Thesis Proposal

Trajectory Inference for Single Cell data

Submitted by
Nazifa Tasnim Hia
Institute of Information Technology
University of Dhaka

Supervised by
Dr. Sumon Ahmed
Associate Professor
Institute of Information Technology
University of Dhaka

Institute of Information technology
University of Dhaka

22th July, 2022

Thesis Proposal

Trajectory Inference for Single Cell data

Author

Nazifa Tasnim Hia
MSSE 0922
Institute of Information Technology
University of Dhaka
Email: bsse0930@iit.du.ac.bd

Supervisor

Dr. Sumon Ahmed
Associate Professor
Institute of Information Technology
University of Dhaka
Email: sumon@iit.du.ac.bd

Table of Contents

List of Figures

List of Tables

1. Introduction

In recent years, the field of functional genomics is advancing very rapidly, where the development of single-cell (sc) assaying techniques is playing a very crucial role. The analysis of single-cell trascriptomics has gained widespread popularity because of its ability to uncover complex biological systems ranging from the microbial ecosystem to the human cancer genomics (Gawad et al., 2016). In single-cell technology expression profile of each cell is measured individually, in contrast to the bulk RNA-seq technology that averages gene expression across a cell population when measuring transcriptomes. When transcriptomes are averaged throughout a cell population, this approach cannot capture the transcriptional heterogeneity present in individual cells. Researchers have found that by studying organisms at the single-cell level, they can get more precise answers to numerous issues (see e.g. Gawad et al., 2016; Hwang et al., 2018). For instance, during the process of differentiation, individual cells determine their own destinies in response to signals received from other cells and external cues. Moreover, all cells in a population do not develop at the same rate, similar alterations in transcriptomes might be observed at various time scales. Accordingly, the actual picture of the developmental and differential processes at the cellular level is not captured by averaging the expression profiles across a population of cells in a bulk study.

Examining expression profiles at single-cell level promises a better understanding of cell function at the genetic and cellular levels (Pasquini et al., 2021). It allows researchers to analyze cellular heterogeneity and transcriptional dynamics at the single-cell level (Qi et al., 2021). Recent developments allow the gene experssion to be profiled in tens of thousands of cells at single-cell level. In some cases, the captured cells are undergoing through developmental and differentiation processes. For instance, tissue development, cellular differentiation, tumor growth, cell cycle, and other biological processes go through transcriptomic phases in a cell-specific manner. Therefore, there is an underlying order of cells (Borcherding et al., 2019; Cacchiarelli et al., 2018; Loeffler-Wirth et al., 2018) behind these transcriptomic phases that has largely gone unexplored. Moreover, all cells are not in the same stage of a biological process, resulting in cell-to-cell variation in gene expression profiles. With single-cell RNA-Seq, researchers can observe this continuum of transcriptomic changes via analysis of individual cells. But there is no time label in the data, this

information is lost at single cell capturing process. Therefore, there is a need of pseudotemporal ordering of cells in order to investigate the gene expression dynamics. The principle behind the pseudotemporal ordering at single-cell level is that a sample of cells corresponds to a time series, where each cell represents a distinct time point along the psedotime trajectory, representing the individual cell's progress through a process of interest.

2. Motivation:

Single-cell data are generally high dimensional. A moderately large dataset may contain 10-15 thousands cells and 30-50 thousands genes (Hwang et al., 2018). Moreover, the analysis of single-cell data is challenging due to inherent biological and technical noise. The common confounding factors are differences in capturing process, sequencing depth, and the occasional failure of capture (Chen & Zhou, 2017). As cells are captured individually, single-cell data is largely made up of 0 readings. In most cases, missing information is indicated by the presence of a 0 in a dataset, but in case of single-cell data, these 0s are instructive, and it is difficult to tell which 0s are essential and which 0s are simply missing. Researchers who specialize in bioinformatics typically focus on capturing highly variable, highly expressed genes. However, it is not sufficient to simply consider variables with large variance or expression to be informative (M Ascensión et al., 2022).

