A Novel Method for Gene Regulatory Network Inference with Pseudotime Data Using Information Criterion

Shuhei Yao, Kaito Uemura, Shigeto Seno, Hideo Matsuda*
Graduate School of Information Science and Technology, Osaka University, Suita, Osaka, Japan.

* Corresponding author. Tel.: +81-6-6879-4390; email: matsuda@ist.osaka-u.ac.jp
Manuscript submitted November 30, 2021; accepted February 11, 2022.
doi: 10.17706/ijbbb.2022.12.3.43-52

Abstract: Trajectory inference has been used to model cellular dynamic processes by using single-cell RNA sequence data. The inference often computes pseudotime representing the progression through the process along the trajectory. Several methods to infer gene regulatory networks have been proposed using the gene expression profiles of the cells ordered with the pseudotime to elucidate the regulatory relationships between genes in a dynamic process. In this paper, we propose a novel method for the inference of such gene regulatory networks. To predict highly accurate gene regulatory relationships in the network, we introduce an edge-scoring scheme with bootstrap sampling. We demonstrate the accuracy of the proposed methods by comparing the results with those of existing methods using synthetic and real single-cell RNA-seq data.

Key words: Gene regulatory network, information criterion, pseudotime analysis, single cell RNA transcriptome.

1. Introduction

Because single-cell RNA sequencing can measure gene expression in individual cells, it has been used to elucidate the mechanisms of cellular dynamic processes such as cell differentiation, cell cycle, and stimulus response[1], [2]. In some cases, such dynamic processes can be modeled as cellular trajectories when cells are classified based on their gene expression profiles. Trajectory inference is often used as a method to represent this dynamic process. In this method, the progression of a dynamic process can be represented as a continuous path, on which cells are placed. The pseudotime is defined as the degree of progression of the process. In this paper, we focus on a cell differentiation process.

Several methods have been proposed for modeling cellular dynamic processes as trajectory inference methods such as Monocle [3], Slingshot [4], and Scanpy [5]. These methods often express the cell states in a process as clusters of cells and explore the trajectory that has the highest score from possible alternatives. Among those methods, Scanpy computes the confidence of each edge between clusters in the trajectory as connectivity. The connectivity score is useful to evaluate the inferred trajectory. However, Scanpy does not compute the pseudotime of cells along the lineages in the trajectory but only calculates it at the edges between clusters. On the other hand, Slingshot computes the pseudotime of all the cells along each lineage in the trajectory. Since we need to obtain time-series gene-expression profiles for inferring gene regulatory networks along each lineage, we propose a method that combines Scanpy and Slingshot.

Pratapa et al. present a systematic evaluation of 12 existing methods for inferring gene regulatory networks from well-defined benchmark datasets [6]. Among the methods, SINCERITIES shows the best AUPRC (area under the precision-recall curve) but the accuracy scores significantly vary depending on the topologies of

the gene regulatory networks and also depending on the numbers of the cells [6]. To improve the network inference accuracy, we introduce a bootstrap-sampling method to obtain a confidence score for each edge of the network by computing the concordance rate among the sampled replicates of gene expression data.

Using the gene expression data of the cells ordered by their pseudotime, we set the time-point of each cell by dividing the pseudotime of the cell by a pre-defined constant and converting it into an integer. The expression data having the same time points are regarded as replicates. Then we perform bootstrap sampling by extracting one of the replicates from each time-point, infer individual gene regulatory networks from the sampled datasets, and compute the confidence score of each regulatory relationship (i.e., each edge of the network) from the inferred networks. For the confidence score, we propose a scoring scheme, called *edge gain*, which was originally implemented in SiGN-BN [7] but has been used first time for the edge score in the network inference from single-cell RNA-seq data. By using the score, we can assess the reliability of each edge of the resulting network. The accuracy of the method is evaluated by comparing the results with an existing method, SINCERITIES, using synthetic data and real cell-differentiation data.

