Asymmetry in indegree and outdegree distributions of gene regulatory networks arising from dynamical robustness

著者	Ichinose Natsuhiro, Yada Tetsushi, Wada Hiroshi
journal or publication title	Physical review E
volume	97
number	6
page range	62315
year	2018-06
権利	(C)2018 American Physical Society
URL	http://hdl.handle.net/2241/00153095

doi: 10.1103/PhysRevE.97.062315

Asymmetry in indegree and outdegree distributions of gene regulatory networks arising from dynamical robustness

Natsuhiro Ichinose
Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan

Tetsushi Yada
Science and Systems Engineering, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka-shi, Fukuoka 820-8502, Japan

Hiroshi Wada
Graduate School of Life and Environmental Sciences, University of Tsukuba, Tennodai, Tsukuba 305-8672, Japan

(Received 4 April 2018; published 28 June 2018)

Although outdegree distributions of gene regulatory networks have scale-free characteristics similar to other biological networks, indegree distributions have single-scale characteristics with significantly lower variance than that of outdegree distributions. In this study, we mathematically explain that such asymmetric characteristics arise from dynamical robustness, which is the property of maintaining an equilibrium state of gene expressions against inevitable perturbations to the networks, such as gene dysfunction and mutation of promoters. We reveal that the expression of a single gene is robust to a perturbation for a large number of inputs and a small number of outputs. Applying these results to the networks, we also show that an equilibrium state of the networks is robust if the variance of the indegree distribution is low (i.e., single-scale characteristics) and that of the outdegree distribution is high (i.e., scale-free characteristics). These asymmetric characteristics are conserved across a wide range of species, from bacteria to humans.

DOI: 10.1103/PhysRevE.97.062315

I. INTRODUCTION

Many biological networks are scale-free (or scale-free-like) networks, where the number of links originating from a given node exhibits a power-law distribution [1,2]. For example, metabolic networks [3,4], protein-protein interaction networks [5,6], and gene coexpression networks [7] can be characterized as scale-free networks. Metabolic networks are generated by the chemical reactions of metabolites in a cell. These networks are represented by directed graphs corresponding to the direction of the chemical reaction. In directed graphs, the indegree and outdegree distributions can differ from each other. However, it is known that there are no differences between the indegree and outdegree distributions of metabolic networks, and both distributions have scale-free characteristics or at least broad-tailed characteristics [2,8]. In protein-protein interaction networks and gene coexpression networks, no differences are observed between the indegree and outdegree distributions because these networks are represented by undirected graphs. Therefore, their scale-free characteristics are also symmetric.

Among such biological networks, gene regulatory networks are unique in that the characteristics of the indegree and outdegree distributions differ from each other [1]. Gene regulatory networks govern the expression of genes, which encode various proteins. The expression of a gene is regulated by transcription factors. Transcription factors bind a promoter, which is the regulatory region of the gene, thereby enhancing (or repressing) the transcription of the gene to a messenger RNA (mRNA). Then, a protein translated from the mRNA performs a biological function in the cell. A transcription factor by itself is a type of protein. Therefore, the regulatory relationship leads to the construction of interconnected networks. Here, the outdegree distribution is determined by the number of promoters bounded by a transcription factor. It is known that the outdegree distribution is a power-law distribution [1,2]. The indegree distribution is determined by the number of transcription factors that bind a promoter. In contrast to the outdegree distribution, in *Escherichia coli* and yeast, the indegree distribution is an exponential distribution, which has lower variance than that of a power-law distribution [9,10].

As compared to scale-free networks that are characterized by a power-law distribution with high variance, networks having a low-variance degree distribution are referred to as single-scale networks [11]. The variance of the power-law distribution infinitely diverges in a certain range of the parameter. Therefore, we cannot have prior information about the scale of the number of connections in a gene. The word "scale-free" is derived from the absence of the scale of the number of connections [12]. Conversely, the variances of Poisson and exponential distributions have linear and square orders, respectively, for the mean values. Such a degree distribution is referred to as "single-scale" [11], because the scale of the number of connections can be estimated from the mean value. In other words, gene regulatory networks have single-scale characteristics for the indegree distribution, and scale-free characteristics for the outdegree distribution.

