Exploration into Single-cell RNA sequencing and Velocity under Different Dimensionality Reduction Approaches

01883444

Compiled: June 12, 2024

Github Repo: https://github.com/1883444DataScience/Exploration-into-Single-cell-RNA-sequencing-and-Velocity.git

1 Project Description

Single-cell RNA sequencing (scRNA-seq) is a powerful technology that allows us to profile the whole transcriptome of a large number of individual cells. However, it also brings many challenges for data analysis. One of them is the curse of dimensionality, since scRNA-seq produces datasets with high dimension in the numbers of cells and genes. Not all genes are informative for the particular tissue sample and the biology question at study. Besides feature selection techniques, we can reduce the dimensions of single-cell data with dimensionality reduction algorithms. Effective dimensionality reduction helps the data analysts understand the dataset through visualization. And a good dimensionality reduction algorithm should preserve both the local and global structures of the original dataset as much as possible. On the another hand, while the dataset sizes continue to grow in this field, efficiency is also quite crucial. In this project, we compare three widely used dimensionality reduction algorithms, namely Principal component analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) on three single-cell datasets which are publicly available. In accordance with previous studies, we found t-SNE and UMAP consistently deliver more meaningful representations than linear methods like PCA. And UMAP is significantly faster than t-SNE as the dataset grows.

2 Assessment Criteria

Technical Competence: Proficiency in data collection, processing, analysis, and coding.

- Python, Jupyter Notebook, ScanPy, AnnData, NumPy, Pandas, Matploblib and more tools and libraries are used in this project. Most of them I learned from scratch.
- To properly process and analyze the dataset, I also studied related biology and single-cell sequencing technologies.

User Interface: Design, functionality, and usability of the final data product.

- Jupyter Notebook is used in this project since it is a interactive web-based Python programming environments. Compared to data analysis with scripts, visualizations and markdown cells are embedded into the notebook for better interactivity and user experience. I get to know this tool from the Kaggle community.

Analysis and Interpretation: Depth of analysis, appropriate use of statistical methods, and meaningful interpretation.

- The design of this study follows a 2019 paper published in Nature Biotechnology. Due to some data availability issues, we use different datasets, which are easier to get access to. We compare three widely-adopted dimensionality reduction algorithms in the field of single-cell analysis, and draw conclusions in accordance with previous studies.
- By visualizing the representations produced by these algorithms, we gain interesting insights about different cell types and their gene expression profiles.

Presentation and Communication: Clarity, organisation and effectiveness of written and visual communication.

- The data analysis workflow and results are well presented in Jupyter Notebooks. Besides comments, we also add markdown cells of background information, to help the user better understand the contexts.

Reproducibility: Clarity and completeness of documentation for result reproducibility.

- A README file is included to help the use setup the Python environment used in this project.
- The code is well documented by comments and markdown cells in Jupyter Notebooks.
- Random seeds are controlled for reproducibility.

Version Control: Effective use of version control systems.

- Git and Github are used in this project for version control and sharing code, data and results.

3 Project Reflection

Reflect on the experience of creating your data product. In 6 bullet points and at most 1 page total, summarise the following.

- *3 things you have learned as part of this process,*
- *2 aspects of the project that you found challenging or would approach differently with hindsight,*
- *1 aspect of the project that you would like to learn more about in the future.*

You may delete this italicised text when filling in the template.

Learnings:

- Besides the statistical and computational tools at hand, it is very important for the data analyst to acquire enough domain knowledge and understand the data well. Only by this way, data analysis will result in meaningful insights (biological insights, in this project).
- Different algorithms and methods has various characteristics, in terms of effectiveness, computational efficiency, robustness. It is important to choose the algorithm wisely depending on the particular scenario, for example, how large the dataset is and how much computational resources I have.
- Another thing I learn from this project is how to produce visualizations of high quality. This is very important for effective communication.

Challenges:

- The most challenging thing in this project is to learn biology contexts used in this project. Fortunately, there are many high-quality articles and tutorials on the Internet.
- Another challenge is the data preprocessing workflow and the choice of hyperparameters for some algorithm. It takes a lot of iterations back and forth.

Further Development:

- Single-cell analysis is still a very active field, which many interesting and challenging open problems to solve. For example, trajectory inference, differential gene expression analysis and regulatory network inference are all intriguing research topics for me to explore.