Evaluating Performance of Clustering Algorithms using Single-Cell RNA-Seq Data

Sarah Baalbaaki
Computational Biology Department
Carnegie Mellon University
sbaalbak@andrew.cmu.edu

Sumitra Lele
Computational Biology Department
Carnegie Mellon University
slele@andrew.cmu.edu

Sarah Oladejo
Biology Department
Carnegie Mellon University
soladejo@andrew.cmu.edu

1 Introduction

1.1 Background & Objective

Single-cell RNA-seq data can be used to answer crucial questions about the differences between individual cells within an environment, and cells across different subtypes, which helps identify sub-populations and areas of interest for a specific problem being tackled. Efficient computational analyses of scRNA-seq data are required to extract inherent biological information from the huge quantity of information that single-cell RNA-Seq technology presents. The integration of single-cell RNA-seq data with other machine learning algorithms has facilitaed these methods of analysis. One of these algorithms is clustering which is applied on such data. It is an upstream step that affects other downstream analysis, and it is therefore imperative to obtain accurate cluster assignment and number of cell types from scRNA-Seq data.

In our project, we compare algorithms that use deep learning for clustering. We apply the algorithms to scRNA-Seq data and compare the different clustering methods and obtained results across different tissue types. The algorithms are benchmarked based on how much they deviate from the true cell type label provided. We also infer the correctness of the algorithm based on the differential expression of established marker genes in those tissue types.

1.2 Problem and Motivation

How do deep learning based clustering methods vary in the way they cluster sc-RNA seq data at a high level, and how do the models compare based on resulting cell annotation and evaluation metrics.

1.3 Previous Literature

Previous literature has compared different algorithms that estimate the number of cell types and those algorithms have been classified into categories. These categories include intra- and inter-cluster similarity, modularity in community detection, stability metrics, and eigenvector-based metrics. Efforts have also been made to build algorithms that cluster and functionally annotate cell types using unsupervised methods. Deep learning methods are not widely used and have not been adequately bench-marked with popular methods for single-cell RNA-Seq analysis. Here, we focus on the annotation of cell types using SEURAT (which is the traditional sc-RNA seq clustering algorithm) in comparison to other deep learning clustering algorithms.

1.3.1 SEURAT [1]

We use SEURAT as it is the traditional method for single-cell RNA sequencing analysis as it is composed of robust and widely accepted algorithms for quality control, normalization, dimensionality reduction, clustering, and cell type identification, which can serve as a solid starting point for our comparison to other methods.

1.3.2 DESC [2]

DESC (Deep Embedding for Single-Cell Clustering) is a deep learning model used to cluster sc-RNA seq data. The model is composed of a deep neural network designed to learn a low-dimensional representation of the scRNA-seq data that also reflects and captures the underlying structure of the cells and genes. The model is composed of a deep neural network autoencoder that maps the high-dimensional scRNA-seq data to a low-dimensional latent space. The algorithm thus initializes parameters obtained from an autoencoder and learns a nonlinear mapping function from the original scRNA-seq data space to a low-dimensional feature space by iteratively optimizing a clustering objective function. The model uses the KL-divergence-based clustering objective loss function. This loss function helps separate and compact latent spaces, which is used to train the encoder and clustering layer simultaneously, and helps improve the accuracy of clustering.

1.3.3 scDeepCluster [3]

This deep learning algorithm is a semi-supervised clustering algorithm. It uses the unsupervised K-Means algorithm to cluster the single cell RNA seq data, and then uses the resulting cluster labels along with the sc-RNA seq data as supervised input into an encoder-decoder architecture. The autoencoder is trained to learn a low-dimensional mapping for representing the scRNA-seq data, and so the latent space embeddings are used as inputs to K-Means clustering, which assigns cell types to each cluster. Thus, K-Means is used as an initialization for the embedding spaces. This method aims to overcome the issues of technical noise, cell-to-cell variability, and complex cell types that are hard to represent in low dimensions due to the curse of dimensionality, which they overcome when they use a deep neural network with ZINB loss.

The ZINB (Zero-Inflated Negative Binomial) loss is a loss function specifically used for sc-RNA seq data analysis since it is a type of negative binomial regression that takes into account the high degree of technical noise and dropouts that are present in scRNA-seq data. The model assumes that the observed counts in scRNA-seq data are generated by a mixture of two processes: a Poisson process that models the true gene expression, and a zero-inflated process that models the technical noise and dropouts. The ZINB loss function is used to optimize the parameters of a deep learning model to predict the underlying gene expression levels from the scRNA-seq data. When used with the autoencoder with the goal of minimizing the ZINB loss (minimizing the discrepancy between the predicted gene expression levels and the observed counts), it helps learn a low-dimensional representation of the scRNA-seq data that captures the underlying structure of the cells and genes.

2 Methods

2.1 Dataset

Single-cell RNA-seq data from Smart-seq2 sequencing of FACS sorted cells [4]

This dataset is composed of scRNA-seq data from Smart-seq2 sequencing of FACS sorted cells of 20 organs from 7 mice. It contains gene expression information for individual cells within a given cell population.

