Effective Clustering of scRNA-seq Data to Identify Biomarkers without User Input

Hussain A. Chowdhury
Department of Computer Engineering and Engineering, Tezpur University, India, 784028
hachowdhury09@yahoo.in

Abstract

Clustering unleashes the power of scRNA-seq through identification of appropriate cell groups. It is considered a prerequisite to performing differential expression analysis, followed by functional profiling to identify potential biomarkers from scRNA-seq data. Most existing clustering methods either integrate cluster validity indices or need user assistance to identify clusters of arbitrary shape. We develop two clustering methods 1) UIFDBC to identify clusters of arbitrary shapes, 2) UIPBC to cluster scRNA-seq data. Neither method integrates a cluster validity index nor takes any user input. However, specialised approaches are used to benchmark the parameters. Both approaches outperform state-of-the-art methods.

Introduction

Evolution of next generation sequencing, in terms of RNA-seq and scRNA-seq, has enabled researchers to capture proteins produced by genes, i.e., gene expressions at particular times both at tissue and cellular levels, facilitating the study of the behaviour of genes and discover of causality, considering different phenotypic variations and diseases (Chowdhury, Bhattacharyya et al. 2020b). scRNA-seq provides new opportunities to understand biological processes at the cellular level through identification of cell groups and characteristics of cell heterogeneity. Identification of potential biomarkers from scRNA-seq is another important task, which can be performed using differential expression analysis on the basis of detected cell groups, followed by functional profiling, topological and integrative analyses. In scRNA-seq, cell groups may be of arbitrary shapes, sizes and densities, which researchers attempt to discover using density-based clustering (Hu, Kim et al. 2019).

Most clustering methods applied to scRNA-seq are developed for other tasks. But, there are a few which are either especially developed for scRNA-seq or adapted from other application. However, most methods require prior preprocessing to correct biases that arise due to droupout events, low read counts, varying transcript sizes and data distributions, zero inflation and noise. The performance of these clustering methods depends on user input, tight integration of cluster quality indices, and quality of the provided data (Chowdhury, Bhattacharyya et al. 2020a). To the best of our knowledge, ClusterDv (Marques and Orger 2019) is the only density-based clustering method which neither integrates cluster validity index nor takes user input to discover clusters of arbitrary shapes and sizes. However, it requires high execution time and memory which limits its application on scRNA-seq. A few other methods developed for scRNA-seq clustering include SC3 (Kiselev, Kirschner et al. 2017), Seurat (Satija, Farrell et al. 2015), CIDR (Lin, Troup, and Ho 2017) and RaceID (Grün, Lyubimova et al. 2015). Except CIDR and RaceID, the other clustering methods require prior external preprocessing. CIDR is the only method which does not require user input. Although especially developed for scRNA-seq, the clustering results produced by these methods are still poor and need further improvement. Detailed background of machine learning and statistical methods on scRNA-seq can be found in (Petegrosso, Li et al. 2020).

Plan of Research

Approaches used in computational biology and bioinformatics have become more accurate than routine wet-lab experiments in the quest to identify biomarkers (Lin, Qian et al. 2019). In order to identify potential biomarkers from scRNA-seq, identification of cell groups is the most crucial step. So, the main objective the author's Ph.D. thesis is to address the limitations of state-of-the-art clustering methods developed for scRNA-seq and to expedite the automation of biomarkers identification, making it user input free. In order to achieve this objective, the thesis plans to carry out following sub-tasks. The proposed workflow of the Ph.D. work can be viewed in Figure 1.

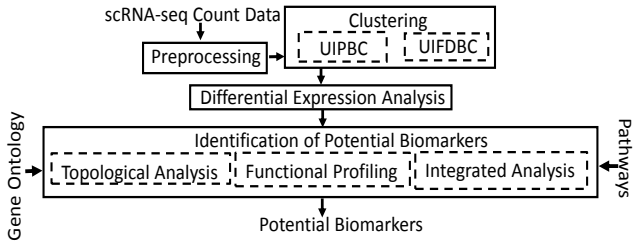

Figure 1: Proposed workflow of the author Ph.D. thesis.

