A Differential Evolution-based Pseudotime Estimation Method for Single-cell Data

Nazifa Tasnim Hia[1], Ishrat Jahan Emu[2], Muhammad Ibrahim[3], Sumon Ahmed[4]*
Institute of Information Technology, University of Dhaka, Dhaka-1000, Bangladesh[1,2,4]
Department of Computer Science and Engineering, University of Liberal Arts Bangladesh, Dhaka-1207, Bangladesh[1]
Department of Computer Science and Engineering, University of Dhaka, Dhaka-1000, Bangladesh[3]

***Abstract*—The analysis of single-cell genomics data creates an intriguing opportunity for researchers to examine the complex biological system more closely but is challenging due to inherent biological and technical noise. One popular approach involves learning a lower dimensional manifold or pseudotime trajectory through the data that can capture the primary sources of variation in the data. A smooth function of pseudotime then can be used to align gene expression patterns through the lineages in the trajectory which later facilitates downstream analysis such as heterogeneous cell type identification. Here, we propose a differential evolution based pseudotime estimation method. The model operates on continuous search space and allows easy integration of the cell capture time information in the inference process. The suitability of the proposed model is investigated by applying it on benchmarking single-cell data sets collected from different organisms using different assaying techniques. The experimental result shows the model's capability of producing plausible biological insights about cell ordering which makes it an appealing choice for pseudoitme estimation using single-cell transcriptome data.**

***Keywords*—Pseudotime estimation; trajectory inference; single-cell; differential evolution; RNA-seq**

I. Introduction

The average expression profile provided by the microarray-based conventional bulk RNA-seq technology fails to accurately capture transcriptome variation in individual cells. Gene expression is intrinsically heterogeneous, even in the same or similar cell types [1]. Averaging expression profiles across a cell population fails to capture the stochastic nature of the gene expression associated to different functionally restricted cell types. Therefore, to comprehend the complex biological processes such as the development and differentiation of different cell types, a precise understanding of transcriptome is necessary for individual cells. In single-cell technology, the expression profile of each cell is measured individually. Increasing evidence suggests that many questions in biology such as cellular function development, cell fate decision, etc. can be answered in a more refined way at single cell level [1, 2].

While analyzing gene expression profiles at the individual cell level holds the potential to uncover novel states of complex biological processes, this task is difficult due to intrinsic challenges of both biological and technical nature. Similar to other RNA-seq technologies like microarray, single-cell assaying approaches are also destructive. Hence, in certain instances, the cells being analyzed are undergoing the process of development and differentiation, however, the data lacks any temporal labels. Gene expression dynamics can be analyzed by employing a pseudotemporal ordering of cells. This ordering is based on the principle that cells can be viewed as a time series, where each cell represents a specific time point along the pseudotime trajectory that corresponds to the progression through a process of interest.

The estimation of pseudotime, known as a crucial aspect of analyzing single-cell data, provides a key role in discovering the complex dynamics of biological processes. It involves placing cells along a trajectory that shows the biological phenomenon's relative activity or growth. This crucial task lets us evaluate normal cellular function and identify potential variations that could cause physiological diseases. Time series investigations that track cell transcriptional dynamics over time may gain from pseudotime estimation.

For presenting pseudotime trajectories, different formalisms have been employed, with early approaches focused primarily on dimension reduction, followed by cell mapping. Popular dimension reduction algorithms that have been used on single-cell data includes linear methods such as Principal Component Analysis (PCA) [3] and Independent Component Analysis (ICA) [4], as well as non-linear methods such as t-Stochastic Neighborhood Embedding (t-SNE) [5], diffusion maps [6, 7, 8], Gaussian Process Latent Variable Model (GPLVM) [9, 10, 11] and more recently Uniform Manifold Approximation and Projection (UMAP) [12].

For creating a pseudotime path, after the initial dimension reduction, graph-based methods such as Monocle [3], Wanderlust [13], Waterfall [14], TSCAN [4], Monocle 3 [12] use a simplified graph or tree for pseudotime estimation, where each node of the graph or tree corresponds to either a individual cell or a group of cells. Finally, these methods use different path-finding algorithms to find a path through the series of nodes representing the temporal position of cells across the pseudotime trajectories. SCUBA [15], Slingshot [16], Trade-Seq [17] use curve fitting to model pseudotemporal ordering of cells. These methods use principal curves to characterize pseudotime trajectory where each cell is assigned a pseudotime point based on its lower dimensional projection on principal curves. On the other hand, the diffusion pseudotime (DPT) framework [6, 7] uses random walk-based inference where all the diffusion components are used to infer pseudotime.

Deep learning methods have also been used for pseudotime estimation. An autoencoder is a neural network consisting of an encoder, bottleneck, and decoder that compresses and recon-

**Corresponding author

structs data to obtain a precise representation in a latent space. Variational Autoencoder (VAE) stands out among its seven different types for pseudotime estimation. VAE finds a probability distribution over input data that has been compressed, allowing for unsupervised learning and data compression. VAE applies a normal distribution on the encoded representation and can generate new data samples by decoding learned distribution samples. As demonstrated by Variational Inference for Trajectory by Autoencoder (VITAE) [18], which integrates VAE and hierarchical mixture models to identify non-linear trends and account for confounding covariates. Probabilistic approaches of pseudotime estimation are also available which focus on the quantification of uncertainty across the inferred trajectory [19]. DeLorean [10] and GrandPrix [11] use GPLVM to project cells on latent dimension. These methods support the incorporation of capture time information when availalbe. Recently, DGP-LVM [20] method is developed that additionally supports the incorporation of RNA velocity [21] in the form of derivatives within the GPLVM framework.

The existing pseudotime estimation algorithms use dimension reduction methods at some point in the inference process. The performance of a model, i.e. the accuracy of estimated pseudotime may largely depend on the dimension reduction algorithm being used and the amount of information lost while converting the original data to the lower dimensional space. For instance, linear methods like PCA and ICA may not capture nonlinear biological processes, whereas nonlinear methods like t-SNE and UMAP are computationally expensive as well as difficult to interpret. Recently, [22] have investigated the effects of dimension reduction on pseudotime estimation. They simulated three-dimensional data under three different settings and then employed five distinct dimensional reduction strategies to assess the extent to which the original data might be preserved. They found that all dimension reduction algorithms fail to clearly depict the temporal structure of the data. Therefore, certain pseudotime estimation methods may fail to approximate the underlying trajectory using lower dimensional representation of data particularly when some genes exhibit typical behaviors such as piece-wise linearity etc.