We hypothesize that these asymmetric characteristics of indegree and outdegree distributions are derived from dynamical

robustness of the gene regulatory networks [1]. Robustness of networks can be categorized into topological and dynamical robustness [1]. As regards topological robustness, the changes of interconnectedness of the networks to a node dysfunction are discussed. The interconnectedness is measured by a network diameter, which is the mean of the shortest path between any two nodes. Thus, networks are topologically robust if the network diameter is relatively unchanged when a node is deleted. Moreover, it is known that scale-free networks are topologically robust [13]. However, networks are dynamically robust if there are less changes of the internal states when a node is perturbed. For example, in gene regulatory networks, dynamical robustness implies the maintenance of an equilibrium state of gene expressions against perturbations, such as gene dysfunction and mutation of promoters [14,15]. The biological role of the gene regulatory networks is to express a specific gene at an appropriate place and time. As long as such appropriate gene expressions are achieved, gene regulatory networks do not necessarily maintain the network topology. Therefore, dynamical robustness is more important than topological robustness to understand gene regulatory networks. Although the two types of robustness are mutually related, the optimal network structures required to attain both types of robustness are not necessarily identical. Therefore, to understand the structure of the gene regulatory networks, we require an alternative approach of dynamical robustness that differs from that of topological robustness.

In developmental biology, dynamical robustness is referred to as "canalization." This term was proposed by Waddington to explain the fact that phenotypic variation is low despite the abundance of genetic and environmental perturbations in the development of organisms [16]. In gene regulatory networks, such phenotypic constancy corresponds to the maintenance of an equilibrium state of gene expressions. Wagner proposed a model of gene regulatory networks to evaluate canalization and showed that highly connected networks are more robust [17]. Based on Wagner's work, several studies focused on the relationship between canalization and properties of gene regulatory networks, such as connection density [18,19], network topology [20,21], and mutation rate [22]. However, most of these studies were conducted based on dynamical simulation of the networks. Therefore, the theoretical mechanism of dynamical robustness or canalization to explain the structural properties of gene regulatory networks remains elusive.

This work attempts to reveal the mechanism underlying the asymmetric characteristics of indegree and outdegree distributions by proposing a mathematical framework to analyze the dynamical robustness. In Sec. II, using a simple neuron-like model, we discuss the effect of the number of connections on the dynamical robustness in a single gene. We show that the single gene is robust if the number of inputs is large and the number of outputs is small. Based on these results, in Sec. III, we discuss the effect of degree distributions on the dynamical robustness in the networks. We show that the networks are robust if the indegree distribution has low variance and the outdegree distribution has high variance. In Sec. IV, we analyze the experimental distributions of *E. coli*, fruit fly, and human genes, and we show that the asymmetric characteristics are conserved across a wide range of species.

II. ROBUSTNESS OF A SINGLE GENE

A. Gene regulatory network model

Based on Wagner's work [17], we adopt the following model for gene regulatory networks:

$$s_i(t+1) = \mathrm{sgn}[x_i(t)], \quad x_i(t) = \sum_{j=1}^{N} w_{ij} s_j(t), \tag{1}$$

where t is the discrete time, N is the number of genes, s_i is the expression state of the ith gene, x_i is the internal state of the ith gene, w_{ij} is the connection weight from the jth gene to the ith gene, and sgn is the signum function:

$$\mathrm{sgn}(x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x \geqslant 0 \end{cases}. \tag{2}$$

The expression state of a gene can have two values corresponding to gene expression ($s_i = 1$) and resting ($s_i = -1$). We refer to the vector of the expression states $s = (s_1, s_2, \ldots, s_N)^T$ as the expression pattern. The connection weight represents the strength of transcriptional regulation, in which a positive value ($w_{ij} > 0$) indicates transcriptional activation and a negative value ($w_{ij} < 0$) implies repression. In particular, $w_{ij} = 0$ implies that there is no connection from the jth gene to the ith gene.

In actual gene regulatory networks, the connection weight w_{ij} indicates that the ith gene has the binding motif of the jth gene in the regulatory region (promoter) in the genome. Moreover, the genomes of all cells in an individual are (almost) identical; therefore, the gene regulatory networks are essentially identical (however, connection strength can change by the effect of epigenetics [23]).

