Chem-Bio Informatics Journal, Vol.22, pp.88-109 (2022)

Inference of genetic networks using random forests: Performance improvement using a new variable importance measure

Shuhei Kimura[1*], Yahiro Takeda[2], Masato Tokuhisa[1], and Mariko Okada[3]

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

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

(Received July 05, 2022; accepted December 11, 2022; published online December 29, 2022)

Abstract

Among the various methods so far proposed for genetic network inference, this study focuses on the random-forest-based methods. Confidence values are assigned to all of the candidate regulations when taking the random-forest-based approach. To our knowledge, all of the random-forest-based methods make the assignments using the standard variable importance measure defined in tree-based machine learning techniques. Therefore, the sum of the confidence values of the candidate regulations of a certain gene from the other genes, that are computed from a single random forest, is always restricted to a value of almost 1. We think that this feature is inconvenient for the genetic network inference that requires to compare the confidence values computed from multiple random forests. In this study we therefore propose an alternative measure, what we call ``the random-input variable importance measure,'' and design a new inference method that uses the proposed measure in place of the standard measure in the existing random-forest-based inference method. We show, through numerical experiments, that the use of the random-input variable importance measure improves the performance of the existing random-forest-based inference method by as much as 45.5% with respect to the area under the recall-precision curve (AURPC).

Key Words: Genetic network inference; Random forest; Variable importance measure; Random-input variable importance measure

Area of Interest: AI technologies and their applications

Licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0)

1. Introduction

Advancements in high-throughput measurement technologies, such as DNA microarrays and RNA-seq using next-generation sequencers, have led to a huge amount of gene expression data. There is a great value in developing computational methods capable of extracting useful information from these data. The inference of genetic networks is conceived as one promising means for this purpose. In a genetic network inference problem, mutual regulations among genes are inferred from the measured gene expression data. Knowledge of the regulatory structure of the genetic network enables us to understand biological mechanisms.

Various models, such as the boolean network model, the Bayesian network model, the Petri net model, the differential-equations-based model, and so on, have been used to describe genetic networks, and the inference methods based on individual models have been proposed [2, 6, 7, 14, 17, 23, 25, 35, 38, 40]. With recent advances in machine learning techniques, several researchers have become interested in the inference methods based on the machine learning techniques [1, 22, 31]. In this study, we narrow the focus to genetic network inference that uses the random forest, a kind of machine learning technique [4]. Huynh-Thu and colleagues first introduced the random forest into the field of the genetic network inference [15]. Their method, GENIE3, won the DREAM4 *in silico* multifactorial challenge (https://dreamchallenges.org/). Researchers, inclusive of the author and colleagues, have been attracted by the possibilities of GENIE3, and have therefore developed its extensions [16, 19, 20, 27, 29].

The random-forest-based inference methods train multiple random forests, each of which corresponds to each gene. Based on these random forests, the methods assign confidence values to all of the candidate regulations. Specifically, by analyzing the n-th random forest corresponding to the n-th gene, the inference methods compute the confidence values of the regulations of the n-th gene from the other genes. The random-forest-based methods then rank all of the candidate regulations according to their confidence values. To obtain the confidence values, all of the existing random-forest-based inference methods use the standard variable importance measure defined in tree-based machine learning techniques. In this measure, the sum of the confidence values of the regulations of a certain gene from the other genes, that are computed by one of the random forests trained, is restricted to a value of almost 1. Note that this condition is always satisfied regardless of the number of regulating genes. We thus believe that this restriction makes the importance scores relative, rather than absolute, within each random forest and is therefore inadequate for comparing the confidence values obtained from multiple random forests.

As an alternative to the standard variable importance measure, this study proposes what we call the ``random-input variable importance measure,'' a new measure that is free from the abovementioned restriction. The removal of the restriction could make the importance scores absolute and the proposed measure is therefore suitable for comparing the confidence values obtained from multiple random forests. In this study, we thus use the proposed random-input variable importance measure in lieu of the standard one to compute the confidence values of all of the candidate regulations in the existing random-forest-based inference method. Last, we perform numerical experiments with artificial and real genetic network inference problems to confirm that the proposed measure can be effectively applied in the random-forest-based inference method. In this study, we use our measure only for the inference of genetic networks. However, we think that the proposed measure is capable of extracting more reliable information in other applications of random forests.

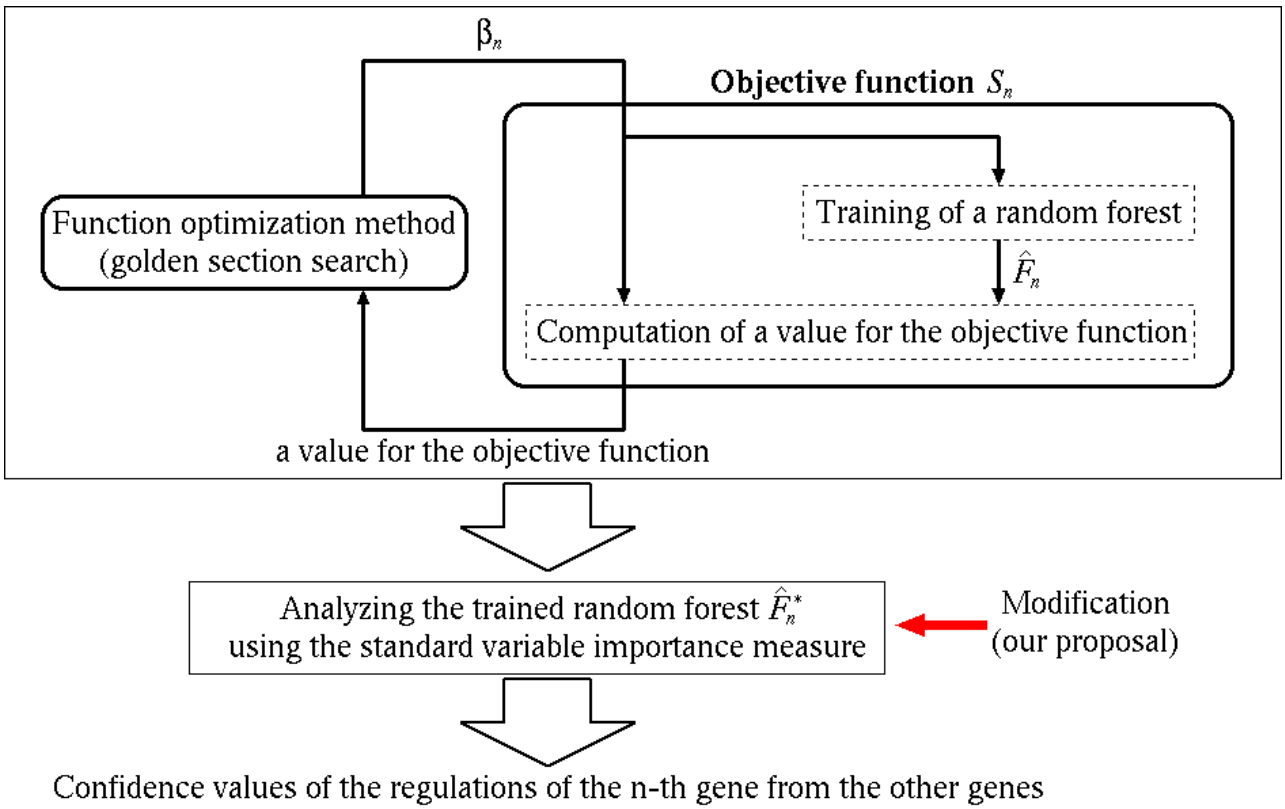

Figure 1. The framework of the random-forest-based inference method [19] and our proposal

2. Random-forest-based inference method

This study applies a new variable importance measure to the existing random-forest-based inference method. While any random-forest-based inference method could be used for this purpose, here we use an inference method we have proposed in an earlier paper [19]. We will begin, in this section, by explaining the method. Readers can find more detailed information on this method in our earlier paper [19] (see also Fig. 1).

2.1 Model for describing genetic networks

The method we use in this study represents a genetic network as a set of differential equations of the form

$$\frac{dX_n}{dt} = F_n\left(\mathbf{X}_{-n}\right) - \beta_n X_n, \quad (n = 1, 2, \cdots, N), \tag{1}$$

where $\mathbf{X}_{-n} = \left(X_1, \cdots, X_{n-1}, X_{n+1}, \cdots, X_N\right)$, X_m is the expression level of the m-th gene, N is the number of genes contained in the target network, β_n (> 0) is a constant parameter, and F_n is a function of arbitrary form.

The inference of a genetic network is achieved by obtaining a function F_n and a parameter β_n ($n = 1, 2, \cdots, N$) that will produce time-courses consistent with the observed gene expression levels. The inference method we are using obtains them in the manner described below.

2.2 Obtaining F_n and β_n

The inference method [19] divides an inference problem of a genetic network consisting of N genes into N subproblems, each of which corresponds to one gene. By solving the n-th subproblem, the method obtains a reasonable approximation of the function F_n and a reasonable value for the parameter β_n. The remainder of this section will describe a method for obtaining F_n and β_n.