2. Method

2.1. Outline of the Proposed Method

This method is divided into three main steps, which are outlined in Fig. 1. As input data, we use the expression data of cells sorted by pseudotime. In Step 1, bootstrap sampling is performed N times on this data to obtain N datasets. Step 2 is to infer the network, using the dynamic Bayesian network (DBN) model with nonparametric analysis by SiGN-BN and the greedy hill-climbing method as a search algorithm. In this step, the Bayesian network and nonparametric regression criterion (BNRC) are calculated for each network [8]. The better gene-regulatory network that well reflects given gene expression data exhibits a smaller BNRC value.

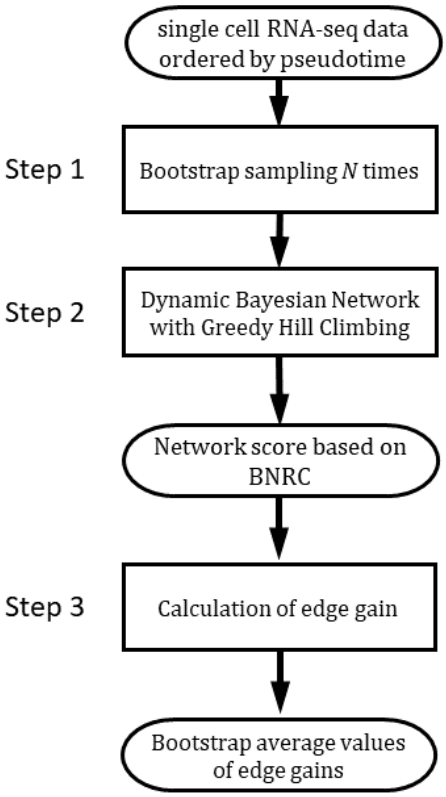


Fig. 1. Outline of the proposed method.

The bootstrap probability that is the concordance rate of each edge among a set of the inferred networks according to the bootstrap-sampled datasets has been used as a confidence score of the edge. However, we introduce another scoring scheme, called edge gain. The edge gain is calculated as follows. Let G and $S(G)$ be a graph representing a gene-regulatory network and the BNRC score of the graph, respectively. We define the contribution of edge e to the network score $S(G)$ as in (1). Since the BNRC score of a network is minimum when the network optimally reflects the given gene-expression profile, the gain of an edge becomes maximum if the edge maximally contributes to the network score. Our method computes the average value in the bootstrap samples as the score for each edge in Step 3.

$$\text{edge gain}(e) = S(G - \{e\}) - S(G). \quad (1)$$

2.2. Calculation of Pseudotime

The data of hematopoietic stem cells used in the Scanpy tutorial was imported into a Seurat object, and the differentiation lineage of each blood cell was inferred from the graph structure obtained by PAGA (partition-based graph abstraction), starting from pluripotent progenitor cells [5], [9]. A PAGA graph is obtained by connecting each cell cluster with weighted edges to represent the connectivity between the clusters in the kNN graph by choosing a suitable low-dimensional representation, e.g., PCA-based representations with Euclidean distance [5], [9]. Pseudotime was calculated by Slingshot based on the obtained differentiation lineage [4]. Expression plots of neutrophil/monocyte markers were drawn by the plotSmoothers function in the tradeSeq package [10].

