Memory usage analysis of binary clustering algorithm

What is the gain in peak memory usage of the binary clustering algorithm compared to current state-of-the-art clustering methods?

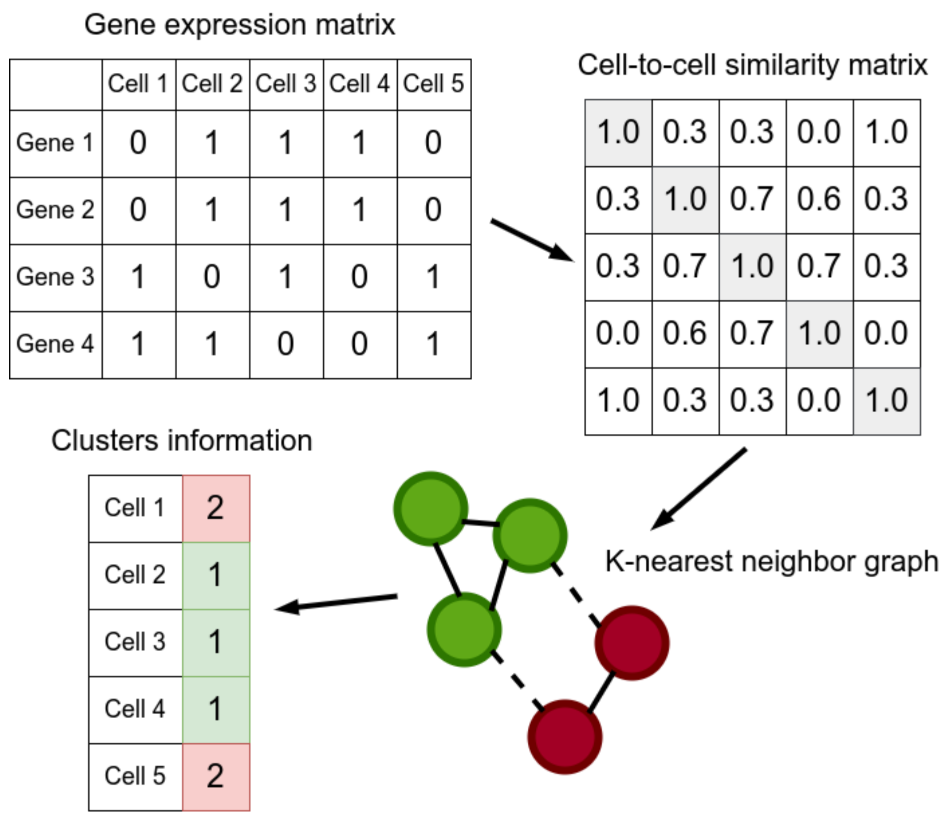
Author: Pavel Verigo (paul.verigo@gmail.com)

Supervisor: Gerard Bouland
Responsible Professor: Marcel Reinders

June 28, 2023

INTRODUCTION

This research aims to enhance the efficiency of clustering techniques for analyzing single-cell RNA sequencing (scRNA-seq) data, which poses scalability challenges due to increasing size input. The primary objective is to evaluate the potential memory efficiency and computational speed gains of a binary clustering algorithm, which were proposed in a paper by Bouland et al. [1]. The idea involves using a boolean value to represent gene expression in a cell. We illustrate the common clustering workflow below.



Gene expression matrix

	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5
Gene 1	0	1	1	1	0
Gene 2	0	1	1	1	0
Gene 3	1	0	1	0	1
Gene 4	1	1	0	0	1

