Proceedings of the 2nd International Conference on Green Energy, Environment and Sustainable Development (GEESD2021), D. Dobrotă and C. Cheng (Eds.)

doi:10.3233/ATDE210335

A Performance Evaluation Model of Single-Cell Pseudotime Trajectory Inference Algorithms

Jiuru ZHU, Jiaqing CHEN[1], Peizheng LI and Yuanze CHEN
School of Science, Wuhan University of Technology, Wuhan 430070, China

Abstract. The study on single-cell pseudotime trajectory is of great significance to the exploration of the environmental factors of life and diseases. The large scale and complexity of single-cell data make the single-cell pseudotime trajectory algorithms face great challenges. A performance evaluation model is proposed to measure the performance of existing pseudotime trajectory inference algorithms and mine the problems existing in the inference algorithms in order to promote the improvement of the inference algorithms. Under the condition of given original single-cell data, the model uses the Spearman correlation coefficient to evaluate the performance of the inference algorithms from noise resistance and robustness. Experiments on the algorithms Monocle2 and Scout were conducted to analyze the application effect of the model.

Keywords. Single-cell data, pseudotime trajectory algorithm, performance evaluation

1. Introduction

The scale and complexity of single-cell data make it a major challenge to analyze and interpret it. How to use multi-dimensional data mining method to establish a performance evaluation model of single-cell pseudotime trajectory inference (TI) algorithms and to further promote environmental governance is a problem worth analyzing and studying.

Inference of pseudotime trajectory based on single-cell data is currently an active research field in single-cell sequencing data analysis [1]. Although many tools and algorithms for intelligent analysis of single-cell sequencing data have emerged in recent years, data analysis tools for single-cell sequencing are still far behind the speed of experimental technology development. Researches on the TI analysis model of single-cell data can better help researchers to analyze the TI methods and dig out the existing problems and causes of the TI methods, so as to promote the improvement of the TI methods. This paper focuses on the performance evaluation model of single-cell pseudotime TI algorithms.

[1] Corresponding Author, Jiaqing CHEN, School of Science, Wuhan University of Technology, Wuhan 430070, China; Email: jqchenwhut@163.com.

2. Related Works

In recent years, several single-cell pseudotime TI algorithms have been developed [2-10], which provide a basis for the study of cell fate transformation in a complex cell ecosystem. Monocle [2] is the first one. It was proposed by Trapnell et al. in 2014. It firstly projected all cells into two-dimensional space through ICA dimension reduction algorithm, then connected all landmark cells with minimum spanning tree, and finally projected all cells onto landmark cells to construct a pseudo temporal sequence of cells. Monocle2 [3] improves on Monocle by performing descending and sorting through reverse graph embedding (GRE), allowing it to detect branching events in an unsupervised manner. The method Scout [4] uses Apollonian circle projection (AP) or weighted distance (WD) to determine the time trajectory of a single cell. Other typical TI algorithms include Waterfall, Wishbone, TSCAN, PAGA, etc. By comparing the cell sequencing results of 45 TI methods and comprehensively considering the accuracy, extensibility, stability, and ease of use of the prediction results, Wouter et al. [7] proposed guidances for users to select an appropriate single-cell TI method according to the data dimension and trajectory topology structure.

3. Proposed Model

3.1. Definitions

Assume the original single-cell dataset contains *n* cells, the expression of *m* genes per cell, and the stages of each cell. The stages of each cell are arranged in sequence, from 1 to *t*.

Definition 1: Original expression matrix. Let p_{ij} be the *j*th expression value of the *i*th gene, then the original expression matrix can be expressed as $(p_{ij})_{n \times m}$.

Definition 2: Standard sequence. The stages of each cell in the original single-cell dataset are arranged in order from the smallest to the largest, which can be regarded as the standard sequence.

Definition 3: Predictive sequence. The trajectory inference is carried out based on the expression matrix using a certain inference method, and the pseudotime of each cell is obtained. The sequence formed by ordering all cells with pseudotime increment and the corresponding primitive stage of each cell is called the predictive sequence.

