CS690 : Computational Genomics
Clustering of Spatial Transcriptomics Data
Group Assignment 2022-23
(Group 1)

Indian Institute of Technology, Kanpur

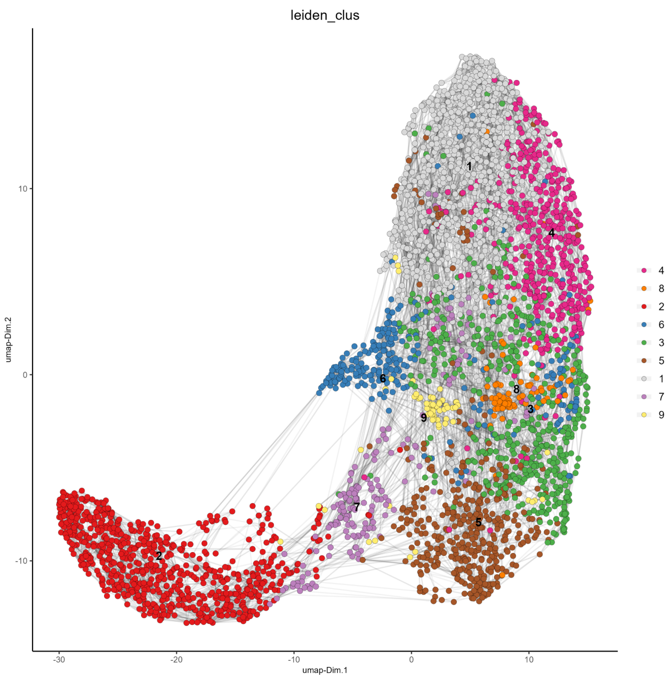
TASK 1: Leiden Clustering

- Algorithm Description -

  Leiden Clustering was developed in 2008 and most widely used in the subsequent years for clustering. Its development overcame the major disadvantages of the earlier used Louvain Clustering method.

  Louvain Clustering swapped nodes randomly to find out well connecting clusters, and once it could not find any better clusters, each cluster would be grouped into a node and then these nodes would be grouped further to make nodes. This random movement of nodes can lead to swapping out of one of the most well connected nodes which would break the integrity of the cluster it belonged to. This also leads to grouping of nodes/clusters which are not very closely linked. The continuous movement of nodes also made the process time consuming.

  Leiden Clustering is not only able to group clusters, but also split them. If the process is run many times, it can lead to subset optimal nodes, which would not undergo significant quality changes if a node is swapped. It also moves only those nodes which are unstable. This results in more optimal clusters than Louvain Clustering, in a lesser time. Thus, Leiden Clustering is comparatively more efficient. We have employed it to perform spatial clustering of the given data.
- Implementation -
  1. Raw data normalisation
  2. HVG(Highly variable genes - to remove abnormal data) and PCA dimensional reduction
  3. Nearest Network(NN) Algorithm (for Domain Number Selection)
  4. Leiden Algorithm
- Spatial Feature Plot -



TASK 2: PCA(Principal Component Analysis) and HMRF(Hidden Markov Random Fields) - Giotto Package

- Algorithm Description -

  Leiden Clustering is an algorithm which is not built for recognizing spatial relationships in particular. HMRF by Giotto Suite was developed in 2021 and is capable of bringing a smoothness constraint specifically for spatial clustering. This method requires pre-processing(normalisation and dimensional reduction) and an input to specify the number of clusters (Domain Number Selection) which is a shared feature with Leiden Clustering. HMRF is able to detect spatial domains by systematically comparing the gene expression of each cell with its surrounding cells to generate coherent patterns. First, a spatial representation of cells is created in the form of an undirected graph and then the domain state of the cell(which cluster it belongs to) is found based on its own gene expression pattern and the domain states of the neighbouring cells.

  It is based on the Markov property, which allows reduction of spatial constraints by only considering correlation between immediate neighbouring nodes. The "hidden" refers to inferring the spatial pattern indirectly from directly measured variables.

