Vol.2009-MPS-76 No.20
Vol.2009-BIO-19 No.20
2009/12/17

Comparative Study of Score Functions for Edge Orientation in Genetic Network Estimation

Hitoshi AFUSO†, Takeo OKAZAKI‡, Morikazu NAKAMURA‡
†: Graduate School of the University of the Ryukyus, Information Engineering,
‡: University of the Ryukyus, Information Engineering

Abstract

In this paper, we compare the score functions for *edge orientation problem* in estimation of genetic network from DNA microarray data. We focused four score functions, Standard Bayesian Metric(SBM), Bayesian Dirichlet Metric(BDM), K2 Metric and PageRank Orientation Metric(PROM). To compare and evaluate the performance of each score function in various situations, we utilized the generation method of artificial genetic networks and DNA microarray data and used those artificial data. To generate the networks that have certain network property, such as scale-free property, we used certain network generation models such as Barabasi-Albert Model. In the experiments, the number of edges such that orientated incorrectly was used to evaluate the performance of the score function.

1 Introduction

Inside life-form cells, many genes interact each other to utilize biological functions. The network that represents these interactions among genes is called as *genetic network*. Construction of genetic networks from DNA microarray data is an challenging topic in bioinformatics area[1].

Several methods to estimate a genetic network from DNA microarray data have been proposed, such as Boolean networks[2], differential equation model[3], Petri-net[4] and Bayesian network. The approach based on Bayesian network especially have been studied and shown successful results. However, traditional Bayesian network approach cannot handle the network that contains cyclic structures. Then we can say that traditional Bayesian network approach has difficulty to be applied to actual expression data from DNA microarray experiments. To address this problem, Afuso *et al*[5] proposed the estimation method that is constructed from two phases, *Directly Related Path Detection Phase* and *Edge Orientation Phase*[5]. In this method, interactions among genes are detected as undirected path in the graph where each node and path represent a gene and a interaction, respectively. As next step, searching the their orientation that maximize certain score function is done. Then, we can obtatin the genetic network from DNA microarray data. However, there are several score function for edge orientation phase, Standard Bayesian Metric(SBM), Bayesian Dirichlet Metric(BDM), K2 Metric, PageRank Orientation Metric that proposed in Afuso[5] and so on, although, in actual case, we cannot say that which score function is more effective for each target genetic network in *Edge Orientation Phase*.

In this paper, we compared the score functions for edge orientaion to determine which score function can lead more accurate edge orientation for each target genetic network.

To compare those score functions in various situation of target genetic network, we need the varied patterns of DNA microarray data and target genetic networks. But it is difficult to collect such data of actual . To this end, the artificial genetic network and artificial DNA microarray data were generated. To generate the networks that have certain property that corresponding to actual genetic network[6], we utilized network generation methods such as Barabasi-Albert model[7].

To evaluate the effectiveness of the score function, the number of the edges that oriented incorrectly was used. After the comparison, we made some discussion about the performance of each score function.

2 Edge Orientation Problem

In this paper, we focused the problem that corresponds to *Edge Orientation Phase* in Afuso[5]. The problem is called as Edge Orientation Problem. Edge orientation problem is to give the orientatin to each edge in given undirected graph such the orientation maximize certain score function.

The problem is formulated as follows.

INPUT:

1. Undirected graph G that represents directly related interaction among genes.
2. DNA microarray data matrix Dt.

$$Dt \in \{\log(\frac{R}{G})\}^{(m,n)}$$

   where m represent the sampling number, such as knock-out genes. And n is the number of vertices in given graph. R and G are color strength that observed in DNA microarray experiments, Red and Green.
3. Score function *score* from directed graph G' to real value v that denotes fitness between G' and given DNA microarray data Dt.

OUTPUT:

Directed graph G'_{max} that maximize score function *score*.

Focusing this problem, we compare the performance of score functions for edge orientation.