The method used in this study obtains an approximation of the function F_n and a value for the parameter β_n through the optimization of the one-dimensional function,

$$S_n(\beta_n) = \sum_{k=1}^{K_T} \frac{w_k^T}{\beta_n} \left[\left. \frac{dX_n}{dt} \right|_{t_k} - \hat{F}_n\left(\mathbf{X}_{-n}\big|_{t_k} ; \beta_n \right) + \beta_n X_n\big|_{t_k} \right]^2 + \sum_{k=1}^{K_S} \frac{w_k^S}{\beta_n} \left[\left. \frac{dX_n}{dt} \right|_{s_k} - \hat{F}_n\left(\mathbf{X}_{-n}\big|_{s_k} ; \beta_n \right) + \beta_n X_n\big|_{s_k} \right]^2, \quad (2)$$

where $\mathbf{X}_{-n}\big|_{t_k} = \left(X_1\big|_{t_k}, \cdots, X_{n-1}\big|_{t_k}, X_{n+1}\big|_{t_k}, \cdots, X_N\big|_{t_k} \right)$, $\mathbf{X}_{-n}\big|_{s_k} = \left(X_1\big|_{s_k}, \cdots, X_{n-1}\big|_{s_k}, X_{n+1}\big|_{s_k}, \cdots, X_N\big|_{s_k} \right)$, and $X_m\big|_{t_k}$ and $X_m\big|_{s_k}$ are the expression levels of the m-th gene at the k-th measurement of time-series and steady-state experiments, respectively. K_T (≥ 2) and K_S (≥ 0) are the numbers of measurements performed in the time-series and steady-state experiments, respectively. Note that, in the genetic network inference, $X_m\big|_{t_k}$'s and $X_m\big|_{s_k}$'s are measured using biochemical techniques. $\left. \frac{dX_n}{dt} \right|_{t_k}$ and $\left. \frac{dX_n}{dt} \right|_{s_k}$ are the time derivatives of the expression level of the n-th gene at the k-th measurement of the time-series and steady-state experiments, respectively. The time derivatives in the time-series experiments, $\left. \frac{dX_n}{dt} \right|_{t_k}$'s, are directly estimated from the measured time-series of the gene expression levels using some smoothing technique, while the time derivatives in the steady-state experiments, $\left. \frac{dX_n}{dt} \right|_{s_k}$'s, are all set to zero. w_k^T and w_k^S are weight parameters for the k-th measurements in the time-series and steady-state experiments, respectively. Our earlier paper [19] showed that our random-forest-based inference method performs better when the constant parameters w_k^T's and w_k^S's are appropriately set. In order to determine these values, then, the methods that utilize the similarity between measurements have been proposed [19, 21].

$\hat{F}_n(\cdot\,; \beta_n)$ is an approximation of the function F_n trained under the given β_n. The computation of the objective function (2) requires an approximation of the function F_n, i.e., $\hat{F}_n$. The inference method [19] obtains an approximation of the function F_n using the random forest [4]. The random forest that approximates the function F_n is trained on the basis of the training data consisting of the following set of input-output pairs,

$$\left\{ \left(\mathbf{X}_{-n}\big|_{t_k}, \left. \frac{dX_n}{dt} \right|_{t_k} + \beta_n X_n\big|_{t_k} \right) \middle| k = 1, 2, \cdots, K_T \right\} \cup \left\{ \left(\mathbf{X}_{-n}\big|_{s_k}, \left. \frac{dX_n}{dt} \right|_{s_k} + \beta_n X_n\big|_{s_k} \right) \middle| k = 1, 2, \cdots, K_S \right\}.$$

Note that, when trying to compute a value for the objective function (2), a value for the parameter

β_n is always given. With this value given, we can train the random forest using the training data described above. Note also that, in order to keep consistency with the objective function (2), the random forest used in the method [19] tries to obtain an approximation of the function F_n that minimizes a weighted sum of the squared errors between the given output values and the output values computed from the model.

A reasonable approximation of the function F_n and a reasonable value for the parameter β_n are obtained through the optimization of the objective function (2). The random-forest-based inference method [19] uses the golden section search [30] to minimize the function (2).

3. Assigning confidence values to regulations

By analyzing the random forests that have been trained, the random-forest-based inference methods assign confidence values to all of the candidate regulations. The inference methods then rank all of the candidate regulations according to their confidence values. The methods obtain the confidence values of the regulations of the *n*-th gene from the other genes by analyzing the *n*-th random forest that approximates the function F_n. Here, the approximation of the function F_n and the value for the parameter β_n obtained through the optimization of function (2) are represented as $\hat{F}_n^*$ and β_n^*, respectively.

The random-forest-based inference methods compute the confidence values using the variable importance measure. By using the variable importance measure, tree-based machine learning techniques such as the random forest, Extra-Trees [11], VR-Trees [24], and so on compute importance scores for all of the input variables. The importance score of a certain input variable represents the degree to which the variable contributes to the prediction of the output values.

To our knowledge, all of the existing random-forest-based inference methods use the standard variable importance measure to compute the confidence values of the candidate regulations. In this section, therefore, we begin by describing the standard method of using the standard variable importance measure to compute the confidence values. We then propose a method that uses a new measure (see also Fig. 1).

3.1 Standard variable importance measure

As mentioned just above, the random-forest-based inference methods use the standard variable importance measure, that is defined only in tree-based machine learning techniques, to compute the confidence values of the candidate regulations. The random forest used by these inference methods consists of multiple regression trees. Each regression tree is trained by dividing the given training dataset into two subsets, each of which is then divided into two sub-subsets, and so forth, until stopping criteria are satisfied. When dividing a dataset into two subsets, the algorithm selects the input variable and the split point from the candidates so that the sum of the variances of the output values in the two subsets is minimized. In the standard variable importance measure, the importance score of a certain input variable represents the mean difference between the variance of the output values in the dataset that is divided by the input variable into two subsets and the sum of the variances of the output values in the subsets. When we use the standard variable importance measure, thus, we can compute the confidence value of the regulation of the *n*-th gene from the *m*-th gene, $C_{n,m}^S$, by

$$C_{n,m}^{S} = \frac{1}{Sq_{w0}} \frac{1}{N_{tree}} \sum_{i=1}^{N_{tree}} \sum_{v \in V_i(m)} I(v), \tag{3}$$

where

$$Sq_{w0} = \sum_{k=1}^{K_T} w_k^T \left(y_{t_k} - \overline{y}_{w0} \right)^2 + \sum_{k=1}^{K_S} w_k^S \left(y_{s_k} - \overline{y}_{w0} \right)^2, \tag{4}$$

$$\overline{y}_{w0} = \frac{1}{N_{w0}} \left[\sum_{k=1}^{K_T} w_k^T y_{t_k} + \sum_{k=1}^{K_S} w_k^S y_{s_k} \right], \tag{5}$$

$$N_{w0} = \sum_{k=1}^{K_T} w_k^T + \sum_{k=1}^{K_S} w_k^S, \tag{6}$$

$$y_{t_k} = \left. \frac{dX_n}{dt} \right|_{t_k} + \beta_n^* X_n \big|_{t_k}, \tag{7}$$

$$y_{s_k} = \left. \frac{dX_n}{dt} \right|_{t_k} + \beta_n^* X_n \big|_{s_k}, \tag{8}$$

$$I(v) = N_w(v) Sq_w(v) - N_w(v_L) Sq_w(v_L) - N_w(v_R) Sq_w(v_R), \tag{9}$$

$$Sq_w(v) = \sum_{k \in T(v)} w_k^T \left[y_{t_k} - \overline{y}_w(v) \right]^2 + \sum_{k \in S(v)} w_k^S \left[y_{s_k} - \overline{y}_w(v) \right]^2, \tag{10}$$

$$\overline{y}_w(v) = \frac{1}{N_w(v)} \left[\sum_{k \in T(v)} w_k^T y_{t_k} + \sum_{k \in S(v)} w_k^S y_{s_k} \right], \tag{11}$$

$$N_w(v) = \sum_{k \in T(v)} w_k^T + \sum_{k \in S(v)} w_k^S, \tag{12}$$

N_{tree} is the number of trees in the random forest $\hat{F}_n^*$, and $V_i(m)$ is a set of nodes that use the expression levels of the *m*-th gene to split the training examples in the *i*-th decision tree of $\hat{F}_n^*$. v_L and v_R are the left and right child nodes of the node v, respectively. T(v) and S(v) are sets of indices of the training examples generated from time-series and static gene expression data, respectively, and are allocated to the node v. Note here that the inference method mentioned in the previous section needs to set values for the weight parameters, w_k^T's and w_k^S's. When computing the confidence values, therefore, the method also considers these values, as described above.

When we use the standard variable importance measure, the sum of the confidence values of the candidate regulations of a certain gene from the other genes, that are computed from one of the random forests trained, is always restricted to a value of almost 1. The restriction could make the importance scores relative, rather than absolute. When we try to compare the confidence values computed from a single random forest, this feature of the standard variable importance measure will not hinder our investigation. We must note however that the random-forest-based inference methods must rank all of the regulations with respect to the confidence values computed from the multiple random forests. The use of the standard variable importance measure might therefore degrade the performance of the random-forest-based inference methods.

3.2 Random-input variable importance measure

If a certain input variable is irrelevant to the output, a change in the variable does not affect the output. When a certain input variable actually affects the output, therefore, the amount of

fluctuation of the output caused by a change in the variable could be larger.