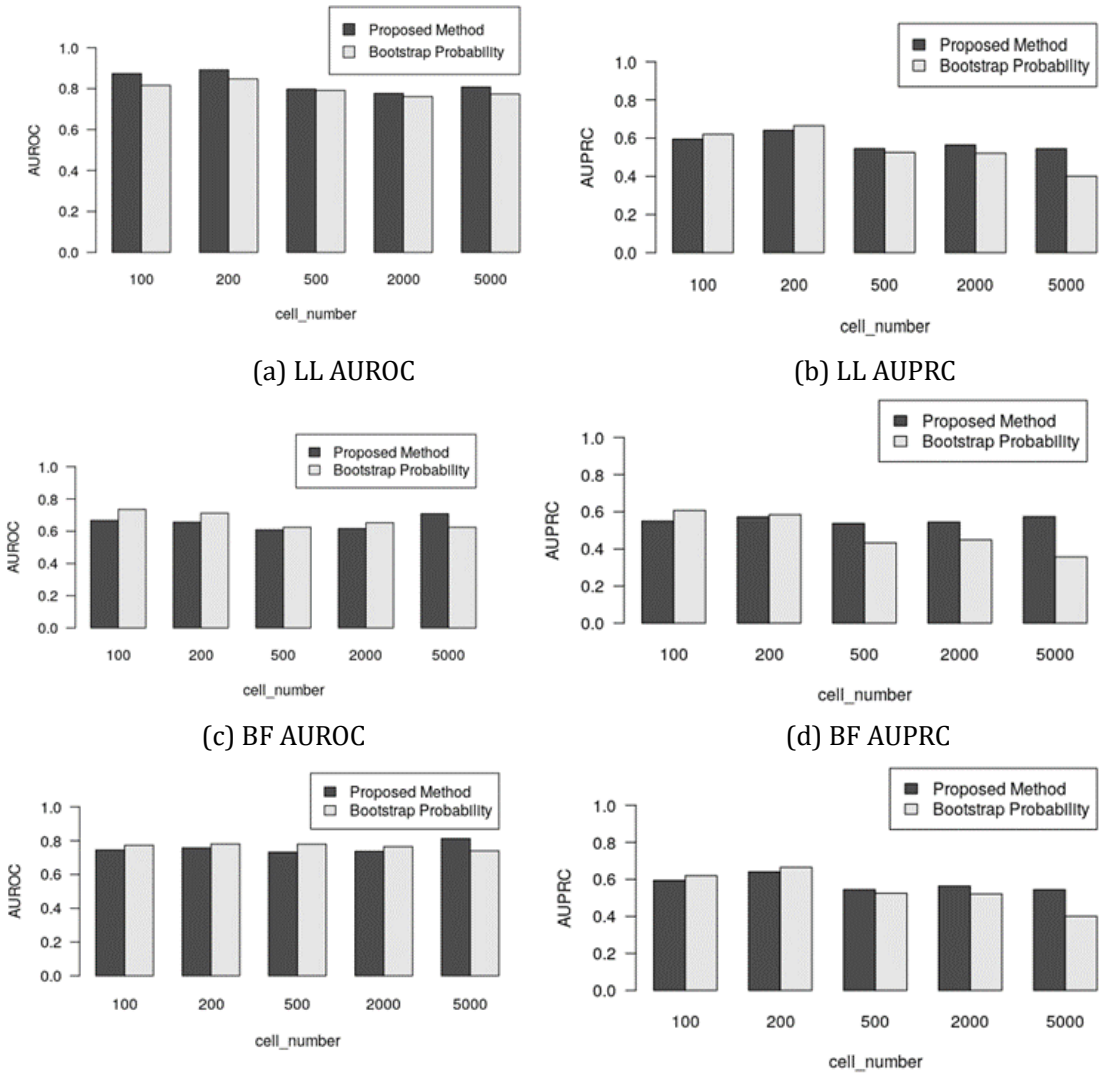
2.3. Network Inference

The synthetic data was generated using BoolODE for three types of network topologies: bifurcating (BF), which is the process by which a cell diverges and changes from one initial state to its final two states through the mutual repression of two genes; bifurcating converging (BFC), means that a state in which one initial state branches into two and then converges to the same state again; and linear long (LL), which is a type in which the regulatory relationships of genes are connected in a long straight line [6]. Each network was generated with five different patterns of cell numbers: 100, 200, 500, 2000, and 5000. In each condition, 10 networks were generated and evaluated as described in [6]. For the actual data, we extracted the expression data of 10 genes from the network of transcription factors experimentally confirmed in the report of [11] from the expression data of cells in the differentiation lineage from pluripotent progenitor cells to monocytes as described above. 315 cells in which at least 3 of the 10 genes were expressed. The same time point was defined as the number truncated after the decimal point of the pseudotime. Using these data as input, network inference was performed using the DBN approach of SiGN-BN, and edge gain was calculated using the approach described above. We performed 100 bootstrap sampling for the synthetic data and 1000 bootstrap sampling for the real data. For each edge of the obtained network, the bootstrap probability, which means the probability that a regulated edge appears in the network inferred from multiple data sets generated by the bootstrap method, was calculated separately from the proposed method. The results obtained by the bootstrap probabilities and SINCERITIES were compared with the proposed method. The AUROC (area under the receiver operating characteristic curve) and AUPRC for each method were computed with the PRROC package [12], and all graphs were plotted in R. For the network visualization, Cytoscape was used.

