Binarized single cell RNA sequencing data clustering

The impact of binarized scRNA-seq data on clustering through community detection algorithms

1 Introduction

Background:
Single-cell RNA sequencing (scRNA-seq) technology has become an important method for deciphering the heterogeneity and complexity of RNA expressions within individual cells. scRNA-seq also reveals the composition of different cell types and functions within highly organized tissues, organs and organisms [1]. A major technique for getting valuable insight from the scRNA-seq data is the clustering of cells based on their gene expression levels.

Problem definition:
The rapid up-scaling of scRNA-seq datasets in recent years requires the clustering algorithm to run in a more time and memory efficient manner.

Potential solution:
Binarize the scRNA-seq data so that it can be stored in a binary format to reduce memory usage and use binary clustering methods with improved time efficiency. The full impact on resulting clustering quality is still unknown.

Research question:
What community detection algorithm results in clusters most similar to current state-of-the-art clustering methods when applied to binarized scRNA-seq data?

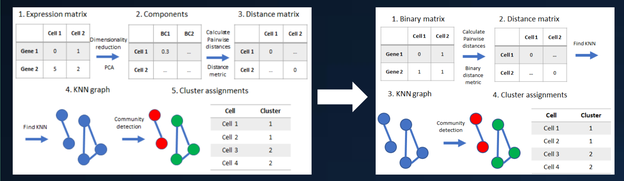
2 Materials

Alzheimer's dataset				
Cells	*Genes*	*Ground Truth*		
11884	867	Jaccard	K=10	#Clusters=6

Xenopus' tail dataset				
Cells	*Genes*	*Ground Truth*		
13199	4469	Cosine	K=10	#Clusters=8

3 Methodology

Step 1. Build a binary clustering pipeline



Conventional clustering pipeline

Binary clustering pipeline

Step 2. Configure the experimental setup

KNN-graph:
Parameters
- K = 10

Distance metrics
- Binary Cosine (Ochiai)

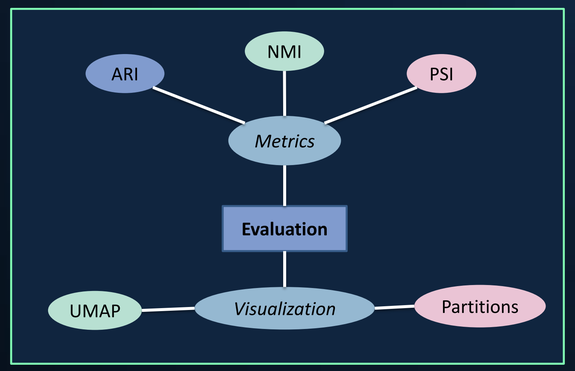
Community detection:
Algorithms
- Leiden
- Louvain

Parameters
- Resolution (γ) = variable
- Randomness (β) = 0.01
- Iterations (n) = infinity

Step 3. Run binary clustering experiments

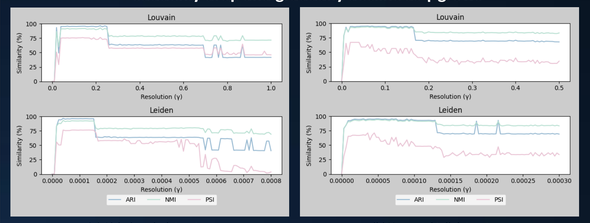
1. Collect 1000 binary clustering solutions for increasing resolution parameter γ to get a general overview of the behavior of the Louvain and Leiden community detection algorithms.
2. Evaluate the binary clustering solutions against a conventional ground truth solution in a supervised manner.
3. Find the γ-range for which the evaluation metrics peak
4. Find the binary clustering solutions that are most similar to the conventional ground truth for each evaluation metric.

Step 4. Evaluate the clustering results



4 Results

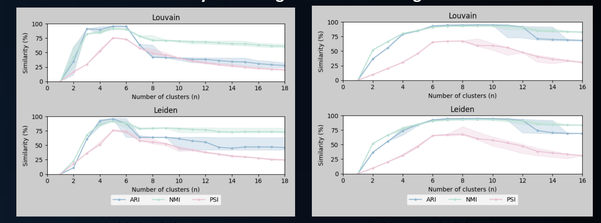
1. Similarity scores for all evaluation metrics peak for low γ and then suddenly drop and gradually decline as γ grows



(a) Alzheimer's dataset (b) Xenopus' tail dataset

Figure 1: Plot of the ARI, NMI and PSI similarity scores for binary clusterings resulting from the Louvain and Leiden community detection algorithms relative to the ground truths for increasing resolution (γ) for (a) the Alzheimer's dataset and (b) the Xenopus' tail dataset. The x-axis represents the resolution parameter γ, the y-axis represents the similarity score and the colors represent the different cluster evaluation metrics.

