Chem-Bio Informatics Journal, Vol.17, pp.53-71 (2017)

Using A Priori Knowledge after Genetic Network Inference: Integrating Multiple Kinds of Knowledge

Shuhei Kimura[1*], Koji Kitazawa[1], Masato Tokuhisa[1], Mariko Okada-Hatakeyama[2]

[1]*Graduate School of Engineering, Tottori University, 4-101, Koyama-Minami, Tottori 680-8552, Japan*
[2]*Institute for Protein Research, Osaka University, 3-2, Yamadaoka, Suita, Osaka 565-0871, Japan*

**E-mail: kimura@eecs.tottori-u.ac.jp*

(Received March 31; accepted June 3; published online June 17, 2017)

Abstract

Several researchers have focused on the inference of genetic networks as a process for extracting useful information from gene expression data. Their work has led to the proposal of a number of methods for genetic network inference. Yet the genetic networks inferred by these methods often contain large numbers of false-positive regulations along with the true-positives. One effective way to reduce the number of erroneous regulations is to apply inference methods that use a priori knowledge on the properties of the genetic networks. The existing inference methods adopting this approach generally use a priori knowledge and the observed gene expression data simultaneously to determine whether or not the target genetic network actually contains each of the candidate regulations. In this study, we establish a new framework for "using a priori knowledge after genetic network inference." The framework uses a priori knowledge only to modify the genetic network that has already been inferred by the other inference method. Based on this framework, we propose a new inference method that uses multiple kinds of a priori knowledge about genetic networks. The proposed method effectively combines multiple kinds of knowledge and computes the confidence values of regulations. Here, we confirm the effectiveness of the proposed method by applying it to artificial and actual genetic network inference problems. While only a small improvement is gained from the use of multiple kinds of a priori knowledge, we can improve the performance of many other existing inference methods by combining them with the method we propose here.

Key Words: Inference of Genetic Networks, A Priori Knowledge, Using A Priori Knowledge after Genetic Network Inference, Function Optimization

Area of Interest: Bioinformatics and its applications in medicine

1. Introduction

Several researchers have focused on the inference of genetic networks as a means for extracting useful information from gene expression data. The inference of genetic networks is a problem in which mutual regulations between genes are inferred from the observed time-series of gene expression data. The inferred models are conceived as ideal tools to help biologists generate hypotheses and design experiments. While numerous methods have been proposed for the inference of genetic networks [1-3], the models inferred by these methods often contain false-positive regulations along with the true-positives. One approach to remove these erroneous regulations from the inferred models is to utilize a priori knowledge about the target networks. Methods have been developed to reduce the number of erroneous regulations using a priori knowledge about the properties of genetic networks such as their sparseness [4-6], scale-free structure [7], etc.

Our group recently proposed an inference method that uses another kind of a priori knowledge, i.e., a hierarchical structure [8]. The first step in this hierarchy-based method [8] is to obtain multiple genetic networks from the given gene expression data using the BS-LPM inference method [9], a combination of the existing inference method [10] and a bootstrap method [11]. The second step is to detect a hierarchical structure that is consistent with most of the inferred networks. The hierarchical structure obtained is then used to compute the confidence values of all of the candidate regulations.

The existing inference methods generally use a priori knowledge about genetic networks and the observed gene expression data simultaneously in order to determine whether or not the target genetic network actually contains each of the candidate regulations. For example, when estimating model parameters that represent regulations of genes, PEACE1 [4] forces most of these parameters down to zero in order to utilize the sparseness of genetic networks. In this study, we call this framework “using a priori knowledge while inferring genetic networks.” The inference method based on the hierarchical structure, meanwhile, uses a priori knowledge only to modify the genetic network that has been already inferred by the other inference method. We could therefore say that the method uses the a priori knowledge after the genetic network inference. Our trial runs with this method proved that the use of a priori knowledge can improve the quality of the inferred networks even after the genetic network inference.

We believe that the new framework for “using a priori knowledge after genetic network inference” allows us to utilize several kinds of a priori knowledge that have never previously been used for the inference of genetic networks. Based on this framework, we therefore propose a new inference method that uses multiple kinds of a priori knowledge about genetic networks. The proposed method effectively combines multiple kinds of knowledge and then computes the confidence values of the regulations. In this study, we report the proposed method and confirm its effectiveness by applying it to various artificial and actual genetic network inference problems. We should note here that the proposed method does not infer regulations but assigns confidence values to all of the candidate regulations. However, we can construct genetic networks by gathering regulations whose confidence values exceed a threshold. Moreover, when biologists try to perform experiments for confirming the inferred regulations of genes, the confidence values could be used to determine the order of the experiments.

2. Using A Priori Knowledge after Genetic Network Inference

According to the framework for "using a priori knowledge after genetic network inference," the method proposed in this study utilizes multiple kinds of a priori knowledge about the properties of genetic networks. This section establishes the procedure for this framework.

According to the following procedure, the proposed method integrates multiple kinds of a priori knowledge and computes the confidence values of all of the regulations.

1. Infer multiple genetic networks from the observed gene expression data. For this purpose, we can use any inference method that is capable of producing multiple genetic networks. As its computational cost is quite low, however, this study uses the BS-LPM inference method [9]. First, the BS-LPM inference method constructs multiple gene expression datasets on the basis of a bootstrap method [11]. The BS-LPM inference method then infers a genetic network from each of the constructed dataset by using the existing inference method, .i.e., the LPM-based inference method [10]. The LPM-based inference method is a fast method that infers genetic networks by solving linear programming problems.
2. Use the genetic networks obtained to compute the confidence values of all of the candidate regulations. The confidence value of the regulation of the *n*-th gene from the *m*-th gene, $p^B_{n,m}$ (m,n = 1, 2, ⋯ , N), is computed by
$$p^B_{n,m} = \frac{N^B_{n,m}}{N_B}, \tag{1}$$
where N_B is the number of the networks inferred in the step 1, $N^B_{n,m}$ is the number of inferred networks that contain the regulation of the *n*-th gene from the *m*-th gene, and N is the number of genes contained in the target network.
3. Compute the confidence values of the regulations of the inferred networks by checking whether they are consistent with each kind of a priori knowledge. Based on the *i*-th kind of a priori knowledge, this study computes the confidence value of the regulation of the *n*-th gene from the *m*-th gene as $p^{(i)}_{n,m}$ (i = 1, 2, ⋯, N_{AP}, m,n = 1, 2, ⋯, N), where N_{AP} is the number of kinds of a priori knowledge applied. In this study, we assume that the value of $p^{(i)}_{n,m}$ increases from 0 to 1 with an increasing degree of confidence. Here, we use three kinds of a priori knowledge. The section 3 will describe a way to compute confidence values based on each kind of knowledge.
4. Modify the confidence values by integrating all of the confidence values obtained in the previous steps. The modified confidence value of the regulation of the *n*-th gene from the *m*-th gene, $p_{n,m}$ (m,n = 1, 2, ⋯, N), is defined as
$$p_{n,m} = \frac{p^B_{n,m} + \sum_{i=1}^{N_{AP}} w_i g\left(p^{(i)}_{n,m}; \alpha_i, \beta_i\right)}{1 + \sum_{i=1}^{N_{AP}} w_i}, \tag{2}$$
where
$$g(x;\, \alpha, \beta) = \begin{cases} 0, & (\text{if } x < \alpha), \\ \frac{x-\alpha}{\beta-\alpha}, & (\text{if } \alpha \le x < \beta), \\ 1, & (\text{otherwise}), \end{cases} \tag{3}$$
and w_i, α_i and β_i ($0 \le \alpha_i \le \beta_i \le 1$; i = 1, 2, ⋯, N_{AP}) are constant parameters. The section 4 will describe a method to find the optimal values for these constants. Note that reference [8] simply combines the confidence values. Here, on the other hand, we combine them using the non-linear function g.

5. Output the modified confidence values of all of the candidate regulations, $p_{n,m}$'s.

Note here that our framework for "using a priori knowledge after genetic network inference" can be combined with the framework for "using a priori knowledge while inferring genetic networks." In our framework, the method tries to improve the genetic networks using the structural properties of the networks that have been already inferred by the other inference method applied. Therefore, the improvement in the performance of our method depends much on the quality of the genetic networks inferred by the applied inference method. On the other hand, when we use the framework for "using a priori knowledge while inferring genetic networks," we can directly reflect the knowledge in the inferred networks. Thus, when we can introduce some kind of a priori knowledge on the basis of both of the frameworks, we should introduce it based on the framework for "using a priori knowledge while inferring genetic networks."

