Evolutionary Multi-Objective Clustering of Single-Cell RNA Sequencing Data

Konghao Zhao and Natalia Khuri
DataMine Research Group, Department of Computer Science, Wake Forest University

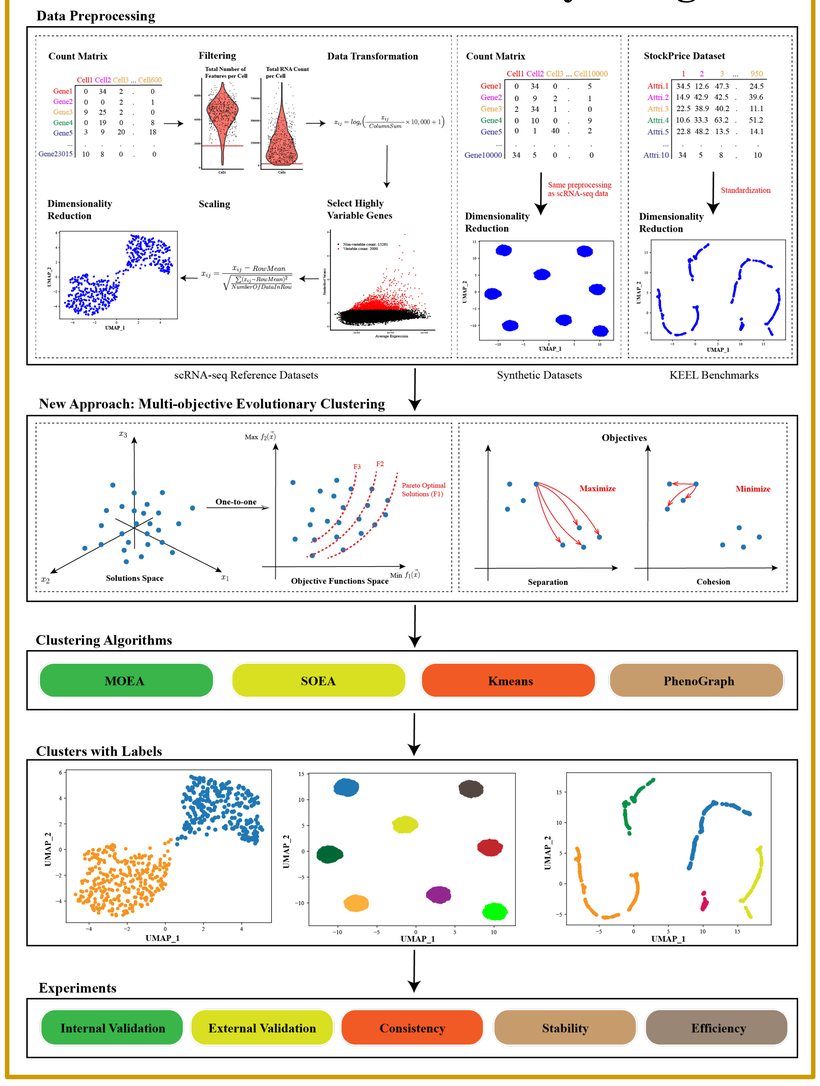
Background

- Cells are the basic building blocks of the human body and are significant to human health.
- Single-Cell RNA sequencing (scRNA-seq) produces data that measure gene expression in individual cells.
- Clustering is an essential method for the discovery of cell types.
- Most clustering approaches focus on one objective only, which may not produce meaningful results in sc-RNA seq data.

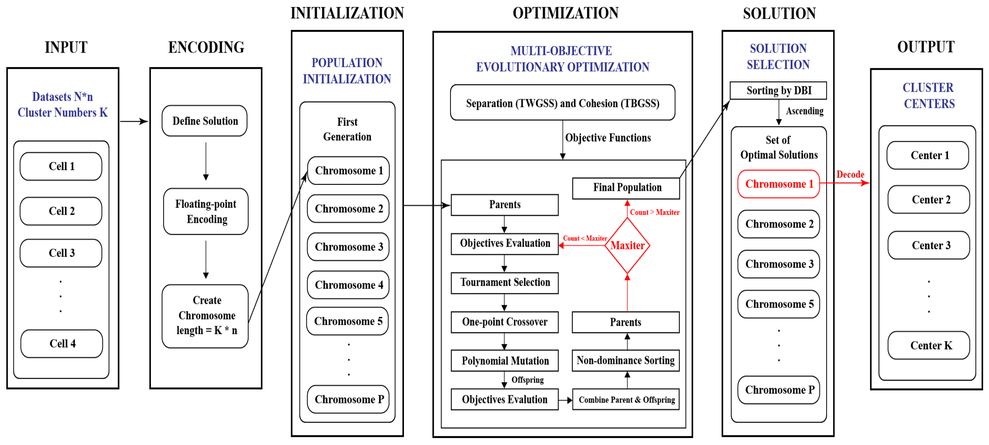
Research Objectives

1. Design, implement, and test multi-objective evolutionary clustering algorithm (MOEA) for scRNA-seq data.
2. Perform systematic comparison and evaluation with baseline clustering algorithms (Kmeans and PhenoGraph) and single-objective evolutionary algorithm (SOEA).