Based on this idea of using the fluctuation of the output to evaluate an input, we propose our new measure, the random-input variable importance measure, in this study (see Fig. 2). When using this random-input variable importance measure, the confidence value of the regulation of the *n*-th gene from the *m*-th gene, $C_{n,m}^{R}$, is computed by

$$C_{n,m}^{R} = \frac{1}{Sq_{w0}}\left(WSE_R - WSE_0\right), \tag{13}$$

where

$$WSE_R = \sum_{k=1}^{K_T} w_k^T \left[\hat{F}_n^*\left(\mathbf{X}_{-n}\big|_{t_k}^{(m)}\right) - y_{t_k}\right]^2 + \sum_{k=1}^{K_S} w_k^S \left[\hat{F}_n^*\left(\mathbf{X}_{-n}\big|_{s_k}^{(m)}\right) - y_{s_k}\right]^2, \tag{14}$$

$$WSE_0 = \sum_{k=1}^{K_T} w_k^T \left[\hat{F}_n^*\left(\mathbf{X}_{-n}\big|_{t_k}\right) - y_{t_k}\right]^2 + \sum_{k=1}^{K_S} w_k^S \left[\hat{F}_n^*\left(\mathbf{X}_{-n}\big|_{s_k}\right) - y_{s_k}\right]^2, \tag{15}$$

$\mathbf{X}_{-n}\big|_{t_k} = \left(X_1\big|_{t_k}, \cdots, X_{n-1}\big|_{t_k}, X_{n+1}\big|_{t_k}, \cdots, X_N\big|_{t_k}\right)$, and $\mathbf{X}_{-n}\big|_{s_k} = \left(X_1\big|_{s_k}, \cdots, X_{n-1}\big|_{s_k}, X_{n+1}\big|_{s_k}, \cdots, X_N\big|_{s_k}\right)$. $\mathbf{X}_{-n}\big|_{t_k}^{(m)}$ and $\mathbf{X}_{-n}\big|_{s_k}^{(m)}$ are vectors constructed by changing the expression levels of the *m*-th gene in $\mathbf{X}_{-n}\big|_{t_k}$ and $\mathbf{X}_{-n}\big|_{s_k}$, respectively. Values for the expression levels of the *m*-th gene in $\mathbf{X}_{-n}\big|_{t_k}^{(m)}$ and $\mathbf{X}_{-n}\big|_{s_k}^{(m)}$ are randomly drawn from [L_m, R_m], where

$$L_m = \min S_m, \tag{16}$$

$$R_m = \max S_m, \tag{17}$$

$$S_m = \left\{X_m\big|_{t_k} \,\middle|\, k = 1, 2, \cdots, K_T\right\} \cup \left\{X_m\big|_{s_k} \,\middle|\, k = 1, 2, \cdots, K_S\right\}. \tag{18}$$

Note that the confidence values computed according to the random-input variable importance measure depend strongly on the random numbers used. In order to reduce the effect of random numbers, the confidence values, $C_{n,m}^{R}$'s, are computed N_{rnd} times by changing the random numbers, and their averages are used to rank the regulations. As the equations (16), (17) and (18) show, on the other hand, the proposed measure determines the ranges of the random variables according to the distribution of the gene expression levels. In this study, we assume that the gene expression data contain neither erroneous large nor small values. These erroneous values might make importance scores computed by the equation (13) unreliable. When using the proposed measure, thus, we should remove these erroneous values in advance.

The sum of the confidence values of the regulations of a certain gene from the other genes, that are calculated according to the equation (13), is not restricted to 1. The removal of the restriction

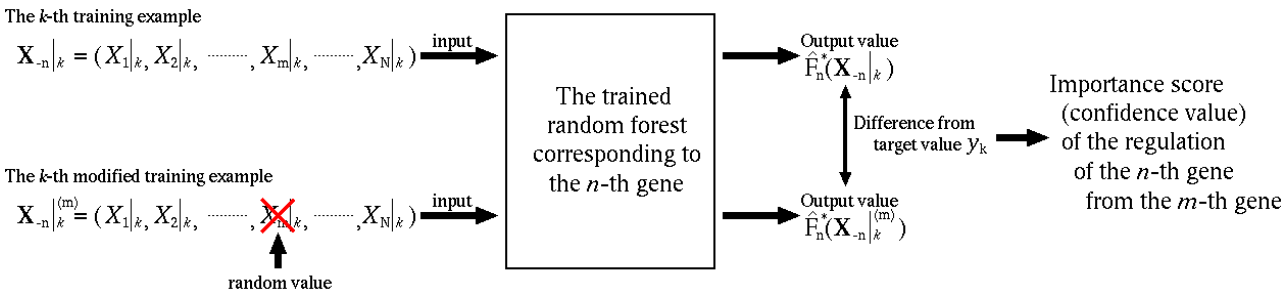

Figure 2. The concept of the random-input variable importance measure

could make the importance scores absolute and therefore the importance measure proposed in this study is suitable for comparing the confidence values obtained from multiple random forests. In this study, we thus use the random-input variable importance measure in place of the standard measure for the computation of the confidence values. The importance measure that uses permuted values, instead of using random values, in the proposed measure is equivalent to the permutation-based importance measure [10]. We should note here that, although we combined the random-forest-based inference method with the permutation-based importance measure in the section 4.4, it did not always outperform the original inference method. The poor performance would be caused by a reason that the permutation-based importance measure depends too much on the distribution of input variables. In the genetic network inference, what we try to know is a nature of the target function F_n. As the nature of the function F_n is independent of the distribution of input variables, we should not depend much on the input distribution.

4. Experiments with artificial gene expression data

This section describes experiments we conducted with artificial genetic network inference problems to evaluate the performance of the proposed approach.

4.1 Analysis of DREAM3 networks

In this experiment, we compared the original random-forest-based inference method [19] with a modified version of the method that computes the confidence values of the regulations using the random-input variable importance measure proposed in this study.

4.1.1 Experimental setup

The two inference methods were applied to 5 artificial genetic network problems obtained from the DREAM3 *in silico* network challenges (http://dreamchallenges.org/): Ecoli1, Ecoli2, Yeast1, Yeast2 and Yeast3. The target networks of these problems consisted of 100 genes ($N = 100$).

Each problem used here contained both time-series and static expression data of all 100 genes. The time-series data consisted of 46 datasets of gene expression levels obtained by solving a set of differential equations on the target network [32], and were polluted by internal and external noise. The datasets began from randomly generated initial values, and each gene in each set was assigned 21 observations. The static data consisted of wild-type, knockout, and knockdown data. The wild-type data contained the steady-state gene expression levels of the unperturbed network. The knockout and knockdown data contained the steady-state expression levels of every single-gene knockout and every single-gene knockdown, respectively. When trying to solve the n-th subproblem corresponding to the n-th gene, however, we removed the static data of the knockout and the knockdown of the n-th gene. The number of measurements of the time-series experiment, K_T, was therefore $46 \times 21 = 966$, while that of the steady-state experiment, K_S, was $1+100+100-2 = 199$. Noisy time-series data were provided as the observed data in the problems, so they were smoothed using a local linear regression [8], a data-smoothing technique. The same smoothing technique was used to estimate the time derivatives of the gene expression levels. This study inferred the genetic network only from the smoothed time-series of the gene expression levels, their estimated time derivatives, and the static gene expression data.

The number of trees in the random forest, the number of input variables to be considered in

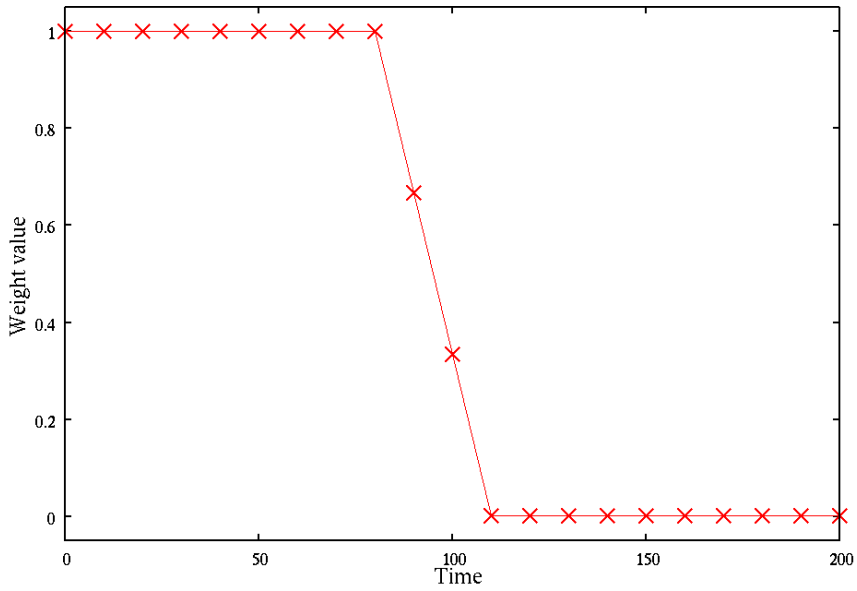

Figure 3. The values for the weight parameters, w_k^T's, corresponding to each of the time-series datasets of the DREAM3 problems

each internal node of each tree, and the maximum height of each tree were set to 1000, $\left\lceil \frac{N-1}{3} \right\rceil$, and 32, respectively. These settings were determined according to the recommended values for the random forest [4], and our earlier study [19] also used them. Because the parameter to be estimated, β_n, was positive, we searched for its optimum value in a logarithmic space. The search area of $\log \beta_n$ was $[-10, 5]$. Note that, in order to infer genetic networks, we must assign values to the weight parameters w_k^T's and w_k^S's. The weight parameters for the measurements in each of the 46 time-series datasets were set at the values used in our earlier paper [19], namely, 0.6674 for the 10th measurement, 0.3348 for the 11th measurement, and 0.002174 for the last 10 measurements. The weight parameters for the other measurements in the time-series datasets and for the measurements in the static dataset were set to 1.0 and 1.1, respectively. Our earlier paper determined these values according to the guidelines for determining weight parameters [19]. The weight values assigned for each of the time-series datasets are shown in Fig. 3. The number of iterations required for statistically evaluating the confidence values computed based on the random-input variable importance measure, N_{rnd}, was set to 100. As the inference methods used here were stochastic, we performed 10 trials on each of the 5 problems by changing the seed for pseudo-random numbers.