Definition 4: Spearman correlation coefficients. Let *X* and *Y* be two sequences, each containing *n* elements, denoted as $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$, and order the elements in and from large to small, respectively. The rank of $x_i$is denoted as $r(x_i)$, and the rank of y_i is denoted as $r(y_i)$, let $d_i = r(x_i) - r(y_i)$, then the Spearman correlation coefficient of the two sequences can be expressed as:

$$\rho = 1 - \frac{6 \times \sum_{i=1}^{n} d_i^2}{n^3 - n} \tag{1}$$

Two sequences with the same number of elements can be measured by Spearman correlation coefficient ρ. ρ is within the range of [-1,1]. For a certain inference method, the ρ value between the predicted sequence and the standard sequence is calculated. The closer the ρ is to 1, the closer the predicted sequence is to the standard sequence, the higher the accuracy of the inference method is.

3.2. Proposed Model

The proposed model consists of two parts: the noise resistance model and the robustness model.

3.2.1. Noise Resistance Model

Influenced by the experimental environment and other factors, the measurement error of gene expression level often exists in single-cell sequencing methods, which may affect the performance of different TI algorithms. Since such errors are not regular, the measurement error can be simulated by adding noise to the original dataset to evaluate the noise resistance level of the TI algorithms. The specific steps are as follows:

Step 1: For all non-0 elements in the original expression matrix $(p_{ij})_{n\times m}$, a white noise sequence $WN(0,\sigma_j{}^2)$ is added by formula (2) and a matrix $(p'_{ij})_{n\times m}$ is obtained. Where, n_j represents the number of non-0 elements in column j, k is the noise coefficient, and k will get the value of 1%, 2%, 5%, 8% respectively.

$$\sigma_j{}^2 = k \times \frac{\sum_{i=1}^{n} p_{ij}}{n_j} \tag{2}$$

Step 2: For different k, trajectory inferences are carried out based on $(p'_{ij})_{n\times m}$ using the selected TI algorithm.

Step 3: The corresponding Spearman correlation coefficients are calculated by equation (1), denoted as ρ_1, ρ_2, ρ_3, ρ_4 respectively.

Step 4: The performance of the method with increasing noise was analyzed. When the k increases, the less significant the decrease of the ρ is, then the stronger the noise resistance can be considered.

3.2.2. Robustness Model

Accidental loss of data may occur during data collection, so it is very important to evaluate the robustness of the TI algorithms. To analyze the robustness of the TI methods in the absence of cell data, the robustness evaluation is processed as follows.

Step 1: Delete all cell data with stage 1 in the original expression matrix $(p_{ij})_{n\times m}$ and calculate the ρ value based on this new matrix using the selected TI algorithm.

Step 2: Delete all cell data with stage t in the original expression matrix $(p_{ij})_{n\times m}$ and calculate the ρ value based on this new matrix using the selected TI algorithm.

Step 3: Delete all cell data with stage i $(i\neq 1$ and $i\neq t)$ in the original expression matrix $(p_{ij})_{n\times m}$ and calculate the ρ value based on this new matrix using the selected TI algorithm.

Step 4: The proportion of cells in the original expression matrix $(p_{ij})_{n \times m}$ is randomly selected and the ρ value based on this new matrix using the selected TI algorithm is calculated.

4. Application of the Proposed Model

4.1. Selected Dataset and Algorithms

The public dataset "human embryonic stem cell RNA sequencing (sRNA-seq) dataset", commonly used in the single-cell pseudotime trajectory inference, was selected for the experiments as the original dataset. The dataset contained the expression levels of 100 genes in 758 single cells, in which the cells were at the 0, 12, 24, 36, 72, and 96 hours, respectively. The entire dataset describes the development of embryonic stem cells from the initial state to the terminal endoderm. The dataset contains the standard sequence and original expression matrix of human embryonic stem cells. In the experiments, a widely used TI method called Monocle2 and a newly proposed TI method Scout were selected. The inference accuracy of the algorithm Monocle2 and Scout are very high on their default parameters with the original dataset, and the ρ values are all close to 1.

