Clustering scRNA-Seq Data using TF-IDF

Marmar Moussa and Ion Măndoiu

Computer Science & Engineering Department
University of Connecticut, Storrs, CT, USA
{marmar.moussa,ion}@engr.uconn.edu

Abstract. Single cell RNA sequencing (scRNA-Seq) is critical for understanding cellular heterogeneity and identification of novel cell types. We present novel computational approaches for clustering scRNA-seq data based on the TF-IDF transformation.

Introduction

In this abstract, we propose several computational approaches for clustering scRNA-Seq data based on the Term Frequency - Inverse Document Frequency (TF-IDF) transformation that has been successfully used in the field of text analysis. Empirical evaluation on simulated cell mixtures with different levels of complexity suggests that the TF-IDF methods consistently outperform existing scRNA-Seq clustering methods.

Methods

We compared eight scRNA-Seq methods, including three existing methods and five proposed methods based on the TF-IDF transformation. All methods take as input the raw *Unique Molecular Identifier (UMI)* counts generated using 10X Genomics' CellRanger pipeline [4]. *Existing scRNA-Seq clustering methods* are: the recommended workflow for the Seurat package [3], the Expectation-Maximization (EM) algorithm implemented in the mclust package [2], and a K-means clustering approach similar to that implemented in the CellRanger pipeline distributed by 10X Genomics [1]. Two types of *TF-IDF based methods* were explored. In first type of methods, TF-IDF scores were used to select a subset of the most informative genes that were then clustered with EM and spherical K-means. In the second type all genes were used for clustering, but the expression data was first binarized using a TF-IDF based cutoff. The binary expression level signatures were clustered using: hierarchical clustering with Jaccard distance, and hierarchical clustering with cosine distance with or without an additional cluster aggregation step.

Experimental setup and results

To assess accuracy we used mixtures of real scRNA-Seq profiles generated from FACS sorted cells [4]. We selected five cell types: CD8+ cytotoxic T cells (abbreviated as C), CD4+/CD45RO+ memory T cells (M), CD4+/CD25+ regulatory

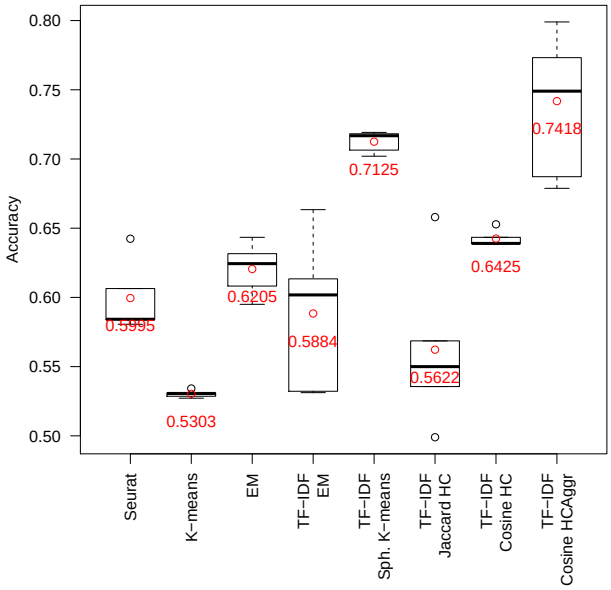

Fig. 1. Accuracy for the B:R:H:M:C datasets with 1:1:1:1:1 ratio.

T cells (R), CD4+ helper T cells (H), and CD19+ B cells (B). We generated mixtures comprised of 5,000 cells sampled from all five cell types in equal proportions. Box-plots of classification accuracy achieved by the eight compared methods are shown in Figure 1. TF-IDF based hierarchical clustering with cosine distance and cluster aggregation performs better than all other methods, with a mean accuracy of 0.7418, followed by the TF-IDF based spherical K-means, with a mean accuracy of 0.7125.

Acknowledgements

This work was partially supported by NSF Award 1564936 and a UConn Academic Vision Program Grant.

Bibliography

[1] Cell Ranger R Kit Tutorial, http://s3-us-west-2.amazonaws.com/10x.files/code/cellrangerrkit-PBMC-vignette-knitr-1.1.0.pdf
[2] Fraley, C., Raftery, A., Murphy, T., Scrucca, L.: mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. 2012. University of Washington: Seattle
[3] Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., Regev, A.: Spatial reconstruction of single-cell gene expression data. Nature biotechnology 33(5), 495–502 (2015)
[4] Zheng, G.X.Y., et al.: Massively parallel digital transcriptional profiling of single cells. Nature Communications 8, 14049 (2017)