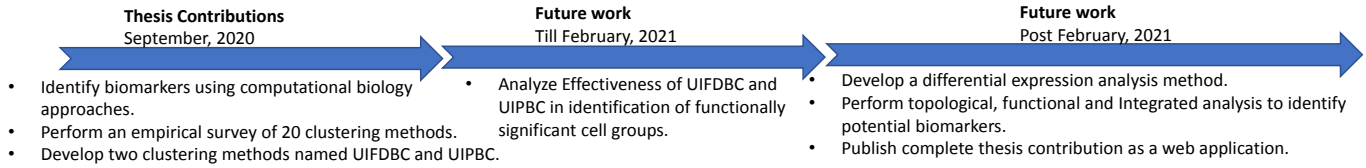

Figure 2: Brief timeline of my research

- Develop an effective density based clustering method, which can detect clusters of arbitrary shapes and sizes without taking any user input.
- Develop a specialised clustering method to identify cell groups from scRNA-seq without user input.
- Develop a differential expression analysis method to identify differentially expressed genes (DEGs).
- Develop a web tool to identify potential biomarkers for a range of diseases using functional profiling, topological analysis, and integrated data analysis of identified DEGs.

Thesis Contributions

The following milestones have been accomplished towards the completion of the thesis.

- Completed literature review and developed a pipeline which can generate count data from raw RNA-seq and scRNA-seq FASTQ file for input.
- Developed a pipeline to perform preprocessing of scRNA-seq count data using common preprocessing techniques.
- Evaluated the performance of 20 state-of-the-art clustering methods for scRNA-seq to understand their effectiveness.
- Developed a fast and effective user-input-free density-based clustering method named UIFDBC[1] to discover arbitrary shaped clusters without integrating a cluster validity index. The method was tested and verified on 15 synthetic and real world datasets. The performance of the method is consistently high, and it outperformed 8 recent state-of-the-art density-based clustering methods.
- Developed a specialised clustering method named UIPBC[2] for scRNA-seq data to identify cell groups. The method integrates required preprocessing steps for data cleaning, followed by a robust clustering of scRNA-seq data. The method does not require any user input. After comparing the proposed method with 12 recent scRNA-seq clustering methods on 12 datasets, we observed that UIPBC is able to perform consistently well on all datasets, being the best performer in 85%, 89%, 77% and 58% in exhaustive comparison experiments performed, in terms of NMI, Purity, ARI and running time.

[1]UIFDBC: An Effective Density Based Clustering to Find Clusters of Arbitrary Shapes without User Input (to be published)

[2]UIPBC: An Effective Clustering for scRNA-seq Data Analysis without User Input (to be published)

Future Work

In order to achieve the main goal of the Ph.D. thesis, the following sub-tasks remain. A brief timeline of the research is given in Figure 2.

- Develop a method to identify DEGs across cell groups.
- Perform functional profiling on identified DEGs to understand the effectiveness of UIPBC in identification of functionally significant cell groups.
- Identify potential biomarkers based on topological analysis, functional profiling and integrated analysis.
- Develop an automated web application, which will accept scRNA-seq count data in input and give a list of potential biomarkers in output.

References

Chowdhury, H. A.; Bhattacharyya, D. K.; et al. 2020a. (Differential) Co-Expression Analysis of Gene Expression: A Survey of Best Practices. *IEEE/ACM TCBB* 17(4): 1154–1173.

Chowdhury, H. A.; Bhattacharyya, D. K.; et al. 2020b. Differential Expression Analysis of RNA-seq Reads: Overview, Taxonomy, and Tools. *IEEE/ACM TCBB* 17(2): 566–586.

Grün, D.; Lyubimova, A.; et al. 2015. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* 525(7568): 251–255.

Hu, M.-W.; Kim, D. W.; et al. 2019. PanoView: An iterative clustering method for single-cell RNA sequencing data. *PLoS Computational Biology* 15(8): e1007040.

Kiselev, V. Y.; Kirschner, K.; et al. 2017. SC3: Consensus clustering of single-cell RNA-seq data. *Nature Methods* 14(5): 483–486.

Lin, P.; Troup, M.; and Ho, J. W. 2017. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biology* 18(1): 59.

Lin, Y.; Qian, F.; et al. 2019. Computer-aided biomarker discovery for precision medicine: Data resources, models and applications. *Briefings in Bioinformatics* 20(3): 952–975.

Marques, J. C.; and Orger, M. B. 2019. ClusterDv: A simple density-based clustering method that is robust, general and automatic. *Bioinformatics* 35(12): 2125–2132.

Petegrosso, R.; Li, Z.; et al. 2020. Machine learning and statistical methods for clustering single-cell RNA-sequencing data. *Briefings in Bioinformatics* 21(4): 1209–1223.

Satija, R.; Farrell, J. A.; et al. 2015. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* 33(5): 495–502.