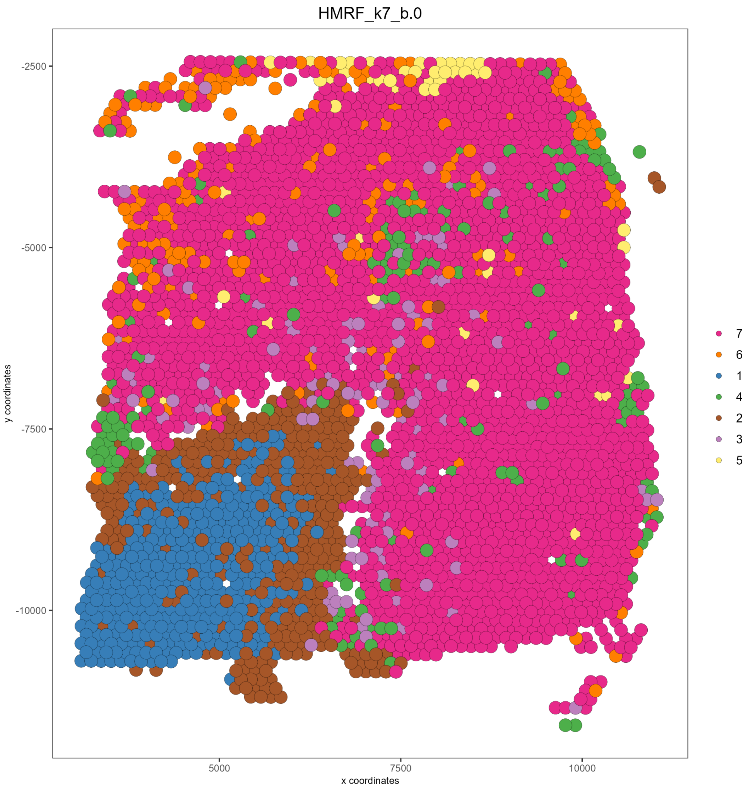
  Mathematically, the contribution of neighbouring cells is represented by an energy field and the optimal solution is represented by the equilibrium of the field.

  It makes use of a multivariate Gaussian random variable model to generate the distribution of the spatial information in the form of cluster patterns and gene expression in the form of signals. 4 steps are - 1) Neighbouring graph representation. 2) Gene selection. 3) Domain number selection, and 4) Implementation and model inference.

  Newer methods may aim to make the pipelines in the form of an integrated workflow, such that the number of parameters which need to be defined manually and non intuitively are reduced at each processing step.
- Several variables can be chosen manually, based on individual data biases and platform biases -
  - Number of differentially expressed genes(lowly and highly expressed, based on number of copies per cell) across different cell types to remove transcriptomic redundancy
  - Number of cells to be excluded which can not be designated a single cell type (based on the probability of belonging to any cell type)
  - The number of top genes to be used as input for HMRF based on a spatial coherence score
  - Filtering out genes which are highly specific to a single cell type to improve the resolution
  - Number of bins to collectively average with during normalisation(PCA)
  - C - regularisation pattern that trades off misclassification due to overfitting against simplicity of the decision function. A lower C increases the ability of the model to generalise to unseen data at a cost of larger fitting error - to maximise cross platform correlation between cell type specific gene expression profiles
  - Beta - reflects the strength of interactions in HMRF
- Implementation -

1. Data was z score transformed to make the dynamic range compatible and normalised to equalise statistical distribution between spatial and expression data
2. HVF was used to filter out outlier data
3. Dimensional Reduction by PCA, UMA and tSNE was performed
4. Number of clusters/domains required was calculated from K-means (k=7 was taken)
5. Top 100 genes were selected based on their features
6. HMRF algorithm was applied

- Spatial Feature Plot -



WORK DISTRIBUTION

We worked as a team and each & every Task.
Aman Dixit (190103)
Shreya Vaish (190822)
Tanishq Gupta (190894)

REFERENCES

https://web.archive.org/web/20220522082946id_/https://www.biorxiv.org/content/biorxiv/early/2021/11/11/2021.10.27.466045.full.pdf
https://www.cwts.nl/blog?article=n-r2u2a4
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6488461/