To overcome the issues with dimension reduction, pseudoGA [22] proposes a Genetic Algorithm(GA)-based method that directly uses the original data for pseudotime estimation. PseudoGA employs gene expression value ranking prior to entering the main procedure, assigning the average value in cases where values are identical. Applying GA for optimization requires finding a suitable representation of the candidate solution. PseudoGA assumes the search space is discrete, i.e. the goal of the model is to find the best permutation of cells that can explain the transcriptomic change of gene expression levels along the corresponding trajectory. Therefore, the model uses the permutation representation of cell ordering. Cells are indexed from 1 to n, where n is the number of cells. The algorithm randomly populates different permutations from 1 to n, each representing a candidate pseudotime ordering. This representation of the candidate solutions enables the algorithm to apply genetic operators, i.e. recombination, mutation, and selection on a population to generate a new one. PsedoGA uses a cubic polynomial function and Bayesian Information Criterion (BIC) to evaluate the fitness of each candidate solution and selects the fittest ones for the next generation.

Although the genetic algorithm provides an appealing solution for pseudotime estimation that does not require any dimension reduction, it demands the search space to be discrete. Therefore, PseudoGA only considers the ordering of cells and ignores the absolute position of cells on the estimated trajectory. The paper argues that there may be no physical meaning to the quantitative location of cells on a pseudotime trajectory. For discrete representation cell to cell, distance is the same for all cells of the system. From a biological point of view, this does not seem right. During development, cells receive signals from other cells and stimuli and define their fate decisions. Therefore, all cells do not progress at the same rate hence creating the cell ordering. A cell may be in close proximity to a group of cells and relatively far from other cells. The absolute position of a cell across pseudotime trajectory reflects the cell progression through the underlying biological system. The discrete representation of pseudotime ordering fails to capture these dynamics. The discrete pseudotime ordering forces the Pseudo cost function to use the rank values rather than the actual gene expression values [22]. While the authors claim that the use of ranks aids in the model's ability to avoid the specific effects of any particular functional form of gene expression, it may endanger the model losing valuable information.

Moreover, in some cases, cell capture time is available along with the single-cell RNA-seq data. This capture time information is informative [10, 11]. As capture times are real values, the discrete representation of pseudotime does not allow the incorporation of this information within the inference process, although PseudoGA has used capture time to validate the estimated pseudotimes. But [10, 11] have shown that the incorporation of capture time information within the inference process helps the model significantly, even the model can identify specific features of interest such as cell cycle or other sources of variations such as branching dynamics. In this contribution, we present a new efficient pseudotime estimation algorithm based on differential evolution (DE). Differential evolution is a metaheuristics optimization algorithm that has a long legacy in bioinformatics applications [23, 24, 25]. DE optimization operates on continuous search space hence facilitating the smooth integration of capture time information within the inference process. The model obviates the necessity for dimensionality reduction techniques and the estimated pseudotime represents the ordering of cells as well as cell progression through the dynamic biological process.

The rest of the paper is divided into a set of sections, each developing a part of the research. Section II discusses the proposed approach and its specific workings. Section III outlines the experimental results. Sections IV and V include the Result Analysis and Discussion respectively, analyzing the study's significant outcomes and implications. Finally, Section VI concludes the current study with possible future directions.

II. Proposed Method

Differential Evolution (DE) is a widely used metaheuristics optimization algorithm that can be easily adapted for pseudotime estimation. The algorithm iteratively tries to improve a candidate solution based on a quantity known as the fitness score. The algorithm proceeds by generating an initial population of candidate solutions known as chromosomes or individuals. Each chromosome represents a pseudotemporal ordering of cells under consideration. By combining the existing candidate solutions, the optimization process generates new candidate solutions and keeps whichever candidates possess a better fitness score. In this way, DE maintains a population of the fittest candidate solutions from one generation to the next. This process continues until a termination criterion is met. The outline of the proposed algorithm is shown in Fig. 1.

Since the expression profiles of all genes do not contribute equally to the inference process, it is recommended to perform a preliminary gene selection to improve the accuracy of pseudotime estimation. As cells are ideally clustered into two or more clusters, therefore, genes that are differentially expressed among clusters are chosen for the inference process. Other feature selection approaches for single-cell data such as the selection of highly variable genes [26, 27], and dropout-based feature selection [28] can also be employed.

A. Feature Selection

We use the Wilcoxon rank sum test [29] to compare the expression levels of the transcriptomics dataset of individual genes between pairs of clusters.

The Wilcoxon rank sum test, also known as the Mann-Whitney U test, is a nonparametric statistical test used to compare the differences between two independent groups or samples. It involves ranking all the observations from both groups together and calculating the sum of ranks for each group. The test statistic is calculated as the smaller of the two sums of ranks, which represents the probability that a randomly chosen observation from one group is smaller than a randomly chosen observation from the other group. It compares the differences between groups without making assumptions about the underlying distribution of the data.

Therefore, to select interesting genes, at first, a cluster (cluster i) is selected and has been compared with other clusters. Then submatrices are created containing only the samples from cluster i and the second cluster being compared (cluster j). The resulting p-values are stored in a vector, which is then sorted to identify the genes that are most differentially expressed between the two clusters. This process is repeated for all pairs of clusters, with the resulting vectors of differentially expressed genes and cluster indices being stored to be used for pseudotime estimation.

B. Representation of Pseudotime and Incorporation of Cell Capture Time

Differential evolution operates in a continuous search space. Therefore, the chromosomal representation of pseudotime is straightforward. Any collection of n real numbers can be a candidate chromosome where n is the number of cells. Formally, each individual X is represented as,

$$X = \{x_1, x_2, ..., x_n\}, \tag{1}$$

where each x_j corresponds to the pseudotime point of cell j.

However, the critical assumption of the proposed model is that the available cell capture times are informative to model the biological dynamics of interest. Therefore, at the time of population initialization, for each chromosome, the pseudotime value x_j of cell j is drawn from a normal distribution centered on the capture time c_j of cell j,

$$x_j = N(c_j, \sigma^2), \tag{2}$$

where σ^2 represents the variance of pseudotime around the cell capture time.

C. Cost Function

The extent to which a pseudotime trajectory interprets specific changes in gene expression level can also be described in terms of a cost function. Fitting a smooth curve with the expression values as the dependent variable and the pseudotime values as the explanatory variable yields this cost or penalty. The hypothesis for this cost function is to find out which individual is better at explaining the behavior of gene expression.

Gene expression along pseudotime exhibits three distinct patterns: (i) monotonic increase or decrease, (ii) peak or dip followed by a reversal, and (iii) peak or dip followed by a secondary change in expression. To capture these patterns, our algorithm assumes that gene expression values can be modeled by a polynomial of degree up to 3 as in [22]. This flexibility accommodates even cyclic behavior in specific genes throughout the pseudotime trajectory, encompassing all three expression patterns mentioned.

For each gene j in cell i, the expression level $y_{i,j}$ is modeled using a cubic polynomial,

$$y_{i,j} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_{i,j}, \tag{3}$$

where x_i represents the pseudotime for cell i and $\epsilon_{i,j}$ is associated noise. Therefore, cost of a chromosome X for gene j can be defined by the mean square error (MSE),

$$\text{MSE}_{X,j} = \frac{1}{n} * \sum_{i=1}^{n} \left(y_{i,j} - \widehat{y}_{X,i,j} \right)^2, \tag{4}$$

where n is the number of cells and $\widehat{y}_{X,i,j}$ represents the calculated expression level of gene j using the pseudotime value of the cells i according to chromosome X. Now, the error or fitness score of chromosome X is,

$$\text{MSE}_X = \sum_{j=1}^{D} \text{MSE}_{X,j} \tag{5}$$

where D is the number of genes being used for pseudotime estimation.

