1 Title

Using “whole organism” single cell datasets to perform a comparative analysis of clustering methods.

2 Group Members

Justin Do and Kaki Ryan

3 Abstract

Our project will provide a comparative analysis of clustering methods in the biological context, using datasets provided by the Seattle Organismal Molecular Atlases (SOMA) [1]. We will study five clustering algorithms: K-means, spectral, nearest neighbors, hierarchical clustering and multivariate Gaussian mixture and evaluate how they perform relative to each other in inferring biologically interpretative structure from the single-cell data.

4 Formal Statement of the Problem

Clustering algorithms typically involve many parameters, operate in high dimensional spaces and are dealing with noisy or incomplete data. For these reasons choosing a suitable clustering method for a given dataset or problem is a nontrivial task and crucial for achieving good results. We will be trying to identify what factors are most important to take into account when clustering with single cell datasets for a whole organism.

5 Related Work

There has been extensive work in the realm of using real world data sets for clustering analysis. For example, in [5] the authors aim to evaluate different methods applied to several large credit risk and bankruptcy data sets. Similarly in [6], a comparative evaluation was done on these same methods in the context of speaker-verification tasks. Other work has been done that is not domain-specific, too, such as in [4] and [7].
It will also be interesting to read about the papers that came out of experiments done on the datasets we plan to work with. In [2], the authors use single-cell combinatorial indexing assay for transposase accessible chromatin with sequencing (sci-ATAC-seq) to evaluate the chromatin of *Drosophila melanogaster* (common fruit fly) nuclei at various lengths of time following the egg laying. It was found that spatially distinct cell populations can be observed in clusters after just 2-4 hours, and after 6-8 hours cell types can be inferred from chromatin accessibility. We expect that these data will be an interesting litmus test to compare our methods because it seems as if clustering on gene expression will perform better as the embryos age. In [3], the authors use single-cell combinatorial indexing RNA sequencing (sci-RNA-seq) to profile cells from *Caenorhabditis elegans* (roundworm) larvae. *C. elegans* is unique in the sense that it is the only multicellular organism with all cells and cell types defined. Additionally in [2], the authors state that *Drosophila* endoderm and mesoderm bear resemblance to that of *C. elegans*. Thus, it may be possible that similar patterns may appear across the two datasets that would be accessible for us to observe.

6 Contributions

Our main contributions and goals are to show what clustering methods perform the best on these datasets and why. We will be performing a quantitative analysis of the strengths and weaknesses of the various clustering methods applied to single-cells datasets. It would be nice if were able to provide (or replicate) any insights about the methods that are more generalize-able.

7 Datasets

We plan to use the fly and worm data from the Seattle Organismal Molecular Atlases (SOMA) [1].

8 Intended Experiments

We plan to cluster the fly and worm data using the following methods: K-means, spectral, nearest neighbors, hierarchical clustering and multivariate Gaussian mixture.
We will evaluate each method on the quality of the clusters formed using metrics like NMI, area under the ROC curve and cosine similarly. We will also measure the scalability of each approach by taking performance measurements, comparing the run-times of the algorithms with different numbers of cells and features.

9 Expected Challenges

One challenge will be identifying meaningful biological implications (or lack thereof) in our results. Neither of us have a strong biological background, so doing some reading on and background research on things to be on the lookout for with these particular organism will be important.

10 Implementation

At the end we will provide a Github repo with all of our code and documentation. The inputs will be the pre-processed fly and worm datasets (need to add more here) We would provide the outputs of each of the methods individually, providing some plots/graphics and in addition to metrics capturing how well it performed (ex: normalized mutual information, area under the ROC curve) We will also provide the code for performing the different runtime experiments and the associated charts we produce.

11 Preliminary Results

As of right now, we don't have any preliminary results to share. The proposed timeline for the remainder of our project is as follows:
03/14/21: Get comfortable with the 2 datasets by reading the papers, tutorials and documentations provided by the authors.
03/21/21: Complete any necessary pre-processing of the fly and worm data.
03/28/21: First two clustering methods done.
04/14/21: All experiments done.
04/21/21: First draft of project paper done, including the comparisons between the different methods.

References

[1] https://atlas.gs.washington.edu/hub/

[2] Cusanovich, D., Reddington, J., Garfield, D. et al. The cis-regulatory dynamics of embryonic development at single-cell resolution. Nature 555, 538–542 (2018). https://doi.org/10.1038/nature25981

[3] Cao J, Packer JS, Ramani V, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science. 2017;357(6352):661-667. doi:10.1126/science.aam8940

[4] Abu Abbas, Osama. (2008). Comparisons Between Data Clustering Algorithms. Int. Arab J. Inf. Technol.. 5. 320-325.

[5] Kou G, Peng Y, Wang G. Evaluation of clustering algorithms for financial risk analysis using MCDM methods. Information Sciences. 2014;275:1–12.

[6] Kinnunen T, Sidoroff I, Tuononen M, Fränti P. Comparison of clustering methods: A case study of text-independent speaker modeling. Pattern Recognition Letters. 2011;32(13):1604–1617.

[7] Rodriguez MZ, Comin CH, Casanova D, et al. Clustering algorithms: A comparative approach. PLoS One. 2019;14(1):e0210236. Published 2019 Jan 15. doi:10.1371/journal.pone.0210236