Many biological factors except transcription factors are involved in gene expressions, such as transcription initiation, transcription termination, translation, and post-translational modifications. In addition, other molecules such as metabolites and microRNAs can also regulate gene expressions. However, in our model, we focus only on the networks of transcription factors. This is because our purpose is to analyze the degree distributions of transcription-factor regulatory networks. The other important point is that we assume equilibrium states as dynamics of gene regulatory networks, as shown in the subsequent section. In an equilibrium state, it is expectable that each factor is fixed at a certain level. In this case, we can exclude the factors except transcription factors from the analysis because their effects on the dynamics are constant. Therefore, we can focus only on transcription factors to analyze the dynamics of gene regulatory networks.

B. Modelling equilibrium states

Gene regulatory networks have multistable equilibrium states [24]. The multistability of the gene regulatory networks can explain cell differentiation to distinct cell types in multicellular organisms. In unicellular organisms, it is known that the system of lactose metabolism (the *lac* operon) has bistable states [25]. As regards the multistability of networks, there are several studies in the field of neural networks such as associative memory [26–29]. In the associative memory model, based on Hebbian learning [30], the connection weights are

determined by correlation learning, which uses *a priori* known patterns. However, in gene regulatory networks, methods determining the connection weights based on specific patterns cannot be adopted because gene regulatory networks do not "learn" expression patterns. Instead of specific-pattern-based methods, in gene regulatory networks, several works have randomly determined the connection weights [18]. However, in this case, we cannot obtain the expression pattern in the equilibrium state directly from a specific set of the connection weights. Thus, we are required to obtain the equilibrium state by iterating Eq. (1) from a certain initial pattern. To avoid this problem, we consider the correlation between the connection weight and the expression pattern in an assumed equilibrium state.

Let $s^* = (s_1^*, s_2^*, \ldots, s_N^*)$ be the expression pattern in the equilibrium state, in which the following equations hold:

$$s_i^* = \mathrm{sgn}(x_i^*), \quad x_i^* = \sum_{j=1}^{N} w_{ij} s_j^*. \tag{3}$$

We consider the following characteristic value u_i:

$$u_i = s_i^* x_i^* = s_i^* \sum_{j=1}^{N} w_{ij} s_j^*. \tag{4}$$

By the definition of sgn, x_i^* and s_i^* have the same sign. Therefore, u_i is always nonnegative ($u_i \geqslant 0$) [31]. We decompose the summation in Eq. (4) into each element:

$$v_{ij} = s_i^* w_{ij} s_j^*. \tag{5}$$

Since u_i is nonnegative, v_{ij} tends to be nonnegative with high probability if there is no excessive outlier in w_{ij}. The correlation between the connection weight and the expression pattern is defined by the probability ρ that v_{ij} is nonnegative:

$$\rho = P(v_{ij} \geqslant 0 | w_{ij} \neq 0). \tag{6}$$

As mentioned earlier, we cannot obtain the expression pattern in the equilibrium state directly from the given connection weights. In this work, instead of the expression pattern, we assume that the correlation ρ is known. If ρ is known, then we can evaluate the equilibrium state without obtaining the expression pattern, as will be in the subsequent sections.

Let the connection weight be generated by the uniform random distribution in the closed interval $[-1,1]$. We assume that the sign of v_{ij} is approximately independent of $|w_{ij}|$. Then, the density distribution of v_{ij} is represented by the mixture distribution with the probability ρ:

$$P(v_{ij} | w_{ij} \neq 0) \approx \begin{cases} 1-\rho & \text{if } -1 \leqslant v_{ij} < 0 \\ \rho & \text{if } 0 \leqslant v_{ij} \leqslant 1 \\ 0 & \text{otherwise} \end{cases}. \tag{7}$$

The mean μ_v and the variance σ_v^2 of v_{ij} are represented as follows:

$$\mu_v = \rho - \tfrac{1}{2}, \quad \sigma_v^2 = \tfrac{1}{12} + \rho - \rho^2. \tag{8}$$

u_i is the sum of v_{ij}, i.e., $u_i = \sum_{j=1}^{N} v_{ij}$. Let k_{in} be the number of input connections that have nonzero w_{ij}. By applying normal approximation, the density distribution of the sum of v_{ij} can be represented by the normal distribution with mean $k_{\mathrm{in}}\mu_v$ and variance $k_{\mathrm{in}}\sigma_v^2$. However, by the condition of the equilibrium state, it is known that $u_i \geqslant 0$. Thus, in all possible sums of v_{ij}, we observe only the positive realization values of u_i. Therefore, the density distribution of u_i is approximately represented by the following truncated normal distribution:

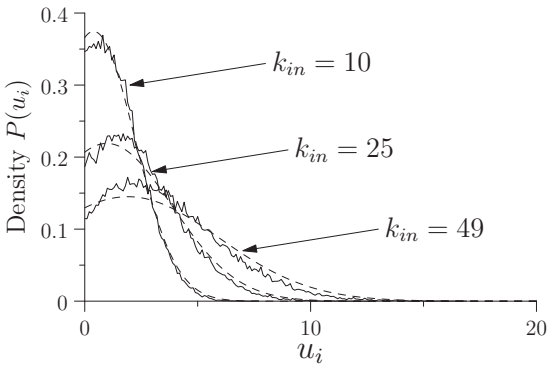

FIG. 1. Density distribution of u_i. The solid lines indicate the simulated distributions, and the dashed lines indicate the theoretical distributions by Eq. (10). Each simulated distribution is calculated by 1000 networks. Here, $N = 50$, where N is the number of genes. We show the distributions for the following numbers of input connections $k_{\mathrm{in}} = 10$, 25, and 49. The correlation ρ is estimated by the mode of the simulated distribution with $k_{\mathrm{in}} = 49$ ($\rho = 0.539$).

$$P(u_i) = P\left(\sum_{j=1}^{N} v_{ij} \middle| \sum_{j=1}^{N} v_{ij} \geqslant 0 \right) \tag{9}$$

$$\approx \begin{cases} 0 & \text{if } u_i < 0 \\ \dfrac{\phi\left(u_i; k_{\mathrm{in}}\mu_v, k_{\mathrm{in}}\sigma_v^2\right)}{1 - \Phi\left(0; k_{\mathrm{in}}\mu_v, k_{\mathrm{in}}\sigma_v^2\right)} & \text{if } u_i \geqslant 0 \end{cases}, \tag{10}$$

where $\phi(u; \mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2, and Φ is the cumulative density distribution of ϕ.

We evaluate the approximation accuracy of the density distribution $P(u_i)$ by simulation. In the simulation, we randomly select k_{in} of $(N-1)$ possible connections in each gene (not including the self-connection w_{ii}). We determine the value of the selected connection w_{ij} by the uniform random in $[-1,1]$. We obtain an equilibrium state by iterating Eq. (1) from a random initial state. The networks in which an equilibrium state cannot be obtained in 1000 iterations are excluded. The simulated distribution of u_i is calculated by using the 1000 networks obtained by this procedure.

The correlation ρ used in the theoretical distribution of Eq. (10) is estimated from the simulated distribution. Note that ρ cannot be observed directly from simulated v_{ij} although ρ is the probability of nonnegative v_{ij}. This is because simulated v_{ij} has a bias because $u_i \geqslant 0$. Since the mode of the density distribution is equivalent to $k_{\mathrm{in}}\mu_v = k_{\mathrm{in}}(\rho - 1/2)$ as the property of the truncated normal distribution (if $k_{\mathrm{in}}\mu_v \geqslant 0$), we estimate ρ from the mode of the simulated distribution with full connections. As shown in Fig. 1, the theoretical distribution of Eq. (10) sufficiently follows the simulated distribution with changes of k_{in}.

Despite the continuous-time nature of gene regulatory networks, we adopt the discrete-time model. However, in equilibrium states, both the models have equivalent characteristics. For example, let us consider the following continuous-time model:

$$\frac{ds_i(t)}{dt} = -\gamma s_i(t) + f[x_i(t)], \quad x_i(t) = \sum_{j=1}^{N} w_{ij} s_j(t), \tag{11}$$

where t is the continuous time, γ is the degradation rate, and f is a nonlinear function. Then, in an equilibrium state (i.e., $\frac{ds_i(t)}{dt} = 0$), we obtain the following relation:

$$s_i^* = f(x_i^*)/\gamma, \quad x_i^* = \sum_{j=1}^{N} w_{ij} s_j^*. \tag{12}$$

In other words, the characteristic of the continuous-time model is equivalent to that of the discrete-time model shown in Eq. (3) if $f/\gamma = \mathrm{sgn}$. Therefore, the following analysis can be applied to both the discrete-time and continuous-time models.

