[BIOINFO 2023]

Title: Single Cell Data Analysis using Hierarchical Clustering Based on PCA with Fuzzy C-Means

Vikas Singh[1], and Sunjae Lee[2]

[1, 2] *School of Life Sciences, Gwangju Institute of Science and Technology (GIST)*
**Corresponding author: vikkyak07@gist.ac.kr*

Abstract: Unsupervised clustering of single-cell RNA sequencing data is essential in disclosing the biological information to identify the complex cellular states and tissue composition. In this work, we have tried to develop a novel method based on agglomerative clustering that combines principal component analysis (PCA) with fuzzy c-means clustering (FCM) to generate a cell state hierarchy. The data is transformed using the PCA, and the initial clusters of the FCM are associated with the number of principal component variations. The FCM provides fewer clusters and combines in an agglomerative way to distinguish cell states. The present approach is validated on the single-cell RNA-Seq dataset with 300 cells whose transcriptional measurements are driven across 8,686 distinct genes [1]. This dataset is obtained from 11 different cell types, i.e., K562 – myeloid (chronic leukemia), HL60 – myeloid (acute leukemia), CRL-2339 – lymphoblastoid; iPS – pluripotent; CRL-2338 – epithelial, BJ – fibroblast (from human foreskin), Kera – foreskin keratinocyte; NPC – neural progenitor cells, GW(16, 21, 21+3) – gestational week (16,21, 21+3 weeks), fetal cortex [2]. These cell types can also be classified into four different disparate tissues: blood, skin, stem, and neural tissues. We are describing these problems as the cell-level and tissue-level classifications and using the levels as actual classes in our performance measurement; i.e., we have focused on data partitions with K=11 (cell-level) and K=4 (tissue-level) clusters. The present approach is also compared with pcaReduce [2] and Hierarchical clustering [3] regarding performance assessment.

References:

1. Pollen, Alex A., Tomasz J. Nowakowski, Joe Shuga, Xiaohui Wang, Anne A. Leyrat, Jan H. Lui, Nianzhen Li, et al. "Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex." Nature Biotechnology 32, no. 10 (2014): 1053-1058.
2. Žurauskienė, Justina, and Christopher Yau. "pcaReduce: hierarchical clustering of single-cell transcriptional profiles." BMC Bioinformatics 17 (2016): 1-11.
3. Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. Sincera: A pipeline for single-cell rna-seq profiling analysis. PLoS Comput Biol. 2015; 11(11):e1004575.