Artificial Data Generation Scheme Based on Network Alignment for Evaluation Considering Structural Diversity

***Hitoshi AFUSO*†*, Takeo OKAZAKI*††*, and Morikazu NAKAMURA*††**

†*Graduate School of Information Engineering in University of the Ryukyus, Okinawa, Japan*
††*Faculty of Information Engineering, University of the Ryukyus, Okinawa, Nishihara Senbaru 1, Japan*

Summary

Estimation of transcriptional regulatory networks (TRNs) is the one of most challenging area in post genomic era. While various methods to estimate TRNs, evaluation for such methods, based on generation of artificial TRNs and corresponding artificial gene expression profile data, has been received attentions. However, traditional artificial data generation method does not confirm the structural diversity of generated TRNs. Then, The results of evaluation for estimation methods may be biased. On the other hand, to extract the equivalent subnetwork between two different networks, network alignment methods have been proposed. In this paper, we proposed the artificial data generation scheme for evaluation of network estimation methods so that one can confirm structural diversity in generated TRNs. And also, as a example for application, we compared four score functions for edge orientation problem that one part of network estimation problem, according to proposed data generation scheme.

Key words:

Artificial TRNs generation method, *Confirmation of structural diversity, Network alignment*

1. Introduction

Inside life-form cells, many genes or proteins interact each other, and these complex interactions utilize certain biological functions. Recently years, it has been received attentions to understand these interactions. The technique that can observe the expressions of large amount of genes at a time gives us massive information about the internal connections among genes in life-form cells. As a instance of such technique, we can see DNA microarray[1]. DNA microarray data is also called *gene expression profile data* in the sense of that it denotes the expression levels of corresponding genes. As of now, various studies using DNA microarray data has carried out[2]. In those studies, the estimation of transcriptional regulatory networks(TRNs) is the one of most challenging topic in post-genomic era. TRNs represent the regulatory relationships among genes as a directed and edge-labeled graph in which each node represents a gene and edge denotes the existence regulatory relationship among genes. The labels on each edge correspond to whether the regulation is positive or negative regulation. Recently years, various methods to estimate TRNs from DNA microarray data have been proposed. In the rest of this paper, we referred the method to estimate TRNs as *network estimation methods*, for short.

As increasing the number of network estimation methods, the importance of evaluation and comparison of them have been emphasized. By such analysis, not only it becomes clear that strength and weakness of each network estimation method, but also we can build coming strategy for increasing the accuracy of estimation. As major instance of analysis, we can cite DREAM project[3]. For valid evaluation of network estimation methods, various types of TRNs their structures are already known in advance and corresponding gene expression profile data are required. However it is difficult to collect such dataset in practice. Then, artificially generated data has been used for evaluation. As an example of such artificial data generation methods, we can see GeneNetWeaver[4]. GeneNetWeaver can generate realistic artificial TRNs by extracting subnetwork from large real ones. Although artificial TRNs generated by GeneNetWeaver are plausible in biological sense, there is no discussion about the structural diversity among obtained artificial TRNs in GeneNetWeaver[4]. Confirmation of the structural diversity is important to clarify the connections between the performance of each network estimation method and the structure of true TRNs. In Afuso[5], authors tried to confirm structural diversity using global network characteristics. However, it was not enough because comparing the TRNs with global network characteristics cannot capture the partial similarity or equivalence of given TRNs.

On the other hand, to measure the similarity among directed graphs, the methods, so-called *network alignment*, have been proposed[6][7]. Using network alignment methods, we can extract the equivalent subnetwork between two given networks.

In this paper, we proposed new artificial data generation scheme so that one can confirm the structural diversity among generated TRNs. And also, we compared four score functions for Edge Orientation Problem[8], that is a part of network estimation problem, using artificial dataset generated according to proposed scheme.

