Fundamental Research 4 (2024) 770–776

Contents lists available at ScienceDirect

Fundamental Research

journal homepage: http://www.keaipublishing.com/en/journals/fundamental-research/

Article

An increment of diversity method for cell state trajectory inference of time-series scRNA-seq data

Yan Hong[1], Hanshuang Li[1], Chunshen Long, Pengfei Liang, Jian Zhou, Yongchun Zuo*

State Key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock, Institutes of Biomedical Sciences, College of Life Sciences, Inner Mongolia University, Hohhot 010020, China

ARTICLE INFO

Article history:
Received 4 July 2023
Received in revised form 29 August 2023
Accepted 3 January 2024
Available online 9 February 2024

Keywords:
Increment of diversity
Time-series scRNA-seq data
Cell state trajectory inference
Topology similarity
Branching accuracy

ABSTRACT

The increasing emergence of the time-series single-cell RNA sequencing (scRNA-seq) data, inferring developmental trajectory by connecting transcriptome similar cell states (i.e., cell types or clusters) has become a major challenge. Most existing computational methods are designed for individual cells and do not take into account the available time series information. We present IDTI based on the Increment of Diversity for Trajectory Inference, which combines time series information and the minimum increment of diversity method to infer cell state trajectory of time-series scRNA-seq data. We apply IDTI to simulated and three real diverse tissue development datasets, and compare it with six other commonly used trajectory inference methods in terms of topology similarity and branching accuracy. The results have shown that the IDTI method accurately constructs the cell state trajectory without the requirement of starting cells. In the performance test, we further demonstrate that IDTI has the advantages of high accuracy and strong robustness.

1. Introduction

Cell development and differentiation is a dynamic process, which is the basis of studying ontogenesis in multicellular organisms [1]. The scRNA-seq is an excellent technique for studying cell fate, allowing transcription analysis to reveal the underlying developmental dynamics, cell communication, gene regulation and disease development [2]. The analysis of trajectory inference can verify known cell differentiation relationships and reveal cell development trajectories. In particular, reconstructing cell state trajectories between adjacent time points is key to analyzing transcriptional dynamics over time [3,4]. At present, it remains a challenge to accurately infer the cell state trajectory of time-series scRNA-seq data.

In recent years, a series of trajectory inference methods based on scRNA-seq data have been developed [5–9]. In 2014, Trapnell et al. proposed Monocle to construct Minimum Spanning Tree (MST) based on transcriptome similarity to infer cell trajectory, which was a pioneering trajectory inference method [10–12]. La Manno et al. proposed RNA velocity to infer the direction and speed of cell differentiation based on the spliced and unspliced mRNAs [13]. Schiebinger et al. developed the landmark work Waddington-OT based on the principle of using the optimal transport framework to model cell development in dynamic processes [14]. Setty et al. and Stassen et al. presented Palantir and VIA respectively, both of whom applied Markov chain to single cell pseudotime analysis [15,16]. Saelens et al. developed Dyno to integrate and evaluate more than 70 trajectory inference methods as of 2019 [17]. These computational methods have emerged to meet different needs. However, most of the existing trajectory inference methods have been designed for individual cells, ignoring the importance of cell state trajectory inference, and forgetting the available time series information. In the last several years, there have also been approaches to infer cell state trajectories by combining temporal information. For example, CSHMM utilized a continuous state Hidden Markov Model (HMM) to reconstruct continuous cell state trajectory [18]. Tempora combined biological pathways to identify cell types and incorporated temporal information to infer cell state trajectories [19]. CStreet constructed *k*-nearest neighbor connections of cells within each time point and between adjacent time points, and then used force-directed graphs to estimate the connection probability of cell states [20]. GraphFP is a nonlinear Fokker-Planck equation based on graph model and dynamic inference framework, which can reconstruct the cell state transition potential energy landscape [21].

Here, we present IDTI, which for the first time utilizes increment of diversity to cell state trajectory inference. It develops for time-series scRNA-seq data, so gene expression matrix with time series information

* Corresponding author.
E-mail address: yczuo@imu.edu.cn (Y. Zuo).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.fmre.2024.01.020

is used as input. IDTI trajectory inference includes identification of cell states, sectionalization data based on time points, calculation of increment of diversity, determination of the relationship between cell states, visualization of the inferred trajectory and so on. Through application and comparison, we conclude that IDTI method has high accuracy and robustness. Thus, the trajectories inferred by IDTI can reflect real developmental relationships and help to understand and explain the process of cellular identity transformation.

2. Materials and methods

2.1. Data collection and preprocessing

We tested IDTI on simulated and several real time-series scRNA-seq datasets. The simulated dataset has been generated by Splatter [22], an *R* package for the simple simulation of scRNA-seq data. The real datasets have been available in the Gene Expression Omnibus (GEO) database under accession code GSE98150 [23], GSE90047 [24] and GSE107122 [25], which are mouse early embryonic development, mouse hepatoblast differentiation and mouse cerebral cortex development respectively. The gene expression matrix with time series information as an input to IDTI needs to be prepared in h5ad file format. Cell state labels have been given for the datasets used in our study; if not, which can be obtained from *cell_clusters* function of IDTI. The data preprocessing includes: filtering low-count genes and cells by function *sc.pp.filter_genes* and *sc.pp.filter_cells*, and normalizing data by function *sc.pp.normalize_total* and *sc.pp.log1p*. Here, we have tried to select different amounts of highly variable genes using function *sc.pp.highly_variable_genes* for downstream analysis. Then, the data have been normalized by function *MinMaxScaler*, which was scaled to a positive value between 0 and 1 to facilitate the calculation of the logarithmic function in subsequent analyses. See the code section at https://github.com/hy-1994/IDTI for specific parameters.

