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

Shuhei Kimura (kimura@tottori-u.ac.jp)
Tottori University: Tottori Daigaku https://orcid.org/0000-0002-6246-2636
Yahiro Takeda
Tottori University: Tottori Daigaku
Masato Tokuhisa
Tottori University: Tottori Daigaku
Mariko Okada
Osaka University: Osaka Daigaku

Research

Keywords: genetic network inference, random forest, variable importance measure

Posted Date: July 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-737867/v1

RESEARCH

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]

[*]Correspondence:
kimura@tottori-u.ac.jp
[1]Faculty of Engineering, Tottori University, Tottori, Japan
Full list of author information is available at the end of the article

Abstract

Background: 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. We think however that this measure has drawbacks in the inference of genetic networks.

Results: 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).

Conclusion: This study proposed the random-input variable importance measure for the inference of genetic networks. The use of our measure improved the performance of the random-forest-based inference method. In this study, we checked the performance of the proposed measure only on several genetic network inference problems. However, the experimental results suggest that the proposed measure will work well in other applications of random forests.

Keywords: genetic network inference; random forest; variable importance measure

Introduction

Gene expression data have become prevalent with the development of high-throughput techniques in molecular biology such as DNA microarrays and RNA-seq using next-generation sequencers. These data implicitly contain enormous amounts of information on biological systems. Several researchers have taken an interest in the inference of genetic networks as a means of extracting useful information from the observed gene expression data. In a genetic network inference problem, mutual regulations among genes are inferred from the measured gene expression data. The inferred model of a genetic network is conceived as an ideal tool to help biologists generate hypotheses and facilitate the design of their experiments. The development of these methodologies has thus become one of the major topics in systems biology.

A number of genetic network inference methods have been proposed [1, 2, 3]. In this study, we narrow the focus to random-forest-based inference, a type of method confirmed to perform exceptionally well [4, 5, 6, 7, 8]. Some random-forest-based methods also have a useful ability to analyze both time-series and static gene expression data [5, 6, 8]. The 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 is 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. We then 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 uses 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.

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 [6]. We will begin, in this section, by explaining the method. Readers can find more detailed information on this method in our earlier paper [6].

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

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

where $\mathbf{X}_{-n} = (X_1, \cdots, X_{n-1}, X_{n+1}, \cdots, X_N)$, 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.

Obtaining F_n and β_n

The inference method [6] 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,

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

where $\mathbf{X}_{-n}|_{t_k} = (X_1|_{t_k}, \cdots, X_{n-1}|_{t_k}, X_{n+1}|_{t_k}, \cdots, X_N|_{t_k})$, $\mathbf{X}_{-n}|_{s_k} = (X_1|_{s_k}, \cdots, X_{n-1}|_{s_k}, X_{n+1}|_{s_k}, \cdots, X_N|_{s_k})$, and $X_m|_{t_k}$ and $X_m|_{s_k}$ 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, $X_m|_{t_k}$'s and $X_m|_{s_k}$'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 [6] 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.

$\hat{F}_n (\cdot ; \beta_n)$ 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 [6] obtains an approximation of the function F_n using the random forest [9]. 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,

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

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 method [6] 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.

The inference method described here uses the golden section search [10] to minimize the objective function (2).

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 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 [12], 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.

Standard variable importance measure

When we use the standard variable importance measure, we can compute the confidence value of the regulation of the n-th gene from the m-th gene, $C_{n,m}^S$, by

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

where

$$
\begin{aligned}
Sq_{w0} &= \sum_{k=1}^{K_T} w_k^T (y_{t_k} - \overline{y}_{w0})^2 + \sum_{k=1}^{K_S} w_k^S (y_{s_k} - \overline{y}_{w0})^2, && (4)\\
\overline{y}_{w0} &= \frac{1}{N_{w0}} \left[\sum_{k=1}^{K_T} w_k^T y_{t_k} + \sum_{k=1}^{K_S} w_k^S y_{s_k} \right], && (5)\\
N_{w0} &= \sum_{k=1}^{K_T} w_k^T + \sum_{k=1}^{K_S} w_k^S, && (6)\\
y_{t_k} &= \left. \frac{dX_n}{dt} \right|_{t_k} + \beta_n^* \, X_n|_{t_k}, && (7)\\
y_{s_k} &= \left. \frac{dX_n}{dt} \right|_{s_k} + \beta_n^* \, X_n|_{s_k}, && (8)\\
I(\nu) &= N_w(\nu) Sq_w(\nu) - N_w(\nu_L) Sq_w(\nu_L) \\
&\quad - N_w(\nu_R) Sq_w(\nu_R), && (9)\\
Sq_w(\nu) &= \sum_{k \in T(\nu)} w_k^T \left[y_{t_k} - \overline{y}_w(\nu) \right]^2 \\
&\quad + \sum_{k \in S(\nu)} w_k^S \left[y_{s_k} - \overline{y}_w(\nu) \right]^2, && (10)\\
\overline{y}_w(\nu) &= \frac{1}{N_w(\nu)} \left[\sum_{k \in T(\nu)} w_k^T y_{t_k} + \sum_{k \in S(\nu)} w_k^S y_{s_k} \right], && (11)\\
N_w(\nu) &= \sum_{k \in T(\nu)} w_k^T + \sum_{k \in S(\nu)} w_k^S, && (12)
\end{aligned}
$$