D. Inference Algorithm

Then crossover and mutation operations of DE are applied to these individuals to generate a new population of NP offspring individuals, where NP is 4 to 10 times greater than the size a single chromosome. These newly generated offspring individuals are combined with the old parent individuals to create a combined population of size $2.NP$.

Crossover is a key element of the Differential Evolution algorithm, as it permits the combination of data from various individuals to generate new candidate solutions. This process entails the exchange or recombination of parent solution

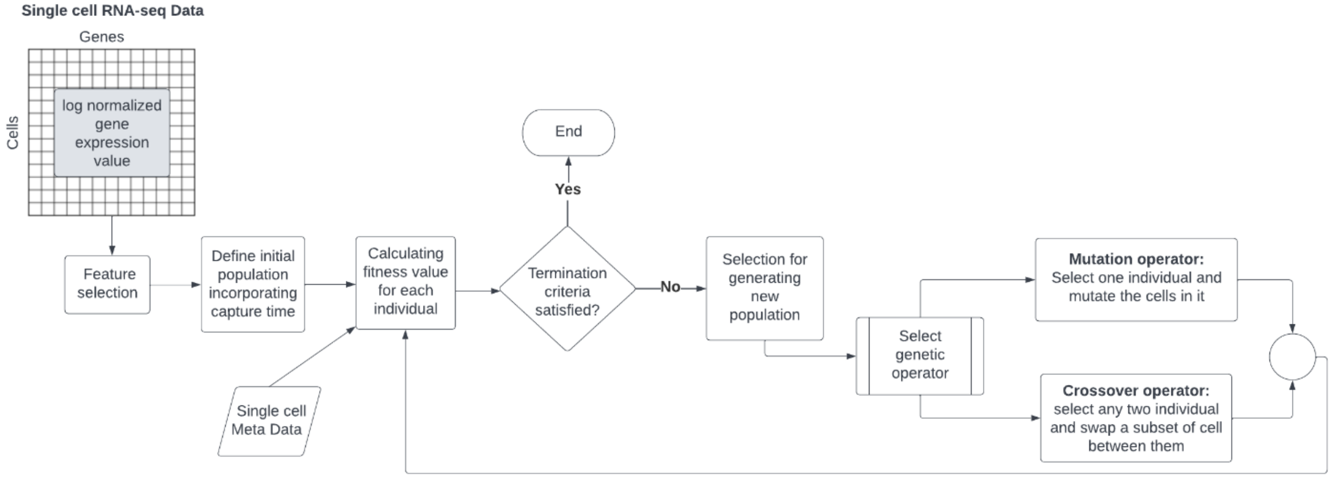
Single cell RNA-seq Data
Genes
Cells
log normalized gene expression value
Feature selection
Define initial population incorporating capture time
Single cell Meta Data
Calculating fitness value for each individual
Termination criteria satisfied?
Yes
End
No
Selection for generating new population
Select genetic operator
Mutation operator: Select one individual and mutate the cells in it
Crossover operator: select any two individual and swap a subset of cell between them

Fig. 1. Framework of the proposed methodology.

parameters or variables. In our algorithm, we employ the crossover operation to produce diverse offspring solutions and potentially enhance the population's overall performance. In our work, we have used a standard single-point crossover strategy. However, the specific crossover strategy and parameters employed depend on the problem domain and optimization process objectives.

The solution space of the problem contains many local optima that may lead the search algorithm to the wrong direction and, eventually, the global solution may remain undetected [23]. Thus, for locating the global optimal solution in such a search space, population diversity has to be maintained. Mutation is the operator that has traditionally been used in differential evolutions for introducing diversity in the population. As our search space is continuous, the mutation operation updates the pseudotime value of a cell with a new value drawn from a normal distribution centered at the current pseudotime value of that particular cell. Each of the $2NP$ chromosomes is chosen and based on a mutation probability, a mutation operation is applied to them. This gives us an augmented set of chromosomes of size $4NP$.

The optimization procedure evaluates the quality of all $4NP$ solutions using a cost function and the top 25% chromosomes are selected for the next generation. This way, the algorithm iteratively enhances the population through multiple iterations of generating new offspring, evaluating their fitness using the cost function, and updating the population based on selection criteria. This method permits the exploration and refinement of candidate solutions until an optimal or near-optimal solution is reached.

III. Experimental Results

We examine our proposed model's performance by employing it on multiple datasets with different sizes and characteristics that have been collected from distinct organisms using various approaches. Table I contains a brief description of the datasets.

```
Algorithm Pseudotime Estimation
  Input: Cell by gene matrix obtained from single cell RNA-seq data. Choose an $\epsilon$, a small preassigned positive quantity.

  Output: Near optimum pseudotime of cells.
  Construct $X^0 = \{X_1, \dots, X_{NP}\}$: initial set of chromosomal representing of pseudotemporal ordering of cells.
  while Minimum cost function over the population converges do
    Step 1: Perform crossover on $X^0$ to generate offsprings. Set of chromosomes becomes $X^1 = \{X_1, \dots, X_{NP}, X_1^{(o)}, \dots, X_{NP}^{(o)}\}$, where $\{X_1^{(o)}, \dots, X_{NP}^{(o)}\}$ are the offspring from $\{X_1, \dots, X_{NP}\}$ due to crossover. Here $C(X^1) = 2NP$, where $C(A)$ is the cardinality (number of elements) of a set $A$.
    Step 2: Perform Mutation on each element of $X^1$ to find a new augmented set of chromosomes $X^2 = \{X^1, X^{(m)}\}$. $X^{(m)} = \{X_1^{(m)}, \dots, X_{NP}^{(m)}, X_1^{(mo)}, \dots, X_{NP}^{(mo)}\}$, where $X_i^{(m)}$ and $X_i^{(mo)}$ are new chromosomes due to mutation from $X_i$ and $X_i^{(o)}$ respectively for each $i = 1, \dots, NP$. Clearly $C(X^2) = 4NP$.
    Step 3: Calculate cost for each chromosome in $X^2$ and order them as $C_{(1)}, \dots, C_{(4NP)}$, where $C_{(r)}$ is the $r$-th ordered value of $\{C_{(1)}, \dots, C_{(4NP)}\}$. Selection is based on choosing the best $NP$ chromosomes, i.e. chromosomes corresponding to $\{C_{(1)}, \dots, C_{(NP)}\}$. Denote this new set of chromosomes as $X^1$ obtained after first iteration.
    Step 4: Go back to Step 1 - 3 until $|C_{\text{new}(1)} - C_{(1)}| < \epsilon$
```