3 Score Functions for Edge Orientation

In *Edge Orientation Phase*, there are four alternative score functions.

Standard Bayesian Metric(SBM) is most popular one of Bayesian approach. In this score function, all variables in the network are assumed that they are multinomial distributed. This score function is based on the maximization of posterier probability. The SBM score$SBM(S)$ of candidate network S can be calculated by following formula.

$$\begin{aligned} SBM(S) &= \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i}(N_{ijk}+\alpha_{ijk}-1) \\ &\quad \times \log\frac{N_{ijk}+\alpha_{ijk}-1}{N_{ij}+\alpha_{ij}-r_i} \\ &\quad -Dim(S)log(N) \end{aligned}$$

where n, q_i and r_i denote the number of the variables, the number of value configuration of parents' of variable i and the number of value configuration of variable i, respectively. In this formula, N_{ijk} corresponds to the frequency of the cases that the variable i's value is in the k-th configuration and i's parents' configuration is in the j-th configuration. α represents the prior information. The term $Dim(S)log(N)$ is corresponding to penalty for network conplexity. From these terms, this score function can lead more simpler network.

Bayesian Dirichlet Metric(BDM) is another type of Baysian scor function. In this score function, all variables in the network are assumed that Dirichlet distributed. The Dirichlet distribution is obtained by extention of multi binomial distribution. The BDM score $BDM(S)$ is obtained by calculation of following formula.

$$\begin{aligned} BDM(S) &= \sum_{i=1}^{n}\{\sum_{j=1}^{q_i}\{\log\frac{\Gamma(\alpha_{ij})}{\alpha_{ij}+N_{ij}} \\ &\quad +\sum_{k=1}^{r_i}\log\frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})}\}\} \end{aligned}$$

where α and Γ represent Dirichlet prior parameters and gamma function.

As special case of BDM, there is another score function, that called K2 Metric(K2). In the $K2$ metric, all variables are also assumed those are Dirichlet distributed. However, in this score function, each prior parameter α are treated as they have same constant value. The value of K2 metric $K2(S)$ can be obtained from following formula.

$$K2(S) = \prod_{j=1}^{q}\frac{(r-1)!}{(N_j+r-1)!}\prod_{i=1}^{r}N_{ij}!$$

Previous three score functions are based on probablistic approach. On the other hand, in Afuso[5], the score function is based on network structural approach had been proposed. This score function is called as PageRank[8] Orientation Metric(PDM). To calculate this score function, at first, we estimate the PageRank of the target genetic network. Next, the PageRank value of candidate genetic network is also calculated. And finally, these two PageRank values are compared and if these values are similar, then candidate genetic network and target genetic network are also similar. To estimate the PageRank value from DNA microarray data, the sum of the values for each DNA microarray experiments are calculated and normalized by number of experiments.

$$\hat{pr}_i = \frac{\sum_i exp_{ij}}{N}$$

The PageRank value is relative value, so the $POM(S)$ of the network S is calculated by following formula.

$$POM(S) = Cor(epr, cpr)$$

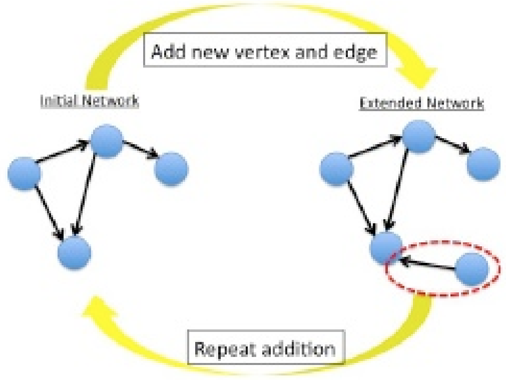
In this formula, Cor denotes Spearman's correlation function. And *epr* and *cpr* represent estimated PageRank and calculated PageRank, respectively.