We decided to use five of these tissue samples from this dataset, related to immunity:

- Spleen

- Thymus
- Marrow
- Brain Neurons
- Brain Microglia

We decided to cluster the genes, with the aim of correctly identifying different cell populations.

2.2 Data Preparation and Pre-processing

Each cell type in the cell annotations had its own corresponding barcode, so we integrated the gene counts with the cell annotations based on these barcodes, and then combined the data across 5 tissues based on the corresponding genes, which create our integrated dataset. The dataset itself was mostly pre-processed and complete, but we further filtered the data. Basic QC was performed to remove mitochondrial genes, and then cells with less than 200 genes were filtered out, and genes expressed in less than three cells were also dropped. The downstream analysis was performed with 18501 cells and 21265 genes which were then normalized based on log normalization.

2.3 Implementation

2.3.1 SEURAT

To implement SEURAT, we had to find variable features, this helps to identify subset of features that show a high cell-to-cell variation in the dataset. These features are highly expressed in some cells, and expressed at a low rate in other cells. Next we scale the dataset using a linear transformation which shifts the expression of each gene so that the mean expression across cells is 0. This transformation also ensures the variance across cells is 1, giving equal weight to samples , so that highly-expressed genes do not dominate in downstream analysis. After scaling, we run the principal component analysis for dimension reduction. Cells are further clustered based on their PCA scores using a graph-based clustering approach. A KNN graph based on the euclidean distance in PCA space is constructed and edge weights between any pair of cells are refined based on the shared overlap in their local neighborhoods. The louvain modularity optimization algorithm is then used to iteratively group cells together. We finally plot cells in a low dimensional spacing using UMAP to learn the underlying manifold of the data which helps to place similar cells together in the low dimensional space.

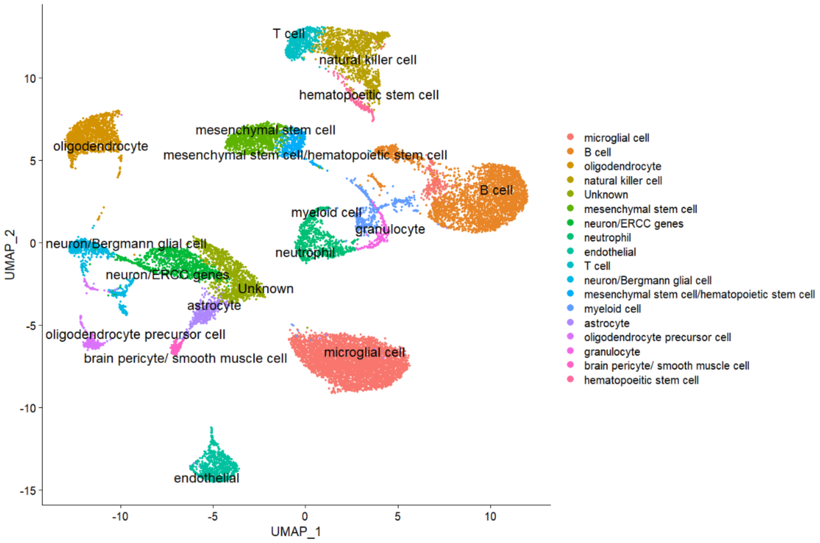

Figure 1: Annotation using the Seurat Pipeline

2.3.2 DESC

To implement DESC, we trained the model on our dataset, built a low-dimensional representation of the datasets, applied soft-clustering to assign cells, and then further annotated the cell types based on the expression of certain marker genes. We used already established and known marker genes (REFERENCE) to help us identify and annotate the cell clusters.

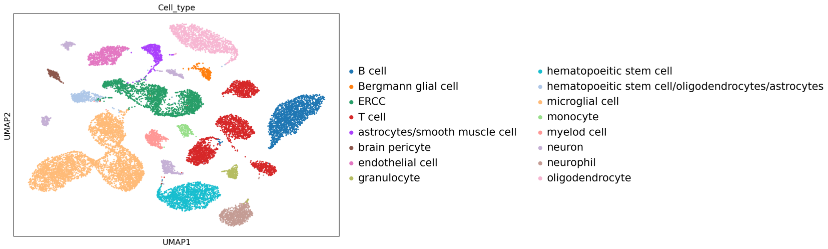

Figure 2: Annotation using the DESC Pipeline

2.4 Results and Conclusion

We obtained the results in Figures 1 and 2 above for the annotations using the Seurat and DESC pipelines. We also annotated the clusters using known marker genes, with an example as seen below, in Figures 3 and 4 for Hematopoietic Stem Cells, and B-Cells.