TABLE I. Datasets in Detail

Dataset Name	Samples	Features	Capture Time
Whole-leaf Microarrays of *Arabidopsis Thaliana* Data [30]	24	100	4
Human Preimplantation Embryos Data [31]	90	500	7
Human Acinar Cell Data [32]	271	500	4
Human Skeletal Muscle Myoblasts (HSMM) [33]	312	500	8
Mouse Embryonic Fibroblast [34]	315	500	5

A. Whole-leaf Microarrays of Arabidopsis Thaliana

Windram et al. [30] studied a high-resolution time series of gene expression profiles from a single leaf of Arabidopsis thaliana during infection by Botrytis cinerea. Using time series measurements, they compared infected samples to control conditions over 48 hours. The study found that about one-third of the Arabidopsis genome showed differential expression during the first 48 hours after infection. The data included 24 distinct time points, with measurements conducted every two hours. For our experiment, we divide these time points into four groups, each containing six consecutive time points that are used to initialize the model. For pseudotime inference, 100 genes out of 150 described genes are used, with the remaining 50 genes being held out for validation.

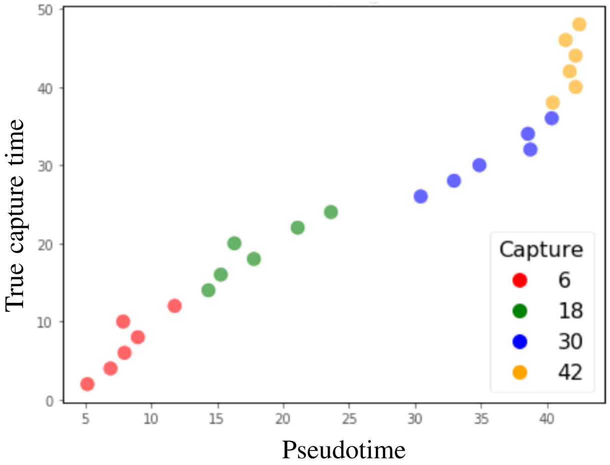

Fig. 2. *Arabidopsis thaliana* microarray data [30]: Pseudotime (horizontal axis) versus true capture time. Colors represent the prior information utilized for the inference process.

We plot the estimated pseudotime against the actual cell capture times to examine their correspondence, as shown in Fig. 2. Each point on the plot represents a specific time point, with colors indicating the synthesized cell capture time.

B. Human Preimplantation Embryos Data

The Human embryo development data [31] includes embryos at seven preimplantation stages, including oocyte, zygote, 2-cell, 4-cell, 8-cell, morula, and late blastocyst at the hatching stage. The dataset also includes individual blastomeres of three 2-cell, three 4-cell, and two 8-cell embryos for analysis. Before pseudotime estimation, gene filtering improves the algorithm's accuracy. We select the top 500 differentially expressed genes for our experiment by dividing all genes into two clusters. The detailed process is described in Section II.

The analysis [31] shows that all cells grouped together are from the same stage of development, except for two blastomeres from a morula stage embryo that were grouped with blastocysts. This finding is also consistent with our findings shown in Fig. 3.

Based on the information presented in Fig. 3, we can see that the cells of each stage have formed distinct regions, making them readily identifiable. According to the source [35, 36], during the transition from the 4-cell to the 8-cell stage, the most significant variations in gene expression were observed. Our findings reflect this pattern of behavior. From the oocyte to the 4-cell stage, they maintain a shared pseudotime range, and from the 8-cell stage to late blastocytes, some cells exhibit a similar pseudotime. However, there are hardly any shared pseudotime-contained cells between the 4 and 8-cell stages.

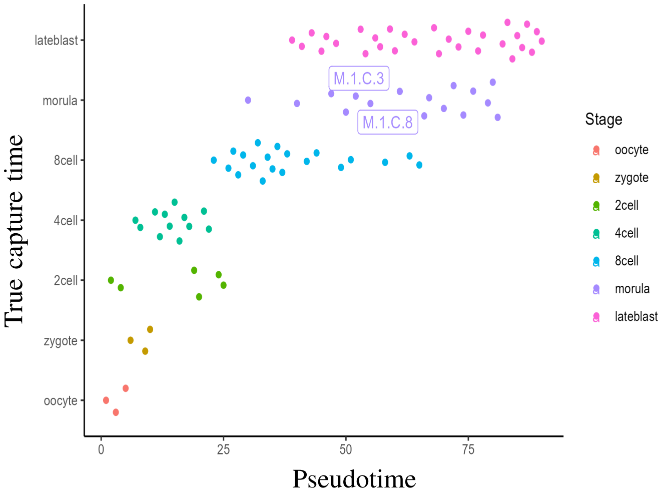

Fig. 3. Analysis of Human embryo developmental data [31]. X-axis is the the ranks of pseudotime; Y-axis represents the developmental stage, with discrete values for each point of each stage.

ACCSL, *C21ortf*, *ALOX15*, *C10orf82* and *RSPO2* are the top five genes that have the highest linear rank correlation with the estimated pseudotime. Fig. 4 plots the profiles of the top genes with the estimated pseudotime. On the x-axis, we observe pseudotime projections for each cell, and the y-axis displays smoothed, z-scored log gene expression. The color scheme indicates label order.

The conclusion drawn from the aforementioned biological validation is that the result obtained by our model is capable of seeing the latent pattern of data.

C. Human Acinar Cell Data

The study [32] examines the changes in the pancreas with age and diabetes development using single-cell RNA sequencing from 28 human volunteers aged 1 to 75. The age-dependent mutational signature in the endocrine pancreas is caused by reactive oxygen species and consists of high rates of $C > A$ and $C > G$ changes. The accumulation of epigenetic errors may explain the decline in fitness and organ function with age.

The initial dataset contains 411 cells. For our experiment, we select 312 samples using random sampling. A population of 1248 potential solutions is generated using donor age as a starting point for optimization. Each individual is created from a normal distribution with a mean equal to the donor's age and a standard deviation of 4. The optimization process involves 100 iterations, with promising individuals chosen based on the objective function. Promising individuals are used to generate improved solutions through crossover and mutation. Fig. 5 shows the estimated pseudotime for individual cells

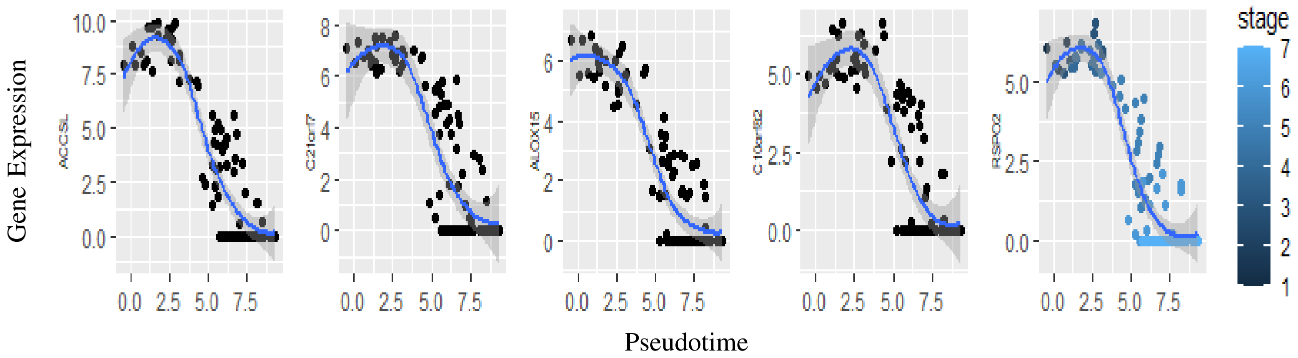

Fig. 4. Exploring the intricate profiles of highly variable genes within Embryo data [31]. Visualizing the five genes characterized by the highest absolute coefficients against the pseudotime generated through our model. The line depicts a geom smoothed curve, crafted using the ggplot2 R package.

