A Literature Review on Pseudotime Estimation Using Single-Cell Data

Nazifa Tasnim Hia
Institute of Information Technology
University of Dhaka
Dhaka, Bangladesh

Sumon Ahmed
Institute of Information Technology
University of Dhaka
Dhaka, Bangladesh

Abstract:- Transcriptomics and lineage tracing have enabled the development of computational methods for determining developmental trajectories, allowing for high-resolution examination of cellular dynamics. Since 2014, numerous methodologies have been proposed in the field of pseudotime estimation. Although these algorithms are potent, they are still in their infancy, and consideration must be taken when employing their strengths and weaknesses. This paper describes prominent computational approaches in depth for pseudotime estimation based on the information they utilize, as well as describing future challenges and opportunities for the development of new methods.

Keywords:- *Pseudotime Estimation; Sigle-Cell Data; Literature Review;*

I. INTRODUCTION

Functional genomics has been rapidly advancing in recent years, thanks to the development of single-cell (sc) assaying techniques. These techniques are crucial for capturing cells that are undergoing developmental and differentiation processes, such as tissue development, cellular differentiation, tumor growth, and cell cycle[1]–[3]. During these processes, cells go through transcriptomic phases in a cell-specific manner, and there exists an underlying order of cells that has not been extensively explored. Due to variations in gene expression profiles among cells in different biological process stages, it becomes necessary to infer a pseudo temporal ordering of cells. Pseudotime inference involves placing cells along a trajectory that represents the relative activity or progression of the underlying biological process. However, since the time label is lost during the single-cell capturing process, it is important to reconstruct this lost time series information, known as pseudotime. In this paper, we will discuss about some popular pseudotime estimation methods.

II. LITERATURE REVIEW

There are a variety of approaches for calculating pseudotime, including unsupervised, supervised and probabilistic techniques. Moreover, it is an emerging field. Hence, literature review is essential for comprehending this particular research subject and determining where to concentrate in order to get inspiration to discover a new strategy. Several pseudotime estimation methods are described in detail in this review.

A. Exploring Unsupervised Approaches in Depth

For Pseudotime estimation, unsupervised approaches are mainly graph and tree based. In Graph based method dimensionality reduction is followed by the construction of a minimal spanning tree [4], principal curve [5] or reverse graph embedding [6] to discover a primary tree in the data and generate a pseudotime route. The nodes of the graphs can correspond to individual cells [7] or groups of cells [4], [8] in the graph.

- *TSCAN*

TSCAN [4] is a Minimum Spanning Tree (MST) based Linear model for ordering cells. Due to the fact that single-cell data is both highly dimensional and chaotic. Dimensionality reduction is required for working with data from a single cell. Although there are a variety of Dimensionality reduction strategies available. However, each methodology has a distinct application. In TSCAN, genes with equivalent expression levels are initially clustered. The expression levels of clustered genes are then averaged across each cell and cluster. This phase assisted in decreasing the variance and leveling the values. The data still exists in a high-dimensional space. Therefore, PCA is utilized, followed by data normalization.

Then the cells with similar gene expression are clustered using mclust [9] package which is a model-based clustering approach. The number of clusters are chosen using the Bayesian Information Criterion (BIC). Though there's available opportunity for the user to select the cluster number. After that, a minimum spanning tree is constructed connecting the centers of the clusters. The path which will contain largest number of clusters of the MST is flagged as the main trajectory. The root is selected randomly by choosing any one of the two ends of the main branch. The other branches will be enumerated from the root of the trajectory.

Finally, the cells are ordered by considering three different scenarios. Each cluster's cells are projected on the same edge of that cluster. The order of cells within a cluster that are projected onto two distinct edges is determined by the order of the edges. The order of clusters determines the arrangement of cells within each cluster. As a result, a pseudotime is assigned to every cell and a trajectory is also generated.

- *Slingshot*

Slingshot [5] is a multiple lineage detection algorithm and does not require the number of lineages to be pre specified. Slingshot uses Euclidean Distance for curve fitting which may fail in a high dimensional space. Thus they use dimensionality reduction approach before applying the algorithm. A prerequisite for using this algorithm is that the data must contain clustering level. The trajectory that is going to construct is dependent on the clusters.

At first, it builds a minimum spanning tree of clusters. After that according to the specified root node the clusters are arranged in a certain sequence. A global Lineage structure is created in which all the lineages have the same starting cluster in common. However, each Lineage has its own terminating cell cluster.

Secondly, Slingshot determines the pseudotime ordering of each lineage's events. In order to accomplish this, simultaneous principal curves are utilized in order to smooth out each lineage[10]. The attribution of each cell to a lineage is improved as a result of this. The final product is a pseudo temporal ordering that is particular to each lineage, as well as assignment weights that indicate individual cells that belong to different lineages. Slingshot's capacity to distinguish between a single lineage and several lineages was demonstrated by applying it to several previously published datasets [2], [8].

