from professor:

1. Protein-RNA/DNA interaction
    i. Motif analysis
2. Artificial intelligence VS Machine learning VS Deep learning
3. Deep neural networks

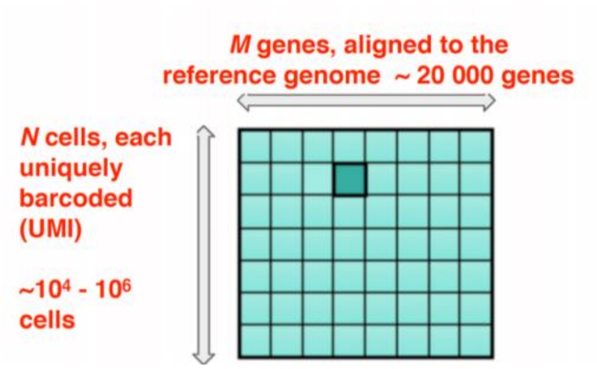
Lecture Progress:

<u>Single-cell data analytics (Remind from last lecture)</u>

1. Challenges of single-cell data analytics

- Noise: (How to denoise?...
- Doublet: Not perfect(In cell-isolation process), especially the data Needs to remove duplicates
- Dropout: About missing value in the data matrix analysis
- Batch effect: Artefacts from different experiments Wet lab: Different results by different students. Difference induced by different environment

2. Gene expression matrix:


M genes, aligned to the reference genome ~ 20 000 genes
N cells, each uniquely barcoded (UMI)
~10^4 - 10^6 cells

CPM(counts per million) = 106 * Xi / N

Since the counts in each cell is too small, we need to time a 1 million.

$$\mathrm{CPM}_i = \frac{X_i}{\frac{N}{10^6}} = \frac{X_i}{N} \cdot 10^6$$

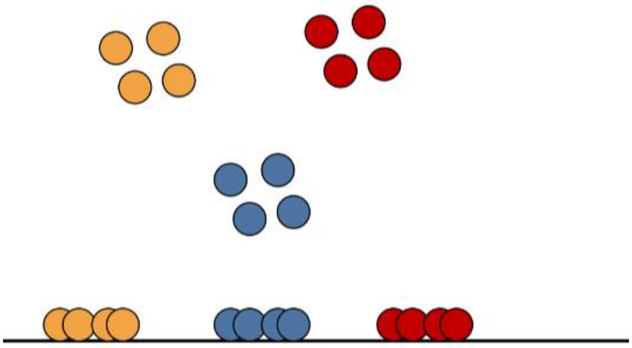
Visualization (Remind from last lecture)

1. PCA VS t-SNE

In the previous lecture, we have learnt PCS before. Although it can reduce the dimensions, it has a problem. It cannot preserve the original cluster information. Because of that we should use another method called t-SNE.

2. t-SNE

- Non-linear stochastic dimension reduction technique
- To map high dimensional data into low dimensional space
- Random initialization→ For each point, update the position a little bit→ Repeat the procedure→Until no more update



Compare the cluster to the original cluster. The points from the same cluster attract each other. The points from different clusters push apart each other

- T-SNE result will preserve the cluster while scarifying the physical meaning of cluster distance

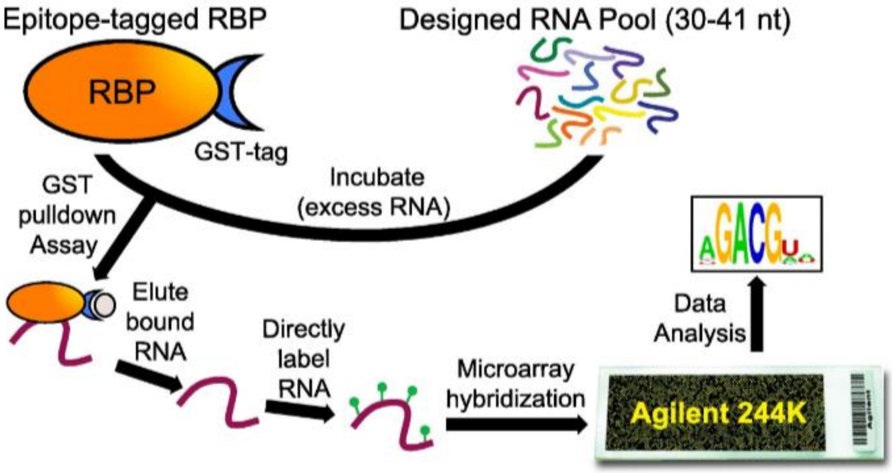
Protein-RNA/DNA Interaction

1. Protein binding analysis

It shows that the nucleotide has the highest possibility to bind with a protein since there is a binding preference.

