Reconstruction of Dynamic Gene Regulatory Networks for Cell Differentiation by Separation of Time-course Data

T. Nakayama[1], **H. Daiyasu**[1], **S. Seno**[1], **Y. Takenaka**[1], **and H. Matsuda**[1]
[1]Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, 1-5, Yamadaoka, Suita, Osaka, Japan

Abstract— *Recently, dynamic Bayesian network (DBN) model is widely used for estimating gene regulatory networks (GRNs) from time-course gene expression data. Ordinary DBNs estimate only a single network using the whole time-course data. However, some GRNs, such as cell differentiation, dynamically change their network structures due to chromatin remodeling. In this papers we present a method to estimate such dynamic GRNs that follow the dynamic changes of the regulations in adipocyte differentiation by separating time-course data. We analyzed the estimated GRNs and confirmed that the GRNs showed the dynamic changes in adipocyte regulation. The result shows that our method can identify the regulatory relationships of the genes that are dynamically changing during adipocyte differentiation by separating the time-course data.*

Keywords: cell differentiation, adipocyte, dynamic Bayesian network model, time-course data separation

1. Introduction

Reconstruction of gene regulatory networks (GRNs) from gene expression data is a fundamental but challenging task in bioinformatics area. A number of methods have been developed for reconstructing GRNs. Among the methods, dynamic Bayesian network (DBN) model is widely used for estimating GRNs from time-course gene expression data [1]. However, ordinary dynamic Bayesian networks estimate a single network using whole time-course data, while some GRNs (e.g., GRNs in cell differentiation) dynamically change their network structures at their observed time points [2]. In this paper, we present a method to estimate the dynamic GRNs by separating time-course data.

Adipocyte differentiation is the one of the processes that is controlled by a complex network of transcription factors acting at different stages of differentiation due to chromatin remodeling [3]. It have been suggested that the four important adipogenic genes act at different stages [3][4]. During the early stages of adipogenesis, C/EBPβ and C/EBPδ activate expression of PPARγ, C/EBPα and probably other adipogenic genes. And then, PPARγ and C/EBPα activate expression of adipocyte specific genes. Furthermore recent studies have been revealing a complex transcriptional cascade controlling adipocyte differentiation [5][6][7][8].

The node-set separation method (NSS) [2] tries to capture different sub-networks that have high activity at their observed time points. This method estimates a GRN from whole timecourse data by using DBN, and then represents the dynamics of the GRN as transition of the regulations among the genes that are in active gene sets. An active gene set is determined as a set of differentially expressed genes comparing with the controls for each time point. Regulations among the genes in the active gene sets from consecutive two time points show the activity of the GRN at the time. In whole time-course data, the activities are changed at each time point. The transitions of activities of the GRN are regarded as the dynamics of the GRN.

There is matter that the method like the NSS uses the whole time-course gene expression data to estimate GRNs. It is suggested that the estimations with whole time points cannot identify the regulations that only exist in short span. Such short-term dynamic transcription controls are caused by chromatin remodeling [3]. Recently, experiments of microarray and updated methods, like RNA-Seq, that enable us easily to acquire high resolution time-course data. However, ordinary DBN-based methods evaluate the overall change of the gene expressions rather than the expressions represent the regulation change during short-term time intervals. In this paper, we estimate dynamic GRNs by DBN from separating the timecourse data of adipocyte differentiation, and present our proposed method can estimate some experimentally-confirmed regulations that are not detected by the NSS.

2. Materials and Method

Our method needs a time-course data with more time points than the NSS to estimate the dynamic GRN. It means that the data need to have the many time points enough to estimate a GRN if we separate the data. In addition, the estimation costs a computational time because the data need to have the many genes enough to estimate the relationships among genes that are concern of adipocyte differentiation. In this study, we used parallelized software on massively parallel systems for estimating the dynamic GRNs of adipocyte differentiation.