3. Result

3.1. Accuracy Evaluation Using the Synthetic Networks

The accuracy evaluation of the proposed method, based on the edge gain, compared with the same network inference method based on the bootstrap probability as the edge score using the synthetic network datasets is shown in Fig. 2. Fig. 2 shows their accuracy scores according to the evaluation in [6].



(a) LL AUROC (b) LL AUPRC

(c) BF AUROC (d) BF AUPRC

(e) BFC AUROC (f) BFC AUPRC

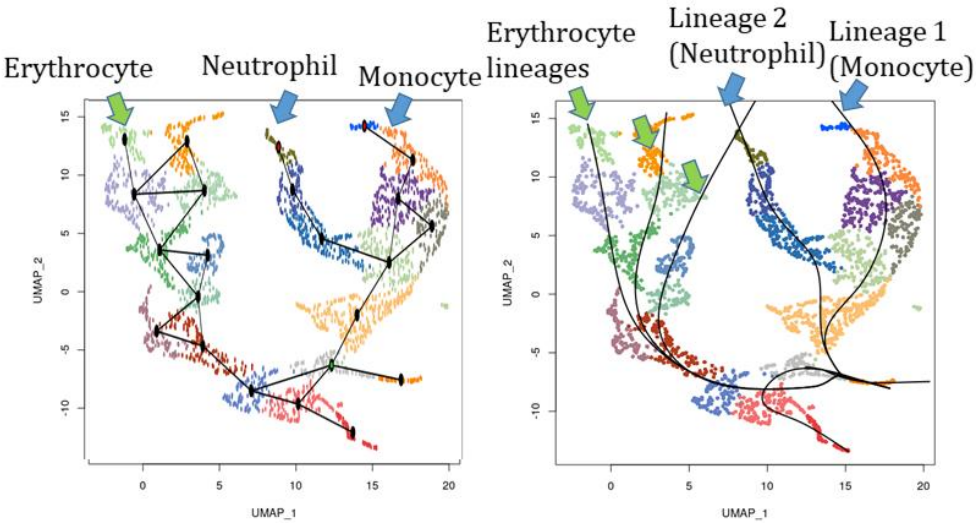
Fig. 2. Performance of the proposed method and bootstrap-probability scoring method. LL, BF, BFC, AUROC, and AUPRC denote linear long, bifurcating, bifurcating converging, the area under the receiver operating characteristic curve, and the area under the precision-recall curve, respectively.

In Fig. 2, the AUROC of bootstrap probability ranged from 0.7 to 0.9 for all network topologies, indicating that they do not change according to the number of cells. By contrast, the AUPRC of the bootstrap probability tends to decrease when the number of cells exceeds 500. On the other hand, the results of our edge-gain-based method (the proposed method) show that they do not change according to the number of cells in BF and BFC, while they tend to decrease when the number of cells exceeded 500 in LL. Overall, the proposed method mostly achieves better accuracy than the bootstrap probability results where the number of cells increases, typically 5000 and more.

Compared with the results of SINCERITIES described in the benchmarking paper [6], although SINCERITIES exhibits better accuracy in a part of the cases (the proposed method ranges from 0.6 to 0.7 and SINCERITIES is around 0.8 in BF AUROC), the proposed method mostly achieves better or similar accuracy than SINCERITIES, especially the proposed method ranges from 0.6 to 0.7 in AUPRC regardless of the network topologies and the number of cells as shown in Fig. 2 whereas the AUPRC of SINCERITIES significantly varies ranging from 0.1 to 0.7 in LL and from 0.2 to 0.6 in BF and BFC depending on the number of the cells.