- *Monocle3*

Monocole3 [11] is a Partitioned Approximate Graph Abstraction (PAGA) based method [12] that considers trajectories as a forest rather than a single tree.

At first data is projected onto the top 50 Principal components followed by normalization to reduce the noise. Then it uses the UMAP [13] the non-linear dimension reduction method. UMAP preserves the global structure of the data by placing related cell types close to each other.

Partitioned Approximate Graph Abstraction (PAGA) method is a recent approach that helps to organize the single cell data. This approach initially constructs a K-nearest neighbor graph on cells. Then it applies Louvain Community Detection method to identify communities. Then it tests each pair of community for a significance between them in terms of cells. The communities that have more links than hypothesis they remain connected. This will create one or more component. The resulting components will form trajectories. The technique of reversed graph embedding is used to organize cells into trajectories. It learns a Principal graph that matches the data and projects each cell onto it. The algorithms then require the user to designate one or more points as the tree's root node(s). The pseudotime of each cell, which is the closest distance between a cell and the graph's starting points, is then calculated.

B. *Exploring Supervised Approaches in Depth*

Supervised approaches have been used as a method for modeling developmental processes for several years now. In supervised learning, we need correctly labeled datasets. These datasets are analyzed by the model to reveal hidden patterns. After the model has been trained, it can make predictions about the results of previously unseen datasets. Supervised learning seeks to reduce the deviation of a model from the true output as much as possible.

- *Psupertime*

Psupertime [14] is a supervised logistic regression-based approach for estimating pseudotime. It actually identifies a small subset of relevant features.

In this methodology for denoising and correcting the dropout issue, data goes through pre-processing steps. At first, data are normalized in log scale. As the data has missing values and zero reads, those values are replaced with the average of the neighboring cells values. Then for further analysis, as all the values are not informative, only highly variable genes are selected. As this approach is supervised, sequential levels (day 1, day 3, etc.) are needed for corresponding cells. Not necessarily all the cells need to be leveled; this approach can work even if the subsets are labeled.

This methodology is mainly based on ordinal logistic regression, and here the orders are the sequence level. Suppose 1–K are the sequence levels. Progressively, this approach will learn a set of weighted averages of gene expression values that will indicate them separately. It can be thought of as a combination of k-1 simultaneous logistic regressions. Every one of these separates level i from level i+1...k. To increase the interpretability of the results, an informative subset of genes needs to be selected. To fulfill the goal, L1 regularization is applied. It gives non-zero coefficients for only the genes that have a significant impact on the segmentation of levels. As a result, this model gives:

- A small and therefore interpretable set of genes with non-zero coefficients.
- A pseudotime value for each individual cell

As this methodology incorporates the available time series information. It claims that there are mainly three types of transcriptomic variations at the data level. In brief, one is gene mean expression that varies with time series, another is level-specific gene expression, and lastly, there are constant mean expressions.

This model is built based on the above hypotheses, so to validate the claim, they simulated an experiment. As this is mainly a classifier. For verification, they compared this with other renowned classifiers. Those are random forest and multinomial logistic regression. At first, they simulated a dataset consisting of genes that followed the three possible profiles: time series, label-specific, and constant. They varied the proportions of these types of genes. They did 20 simulations from 20 different seeds for each triplet (time-series, label-specific, and constant). Then they trained the model and found some cofactors for important genes. Then, based on how well these values identified the real time-series genes, they constructed precision-recall curves for each classifier.

The drawbacks of these classifiers are that they extract genes that are mostly level-specific. For this reason, Psupertime outperforms them when the trained model is applied to the test data.

- *Tempora*

Tempora [15] focuses on the relationship between time series and cell types (clusters). Time series experiments are an informative resource for sorting cells from earlier to later time points. Tempora enhances the accuracy and speed of cell trajectory inference validated by reconstructing known developmental trajectories from three different data sets. This is achieved by introducing a novel use of prior biological pathway information.

The trajectory is identified based on the clusters. For drawing the edges between clusters, they need to compare and retrieve the similarity between each pair. As a preliminary step, the centroids and marker genes of the clusters are identified. Using those redundancy-based biological pathway enrichment profiles, enrichment profiles are generated for every cluster.

Now that pathways are redundant, PCA is applied to focus on the significant and non-redundant pathway information. Selected pathways are used to construct lineages in the next step. Mutual information between each cluster is also calculated along with the pathway information. Using this information, edges are drawn. For inferring the direction of the edges, temporal information from the meta data is used. The density of each time point in a cluster is identified. The clusters that have more cells from the earlier time points are assigned at the start. Similarly, the clusters, which consist of later time points, are assigned at the end. And the direction goes from the start to the end.