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 $\hat{F}_n^*$. ν_L and ν_R are the left and right children nodes of the node ν, respectively. $T(\nu)$ and $S(\nu)$ are sets of indices of the training examples generated from time-series and static gene expression data, respectively, and are allocated to the node ν. 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. When we try to compare the regulations whose confidence values are computed from a single random forest, this restriction 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.

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. Thus, the amount of fluctuation of the output caused by a change in a certain input variable could be used to evaluate the degree to which the variable contributes to the prediction of the output values.

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. When using this random-input variable importance measure, the confidence value of the regulation of the n-th gene from the m-th gene, $C^R_{n,m}$, is computed by

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

where

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

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

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

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

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

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

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, $C^R_{n,m}$'s, are computed N_{rnd} times by changing the random numbers, and their averages are used to rank the regulations.

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. With the sum of the confidence values unrestricted, we use the random-input variable importance measure in place of the standard measure for the computation of the confidence values.

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.

Analysis of DREAM3 networks

In this experiment, we compared the original random-forest-based inference method [6] 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.

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 [13], 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 [14], 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 each internal node of each tree, and the maximum height of each tree were set to 1000, $\lceil \frac{N-1}{3} \rceil$, and 32, respectively. We selected these parameter values according to those for the original inference method [6]. 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 [6], 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. 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.

Results

Table 1 lists performances of the original inference method [6] 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). The recall-precision curve of an algorithm was obtained by checking the recalls and precisions. The recall and precision are defined as

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

where TP, FP, and FN are the numbers of true-positive, false-positive, and false-negative regulations, respectively. The recall and precision were computed by constructing a network of regulations whose confidence values exceeded a threshold, and then comparing it with the gold-standard network. Next, the recall-precision curve of the algorithm was obtained by changing the threshold for the confidence value. 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.

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 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. In this study, we used the proposed measure only in the genetic network inference problems. If this feature also appears in other kinds of problems, the random-input variable importance measure would be a more useful technique for extracting information from a random forest.

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.

Experimental setup

In this section, we describe the application of the proposed approach to 5 problems from the DREAM4 *in silico* network challenges (http://dreamchallenges.org/). 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 [13].

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 regression [14] 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 [6], 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.

Results

In this experiment, we also used the area under the recall-precision curve (AURPC) to quantify the performance of the inference method. Although we inferred the regulations of the 100 genes from these genes and the 10 additional elements representing 10 perturbations, the regulations of the genes from the additional elements were disregarded in the evaluation of the performance. Further, the auto-regulations/auto-degradations were also disregarded. Table 2 shows the AURPCs of the proposed approach on the 5 problems, along with the AURPCs of the original random-forest-based inference method [6], dynGENIE3 [5], MCZ [15], a combination of dynGENIE3 and MCZ, and iRafNet [8]. The values of the AURPCs of dynGENIE3, MCZ, and the combination of dynGENIE3 and MCZ were taken from Huynh-Thu et al. [5], and the values of the AURPCs of iRafNet were taken from Petralia et al. [8].

The table shows that the proposed approach outperformed the original random-forest-based method [6] 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 [5] 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. 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 can improve the performances of the other random-forest-based inference methods.

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.

Experimental setup

In this experiment, we used the Vohradský's model [16] to describe target networks. The Vohradsky's model is a set of differential equations of the form

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

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 [6] on the constructed problems. These target networks were randomly constructed according to the procedure described in Kimura et al. [17]. 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 [6]: The weight values for the last 6 measurements in each of the time-series datasets were set to $1/(6 \times 10) \simeq 0.01667$; 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.

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.

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

Experiment with real gene expression data

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

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/) [18]. 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. Fig. 1 shows a sample of the time-series datasets. 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 102 and 10, 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 corresponded to a stimulus applied to the cells. According to Kimura et al. [19], 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.