2.2. Methods

2.2.1. The measure of diversity

As early as 1978, Laxton proposed the concept of measure of diversity [26], which was applied in the geographical distribution of biological species. For the high-dimensional gene expression space $S = \{X_1, X_2, ..., X_n\}$, which is composed of n cell states. Let $X \in S$, x_i denotes the sum of gene expression values of i^{th} dimension of cell state X. The measure of diversity of $X : [x_1, x_2, ..., x_m]$ is defined as

$$D(X) = N_X \log_b N_X - \sum_{i=1}^{m} x_i \log_b x_i \tag{1}$$

where $N_X = \sum_{i=1}^{m} x_i$ is sum expression values of each x_i in X; b is the given base of logarithm, which is e; if $x_i = 0$, then $\log_b x_i = 0$. Similarly, when we have another cell state $Y : [y_1, y_2, ..., y_m]$, $D(Y)$ can be defined as Eq. 2, where $N_Y = \sum_{i=1}^{m} y_i$ is also sum expression values of every y_i in Y.

$$D(Y) = N_Y \log_b N_Y - \sum_{i=1}^{m} y_i \log_b y_i \tag{2}$$

Here, we can see that the measure of diversity is highly similar to information entropy [27,28], both are descriptions of state space from the perspective of information, and the basis of measurement is the logarithmic function measured according to information. However, the meanings between them are different: Information entropy is a description of state uncertainty or disorder; while the measure of diversity is a description of the overall uncertainty. Greater information entropy implies a large degree of uncertainty, but not necessarily a large measure of diversity. Conversely, a higher measure of diversity does not necessarily indicate greater disorder.

2.2.2. The increment of diversity

Furthermore, the measure of diversity is extended to the concept of the increment of diversity (ID), which can quantitatively represent biological similarity. Subsequently, the ID has been widely used in the field of bioinformatics, especially in the study of biological classification [29–32]. The ID between $X : [x_1, x_2, ..., x_m]$ and $Y : [y_1, y_2, ..., y_m]$ is calculated as

$$\mathrm{ID}(X, Y) = D(X + Y) - D(X) - D(Y) \tag{3}$$

The smaller the value of $\mathrm{ID}(X, Y)$, the higher the similarity between X and Y. $D(X + Y)$ can be calculated as

$$D(X + Y) = (N_X + N_Y)\log_b(N_X + N_Y) - \sum_{i=1}^{m} (x_i + y_i)\log_b(x_i + y_i) \tag{4}$$

We then apply ID to cell state trajectory inference of time-series scRNA-seq data, on the premise that we need to determine standard sources. Given the available time information, we always regard cell states at the previous moment as the standard sources of cell states at the later moment. For example, there are standard sources STD_k (k = 1, 2, …, K) at time point T_i, where K is the number of standard sources, then the relationship between standard sources and cell states $Z \in S$ at time point T_{i+1} is determined by the minimum increment of diversity algorithm, and the decision principle is

$$\mathrm{ID}(Z, STD_k) = min\{\mathrm{ID}(Z, STD_1), \mathrm{ID}(Z, STD_2), ..., \mathrm{ID}(Z, STD_K)\} \tag{5}$$

2.2.3. Calculation of graph edit distance and F_1 score

Here, we evaluate IDTI by comparing its inferred trajectory to known trajectory manually curated from the literature. Two approaches are used to evaluate the topology similarity and branching accuracy between trajectories: graph edit distance (GED) score and F_1 score.

The GED score is degree of similarity between two graphs G_1 and G_2 [33], which is defined as follows:

$$\mathrm{GED}(G_1, G_2) = min\left\{ \sum_{e_j \in \gamma(G_1, G_2)} c(e_j) \right\} \tag{6}$$

where $\gamma(G_1, G_2)$ denotes all the complete edit paths from graph G_1 to G_2, and $c(e_j)$ denotes the edit cost of the edit operation e_j. The deletion of nodes V and edges E, and the substitution of nodes V constitute a complete edit path γ_i, and the score of each edit operation is defined as 1. If the sum of the cost values of this path is the smallest, this cost is the edit distance between graphs. The GED score is calculated using the function *nx.graph_edit_distance*(G_1, G_2) in *NetworkX* [34]. The closer the GED score between two trajectories is to 0, the higher the similarity between them.

The F_1 score is the harmonic mean of precision and recall of trajectory directed edge identification [35,36]. A true positive (TP)/false positive (FP) edge in the inferred trajectory is an edge that actually exists/does not exist in the gold trajectory. A False Negative (FN) edge in the inferred trajectory is when there is an edge in the gold trajectory between cell states that is absent in the inferred trajectory. The precision is calculated as the ratio of the number of TP edges to the total number of predicted edges (the sum of TP edges and FP edges). The recall is calculated as the ratio of the number of TP edges to the number of all real edges (the sum of TP edges and FN edges) [37]. The range of F_1 score is [0, 1], and the higher the F_1 score between two trajectories indicates the higher branching accuracy.