3.2. Gene Regulatory Network Inference Using Real Dataset

To infer a network from a real single-cell RNA-seq dataset using the proposed method, pseudotime analysis was conducted against hematopoietic stem cells data [13]. In the paper [13], to characterize the state of myeloid progenitor cells, the mRNA of bone marrow cells were sequenced and separated by flowcytometry into the common myeloid progenitors, megakaryocyte/erythrocyte progenitors (MEP), and granulocyte/macrophage progenitors (GMP). GMP and MEP differentiate into neutrophils/monocytes and erythrocytes, respectively. The gene expression data of the myeloid progenitors were obtained according to the Scanpy tutorial. The data were projected into UMAP space and conducted graph abstraction by PAGA as described in Section 2.2. The result is shown in Fig. 3(a). Colors mean clusters classified by Louvain in the Scanpy tutorial. Due to the character of PAGA as a graph partitioning method, some unnecessary paths were shown here. As a result of narrowing down the differentiation lineage to a single path using PAGA connectivity as an indicator, we found two major types of paths: one is from the cluster of multipotent progenitors to erythrocytes, and the other is to monocytes and neutrophils as shown in Fig. 3(a).



(a) Graph paths in PAGA (b) Lineages by Slingshot

Fig. 3. Differentiation lineage of hematopoietic stem cells using PAGA and slingshot. These plots show dimensional reduction by UMAP. Each color means cluster by Louvain in Scanpy.

Furthermore, we were able to obtain a clear divergence in the paths to monocytes and neutrophils. This path can be represented as a cluster, and by specifying this path, we performed a pseudotime series analysis using Slingshot (Fig. 3(b)). We were able to determine the position of the cells in the lineage.

To demonstrate the validity of this lineage as a monocyte or neutrophil lineage, we examined the expression of markers (Fig. 4). Plot color means each lineage. Yellow is monocyte lineage and dark blue is neutrophil lineage. Other colors indicate other lineages. The X-axis is pseudotime and Y-axis is the logarithm of the expression value. The monocyte markers Irf8 (Fig. 4(a)) and Csf1r (Fig. 4(b)) were upregulated as the pseudotime progressed in the monocyte lineage. On the other hand, Cebpe (Fig. 4(c)) and Gfi1 (Fig. 4(d)),

which are important transcription factors in the neutrophil lineage, were also upregulated in the neutrophil lineage detected in this study. The pseudotime data obtained using real biological samples for the proposed method showed biological validity.