4 Artificial DNA Microarray Data Generation

To compare the score functions in varied situations, we need various type of genetic networks that are known whole structure in advance and DNA microarray data corresponding those networks. However, it is difficult to collect such actual data. Then, the artificial genetic networks and DNA microarray data were substituted. In the generation of those, to produce the networks that have certain property such that actual genetic networks have[6], we utilized four network generation models, Barabasi-Albert model(BA), modified BA model(BA*), YB model(YB)[9] and modified YB model(YB*).

As network generation model to construct the networks that have scale-free property, Barabasi-Albert(BA) model was utilized. This network generation model is based on two aspects, preferential selection and network evolution. As another possible network property, we can see the small-world property. BA model cannot generate the networks that have small-world property, so we used another network generation model, that called YB model[9]. By using YB model, we can generate the networks that have both scale-free and small-world property. These BA and YB model can't produce the network that contain cyclic structure inside network. However, inside life form cells, it can be considered that genetic network contains cyclic structure, such as metabolic system for glucose. To evaluate the influence of cyclic structure to performance of edge orientation, we modified these BA and YB model to be able to generate cyclic network.

Figure 1: Basic Steps of Artificial Genetic Network Generation



After generation of artificial genetic network, artificial DNA microarray data was generated. In Bayesian approach, continuous DNA microarray data was discritized by utilizing certain thresholding. In this paper, we generated already discritized artificial DNA microarray data into 0 and 1. To generate artificial DNA microarray data, at first, binary initial expression vector $v_{initial}$ was generated randomly.

$$v_{initial} \in \{0,1\}^n$$

The value of the elementss of vector $v_{initial}$ is set to 1 if corresponding genes were expressed, and to 0 else. Next, by multiplying transition matrix $\mathbf{T}$ that is obtained by transposing the adjacency matrix $\mathbf{A}$ corresponding to target genetic network, we obtained one experiment data sample exp_i.

$$\mathbf{A} \in \{0,1\}^{(n,n)}$$

$$\mathbf{T} \;=\; {}^{t}\mathbf{A}$$

The value of each element(i, j) in matrix $\mathbf{A}$ is set to 1 if there is connection between gene i and j, and to 0 else. Representing generation of artificial DNA microarray data mathematically, we can lead following formula.

$$exp_i \;=\; \mathbf{v}_{initial} + T\mathbf{v}_{initial} + \cdots + T^l\mathbf{v}_{initial}$$

where l denotes depth that gene expression reaching. Repeating these process, we generated the artificial DNA microarray data Dt^* from artificial genetic network.

$$Dt^* \;=\; \begin{pmatrix} exp_1 \\ exp_2 \\ \vdots \\ exp_n \end{pmatrix} \qquad (1)$$

In Formula.(1), n denotes the number of artificial DNA microarray experiments.

The example of generation of artificial DNA microarray data is shown in Fig.4.

5 Experiments and Results

In comparative experiments, to evaluate the performance of score functions, we need enough number of network types that cover the almost actual cases of genetic network. In this paper, probablistic artificial genetic network generation methods were used. So, it is difficult to control network type of generated artificial genetic network. Then, we calculated network characters for each generated network to confirm the type of network.

However, in the generated artificial genetic networks, there might be some equivalent network structure or property. So, we devided these generated genetic networks into some categories considering the value of network characters. After that, we executed the searching of maximal orientaion using four score functions for each network type, although these results might contain some statistically unsignificant results. To prevent these results, we applied multiple comparison method to obtained results.

In artificial DNA microarray data generation, we fixed the number of vertices in artificial genetic network. The number of vertices in the generated genetic network was 20. 120 artificial genetic networks were generated for each network generation model.