The rest of this paper were organized as follows. First, we

Manuscript received November 5, 2012
Manuscript revised November 20, 2012

gave the explanation about traditional artificial data generation method and showed its problem. After that, we showed the outline of new data generation scheme considering structural diversity based on certain similarity between networks. Next, concrete definition of distance between networks that used in our proposed data generation scheme. Subsequently, to show its capability, we compared generation method for artificial TRNs used in Afuso[5] and proposed scheme in view of diversity of structural similarity. Finally we concluded our study.

2. Artificial Data Generation Scheme based on Structural Diversity

In this section, we showed the outline of traditional artificial data generation method, named GeneNetWeaver, and pointed out its problem on the evaluation with consideration of structural diversity.

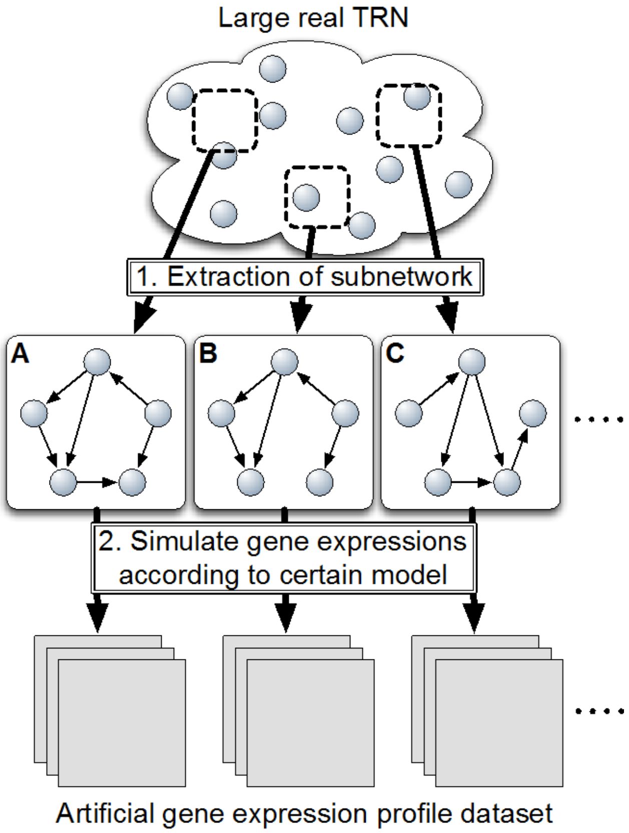

Large real TRN
1. Extraction of subnetwork
A
B
C
....
2. Simulate gene expressions according to certain model
....
Artificial gene expression profile dataset

Fig 1 Data generation steps of GeneNetWeaver.

The outline of generation of artificial dataset in GeneNetWeaver was shown in Fig.1. GeneNetWeaver consists of two steps, generation of artificial TRNs and one of artificial gene expression profile data. The method uses large real TRNs of S.cerevisiae[9] or E.coli[10] to generate artificial TRNs. At the first step, extraction of subnetwork in large real TRN is utilized using certain index, so-colled *modularity*. Extracted subnetworks are called *modules* and used as artificial TRNs. At the second step, GeneNetWeaver generates artificial gene expression profile data by simulating gene expressions according to the interactions among genes in each artificial TRN. To simulate gene expressions, GeneNetWeaver models the regulatory relationships as stochastic differential equations. For more details, see [4] and [11].

In the generation of artificial TRNs, biological plausibility of generated TRNs is confirmed by the fact that the generation is based on subsampling of real TRN. In [], it was shown that extracted subnetworks have certain structural property similar to real TRN. However, in GeneNetWeaver, there is no way to avoid the extraction of similar subnetworks, such like subnetwork A and B in Fig.1. In other words, it is difficult to confirm that generated TRNs contain diverse structures. The confirmation of structural diversity in the artificial TRNs is important for valid evaluation of network estimation methods.