2.1 Microarray Data of Adipocyte Differentiation

We collected RNAs from Mouse ST2 Bone marrow stroma cell-derived stem cell (RCB0224) from RIKEN BioResource Center (BRC, Tsukuba, Japan) for adipocyte cell differentiation. The ST2 cell was induced by changing the medium from RMPI1640 to DMEM supplemented with 10% FBS, 0.5 mM 3-isobutyl-1-methlxanthine (MIX), 0.25μM DEX, and insulin-transferrin-selenium-X supplement containing 5μg/ml of insulin and 1μM rosiglitazone. After 48 hours, the differentiation medium was replaced with DMEM supplemented with 10% FBS.

The collected RNAs were analyzed with Affymetrix GeneChip Mouse Genome 430 2.0 Array, which generated transcript expression profiles at the time points: 5, 15, 30 and 45 minutes, 1 to 30 hours for every hour, 36 to 192 hours for every 6 hours after adipogenesis induction. Each time-course data was background-subtracted and normalized with the robust multi-array analysis (RMA) [9] using affy package from the Bioconductor version1.8.1. The transcript expression profiles are available from Genome Network Platform (http://genomenetwork.nig.ac.jp). We also calculated expression rate in each gene by Z-score. The data are converted to a common scale with an average of zero and standard deviation of one. This normalization was to emphasize the changing behavior of the gene expressions of the data rather than the value of the gene expressions.

In adipocyte differentiation, it is well-known that some significant transcription genes act a key regulator of adipocyte development [5] (see Fig. 1). These genes expressed enough value in our observed data. The network represents 23 regulations among 14 adipogenic genes.

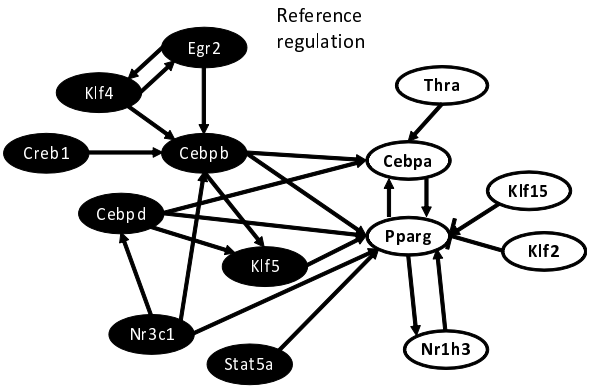

Fig. 1: Reference gene regulatory network [5]. Black and white circles represent the genes that are regulated at early stage and at late stage, respectively. Arrow edge represents upregulation and T-shaped edge, which exist on the relationship among Klf2-Pparg, represents down-regulation.

2.2 Separating the Time-course Data

We separate the time-course data to describe the changes of the gene regulations. If we estimate the network using whole time-course data of adipocyte differentiation, the result of the estimation describes the relationships between genes that regulate the other genes at any time throughout the whole differentiation. Other studies suggest that the gene regulatory relationships in cell differentiations are changing dynamically [3][6][10]. We generate subsequences from the gene expression data to make sure of the changes and estimated networks from each subsequence.

The node set separation method [2] is one of the methods to make subsequences. This method defines an active gene set for each time point and estimates GRN with each continuous couple of the time points at the active genes. In the method concept, the sub-networks that are constructed from the active genes have high activity and transmit information of external signals to other sub-networks.

We separate data by time-course, inspired by the NSS algorithm. In contrast to the NSS, the subsequences have some continuous time-courses at least 10 time points and all genes of input data (see Fig. 2). The NSS uses only active genes at consecutive two time points, while this method takes many time points to clear the causal relationships between two genes.

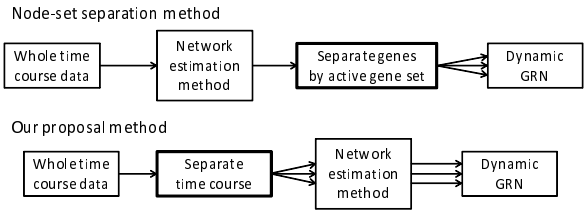

Fig. 2: Summary of the methods

Our method separates input time-course data into equal intervals with overlap. We formalized separated subsequences Z, that is

$$Z_i = (X_{(i-1)S+1}, X_{(i-1)S+2}, X_{(i-1)S+W}) \quad i = 1, 2, \ldots, 1 + (T - W)/S \tag{1}$$

where $X = (X_1, X_2, \ldots, X_T)^t$ is the input time-course data and T is the number of the time points of input data. W is the size of the interval, which is "window size", and S is the value of shifted time points, which is "sliding width".