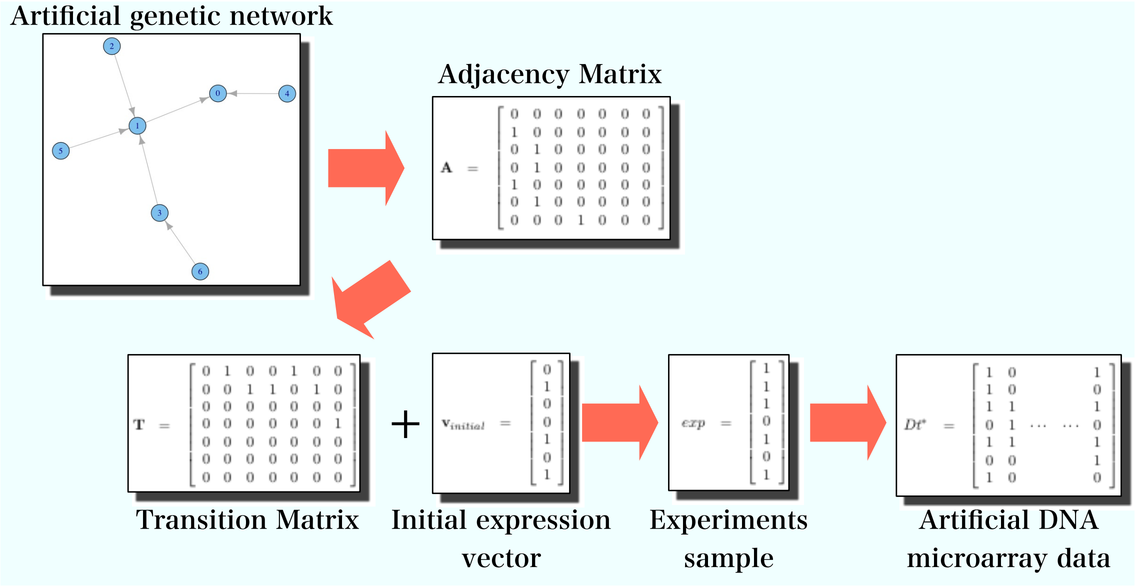
Next, to categorize generated genetic network, the values of network characters were calculated. As network character, we used the mean and variance of path legth, in-degree of each node, the number of cycles and the length of each cycles, respectively. Mean and variance of path length are corresponding the unevenness of connectivity among genes. The mean and variance of in-degree of each node denotes density of genetic network. After that, based on the miximization of BIC[10], we categorized the value of each network character into some levels.

After the categorization, we obtained 217 unique combination of value. We regarded these unique combination as network type. To cut-back the number of unique network types more, hierarchial clustering was utilized by representing each generated artificial genetic network as 8 dimensional vector. We used normalized manhattan distance as distance function and Ward method as clustering method. After the clustering, 38 unique network types without cyclic structure and 39 with cyclic structure were obtained. From generated 78 artificial genetic network types, we selected typical network by focusing similarity and drawing network graphically.

For each artificial genetic network we selected, we generated 30 artificial DNA microarray data. Using that artificial DNA microarray data, the trials of searching the optimal orientation were executed with genetic algorithm(GA). In GA searching, the number of chrosomes was 100 and the number of generations was 500. we represented direction of each edge as binomial value in corresponding digit in chrosome. Chrosome selection method was combination of 10% elite and roulette. Selected chrosomes were mixed by uniformed cross-over method. To mutate chrosomes, we selected 10 chrosomes randomly and changed one digit.

After searching optimal, we determined the correctness of each direction of the edge and counted the number of miss directions(MD). In the experiments, selected unique genetic networks had different number of edges. Then, we normalized the MD