To solve this problem, we extended data generation scheme of GeneNetWeaver as shown in Fig.2. In proposed data generation scheme, clustering of generated TRNs and selection of representatives for each cluster were used to remove similar artificial TRNs. The proposed data generation scheme consists of four steps. First, subnetwork extraction is carried out same to GeneNetWeaver. Next, to pull together the similar TRNs, clustering is utilized based on similarity of the structures in generated TRNs. After that, the representatives are determined for each cluster as test subject networks. And finally, same to GeneNetWeaver, the gene expressions are simulated according to certain model.

Following these steps, we can generate artificial TRNs and gene expression profile data with confirmation of structural diversity among the obtained TRNs.

3. Distance between Networks based on Network Alignment

In previous section, we proposed new data generation scheme using clustering of generated TRNs. To utilize the clustering of TRNs, the distance between them is required. In this section, we gave the definition of the distance between networks.

To define the distance between networks, we used the extraction of equivalent subnetworks between given two networks. If the extracted equivalent parts is large comparing to the size of given networks, then corresponding two networks can be considered similar each other. Generally, it is said that the extraction of equivalent parts between two networks is NP-hard. Various methods has been proposed to extract equivalent parts from undirected graphs[6][7]. The methods to extract equivalent parts of given networks is called *network alignment*. Traditional network alignment methods focused on undirected graphs. On the other hand, Afuso[12] proposed the method to extract them from directed graphs, named DiAliNe. DiAliNe calculates similarities among vertices in different two directed graph using network characteristics, such as clustering coefficient[13].

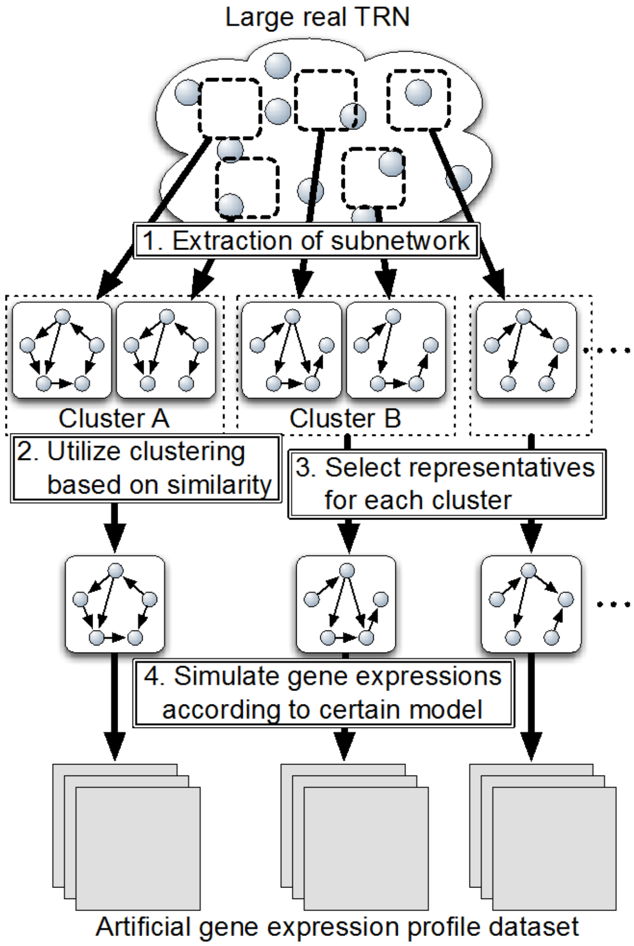

Large real TRN
1. Extraction of subnetwork
Cluster A
Cluster B
2. Utilize clustering based on similarity
3. Select representatives for each cluster
4. Simulate gene expressions according to certain model
Artificial gene expression profile dataset

Fig 2 Outline of proposed data generation scheme

After the calculation of similarity, the method finds the optimum matching f between vertex sets V_1 and V_2 in different two directed graphs $G_1(V_1,E_1)$ and $G_2(V_2,E_2)$ based on certain heuristics.

$$f : V_1 \to V_2 \quad (1)$$

For more details, see [12]. Using DiAliNe, we can extract the equivalent subnetworks between two networks and measure the sizes of them.