Therefore, it is often useful to reduce the dimensionality of single-cell data via feature selection or dimension reduction algorithms. There are a number of cell type identification methods for single cell data, including SingleCellNet (Tan and Cahan, 2019), ACTINN (Ma and Pellegrini, 2020), Moana (Wagner and Yanai, 2018), and scPred (Alquicira-Hernandez et al., 2019) which used feature selection prior to training phase. This review paper (Theunissen, 2021) conducted an experiment and found that selecting features improved classification accuracy.

The most Popular methods for dimensionality reduction that have been applied to single-cell data are linear methods such as Principal and Independent Components Analysis (P/ICA) (Trapnell et al., 2014; Ji and Ji, 2016), as well as non-linear methods such as t-stochastic neighbourhood embedding (tSNE) (Maaten and Hinton, 2008; Becher et al., 2014), diffusion maps (Haghverdi et al., 2015, 2016), Uniform manifold approximation and projection (UMAP) (McInnes et al., 2018;

Cao et al., 2019). In some cases the dimension is reduced to a single dimension called *pseudotime* which represents the trajectory of cells undergoing some dynamic process such as differentiation or cell division. The trajectory may be linear, branching or even cyclic depending on the underlying process (Ahmed et al., 2019).

Pseudotime inference or trajectory inference is the process of positioning cells along the trajectory that quantifies the relative activity or progression of the underlying biological process. Single cell trajectory analysis is a way to sort cells along a pseudotime axis using a computational approach. Using trajectory analysis, it is possible to make assumptions about the outcomes at different periods in time by adjusting for the group's behavior. For instance, normal cellular function and potential perturbations that might give rise to physiological disorders can be determined with a better understanding of the transcriptional dynamics that controls fundamental, dynamic biological processes. Time series experiments, wherein the characteristics of individual cells are monitored over time, are ideal for such studies (Campbell & Yau, 2016). Therefore, there is a growing need of developing trajectory inference methods for single cell data that can mimic the progression dynamics of biological process under consideration.

Figure 1 helps to understand the whole process from capturing cells' expression to pseudotime estimation to trajectory inference. The idea behind these methods is that each single-cell RNA sequencing experiment is a time series in which each cell represents a different point in time along a continuum that shows how much time has passed *Figure 1*(A). During the single cell capture process, the true temporal label that tells what stage the cell is in is lost *Figure 1*(B). This means that these parameters become latent, unobserved quantities that must be statistically inferred from the collection of single cell expression profiles *Figure 1*(C). The pseudotimes