2.3 GRN Estimation

We estimated GRNs by the DBN model [1][11] using SiGN [11][12], which is the software that implements the DBN and works at high speed in parallel for supercomputer systems. The DBN model is able to construct cyclic regulation and is based on time-course data. In general, the DBN is estimated by an approximate search (greedy hill climbing)

algorithm because the DBN model takes a large amount of computational time as increasing the number of genes.

3. Result

We present our separation method is suitable for high resolution time-course data of adipocyte differentiation than the NSS.

In this study, we separate the above time-course data into 10 subsequences. We set the parameters of (1) to $W = 15$ and $S = 5$. It means that each subsequence has 15 time points and the first time point of the subsequences a five time point time lag between two continuous sub sequences. We estimated the DBNs by SiGN with the 10 subsequences that have 15 time points and all time points for comparison. The network N_t where $t = 1, \ldots, 10$ is estimated from Z_t and N is estimated with all time points X. The NSS is applied to N. We set the threshold of active gene to zero. It means that the gene is assumed active if the expression value of the gene is greater than mean of the gene expression value.

In this work, the computational environments of the estimation are the Human Genome Center (HGC) super-computer system, the University of Tokyo, and K computer (Advanced Institute for Computational Science, RIKEN). We used SiGN to estimate the DBN networks. The parameters we set is below; the number of bootstrap = 10,000, bootstrap replication = 3, bootstrap threshold = 0.05, hyperparameters of the BNRC score function hn=2, hb= 1.0 and hi=2.0. The other parameters were set to their default values. We decided these parameters by repeating small experiments with changing the parameters. This parameter set makes SiGN repeat network estimation 10,000 times to determine one network for bootstrapping, and output a network consisting of the regulations that appear on at least five percent of the 10,000 networks.

Figure 3 shows estimation accuracy of the each 10,000 estimated networks. SiGN uses an informatic criterion named BNRC [11]. The optimal network is chosen such that the BNRC is minimal. BNRC depends on the number of time points. In this study, BNRC of the estimated network was divided by the number of time points of the input data for the bias correction.

The overall network N is shown in Fig. 4. N has 62 edges among 16 genes. The number of estimated networks by NSS method is 60 because active gene set are determined at each time point. We show parts of the networks in Fig. 5. Our proposed method estimated 10 networks. For comparison with the results of NSS, N_2, N_4, N_7, N_8, and N_9, which are the result from 6th time point to 21th time point, 16th to 30th, 31st to 45th, 36th to 50th, and from 41th to 55th, respectively, are shown in Fig. 6. These networks in figures were arranged by force directed algorithm using Cytoscape (http://www.cytoscape.org), which is a visualization and analysis tool for biologic network.

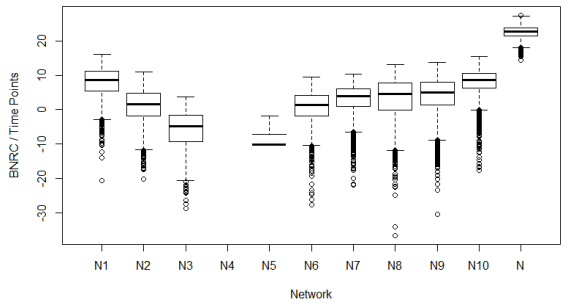

Fig. 3: This box plot shows the network estimation accuracy. The lower BNRC the network has, the higher accuracy the estimation of the network is. $N_1, \ldots, N_{10}$ are estimated by our proposed method, and N is estimated by NSS. N_4 has no box in the box plot because the results of N_4 were too low to draw in this graph.