4.2. Experiments on Algorithm Monocle2

4.2.1. Analysis of Noise Resistance

When the noise coefficient k is 1%, 2%, 5%, and 8% respectively, the pseudotime trajectories of Monocle2 are shown in figures 1a-1d respectively.

The Spearman correlation coefficients under the four types of noise are ρ_1=0.9659, ρ_2=0.9627, ρ_3=0.9653, ρ_4=0.9056. It can be found that compared with the ρ=0.9666 of the original data, ρ_1, ρ_2 and ρ_3 have no significant changes, but ρ_4 is negative and its absolute value is close to 0.9. This shows that when the noise is small, the accuracy of Monocle2 is hardly affected, but when the noise coefficient k>5%, the noise may cause the trajectory to develop in the opposite direction.

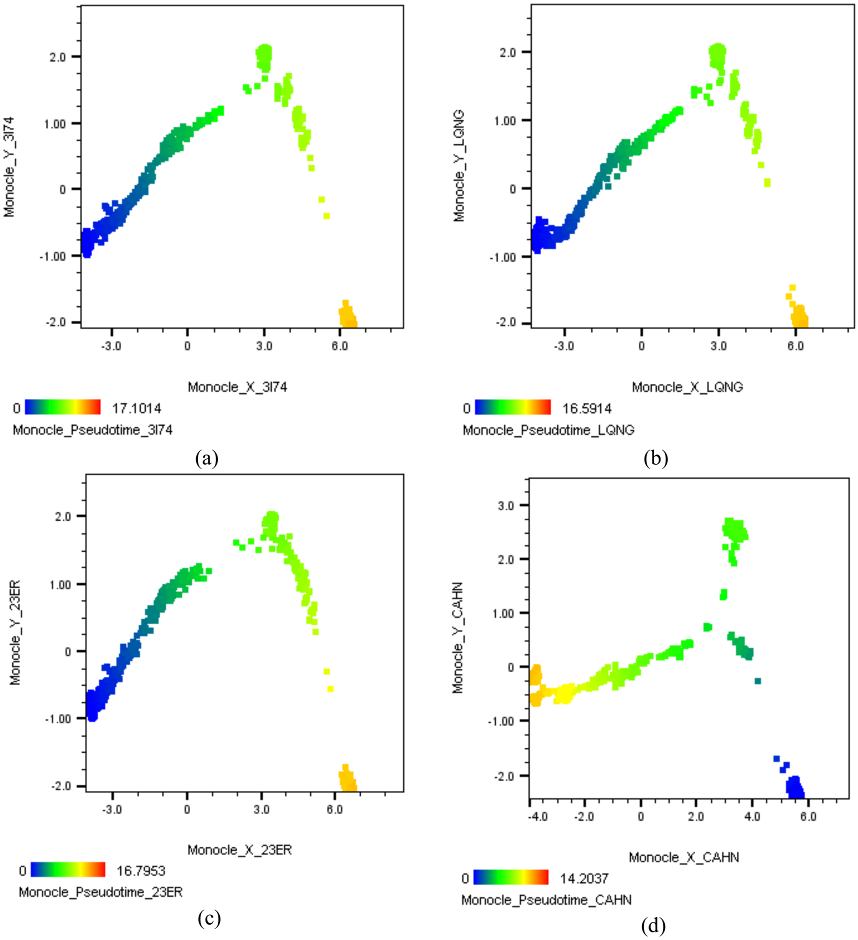
4.2.2. Analysis of Robustness

For the data set of scRNA-seq, according to the description in section 3.2.2, the data of different stages of cells were deleted respectively.

(1) Lack of cell data in the initial or final stage. Delete all the cells in the first stage and the pseudo-time trajectory result is obtained in figure 2. Delete all the cells in the final stage and the trajectory result is shown in figure 3.

As you can see, the result of removing the cell data in the first stage doesn't change much. However, the lack of cell data in the final stage may lead to wrong selection when selecting the root node, and the prediction sequence is completely wrong. The pseudotime trajectory of the deletion of the final stage is almost opposite to the original data. Therefore, Monocle2 is very sensitive to the integrity of cell stages.