3. A Priori Knowledge about Genetic Networks

In this study, we seek to infer more reasonable genetic networks by introducing into the proposed method three kinds of a priori knowledge that have received little attention in genetic network inference. However, note that it would be still possible for the proposed method to use other kinds of a priori knowledge. This study used three kinds of the knowledge just because they seemed easy to introduce.

None of the three kinds of knowledge described in this section distinguishes the regulation of the *n*-th gene from the *m*-th gene and vice versa. When we evaluate the confidence values of these regulations according to each kind of knowledge, the values are therefore always the same, i.e., $p_{n,m}^{(i)} = p_{m,n}^{(i)}$. In contrast, the BS-LPM inference method used in this study distinguishes the regulation of the *n*-th gene from the *m*-th gene and vice versa, hence $p_{n,m}$ and $p_{m,n}$ computed according to the equation (2) are not always the same. The BS-LPM inference method is also capable of inferring an auto-regulation/auto-degradation, i.e., a regulation of a gene by itself. Here, however, we disregard auto-regulations/auto-degradations, as the a priori knowledge used in this study cannot cope with them. Inferred networks usually contain auto-regulations/auto-degradations, because inference methods often infer the degradation of gene transcripts as a regulation of the gene by itself. We would not always need to search for regulations that usually exist. Thus, we doubt that the inference of auto-regulations/auto-degradations is always essential for the inference of actual genetic networks.

3.1 Hierarchical structure

Biochemical networks are known to exhibit hierarchical structures [12]. The hierarchical structure of a network is composed of vertices that cluster together into groups, which in turn form into groups of groups, and so forth. If we know the hierarchical structure in a target network, we can improve a network inferred by an inference method. That is, we can conclude that inferred regulations are unreasonable if they are inconsistent with the hierarchical structure.

A hierarchical random graph model is capable of representing a hierarchical structure in a network [12]. This study obtains a hierarchical random graph model consistent with most of the genetic networks inferred in the step 1 of the procedure described in the previous section. The hierarchical random graph model obtained approximately represents the hierarchical structure of the actual target network and provides the probabilities that the target network has interactions between genes. We thus use these probabilities as the confidence values. This study represents the

confidence values evaluated based on the hierarchical structure as $p_{n,m}^{(1)}$'s. In reference [8], readers can find a detailed algorithm to extract a hierarchical structure from multiple genetic networks.

3.2 Common neighbors

Interactions among functionally related genes are reported to frequently occur [13]. If two genes share a large number of interacting genes, the two genes can therefore be assumed to interact with each other.

The interacting genes shared by two genes are referred to as common neighbors of the two genes [14]. In this study, we use the knowledge described here by counting the common neighbors of genes. Based on the common neighbors, we compute a confidence value of the regulation of the *n*-th gene from the *m*-th gene, $p_{n,m}^{(2)}$, by

$$p_{n,m}^{(2)} = \frac{1}{N}\frac{1}{N_B}\sum_{j=1}^{N_B}\left|C_j(n) \cap C_j(m)\right|, \quad (4)$$

where $C_j(n)$ is a set of genes that directly interact with the *n*-th gene in the *j*-th network obtained in the step 1 of our procedure, and *N* is the number of genes contained in the target network. As described previously, our method disregards auto-regulations/auto-degradations. Therefore, $C_j(n)$ does not contain the *n*-th gene.

3.3 Degree correlation

Biochemical networks show negative degree correlations [15]. This means that a gene interacting with a large number of genes tends to interact with a gene interacting with a small number of genes.

Interactions between genes interacting with many genes and those between genes interacting with few genes contradict the knowledge of the negative degree correlation. When we use this knowledge, therefore, we assign low confidence values to these interactions. This study thus defines $p_{n,m}^{(3)}$, the confidence value of the regulation of the *n*-th gene from the *m*-th gene, a value determined based on the degree correlation,

$$p_{n,m}^{(3)} = \frac{1}{1+|d_{n,m}|}, \quad (5)$$

where

$$d_{n,m} = D(n) + D(m) - D_{max} - D_{min}, \quad (6)$$

$$\mathrm{D}(n) = \frac{1}{N_B}\sum_{j=1}^{N_B}\left|C_j(n)\right|, \quad (7)$$

$$D_{max} = \max\{D(1), D(2), \cdots, D(N)\}, \quad (8)$$

$$D_{min} = \min\{D(1), D(2), \cdots, D(N)\}, \quad (9)$$

and max *S* and min *S* are operators that respectively return maximum and minimum values of the members of the set *S*.

4. Adjustment of the Constant Parameters

A confidence value evaluated based on some kind of a priori knowledge would not always be proportional to a degree of confidence. Even when we simply combine multiple confidence values, it could be difficult to improve the quality of the inferred genetic networks. Instead, this study combines the confidence values transformed non-linearly by the function (3).

In order to compute the modified confidence values $p_{n,m}$'s using the equation (2), we must determine values for the constant parameters $\mathbf{w} = (w_1, w_2, \cdots, w_{N_{AP}})$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_{N_{AP}})$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_{N_{AP}})$. For this purpose, we use several inference problems of artificial genetic networks. This study adjusts the parameters with a view to improving the performance of the proposed method on these problems. We thus define the adjustment of the constant parameters as a minimization problem of the following function.

$$f(\mathbf{w}, \mathbf{s}, \mathbf{t}) = \sum_{k=1}^{N_p} \max\{(1 + C) \times AURPC_0(k) - AURPC_{propose}(k; \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}), 0\}, \quad (10)$$

where $\mathbf{s} = (s_1, s_2, \cdots, s_{N_{AP}})$, $\mathbf{t} = (t_1, t_2, \cdots, t_{N_{AP}})$,

$$\alpha_i = \begin{cases} 0, & (\text{if } s_i < 0), \\ s_i, & (\text{if } 0 \le s_i < 1), \\ 1, & (\text{otherwise}), \end{cases} \quad (11)$$

$$\beta_i = \begin{cases} \alpha_i, & (\text{if } t_i < 0), \\ \alpha_i + (1 - \alpha_i)t_i, & (\text{if } 0 \le t_i < 1), \\ 1, & (\text{otherwise}), \end{cases} \quad (12)$$

N_p is the number of the artificial genetic network inference problems applied, and C (≥ 0) is another constant parameter. As mentioned previously, the parameters α_i and β_i must satisfy the condition $0 \le \alpha_i \le \beta_i \le 1$. Here, however, we face the difficulty of solving constrained optimization problems. We therefore remove the constraints by searching for the parameters **w**, **s** and **t** instead of searching for the parameters **w**, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

$AURPC_{propose}(k; \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is the area under the recall-precision curve (AURPC) of the proposed method with the parameters **w**, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ on the *k*-th genetic network inference problem. $AURPC_0(k)$ is the AURPC of the inference method based on the hierarchical structure [8]. Note that, when the hierarchical structure is the only a priori knowledge used and the constant parameters w_1, α_1 and β_1 are set to $\frac{1}{N_B - 1}$, 0 and 1, respectively, the proposed method is equivalent to the inference method based on the hierarchical structure. The AURPC is a performance measure that increases from 0 to 1 as the performance of an algorithm improves. The AURPC of an algorithm is obtained by checking its recalls and precisions. The recall and the precision are defined as

$$\text{recall} = \frac{TP}{TP + FN}, \qquad \text{precision} = \frac{TP}{TP + FP},$$

where *TP*, *FP* and *FN* are the numbers of true-positive, false-positive and false-negative regulations, respectively. We compute the recall and precision by constructing a network of regulations whose confidence values exceed a threshold and then comparing it with the target network. Next, we obtain the recall-precision curve of the algorithm by changing the threshold for the confidence value. Note here that, as described previously, our method does not assign confidence values to auto-regulations/auto-degradations. We therefore disregard these regulations in the evaluation of the recalls and precisions.

By optimizing the objective function (10), this study tries to obtain a method that performs, if not always much better, at least not worse than the inference method based on the hierarchical structure. We propose this objective function in order to obtain parameters that will not cause the over-learning, a phenomenon to be avoided in the machine learning field. Any function optimization algorithm can be used to optimize the objective function (10). As it seems to be multimodal, however, this study uses an evolutionary algorithm, AGLSDC [16], to minimize the function.

The overall workflow of the proposed approach is shown in Figure 1.