A. Overview of the Study Design



B. General Architecture of MOEA



scRNA-seq data challenges:

- Large size
- High dimensionality
- Sparsity
- Technical noise

HPC Resources: DEAC

- Nodes or Blades : 94 nodes
- Processors : 4,224 cores
- GPU cores : 68,608 cores
- Memory : 18.67TB
- Storage : 221TB

C. Parameter Tuning

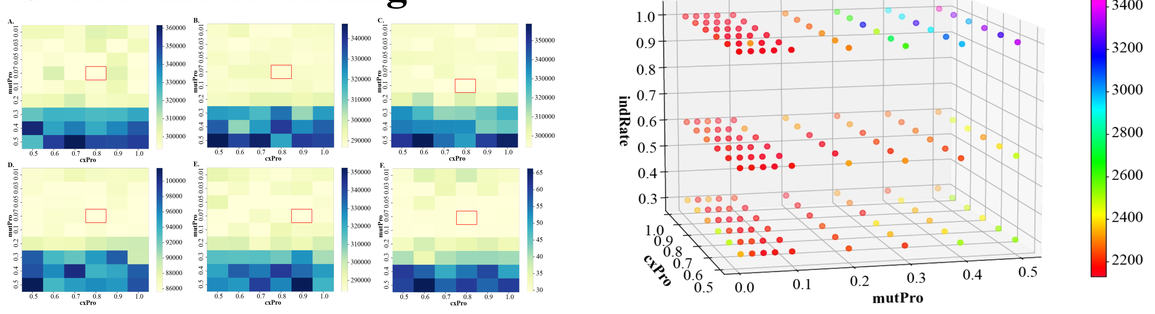


Figure 1: Hyperparameter tuning of mutation rate and probability, and crossover probability of SOEA with KEEL benchmarks.

D. Running Time Analysis

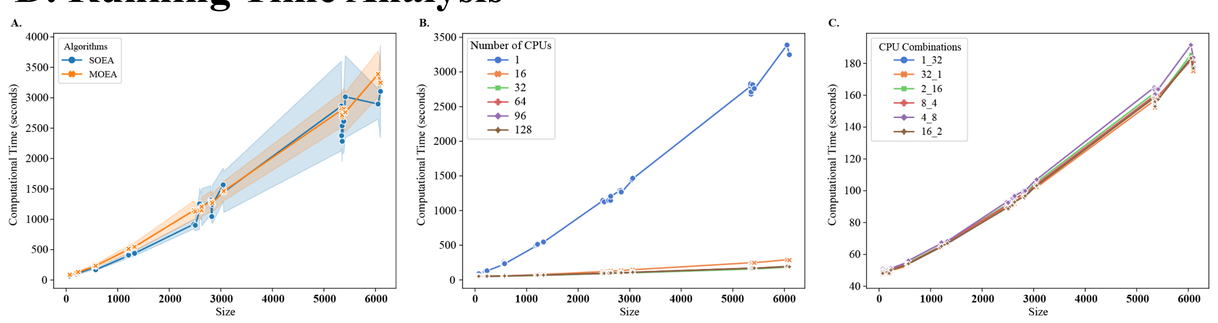


Figure 2: (A.) Time comparison of SOEA and MOEA, (B.) MOEA with 6 different numbers of CPUs, and (C.) MOEA with 32 CPUs with different combinations.

E. Results

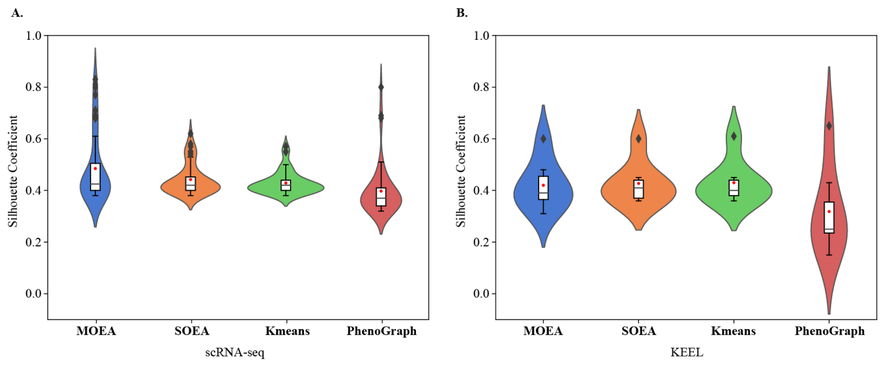


Figure 3: Internal validation of MOEA, SOEA, Kmeans, and PhenoGraph.