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{8}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{9}$$

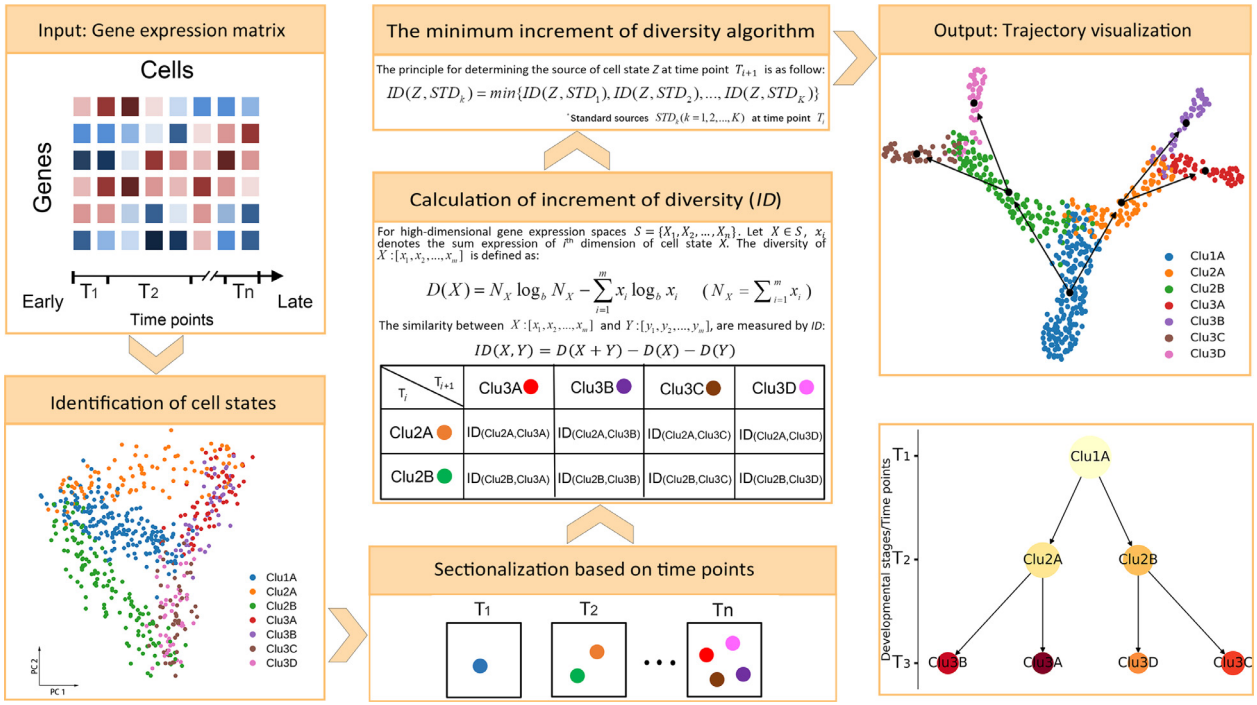

Fig. 1. The workflow of IDTI.

3. Results and discussion

3.1. Overview of IDTI

IDTI infers the cell state trajectory from the expression matrix of time-series scRNA-seq data. IDTI first performs the identification of cell states and then sectionalizes the data based on available time information. Most importantly, the calculation of the ID and the inference of developmental relationships between cell states. We calculate the ID between cell states at adjacent time points to represent the similarity between cell states, and then determine the development trajectory by the minimum increment of diversity algorithm. In the end, we visualize trajectories through Uniform Manifold Approximation and Projection (UMAP) plot and directed graph, of which the UMAP shows the relationships between cell states in the form of individual cells, and the directed graph shows the hierarchical structure of evolutionary relationships (Fig. 1).

3.2. Application of IDTI on simulated dataset

We used function *splatsimulatePath* in Splatter to generate the simulated time-series scRNA-seq data with continuous trajectory. We have gotten a simulated dataset including 600 cells and 10,000 genes at three time points (T_1, T_2, T_3). These cells can be classified into seven categories, of which the T_1 stage contains the cell type *Clu1A*, the T_2 stage contains two cell types *Clu2A* and *Clu2B*, and the T_3 stage contains four cell types *Clu3A, Clu3B, Clu3C* and *Clu3D* (Fig. 2a). The known development trajectories are *Clu1A–Clu2A–Clu3A, Clu1A–Clu2A–Clu3B, Clu1A–Clu2B–Clu3C, Clu1A–Clu2B–Clu3D*, and we manually draw the trajectories (Fig. 2b). As shown in UMAP plot and directed graph, the IDTI can accurately reconstruct the four developmental trajectories for the simulated data, which are same as the developmental trajectory of the simulation (Fig. 2c,d).

3.3. Application of IDTI on real time-series scRNA-seq datasets

First, we applied IDTI on the time-series scRNA-seq dataset of mouse early embryonic development, which contains 40 single cells from eight stages on embryos from MII Oocyte to embryonic day 6.6 (E6.6), and we manually mapped the gold standard developmental trajectory (Fig. 3a). IDTI was able to successfully predict the different trajectories of trophoblast ectoderm (TE) and inner cell mass (ICM), and the ICM continued to gradually differentiate into extraembryonic ectoderm (Exe) and epiblast (Epi) (Fig. 3b,c).

Next, we applied IDTI on the time-series scRNA-seq dataset of mouse hepatoblast differentiation, consisting of 447 single cells collected at embryos (E10.5-E17.5). Here, we used the strategy proposed by Yang, et al. [24] to annotate the cells, labeling as “hepatoblast” at early time points (E10.5, E11.5), “hepatoblast/hepatocyte” at intermediate time points (E12.5, E13.5, E14.5), “hepatocyte” and “cholangiocyte” at late time points (E15.5, E17.5). We also manually mapped the gold standard developmental trajectory (Fig. 3d). IDTI also successfully inferred developmental trajectories of hepatoblast differentiation through intermediate cells into hepatocyte and cholangiocyte cells (Fig. 3e,f).