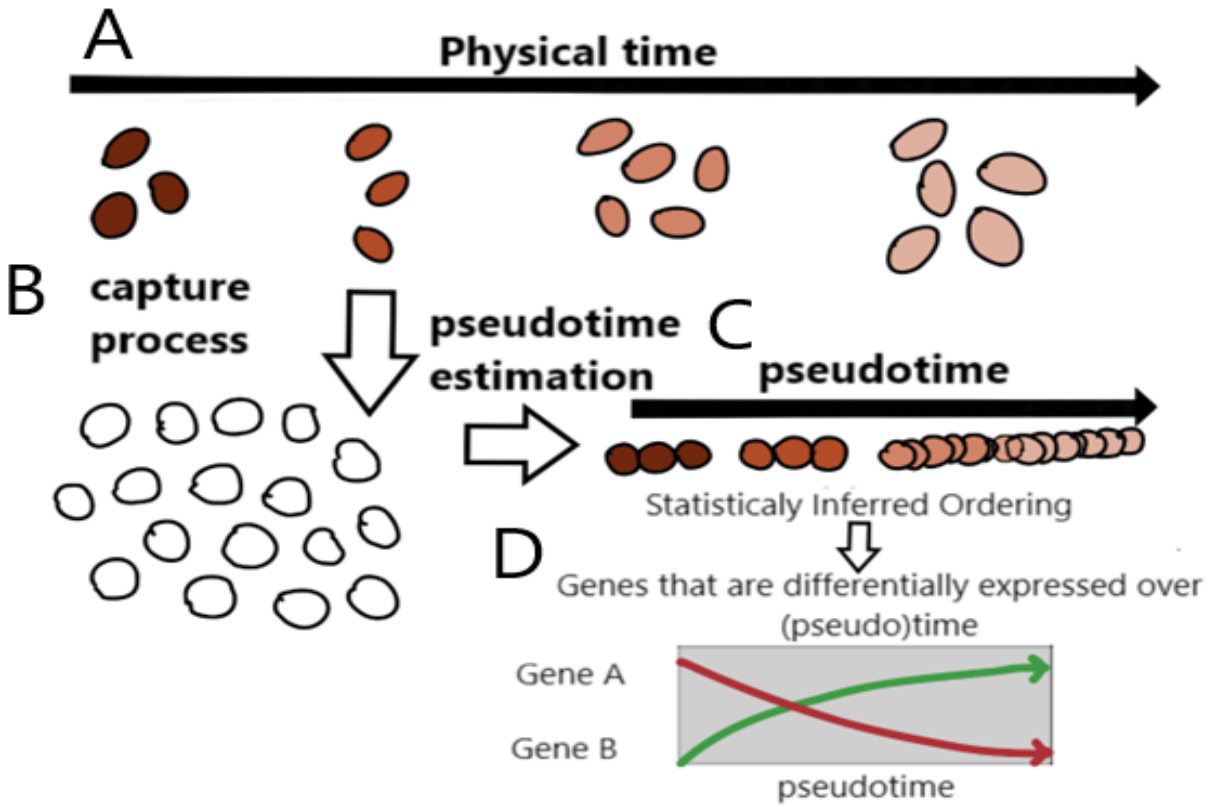


Figure 1: The problem of estimating pseudotime for a single cell. (A)Single cells at different stages in the process of growing and changing. (B)When a single cell is captured, the information about how it was labeled in time is lost. (C) Statistical pseudotime estimation algorithms try to figure out how the cells should be ordered in terms of time. (D) The estimates of pseudotime can be used to find genes whose expression changes over pseudotime. (redrawn from (Campbell & Yau, 2016)*.*

could then be used to find genes that are expressed differently across pseudotime *Figure 1*(D), which would show how transcription programming has changed over time. (Campbell & Yau, 2016)

3. Issues in the State-of-Art Approaches:

Different formalisms are used to represent a pseudotime trajectory. Methods available in the literature concentrate mostly on dimension reduction followed by mapping of cells to a trajectory. Dimensionality reduction is followed by the construction of a minimal spanning tree (Ji & Ji, 2016), principal curve (Street et al., 2018), or reverse graph embedding (Trapnell, 2018) to discover a primary tree in the data and generate a pseudotime route. There are also probabilistic

methods to infer pseudotemporal ordering of cells within the Bayesian framework that considers associated uncertainty in pseudotime estimation (Campbell and Yau, 2016; Reid and Wernisch, 2016; Ahmed et al., 2019).

Currently available pseudotime construction methods rely heavily on dimensionality reduction to extremely low dimensional subspaces, which result in data loss. The accuracy of a method may be affected by the dimensionality reduction method employed in the initial phase, as well as the amount of information loss when transforming original data to lower dimensions. More recently, an evolutionary approach based method PseudoGA (Mondal et al., 2021) is developed for trajectory inference using single-cell RNA-seq data. PseudoGA is a genetic algorithm-based technique for arranging cells. It was built on the premise that changes in gene expression occur along a smooth curve in pseudotime (Mondal et al., 2021). PseudoGA was designed to take advantage of this relationship. Due to the fundamental underlying assumption that it operates under and its independence from dimensionality reduction approaches, the pseudoGA method is a trustworthy tool for calculating pseudotime from single cell transcriptome data (Mondal et al., 2021). However, being a genetic algorithm based approach, PseudoGA does not provide any real value for pseudotime point rather only the ordering of cells. This limitations may hinder the strength of the developed methods. For instance, in some cases, there is some capture time information available which is found informative in literature (Reid and Wernisch, 2016; Ahmed et al., 2019).

4. Aims & Objectives:

The Aim is to provide a trajectory inference method that can use the real-world measurements of gene expression. As capture time information is found informative, the method will be able to incorporate it whenever available. A feature selection method can also be investigated to find out the most relevant genes.

So the aims and objectives in brief -

- Developing a trajectory inference method for single-cell data.
- Incorporating capture time information in the inference process.
- Investigate feature selection techniques for single-cell data.