Based on above consideration, we defined certain index such that, when the sizes of equivalent subnetworks between two given networks is large, its value becomes small. We used such index as distance between given two networks. The distance $d(G_1,G_2)$ between extracted subnetworks G_1 and G_2 was defined as follows.

$$d(G_1,G_2) = \frac{1}{EC(G_1,G_2)} - 1 \quad (2)$$

where $EC(G_1,G_2)$ denotes edge coverage rate between directed graphs G_1 and G_2. Its value is calculated with Eq.3.

$$EC(G_1,G_2) = \frac{|(u,v) \in E_1 \wedge (f(u),f(v)) \in E_2|}{\min(|E_1|,|E_2|)} \quad (3)$$

where u and v are vertices in G_1. $f(u)$ denotes the vertex that is matched to vertex u with matching f.In Eq.2, we subtracted 1 because maximum of EC value is 1. By this subtraction, we forced the distance to oneself to be zero. Edge coverage rate reflects how many edges in smaller subnetwork G_1 is preserved in larger subnetwork G_2 by matching f.

Using Eq.2 and .3, we can calculate the distances among generated subnetworks. Using calculated distances, clustering could be utilized. To generate test subject networks, it is required to select the representatives for each cluster. In this paper, we determined representatives considering distances of intra-cluster. In each cluster, the subnetwork that is the closest to other subnetworks in a cluster can be considered as a centroid of corresponding cluster. From above consideration, we selected subnetworks that have minimum sum of distance to other subnetworks in same cluster as representatives of clusters.

Using the distance defined as Eq.2 and selection of representatives for each cluster, we can generate artificial TRNs with confirmation the structural diversity.

4. Comparison of Artificial TRNs in Traditional Research and Proposal

In this section, to show capability of confirmation of structural diversity, we compared artificial TRNs generated by the method in Afuso[5] and one by proposed scheme. In that paper, authors tried to confirm the structural diversity among generated artificial TRNs using global network characteristics. However, global characteristics may not reflect local equivalent structure in artificial TRNs. Then it was not enough to confirm the structural diversity. We compared two dataset of artificial TRNs using following steps.

(i) Using subnetwork extraction from large real TRN of E.coli[10], 160 subnetworks each one has 100 vertices, were extracted so that each vertex in real TRNs would be extracted ten times, averagely.
(ii) Calculate global network characteristics and represent each subnetwork as a vector that each element is the value of global network characteristic.
(iii) Using vector representation with global network characterisitcs, calculate the distances between subnetworks.
(iv) Utilize the clustering and determine the representatives for each cluster as test subject TRNs, by selecting the subnetwork nearest to the centroid of the cluster. We referred these TRNs as dataset A.
(v) Using DiAliNe, calculate another distances between subnetworks.
(vi) Using another distances, utilize the clustering and select representatives for each cluster according to our scheme. We referred these TRNs as dataset B.
(vii) Calculate EC values of each pair of TRNs in dataset A and one of each pair in dataset B.

In step.(iv) and step.(vi), we constructed nine clusters considering the dendrogram obtained by Ward method. Dendrogram obtained by global network characteristics and by network alignment were shown in Fig.3. and .4. In Fig.3 and .4, horizontal red line denotes the dividing line for construction of clusters. In Afuso[5], the subnetwork nearest to centroid was selected as a representative. And in our proposal, one has smallest sum of distances to other subnetworks in a cluster was selected as a representative. For each case, we calculated EC values among obtained nine representatives. The summary of calculations were shown in Table.1 and Table.2, respectively.