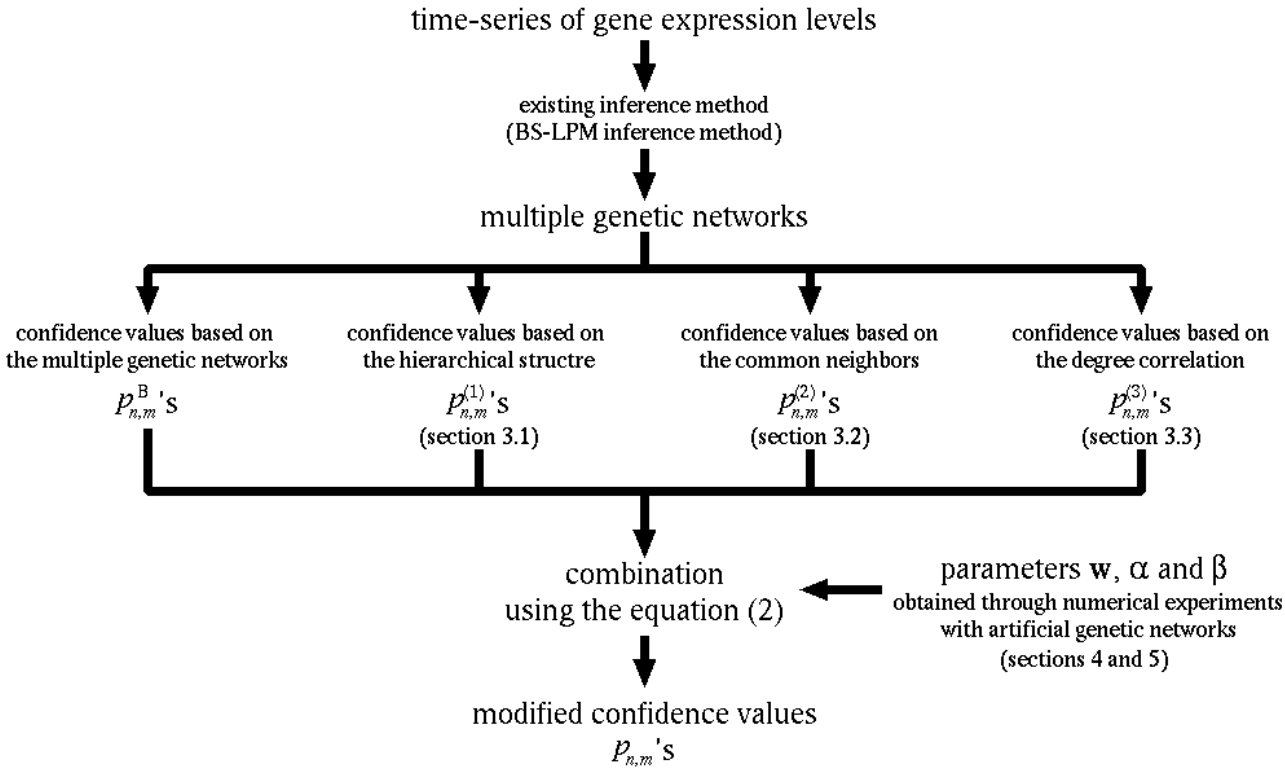

Figure 1. The overall workflow of the proposed approach

5. Experiments on Artificial Problems

In this section, we check the performance of the proposed method on artificial genetic network inference problems and then determine the constant parameters.

5.1 Genetic network inference problems

This study used 50 artificial genetic network inference problems constructed earlier for the evaluation of the inference method based on the hierarchical structure [8]. Each 10 problems had target networks whose topologies are identical to those of each of the five networks provided by the DREAM3 *in silico* network challenges, i.e., there were ten each of Ecoli1, Ecoli2, Yeast1, Yeast2 and Yeast3 [17] (Figure 2). Every target network consisted of 100 genes (N = 100). The network designs are based on actual biochemical networks and therefore reflect the actual topological properties. Note here that all kinds of a priori knowledge used in this study assign confidence values to regulations only on the basis of the topologies of the inferred networks. Thus, although the target networks are artificial, the experiments we describe here are capable of proving the effectiveness of the proposed approach.

The target networks of the inference problems were described using an S-system model [18]. The S-system model is a set of differential equations of the form

$$\frac{dX_n}{dt} = \alpha_n \prod_{m=1}^{N} X_m^{g_{n,m}} - \beta_n \prod_{m=1}^{N} X_m^{h_{n,m}}, \quad (n = 1, 2, \cdots, N), \tag{13}$$

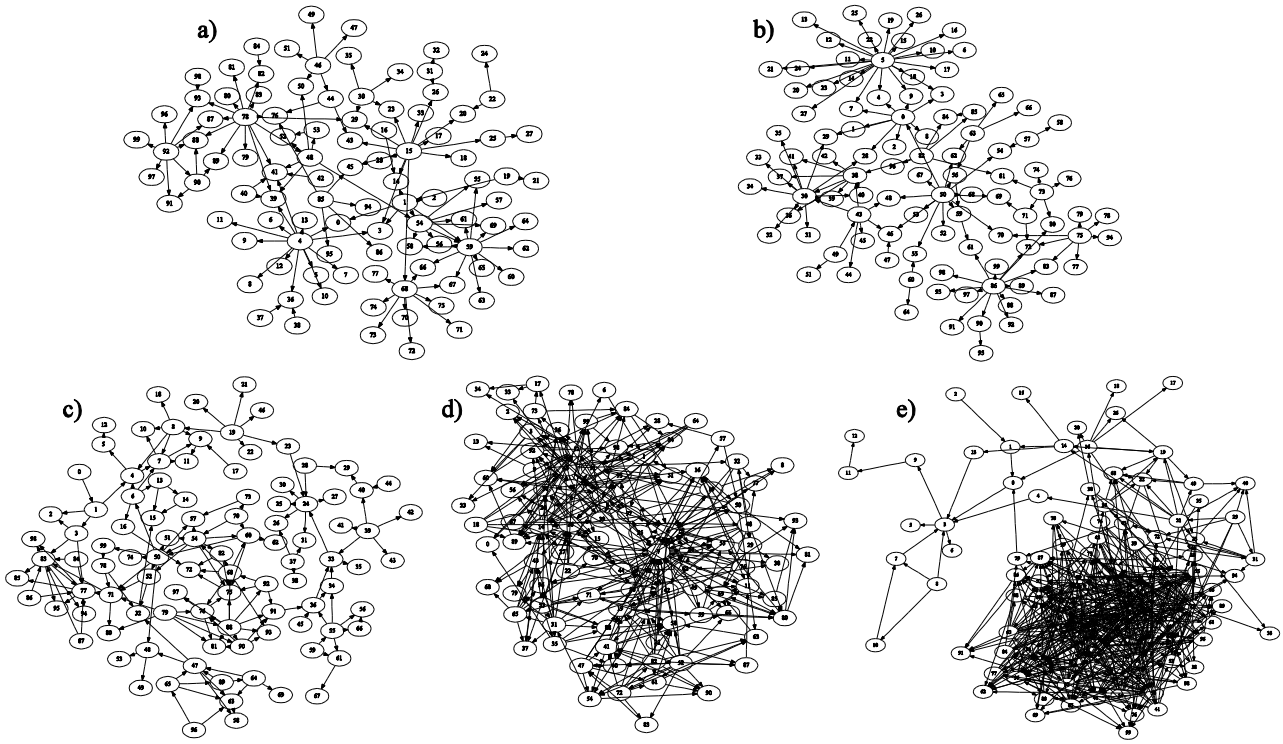

Figure 2. The network structures of a) Ecoli1, b) Ecoli2, c) Yeast1, d) Yeast2 and e) Yeast3
These networks are obtained from the DREAM3 *in silico* network challenges [17].

where X_n represents the expression level of the *n*-th gene, *N* represents the number of genes contained in the target network, and α_n (> 0), β_n (> 0), $g_{n,m}$ and $h_{n,m}$ are model parameters. The $g_{n,m}$ and $h_{n,m}$ parameters determine the topology of the network. These values were set according to the following rules: The value for $g_{n,m}$ is randomly selected from $[-1,-0.5] \cup [0.5,1]$ if the original DREAM3 network has the regulation of the *n*-th gene from the *m*-th gene, and is otherwise set to 0.0; The parameter $h_{n,n}$ is set to 1.0 in order to simulate the auto-degradation, and the other $h_{n,m}$ $(n \neq m)$ values are set to 0.0. The parameters α_n and β_n are all set to 1.0. The inference problems used in this section had target networks with different model parameters, even when their topologies were the same.