The weight values in this experiment were set according to our earlier paper [6]. Specifically, the weight value settings were as follows: those corresponding to the 11th, 12th, 13th, 14th, 15th, and 16th measurements in the time-series dataset for the lymphatic endothelial cells were set to 0.75, 0.5, 0.25, 0.25, 0.25 and 0.25, respectively; those for the 8th, 9th, 10th, and 11th measurements in the time-series dataset for the Saos-2 cells, and for the 7th, 8th, 9th and 10th measurements in the two time-series datasets for the aortic smooth muscle cells, were commonly set to 0.8333, 0.6667, 0.5, and 0.5, respectively; those for the two measurements in the steady-state experiments with Saos-2, MCF-7, mesenchymal stem, and aortic smooth muscle cells were respectively set to 0.55; those for the other measurements in the time-series and static datasets were set to 1.0 and 1.1, respectively. The other experimental settings were identical to those used in the previous experiments.

Results

Table 3 shows the top 20 regulations with respect to the confidence values assigned by the proposed approach and the original inference method. As the random forest used in the inference methods is a stochastic algorithm, the confidence values assigned by the methods were slightly different in every trial. The table therefore shows the regulations ranked with respect to the confidence values averaged over 10 trials.

The correct structure of the target network however remains unknown. As a workaround, we compared the inferred regulations with protein-protein interactions obtained from the STRING database (https://string-db.org/) [20]. The comparison suggests that 11 of the 20 regulations (boldface in the table) are reasonable, as the interactions between the proteins corresponding to the genes have been confirmed in human and/or other species. The regulation of ATF3 from the external stimuli of 'TGF-β and TNF-α' (italic font in the table) also seems to be reasonable, as TGF-β has been confirmed to induce ATF3 [21]. The regulations of FOS and EGR2 from 'HRG' (italic font in the table) appeared to be reasonable, as well, given the suggestions from earlier reports [22, 23].

As shown in the table, the top 20 regulations obtained from the proposed approach and the original inference method were similar to each other. 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., [24, 25]). In addition, the interaction between the cAMP and MAPK signaling pathways has been confirmed [26]. 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.

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 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.

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 (e.g., [27, 28]).

Although we checked the performances of the existing measures on several genetic network inference problems, they sometimes failed to outperform the standard variable importance measure. While the random-input variable importance measure proposed in this study seems to be more effective, 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.

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

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
SK proposed the algorithm and performed the experiments. YT and MT contributed to design some parts of the proposed algorithm. MO supervised the biological aspect of this work. All authors read and approved the manuscript.

Author details
[1]Faculty of Engineering, Tottori University, Tottori, Japan. [2]Graduate School of Sustainability Sciences, Tottori University, Tottori, Japan. [3]Institute for Protein Research, Osaka University, Suita, Japan.