Finally, in order to evaluate the performance of IDTI with multiple cell states at each time point, we applied IDTI to the time-series scRNA-seq dataset during mouse cerebral cortex development, which contains 6316 cells collected at E11.5, E13.5, E15.5 and E17.5. These cells have covered a wide range of neuronal development, from early precursors (apical precursors (APs) and radial precursors (RPs)) to intermediate precursors (IPs) and differentiated cortical neurons. Tran et al. [19] used GSVA and the marker genes of APs, RPs, IPs, young neurons and neurons to automatically annotate the seven clusters. Meanwhile, we manually mapped the gold standard developmental trajectory through literature search [21] (Fig. 3g). IDTI inferred two trajectories rooted in APs/RPs, one trajectory branching into young neuron cells and neuron cells via IPs, and the other trajectory converging in the cluster of neuron cells. Unfortunately, it was not able to infer trajectories of APs/RPs to young neuron cells and neuron cells to young neuron cells (Fig. 3h,i).

3.4. Comparison of IDTI with other trajectory inference methods

Here, we compared IDTI with six other trajectory inference methods (i.e., Monocle 2, TSCAN [38], Slingshot [39], PAGA [40], Tempora and CStreet) on the simulated dataset, the mouse early embryonic de-

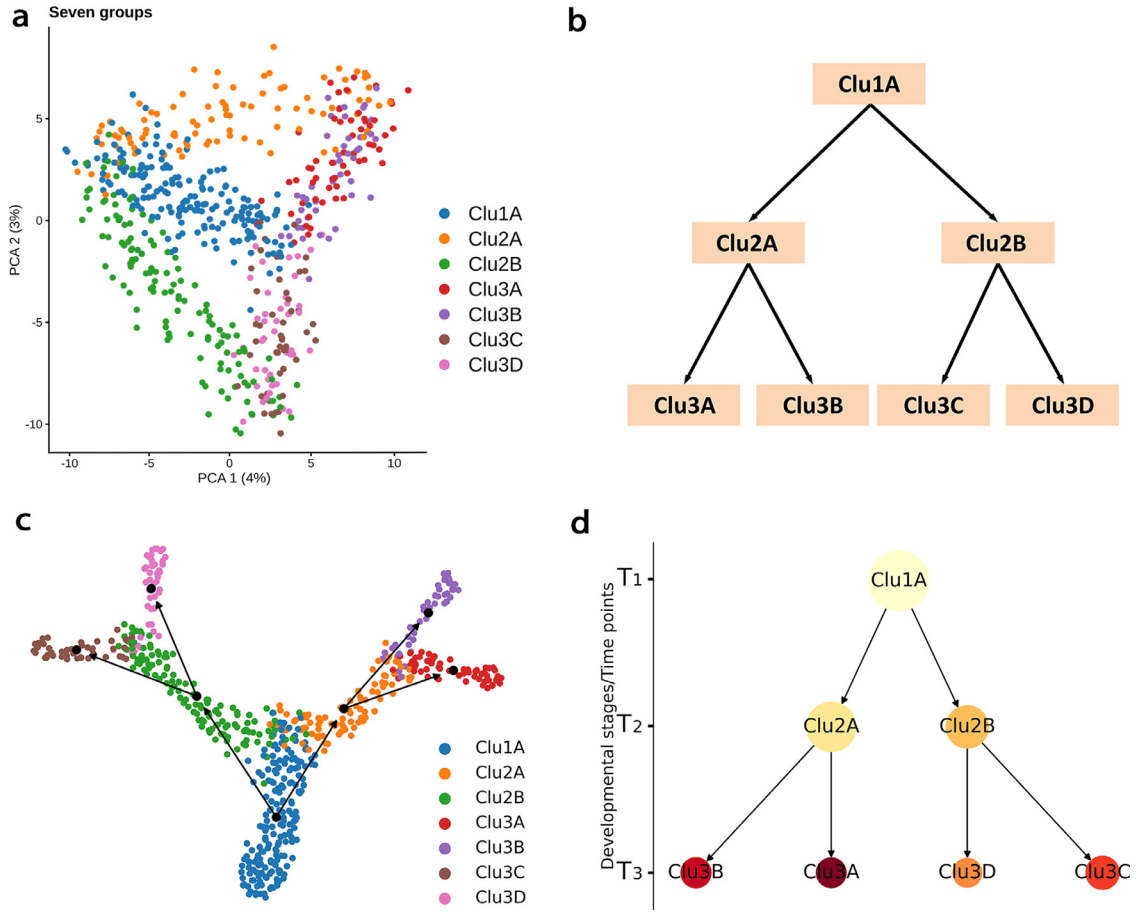

Fig. 2. IDTI analysis of the simulated dataset. (a) Scatter plot showing the visualization of Principal Component Analysis (PCA) dimensional reduction of the simulated data. Different cell states are plotted by different colors. (b) The gold trajectory of the simulated data is used to evaluate the accuracy of the inferred trajectories. (c) UMAP plot showing the cell state trajectory inferred from the simulated dataset using IDTI. Each node represents a single cell, and which are colored by cell states. The black nodes indicate the center of cell states, and the arrows connecting them represent the cell state trajectory. (d) Directed graph showing the hierarchy of cell state trajectory of the simulated dataset using IDTI. The timeline on the left represents developmental stages or time points. Circles represent cell states, of which the relative size represents cell population and the relative color depth represents the increment of diversity between the cell states at adjacent time points.

Table 1
Comparison between IDTI and other methods on the topology similarity.

Methods	GED score			
	Simulated data	Mouse Embryo data	Mouse Hepatoblast data	Mouse Cerebral Cortex data
IDTI	0	0	0	2
Monocle 2	5	8	2	4
TSCAN	2	9	0	3
Slingshot	5	10	-[a]	2
PAGA	5	9	2	3
Tempora	-[a]	-[a]	2	1
CStreet	1	-[a]	0	0

[a] - represents that the result of the corresponding method is not available.

velopment dataset, the mouse hepatoblast differentiation dataset and the mouse cerebral cortex development dataset. In order to facilitate comparison, we formalized all the inferred trajectories to graph (or network), of which nodes represent cell states and edges represent the relationship between two cell states. In addition to IDTI and CStreet, other methods need to manually determine the starting cells of the trajectory. Here, we evaluated the results using two metrics: the GED score, which was used to evaluate the similarity between the inferred trajectory and the gold trajectory, and the F_1 score, which was used to evaluate the branching accuracy between the inferred trajectory and the gold trajectory.