C. Robustness for the number of input connections

An expression flip derived from a gene dysfunction and a change of the connection weight derived from a mutation of promoter are considered as perturbations to the gene regulatory networks. In the former, when the expression of the jth gene with nonzero w_{ij} is flipped, the value of v_{ij} changes to $-v_{ij}$. Therefore, the characteristic value u_i' after the perturbation is represented by the following equation [14]:

$$u_i' = u_i - 2v_{ij}. \tag{13}$$

In the latter, when the nonzero connection w_{ij} is lost, the value of v_{ij} changes into zero. Therefore, the characteristic value u_i'' after the perturbation is represented as follows:

$$u_i'' = u_i - v_{ij}. \tag{14}$$

Since Eqs. (13) and (14) have a similar tendency, we focus on Eq. (13), which has a greater perturbation than Eq. (14).

The robustness for the number of input connections can be defined as the probability that the expression pattern does not change after the perturbation. This implies that u_i' is nonnegative similar to u_i. The probability $P(u_i' \geqslant 0)$ is dependent on k_{in} and its higher value implies that the gene is more robust. We assume that u_i and v_{ij} are approximately independent of each other. Then, using Eqs. (7) and (10), the robustness $r_{\mathrm{in}}(k_{\mathrm{in}}) = P(u_i' \geqslant 0)$ is represented by the integral of the convolution between u_i and v_{ij}, which is given as follows:

$$r_{\mathrm{in}}(k_{\mathrm{in}}) \approx \int_{u_i' = u_i - 2v_{ij} \geqslant 0} P(u_i) P(v_{ij} | w_{ij} \neq 0) du_i dv_{ij}, \tag{15}$$

$$= \int_{-\infty}^{\infty} \int_{2v_{ij}}^{\infty} P(u_i) P(v_{ij} | w_{ij} \neq 0) du_i dv_{ij}, \tag{16}$$

$$\approx \int_{0}^{1} \int_{2v_{ij}}^{\infty} \rho P(u_i) du_i dv_{ij} + \int_{-1}^{0} (1-\rho) dv_{ij}, \tag{17}$$

$$\approx \frac{\rho}{1-\Phi_v(0)} \left[1 - \int_0^1 \Phi_v(2v_{ij}) dv_{ij} \right] + (1-\rho), \tag{18}$$

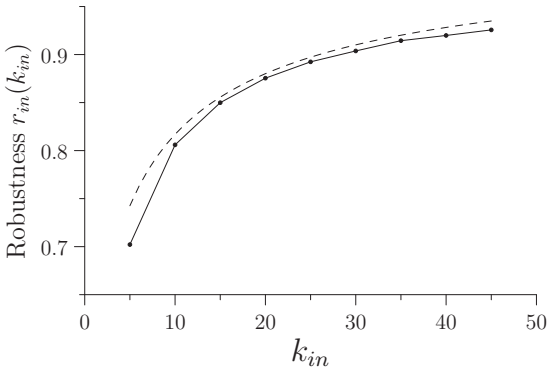

FIG. 2. Robustness $r_{\mathrm{in}}(k_{\mathrm{in}})$ for the number of input connections k_{in}. The solid line indicates the simulation results and the dashed line indicates the theoretical values of Eq. (18). Each measure is calculated by 1000 networks ($N = 50$, $\rho = 0.539$).

where $\Phi_v(u) = \Phi(u; k_{\mathrm{in}}\mu_v, k_{\mathrm{in}}\sigma_v^2)$. Equations (17) and (18) are obtained by applying Eqs. (7) and (10) to Eq. (16), respectively. In the second term of Eq. (17), we use $\int_v^{\infty} P(u_i) du_i = 1$ (if $v \leqslant 0$). The integral part of Eq. (18) is represented as follows:

$$\int_0^1 \Phi_v(2v_{ij}) dv_{ij} = \frac{1}{2} \{ 2\Phi_v(2) - k_{\mathrm{in}}\mu_v [\Phi_v(2) - \Phi_v(0)] + k_{\mathrm{in}}\sigma_v^2 [\phi_v(2) - \phi_v(0)] \}, \tag{19}$$

where $\phi_v(u) = \phi(u; k_{\mathrm{in}}\mu_v, k_{\mathrm{in}}\sigma_v^2)$.