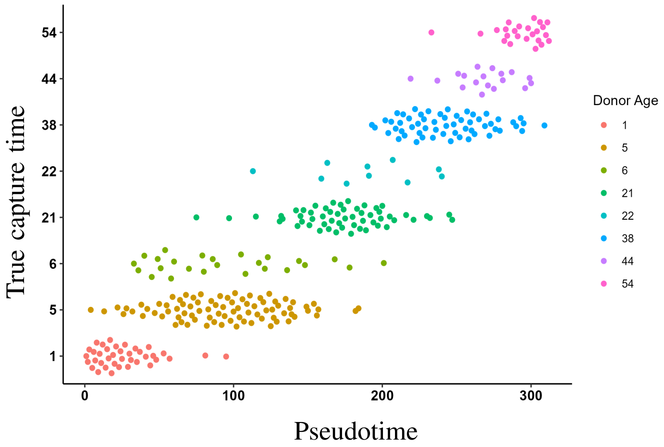

Fig. 5. Analysis of Human Acinar Cell [32]. Pseudotime (horizontal axis) versus true capture time. Colors represent the prior information utilized for the inference process.

corresponds to different donor ages. The expression profiles of the most correlated genes with pseudotime are plotted against the estimated pseudotime in Fig. 6. Clusterin (CLU) is important in pancreatic regeneration and is expressed in chronic pancreatitis [37]. Amylase (AMY2B) is characteristic of mature acinar cells and encodes a digestive enzyme [38]. ITM2A is significantly differently regulated in a model of chronic pancreatitis.

The study reveals molecular changes in the pancreas as we age, with somatic mutations potentially contributing to pancreatic diseases like cancer and diabetes. The research uses single-cell RNA sequencing data on primary cells to understand genetic and transcriptional processes in human tissue as it ages. This allows studying traits in arbitrary cell populations from primary tissue, regardless of cell division ability. The results could guide future research on age-related diseases and develop effective treatments. The analysis reveals specific gene expression alterations associated with aging in the pancreas of humans, with pseudotime for close age data falling within a roughly identical range. Cells captured from later age groups contain diverse and distinct ranges from earlier age groups.

D. Human Skeletal Muscle Myoblasts (HSMM)

Primary Human Skeletal Muscle Myoblasts (HSMM) [33] are the first myoblast cells isolated from human skeletal muscle tissue. These cells can proliferate and multiply and these were cultured in mitogen-rich environments to promote growth and division. After proliferation, they undergo differentiation, which transforms undifferentiated cells into specialized or mature cells. To induce differentiation, myoblasts are transferred to a culture medium with minimal mitogen concentrations. RNA-seq libraries were collected from several hundred serum-induced differentiated cells over an extended period of time. The data were collected from 271 cells at 0, 24, 48, and 72 hours after differentiation conditions. Myoblasts, intermediates, myotubes, fibroblasts and undifferentiated cells were annotated using Gene Set Variation Analysis (GSVA) [39] based on known gene markers.

In this experiment, we optimize an initial population of 1084 viable solutions, using capture time as a baseline. We generate each individual by randomly selecting data from a normal distribution, using the capture time as the mean and within three standard deviations. The model requires one hundred iterations, each selecting individuals according to the objective function. We employ survivals to develop improved solutions through crossover and mutation, with a probability of 0.95 and 0.1, respectively. Fig. 7 illustrates the relationship between the resultant pseudotime and the capture time.

Our model's estimated pseudotime is consistent with the findings of Tran and Bader [40]. There is a shared pseudotime range between cells from 0H to 24H, as well as a shared range between the other three stages. The result shows an increasing trend and aligns with the known biology of myotube development. Our model's pseudotime has a Pearson correlation of 0.943 with the collection time of the data sets.

E. Mouse Embryonic Fibroblast

The dataset reveals the transcription changes that occur when MEFs are converted into neurons using transcription factors Ascl1, Brn2, and Myt1l (BAM). Researchers examined transcriptomes of single cells at multiple time points during the direct conversion of MEFs into induced neuronal cells. The data was extracted at Day 0 (starting point), Day 2 (Ascl1-only cells), Day 5 (purifying Tau–eGFP+ and Tau–eGFP-

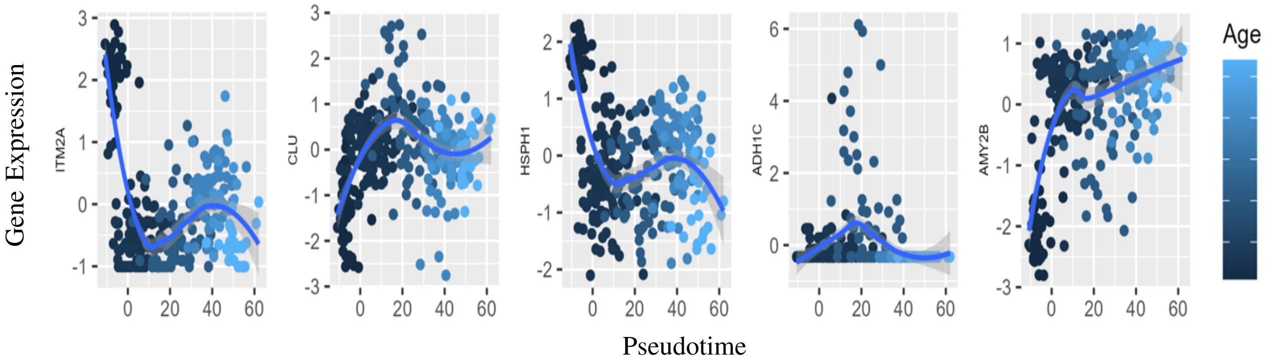

Fig. 6. Profiles of the highly variable genes in acinar cells [32]. Plotting the five genes with the highest absolute coefficients against the pseudotime generated by this model. The line represents a geom smoothed curve as determined by the ggplot2 R package.