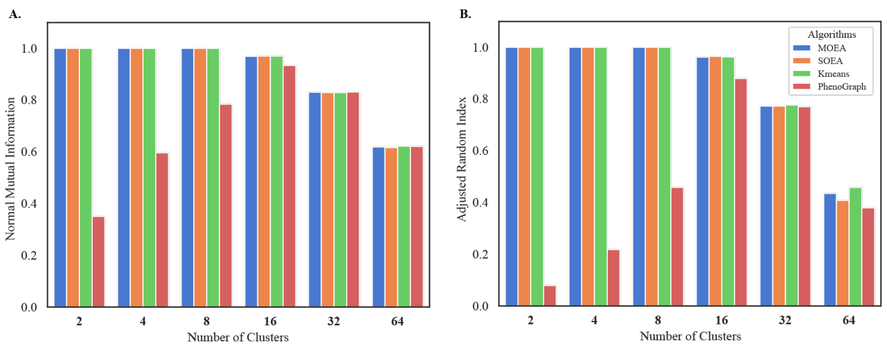


Figure 4: External validation of MOEA, SOEA, Kmeans, and PhenoGraph.

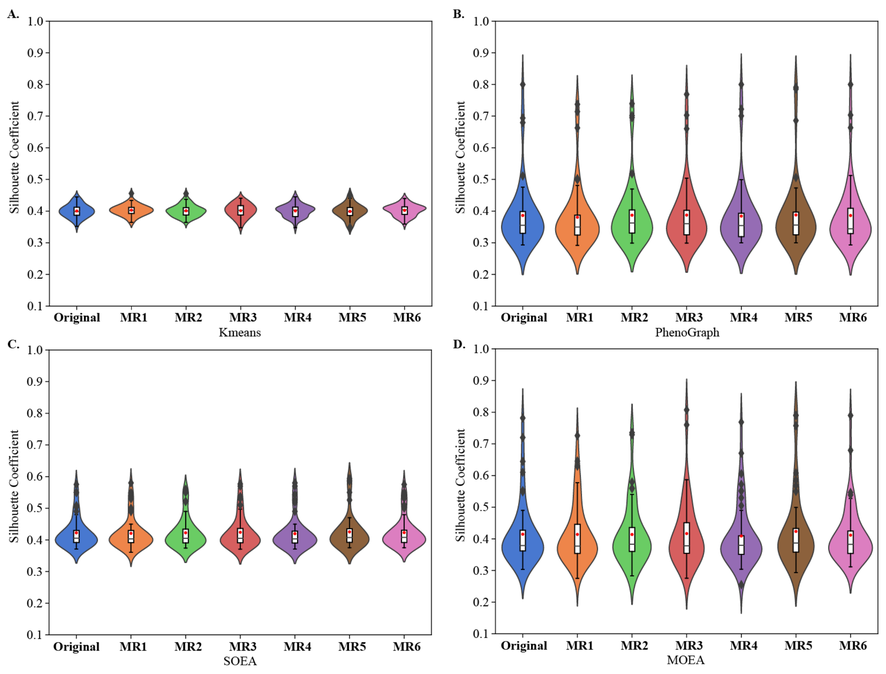


Figure 5: Metamorphic testing of stability of MOEA, SOEA, Kmeans, and PhenoGraph.

MR1: Permute the order of cells
MR2: Modify gene counts
MR3: Duplicate 1 cell
MR4: Permute the order of genes
MR5: Add zero gene counts
MR6: Negate gene counts

Metrics:

$$Sil = \frac{1}{n}\sum_{i=1}^{n} S_i,\ where\ S_i = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}}$$

$$NMI(L,C) = \frac{2 \times I(L;C)}{H(L) + H(C)},\ where$$

$$I(L;C) = H(L) - H(L|C),\ H(x) = -\sum_{i \in x} P_i * log_2 P_i$$

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{n_i}{2}\sum_j \binom{n_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{n_i}{2}\sum_j \binom{n_j}{2}] - [\sum_i \binom{n_i}{2}\sum_j \binom{n_j}{2}]/\binom{n}{2}}$$

Acknowledgement:

- Wake Forest Research Fellowship from URECA Center of Wake Forest University
- DataMine Research Group, Cody Stevens, Nathan Whitener

Conclusion

- We designed, implemented, and tested a MOEA and a SOEA for clustering of scRNA-seq data.
- We systematically tested algorithms' performance, accuracy, stability, consistency, and efficiency.
- We showed that the performance and accuracy of the MOEA and SOEA are stable, consistent, and on par with or better than baseline algorithms.
- In the future, we plan to implement an adaptive selection of hyperparameters.

WAKE FOREST UNIVERSITY