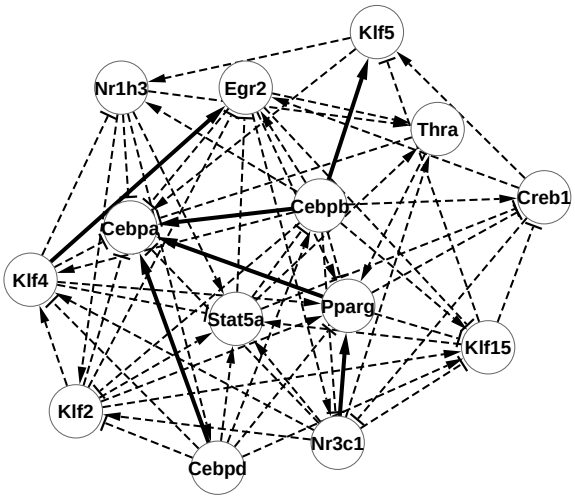

Fig. 4: The result of estimation with whole time-course data. Solid arrows show regulations that match with known regulations shown in Figure 1, and dash arrows show those that do not match with them.

Summary of these networks is shown in Fig. 7. Figure 7 shows distribution of F-measure in estimated networks by NSS and proposed method. F-measure, which is calculated by Eq. (2), is a measure of a estimation accuracy to compare the result with a reference. The best score of F-measure becomes 1, and the worst score of F-measure becomes 0.

$$
\begin{aligned}
F - measure &= 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \\
Precision &= \frac{true\ positive}{true\ positive + false\ positive} \\
Recall &= \frac{true\ positive}{true\ positive + false\ negative}
\end{aligned}
\tag{2}
$$

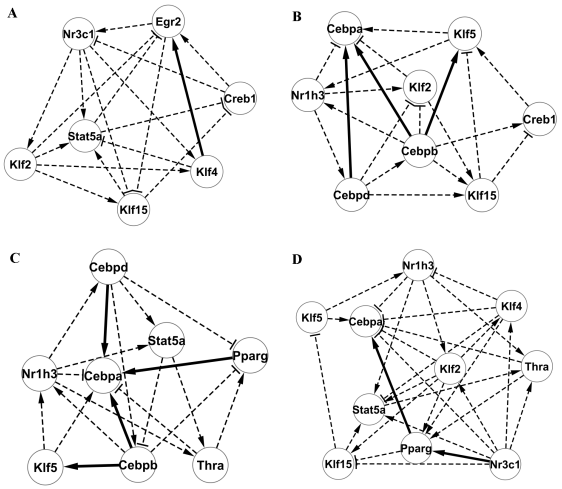

Fig. 5: A part of the results of estimation by NSS. As the same as Fig. 4, Solid and dash arrows show the regulations that match and do not match with known regulations shown in Fig. 1, respectively. Network A extracts an active gene set at the first time point from the network shown in Fig. 1. Similarly, networks B, C, and D extract active gene sets from the 15th and the 16th, from the 30th and the 31th, and from the 45th and the 46th time points, respectively.

Figure 8 is the network represents the result of comparing the reference network shown in Fig. 1 with the estimated networks. The number of matched edges that estimated only by NSS is one, and estimated only by the proposed method is five. Five edges are commonly appeared in both methods.