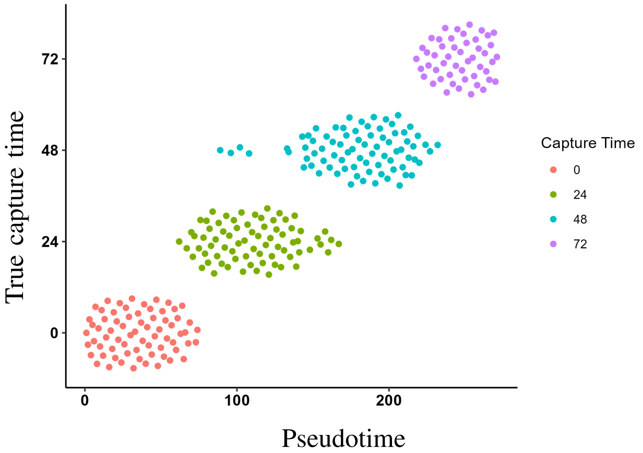

Fig. 7. Human Skeletal Muscle Myoblasts (HSMM) [33]. Pseudotime (horizontal axis) versus true capture time. Colors represent the prior information utilized for the inference process.

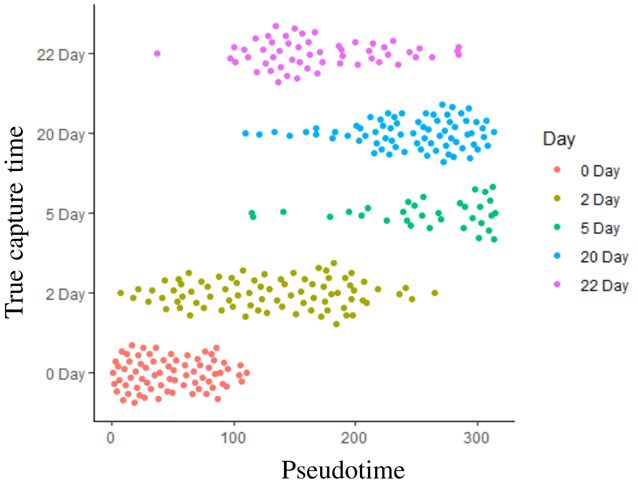

Fig. 8. Analysis of Mouse Embryonic Fibroblast (MEF) data [34]. Pseudotime (horizontal axis) versus true capture time. Colors represent the prior information utilized for the inference process.

cells), Day 20 (late stage of reprogramming), and Day 22 (BAM-mediated reprogramming). The researchers used principal component analysis (PCA) to identify three distinct clusters within Ascl1-only cells based on their expression level. On Day 20, they analyzed a subset of Tau–eGFP+ cells, representing the late stage of the reprogramming process. On Day 22, they analyzed both Ascl1-only cells and cells reprogrammed using all three BAM factors, comparing the transcriptional profiles of cells reprogrammed with different factor combinations. This data provides insights into the heterogeneity and limitations of the reprogramming process.

The optimization process involves generating an initial population of 1260 potential solutions using the collection time as a basis. Each solution is created from a normal distribution with a mean equal to the donor's age and a standard deviation of 3. The process involves 150 iterations, with promising individuals chosen based on the objective function. Promising individuals are used to generate improved solutions through crossover and mutation. The pseudotime results are plotted against the capture time in Fig. 8, and expression values of the most correlated genes are drawn.

Expression values of the most correlated genes with pseudotime are drawn in Fig. 9, x-axis is pseudotime value learned for each cell; y-axis is z-scored log2 gene expression values.

IV. Result Analysis

A. Tracing Gene Expression Changes through Pseudotime

Throughout the processes of cellular development, proliferation and the other similiar activities, individual genes manifest distinct behavioral patterns. As per existing literature and our formulated hypothesis, these behaviors can broadly be categorized into three distinct patterns: (i) a monotonic increase or decrease, (ii) a peak or dip followed by a reversal, and (iii) a peak or dip succeeded by a secondary change in expression.

The calculation of pseudotime relies primarily on understanding the intrinsic behaviour of the genes involved. To evaluate the accuracy of the derived pseudotime, a crucial procedure is generating a graphical representation that aligns gene profiles with the estimated pseudotime. In these plots,

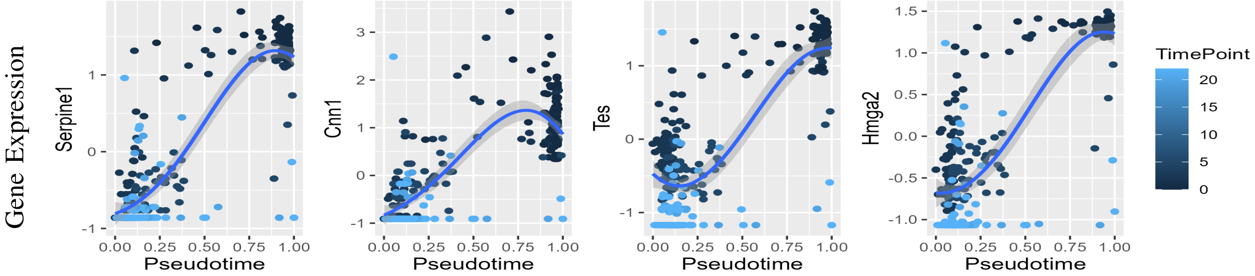

Fig. 9. Profiles of highly variable genes from MEF data [34]. Plotting the five genes with the highest absolute coefficients against the pseudotime generated by this model. The line depicts a geom-smoothed curve generated by the ggplot2 R package.

a discernibly smooth curve serves as an indicator that the resulting pseudotime adeptly captures the intricate behavior of the genes.

In Fig. 4, 6, 9 we observe the preeminent correlation of genes with pseudotime. These plots notably exhibit a remarkably smooth curve, consistently adhering to the anticipated and hypothesized patterns. It thus validates our estimated pseudotime.

B. Roughness Statistics

To validate our results, we utilize a technique describe in [10]. This method focuses on assessing the uniformity of expression profiles for excluded genes throughout the estimated pseudotime.

The statistical process is used to capture the smoothness of the gene expression values x_g, c' across cells $1 \leq c \leq C$, pseudotime $\tau 1 .....\tau C$, and ordering $z1......zC$ satisfying the condition $\tau z1 \leq .... \leq \tau zC$. The roughness of the genes is determined by the disparities between successive expression measurements in pseudotime ordering.

$$\mathrm{R}_g(z) = \frac{1}{\sigma_g}\sqrt{\frac{1}{C-1}\sum_{c=1}^{C-1}\left(x'_{g,z_c} - x'_{g,z_{c+1}}\right)^2} \qquad (6)$$

TABLE II. The Roughness Statistics Values for the Datasets using the Pseudotime Generated by the Proposed Method, in Comparison to Three other Widely Recognized Methods

Datasets / Models	1	2	3	4	5
Our Model	0.71	0.55	0.63	1.10	0.57
PseudoGA	0.82	0.44	0.59	1.39	0.61
slingshot	0.77	0.46	1.10	0.86	0.41
Monolcle3	0.92	0.53	1.06	0.83	0.40

In Eq. (6), σ_g represents the standard deviation of the expression measurements in this context. Lower R_g values imply smoother expression profiles, whereas higher values indicate rougher expression profiles. However, there's an acceptance range for this value, which is valid if the value falls within two standard deviations $(2 * \sigma_g)$ of the gene expression values. Values of R_g within this range are considered to be acceptable for assessing the uniformity of expression profiles for excluded genes.

