Short Report on Clustering Analysis of Single-cell RNA-seq Data

Name: Sikha.Vamsi
Roll Number: 221058

1. Introduction

This report outlines the approach taken to cluster single-cell RNA-seq data using various algorithms and techniques. The dataset consists of gene expression profiles for single cells, and the goal is to identify clusters within this high-dimensional data.

2. Data Preprocessing

- **Loading Data:** The dataset was loaded using `scanpy`, with gene names provided in a separate CSV file. Gene names were assigned to the dataset's variables, ensuring that they matched the number of features in the dataset.
- **Quality Control:** Cells with fewer than 200 genes and genes present in fewer than 3 cells were filtered out.
- **Normalisation:** The data was normalised to a total count of 10,000 reads per cell and log-transformed.

3. Feature Selection and Scaling

- **Highly Variable Genes:** Selected the top 3,000 highly variable genes for further analysis to reduce dimensionality while retaining significant biological variance.
- **Scaling:** The data was scaled to a maximum value of 10 to standardize the gene expression values.

4. Dimensionality Reduction

- **Principal Component Analysis (PCA):** PCA was performed to reduce dimensionality, retaining 50 principal components for subsequent analyses.
- **Neighbours Calculation:** Computed the neighbourhood graph based on 50 PCA components with 15 neighbours.

5. Clustering and Resolution Optimization

- **Leiden Clustering:** Evaluated clustering resolutions ranging from 0.1 to 2.0. The Silhouette Score, used to assess cluster separation, indicated that a resolution of 0.5 provided the best results.
- **Best Resolution:** The Leiden clustering was performed with the optimal resolution of 0.5.

6. Visualisation and Validation

- **UMAP Visualization:** Generated UMAP plots to visualize the clustering results. The clusters were color-coded based on the Leiden clustering labels. The plot is saved as `_leiden_clusters.png`.
- **Marker Gene Identification:** Differential expression analysis was conducted to identify marker genes for each cluster. The top 25 marker genes were plotted and saved as `_marker_genes.png`.

7. Submission Preparation

- **Cluster Labelling:** The final cluster labels were adjusted to start from 1, as required for submission.
- **Submission File:** Created a CSV file with the format 'Id, Label', where 'Id' corresponds to the cell identifiers and 'Label' corresponds to the cluster assignments. The submission file is named `submission.csv`.

8. Conclusion

The clustering analysis was successfully conducted using Leiden clustering with optimal resolution. Visualisations and differential expression analysis provided insights into the identified clusters. The final results were prepared for submission in the required format.