Figure 2: The Example of Generation of Artificial DNA Microarray Data


Artificial genetic network
2
0
4
1
5
3
6
Adjacency Matrix
$\mathbf{A} = \begin{bmatrix} 0&0&0&0&0&0&0 \\ 1&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0 \\ 0&1&0&0&0&0&0 \\ 1&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0 \\ 0&0&0&1&0&0&0 \end{bmatrix}$
$\mathbf{T} = \begin{bmatrix} 0&1&0&0&1&0&0 \\ 0&0&1&1&0&1&0 \\ 0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&1 \\ 0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0 \end{bmatrix}$
+
$\mathbf{v}_{initial} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$
$exp = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$
$Dt^{*} = \begin{bmatrix} 1&0&&&1 \\ 1&0&&&0 \\ 1&1&&&1 \\ 0&1&\cdots&\cdots&0 \\ 1&1&&&1 \\ 0&0&&&1 \\ 1&0&&&0 \end{bmatrix}$
Transition Matrix
Initial expression vector
Experiments sample
Artificial DNA microarray data

value with the number of edges. For each artificial genetic network, the mean, best, worst and variance of MD value among 30 trials were calculated. The mean of MD for 38 network types that contain no cyclic structure are shown in Table.2.

To obtaine the statistically significant results, the multiple comparison method, Tukey-Welsch's method[11], was applied to the results. In the multiple comparison, each null-hypothesis was set to that comparing score functions are all same. The results of multiple comparison are shown in Table.1.

6 Discussion

For the almost network types, PROM led more accurate edge orientation. BDM can obtain the accurate orientation next to PROM, although, PROM's variance of results were high. This result shows that the PROM has less confidence of its results than another Bayesian approach score functions. The results of from type01 to type04 shows that Bayesian approaches are more effective in those network types. Those networks are very simpler and very sparse. The results can be led by complex network structure.

In artificial DNA microarray data generation method, the value was already discritized 0 and 1. Using generation method for artificial DNA microarray data in this paper, if generated artificial genetic network contains such structure, then almost variables in target genetic network may have same reaction. In other words, almost rows in artificial DNA microarray data may have similar pattern. By the similarity of rows, it can be blured whether focusing variables have some connection or not. However, actual DNA microarray data contains continuous values and they are descritized by using certain threshold. For more detailed discussion, another method for artificial DNA microarray data generation considering such threshold is required.

References

[1] L.J.Steggles, R.Banks, A.Wipat, "Modeling and Analysing Genetic Networks: From Boolean Networks to Petri Nets", Computational Methods in System Biology, vol4210, pp.127-141, 2006

[2] T.AKutsu, S.Kuhara, O.Maruyama, S.Miyano, Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions, Proc 9th ACM-SIAM SODA, pp.695-702, 1998

[3] E.V.Someren, L.Wessels, and M.Reinders, Linear modeling of genetic networks from experimental data, ISMB, vol.18, pp.355-366, 2002

[4] T.Hayashi, M.Nakamura, T.Okazaki, A Petri Net Model of Gene Networks and Its Identification., Proceedings of the Society Conference of IEICE, pp.162, 2002

[5] H.Afuso, M.Nakamura, T.Okazaki, "Genetic Network Estimation with Covariance Selection and Score Function based on PageRank", IPSJ SIG, 58, pp.5-8, 2008

[6] Guelzim N, Bottani S, Bourgine P, et al. Topological and causal structure of the yeast transcriptional regulatory network Nat Genet, 31(1), pp.60-63, 2002

[7] A.L. Barabasi, Z.N. Oltvai, Network biology: understanng the cells functional organization, Nat Rev Genet, vol5, pp.101-113, 2004

[8] S.Brin, L.Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", COMPUTER NETWORKS AND ISDN SYSTEMS, vol30, pp.107-117, 1998

[9] K.Yabuki, T.Yabuki, "Emergent Model Reproducing the Global Features of Real Networks", IPSJ-ICS, vol85, pp.211-218, 2004

[10] E.Castilli, J.M.Gutierrez, and A.S.Hadi, "Expert Systems and Probablistic Network Models", Springer-Verlag, NewYork, 1997

[11] Y.Nagata, M.Yoshida, "Basic of Statistical Multiple Comparison Method", Scientist-sha, 1997