(2) Lack of cell data at a certain stage in the middle. Here we choose to delete all cells in stage 4, and perform trajectory inference on the remaining cell data, and the result is shown in figure 4.



(a)

(b)

(c)

(d)

Figure 1. (a) k=1%; (b) k=2%; (c) k=5%; (d) k=8% respectively on Monocle2.

It shows that when stage 4 is deleted, the corresponding ρ value is -0.9028, and the ρ value is negative and the absolute value is close to 0.9, which indicating that the root node is misselected, and it verifies that Monocle2 is very sensitive to the integrity of the cell stage again. Comparing figure 4 with the trajectory of the final stage deletion in figure 3, it is found that the two trajectories have certain similarities, but the trajectory of the stage 4 deletion obviously lacks a lot of light green cell points while increasing a lot of orange cell points. That phenomenon is consistent with the difference between the two missing stages.

(3) Random missing of cell data at any stage. Randomly sample a proportion of certain percent cells from all cells, and select the value of q from 10% to 90% at 10% intervals, repeat the experiment 10 times for each proportion, and obtain 90 sampling datasets. Calculate the Spearman correlation coefficient ρ of each group, and the results are shown in table 1.

It can be found that the inference is completely wrong in more than half of the samples, and the number of times $\rho > 0$ does not change with the increase of q. This shows that the Monocle2 algorithm has high requirements for the continuity of the cells.

A slight discontinuity of the cells is likely to lead to the wrong selection of the root node and infer the opposite trajectory. Under the premise of $\rho > 0$, it remains around 0.9, which indicates that as long as the root node is selected correctly, the inference effect is still very good.

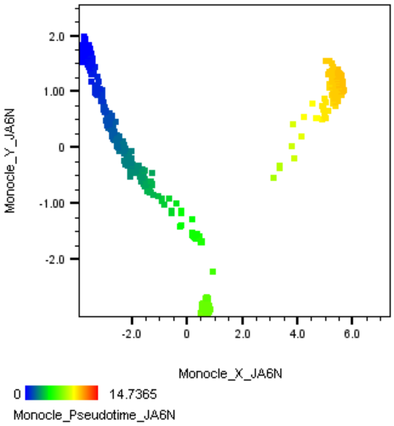


Figure 2. Lack of initial stage on Monocle2.

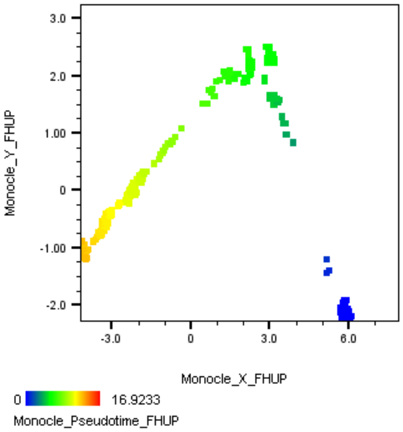


Figure 3. Lack of final stage on Monocle2.

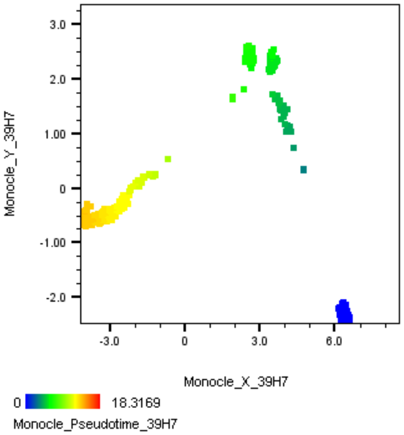


Figure 4. Lack of stage 4 on Monocle2.

Table 1. Values obtained by Monocle2 for 90 groups of sampled data.