In Table II, the roughness statistics values corresponding to the datasets employed in this study are presented. With the exception of HSMM dataset, all values fall within one standard deviation. Notably, while acknowledging that the acceptable range for roughness values extends up to two standard deviations, the observed values affirm a coherent relationship between gene expression and the estimated pseudotime.

V. Discussion

With the emergence of single-cell transcriptomics, the field of functional genomics has made significant progress, which enables an in-depth analysis of cellular processes such as tissue development and cellular differentiation. The first step towards analyzing single-cell data obtained from a developmental biological system is to project cells on a pseudotemporal trajectory representing the ordering of cells based on their cellular development. This ordering of cells can be viewed as the restoration of the time series information that was lost at the time of the cell capture process.

To estimate pseudotime trajectories, a number of methods have been developed in the existing literature. Most of these methods construct the pseudotime based on the lower dimensional representation of the original data. While dimension reduction algorithms aim to identify major trends within the underlying data, recent studies [22] have shown that it is susceptible to losing valuable information. Therefore, certain methods may find it difficult to approximate a temporal trajectory while using reduced dimensional data. A genetic algorithm-based model PseudoGA has been developed in [22] that does not employ any dimension reduction for pseudotime inference. Through a number of experiments, this work has been aimed to tackle the challenges associated with the lower dimensional representation of the data as well as the proposed model's applicability of pseudotime estimation while using the original data.

However, being a GA-based algorithm, PseudoGA is forced to use a discrete chromosomal representation which greatly hinders the flexibility and usability of the proposed model. First, the discrete representation assumes that all cells maintain an equal pseudotemporal distance from one another. The model only provides a pesudotime ordering of cells and ignores the physical interpretation of cells' pseudotime values across the trajectory. Therefore, a cell's progression through development processes compared to other cells can not depicted. Second, PseudoGA needs to use gene rank values instead of actual gene expressions, which in the long run may affect the quality of estimated pseudotime. Third, applying genetic operators, i.e. crossover and mutation on the discrete pseudotime representation demands special consideration. Otherwise, more than one cell may try to occupy the same pseudotime location. This collision is evident and PseudoGA needs to employ special treatments to avoid this [22]. Finally, and most importantly, the discrete chromosomal representation of pseudotime does not allow the incorporation of the cell capture time when available. Finally, and most importantly, the discrete chromosomal representation of pseudotime does not allow the incorporation of the cell capture time when available. This capture time information is informative and its incorporation within the inference process helps the model to find more biologically plausible pseudotime estimation [10, 11].

In this study, we introduce a new computational model, which provides some notable advantages. At the core of our model is the differential evolution algorithm, which operates on continuous search space. The model obviates the necessity for dimensionality reduction techniques and facilitates the smooth integration of capture time information during the population initialization stage. Because of the simple chromosomal representation (see Section II-B), the implementation of crossover and mutation is straightforward and does not require any special attention. The model uses the actual gene expressions which further strengthens the model's ability of pesudotime estimation, especially in the presence of genes having particular expression profiles. Finally, the estimated pseudotime not only provides the cell ordering but also depicts the cellular progression of undergoing biological system. We assessed the performance of our proposed model on multiple datasets of varying sizes and derived from different organisms using different single-cell assaying techniques. Five different datasets have shown consistent results from our approach, which demonstrates its reliability. Through extensive experimentation, we demonstrate that our proposed model can be used to effectively estimate the pseudotime, a significant factor in temporal analysis of single-cell data, with similar or even greater precision. This improvement could enhance our comprehension of complicated biological processes in a dynamic setting by enabling us to analyze single-cell information and extract relevant temporal dynamics.

VI. Conclusion

The analysis of single-cell transcriptomics and pseudotime inference methods provide intriguing possibilities for understanding complex dynamics of cellular processes where the generation of time course experiments is challenging or technically impossible. As single-cell data are becoming increasingly available in larger volumes, therefore, simple yet rigorous approaches such as the differential evolution we have presented will become ever more relevant. Differential evolution is inherently parallel. The flexibility of the proposed approach can further leverage the parallel execution of the model for larger sample data as well as analysis of the connection between pseudotime and lineage or branching structures; with the potential for future refinement and expansion.

Supplementary Source code and data are available at https://github.com/sumonahmedUoM/PseudoDE

Acknowledgment

This work is funded by a research grant (2022-23) from the University of Dhaka, Bangladesh. NTH is supported by the Information and Communication Technology Division, Ministry of Telecommunication and Information Technology, Government of the People's Republic of Bangladesh, ICT Fellowship No. 56.00.0000.052.33.004.22-38.

References

[1] C. Gawad, W. Koh, and S. R. Quake, "Single-cell genome sequencing: current state of the science," *Nature Reviews Genetics*, vol. 17, no. 3, pp. 175–188, 2016.

[2] B. Hwang, J. H. Lee, and D. Bang, "Single-cell rna sequencing technologies and bioinformatics pipelines," *Experimental & molecular medicine*, vol. 50, no. 8, pp. 1–14, 2018.

[3] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn, "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells," *Nature biotechnology*, vol. 32, no. 4, pp. 381–386, 2014.

[4] Z. Ji and H. Ji, "Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis," *Nucleic acids research*, vol. 44, no. 13, pp. e117–e117, 2016.

[5] B. Becher, A. Schlitzer, J. Chen, F. Mair, H. R. Sumatoh, K. W. W. Teng, D. Low, C. Ruedl, P. Riccardi-Castagnoli, M. Poidinger *et al.*, "High-dimensional analysis of the murine myeloid cell system," *Nature immunology*, vol. 15, no. 12, pp. 1181–1189, 2014.

[6] L. Haghverdi, F. Buettner, and F. J. Theis, "Diffusion maps for high-dimensional single-cell analysis of differentiation data," *Bioinformatics*, vol. 31, no. 18, pp. 2989–2998, 2015.

[7] L. Haghverdi, M. Büttner, F. A. Wolf, F. Buettner, and F. J. Theis, "Diffusion pseudotime robustly reconstructs lineage branching," *Nature methods*, vol. 13, no. 10, pp. 845–848, 2016.

[8] M. Setty, M. D. Tadmor, S. Reich-Zeliger, O. Angel, T. M. Salame, P. Kathail, K. Choi, S. Bendall, N. Friedman, and D. Pe'er, "Wishbone identifies bifurcating developmental trajectories from single-cell data," *Nature biotechnology*, vol. 34, no. 6, pp. 637–645, 2016.