As the observed gene expression patterns, each of the inference problems had 100 sets of time-series data, each covering all 100 genes. We obtained them by solving a set of differential equations (13) on the target model of the problem. The sets began from randomly generated initial values in [0.0, 2.0], and 11 observations with 0.4 time intervals between two adjacent observations were assigned to each gene in each set. We simulated measurement noise by adding 10% Gaussian noise to the computed time-series data. The purpose of each of the genetic network inference problems here is to infer a structure of the target network only from the generated gene expression data.

5.2 Experimental setup

As described previously, we used the BS-LPM inference method [9] to obtain multiple genetic networks. We constructed 100 genetic networks in this study (N_B = 100). For the parameters of the BS-LPM inference method, we used the recommended values; σ = 0.15, $C_1 = \frac{200}{N\sqrt{K}}$, $C_2 = 0.4C_1$ and δ = 0.05, where *N* is the number of genes contained in the target network and *K* is the number

of measurements. Thus, N = 100 and K = 100×11 = 1100 in the experiments described here.

As described earlier, we try to infer reasonable networks by selecting three kinds of a priori knowledge (N_{AP} = 3), i.e., the hierarchical structure, the common neighbors and the degree correlation. To use this knowledge, we must determine values for the parameters $\mathbf{w} = (w_1, w_2, w_3)$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$. We did so by applying the evolutionary algorithm, AGLSDC [16], into the adjustment problem of these parameters. According to the recommendation, we set the AGLSDC parameters to the following values; the population size, n_p, was 3×d, the number of the children generated per selection, n_c, was 10, and the parameters γ, $n^P_{\max}$ and $n^G_{\max}$ for the local search method were 0.3, 10 and 30, respectively, where d is the dimension of the search space and d = 3× N_{AP} = 9. As the parameters obtained might depend much on the performance of the optimization algorithm, we solved every adjustment problem 10 times by changing the seed for pseudo-random numbers used in AGLSDC.

This study searched for the parameters **w**, **s** and **t** instead of directly searching for the parameters **w**, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The search areas of s_i's and t_i's were all set to $[-0.1,1.1]$ in this study. In this study, on the other hand, we searched for the w_i values in the logarithmic space. According to the weight parameter of the inference method based on the hierarchical structure [8], we set the search areas of the $\log w_i$ values to $[-10 - \log N_B, -\log N_B]$. When trying to evaluate the confidence values according to the a priori knowledge, the proposed method uses multiple genetic networks inferred by the BS-LPM inference method. We know, however, that the inferred networks often contain erroneous regulations. It would be misguided to rely too heavily on the confidence values evaluated according to the a priori knowledge. Our solution, in this study, is to set the search areas for the $\log w_i$ values at levels that make the w_i values sufficiently small.

5.3 Determination of a parameter *C*

As described previously, we need to know the values for the constant parameters **w**, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in order to compute the modified confidence values, $p_{n,m}$'s, according to the equation (2). We obtain these values by optimizing the function (10). We can only succeed in finding the reasonable solution, however, by carefully selecting a value for the parameter C contained in the objective function. To set the parameter C to a reasonable value, we first divided our 50 inference problems into five groups, consisting of ten each of the Ecoli1, Ecoli2, Yeast1, Yeast2 and Yeast3 problems. We then obtained the parameters **w**, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by optimizing the function (10) with the inference problems with each of the groups excluded (N_p = 40). Finally, we tested the performance of the proposed method with the obtained parameters **w**, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ on the 10 problems from the excluded group. This study quantified the performance of the algorithm using the area under the recall-precision curve (AURPC).

We compared the proposed method with different values for the parameter C with the inference method based on the hierarchical structure. In Figure 3, the averaged improvements in AURPC of the proposed method upon the inference method based on the hierarchical structure on the training problems and test problems are plotted against the parameter C. As the figure shows, when the value for the parameter C is set between 0.001 and 0.01, the proposed method averagely outperforms the inference method based on the hierarchical structure even in inference problems with target networks never before seen. Moving forward from this section, we thus set the parameter C to 0.002. However, note that the optimum value for C could depend on the inference method used to generate multiple genetic networks, the a priori knowledge applied, and so on.

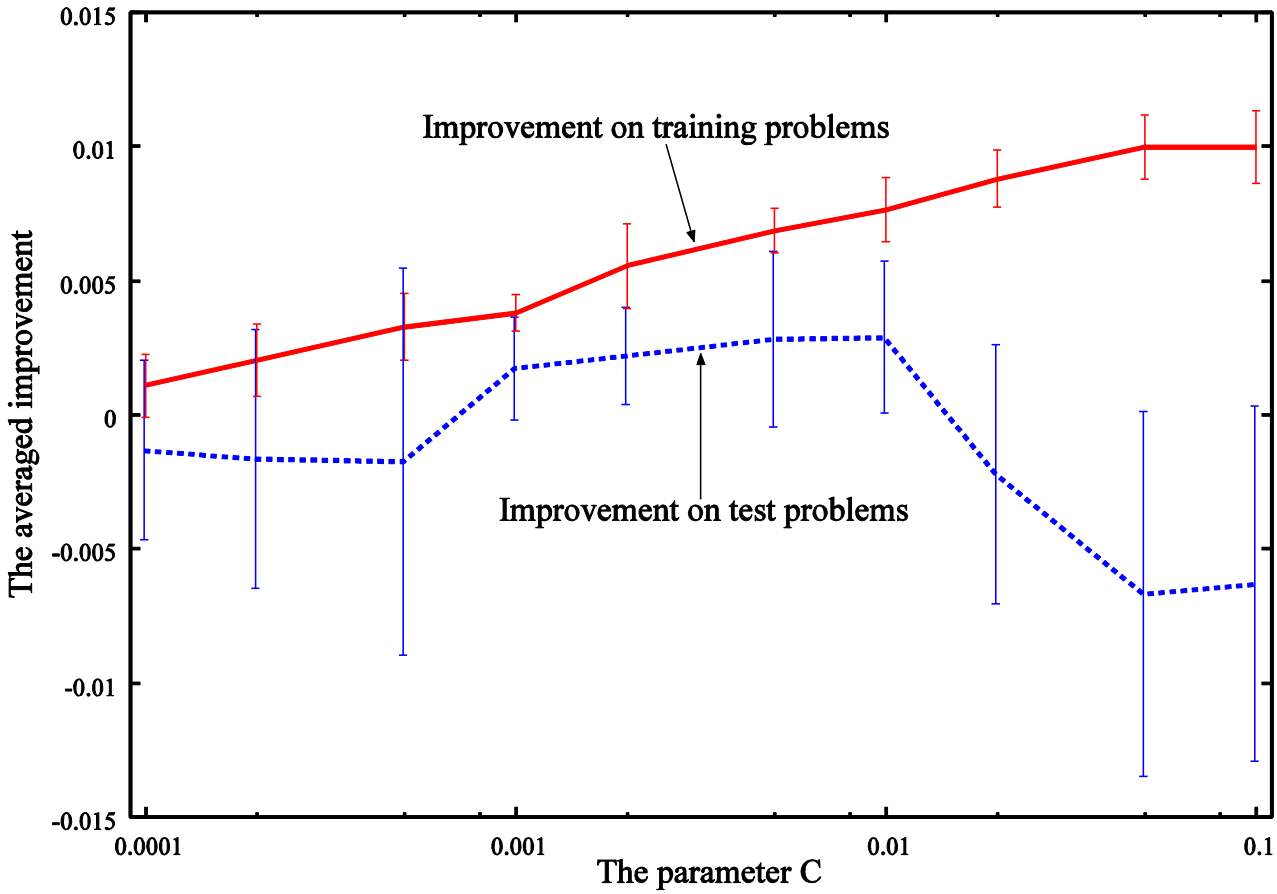

Figure 3. The averaged improvements in AURPC of the proposed method upon the inference method based on the hierarchical structure on the training problems (solid line) and the test problems (dotted line) plotted against the parameter *C*

Table 1. The performances of the proposed method with C = 0.002, the inference method based on the hierarchical structure, and the BS-LPM inference method on the problems of Ecoli1, Ecoli2, Yeast1, Yeast2 and Yeast3

AVG and STD represent the averaged AURPC and its standard deviation, respectively.