To validate Tempora's performance and compare it to other available methodologies, three known gold standard cell trajectories (muscle, neural cortex, and cerebellum) were manually collected. For comparing the results, all trajectories (predicted and known) were converted to graphs. Nodes represent cell types, and the directed edges represent parent-child relationships. Two performance scores were measured. Mismatch score, which is the count of the nodes and edges needed to add or remove for transforming the result to the gold standard. Another one is the harmonic mean of the precision and recall of the gold standard.

To measure the performance, HSMM (human skeletal muscle myoblast) data was curated from literature from growth to differentiation media. Human myoblasts transition through intermediate states after exiting the cell cycle before differentiating into myotubes [16], [17]. Because myoblasts have different differentiating potentials and rates, some will differentiate into myotubes while the rest will remain undifferentiated, i.e., they do not or have yet to express myogenic transcription factors such as MYOG, resulting in two possible branches from the intermediate state. The known lineage and Tempora's predicted lineage are very similar, with the exception that Tempora links the myotube cluster to the myoblasts rather than the intermediate state. As a result, Tempora's output has a mismatch score of 1, which indicates that one edge needs to be changed in order for it to comply with the benchmark. For murine cerebral cortex data, Tempora also has a low mismatch score and high accuracy.

C. Exploring Probabilistic Approaches in Depth

Probabilistic techniques have been utilized ever since the very first experiment as a way for modeling developmental processes or pseudotime estimation. Uncertainty in pseudotime or cell identification at the level of an individual trajectory branch can be estimated by probabilistic approaches.

- *GrandPrix*

GrandPrix [18] method is inspired by the DeLorean [19] approach. According to the study [20], an experiment was conducted to investigate the pseudotime estimation of Monocle. In this experiment, the pseudotime were reestimated multiple times for different random subsets of the data. They have demonstrated that the pseudotime values assigned by Monocle to the same cell can vary substantially across random subsets. This uncertainty in pseudotime assignment has motivated Probabilistic analysis techniques. It leverages cell capture time as a parameter to establish a prior distribution over the pseudotime variable. By adopting a probabilistic framework, GrandPrix can not only estimate pseudotime but also quantify the uncertainty associated with the estimation.

The core technique employed by GrandPrix is the Gaussian Process Latent Variable Model (GPLVM). The GPLVM is a non-linear probabilistic model for dimension reduction [21], [22] that has been extensively applied to single-cell data analysis. This model utilizes a Gaussian process (GP) to define a stochastic mapping between a latent space representing pseudotime and an observed space representing gene expression. A Gaussian process is a multivariate normal distribution in an infinite-dimensional space. It is characterized by a mean function and a covariance function. In the GPLVM, the mean function describes the expected mapping from the latent dimension to the observed gene expression data, while the covariance function specifies the covariance between the mappings functions evaluated at any two arbitrary points in the latent space.

III. CONCLUSION

This study's objective is to review the current state of pseudotime estimation methodologies, with a focus on articles published between 2014 and 2023. The analysis revealed that there is no single method that can effectively infer trajectories in every scenario. 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.

However, promising techniques such as probabilistic models, clustering algorithms, and deep learning-based methods were identified. Recently some supervised approaches for pseudotime estimation are developed, they use time series information. Tempora, Psupertime are such methods that work based on the relationship between time series and cell type or cell progress. Another prior study [19] showed the importance of capture time in estimating pseudotime.

The study provides a comprehensive review and identifies potential solutions and recommendations for researchers and practitioners in this field. Additional research is needed to develop more precise and robust methods to handle complex scenarios.

REFERENCES

[1]. N. Borcherding, A. P. Voigt, V. Liu, B. K. Link, W. Zhang, and A. Jabbari, “Single-cell profiling of cutaneous T-cell lymphoma reveals underlying heterogeneity associated with disease progression,” Clin. Cancer Res., vol. 25, no. 10, pp. 2996–3005, 2019, doi: 10.1158/1078-0432.CCR-18-3309.

[2]. D. Cacchiarelli et al., “Aligning Single-Cell Developmental and Reprogramming Trajectories Identifies Molecular Determinants of Myogenic Reprogramming Outcome,” Cell Syst., vol. 7, no. 3, pp. 258-268.e3, 2018, doi: 10.1016/j.cels.2018.07.006.

[3]. H. Loeffler-Wirth, H. Binder, E. Willscher, T. Gerber, and M. Kunz, “Pseudotime Dynamics in Melanoma Single-Cell Transcriptomes Reveals Different Mechanisms of Tumor Progression,” Biology (Basel)., vol. 7, no. 2, p. 23, Apr. 2018, doi: 10.3390/biology7020023.