4.1.2 Results

Table 1 lists performances of the original inference method [19] and the proposed approach that uses the random-input variable importance measure. The performance of the method was evaluated based on the area under the recall-precision curve (AURPC) (see Fig. 4). Note here that auto-regulations/auto-degradations were disregarded in the evaluation of the performance. The table shows that the use of the random-input variable importance measure in place of the standard measure greatly improved the performance of the random-forest-based method. The improvement achieved by adopting the proposed measure was more than 8% with respect to the AURPC.

Table 1. The performances of the proposed approach and the original random-forest-based inference method [19] on the DREAM3 problems

	Ecoli1	Ecoli2	Yeast1	Yeast2	Yeast3
	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD
Inference method using random-input variable importance measure	0.61037 ± 0.00711	0.59094 ± 0.00559	0.60051 ± 0.00283	0.44873 ± 0.00368	0.34937 ± 0.00242
Random-forest-based inference method [19]	0.41918 ± 0.00388	0.54477 ± 0.00586	0.50083 ± 0.00285	0.39482 ± 0.00344	0.31291 ± 0.00223

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

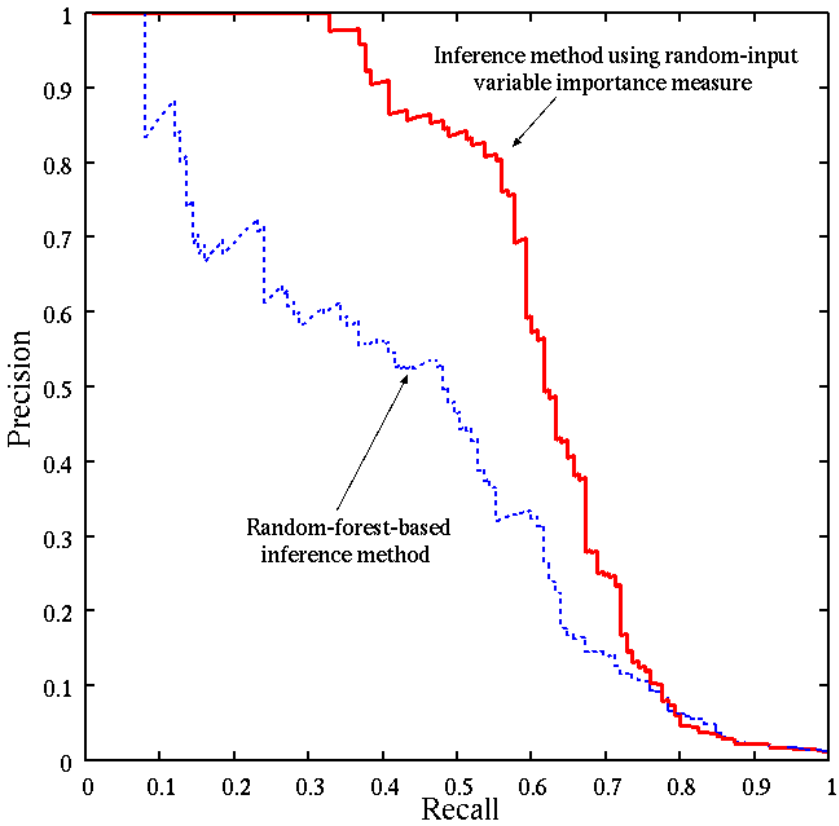

Figure 4. A sample of the recall-precision curves obtained from the inference method using the random-input variable importance measure (red bold line) and the random-forest-based inference method (blue dotted line) on the Ecoli1 problem

The DREAM3 networks contain several genes that are not regulated by any gene. When we used the random-input variable importance measure, the sums of the confidence values computed from the random forest corresponding to a gene not regulated by other genes averaged about 1.5418 ± 0.1799, 1.5892 ± 0.1534, 1.6327 ± 0.1636, 1.5712 ± 0.1550 and 1.7345 ± 0.2474 on Ecoli1, Ecoli2, Yeast1, Yeast2, and Yeast3, respectively. When a gene was regulated by one or more other genes, on the other hand, the sums of the confidence values obtained from the random forest corresponding to the gene averaged about 1.9191 ± 0.8383, 2.0056 ± 0.5979, 2.0400 ± 0.7733, 1.9814 ± 0.5858 and 1.8247 ± 0.4866 on Ecoli1, Ecoli2, Yeast1, Yeast2, and Yeast3, respectively. This finding indicates that the confidence value of the candidate regulation of an unregulated gene

Table 2. The effect of a value for N_{rnd} in the proposed approach on the Ecoli1 problem

Random-forest-based inference method [19]	Inference method using random-input variable importance measure				
	$N_{rnd} = 10$	$N_{rnd} = 20$	$N_{rnd} = 50$	$N_{rnd} = 100$	$N_{rnd} = 200$
Rank diff. ± STDr	Rank diff. ± STDr	Rank diff. ± STDr	Rank diff. ± STDr	Rank diff. ± STDr	Rank diff. ± STDr
358.96 ± 7.53	357.63 ± 9.54	355.99 ± 9.46	354.84 ± 9.34	354.51 ± 9.46	354.36 ± 9.45

The result of the random-forest-based inference method is also shown. The averaged difference between the place of each regulation ranked by the inference method in each trial and that obtained from the ranking with respect to the confidence values averaged over the 10 trails (Rank diff.) and its standard deviation (STDr) are shown.

tends to be smaller. Note here that, when the standard variable importance measure is used, the confidence values computed from a single random forest always sum up to almost 1. The removal of the restriction imposed on the standard variable importance measure may help partly explain why the proposed approach outperformed the original inference method. Given this feature of the random-input variable importance measure, we believe that the measure is a more appropriate tool for comparing the confidence values obtained from the multiple random forests.

Our experimental results indicate that, when the random-input variable importance measure is used, the ranking of the candidate regulations with respect to the confidence values computed from the multiple random forests is better. On the other hand, the ranking of the candidate regulations also seems to be slightly better when it is obtained from the confidence values of a single random forest. In each ranking obtained from each of the random forests trained, our approach averagely ranked the regulations actually contained in the gold-standard network as follows: 11.7th, 8.7th, 10.9th, 21.4th, and 28.1th on Ecoli1, Ecoli2, Yeast1, Yeast2 and Yeast3, respectively. Similarly, the original inference method averagely ranked the regulations as follows:12.9th, 8.8th, 11.8th, 22.1th, and 29.3th on Ecoli1, Ecoli2, Yeast1, Yeast2 and Yeast3, respectively. This feature of the random-input variable importance measure probably also contributed to the better performance of the proposed approach.

As mentioned in the section 3.2, in order to reduce the effect of a stochastic nature of the variable importance measure proposed in this study, our inference method computes the importance scores N_{rnd} times. In order to confirm the validity of our setting, i.e., N_{rnd} = 100, we performed another experiment here. In this experiment, we applied the proposed inference method with N_{rnd} =10, 20, 50, 100 and 200 to the Ecoli1 problem. We performed 10 trials on each N_{rnd} setting by changing the seed for pseudo-random numbers. Because of the stochastic nature of the inference method, the obtained results were slightly different from each other. Table 2 shows the averaged difference between the place of each of the regulations in the ranking with respect to the confidence values averaged over the 10 trials and that in the ranking of each trial. As the table shows, the averaged difference of the places of the regulations decreases as a value for N_{rnd} increases. This fact indicates that the effect of the randomness in the proposed inference approach decreases with an increase in N_{rnd}. The table also lists the result of the original random-forest-based inference method [19]. The averaged difference of the places of the regulations in the proposed approach was smaller than that in the original inference method. Note here that, while the output of the original inference method is affected by the randomness in the random forest, the output of the proposed approach is affected by the randomness in both the random forest and the random-input variable importance measure. The experimental results thus show that, when N_{rnd} is set to 100, the fluctuation in the output caused by the random-input variable importance measure is almost

negligible.

4.2 Analysis of DREAM4 networks

Next, we compared the performance of the proposed approach with the performances of other genetic network inference methods on the DREAM4 problems.

4.2.1 Experimental setup

In this section, we describe the application of the proposed approach to 5 problems from the DREAM4 *in silico* network challenges. Similar to the DREAM3 problems, each of the target networks of these problems consisted of 100 genes. These networks were described using a model identical to the model in the DREAM3 networks [32].