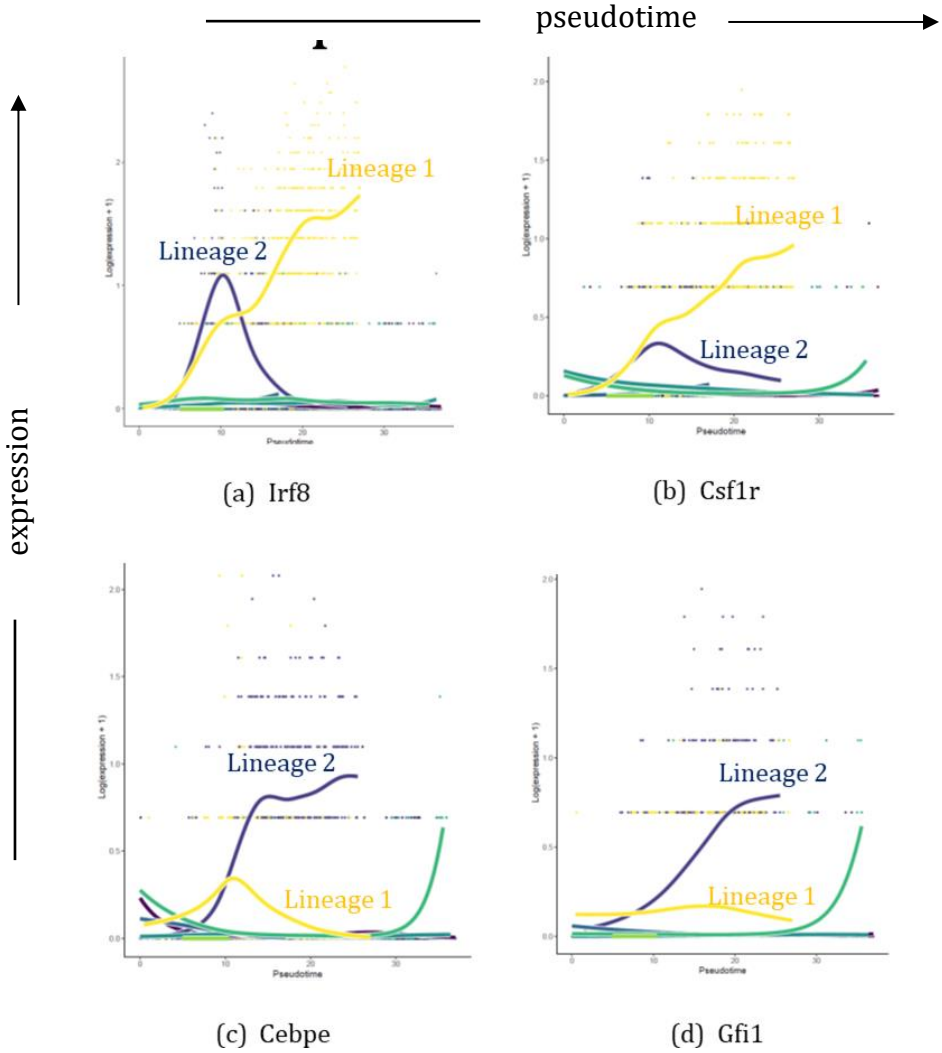
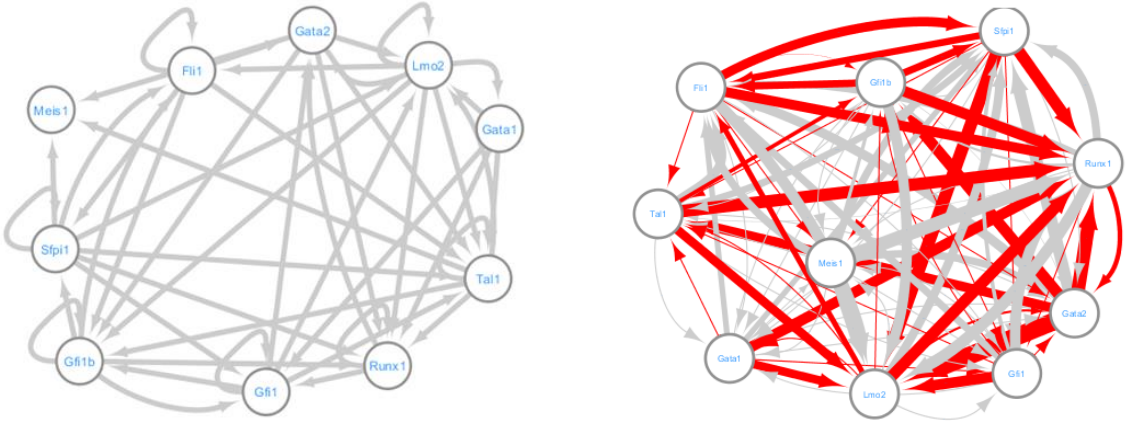


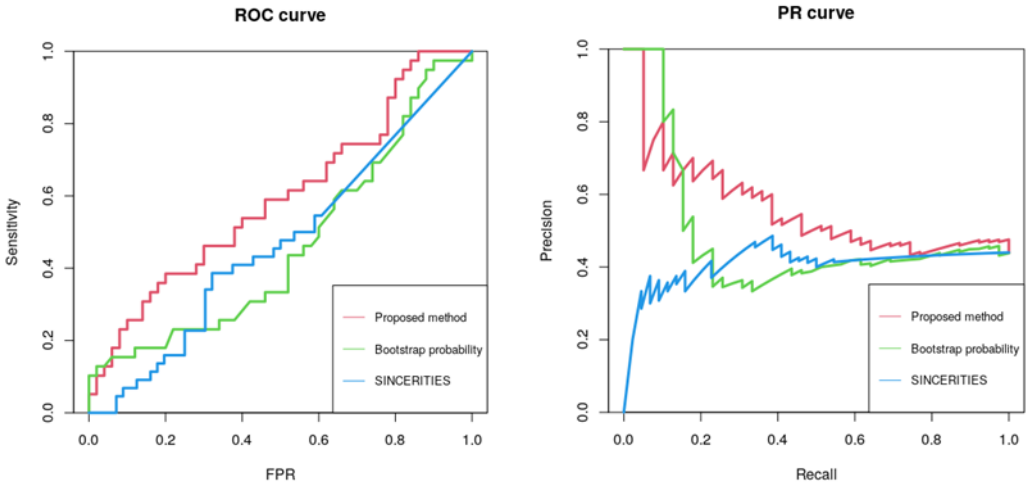
Fig. 4. Expression of monocyte and neutrophil markers along with their lineages. These plots were shown by plotsmoother function in tradeSeq package. X-axis and Y-axis are pseudotime and expression value, respectively. The color of lines indicates the expression of the markers in each lineage (yellow, monocyte; navy, neutrophil and green, erythrocyte).