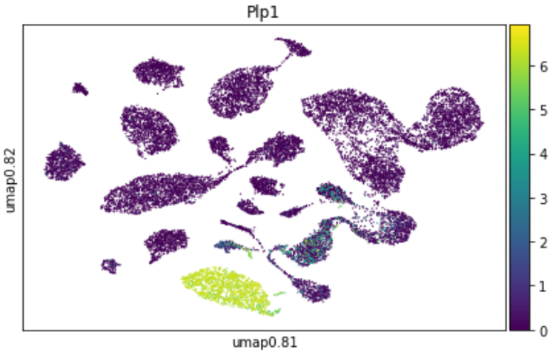

Figure 3: Cell type Assignment based on gene expression (Hematopoietic Stem Cell)

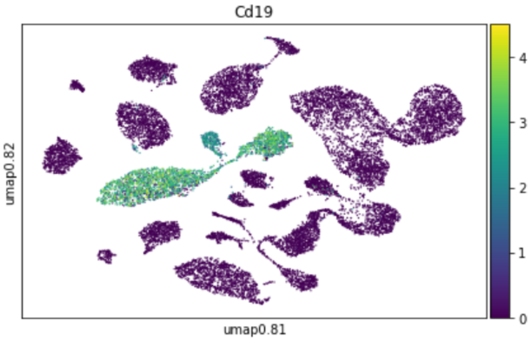

Figure 4: Cell type Assignment based on gene expression (B-Cells)

We then benchmarked both algorithms using the adjusted rand index and normalized mutual information, as seen in Figure 5.

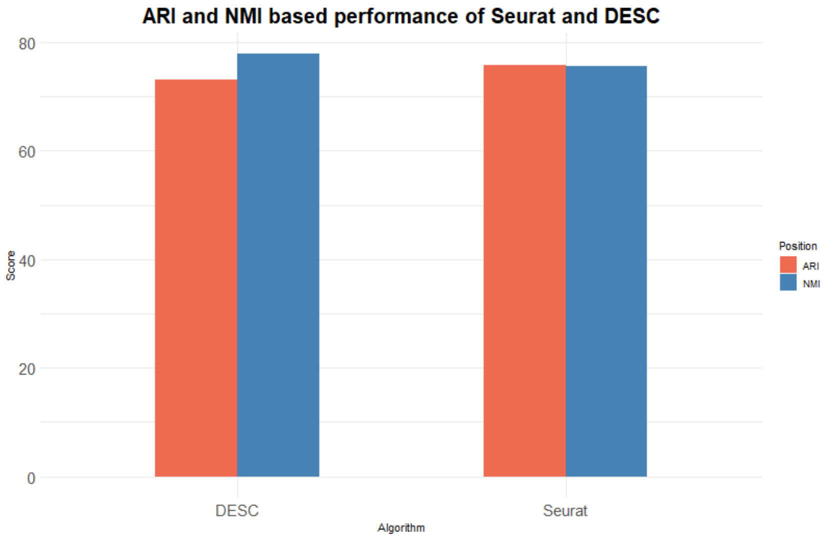

Figure 5: Benchmarking Algorithms using ARI and NMI

From the results shown in Figure 5, we can see that SEURAT and DESC are comparable in their performance, with both achieving an accuracy greater than 70 percent. Seurat was better at differentiating T cells and natural killer cells than DESC. However, DESC was able to clearly cluster the ERCC control transcripts while SEURAT could not. Overall, based on the datasets used, SEURAT performed better than DESC.

2.5 Future Work

To further improve our analysis, it would be crucial to perform the benchmarking with more datasets, in addition to using other benchmarking metrics like Jaccard, Silhouette width, etc.

Another approach would be to implement and further test out other deep learning algorithms, like scDeepCluster, which we mentioned above. scDeepCluster would serve as a combination of both methods since it is a semi-supervised approach to clustering, that includes both a traditional clustering method, in addition to using a neural network autoencoder architecture, and it would be interesting to assess the performance of this algorithm in comparison to SEURAT and DESC.

References

[1] Yuhan Hao et al. “Integrated analysis of multimodal single-cell data”. In: *Cell* (2021). DOI: 10.1016/j.cell.2021.04.048. URL: https://doi.org/10.1016/j.cell.2021.04.048.

[2] Xiangjie Li et al. “Deep learning enables accurate clustering and batch effect removal in single-cell RNA-seq analysis: Supplementary material”. In: (Jan. 2019). DOI: 10.1101/530378.

[3] Tian Tian et al. “Clustering single-cell RNA-seq data with a model-based deep learning approach”. In: *Nature Machine Intelligence* 1 (Apr. 2019), p. 191. DOI: 10.1038/s42256-019-0037-0.

[4] Valentine Svensson and Sarah A. Teichmann. *Single-cell RNA-seq data from Smart-seq2 sequencing of FACS-sorted cells v2*. Apr. 2018. DOI: 10.6084/m9.figshare.5829687.v2. URL: https://figshare.com/articles/dataset/Single-cell_RNA-seq_data_from_Smart-seq2_sequencing_of_FACS_sorted_cells_v2_/5829687.