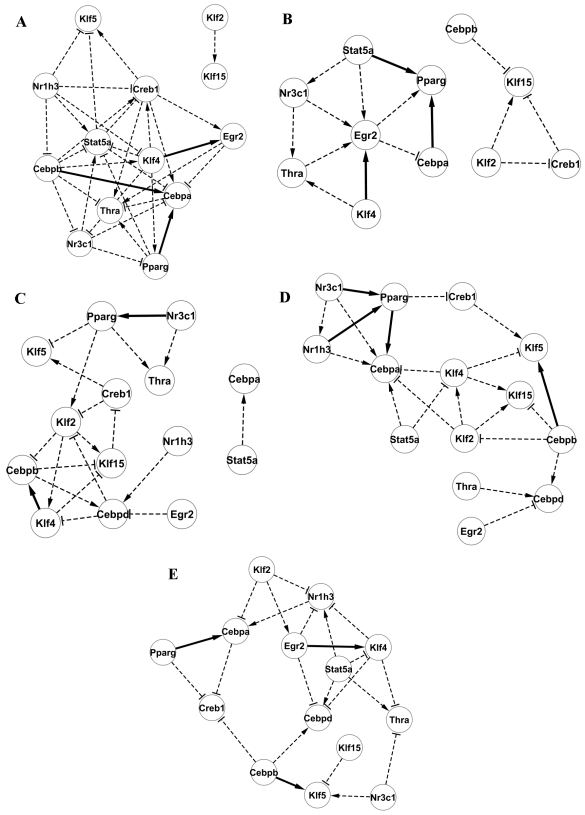

Fig. 6: A part of the results of estimation by our proposed method. Solid and dash arrows and their width mean the same as in Fig. 5. Network A is estimated from the 6th time point to the 21th. Similarly, networks B, C, D and E are estimated from the 16th to 30th, from the 31th to 45th, from the 36th to 50th, and from the 41th to 55th time points, respectively.

4. Discussion

In this paper we proposed a time-separation method for GRN estimation method with high time-resolution data of adipocyte differentiation. Our method has an advantage of tracing dynamic GRN changes over other methods that estimate GRN with whole time-course data. The networks of proposed method capture the gene regulations that are not in entire span of adipocyte differentiations but in short span. This method is applicable to estimate GRN from the mechanisms at what expressions of genes change vary widely for a small amount of time such as adipocyte differentiations.

Figure 3 showed that the BNRC of the all networks estimated by our method is lower than the result of NSS. It means that the estimation accuracy of our method becomes higher than NSS. Furthermore, Figure 8 shows that the proposed method estimated more correct regulations than NSS. Moreover, Figure 7 shows the accuracy of each estimated network is more of the same. It suggests that our method does not decline estimation accuracy in spite of using lesser time points than NSS, and captured the regulations of adipocyte differentiation in short span. Our method focuses on the change of regulation in short span. In contrast, the networks that are estimated by NSS is based on the whole time-course data. Therefore, the regulations are mainly appeared from the entire differentiation behavior. Several studies have reported that various genes regulate other genes like a cascade in short term in adipocyte differentiation. For the reason, the proposed method is more suitable than NSS in this study.

The parameters of the proposed method, which are "window size" and "sliding width", are not optimized. If we could fully optimize the parameters, we would get more favorable performance.

4.1 Author's Contributions

TN developed software and made computational experiments. HM supervised the research. TN, HD, SS, YT and HM wrote the manuscript. All authors read and approved the final manuscript.

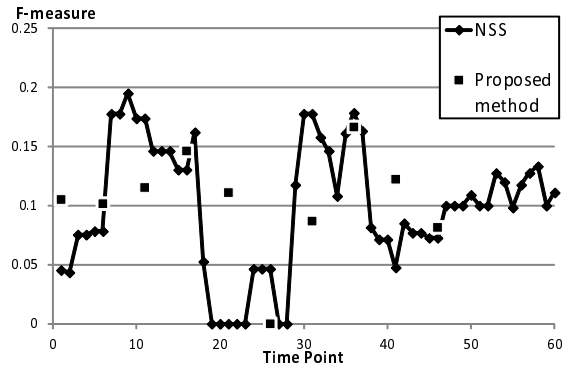

Fig. 7: This graph shows the distribution of F-measure in estimated networks by NSS and proposal method. Each point represents one network estimated respective methods.

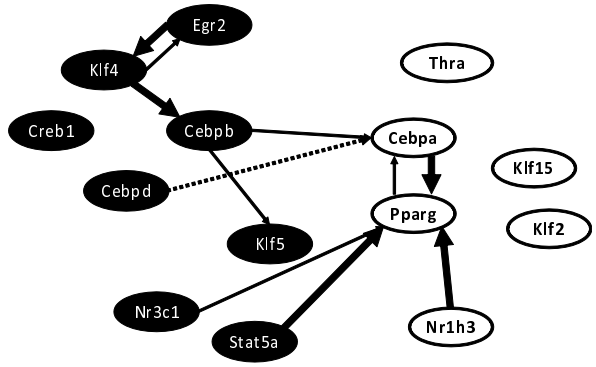

Fig. 8: This network represents the result of comparing the reference network with the estimated networks. Thin edge is the correct edge that is appeared in both networks in common. Thick edge and dot edge mean that the correct edge is appeared in the separated networks and the network estimated by NSS, respectively.