We evaluated the accuracy of the proposed method by using the pseudotime series of cell differentiation from pluripotent progenitor cells to monocytes as time-series data for gene regulatory network analysis. Reference/Inferred network visualized from Cytoscape is shown in Fig. 5. In Fig.5(b), red edges indicate the edges existing in the reference network and their edge widths correspond to the edge gain scores. According to the ROC and precision-recall curves (Fig. 6(a) and (b)), the proposed method was highly accurate than bootstrap probability or SINCERITIES. The AUC of the ROC curve was 0.593, and the AUC of the PR curve was 0.563 as shown in Table 1.



(a) Reference network referred to [10]. (b) Inferred network by the proposed method.

Fig. 5. Reference and Inferred networks of cell differentiation in hematopoietic stem cells. The color of edges denotes true positive (red) and false positive (grey). Edge width indicates edge gain scores.



(a) ROC curve (b) Precision-recall curve

Fig. 6. Accuracy of the proposed method, SINCERITIES, and bootstrap probability.

Table 1. Accuracy for Inferring Networks from a Real Dataset

	Proposed Method	Bootstrap Probability	SINCERITIES
AUROC	0.593	0.463	0.469
AUPR	0.563	0.488	0.405

4. Discussion

The systematic evaluation of 12 existing methods in the benchmarking paper [6] indicates that the accuracy rates of the methods often widely vary from low to high depending on the network topologies, e.g., linear long (LL), bifurcating (BF), and bifurcating converging (BFC), and also the number of cells in their synthetic datasets. Actually, in SINCERITIES, a highly accurate existing method, the AUPRC rates significantly vary from 0.1 to 0.7 in LL and from 0.2 to 0.6 in BF and BFC depending on the number of the cells. By contrast, the accurate evaluation of the proposed method showed its AUPRC rates ranged only between 0.6 and 0.7 regardless of the network topologies and the number of cells. Also in the real dataset for cell differentiation, the proposed method achieved higher accurate results than SINCERITIES both in AUROC and AUPRC.

However, as shown in Fig. 5(b), the result of the proposed method cannot detect several edges in the reference network. Also in Fig. 6(b), the precision rate of the proposed method rapidly decreases according to the increase of the recall rate. This suggests that the inferred results of the proposed method may include several false positives, i.e., spurious or indirect regulatory relationships. Thus although the accuracy of the inference in the proposed method is better than that of the current existing method, the proposed method is still required to improve the accuracy of the inference.

5. Conclusion

In this study, we have proposed a novel method for inferring gene-regulatory networks from single-cell RNA-seq data. The method infers the networks by using time-series gene expression profiles from pseudotime-ordered cells generated by a trajectory-inference method. In the method, we have introduced a new scoring scheme called edge gain, which is the gain to the inference score if an edge is added to the resulting network. We compute the average of the gains from the bootstrap-sampled replicates in the gene expression data, and we use the average gain as a confidence score for each edge in the inferred network.