On the simulated dataset, IDTI can accurately infer all the four development trajectories (Fig. 2c,d). Here, we assigned *Clu1A* as the trajectory starting cells for the other methods. Monocle 2 inferred pseudotime trajectories based on individual cells, but its cells showed confusion (Fig. 4a). TSCAN can infer two main trajectories, unable to correctly infer the development trajectory of the terminal cell states (Fig. 4b). Slingshot cannot construct the real bifurcated trajectory, only linear trajectory (Fig. 4c). However, PAGA constructed a coarse-grained diagram, which contains six connections, and the results couldn't construct the main development trajectory (Fig. 4d). CStreet was relatively accurate, except that the trajectory *Clu1A–Clu2A* cannot be inferred

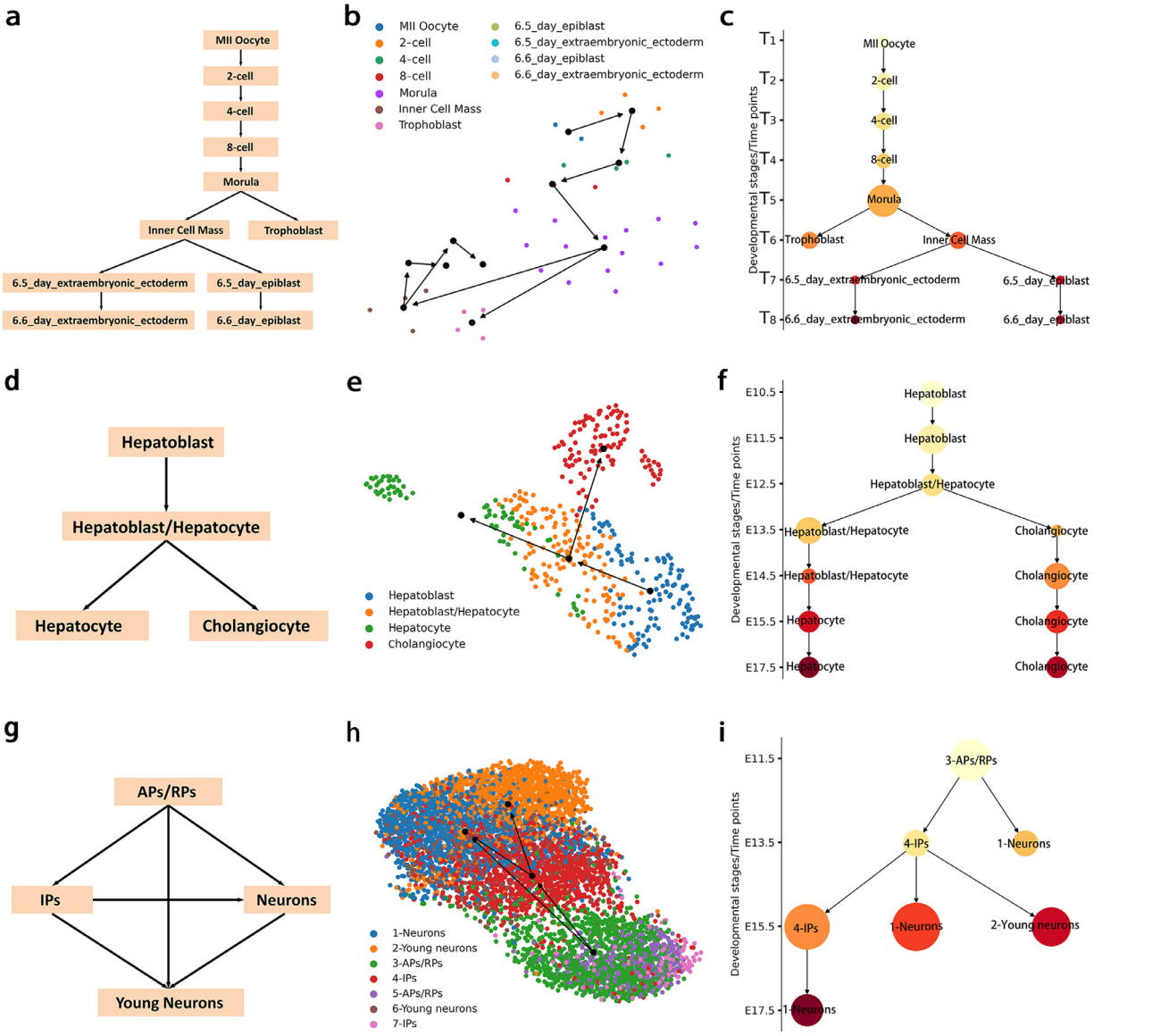

Fig. 3. IDTI applies of the real time-series scRNA-seq datasets. The mouse early embryonic development time-series scRNA-seq dataset: (a) the gold trajectory (b) UMAP plot and (c) directed graph showing cell state trajectory inferred by IDTI. The mouse hepatoblast differentiation time-series scRNA-seq dataset: (d) the gold trajectory (e) UMAP plot and (f) directed graph showing cell state trajectory inferred by IDTI. The mouse cerebral cortex development time-series scRNA-seq dataset: (g) the gold trajectory (h) UMAP plot and (i) directed graph showing cell state trajectory inferred by IDTI.