2. Motif (the preference of protein binding)

- How to get the bind motif: RNA bind with the epitope-tagged RBP→GST pulldown→sequencing



- From aligned sequences to motif
  - i. Align the sequences
  - ii. Convert to position count matrix
  - iii. Convert to position probability matrix
  - iv. Convert to motif

- Example from lecture notes

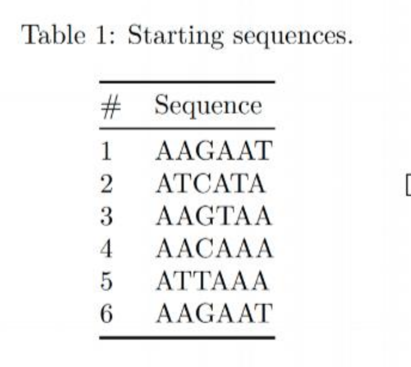


Table 1: Starting sequences.

#	Sequence
1	AAGAAT
2	ATCATA
3	AAGTAA
4	AACAAA
5	ATTAAA
6	AAGAAT

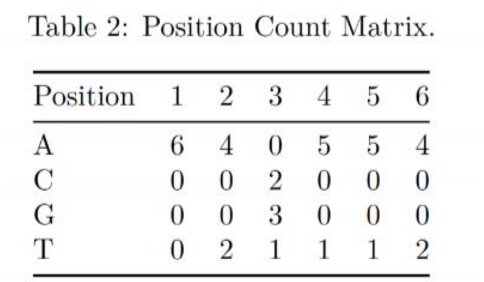


Table 2: Position Count Matrix.

Position	1	2	3	4	5	6
A	6	4	0	5	5	4
C	0	0	2	0	0	0
G	0	0	3	0	0	0
T	0	2	1	1	1	2

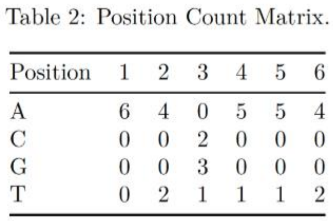


Table 2: Position Count Matrix.

Position	1	2	3	4	5	6
A	6	4	0	5	5	4
C	0	0	2	0	0	0
G	0	0	3	0	0	0
T	0	2	1	1	1	2

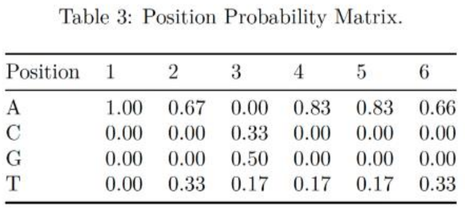


Table 3: Position Probability Matrix.

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

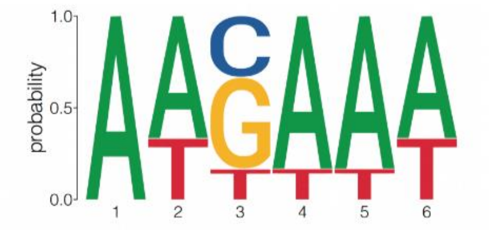
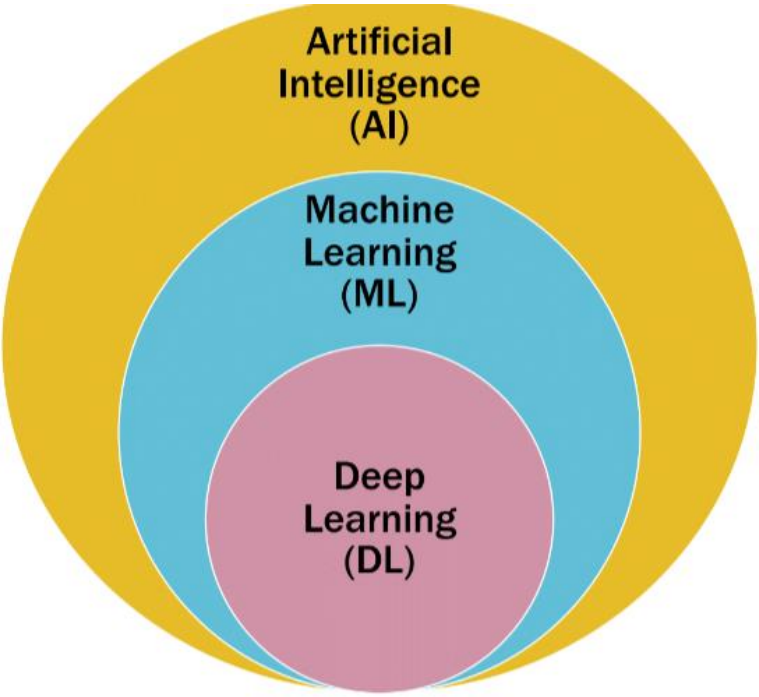


Figure 1: Sequence logo of a Position Probability Matrix

AI vs ML vs DL

1. Artificial intelligence: to mimic human behavior
2. Machine learning: Subset of AI, perform specific tasks without using explicit instructions, only reply on patterns and inference from the data
3. Deep Learning: Subset of ML, takes advantage of multi-layer neural networks


Artificial Intelligence (AI)
Machine Learning (ML)
Deep Learning (DL)