By the performance evaluation using both synthetic and real single-cell datasets, the proposed method achieved higher accurate performance than an existing method, SINCERITIES, which is reported as one of the best-performing methods in a benchmarking paper [6]. The result suggests that our scoring scheme is useful to evaluate the confidence of each edge in the inferred gene-regulatory network. However, the evaluation results of the proposed method indicate that the accuracy of the inference is still required to be improved in the proposed method. It remains our future works.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

SY and HM conducted the research, analyzed the data, and wrote the paper, KM and SS analyzed the data; all authors approved the final version.

Acknowledgment

The authors thank Prof. Yoshinori Tamada for his valuable comments on the scoring scheme of SiGN-BN. This work was supported in part by JSPS KAKENHI Grant Numbers JP18H04124, JP19K22894, JP20H04947, and JP21K19827 Japan.

References

[1] Tatsuoka, H., Sakaniti, S., Yabe, D., Kabai, R., Kato, U., Okumura, T., *et al.* (2020). Single-cell transcriptome analysis dissects the replicating process of pancreatic BETA cells in partial pancreatectomy model. *iScience*, *23(12)*, 101774.

[2] Luginbühl, J., Kouno, T., Nakano, R., Chater, T. E., Sivaraman, D. M., Kishima, M., *et al.* (2021). Decoding neuronal diversification by multiplexed Single-cell RNA-Seq. *Stem Cell Reports, 16(4)*, 810-824.

[3] Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., *et al.*, (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nature. Methods, 14(10),* 979-982.

[4] Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., *et al.* (2018). Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics, 19(1)*, 477.

[5] Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Götttgens, B. (2019). PAGA: Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome*

Biology, *20(59)*, 1-9.

[6] Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, L. A., & Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, *17*, 147-154.

[7] Tamada, Y., Shimamura, T., Yamaguchi, R., Imoto, S., Nagasaki, M., Miyano, S. (2011). Sign: Large-scale gene network estimation environment for high performance computing. *Genome Informatics, 25(1)*, 40-52.

[8] Imoto, S., Goto, T., & Miyano, S. (2002) Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression. *Pac. Symp. Biocomput.*, 7, 175-186.

[9] Wolf, F. A., Angerer, & P., Theis, F. (2018). SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, *19(15)*, 1-5.

[10] Berge, K. V., Be zieux, H. R., Street, K., Saelens, W., Cannoodt, R., Saeys, Y., *et al.* (2020). Trajectory-based differential expression analysis for single-cell sequencing data. *Nature Communications, 11(1)*, 1201.

[11] Goode, D. K., Obier, N., Vijayabaskar, M. S., Lie-A-Ling, M., Lilly, A. J., Hannah, R., *et al.* (2016). Dynamic gene regulatory networks drive hematopoietic specification and differentiation. *Development Cell*, *36(5),* 572-587.

[12] Grau, J., Grosse, I., & Keilwagen, J. (2015). PRROC: Computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics, 31(15)*, 2595-2597.

[13] Paul, F., Arkin, Y., Giladi, A., Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., *et al.* (2015). Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell, 163(7)*, 1663-1677.

Shuhei Yao received his B.Sc. degree in science from Kwansei Gakuin University in 2001, and he received his M.Sc. degree in science from Osaka University in 2003. His research interests include omics data analysis, gene regulatory networks, and genomic data analysis.

Kaito Uemura received his B.E. and M.E. degrees from Osaka University in 2017 and 2019, respectively. He studied gene regulatory network inference from single-cell RNA-seq data at the Department of Bioinformatic Engineering, the Graduate School of Information Science and Technology, Osaka University. His research interests include bioinformatics (gene expression analysis).

Shigeto Seno received his B.E., M.E., and Ph.D. (Information Science) degrees from Osaka University in 2001, 2003, and 2006, respectively. He has been an associate professor in the Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, since 2017. His research interests include data mining, bioinformatics (gene expression analysis), and bioimage informatics. He is a member of IPSJ, JSBi, and MII.

Hideo Matsuda received his B.Sc. degree in physics from Kobe University in 1982, and he received his M.E. and Ph.D. degrees in computer science from Kobe University in 1984 and 1987, respectively. He has been a professor at the Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, since 2002. His research interests include genomic data analysis, gene regulatory networks, and gene expression analysis. He is a member of IPSJ, JSBi, and ISCB.