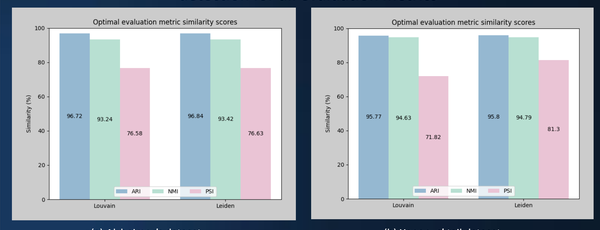
2. Similarity scores for all evaluation metrics are highest when the number of clusters in the binary clustering solution and the ground truth is close



(a) Alzheimer's dataset (b) Xenopus' tail dataset

Figure 2: Plot of the mean, minimum and maximum ARI, NMI and PSI similarity scores by the number of clusters in the binary clustering solutions resulting from the Louvain and Leiden community detection algorithms relative to the ground truths for (a) the Alzheimer's dataset and (b) the Xenopus' tail dataset. The x-axis represents the number of clusters (n) in the binary clustering solution, the y-axis represents the similarity score and the colors represent the different cluster evaluation metrics.

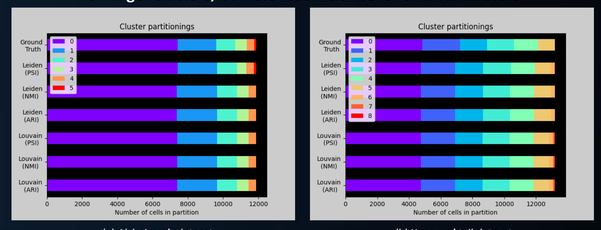
3. Leiden community detection scores higher than Louvain community detection for all evaluation metrics



(a) Alzheimer's dataset (b) Xenopus' tail dataset

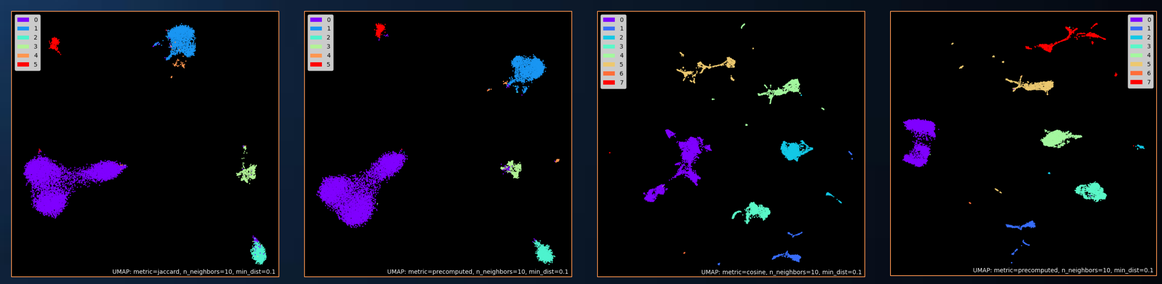
Figure 3: Histogram of the optimal ARI, NMI and PSI similarity scores of the binary clustering solutions resulting from the Louvain and Leiden community detection algorithms relative to the ground truths for (a) the Alzheimer's dataset and (b) the Xenopus' tail dataset. The x-axis represents the community detection algorithm, the y-axis represents the similarity score, the values inside the bars show the exact optimal similarity score and the colors represent the different cluster evaluation metrics.

4. Binary clustering shows great resemblance to conventional clustering for larger clusters, but less resemblance for smaller clusters



(a) Alzheimer's dataset (b) Xenopus' tail dataset

Figure 4: Stacked bar chart showing the cluster partitionings for the ground truth and binary clustering solutions with the optimal ARI, NMI and PSI metric scores for the Louvain and Leiden community detection algorithms for (a) the Alzheimer's dataset and (b) the Xenopus' tail dataset. The x-axis represents the number of datapoints in the partitioning, the y-axis represents the clustering solution and the colors represent the different clusters in the partitioning



Alzheimer's dataset (Ground truth) Alzheimer's dataset (Optimal PSI) Xenopus' tail dataset (Ground truth) Xenopus' tail dataset (Optimal PSI)

5 Conclusion

1. The Leiden community detection algorithm produced results that are closers to the conventional ground truth solutions than the Louvain algorithm for all metrics
2. Binary clustering showed great resemblance to conventional clustering for large clusters, but smaller clusters had little overlap or went completely undetected.

References

[1] Jovic Dragomirka, Liang Xue, Zeng Hua, Lin Lin, Xu Fengping, and Luo Yonglun. Single-cell rna sequencing technologies and applications: A brief overview. *Clin Transl Med.*, 12(3), 2022.

Jurriën Theunisz | J.Theunisz@student.tudelft.nl

TU Delft

Professor: Marcel Reinders
Supervisor: Gerard Bouland