Each problem contained both time-series and static expression data of all 100 genes. The time-series data consisted of 10 sets of time-series of gene expression levels. Each time-series dataset consisted of the expression levels at 21 time points, and was polluted by internal and external noise. A dataset was constructed by applying a perturbation to the network at the first time point and removing the perturbation at the 11th time point. The perturbation affected the transcription rates of a different set of genes in each dataset. To take the perturbations into account explicitly, we added 10 elements to the gene expression data, each corresponding to one of the perturbations. The *i*-th added element had a value of 1 for the measurements between the first and 10th time points in the *i*-th time-series dataset generated by adding the *i*-th perturbation, and a value of 0 for the other measurements. The number of elements, *N*, was therefore 100+10 = 110. The static data consisted of wild-type, knockout, and knockdown data. When trying to solve the *n*-th subproblem corresponding to the *n*-th gene, we also removed the static data of the knockout and the knockdown of the *n*-th gene. The numbers of measurements of the time-series and steady-state experiments, i.e., K_T and K_S, were thus $10 \times 21 = 210$ and $1+100+100-2 = 199$, respectively. The local linear

Table 3. The performances of the inference methods on the DREAM4 problems

	Network1	Network2	Network3	Network4	Network5
	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD
Inference method using random-input variable importance measure	0.53504 ± 0.00331	0.32987 ± 0.00325	0.42130 ± 0.00400	0.40323 ± 0.00291	0.30411 ± 0.00263
Random-forest-based inference method [19]	0.42797 ± 0.00332	0.28656 ± 0.00300	0.33930 ± 0.00397	0.34079 ± 0.00347	0.27199 ± 0.00415
dynGENIE3 [16]	0.34 –	0.22 –	0.32 –	0.34 –	0.22 –
MCZ [12]	0.48 –	0.38 –	0.38 –	0.36 –	0.17 –
dynGENIE3 + MCZ	0.60 –	0.43 –	0.47 –	0.52 –	0.37 –
iRafNet [29]	0.552 –	0.337 –	0.414 –	0.421 –	0.298 –

The AURPCs of the proposed approach, the original random-forest-based inference method [19], dynGENIE3 [16], MCZ [12], a combination of dynGENIE3 and MCZ, and iRafNet [29] are shown.

regression [8] was used to smooth the given time-series data and to estimate the time derivatives of the gene expression levels. We inferred a genetic network using only the smoothed time-series of

the gene expression levels, their estimated time derivatives, and the static gene expression data.

According to our earlier paper [19], the weight values for the 6th, 7th, 8th, 9th, and 10th measurements in each of the time-series datasets were set to 0.2, the weight values for the 17th, 18th, 19th, 20th, and 21st measurements were set to 0.02, and the weight values for the 4th, 5th, 15th and 16th measurements were set to 0.7333, 0.4667, 0.6733 and 0.3466, respectively. The values for the remaining w_k^T 's and w_k^S 's were set to 1.0 and 1.1, respectively. The other experimental conditions were unchanged from those used in the previous experiment.

4.2.2 Results

In this experiment, the performance of the inference method was also evaluated using the area under the recall-precision curve (AURPC). As mentioned previously, we inferred the regulations of the 100 genes from these genes and the 10 additional elements that represent 10 perturbations in this experiment. When we evaluated the performance of the method, however, we disregarded the regulations of the genes from the additional elements. In addition, we also disregarded the auto-regulations/auto-degradations. The AURPCs of the proposed inference method on the 5 DREAM4 problems were listed in Table 3. The table also shows the AURPCs of the original random-forest-based inference method [19], dynGENIE3 [16], MCZ [12], a combination of dynGENIE3 and MCZ, and iRafNet [29]. The values of the AURPCs of dynGENIE3, MCZ, and the combination of dynGENIE3 and MCZ were taken from Huynh-Thu *et al*. [16], and the values of the AURPCs of iRafNet were taken from Petralia *et al*. [29].

The table shows that the proposed approach outperformed the original random-forest-based method [19] even on the DREAM4 problems. In this experiment, the use of the random-input variable importance measure brought about an improvement of more than 11% with respect to the AURPC. As the table shows, on the other hand, the proposed approach performed better than dynGENIE3 and MCZ on most of the 5 problems. We must note here that, while dynGENIE3 was designed based on the random forest, MCZ is based on a very different concept. Huynh-Thu and colleagues [16] mentioned that potential performance improvements could be achieved by combining inference methods designed based on different concepts. The table shows that the combination of dynGENIE3 and MCZ performed quite well. The good performance of iRafNet, another random-forest-based inference method, seems to have resulted from a similar cause. Although the proposed approach did not always outperform iRafNet or the combination of dynGENIE3 and MCZ, we believe that we could improve the approach by combining it with a different kind of inference method, such as MCZ. However, MCZ requires static gene expression data of every single-gene knockout, that can hardly be expected in real experiments. We should also note again that the random-input variable importance measure proposed in this study can be applied to any random-forest-based inference method. By using the proposed measure in place of the standard variable importance measure, we could improve the performances of the other random-forest-based inference methods.

4.3 Analysis of random networks

We then checked the performance of the proposed approach on problems with target networks described by a model different from that of the previous experiments.

4.3.1 Experimental setup

In this experiment, we used the Vohradský's model [36] to describe target networks. The Vohradský's model is a set of differential equations of the form

$$\frac{dX_n}{dt} = \frac{k_{1n}}{1+\exp\left(-b_n - \sum_{m=1}^{N} w_{n,m} X_m\right)} - k_{2n} X_n, \ (n = 1, 2, \cdots, N), \tag{19}$$

where k_{1n}, k_{2n}, b_n and $w_{n,m}$ ($m, n = 1, 2, \cdots, N$) are model parameters. We can change the structure of the network by changing the values of these parameters, and the structure adopted might influence the inference ability of the inference method used. We thus constructed 10 genetic network inference problems with different target networks and checked the performances of the proposed approach and the original random-forest-based inference method [19] on the constructed problems. These target networks were randomly constructed according to the procedure described in Kimura et al. [18]. Each of the networks consisted of 30 genes ($N = 30$).

Each of the constructed inference problems had time-series and static data. The time-series data consisted of 10 time-series datasets of gene expression levels generated by solving a set of the differential equations (19) on the target model corresponding to the problem. The initial values of these sets were selected randomly from [0.0, 3.0]. Each dataset consisted of the expression levels at 21 time points spaced apart by intervals of 0.2 time units. As the static data, we constructed steady-state gene expression levels for the wild-type and every single-gene knockout. The measurement noise was simulated by adding 10% Gaussian noise to the computed gene expression data. As in the previous experiments, we disregarded the steady-state gene expression levels of the knockout of the n-th gene when trying to analyze the n-th gene. The numbers of measurements contained in the time-series and static data, K_T and K_S, were therefore $10 \times 21 = 210$ and $1+30-1 = 30$, respectively.

We also determined values for the weight parameters according to our earlier paper [19]: The weight values for the last 6 measurements in each of the time-series datasets were set to $1/(6\times10) \approx 0.01667$; the weight values for the 14th and 15th measurements were set to 0.6722 and 0.3444, respectively; and the remaining weight values for the time-series datasets and static dataset were set to 1.0 and 1.1, respectively. The other experimental settings were identical to those used in the previous experiments.

4.3.2 Results

The AURPCs of the proposed approach and the original random-forest-based inference method were 0.70518 ± 0.05380 and 0.68207 ± 0.04622, respectively, on average. Our approach outperformed the original inference method on 8 of the 10 problems. In the two problems in which our approach underperformed the original method, the degradation caused by the use of the random-input variable importance measure was less than 0.44% with respect to the AURPC. Although the inference ability of the proposed approach was better, its computational cost was higher. The proposed approach and the random-forest-based inference method [19] averagely took 99.1 min. and 31.0 min., respectively, on a workstation (Xeon Gold 6150 2.7GHz; a single-core use) to infer each of the networks described here. As it is important to extract more useful information from the limited amount of gene expression data, however, the higher computational cost of the proposed approach would be a little issue for concern.

Our experimental results suggest that the improvement achieved by the proposed measure was independent of the model used to describe the target network. We thus think that the proposed

Table 4. The performances of the inference methods using the different importance measures on the DREAM3 problems

	Ecoli1	Ecoli2	Yeast1	Yeast2	Yeast3
	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD
Inference method using random-input variable importance measure	0.61037 ± 0.00711	0.59094 ± 0.00559	0.60051 ± 0.00283	0.44873 ± 0.00368	0.34937 ± 0.00242
Random-forest-based inference method [19]	0.41918 ± 0.00388	0.54477 ± 0.00586	0.50083 ± 0.00285	0.39482 ± 0.00344	0.31291 ± 0.00223
Inference method using permutation-based importance measure [10]	0.44671 ± 0.00497	0.48297 ± 0.00664	0.50540 ± 0.00412	0.36573 ± 0.00475	0.28741 ± 0.00195
Inference method using random forest permutation accuracy importance [33]	0.31919 ± 0.00309	0.37120 ± 0.00357	0.38605 ± 0.00203	0.29975 ± 0.00340	0.25093 ± 0.00231

Table 5. The performances of the inference methods using the different importance measures on the DREAM4 problems

	Network1	Network2	Network3	Network4	Network5
	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD
Inference method using random-input variable importance measure	0.53504 ± 0.00331	0.32987 ± 0.00325	0.42130 ± 0.00400	0.40323 ± 0.00291	0.30411 ± 0.00263
Random-forest-based inference method [19]	0.42797 ± 0.00332	0.28656 ± 0.00300	0.33930 ± 0.00397	0.34079 ± 0.00347	0.27199 ± 0.00415
Inference method using permutation-based importance measure [10]	0.51522 ± 0.00301	0.33839 ± 0.00256	0.39427 ± 0.00483	0.38847 ± 0.00317	0.33446 ± 0.00313
Inference method using random forest permutation accuracy importance [33]	0.45442 ± 0.00181	0.30109 ± 0.00204	0.34665 ± 0.00427	0.35207 ± 0.00166	0.28978 ± 0.00357

idea may work well even in real genetic network inference problems.