References

1. Chou, I.-C., Voit, E.: Recent developments in parameter estimation and structure identification of biochemical and genomic systems. Mathematical Biosciences **219**, 57–83 (2009)
2. Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., Guthke, R.: Gene regulatory network inference: data integration in dynamic models – a review. BioSystems **96**, 86–103 (2009)
3. Larrañaga, R., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J., Armañanzas, R., Santafé, G., Pérez, A., Robles, V.: Machine learning in bioinformatics. Briefings in Bioinformatics **7**, 86–112 (2006)
4. Huynh-Thu, V., Irrthum, A., Wehenkel, L., Geurts, P.: Inferring regulatory networks from expression data using tree-based methods. PLoS One **5**, 12776 (2010)
5. Huynh-Thu, V., Geurts, P.: dyngenie3: Dynamical genie3 for the inference of gene networks from time series expression data. Scientific Reports **8**, 3384 (2018)
6. Kimura, S., Tokuhisa, M., Okada, M.: Inference of genetic networks using random forests: assigning different weights for gene expression data. J. of Bioinformatics and Computational Biology **17**, 1950015 (2019)
7. Maduranga, D., Zheng, J., Mundra, P., Rajapakse, J.: Inferring gene regulatory networks from time-series expression using random forests ensemble. In: Ngom, A., Formenti, E., Hao, J.-K., Zhao, X.-M., van Laarhoven, T. (eds.) Pattern Recognition in Bioinformatics, pp. 13–22 (2013)
8. Petralia, F., Wang, P., Yang, J., Tu, Z.: Integrative random forest for gene regulatory network inference. Bioinformatics **31**, 197–205 (2015)
9. Breiman, L.: Random forests. Machine Learning **45**, 5–32 (2001)
10. Press, W., Teukolsky, S., Vetterling, W., Flannery, B.: Numerical Recipes in C 2nd Edition. Cambridge University Press, Cambridge (1995)
11. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine Learning **63**, 3–42 (2006)
12. Liu, F., Ting, K., Yu, Y., Zhou, Z.-H.: Spectrum of variable-random trees. J. of Artificial Intelligence Research **32**, 355–384 (2008)
13. Schaffter, T., Marbach, D., Floreano, D.: Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. Bioinformatics **27**, 2263–2270 (2011)
14. Cleveland, W.: Robust locally weight regression and smoothing scatterplots. J. of American Statistical Association **79**, 829–836 (1979)
15. Greenfield, A., Madar, A., Ostrer, H., Bonneau, R.: Dream4: Combining genetic and dynamic information to identify biological networks and dynamical models. PLoS One **5**, 13397 (2010)
16. Vohradský, J.: Neural network model of gene expression. FASEB J. **15**, 846–854 (2001)
17. Kimura, S., Sato, M., Okada-Hatakeyama, M.: Inference of vohradsky's models of genetic networks by solving two-dimensional function optimization problems. PLoS One **8**, 83308 (2013)
18. FANTOM Consortium, RIKEN PMI, CLST: A promoter-level mammalian expression atlas. Nature **507**, 462–470 (2014)
19. Kimura, S., Fukutomi, R., Tokuhisa, M., Okada, M.: Inference of genetic networks from time-series and static gene expression data: combinig a random-forest-based inference method with feature selection methods. Frontiers in Genetics **11**, 595912 (2020)
20. Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K., Kuhn, M., Bork, P., Jensen, L., von Mering, C.: String v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Research **43**, 447–452 (2015)
21. Yin, X., Wolford, C., Chang, Y.-S., McConoughey, S., Ramsey, S., Aderem, A., Hai, T.: Atf3, an adaptive-response gene, enhances tgfβ signaling and cancer-initiating cell features in breast cancer cells. J. of Cell Science **123**, 3558–3565 (2010)
22. Yuan, G., Qian, L., Song, L., Shi, M., Li, D., Yu, M., Hu, M., Shen, B., Guo, N.: Heregulin-β promotes matrix metalloproteinase-7 expression via her2-mediated ap-1 activation in mcf-7 cells. Molecular Cell Biochemistry **318**, 73–79 (2008)

23. Martine-Moreno, M., O'Shea, T., Zepecki, J., Olaru, A., Ness, J., Langer, R., Tapinos, N.: Regulation of peripheral myelination through transcriptional buffering of egr2 by an antisense long non-coding rna. Cell Reports **20**, 1950–1963 (2017)
24. Lu, D., Chen, J., Hai, T.: The regulation of atf3 gene expression by mitogen-activated protein kinases. Biochemical J. **401**, 559–567 (2007)
25. Yue, J., Lai, F., Beckedorff, F., Zhang, A., Pastori, C., Shiekhattar, S.: Integrator orchestrates ras/erk1/2 signaling transcriptional programs. Genes & Development **31**, 1809–1820 (2017)
26. Weisenhorn, D., Roback, L., Kwon, J., Wainer, B.: Coupling of camp/pka and mapk signaling in neuronal cells is dependent on developmental stage. Experimental Neurology **169**, 44–55 (2001)
27. Altmann, A., Tolosi, L., Sander, O., Lengauer, T.: Permutation importance: a corrected feature importance measure. Bioinformatics **26**, 1340–1347 (2010)
28. Archer, K., Kimes, R.: Empirical characterization of random forest variable importance measures. Computational Statistics & Data Analysis **52**, 2249–2260 (2008)

Figures

Figure 1 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.

Tables

Table 1 The performances of the proposed approach and the original random-forest-based inference method [6] on the DREAM3 problems. AVG and STD represent the averaged AURPC and its standard deviation, respectively.

	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 [6]	0.41918 ±0.00388	0.54477 ±0.00586	0.50083 ±0.00285	0.39482 ±0.00344	0.31291 ±0.00223

Table 2 The performances of the inference methods on the DREAM4 problems. The AURPCs of the proposed approach, the original random-forest-based inference method [6], dynGENIE3 [5], MCZ [15], a combination of dynGENIE3 and MCZ, and iRafNet [8] are shown.