Table 1 EC value among representatives determined with the clustering base on global characteristics

Min	3rd Qu	Median	Mean	1st Qu	Max
0.0597	0.3778	0.5266	0.5213	0.7029	0.9524

Table 2 EC value among representatives determined with the clustering base on network alignment

Min	3rd Qu	Median	Mean	1st Qu	Max
0.1226	0.2517	0.3397	0.3745	0.4637	0.7647

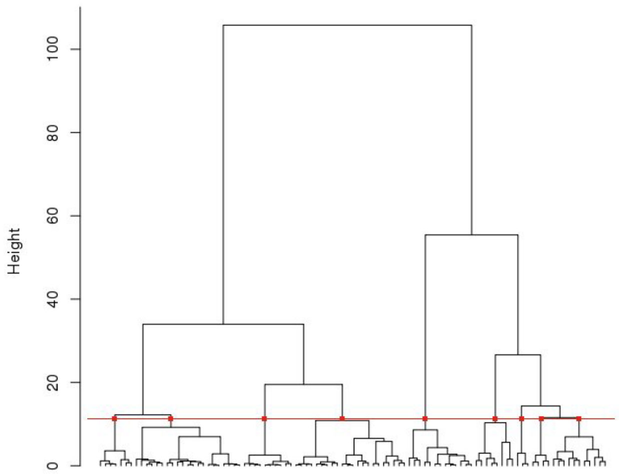


Figure 3 Dendrogram obtained with global characteristics.

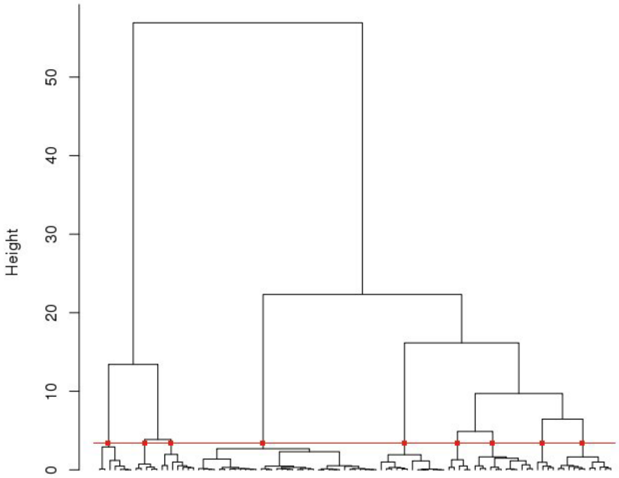


Figure 4 Dendrogram obtained with network alignment

From Table.1 and .2, maximum and mean of EC value among artificial TRNs generated by proposed scheme were smaller than ones generated by the method in Afuso[5]. From the result that maximum similarity among the representatives is smaller, then the structural diversity in obtained artificial TRNs is greater than generated by the one based on global network characteristics. Especially, as shown Table.1, the method based on global network characteristics may result certain pair of artificial TRNs that over 95 % of their structures are equivalent.

From these results, proposed data generation scheme could generate artificial TRNs that structural diversity has been confirmed, that has not discussed in major traditional artificial data generation method. And its capability to generate diverse structure is greater than the method used in Afuso[5].

5. Conclusion

In this paper, we proposed new data generation scheme for evaluation of network estimations. In traditional data generation method, structural diversity has not been discussed, or its discussion was not enough in view of the similarity of network structure. To solve that, we extended traditional major data generation method, GeneNetWeaver to one can confirm the structural diversity of generate TRNs. To do so, the distance beween subnetworks defined based on network alignment method, DiAliNe. Comparing to the subnetworks generated by Afuso[5], it was shown that proposed data generation scheme could construct more diverse subnetworks.

Although proposed data generation scheme successfully constructed artificial TRNs that have more diverse structures, it sometimes resulted similar TRNs in view of EC value. It may depend on the number of cluster that user decided or determination of representatives for each cluster. Then, as future tasks, wee see two subjects. First, including the method to decide the number of clusters automatically to proposed data generation scheme and second, proposal of more sophisticated selection strategy of representatives.