(Fig. 4e). Tempora was not used for simulated data because pathway information is required. The trajectory inferred by the IDTI method was exactly the same as the gold trajectory, of which GED score is 0 and F_1 score is 1. In conclusion, the results on the simulated dataset showed that IDTI outperforms all other methods in the topology and branching accuracy, followed by CStreet, TSCAN, and finally PAGA, Monocle 2, Slingshot (Tables 1, 2).

Similarly, we also made comparisons on the three real datasets. On the mouse early embryonic development time-series scRNA-seq dataset, we assigned MII Oocyte as the starting cells for other methods. The comparison results showed that CStreet and Tempora failed to complete the trajectory construction, and the other methods IDTI did best (Figs. 3c, S1). Hepatoblast was considered as the starting cells on the mouse hepatoblast differentiation time-series scRNA-seq dataset, IDTI, TSCAN and CStreet could accurately infer the real trajectory, in which Slingshot failed to infer the trajectory (Figs. 3f, S2). On the mouse cerebral cortex time-series scRNA-seq dataset, we assigned Aps/RPs as the starting cells. The results displayed that CStreet outperformed optimally, and Tempora failed to infer the trajectory of young neuron cells to neuron cells. The performance of IDTI was second only to CStreet and Tempora, where GED score of IDTI is 2 and F_1 score is 0.8, unable to infer from IPs to neuron cells and young neuron cells to neuron cells (Figs. 3i, S3). The results on real datasets show that IDTI performs best in topology similarity and branching accuracy on the mouse early embryonic development dataset, and the mouse hepatoblast development dataset, and only performs slightly worse on the mouse cerebral cortex dataset, but it is still acceptable (Tables 1, 2).

In summary, the IDTI is the first to utilize the increment of diversity to infer the trajectory for time-series scRNA-seq data, and which can reconstruct relatively accurate trajectories without the need to define the starting cells.

3.5. Evaluation of IDTI performances

To further evaluate the robustness of IDTI, we randomly perturbed the simulated dataset in two ways: different cell sampling rates and different gene dropout rates. Among them, the cell sampling rates had been set at 90%, 80% and 70% respectively, and the selection was performed

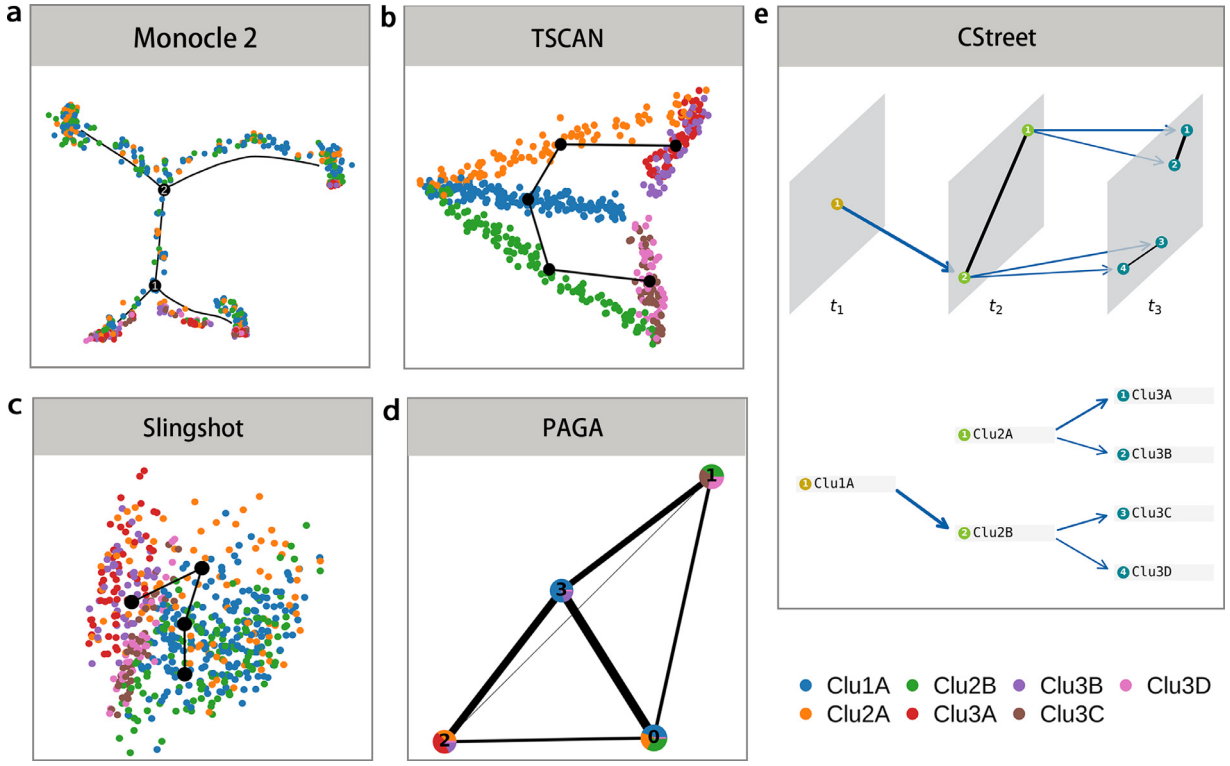

Fig. 4. Comparison of IDTI with other trajectory inference methods on the simulated dataset. (a) Monocle 2 (b) TSCAN (c) Slingshot (d) PAGA (e) CStreet.

Table 2
Comparison between IDTI and other methods on the branching accuracy.