network	the proposed method with C = 0.002	the inference method based on the hierarchical structure [8]	the BS-LPM inference method [9]
	AVG± STD	AVG ± STD	AVG ± STD
Ecoli1	0.90678 ± 0.03861	0.90426 ± 0.04025	0.87920 ± 0.05277
Ecoli2	0.92462 ± 0.02689	0.92330 ± 0.02519	0.91541 ± 0.02075
Yeast1	0.71667 ± 0.05370	0.71404 ± 0.05153	0.68800 ± 0.04908
Yeast2	0.43718 ± 0.02694	0.43302 ± 0.02514	0.41509 ± 0.02068
Yeast3	0.35658 ± 0.04135	0.35630 ± 0.03858	0.35038 ± 0.03685

Table 1 lists the averaged AURPCs of the proposed method with C = 0.002, the inference method based on the hierarchical structure and the BS-LPM inference method on the problems designed based on Ecoli1, Ecoli2, Yeast1, Yeast2 and Yeast3. Note here that, when the hierarchical structure is the only a priori knowledge used and the constant parameters w_1,α_1 and β_1 are set to

$\frac{1}{N_B-1}$, 0 and 1, respectively, the proposed method is equivalent to the inference method based on the hierarchical structure. When, on the other hand, no a priori knowledge is used and the confidence values are computed only on the basis of the equation (1), the proposed method is equivalent to the BS-LPM inference method. Note also that the AURPCs of the proposed method, listed in the table, were evaluated on the test problems. When evaluating the performances of the proposed method, therefore, we set the parameters **w**, **α** and **β** to slightly different values in every trial. The experimental results indicate that the improvement brought by integrating multiple kinds of a priori knowledge is independent of the structure of the target network.

5.4 Determination of parameters w, α and β

Next, we obtained the parameters **w**, **α** and **β** by minimizing the function (10) with all of the 50 genetic network inference problems constructed in this section (N_p = 50). We optimized the objective function 10 times by changing the seed for pseudo-random numbers used in the optimization algorithm. The parameters obtained differed slightly in every trial. However, the shapes of the obtained non-linear functions (3) used to compute the confidence values according to the equation (2) were similar. Note here that the parameters **α** and **β** determine the shapes of the functions (3). Figure 4 shows the shapes of the functions (3) used to integrate the three kinds of a priori knowledge, i.e., the hierarchical structure, the common neighbors and the degree correlation. We see from the figure that the confidence values evaluated based on the hierarchical structure and degree correlation are reliable only when their values are sufficiently large. In contrast, the confidence values evaluated based on the common neighbors seem reliable as long as they are larger than 0.3. This means that a larger confidence value is assigned to the interaction between two genes when the genes are inferred to share more than 30% of the total genes as common neighbors. However, here, the required number of common neighbors seems too large, hence, the confidence values evaluated based on the common neighbors would be almost always 0.0. Figure 5 shows the estimated weight parameters **w**'s. The figure suggests that, as w_3 is much smaller than w_1 and w_2, the degree correlation is less reliable.

As shown in Figure 4, when the confidence values evaluated based on the a priori knowledge are small, the functions (3) transform their values to 0. This behavior indicates that the proposed method often disregards the a priori knowledge applied. In these experiments, the proposed method disregarded the hierarchical structure, the common neighbors and the degree correlation in the evaluations of 97.415%, 99.986% and 99.363% of the candidate regulations, respectively. Note that, while the inference method based on the hierarchical structure uses only the hierarchical structure, the proposed method uses all three kinds of a priori knowledge. As it turns out, however, the common neighbors and degree correlation are considered in the evaluations of only a few of the regulations. This may be one reason why the proposed method performed only slightly better than the inference method based on the hierarchical structure.

The computation time required for the optimization of the objective function (10) averaged about 5.09 h on a personal computer (Core i5-4670). To solve each of the genetic network inference problems, on the other hand, we must infer N_B (= 100) genetic networks using the BS-LPM inference method. The BS-LPM inference method took about 4.12 h to obtain these networks on the same computer. We next evaluated confidence values based on each kind of a priori knowledge. The computation times to evaluate confidence values using the hierarchical structure, the common neighbors and the degree correlation were 4.12 h, 0.44 s and 0.02 s, respectively. The computational cost required for solving each of the genetic network inference problems is therefore high. However, note that, once all of the genetic network inference problems are solved, the adjustment of the parameters **w**, **α** and **β** does not require that the problems be solved again. And once these

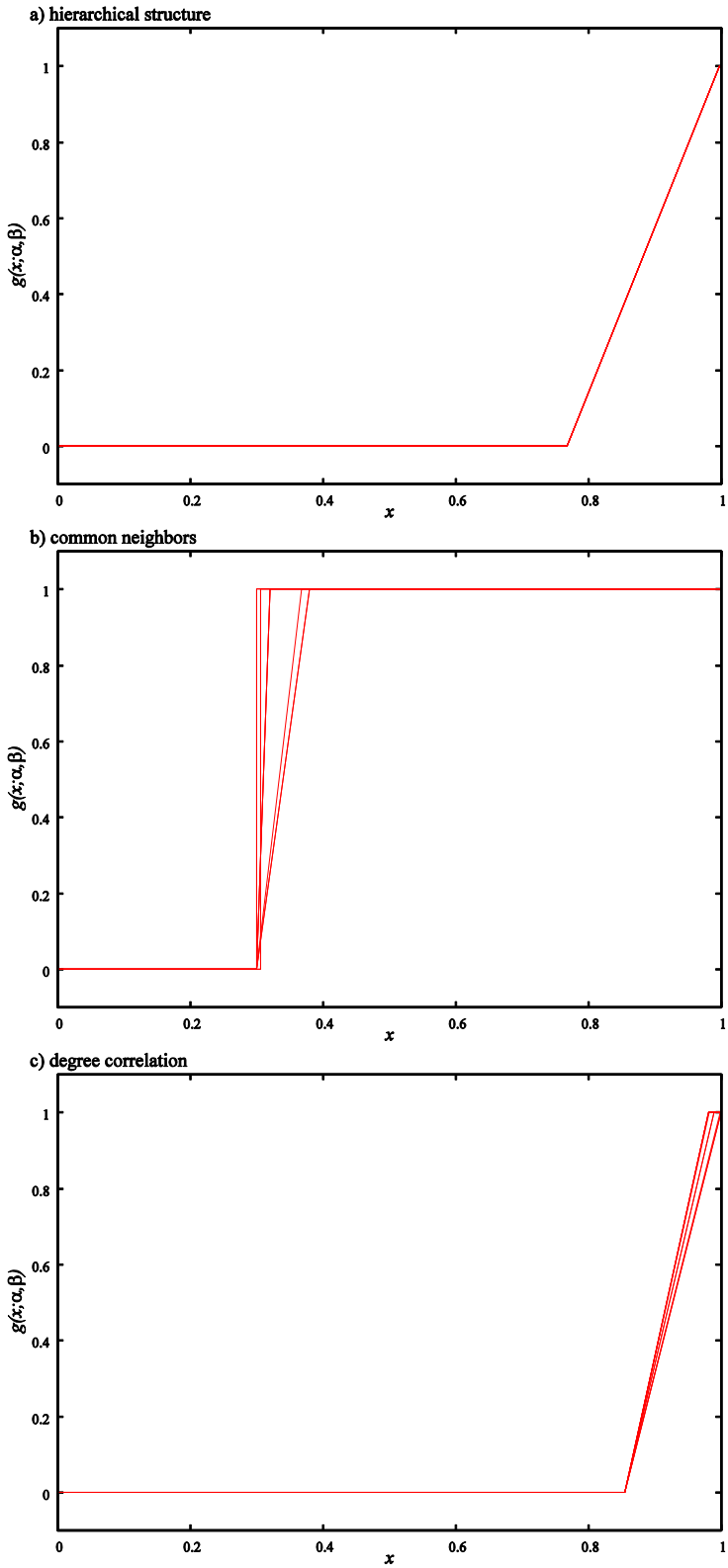

Figure 4. The shapes of the non-linear functions (3) used to introduce a) the hierarchical structure, b) the common neighbors and c) the degree correlation

The shapes determined by the estimated parameters α_i and β_i (i = 1, 2, 3). The results obtained from 10 trials are shown.