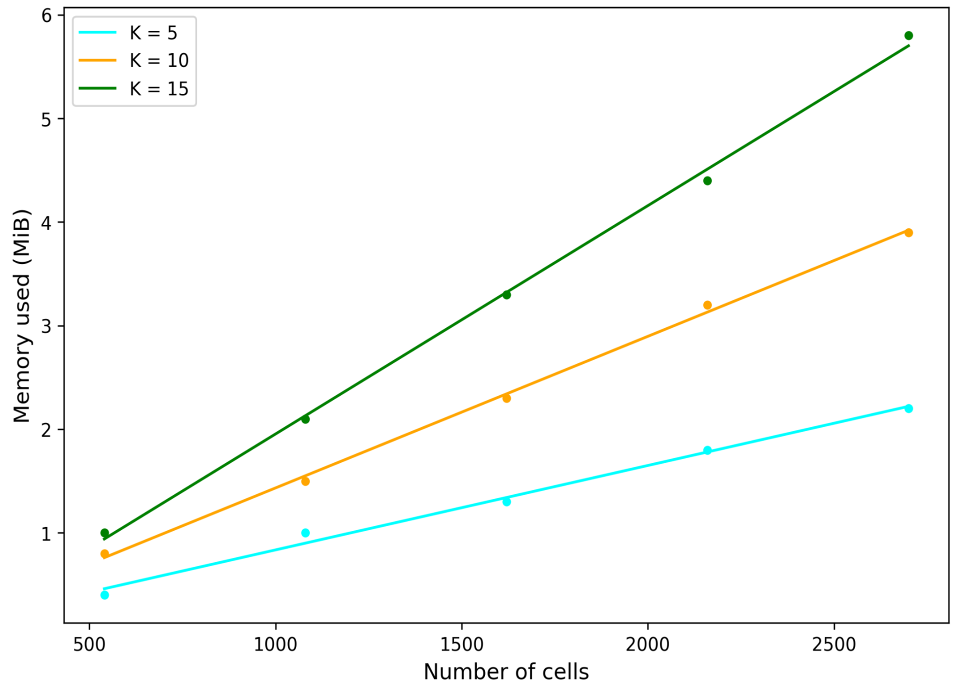
Cell-to-cell similarity matrix

1.0	0.3	0.3	0.0	1.0
0.3	1.0	0.7	0.6	0.3
0.3	0.7	1.0	0.7	0.3
0.0	0.6	0.7	1.0	0.0
1.0	0.3	0.3	0.0	1.0