Methods	F_1 score			
	Simulated data	Mouse Embryo data	Mouse Hepatoblast data	Mouse Cerebral Cortex data
IDTI	1.00	1.00	1.00	0.80
Monocle 2	0.00	0.17	0.67	0.50
TSCAN	0.80	0.31	1.00	0.67
Slingshot	0.00	0.00	-[a]	0.80
PAGA	0.22	0.15	0.67	0.67
Tempora	-[a]	-[a]	0.80	0.91
CStreet	0.91	-[a]	1.00	1.00

[a] - represents that the result of the corresponding method is not available and all values are reserved to two decimal places.

5 times with different random seeds for each number. The gene dropout rates had been set at 10%, 20%, 30%, 40% and 50% respectively, and the selection was performed 3 times with different random seeds for each number. Therefore, we generated a total of 30 perturbed datasets, and constructed trajectories using IDTI. As a result, there are no differences between the trajectories constructed by IDTI on the perturbed and original datasets. Specifically, IDTI still showed reliable results when the cell sampling rate was as low as 70% or the gene dropout rate was as high as 50%. Therefore, changes in cell number and gene dropout rate within a certain range have no effect on IDTI trajectory inference. To sum up, the IDTI is also reliable on datasets with small cell numbers and high gene dropout rates, indicating that the IDTI has high robustness.

4. Conclusion

With the development of sequencing technology, many computational methods of trajectory inference have been proposed. Meanwhile, time series experiments provide available temporal information to trajectory inference. We present IDTI, which makes full use of the time series information, and utilizes increment of diversity for cell state trajectory inference. The IDTI is an effective trajectory reconstruction method, which can reproduce the process of cell state transformation.

The time series information is very important to the performance of IDTI, which provides direction for the trajectory. The IDTI analyses the time-series scRNA-seq data at the level of cell states, not at the individual cell. Compared with inferred trajectories based on single cells, the advantage of cell state trajectory inference is to avoid single cells in the same cell state, and assign to different branches. The IDTI also performs well compared with other six commonly used trajectory inference methods in simulated and real datasets, and IDTI doesn't need to assign the starting cells. Furthermore, the IDTI is highly robust over datasets with different sampling rates and different dropout rates. In summary, the IDTI is a computational method for cell trajectory inference using time-series scRNA-seq data, which provides an easy and accurate way to understand and interpret transition process of cell identity.

Availability

IDTI is written in python and freely available at https://github.com/hy-1994/IDTI.

Declaration of competing interest

The authors declare that they have no conflicts of interest in this work.

Acknowledgments

We thank Prof. Shaorong Gao (Tongji University) for sharing the scRNA-seq data of mouse early embryonic development, Prof. Chengran Xu (Peking University) for sharing the scRNA-seq data of mouse hepatoblast differentiation and Prof. Freda Miller (Toronto University) for sharing the scRNA-seq data of mouse cerebral cortex development in the GEO database. We thank Prof. Yong Zhang (Tongji University) for sharing the python code of the CStreet method on Github. We also thank Prof. Gary Bader (Toronto University) for annotating and preprocessing mouse cerebral cortex development. This work was supported by the National Natural Science Foundation of China (62061034, 62171241), the key technology research program of Inner Mongolia Autonomous Region (2021GG0398) and the Science and Technology Leading Talent Team in Inner Mongolia Autonomous Region (2022LJRC0009).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.fmre.2024.01.020.