5. Research Methodology

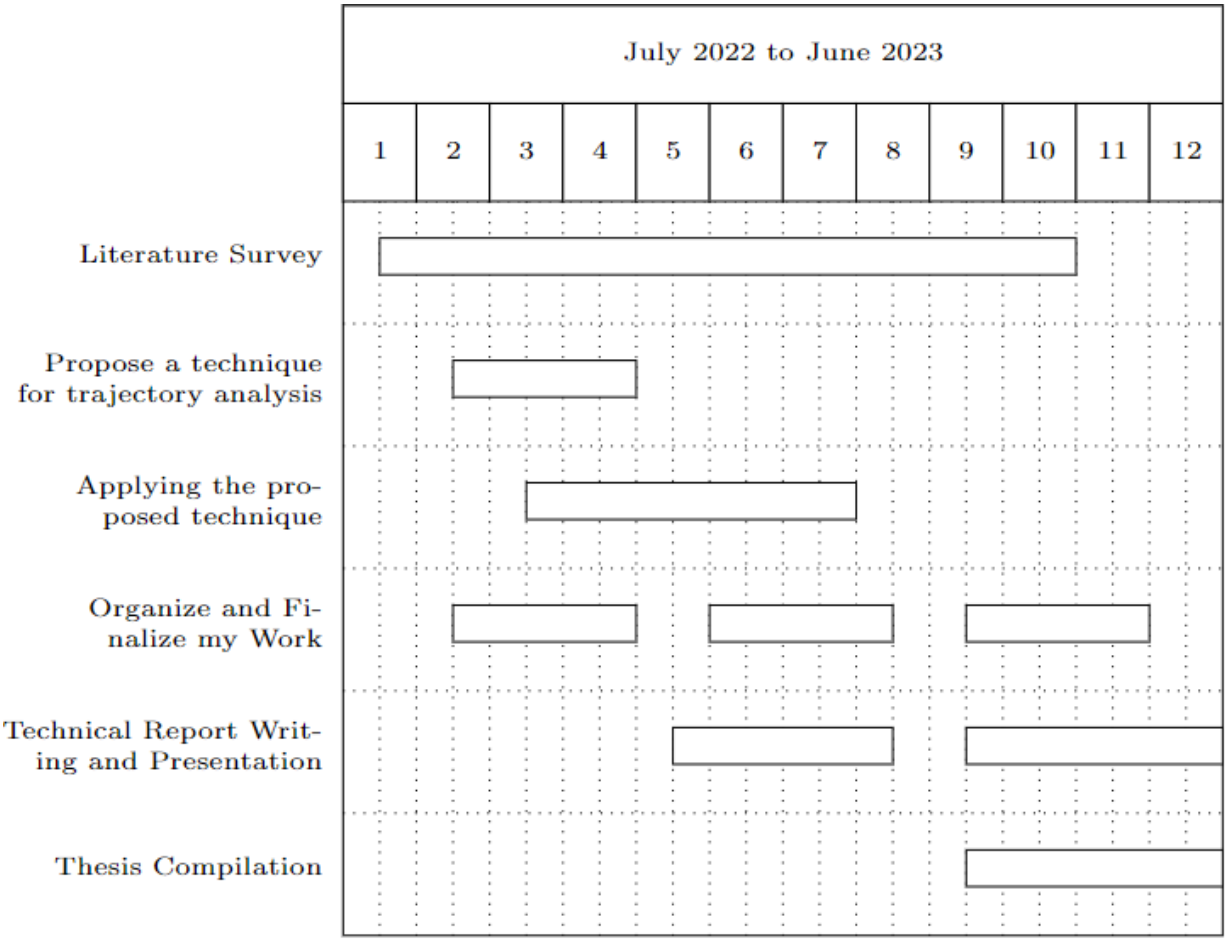
The primary objective of this research is to offer a new trajectory inference criterion to address the aforementioned issue with pseudotime estimation. In order to accomplish this, Table 1 provides a list and brief description of research efforts.