	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 [6]	0.42797 ±0.00312	0.28656 ±0.00300	0.33930 ±0.00397	0.34079 ±0.00347	0.27199 ±0.00415
dynGENIE3 [5]	0.34 —	0.22 —	0.32 —	0.34 —	0.22 —
MCZ [15]	0.48 —	0.38 —	0.38 —	0.36 —	0.17 —
dynGENIE3 + MCZ	0.60 —	0.43 —	0.47 —	0.52 —	0.37 —
iRafNet [8]	0.552 —	0.337 —	0.414 —	0.421 —	0.298 —

Table 3 The top 20 regulations ranked with respect to the confidence values computed by the proposed approach and the original inference method [6]. 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.

Rank	Inference method using random-input variable importance measure	Random-forest-based inference method [6]
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**

Figures

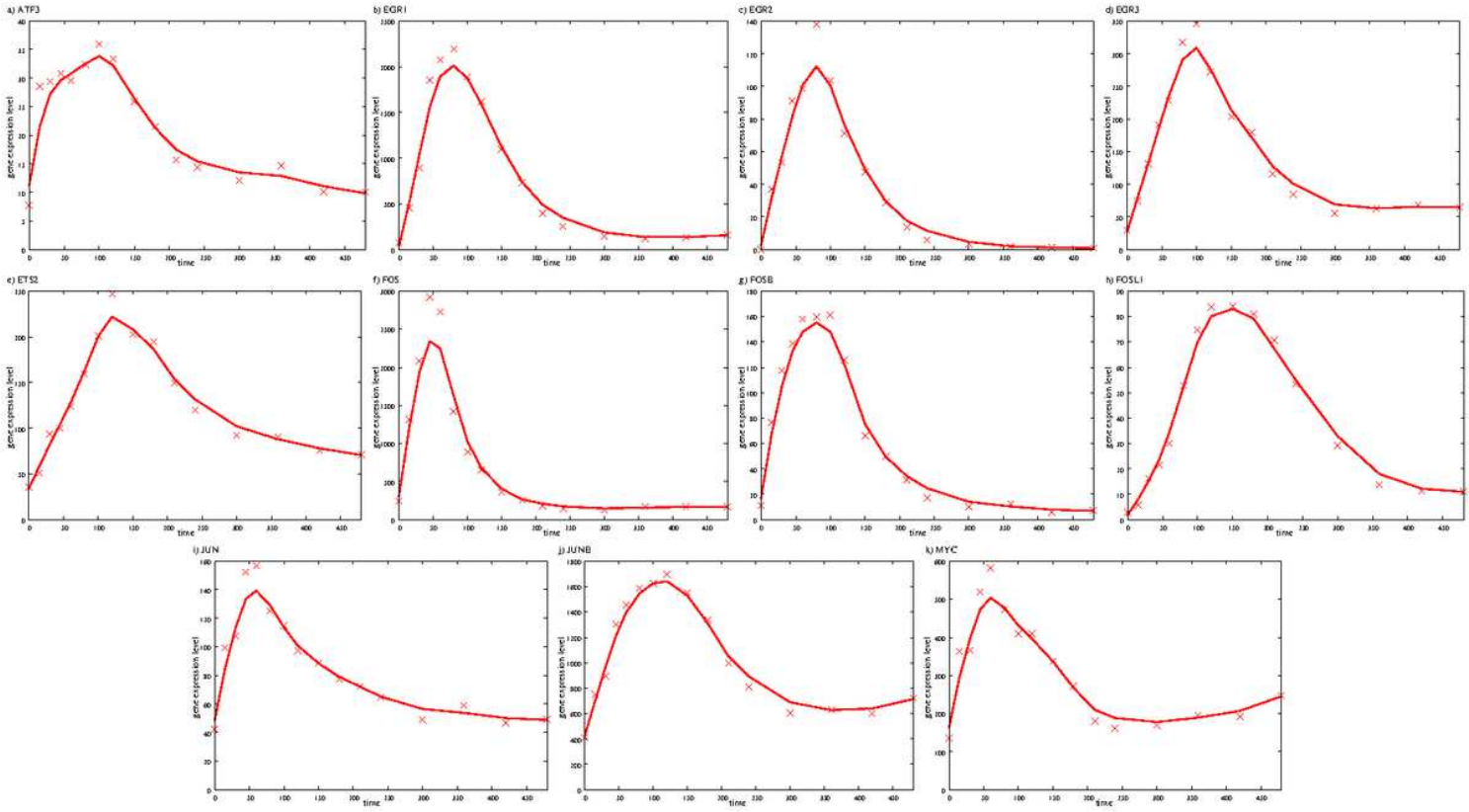

Figure 1

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.