[9] 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. i626–i632, 2012.

[10] J. E. Reid and L. Wernisch, "Pseudotime estimation: deconfounding single cell time series," *Bioinformatics*, vol. 32, no. 19, pp. 2973–2980, 2016.

[11] 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.

[12] J. Cao, M. Spielmann, X. Qiu, X. Huang, D. M. Ibrahim, A. J. Hill, F. Zhang, S. Mundlos, L. Christiansen, F. J. Steemers *et al.*, "The single-cell transcriptional landscape of mammalian organogenesis," *Nature*, vol. 566, no. 7745, pp. 496–502, 2019.

[13] S. C. Bendall, K. L. Davis, E.-a. D. Amir, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan, and D. Pe'er, "Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development," *Cell*, vol. 157, no. 3, pp. 714–725, 2014.

[14] J. Shin, D. A. Berg, Y. Zhu, J. Y. Shin, J. Song, M. A. Bonaguidi, G. Enikolopov, D. W. Nauen, K. M. Christian, G.-l. Ming *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.

[15] E. Marco, R. L. Karp, G. Guo, P. Robson, A. H. Hart, L. Trippa, and G.-C. Yuan, "Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape," *Proceedings of the National Academy of Sciences*, vol. 111, no. 52, pp. E5643–E5650, 2014.

[16] K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, and S. Dudoit, "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics," *BMC genomics*, vol. 19, pp. 1–16, 2018.

[17] K. Van den Berge, H. Roux de Bézieux, K. Street, W. Saelens, R. Cannoodt, Y. Saeys, S. Dudoit, and L. Clement, "Trajectory-based differential expression analysis for single-cell sequencing data," *Nature communications*, vol. 11, no. 1, p. 1201, 2020.

[18] J.-H. Du, M. Gao, and J. Wang, "Model-based trajectory inference for single-Cell RNA sequencing using deep learning with a mixture prior," *bioRxiv*, pp. 1–32, 2020.

[19] M. Rattray, J. Yang, S. Ahmed, and A. Boukouvalas, "Modelling gene expression dynamics with gaussian process inference," *Handbook of Statistical Genomics: Two Volume Set*, pp. 879–20, 2019.

[20] S. Mukherjee, M. Claassen, and P.-C. Bürkner, "Dgp-lvm: Derivative gaussian process latent variable model," *arXiv preprint arXiv:2404.04074*, 2024.

[21] G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan *et al.*, "Rna velocity of single cells," *Nature*, vol. 560, no. 7719, pp. 494–498, 2018.

[22] P. K. Mondal, U. S. Saha, and I. Mukhopadhyay, "Pseudoga: cell pseudotime reconstruction based on genetic algorithm," *Nucleic Acids Research*, vol. 49, no. 14, pp. 7909–7924, 2021.

[23] N. Noman and H. Iba, "Inferring gene regulatory networks using differential evolution with local search heuristics," *IEEE/ACM Transactions on computational biology and bioinformatics*, vol. 4, no. 4, pp. 634–647, 2007.

[24] S. Ahmed, M. Hasan, and N. Noman, "Reconstructing gene regulatory network using linear time-variant model," *Dhaka University Journal of Applied Science and Engineering*, vol. 1, no. 2, pp. 125–129, 2011.

[25] S. Ahmed, M. N. A. Tawhid, K. Sakib, and M. M. Rahman, "A multi-objective evolutionary approach to reconstruct gene regulatory network using recurrent neural network model," *Biojournal of Science and Technology*, vol. 2, pp. 1–11, 2015.

[26] C. A. Vallejos, J. C. Marioni, and S. Richardson, "Basics: Bayesian analysis of single-cell sequencing data," *PLoS computational biology*, vol. 11, no. 6, p. e1004333, 2015.

[27] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle, "Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells," *Nature biotechnology*, vol. 33, no. 2, pp. 155–160, 2015.

[28] T. S. Andrews and M. Hemberg, "M3drop: dropout-based feature selection for scrnaseq," *Bioinformatics*, vol. 35, no. 16, pp. 2865–2867, 2019.

[29] B. Rosner, R. J. Glynn, and M.-L. Ting Lee, "Incorporation of clustering effects for the wilcoxon rank sum test: a large-sample approach," *Biometrics*, vol. 59, no. 4, pp. 1089–1098, 2003.

[30] O. Windram, P. Madhou, S. McHattie, C. Hill, R. Hickman, E. Cooke, D. J. Jenkins, C. A. Penfold, L. Baxter, E. Breeze *et al.*, "Arabidopsis defense against botrytis cinerea: chronology and regulation deciphered by high-resolution temporal transcriptomic analysis," *The Plant Cell*, vol. 24, no. 9, pp. 3530–3557, 2012.

[31] L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan *et al.*, "Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells," *Nature structural & molecular biology*, vol. 20, no. 9, pp. 1131–1139, 2013.

[32] M. Enge, H. E. Arda, M. Mignardi, J. Beausang, R. Bottino, S. K. Kim, and S. R. Quake, "Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns," *Cell*, vol. 171, no. 2, pp. 321–330, 2017.

[33] Cole Trapnell, *Single-cell RNA-Seq for differentiating human skeletal muscle myoblasts (HSMM)*, Bioconductor, 2021. [Online]. Available: https://bioconductor.org/packages/release/data/experiment/html/HSMMSingleCell.html

[34] B. Treutlein, Q. Y. Lee, J. G. Camp, M. Mall, W. Koh, S. A. M. Shariati, S. Sim, N. F. Neff, J. M. Skotheim, M. Wernig *et al.*, "Dissecting direct reprogramming from fibroblast to neuron using single-cell rna-seq," *Nature*, vol. 534, no. 7607, pp. 391–395, 2016.

[35] K. Cockburn, J. Rossant *et al.*, "Making the blastocyst: lessons from the mouse," *The Journal of clinical investigation*, vol. 120, no. 4, pp. 995–1003, 2010.

[36] J. Rossant and P. P. Tam, "Blastocyst lineage formation, early embryonic asymmetries and axis patterning in the mouse," 2009.

[37] S. Lee, S.-W. Hong, B.-H. Min, Y.-J. Shim, K.-U. Lee, I.-K. Lee, M. Bendayan, B. J. Aronow, and I.-S. Park, "c," *Developmental Dynamics*, vol. 240, no. 3, pp. 605–615, 2011.

[38] K. Omichi and S. Hase, "Identification of the characteristic amino-acid sequence for human α-amylase encoded by the amy2b gene," *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, vol. 1203, no. 2, pp. 224–229, 1993.

[39] S. Hänzelmann, R. Castelo, and J. Guinney, "Gsva: gene set variation analysis for microarray and rna-seq data," *BMC bioinformatics*, vol. 14, pp. 1–15, 2013.

[40] T. N. Tran and G. D. Bader, "Tempora: cell trajectory inference using time-series single-cell rna sequencing data," *PLoS computational biology*, vol. 16, no. 9, p. e1008205, 2020.