Table 1: Research methodology

Step No	Title of Activity	Activity Description & Relation to Research Question
1	Literature Survey	Reviewing the existing literature and building the background of the research Writing a research proposal Presenting initial ideas and receiving feedback
2	Study existing inference techniques	Concentrating on a more specific area. This assignment will cover exhaustively the similar or relevant works to our field of interest. Reporting survey results
3	Developing an approach and apply	Determine the limitations of previous studies. Try to change the present one or discover a new inference method. Implementation of the procedure.
6	Technical Report Writing	At the end of each major step, one technical report will be delivered
7	Publications & Presentations	As a pre-requisite for Master's degree, a thesis will be compiled by the end

6. Research Timeline:

Table 2: Research Timeline.



References

1. Ahmed, S., Rattray, M., & Boukouvalas, A. (2019). GrandPrix: Scaling up the Bayesian GPLVM for single-cell data. *Bioinformatics*, *35*(1), 47–54. https://doi.org/10.1093/bioinformatics/bty533
2. Borcherding, N., Voigt, A. P., Liu, V., Link, B. K., Zhang, W., & Jabbari, A. (2019). Single-cell profiling of cutaneous T-cell lymphoma reveals underlying heterogeneity associated with disease progression. *Clinical Cancer Research*, *25*(10), 2996–3005. https://doi.org/10.1158/1078-0432.CCR-18-3309
3. Cacchiarelli, D., Qiu, X., Srivatsan, S., Manfredi, A., Ziller, M., Overbey, E., Grimaldi, A., Grimsby, J., Pokharel, P., Livak, K. J., Li, S., Meissner, A., Mikkelsen, T. S., Rinn, J. L., & Trapnell, C. (2018). Aligning Single-Cell Developmental and Reprogramming Trajectories Identifies Molecular Determinants of Myogenic Reprogramming Outcome. *Cell Systems*, *7*(3), 258-268.e3. https://doi.org/10.1016/j.cels.2018.07.006
4. Campbell, K. R., & Yau, C. (2016). Order Under Uncertainty: Robust Differential Expression Analysis Using Probabilistic Models for Pseudotime Inference. *PLoS Computational Biology*, *12*(11), 1–20. https://doi.org/10.1371/journal.pcbi.1005212
5. Chen, M., & Zhou, X. (2017). Controlling for Confounding Effects in Single Cell RNA Sequencing Studies Using both Control and Target Genes. *Scientific Reports*, *7*(1), 1–14. https://doi.org/10.1038/s41598-017-13665-w
6. Hwang, B., Lee, J. H., & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental and Molecular Medicine*, *50*(8), 1–14. https://doi.org/10.1038/s12276-018-0071-8
7. Ji, Z., & Ji, H. (2016). TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Research*, *44*(13), e117. https://doi.org/10.1093/nar/gkw430
8. Loeffler-Wirth, H., Binder, H., Willscher, E., Gerber, T., & Kunz, M. (2018). Pseudotime dynamics in melanoma single-cell transcriptomes reveals different mechanisms of tumor progression. *Biology*, *7*(2), 1–16. https://doi.org/10.3390/biology7020023
9. M Ascensión, A., Ibáñez-SolCrossed D Sign©, O., Inza, I., Izeta, A., & Araúzo-Bravo, M. J. (2022). Triku: A feature selection method based on nearest neighbors for single-cell data. *GigaScience*, *11*, 1–16. https://doi.org/10.1093/gigascience/giac017
10. Mondal, P. K., Saha, U. S., & Mukhopadhyay, I. (2021). *PseudoGA : cell pseudotime reconstruction based on genetic algorithm*. *49*(14), 7909–7924. https://doi.org/10.1093/nar/gkab457
11. Pasquini, G., Rojo Arias, J. E., Schäfer, P., & Busskamp, V. (2021). Automated methods for cell type annotation on scRNA-seq data. *Computational and Structural Biotechnology Journal*, *19*, 961–969. https://doi.org/10.1016/j.csbj.2021.01.015
12. Qi, R., Wu, J., Guo, F., Xu, L., & Zou, Q. (2021). A spectral clustering with self-weighted multiple kernel learning method for single-cell RNA-seq data. *Briefings in Bioinformatics*, *22*(4). https://doi.org/10.1093/bib/bbaa216
13. Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E., & Dudoit, S. (2018). Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, *19*(1), 1–16. https://doi.org/10.1186/s12864-018-4772-0
14. Trapnell, C. (2018). *trajectories*. *14*(10), 979–982. https://doi.org/10.1038/nmeth.4402.Reversed