4.4 Comparison with the other variable importance measures

When the standard variable importance measure is used in the random-forest-based inference methods, the confidence values of the candidate regulations, that are computed from a single random forest, always sum up to almost 1. The random-input variable importance measure proposed in this study was developed to remove this restriction. This is not to say, that the measure we propose is the only variable importance measure free from the aforementioned restriction. In this section, we thus checked the performances of some other variable importance measures, i.e., the permutation-based importance measure [10] and the random forest permutation accuracy importance [33], that are free from the restriction. As mentioned previously, the permutation-based importance measure is equivalent to the measure that uses permuted values, instead of using random values, in the random-input variable importance measure proposed in this study. The random forest permutation accuracy importance is an algorithm similar to the permutation-based importance measure, but utilizes the feature of the random forest. The random forest permutation accuracy importance is therefore available only for the random forest.

Similar to the proposed approach, we constructed two other inference methods by using the permutation-based importance measure and the random forest permutation accuracy importance, respectively, in place of the standard variable importance measure in the original random-forest-

based inference method [19]. We then applied the constructed inference methods to the DREAM3 and DREAM4 problems. The experimental results are listed in Tables 4 and 5. The experimental results indicate that, even when we simply introduce the importance measure free from the aforementioned restriction into the random-forest-based inference method, we do not always improve the performance of the inference method. The gene expression data often contain the measurements similar to each other. In our earlier study [19, 21], we found that the training examples similar to each other degrade the quality of the importance scores computed according to the standard variable importance measure. The poor performances of the inference methods with the other importance measures would be caused by a similar reason.

5. Experiment with real gene expression data

In the final experiment of this study, we used the proposed approach to analyze real gene expression data.

5.1 Experimental setup

In this experiment, we analyzed the expression data of 11 immediate early genes related to transcription, i.e., ATF3, EGR1, EGR2, EGR3, ETS2, FOS, FOSB, FOSL1, JUN, JUNB, and MYC. The time-series and static gene expression levels were obtained from the FANTOM5 dataset (http://fantom.gsc.riken.jp/5/) [9]. The time-series datasets consisted of sets of gene expression levels in the following cell types, measured at successive time points after exposing the cells to different external stimuli: Saos-2, MCF-7, ARPE-19, lymphatic endothelial, mesenchymal stem, and aortic smooth muscle cells. Table 6 shows detailed information on the time-series datasets used, and Fig. 5 shows a sample of them. As the static data, we used sets of gene expression levels for the Saos-2 and mesenchymal stem cells given as untreated controls. We also used the measurement at time 0 in each of the time-series datasets as static data. The numbers of measurements contained in the time-series and static data in this experiment, K_T and K_S, were thus 102 (= 11+16+16+13+16+10+10+10) and 10 (= 2+8), respectively. To account for the external stimuli explicitly, we added the following 8 elements to the gene expression data: `ascorbic acid and BGP,' `EGF1,' `HRG,' `TGF-β and TNF-α,' `VEGF,' `IBMX, DEX and insulin,' `FGF-2,' and `IL-1B.' Each added element

Table 6. The experimental settings where real time-series datasets were measured

Cell name	Stimulus	Measured time (min.)
Saos-2	Ascorbic acid and BGP	0,15,30,45,60,80,100,120,150,180,240
MCF-7	EGF1	0,15,30,45,60,80,100,120,150,180,210, 240,300,360,420,480
MCF-7	HRG	0,15,30,45,60,80,100,120,150,180,210, 240,300,360,420,480
ARPE-19	TGF-β and TNF-α	0,15,30,45,60,80,100,120,150,180,210, 240,300
Lymphatic endothelial	VEGF	0,15,30,45,60,80,100,120,150,180,210, 240,300,360,420,480
Mesenchymal stem	IBMX, DEX and insulin	0,15,30,45,60,80,100,120,150,180
Aortic smooth muscle	FGF-2	0,15,30,45,60,120,180,240,300,360
Aortic smooth muscle	IL-1B	0,15,30,45,60,120,180,240,300,360

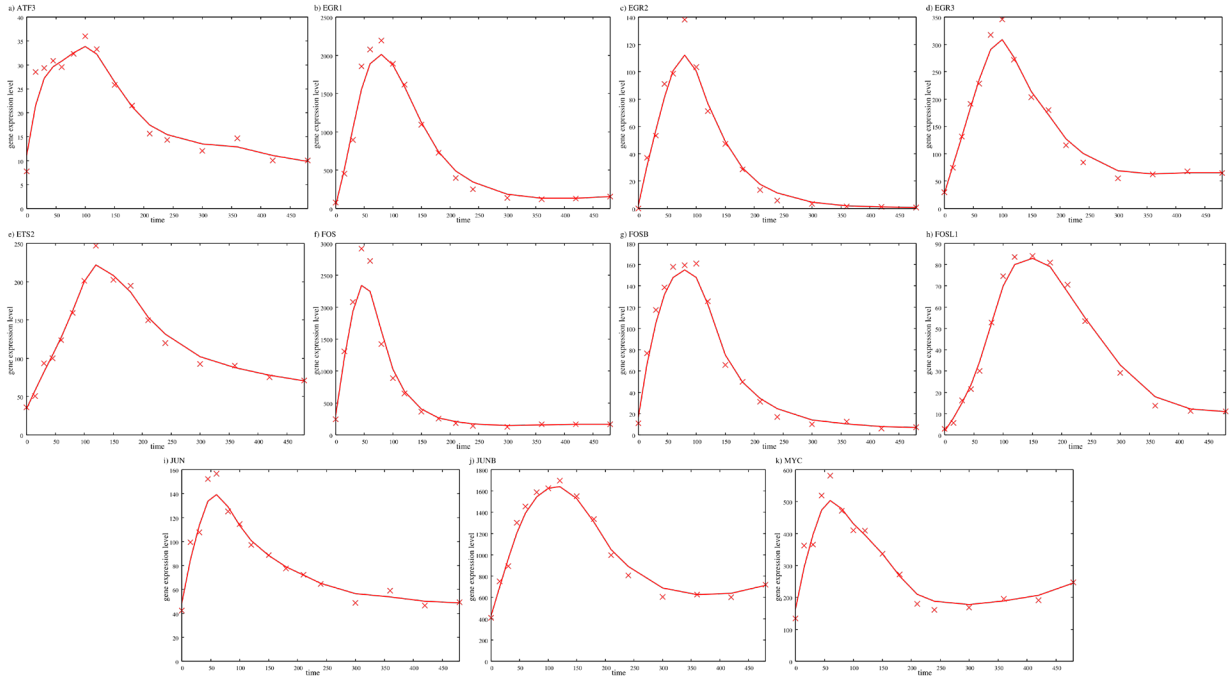

Figure 5. The time-series of expression levels of a) ATF3, b) EGR1, c) EGR2, d) EGR3, e) ETS2, f) FOS, g) FOSB, h) FOSL1, i) JUN, j) JUNB and k) MYC in MCF7 cells stimulated by HRG
Solid line: smoothed expression data used for inferring genetic networks. Plus symbol: measured gene expression data.

corresponded to a stimulus applied to the cells. According to Kimura et al. [20], we considered the decomposition of the biochemical compounds used for stimulating the cells. One added element thus had a value of $0.9^{\frac{t}{48}}$ for the measurements in the time-series dataset obtained by applying the stimulus corresponding to the element, where t was the time (min.) elapsed after the cell stimulation. A value of 0 was assigned to the added element for the other measurements. The number of the total elements, N, was 11+8 = 19. By applying the proposed approach to the gene expression data described here, we inferred regulations of the 11 selected genes from these genes and the 8 additional elements.

5.2 Results

The top 20 regulations with respect to the confidence values assigned by the proposed approach are listed in Table 7. Because of a stochastic nature of the inference methods applied here, the confidence values assigned by the methods were slightly different every trial. In this study, we therefore ranked the regulations using the confidence values averaged over 10 trials. On average, however, 94.5% of the top 20 regulations obtained in each trial were the same as those obtained from the averaged confidence values. According to the STRING database (https://string-db.org/) [34], the regulations written in a bold face font in the table seem to be reasonable, since the interactions between the proteins corresponding to the genes have reportedly been confirmed in human and/or other species. The regulations written in an italic font also appeared to be reasonable, given the suggestions from earlier reports [28, 39, 41]. Figs. 6 and 7 show the networks of the top 30 regulations ranked by the proposed approach and the original inference method, respectively.