Table 1: The Results of Multiple Comparison

	BEST	WORST
Type02	BDM, PROM	BIC, K2
Type03	BDM	
Type04	PROM	
Type05	PROM	K2
Type06	PROM	K2
Type07		K2
Type08	PROM	K2
Type09	BIC, PROM	BDM,K2
Type10	PROM	K2
Type11	PROM	
Type12	PROM	K2
Type13	PROM	K2
Type14	PROM	K2
Type15	PROM	
Type16	PROM	K2
Type17	BIC, PROM	
Type19	PROM	
Type20	PROM	BDM, K2
Type21	PROM	
Type22	PROM	
Type23	PROM	
Type24	PROM	
Type25	PROM	
Type26	PROM	
Type27	PROM	
Type28	PROM	BDM, K2
Type29	PROM	
Type30	PROM	
Type31	PROM	
Type32	PROM	
Type33	PROM	
Type34	PROM	
Type35	PROM	
Type36	PROM	BDM, K2
Type37	PROM	
Type39	PROM	

Table 2: Mean of Each Score Functions against Network Types

	type01	type02	type03	type04	type05	type06	type07
SBM	0.1122807	0.2614035	0.1894737	0.2105263	0.2421053	0.2859649	0.2052632
BDM	0.0964912	0.1859649	0.1491228	0.2070175	0.2982456	0.2929824	0.1824561
K2	0.1210526	0.2789474	0.2157895	0.2175439	0.3736842	0.3631579	0.2649123
PROM	0.1210526	0.2000000	0.2157895	0.1842105	0.1315789	0.1771930	0.1912281

	type08	type09	type10	type11	type12	type13	type14
SBM	0.2701754	0.2400000	0.2900000	0.2983333	0.2158730	0.3111111	0.3476190
BDM	0.2403509	0.2916667	0.3400000	0.2783333	0.2301587	0.3428571	0.3777778
K2	0.3684211	0.3200000	0.4500000	0.3383333	0.3126984	0.4365079	0.4793651
PROM	0.1245614	0.2000000	0.1500000	0.2033333	0.1380952	0.1015873	0.1190476

	type15	type16	type17	type18	type19	type20	type21
SBM	0.2927536	0.3507246	0.1768116	0.4173913	0.6490741	0.6943089	0.7201754
BDM	0.3000000	0.3724638	0.2376812	0.4434783	0.6740741	0.7536585	0.7114035
K2	0.3289855	0.3956522	0.3028986	0.4855072	0.6833333	0.7601626	0.7535088
PROM	0.2101449	0.1434783	0.1666667	0.1449275	0.3796296	0.2861789	0.2956140

	type22	type23	type24	type25	type26	type27	type28
SBM	0.7270270	0.6638095	0.7558333	0.4784946	0.4218750	0.5645833	0.5322222
BDM	0.7000000	0.6600000	0.7616667	0.4989247	0.4364583	0.5437500	0.5722222
K2	0.7081081	0.6542857	0.7500000	0.5161290	0.4718750	0.5812500	0.5811111
PROM	0.4054054	0.2476190	0.3725000	0.1763441	0.2447917	0.1604167	0.2411111

	type29	type30	type31	type32	type33	type34	type35
SBM	0.4989899	0.5382353	0.4737374	0.4448276	0.4379310	0.4525253	0.5000000
BDM	0.5030303	0.5656863	0.5020202	0.4574713	0.4781609	0.4464646	0.5303922
K2	0.5424242	0.5794118	0.5232323	0.5218391	0.5206897	0.4848485	0.4921569
PROM	0.1808081	0.2382353	0.2555556	0.2264368	0.1436782	0.2292929	0.2500000

	type36	type37	type38
SBM	0.3623656	0.4533333	0.4468750
BDM	0.4000000	0.4522222	0.4906250
K2	0.4333333	0.4844444	0.4843750
PROM	0.2000000	0.1533333	0.2489583