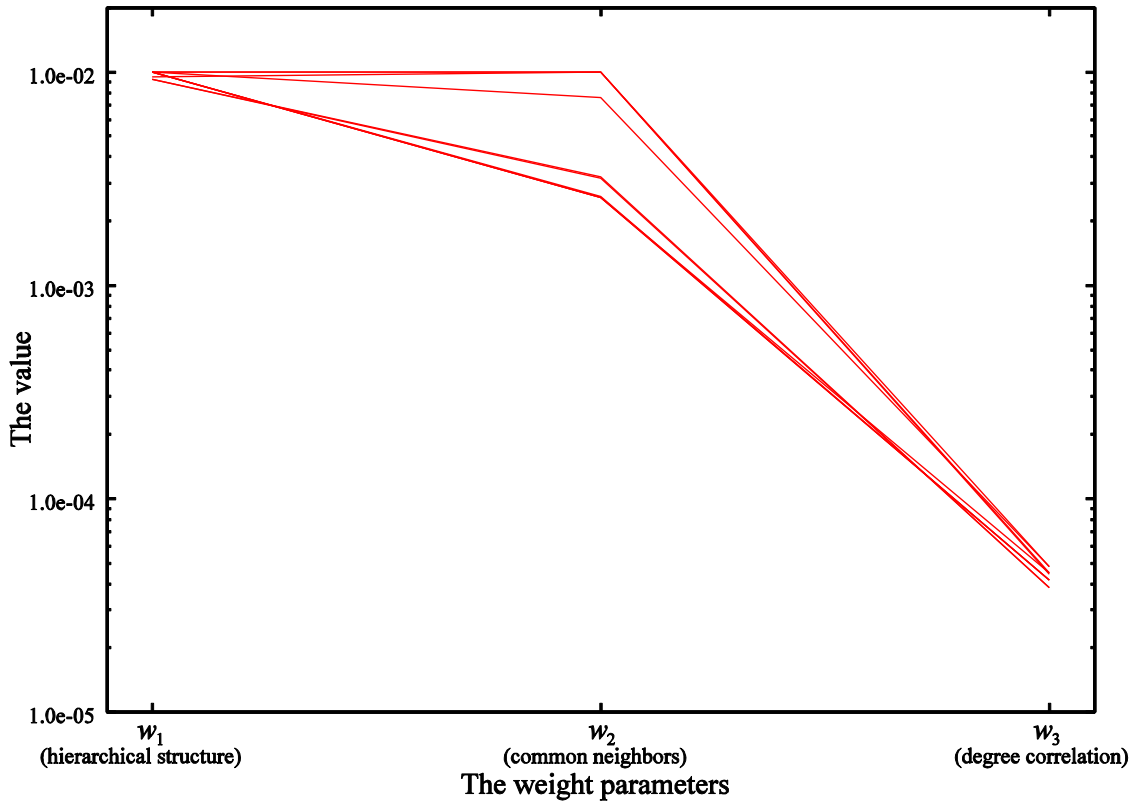

Figure 5. The estimated weight parameters w_1, w_2 and w_3
The results obtained from 10 trials are shown.

constant parameters are determined, the optimization of the function (10) is no longer required. It is currently difficult to measure a sufficient amount of gene expression data because of the measurement cost. Therefore, it would be important to extract useful information as possible from the limited amount of gene expression data. Note moreover that the proposed framework can be combined with a lot of the existing inference methods and would have an ability to improve their performances. Thus, although the improvement in AURPC accomplished by the proposed method is small and its computational cost is not negligible, these drawbacks would not hinder the application of our method.

6. Analysis of Actual Data

In this section, we analyze actual data using the proposed method with the constant parameters obtained in the previous section.

6.1 Experimental setup

In this experiment, we analyzed an ErbB-receptor-mediated regulatory network of transcription factors in normal human epidermal keratinocytes. The network consisted of 29 components, i.e., 3 receptors (EGFR, ErbB2 and ErbB3), 7 signal transducer proteins (ERK, PI3K, AKT, STAT3, PLCg, PKCd and c-SRC), the phosphorylated forms of the 3 receptors and 7 signal transducer proteins (pEGFR(pY845), pEGFR(pY1068), pErbB2, pErbB3, pERK, pPI3K, pAKT, pSTAT3 (pY705), pSTAT3(pS727), pPLCg, pPKCd and pc-SRC), and 7 transcription factors (c-FOS, FRA1,

Table 2. The constant parameters **w**, **α** and **β** used to analyze the actual data

I		w_i	α_i	β_i
1	(hierarchical structure)	9.7777×10^{-3}	7.6798×10^{-1}	9.9721×10^{-1}
2	(common neighbors)	5.4284×10^{-3}	3.0076×10^{-1}	3.2982×10^{-1}
3	(degree correlation)	4.3057×10^{-5}	8.5470×10^{-1}	9.8802×10^{-1}

FRA2, JUNB, c-JUN, JUND and c-MYC). Time-series data consisting of 14 time-points of the 29 components were measured by Saeki and colleagues [19]. The observed data were not always sufficient. When we tried to infer multiple networks using the BS-LPM inference method, we therefore used the following a priori knowledge: i) none of the receptors or signaling proteins are affected by other receptors or signaling proteins; ii) none of the transcription factors are affected by receptors, signaling proteins, or phosphorylated forms of receptors; iii) none of the phosphorylated receptors or phosphorylated signaling proteins are affected by other receptors, signaling proteins, or transcription factors; iv) every component of this system regulates itself; v) every protein regulates its own phosphorylated form. The design of the above knowledge stems from the biological knowledge that the phosphorylated forms of signaling proteins and receptors can form cascades to transduce extracellular signals to transcription factors [20]. We introduced this knowledge into the BS-LPM inference method using the technique proposed in reference [21]. Thus, we introduced the knowledge above on the basis of the framework for "using a priori knowledge while inferring genetic networks." Note that, as described previously, the new framework for "using a priori knowledge after genetic network inference" can be combined with the framework for "using a priori knowledge while inferring genetic networks." In total, 100 networks were inferred by the BS-LPM inference method (N_B = 100).

We used the hierarchical structure, the common neighbors and the degree correlation to evaluate the confidence values of the regulations from the inferred networks. The computed confidence values were then integrated using the equation (2). When computing the modified confidence values $p_{n,m}$'s, we set the constant parameters **w**, **α** and **β** to the average parameter values obtained in the section 5.4 (listed in Table 2).

The other experimental conditions were unchanged from those used in the previous experiment.

6.2 Results

The network of regulations with confidence values exceeding 0.25 is shown in Figure 6. The networks obtained contained 132 regulations. Yet 17 of these regulations seemed to be trivial, as they are regulations of the proteins from their phosphorylated forms or vice versa. Note that this study tried to infer a network consisting of both proteins and their phosphorylated forms. The detailed regulatory relations, however, are still unclear. This study therefore compared the inferred network with a network of protein-protein interactions. Figure 7 shows a network of protein-protein interactions obtained from the STRING database [22, 23]. The comparison results indicate that 74 of the 132 inferred regulations were reasonable, as the interactions between the corresponding proteins have been reportedly confirmed. The number of the reasonable regulations inferred by the proposed method was smaller than that of the inference method based on the hierarchical structure. The inference method based on the hierarchical structure reportedly found 77 reasonable regulations [8]. We must note here that we checked only the regulations with confidence values exceeding 0.25. As these methods often assigns different confidence values even to the same regulation, the numbers of the regulations with confidence values exceeding 0.25 were different. As a result, the numbers of the reasonable regulations were also different.