References

[1] L. Zheng, P. Liang, C. Long, et al., EmAtlas: A comprehensive atlas for exploring spatiotemporal activation in mammalian embryogenesis, Nucleic Acids Res. 51 (D1) (2023) D924–d932.
[2] H. Li, C. Long, Y. Hong, et al., Characterizing cellular differentiation potency and waddington landscape via energy indicator, Research (Wash D C) 6 (2023) 0118.
[3] J. Ding, N. Sharon, Z. Bar-Joseph, Temporal modelling using single-cell transcriptomics, Nat. Rev. Genet. 23 (6) (2022) 355–368.
[4] R. Cannoodt, W. Saelens, Y. Saeys, Computational methods for trajectory inference from single-cell transcriptomics, Eur. J. Immunol. 46 (11) (2016) 2496–2506.
[5] H. Chen, L. Albergante, J.Y. Hsu, et al., Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM, Nat. Commun. 10 (1) (2019) 1903.
[6] Z. Wang, Y. Zhong, Z. Ye, et al., MarkovHC: Markov hierarchical clustering for the topological structure of high-dimensional single-cell omics data with transition pathway and critical point detection, Nucleic Acids Res. 50 (1) (2022) 46–56.
[7] S. Rashid, D.N. Kotton, Z. Bar-Joseph, TASIC: Determining branching models from time series single cell data, Bioinformatics 33 (16) (2017) 2504–2512.
[8] J. Xie, Y. Yin, J. Wang, TIPD: A probability distribution-based method for trajectory inference from single-cell RNA-Seq data, Interdiscip. Sci. 13 (4) (2021) 652–665.
[9] M. Guo, E.L. Bao, M. Wagner, et al., SLICE: Determining cell differentiation and lineage based on single cell entropy, Nucleic Acids Res. 45 (7) (2017) e54.
[10] C. Trapnell, D. Cacchiarelli, J. Grimsby, et al., The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells, Nat. Biotechnol. 32 (4) (2014) 381–386.
[11] X. Qiu, Q. Mao, Y. Tang, et al., Reversed graph embedding resolves complex single–cell trajectories, Nat. Methods 14 (10) (2017) 979–982.
[12] J. Cao, M. Spielmann, X. Qiu, et al., The single-cell transcriptional landscape of mammalian organogenesis, Nature 566 (7745) (2019) 496–502.
[13] G.La Manno, R. Soldatov, A. Zeisel, et al., RNA velocity of single cells, Nature 560 (7719) (2018) 494–498.
[14] G. Schiebinger, J. Shu, M. Tabaka, et al., Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming, Cell 176 (4) (2019) 928–943.
[15] M. Setty, V. Kiseliovas, J. Levine, et al., Characterization of cell fate probabilities in single-cell data with Palantir, Nat. Biotechnol. 37 (4) (2019) 451–460.
[16] S.V. Stassen, G.G.K. Yip, K.K.Y. Wong, et al., Generalized and scalable trajectory inference in single-cell omics data with VIA, Nat. Commun. 12 (1) (2021) 5528.
[17] W. Saelens, R. Cannoodt, H. Todorov, et al., A comparison of single-cell trajectory inference methods, Nat. Biotechnol. 37 (5) (2019) 547–554.
[18] C. Lin, Z. Bar-Joseph, Continuous-state HMMs for modeling time-series single-cell RNA-Seq data, Bioinformatics 35 (22) (2019) 4707–4715.
[19] T.N. Tran, G.D. Bader, Tempora: Cell trajectory inference using time-series single–cell RNA sequencing data, PLoS Comput. Biol. 16 (9) (2020) e1008205.
[20] C. Zhao, W. Xiu, Y. Hua, et al., CStreet: A computed Cell State trajectory inference method for time-series single-cell RNA sequencing data, Bioinformatics 37 (21) (2021) 3774–3780.
[21] Q. Jiang, S. Zhang, L. Wan, Dynamic inference of cell developmental complex energy landscape from time series single-cell transcriptomic data, PLoS Comput. Biol. 18 (1) (2022) e1009821.
[22] L. Zappia, B. Phipson, A. Oshlack, Splatter: Simulation of single-cell RNA sequencing data, Genome Biol. 18 (1) (2017) 174.
[23] C. Wang, X. Liu, Y. Gao, et al., Reprogramming of H3K9me3-dependent heterochromatin during mammalian embryo development, Nat. Cell Biol. 20 (5) (2018) 620–631.
[24] L. Yang, W.H. Wang, W.L. Qiu, et al., A single-cell transcriptomic analysis reveals precise pathways and regulatory mechanisms underlying hepatoblast differentiation, Hepatology 66 (5) (2017) 1387–1401.
[25] S.A. Yuzwa, M.J. Borrett, B.T. Innes, et al., Developmental emergence of adult neural stem cells as revealed by single-cell transcriptional profiling, Cell Rep. 21 (13) (2017) 3970–3986.
[26] R.R. Laxton, The measure of diversity, J. Theor. Biol. 70 (1) (1978) 51–67.
[27] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.
[28] J. Li, X. He, S. Gao, et al., The Metal-binding Protein Atlas (MbPA): An integrated database for curating metalloproteins in all aspects, J. Mol. Biol. 435 (14) (2023) 168117.
[29] M. Lu, S. Liu, A.K. Sangaiah, et al., Nucleosome positioning with fractal entropy increment of diversity in telemedicine, IEEE Access 6 (2018) 33451–33459.
[30] L. Zhang, L. Luo, Splice site prediction with quadratic discriminant analysis using diversity measure, Nucleic Acids Res. 31 (21) (2003) 6214–6220.
[31] C.Y. Wu, Q.Z. Li, Z.X. Feng, Non-coding RNA identification based on topology secondary structure and reading frame in organelle genome level, Genomics 107 (1) (2016) 9–15.
[32] Y.C. Zuo, W. Chen, G.L. Fan, et al., A similarity distance of diversity measure for discriminating mesophilic and thermophilic proteins, Amino Acids 44 (2) (2013) 573–580.
[33] X. Gao, B. Xiao, D. Tao, et al., A survey of graph edit distance, Pattern Anal. Appl. 13 (1) (2010) 113–129.
[34] A. Hagberg, P. Swart, D.S. Chult, Exploring Network Structure, Dynamics, and Function using NetworkX, Los Alamos National Lab.(LANL), Los Alamos, NMUnited States, 2008.
[35] H. Wang, Z. Zhang, H. Li, et al., A cost-effective machine learning-based method for preeclampsia risk assessment and driver genes discovery, Cell Biosci. 13 (1) (2023) 41.
[36] P. Liang, L. Zheng, C. Long, et al., HelPredictor models single-cell transcriptome to predict human embryo lineage allocation, Brief. Bioinform. 22 (6) (2021) bbab196.
[37] H. Wang, P. Liang, L. Zheng, et al., eHSCPr discriminating the cell identity involved in endothelial to hematopoietic transition, Bioinformatics 37 (15) (2021) 2157–2164.
[38] Z. Ji, H. Ji, TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis, Nucleic Acids Res. 44 (13) (2016) e117.
[39] K. Street, D. Risso, R.B. Fletcher, et al., Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics, BMC Genomics 19 (1) (2018) 477.
[40] F.A. Wolf, F.K. Hamey, M. Plass, et al., PAGA: Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells, Genome Biol. 20 (1) (2019) 59.

Yan Hong is a doctoral student of the State key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock, Institutes of Biomedical Sciences, College of Life Sciences, Inner Mongolia University, Hohhot, China. Her research interests include bioinformatics and computational biology.

Hanshuang Li is a doctoral student of the State key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock, Institutes of Biomedical Sciences, College of Life Sciences, Inner Mongolia University, Hohhot, China. Her research interests include bioinformatics, systems biology and developmental biology.

Yongchun Zuo (BRID: 03153.00.08361) is a PhD principal investigator of Bioinformatics. He is the professor of the State Key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock, Institutes of Biomedical Sciences, College of Life Sciences, Inner Mongolia University. Professor Zuo focuses on the computational biology researches, including classification of DNA/Protein sequence, codon optimization of gene and its regulatory sequence, and integration analysis of multi-omics in cell reprogramming.