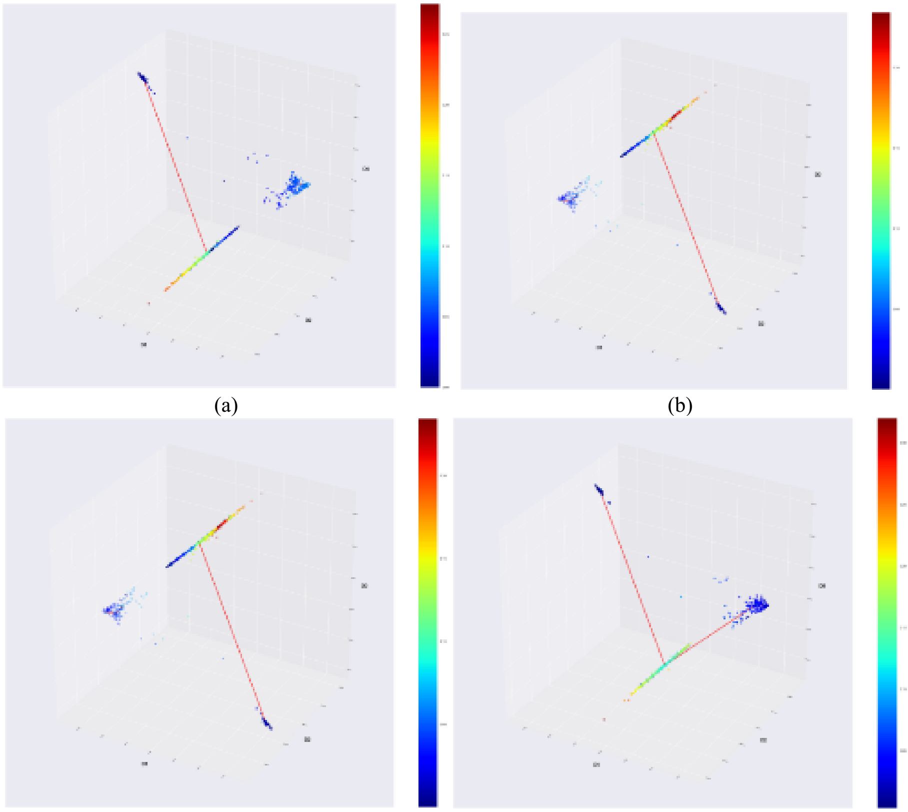
q	1	2	3	4	5	6	7	8	9	10	$\rho > 0$
10%	-0.88	-0.92	0.98	-0.85	-0.90	-0.89	0.83	-0.85	0.92	-0.87	3
20%	-0.89	0.94	-0.87	-0.91	0.87	-0.85	0.89	0.92	0.91	-0.93	5
30%	-0.89	0.92	-0.88	-0.90	0.90	-0.91	0.90	0.92	-0.87	0.90	5
40%	-0.90	0.90	-0.87	-0.88	0.89	0.95	-0.88	0.72	-0.88	0.95	5
50%	-0.89	-0.90	-0.90	-0.90	0.92	-0.88	0.93	0.91	-0.89	-0.89	3
60%	-0.90	0.93	-0.90	-0.91	0.94	-0.89	-0.92	0.95	-0.88	0.94	4
70%	-0.91	0.88	-0.91	0.93	-0.92	-0.91	-0.92	-0.91	-0.91	0.95	3
80%	-0.91	0.92	0.96	-0.92	-0.92	0.95	-0.92	0.94	0.95	0.96	6
90%	-0.90	0.96	-0.92	0.96	-0.92	-0.92	-0.92	-0.92	0.96	-0.91	3

4.3. Experiments on Algorithm Scout

4.3.1. Analysis of Noise Resistance

The process is the same as the description in section 3.2.2. When the noise coefficient k is 1%, 2%, 5%, and 8% respectively, the pseudotime trajectories of Scout are shown in figures 5a-5d respectively.

The Spearman correlation coefficients under the four types of noise are ρ_1=0.7567, ρ_2=0.7221, ρ_3=0.7476, ρ_4=0.5663. It can be found that compared with the original data, the inference effects after adding noise have been significantly reduced, even if the noise is very small. It shows that Scout is very sensitive to noise. When the noise coefficient k is 1%, 2%, and 5%, the difference is not obvious, and they are all greater than 0.7. It indicates that when the noise coefficient is within 5%, the inference method is still meaningful. When the noise is increased to k_4=8%, the value of ρ drops significantly and approaches 0.5, indicating the failure of the inference method.