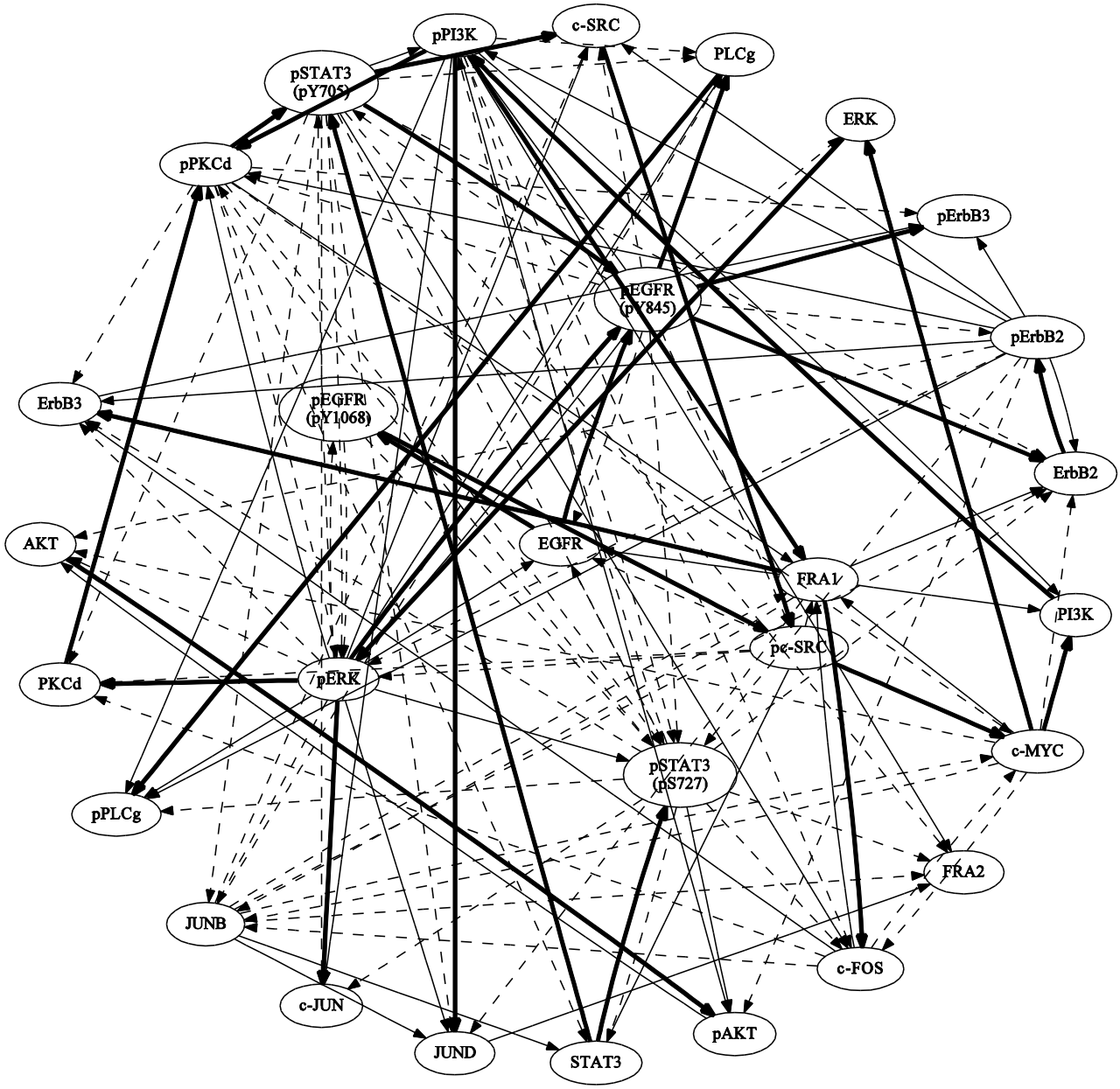

Figure 6. The network of regulations with confidence values exceeding 0.25

Bold, solid and dotted lines represent regulations with confidence values exceeding 0.75, 0.5 and 0.25, respectively. Note that the proposed method constructs a network as a directed graph.

As described previously, in the framework for "using a priori knowledge after genetic network inference," the method tries to improve the genetic networks using only the structural properties of the networks that have been already inferred by the other inference method applied. Therefore, the performance of the method depends much on that of the applied inference method. As a consequence, the results obtained from the proposed method, the inference method based on the hierarchical structure and the BS-LPM inference method were quite similar. Table 3 shows the top 20 regulations with respect to the confidence values assigned by the proposed method, the inference

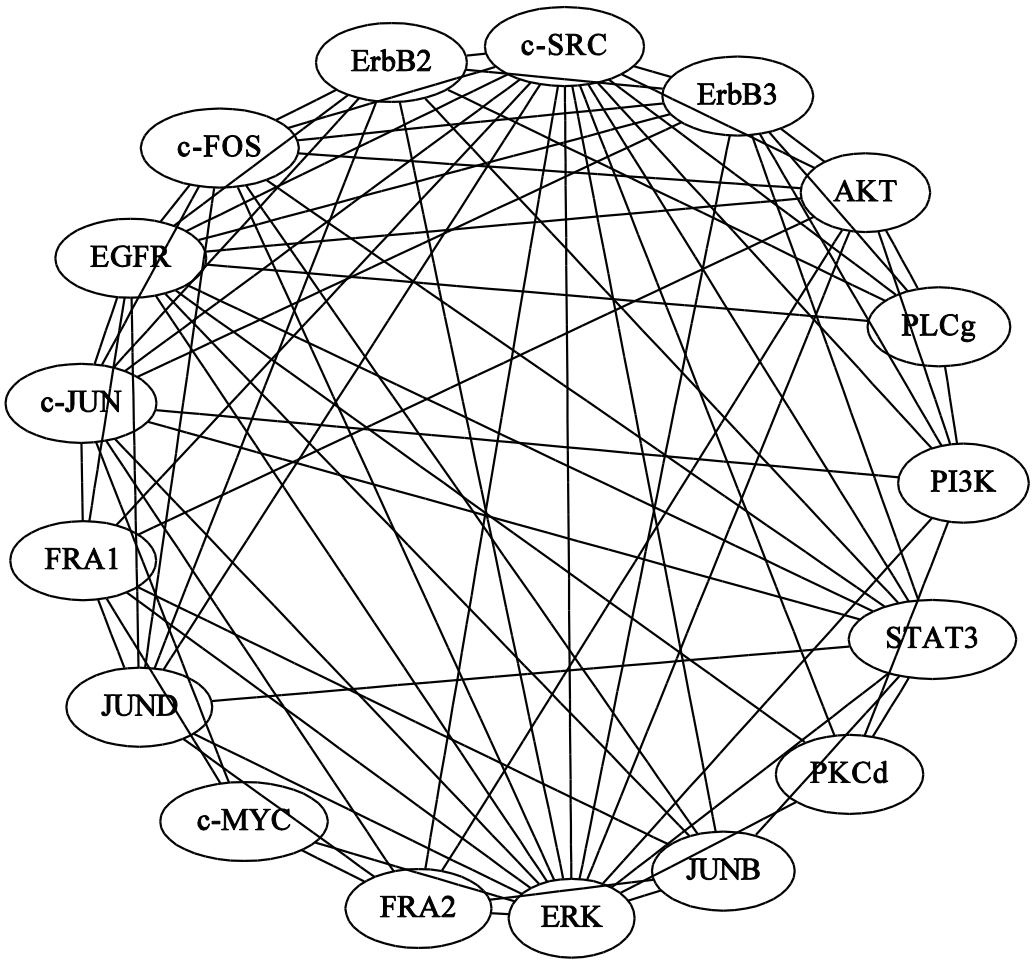

Figure 7. The network of protein-protein interactions obtained from the STRING database [22, 23]
The network obtained from the STRING database is represents as an undirected graph.

method based on the hierarchical structure and the BS-LPM inference method. While the confidence values of the regulations inferred by the BS-LPM inference method were often the same, those of the regulations inferred by the inference method based on the hierarchical structure mostly differed. The confidence values evaluated by the proposed method were the same in an intermediate number of regulations.

As described previously, we detected the hierarchical structure in the target network using the hierarchical random graph model. In order to obtain a hierarchical random graph model consistent with the inferred networks, we used the stochastic search algorithm. When we extracted the hierarchical structure in the target network 10 times by changing the seed for pseudo-random numbers used in the search algorithm, we obtained three different hierarchical random graph models. When we applied these models to the proposed method and the inference method based on the hierarchical structure, we obtained different rankings of the regulations with respect to the confidence value. The proposed method however depended little on the hierarchical random graph model applied. Even when we used different models, the proposed method obtained the same confidence value rankings for the regulations shown in Figure 6. When we applied these models to