Table 7. The top 20 regulations ranked with respect to the confidence values computed by the proposed approach and the original inference method [19]

Rank	Inference method using random-input variable importance measure	Random-forest-based inference method [19]
1	**EGR1 ← FOS**	**EGR1 ← FOS**
2	*ATF3 ← TGF-β and TNF-α*	*FOS ← HRG*
3	**EGR2 ← FOS**	*ATF3 ← TGF-β and TNF-α*
4	**MYC ← FOS**	*EGR2 ← HRG*
5	*FOS ← HRG*	**JUNB ← FOSB**
6	**EGR3 ← FOS**	EGR3 ← EGR2
7	**JUNB ← FOSB**	**EGR3 ← FOS**
8	EGR3 ← EGR2	**FOSL1 ← ATF3**
9	**FOSL1 ← ATF3**	**EGR2 ← FOS**
10	**JUN ← VEGF**	EGR1 ← EGR2
11	*EGR2 ← HRG*	**MYC ← FOS**
12	EGR1 ← EGR2	**JUNB ← EGR2**
13	**FOS ← EGR2**	EGR3 ← EGR1
14	ETS2 ← EGR2	**FOSB ← JUNB**
15	**JUN ← FOSB**	**JUN ← VEGF**
16	**JUNB ← EGR2**	ETS2 ← EGR2
17	EGR3 ← EGR1	**JUN ← FOSB**
18	ATF3 ← FOSB	FOSL1 ← FOSB
19	**FOSB ← EGR2**	**FOSB ← EGR2**
20	FOSL1 ← FOSB	**ATF3 ← JUN**

The regulations written in boldface and italic fonts have reportedly been confirmed in human and/or other species and are accordingly assumed to be reasonable.

Table 7 also shows the top 20 regulations obtained from the original random-forest-based inference method [19]. As shown in the table, the top 20 regulations obtained from the proposed approach and the original inference method were similar to each other. In this experiment, the proposed approach did not always outperform the random-forest-based inference method with respect to the number of the regulations that have been already confirmed. For example, however, while our approach assigned an 18th-place ranking to the regulation of ATF3 from FOSB, the original method ranked it 41st. In spite of this much higher ranking assigned to the regulation of ATF3 from FOSB by the original method, we found no earlier reports proving the existence of this regulation. ATF3 and FOSB are known to be induced by cAMP and MAPK signaling, respectively (e.g., [26, 42]). In addition, the interaction between the cAMP and MAPK signaling pathways has been confirmed [37]. The regulation of ATF3 from FOSB might therefore reflect an indirect interaction between ATF3 and FOSB. Based on the facts just described, however, we think that it would be worthwhile to confirm the existence of a direct regulation of ATF3 from FOSB.

6. Conclusion

Several researchers have focused on random-forest-based inference methods. These methods assign confidence values to all of the candidate regulations. To our knowledge, all of the random-forest-based methods use the standard variable importance measure to assign the confidence values. Our group believes however that the standard variable importance measure is detrimental to the

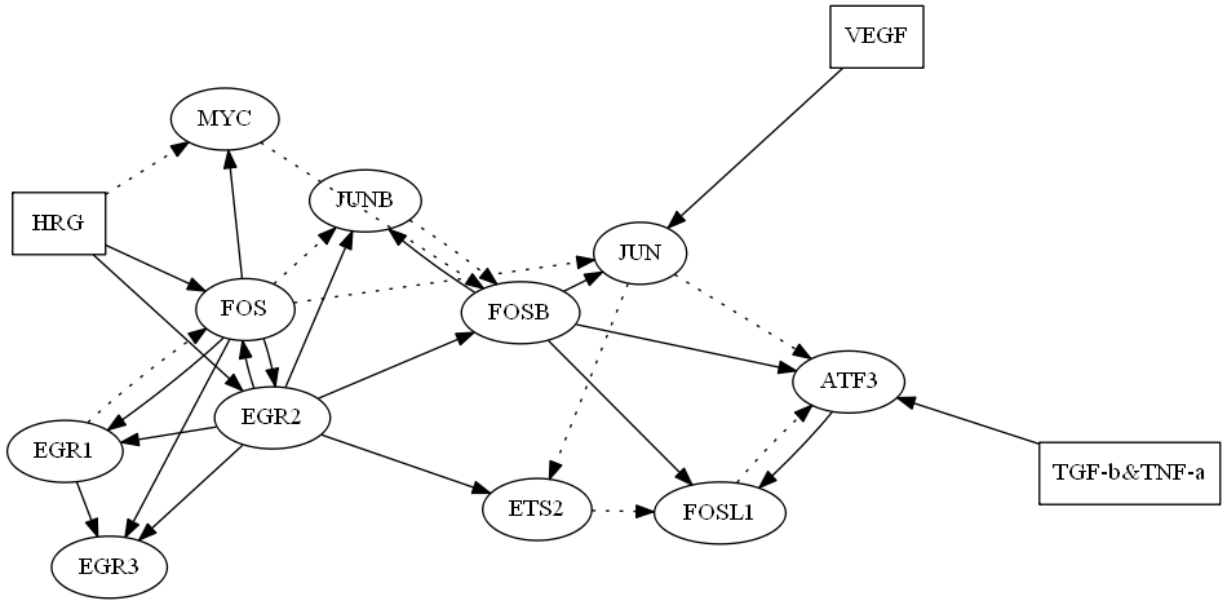

Figure 6. The network of the top 30 regulations obtained from the proposed approach
Solid lines represent the top 20 regulations. Circles and squares represent the genes and external stimuli, respectively.

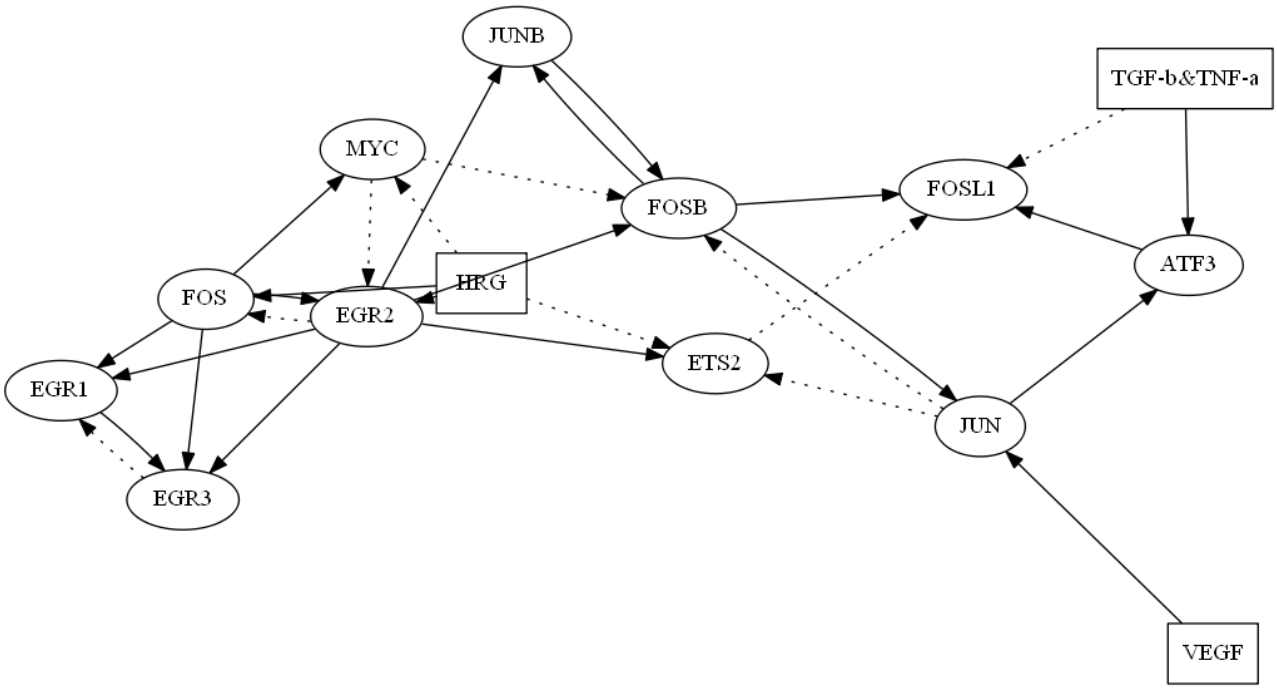

Figure 7. The network of the top 30 regulations obtained from the original method [19]

inference of genetic networks. In this study, we proposed a new measure, i.e., the random-input variable importance measure, as an alternative, and applied it to the existing random-forest-based inference method. We then showed, through numerical experiments, that the use of the random-input variable importance measure in place of the standard measure can improve the performance of the random-forest-based inference method.

Our experimental results suggest that the random-input variable importance measure proposed

in this study works well not only for comparing importance scores computed from multiple random forests but for comparing those computed from a single random forest. However, we have only confirmed its effectiveness on genetic network inference problems so far. In the future work, we plan to confirm its effectiveness on different kinds of problems. On the other hand, we think that the proposed measure relates to the feature selection (e.g., [5, 13]). This study however focused only on the drawbacks of the standard variable importance measure, and then proposed the new measure to overcome them. The use of state-of-the-art feature selection techniques might thus make the performance of the inference method better.

Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Number 18H04031.