Figure 2 shows the theoretical values of Eq. (18) and the simulation results. The simulation procedure is the same as that shown in Fig. 1. In the simulation, we calculate the probability that $u_i' \geqslant 0$ (i.e., $u_i - 2v_{ij} \geqslant 0$) for nonzero w_{ij}. In the simulation results, the inaccuracy of ρ leads to the slight overestimation of the theoretical values. The second term of Eq. (18) indicates that ρ contributes to a bias of r_{in}, and hence r_{in} is over- or underestimated. However, with the exception of the bias, r_{in} sufficiently follows the simulation results for changes of k_{in}. The results indicate that the gene with the larger number of input connections is more robust. Intuitively, this implies that the effect of the perturbation for one connection is smaller if the number of connections are larger.

D. Robustness for the number of output connections

Similar to the input connections, we consider an expression flip by a dysfunction of the jth gene. The flip of s_j^* affects all ith genes with nonzero w_{ij}. Thus, the robustness for the number of output connections can be defined as the probability that the expression states of all genes originating from the jth gene do not change after the perturbation. We assume that the number of input connections of the genes originating from the jth gene is identical (we discuss this assumption in the following section). Let k_{out} be the number of output connections of the gene. Then, the robustness $r_{\mathrm{out}}(k_{\mathrm{out}})$ is represented as follows:

$$r_{\mathrm{out}}(k_{\mathrm{out}}) = \prod_{i \in \{i' | w_{i'j} \neq 0\}} P(u_i' \geqslant 0) = [r_{\mathrm{in}}(k_{\mathrm{in}})]^{k_{\mathrm{out}}}. \tag{20}$$

We evaluate the theoretical values of Eq. (20) using simulation, as shown in Fig. 3. The simulation procedure is the same

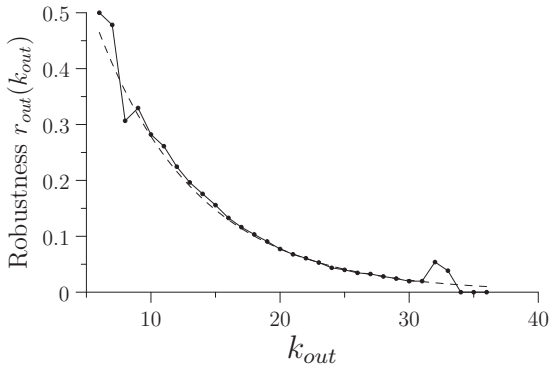

FIG. 3. Robustness $r_{\rm out}(k_{\rm out})$ for the number of output connections $k_{\rm out}$. The solid line indicates the simulation results and the dashed line indicates the theoretical values of Eq. (20). All measures are calculated by 10 000 networks ($N = 50$, $k_{\rm in} = 20$, $r_{\rm in}(20) = 0.880$).

as that illustrated in Fig. 1. However, the number of input connections is fixed at $k_{\rm in} = 20$. Therefore, 10 000 networks are used to increase the number of measures. The genes are sorted based on their $k_{\rm out}$ because the number of output connections has various values as a consequence of random selections. We calculate the probability that $u_i' \geqslant 0$ (i.e., $u_i - 2v_{ij} \geqslant 0$) holds at the same time for all i with nonzero w_{ij} in each $k_{\rm out}$. In Fig. 3, since $k_{\rm out}$ is Poisson distributed, the errors are remarkable at the tails of the distribution where the number of data is low. However, $r_{\rm out}$ sufficiently follows the simulation results in the wide region around the distribution mean.

In contrast to $r_{\rm in}$, the results of $r_{\rm out}$ indicate that the gene with the smaller number of output connections is more robust. Intuitively, this implies that the effect of a gene dysfunction on the other genes is limited if the gene has a small number of output connections. The asymmetric characteristics of the gene regulatory networks are attributed to the fact that the robustness for the number of input connections is contrary to that for the number of output connections.