Table 3. The top 20 regulations with respect to the confidence values assigned by the proposed method, the inference method based on the hierarchical structure and the BS-LPM inference method

the proposed method		the inference method based on the hierarchical structure		the BS-LPM inference method	
rank	regulation	rank	regulation	rank	
1	EGFR→pEGFR(pY1068)	1	EGFR→pEGFR(pY1068)	1	pPI3K → FRA1
1	AKT → pAKT	1	AKT → pAKT	1	EGFR→pEGFR(pY1068)
3	pPI3K → FRA1	3	ErbB2 → pErbB2	1	ErbB2 → pErbB2
3	ErbB2 → pErbB2	4	pPI3K → FRA1	1	AKT → pAKT
5	PLCg → pPLCg	5	PLCg → pPLCg	5	EGFR→ pEGFR(pY845)
6	EGFR→ pEGFR(pY845)	6	EGFR→ pEGFR(pY845)	5	STAT3→pSTAT3(pY705)
6	STAT3→pSTAT3(pY705)	7	STAT3→pSTAT3(pY705)	5	STAT3→pSTAT3(pS727)
6	STAT3→pSTAT3(pS727)	8	STAT3→pSTAT3(pS727)	5	PLCg → pPLCg
9	ERK → pERK	9	PI3K → pPI3K	9	ERK → pERK
10	PI3K → pPI3K	10	ERK → pERK	9	PI3K → pPI3K
11	PKCd → pPKCd	11	PKCd → pPKCd	11	PKCd → pPKCd
12	c-SRC → pc-SRC	12	c-SRC → pc-SRC	12	c-SRC → pc-SRC
13	pSTAT3(pY705) → pEGFR(pY845)	13	pSTAT3(pY705) → pEGFR(pY845)	13	pSTAT3(pY705) → pEGFR(pY845)
14	pPI3K → pPKCd	14	pPI3K → pPKCd	14	pPI3K → pPKCd
15	pEGFR(pY845) → PLCg	15	pEGFR(pY845) → PLCg	15	pEGFR(pY845) → PLCg
16	pPI3K → JUND	16	pPI3K → JUND	16	pPI3K → JUND
17	c-MYC → ERK	17	c-MYC → ERK	17	c-MYC → ERK
18	pERK → pEGFR(pY845)	18	pERK → pEGFR(pY845)	18	pc-SRC → c-MYC
19	pc-SRC → c-MYC	19	pEGFR(pY1068)→pc-SRC	18	pERK → pEGFR(pY845)
19	pEGFR(pY1068)→pc-SRC	20	pc-SRC → c-MYC	18	pEGFR(pY1068)→pc-SRC

the inference method based on the hierarchical structure, on the other hand, 29.4% of the regulations with confidence values exceeding 0.25 were assigned to different rank orders on average. The experimental results indicate that the proposed method not only improves the quality of the inferred network, but also reduces the effect of the randomness in the search algorithm on the rankings of the regulations. Reliable and robust rankings could potentially enable biologists to experimentally validate inferred regulations with less effort.

7. Conclusion

In this study, we established a new framework for "using a priori knowledge after genetic network inference." Based on this framework, we proposed a method that utilizes multiple kinds of a priori knowledge for the inference of genetic networks. The proposed method computes confidence values of all of the candidate regulations by combining multiple kinds of the a priori knowledge. To obtain reasonable networks, the proposed method searches for an effective way to combine the knowledge. Through the numerical experiments, we confirmed that our method improves the confidence values of the regulations and reduces the effect of the randomness in the applied stochastic algorithm on their rankings. As an example, this study used three kinds of a priori knowledge, i.e., the hierarchical structure, the common neighbors and the degree correlation. While the improvement obtained through the use of these three kinds of knowledge was small, our method

can still use other kinds of a priori knowledge that have never been used for genetic network inference. To obtain multiple genetic networks, meanwhile, this study used the BS-LPM inference method. Our proposed method, however, can use any other method capable of producing multiple genetic networks. In addition, the proposed method can be also combined with the inference methods that do not infer multiple genetic networks but just assign confidence values to all of the regulations (see, e.g., the reference [24]). Various combinations of our method with others can be expected to improve the performance of the existing inference method. This study combined three kinds of the a priori knowledge using the equation (2). However, technique developed in the field of statistics might multiple kinds of knowledge more effectively. In our future work, thus, we would like to seek a more effective combination technique.

Acknowledgements

This work has been supported by JSPS KAKENHI Grant Number 26330275.

References

[1] Larrañaga, P.; Calvo, B.; Santana, R.; Bielza, C.; Galdiano, J.; *et al*. Machine Learning in Bioinformatics. *Briefings in Bioinformatics* **2006**, *7*, 86-112.
[2] Chou, I. C.; Voit, E. O. Recent Development in Parameter Estimation and Structure Identification of Biochemical and Genomic Systems. *Math. Biosci.* **2009**, *219*, 57-83.
[3] Hecker, M.; Lambeck, S.; Toepfer, S.; van Someren, E.; Guthke, R. Gene Regulatory Network Inference: Data Integration in Dynamic Models -- a Review. *BioSystems* **2009**, *96*, 86-103.
[4] Kikuchi, S.; Tominaga, D.; Arita, M.; Takahashi, K.; Tomita, M. Dynamic Modeling of Genetic Networks using Genetic Algorithm and S-system. *Bioinformatics* **2003**, *19*, 643-650.
[5] Kimura, S.; Ide, K.; Kashihara, A.; Kano, M.; Hatakeyama, M.; *et al*. Inference of S-system Models of Genetic Networks using a Cooperative Coevolutionary Algorithm. *Bioinformatics* **2005**, *21*, 1154-1163.
[6] Yeung, M. K. S.; Tegnér, J.; Collins, J. J. Reverse Engineering Gene Networks using Singular Value Decomposition and Robust Regression. *Proc. Natl. Acad. Sci. USA.* **2002**, *99*, 6163-6168.
[7] Tominaga, D.; Horton, P. Inference of Scale-free Networks from Gene Expression Time Series. *J. Bioinform. Comput. Biol.* **2006**, *4*, 503-514.
[8] Kimura, S.; Tokuhisa, M.; Okada-Hatakeyama, M. Genetic Network Inference using Hierarchical Structure. *Frontiers in Physiology* **2016**, *7*, 57.
[9] Kimura, S.; Shiraishi, Y.; Okada, M. Inference of Genetic Networks using LPMs: Assessment of Confidence Values of Regulations. *J. Bioinform. Comput. Biol.* **2010**, *8*, 661-677.
[10] Kimura, S.; Nakayama, S.; Hatakeyama, M. Genetic Network Inference as a Series of Discrimination Tasks, *Bioinformatics* **2009**, *25*, 918-925.
[11] Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist.* **1979**, *7*, 1-26.
[12] Clauset, A.; Moore, C.; Newman, M. E. J. Hierarchical Structure and the Prediction of Missing Links in Networks. *Nature* **2008**, *453*, 98-101.
[13] Tong, A. H. Y.; Lesage, G.; Bader, G. D.; Ding, H.; Xu, H.; *et al*. Global Mapping of the Yeast Genetic Interaction Network. *Science* **2004**, *303*, 808-813.
[14] Newman, M. E .J. Clustering and Preferential Attachment in Growing Networks. *Physical Review E* **2001**, *64*, 025102.

[15] Jeong, H.; Tombor, B.; Albert, R.; Oltvai, Z. N.; Barabási, A. L. The Large-scale Organization of Metabolic Networks. *Nature* **2000**, *407*, 651-654.
[16] Kimura, S.; Nakakuki, T.; Kirita, S.; Okada, M. AGLSDC: a Genetic Local Search Suitable for Parallel Computation. *SICE J. of Control, Measurement, and System Integration* **2011**, *4*, 105-113.
[17] Dream project. (http://dreamchallenges.org/).
[18] Voit, E. O. *Computational Analysis of Biochemical Systems*. Cambridge University Press; Cambridge, 2000.
[19] Saeki, Y.; Nagashima, T.; Kimura, S.; Okada-Hatakeyama, M. An ErbB Receptor-mediated AP-1 Regulatory Network is Modulated by STAT3 and c-MYC during Calcium-dependent Keratinocyte Differentiation. *Exp. Dermatol.* **2012**, *21*, 293-298.
[20] Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; *et al*. *Molecular Biology of the Cell*, 5th Edition. Garland Science; New York, 2008.
[21] Kimura, S.; Shiraishi, Y.; Hatakeyama, M. *Inference of Genetic Networks using Linear Programming Machines: Application of A Priori Knowledge*. Proc. of the 2009 Int. Joint Conf. on Neural Networks; Atlanta, June 14-19, 2009, 1617-1624.
[22] Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; *et al*. STRING v10: Protein-protein Interaction Networks, Integrated over the Tree of Life. *Nucleic Acids Res.***2015**, *43*, D447-D452.
[23] STRING database. (http://string-db.org/).
[24] Huynh-Thu, V. A.; Irrthum, A.; Wehenkel, L.; Geurts, P. Inferring Regulatory Networks from Expression Data Using Tree-based Methods. *PLoS One* **2010**, *5*, e12776.