(a) (b) (c) (d)

Figure 5. (a) k=1%; (b) k=2%; (c) k=5%; (d) k=8% respectively on Scout.

4.3.2. Analysis of Robustness

For the data set of scRNA-seq, according to the description in section 3.2.2, the data of different stages of cells were deleted respectively to analyze the robustness of the Scout algorithm when the type of data missing is different.

(1) lack of cell data in the initial or final stage. Delete all the cells in the first stage and the pseudo-time trajectory result is obtained in figure 6. Delete all the cells in the final stage and the trajectory result is shown in figure 7.

After calculation, we got ρ=0.73 when the first stage is deleted, and ρ=0.81 when the final stage is deleted. Although there is a certain gap between the two from the original data that ρ=0.9519, the calculated value is still relatively large, and there is a certain value in using the algorithm for trajectory inference.

(2) Lack of cell data at a certain stage in the middle. Also choose to delete all cells in stage 4, and perform trajectory inference on the remaining cell data, and the resulting pseudo-time trajectory is shown in figure 8. After calculation, when stage 4 is deleted, we got ρ=0.48, which is far from the original data result that ρ=0.9519. In this case, the inference method is completely ineffective. It shows that the lack of cell data at a certain stage in the middle will cause the cells in the stages before and after it to be unable to function normally.

(3) Random missing of cell data at various stages. Sampling is performed in the same way as in section 4.2.2 to obtain 90 datasets of cell sampling data. The results are shown in figures 9-11 respectively.

It can be found that when q <=50%, the corresponding average value of ρ fluctuates around 0.4, indicating that if the number of randomly selected cells is less than 50%, the inference effect is not good. When q >=50%, the corresponding average value of ρ is greater than 0.5, and when the value of q changes from 60% to 80%, the average value of ρ has a gradual upward trend, indicating that when the number of cells extracted is more than 50 %, the inference effect is better, and the more cells, the better the inference effect is. At the same time, the maximum value of ρ in 90 random samples is 0.85, which is a certain gap from the ρ=0.9519 corresponding to the original data, indicating that the Scout algorithm has high requirements for data integrity, and a slight lack of data will have a big difference to the outcome.

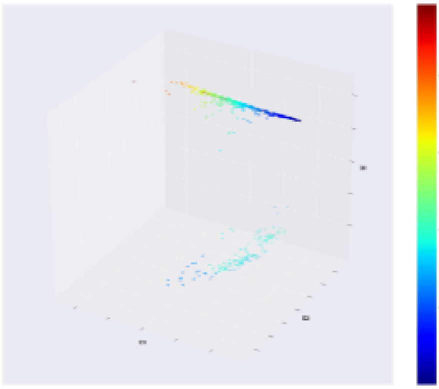


Figure 6. Lack of initial stage on Scout.

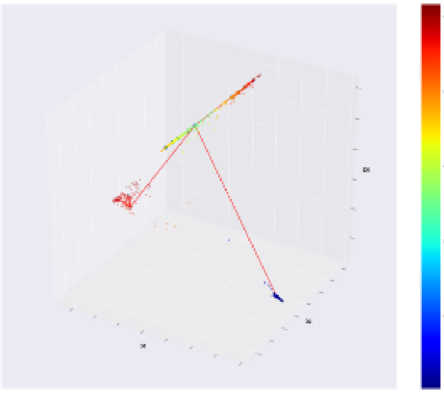


Figure 7. Lack of final stage on Scout.

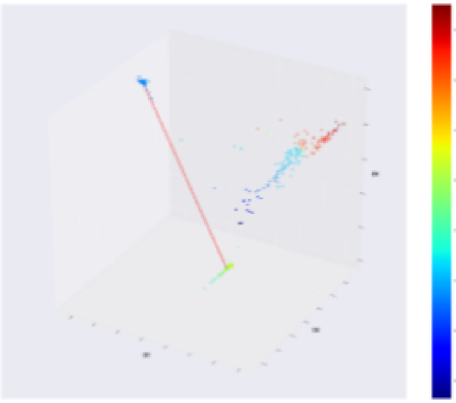


Figure 8. Lack of stage 4 on Scout.