References

[1] DeRisi.J.L, Lyer.V.R, and Brown.P.O, “Exploring the metabolic and genetic control of gene expression on a genomic scale”, Science, 278, pp.680-686, 1997

[2] Simon.R, “Using DNA Microarrays for Diagnostic and Prognostic Prediction”, Vol.3, No.5, pp.587-595, 2003

[3] Columbia University and IBM, “Dialogue for Reverse Engineering Assessments and Methods”, http://wiki.c2b2.columbia.edu/dream/index.php/The_DREAM_Project

[4] Thomas.S, Daniel.M, and Dario.F, “GeneNetWeaver: In Silico Benchmark Generation and Performance Profiling of Network Inference Methods”, Bioinformatics, Vol.27, No.16, pp.2263-2270, 2011

[5] Hitoshi.A, Takeo.O, and Morikazu.N, “Comparative Study of Score Functions for Edge Orientation Problem in Network Estimation”, International Journal of Computer Science and Network Security, Vol.12, No.9, pp.32-38, 2012

[6] Aika.T and Jun.S, “Global Network Alignment using Graph Summarization for Comparison Gene Function”, SIG-BIO, Vol.24, No.12, pp.1-7, 2011

[7] Oleksii.K and Natasa.P, “Integrative Network Alignment Reveals Large Regions of Global Network Similarity in Yeast and Human”, Vol.27, No.10, pp.1390-1396, 2011.

[8] Hitoshi.A, Morikazu.N, Takeo.O, “Genetic Network Estimation with Covariance Selection and Score Function based on PageRank”, IPSJ technical reports, BIO-58, pp.5-8, 2008

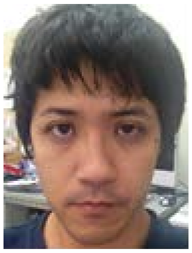
[9] Balaji.S, Babu.M.M, Iyer.L.M, Luscombe.N.M and Aravind.L, “Comprehensive Analysis of Combinatorial Regulation using the Transcriptional Regulatory Network of Yeast”, J. Mol. Biol. 360, pp.213–227, 2006

[10] Gama-Castro.S, et al, “RegulonDB version 7.0: Transcriptional regulation of Escherichia coli K-12 Integrated Within Genetic Sensory Response Units”, Nucleic Acids Res, 39, D98, 2011

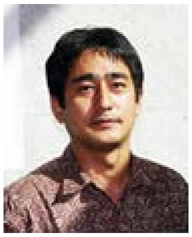
[11] Marabch.D, Schaffter.T, Mattiussi.C and Floreano.D, “Generating Realistic In Silico Gene Networks for Performance Assessment of Reverse Engineering Methods”, Journal of Comput. Biol, Vol.16, No.2, pp.229-239, 2009

[12] Hitoshi.A, Takeo.O and Morikazu.N, “Vertex Similarity based on Network Characteristics for Alignment Directed Graphs”, IPSJ technical reports, BIO-30(7), pp.1-6, 2012

[13] Zhiyu.L, Chen.W, Qiong.Z and Huayong.W, “Clustering Coefficient Queries on Massive Dynamic Social Networks”, Proc. 11th International Conference on Web-age Information Management, pp.115-126, 2010



Hitoshi AFUSO received the B.S. and M.S. degrees in Information Engineering from University of the Ryukyus in 2005 and 2008, respectively. He belongs to doctoral course in same university. His research area is Bioinformatics. Especially, he is interested in stochastic model of genetic expression and alignment of biological network such as TRNs.



Takeo OKAZAKI took B.Sc., M.Sc. from Kyushu University in 1987 and 1989, respectively. He had been a research assistant at Kyushu University from 1989 to 1995. He has been a lecturer at University of the Ryukyus since 1995. His research interests are statistical data normalization for analysis, statistical causal relationship analysis.

Morikazu NAKAMURA took B.E., and M.E. from University of the Ryukyus in 1989 and 1991, respectively. He took ph.D from Osaka University in 1995. He has been a professor at University of the Ryukyus. His research interest includes design and analysis of parallel and distributed algorithms.