**Biological Databases: Theories and Practice (2023 Fall)**

**Assignment II**

1. (30 points) Please provide the introduction of your final project, including the **background**, **motivation**, and **goals** of the project.

Once the network information of all cells in a dataset is compressed into a single NDM with the same dimension as the gene expression profile, we could easily apply existing scRNA-seq methods for further analysis. One of the most common analyses in single-cell studies is the clustering of cell types or cell states. Clustering is a field of unsupervised learning that tries to find structure in unstructured data by creating groups of similar values.

At present, there are numerous tools and algorithms available for clustering analysis, ranging from traditional methods like k-means and hierarchical clustering to more modern methods like Louvain clustering. In addition, dimension reduction methods such as Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) are also commonly applied to compress and extract information and for easy visualization.

Together, UMAP and Louvain clustering provide a powerful framework for analyzing scRNA-seq data and gaining insight into the biological systems under study.

It is also worth noting that we applied logarithm transformation to NDM as in common scRNA-seq studies and standardized each feature (gene) by removing the mean and scaling to unit variance before clustering analysis. Standardization of a dataset is a common prerequisite for many machine learning models (including clustering algorithms); otherwise the models might not perform well as expected.

1. (30 points) Please provide the description about the **data sources** or **materials** for the final project.

In this project, we used scRNA-seq datasets as input to construct single cell networks (SCNs). Specifically, by dataset, we mainly mean its expression profile or gene expression matrix (GEM). A GEM is a gene by cell matrix with each element inside denoting the expression level for a given gene in a given cell.

I downloaded scRNA-seq dataset from https://www.nxn.se/single-cell-studies

In this project, we followed the common single-cell studies and the previous SCN methods to preprocessed the expression profile by the two main steps:

* Applying logarithm transformation (2 based)
* Removing genes expressed in less than 10 cells

In addition to the processing of the data, we also have to re-organize and standardize the cell type information into one universal format for later uses.

1. (40 points) Please provide the **database schema** in terms of **Entity-Relationship diagram** for the final project.