K-nearest neighbor graph

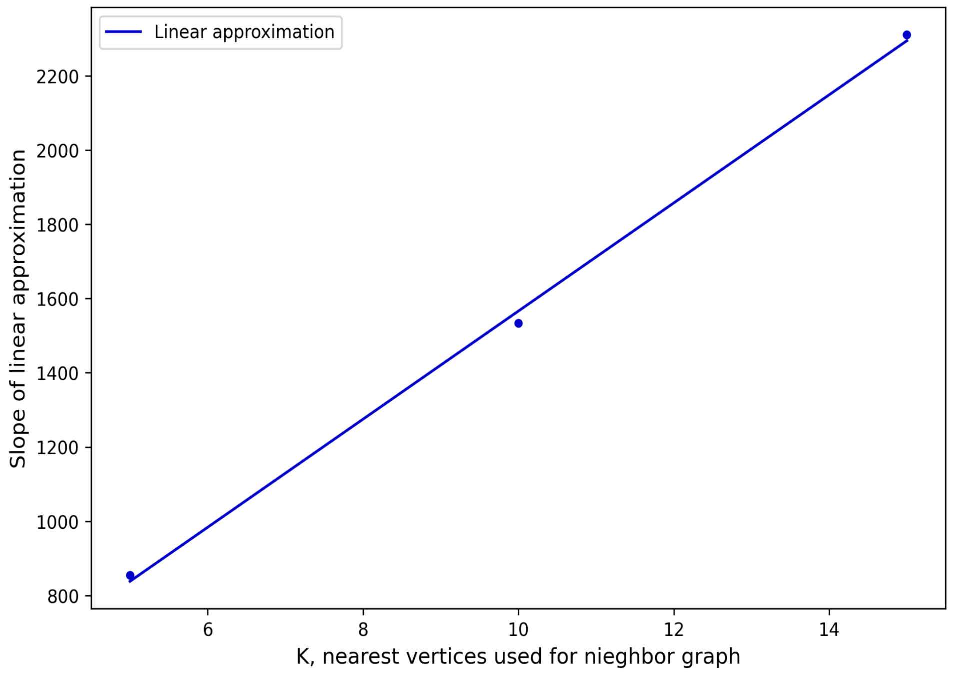
Clusters information

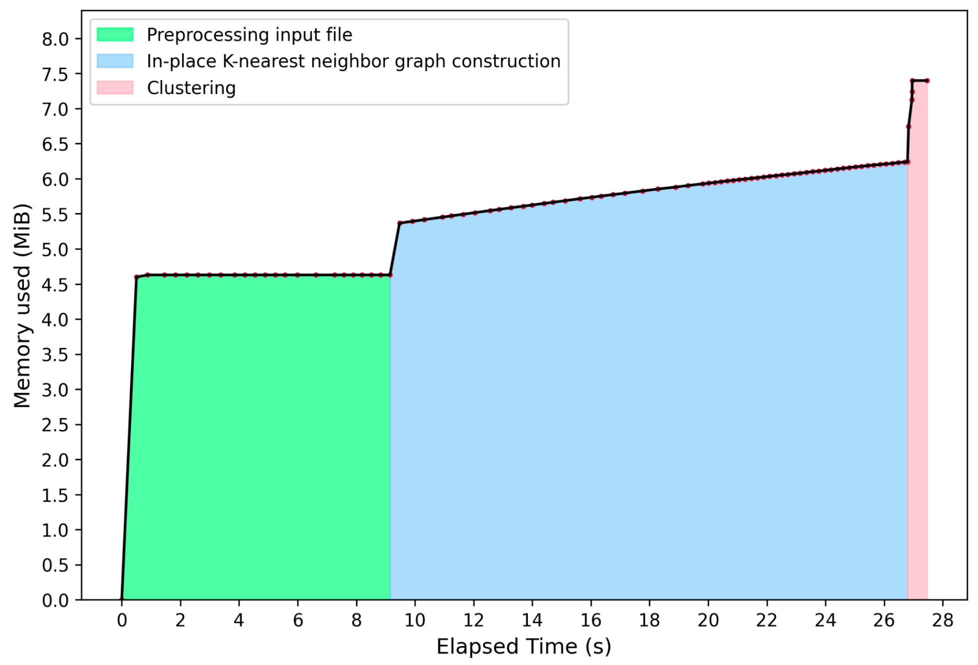
Cell 1	2
Cell 2	1
Cell 3	1
Cell 4	1
Cell 5	2



Q ESTIMATION

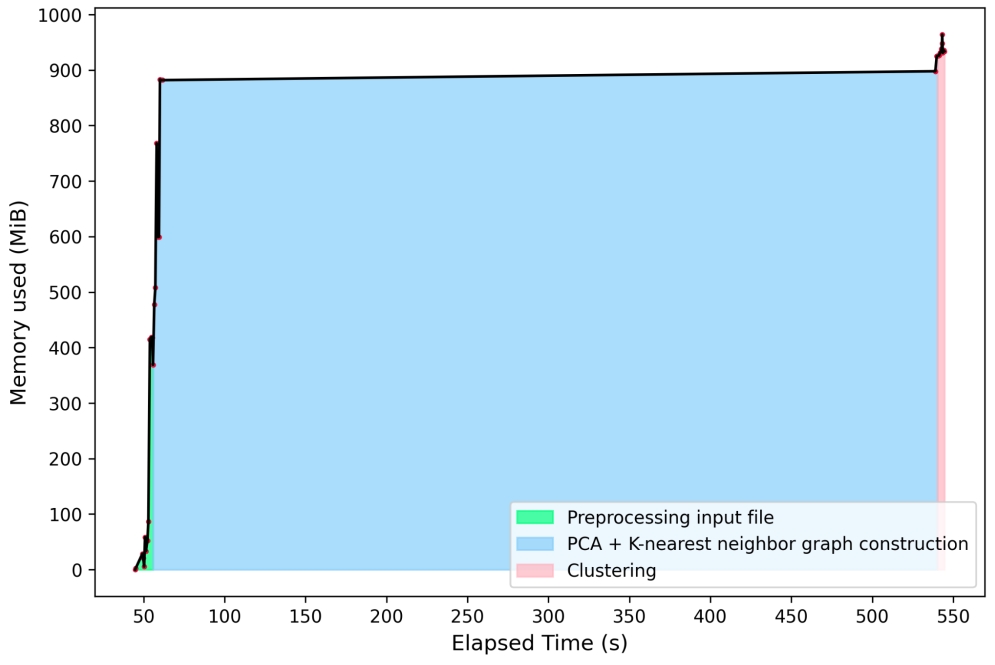
We estimate Q by calculating memory used in workflow, excluding memory that dataset occupies, which should be $Q \times K \times C$. We measured different $Q \times K$ on the left graph using other K. Then Q was approximated to be 145.6 from running linear approximation, depicted on the left





COMPARING TO SEURAT

We tested our implementation (left) and Seurat toolkit (right) on the same dataset. We found, that Seurat uses more memory due to calculating the PCA matrix and unnecessary data duplication.