References

[1] Aalto, A.; Viitasaari, L.; Ilmonen, P.; Mombaerts, L.; Goncalves, J. Gene regulatory network inference from sparsely sampled noisy data, *Nat. Commun.*, **2020**, *11*, 3493. doi: 10.1038/s41467-020-17217-1

[2] Akutsu, T.; Miyano, S.; Kuhara, S. Inferring qualitative relations in genetic networks and metabolic pathways, *Bioinformatics*, **2000**, *16*, 727–734. doi: 10.1093/bioinformatics/16.8.727

[3] Archer, K. J.; Kimes, R.V. Empirical characterization of random forest variable importance measures, *Comput. Stat. Data Anal.*, **2008**, 2249–2260. doi: 10.1016/j.csda.2007.08.015

[4] Breiman, L. Random forests, *Machine Learning*, **2001**, *45*, 5–32. doi: 10.1023/A:1010933404324

[5] Cai, J.; Kuo, J.; Wang, S; Yang, S. Feature selection in machine learning: A new perspective, *Neurocomputing*, **2018**, *300*, 70–79. doi: 10.1016/j.neucom.2017.11.077

[6] Chou, I. C.; Martens, H.; Voit, E. O. Parameter estimation in biochemical systems models with alternating regression, *Theor. Biol. Med. Model.*, **2006**, *3*, 35. doi: 10.1186/1742-4682-3-25

[7] Chou, I. C.; Voit, E. O. Recent developments in parameter estimation and structure identification of biochemical and genomic systems, *Math. Biosci.*, **2009**, *219*, 57–83. doi: 10.1016/j.mbs.2009.03.002

[8] Cleveland, W. S. Robust locally weight regression and smoothing scatterplots, *J. Am. Stat. Assoc.*, **1979**, *79*, 829–836. doi: 10.2307/2286407

[9] FANTOM Consortium; RIKEN PMI; CLST. A promoter-level mammalian expression atlas, *Nature*, **2014**, *507*, 462–470. doi: 10.1038/nature13182

[10] Fisher, A.; Rudin, C.; Dominici, F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously, https://arxiv.org/abs/1801.01489

[11] Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees, *Machine Learning*, **2006**, *63*, 3–42. doi: 10.1007/s10994-006-6226-1

[12] Greenfield, A.; Madar, A.; Ostrer, H.; Bonneau, R. DREAM4: Combining genetic and

dynamic information to identify biological networks and dynamical models, *PLoS One*, **2010**, *5*, e13397. doi: 10.1371/journal.pone.0013397

[13] Guyon, I.; Elisseeff, A. An introduction to variable and feature selection, *J. Mach. Learn. Res.*, **2003**, *3*, 1157−1182. doi.:10.1162/153244303322753616

[14] 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. doi: 10.1016/j.biosystems.2008.12.004

[15] 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. doi: 10.1371/journal.pone.0012776

[16] Huynh-Thu, V. A.; Geurts, P. dynGENIE3: Dynamical GENIE3 for the inference of gene networks from time series expression data, *Scientific Reports*, **2018**, *8*, 3384. doi: 10.1038/s41598-018-21715-0

[17] Kauffman, S.A. Metabolic stability and epigenesis in randomly constructed genetic nets, *J. of Theoretical Biology*, **1969**, *22*, 437−467. doi: 10.1016/0022-5193(69)90015-0

[18] Kimura, S.; Sato, M.; Okada-Hatakeyama, M. Inference of Vohradský's models of genetic networks by solving two-dimensional function optimization problems, *PLoS One*, **2013**, *8*, e83308. doi: 10.1371/journal.pone.0083308

[19] Kimura, S.; Tokuhisa, M.; Okada, M. Inference of genetic networks using random forests: Assigning different weights for gene expression data, *J. Bioinform. Comput. Biol.*, **2019**, *17*, 1950015. doi: 10.1142/S021972001950015X

[20] Kimura, S.; Fukutomi, R.; Tokuhisa, M.; Okada, M. Inference of genetic networks from time-series and static gene expression data: Combining a random-forest-based inference method with feature selection methods, *Frontiers in Genetics*, **2020**, *11*, 595912. doi: 10.3389/fgene.2020.595912

[21] Kimura, S.; Sota, K.; Tokuhisa, M. Inference of genetic networks using random forests: A quantitative weighting method for gene expression data, *Proc. of the 2022 IEEE Conf. on Computational Intelligence in Bioinformatics and Computational Biology*, **2022**, 123−130. doi: 10.1109/CIBCB55180.2022.9863035

[22] Kishan, K. C.; Li, R.; Cui, F.; Yu, Q.; Haake, A. R. GNE: A deep learning framework for gene network inference by aggregating biological information, *BMC Syst. Biol.*, **2019**, *13*, 38. doi: 10.1186/s12918-019-0694-y

[23] Larrañaga, R.; Calvo, B.; Santana, R.; Bielza, C.; Galdiano, J.; *et al*. Machine learning in bioinformatics, *Briefings in Bioinformatics*, **2006**, 7, 86−112. doi: 10.1093/bib/bbk007

[24] Liu, F. T.; Ting, K. M.; Yu, Y.; Zhou, Z. H. Spectrum of variable-random trees, *J. Artif. Intell. Res.*, **2008**, *32*, 355-384. doi: 10.1613/jair.2470

[25] Liu, F.; Zhang, S. W.; Guo, W. F.; Wei, Z. G.; Chen, L. Inference of gene regulatory network based on local Bayesian networks, *PLoS Computational Biology*, **2016**, *12*, e1005024. doi: 10.1371/journal.pcbi.1005024

[26] Lu, D.; Chen, J.; Hai, T. The regulation of ATF3 gene expression by mitogen-activated protein kinases, *Biochemical J.*, **2007**, *401*, 559−567. doi: 10.1042/BJ20061081

[27] Maduranga, D. A. K.; Zheng, J.; Mundra, P. A.; Rajapakse, J. C. Inferring gene regulatory networks from time-series expression using random forests ensemble, *Pattern Recognition in Bioinformatics*, **2013**, 13−22. doi: 10.1007/978-3-642-39159-0_2

[28] Martine-Moreno, M.; O'Shea, T. M.; Zepecki, J. P.; Olaru, A.; Ness, J. K.; *et al*. Regulation of

peripheral myelination through transcriptional buffering of Egr2 by an etantisense long non-coding RNA, *Cell Reports*, **2017**, *20*, 1950−1963. doi: 10.1016/j.celrep.2017.07.068

[29] Petralia, F.; Wang, P.; Yang, J.; Tu, Z. Integrative random forest for gene regulatory network inference, *Bioinformatics*, **2015**, *31*, i197-i205. doi: 10.1093/bioinformatics/btv268

[30] Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. Numerical Recipes in C 2nd Edition, **1995**, Cambridge University Press, Cambridge, UK.

[31] Rubiolo, M.; Milone, D.H.; Stegmayer, G. Extreme learning machines for reverse engineering of gene regulatory networks from expression time series, *Bioinformatics*, **2018**, *34*, 1253−1260. doi: 10.1093/bioinformatics/btx730

[32] Schaffter, T.; Marbach, D.; Floreano, D. GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods, *Bioinformatics*, **2011**, *27*, 2263−2270. doi: 10.1093/bioinformatics/btr373

[33] Strobl, C.; Zeileis, A. Danger: High power! – Exploring the statistical properties of a test for random forest variable importance, *COMPSTAT 2008: Proc. in Computational Statistics*, **2008**, 2, 59−66.

[34] 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. doi: 10.1093/nar/gku1003

[35] Vatsa, D.; Agarwal, S. PEPN-GRN: A Petri net-based approach for the inference of gene regulatory networks from noisy gene expression data, *PLoS One*, **2021**, *16*, e0251666. doi: 10.1371/journal.pone.0251666

[36] Vohradský, J. Neural network model of gene expression, *FASEB J.*, **2001**, *15*, 846−854. doi: 10.1096/fj.00-0361com

[37] Weisenhorn, D.M.V.; Roback, L. J.; Kwon, J. H.; Wainer, B. H. Coupling of cAMP/PKA and MAPK signaling in neuronal cells is dependent on developmental stage, *Experimental Neurology*, **2001**, *169*, 44−55. doi: 10.1006/exnr.2001.7651

[38] 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.

[39] Yin, X.; Wolford, C. C.; Chang, Y. S.; McConoughey, S. J.; Ramsey, S. A.; *et al*. ATF3, an adaptive-response gene, enhances TGFβ signaling and cancer-initiating cell features in breast cancer cells, *J. Cell Sci.*, **2010**, *123*, 3558−3565. doi: 10.1242/jcs.064915

[40] Yu, J.; Smith, V. A.; Wang, P. P.; Hartemink, A. J.; Jarvis, E. D. Advances to Bayesian network inference for generating causal networks from observational biological data, *Bioinformatics*, **2004**, *20*, 3594−3603. doi: 10.1093/bioinformatics/bth448

[41] Yuan, G.; Qian, L.; Song, L.; Shi, M.; Li, D.; *et al.* Heregulin-β promotes matrix metalloproteinase-7 expression via HER2-mediated AP-1 activation in MCF-7 cells, *Mol. Cell. Biochem.*, **2008**, *318*, 73−79. doi: 10.1007/s11010-008-9858-6

[42] Yue, J.; Lai, F.; Beckedorff, F.; Zhang, A.; Pastori, C.; *et al.* Integrator orchestrates RAS/ERK1/2 signaling transcriptional programs, *Genes & Development*, **2017**, *31*, 1809−1820. doi: 10.1101/gad.301697.117