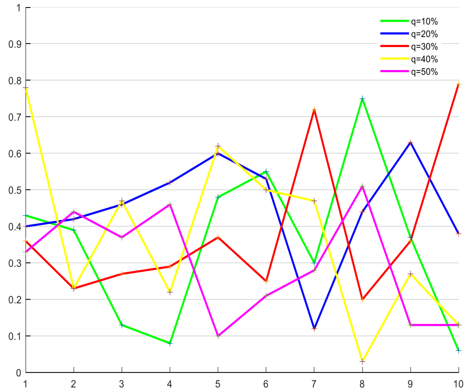
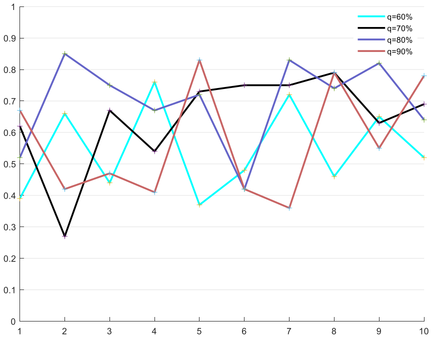


Figure 9. The connection line at q values of 10%-50%.



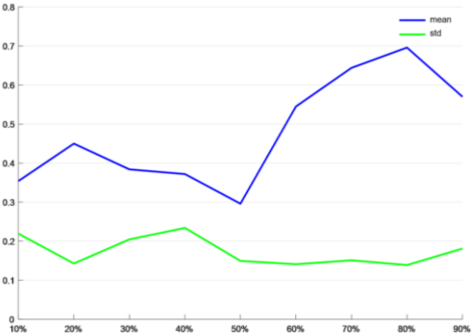


Figure 10. The connection line at q values of 60%-90%.

Figure 11. Variation of the mean value and standard deviation of ρ with the change of q.

5. Conclusions

An evaluation model of pseudotime TI algorithms was proposed based on multi-dimensional data mining. The model evaluates the pseudotime TI methods from two aspects: noise resistance and robustness. Experiments were designed to apply the model on TI methods Monocle2 and Scout, and analyses were carried out on the characteristics of a certain TI method corresponding to the inference results. It was found that the inference effect could be improved by artificial experience. In the next step, deep learning models and algorithms can be designed based on artificial experience to learn the selection method of the root node or the adjustment method of a developmental branch, so as to make the pseudotime TI algorithms more automatic.

References

[1] Skylaki S, Hilsenbeck O and Schroeder T 2016 *Nat. Biotechnol.* **34** 1137-44.

[2] Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon N J, Livak K J, Mikkelsen T S and Rinn J L 2014 *Nat. Biotechnol.* **32** 381-6.

[3] Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner H A and Trapnell C 2017 Reversed graph embedding resolves complex single-cell trajectories *Nat. Methods* **14** 979-82.

[4] Wei J, Zhou T, Zhang X and Tian T 2019 Scout: A new algorithm for the inference of pseudo-time trajectory using single-cell data *Comput. Biol. Chem*. **80** 111-20.

[5] Wolf F A, Hamey F K, Plass M, Solana J, Dahlin J S, Göttgens B, Rajewsky N, Simon L and Theis F J 2019 *Genome Biol*. **20** 59.

[6] Ji Z and Ji H 2016 *Nucleic Acids Res.* **44** e117.

[7] Saelens W, Cannoodt R, Todorov H and Saeys Y 2019 *Nat. Biotechnol.* **37** 547-54.

[8] Lönnberg T, et al. 2017 *Sci. Immunol.* **2** eaal2192.

[9] Haghverdi L, Büttner M, Wolf F A, Buettner F and Theis F J 2016 *Nat. Methods* **13** 845-8

[10] Reid J E and Wernisch L 2016 *Bioinformatics* **32** 2973-80.