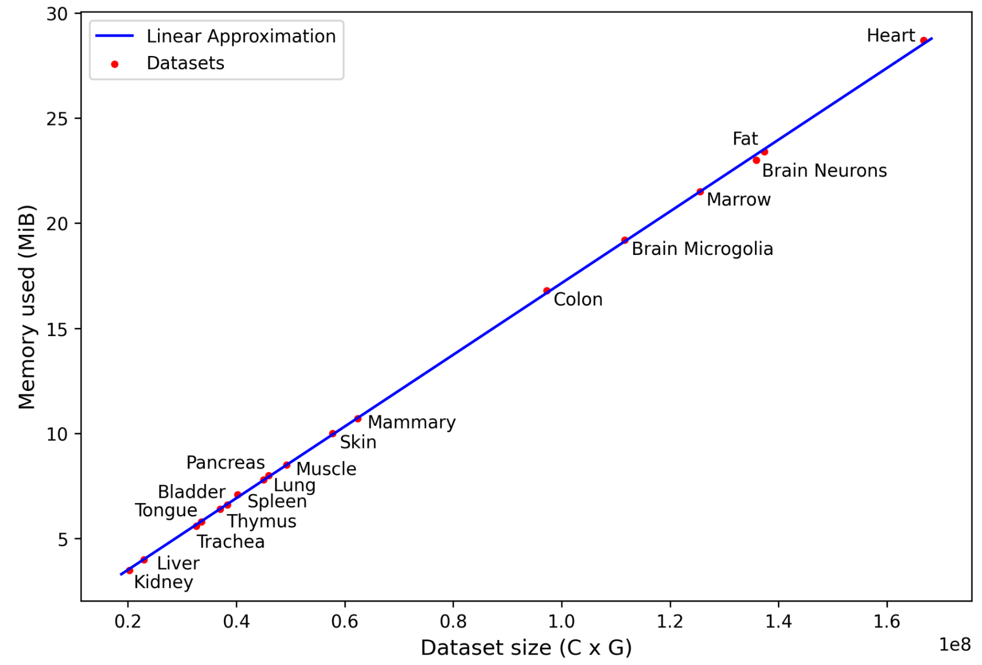
THEORETICAL ANALYSIS

1. The research estimates optimal memory usage in a binary clustering workflow, using parameters like the number of genes G, cells C, and edges per cell K in the neighbor graph.
2. Initial dataset memory usage is estimated as one-eighth of the product of the number of cells and genes, given each gene expression needs only one bit.
3. We propose memory memory efficient computation of the cell-to-cell similarity matrix, by outputting only K closest edges. Next we process this using chosen clustering algorithm which uses memory proportionally to number of edges. Therefore we introduce new coefficient Q to estimate number of bytes per edge.
4. Resulting memory usage is

$$\frac{1}{8} \times C \times G + Q \times K \times C \quad \text{bytes}$$

CONSTANT NUMBER OF GENES

When the number of genes is constant, we expect peak memory to be linear to the number of cells/dataset size. Our profiling of Tabula Muris proves our hypothesis.



DISCUSSION AND CONCLUSION

- Based on $Q \approx 145.6$ and applying formula, we estimate that for datasets where number of genes G exceed 100,000, most of the memory is used by dataset. Therefore binarization technique will majorly win against non-binarized approaches.
- We suggest multiple optimizations for implementation to reduce the Q constant in the clustering workflow by minimizing memory allocation. Not doing reallocation and possibly changing the underlying clustering library can reduce Q below 100.
- We propose a hybrid approach to gene expression data storage that combines binarization and compressed data representation, potentially improving memory efficiency .

[1] Gerard A. Bouland, Ahmed Mahfouz, and Marcel J. T. Reinders. Consequences and opportunities arising due to sparser single-cell rna-seq datasets. Genome Biology, 24:86, April 2023