4.2 Acknowledgment

The authors thank to Drs. Yoshinori Tamada and Satoru Miyano for providing the information on the SiGN software. This work was partially supported by Grant-in-Aid for Scientific Research (22310125) from the Japan Society for the Promotion of Science (JSPS), and MEXT SPIRE Supercomputational Life Science.

References

[1] N. Friedman, K. Murphy, and S. Russell, "Learning the structure of dynamic probabilistic networks". in *Proc. UAI'98*, 1998, pp. 139–147.

[2] Y. Tamada, H. Araki, S. Imoto, M. Nagasaki, A. Doi, Y. Nakanishi, Y. Tomiyasu, K. Yasuda, B. Dunmore, D. Sanders, S. Humphreys, C. Print, DS Charnock-Jones, K. Tashiro, S. Kuhara, and S. Miyano, "Unraveling dynamic activities of autocrine pathways that control drugresponse transcriptome networks," *Pac Symp Biocomput.,* pp. 251–263, 2009.

[3] R. Siersbaek, R. Nielsen, S. John, MH. Sung, S. Beak, A. Loft, GL. Hager, and S. Mandrup, "Extensive chromatin remodelling and establishment of transcription factor 'hotspots' during early adipogenesis" *The EMBO Journal,* vol. 30, pp. 1459–1472, 2011.

[4] R. Siersbaek, R. Nielsen, and S. Mandrup, "PPARγ in adipocyte differentiation and metabolism-novel insights from genome-wide studies," *FEBS Letters,* vol. 584, no. 15, pp. 3242–3249, 2010.

[5] R. Siersbaek, R. Nielsen, and S. Mandrup, "Transcriptional networks and chromatin remodeling controlling adipogenesis," *Trends in Endocrinology and Metabolism,* vol. 23, no. 2, pp. 56–64, 2012.

[6] E. D. Rosen, O. A. MacDougald, "Adipocyte differentiation from the inside out," *Nature Reviews Molecular Cell Biology,* vol. 7, no. 12, pp. 885–896, 2006.

[7] M. I. Lefterova, and M. A. Lazar, "New developments in adipogenesis," *Trends in Endocrinology and Metabolism:TEM,* vol. 20, no. 3, pp. 107–114, 2009.

[8] Q. Q. Tang, and M. D. Lane, "Adipogenesis: from stem cell to adipocyte," *Annual Review of Biochemistry,* vol. 81, pp. 715–736, 2012.

[9] RA. Irizarry, B. Hobbs, F. Collin, YD. Beazer-Barclay, KJ. Antonellis, U. Scherf, and TP. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics,* vol. 4, no. 2, pp. 249–264, 2003.

[10] Y. Tokuzawa, K. Yagi, Y. Yamashita, Y. Nakachi, I. Nikaido, H. Bono, Y. Ninomiya, Y. Kanesaki-Yatsuka, M. Akita, H. Motegi, S. Wakana, T. Noda, F. Sablitzky, S. Arai, R. Kurokawa, T. Fukuda, T. Katagiri, C. Schonbash, T. Suda, Y. Mizuno, and Y. Okazaki, "Id4, a new candidate gene for senile osteoporosis, acts as a molecular switch promoting osteoblast differentiation," *PLoS Genetics,* vol. 6, no. 7, doi: e1001019, 2010

[11] S. Kim, S. Imoto, and S. Miyano, "Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data," *Biosystems,* vol. 75, no. 1–3, pp. 57–65, 2004.

[12] Y. Tamada, T. Shimamura, R. Yamaguchi, S. Imoto, M. Nagasaki, and S. Miyano, "Sign: large-scale gene network estimation environment for high performance computing," in *Genome Inform. '11*, 2011, vol. 25, no. 1, pp. 40–52.