[4]. Z. Ji and H. Ji, “TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis,” Nucleic Acids Res., vol. 44, no. 13, p. e117, 2016, doi: 10.1093/nar/gkw430.

[5]. K. Street et al., “Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics,” BMC Genomics, vol. 19, no. 1, pp. 1–16, 2018, doi: 10.1186/s12864-018-4772-0.

[6]. X. Qiu et al., “Reversed graph embedding resolves complex single-cell trajectories,” Nat. Methods, vol. 14, no. 10, pp. 979–982, 2017, doi: 10.1038/nmeth.4402.

[7]. C. Trapnell et al., “The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells,” Nat. Biotechnol., vol. 32, no. 4, pp. 381–386, 2014, doi: 10.1038/nbt.2859.

[8]. J. Shin et al., “Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis,” Cell Stem Cell, vol. 17, no. 3, pp. 360–372, 2015, doi: 10.1016/j.stem.2015.07.013.

[9]. A. Flynt and M. I. G. Daepp, “Diet-related chronic disease in the northeastern United States: A model-based clustering approach,” Int. J. Health Geogr., vol. 14, no. 1, pp. 1–14, 2015, doi: 10.1186/s12942-015-0017-5.

[10]. Trevor Hastie & Werner Stuetzle Source, “Principal Curves Author (s): Trevor Hastie and Werner Stuetzle Source : Journal of the American Statistical Association , Vol . 84 , No . 406 , (Jun ., 1989), pp . 502 Published by : American Statistical Association Stable URL : http://www.jstor.org/s,” J. Am. Stat. Assoc., vol. 84, no. 406, pp. 502–516, 2008.

[11]. J. Cao et al., “The single-cell transcriptional landscape of mammalian organogenesis,” Nature, vol. 566, no. 7745, pp. 496–502, 2019, doi: 10.1038/s41586-019-0969-x.

[12]. F. A. Wolf et al., “PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells,” Genome Biol., vol. 20, no. 1, pp. 1–9, 2019, doi: 10.1186/s13059-019-1663-x.

[13]. L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” 2018, [Online]. Available: http://arxiv.org/abs/1802.03426.

[14]. W. Macnair, R. Gupta, and M. Claassen, “Psupertime: Supervised pseudotime analysis for time-series single-cell RNA-seq data,” Bioinformatics, vol. 38, pp. I290–I298, 2022, doi: 10.1093/bioinformatics/btac227.

[15]. T. N. Tran and G. D. Bader, “Tempora: Cell trajectory inference using time-series single-cell RNA sequencing data,” PLoS Comput. Biol., vol. 16, no. 9, Sep. 2020, doi: 10.1371/JOURNAL.PCBI.1008205.

[16]. C. F. Bentzinger, Y. X. Wang, and M. A. Rudnicki, “Building muscle: molecular regulation of myogenesis.,” Cold Spring Harb. Perspect. Biol., vol. 4, no. 2, 2012, doi: 10.1101/cshperspect.a008342.

[17]. J. Chal and O. Pourquié, “Making muscle: Skeletal myogenesis in vivo and in vitro,” Dev., vol. 144, no. 12, pp. 2104–2122, 2017, doi: 10.1242/dev.151035.

[18]. S. Ahmed, M. Rattray, and A. Boukouvalas, “GrandPrix: Scaling up the Bayesian GPLVM for single-cell data,” Bioinformatics, vol. 35, no. 1, pp. 47–54, 2019, doi: 10.1093/bioinformatics/bty533.

[19]. J. E. Reid and L. Wernisch, “Pseudotime estimation: Deconfounding single cell time series,” Bioinformatics, vol. 32, no. 19, pp. 2973–2980, 2016, doi: 10.1093/bioinformatics/btw372.

[20]. K. R. Campbell and C. Yau, “Order Under Uncertainty: Robust Differential Expression Analysis Using Probabilistic Models for Pseudotime Inference,” PLoS Comput. Biol., vol. 12, no. 11, pp. 1–20, 2016, doi: 10.1371/journal.pcbi.1005212.

[21]. F. Buettner and F. J. Theis, “A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst,” Bioinformatics, vol. 28, no. 18, pp. 626–632, 2012, doi: 10.1093/bioinformatics/bts385.

[22]. S. Gu and C. H. Ru, “Design of a Novel Piezoelectric Stick-Slip Driving Nanopositioning Stage and Power Supply Circuit,” MARSS 2018 - Int. Conf. Manip. Autom. Robot. Small Scales, vol. 6, pp. 1783–1816, 2018, doi: 10.1109/MARSS.2018.8481225.