III. ROBUST DEGREE DISTRIBUTIONS

A. Network robustness

As mentioned in the previous section, the gene with a large number of input connections and a small number of output connections is robust. However, these results do not directly determine the degree distributions of the gene regulatory networks. To determine the degree distributions, we use the fact that the total number of input connections is necessarily equivalent to that of output connections. This is because an edge comprises a pair of input and output connections. Therefore, the total number of input connections cannot increase or decrease independently of that of output connections, and vice versa. Thus, we assume that the mean values of the numbers of input and output connections ($\mu_{\rm in}$ and $\mu_{\rm out}$) are fixed. $\mu_{\rm in}$ is the mean value for all genes; however, $\mu_{\rm out}$ is defined by the genes that have the number of output connections greater than zero. This is attributed to the fact that all genes have input connections, but the genes, except transcription factors, have no output connections.

We assume that a gene dysfunction or a mutation of promoter occurs independently of the gene properties. This implies that the occurrence probability of these perturbations is uniform for all genes. Therefore, we define network robustness by the expectation value of the robustness for all genes. The network robustness for the indegree and outdegree distributions ($R_{\rm in}$ and $R_{\rm out}$) are represented as follows:

$$R_{\rm in} = E[r_{\rm in}(k_{\rm in})], \quad R_{\rm out} = E[r_{\rm out}(k_{\rm out})], \tag{21}$$

where $k_{\rm in}$ and $k_{\rm out}$ are randomly distributed with mean values $\mu_{\rm in}$ and $\mu_{\rm out}$, respectively.

B. Optimal distributions for network robustness

In the previous section, $r_{\rm in}$ and $r_{\rm out}$ are characterized as monotonically increasing and decreasing functions, respectively. At the same time, $r_{\rm in}$ and $r_{\rm out}$ are also concave and convex functions, respectively. $r_{\rm in}$ is a concave function if the following inequality holds:

$$r_{\rm in}(k-1) + r_{\rm in}(k+1) \leqslant 2r_{\rm in}(k). \tag{22}$$

$r_{\rm out}$ is a convex function if the following inequality holds:

$$r_{\rm out}(k-1) + r_{\rm out}(k+1) \geqslant 2r_{\rm out}(k). \tag{23}$$

$r_{\rm out}$ is a convex function because it can be regarded as an exponential function with a positive base. In a certain region of parameters ($\rho > 0.5$, $k_{\rm in} \leqslant 10^5$), we numerically confirm that $r_{\rm in}$ is a concave function. When determining the optimal distribution for network robustness, the convexity of $r_{\rm in}$ and $r_{\rm out}$ is more important than the monotonicity of $r_{\rm in}$ and $r_{\rm out}$.

The network robustness is defined by the expectation values of $r_{\rm in}$ and $r_{\rm out}$. Jensen's inequality describes the relationship between an expectation value and convexity [32,33]. For the concave function $r_{\rm in}$, Jensen's inequality is represented as follows:

$$R_{\rm in} = E[r_{\rm in}(k_{\rm in})] \leqslant r_{\rm in}(\mu_{\rm in}) = R_{\rm in}^{\rm opt}. \tag{24}$$

This inequality indicates that the network robustness $R_{\rm in}$ for the indegree distribution cannot exceed the robustness $r_{\rm in}(\mu_{\rm in})$. In addition, we assume that $\mu_{\rm in}$ is fixed, and hence $r_{\rm in}(\mu_{\rm in})$ is also fixed. Therefore, $r_{\rm in}(\mu_{\rm in})$ is the optimal robustness $R_{\rm in}^{\rm opt}$. The optimal robustness $R_{\rm in}^{\rm opt}$ is obtained by the networks in which the numbers of input connections for all genes are identical. Consequently, the optimal indegree distribution can be represented as follows:

$$P(k_{\rm in}) = \begin{cases} 1 & \text{if } k_{\rm in} = \mu_{\rm in} \\ 0 & \text{otherwise} \end{cases}. \tag{25}$$

In Sec. II D, we derive the robustness $r_{\rm out}$ of a single gene based on the assumption that $k_{\rm in}$ is fixed. This assumption is appropriate because it is equivalent to the optimal indegree distribution of Eq. (25).

In the network robustness for the outdegree distribution, we obtain the optimal value $R_{\rm out}^{\rm opt}$ of the convex function $r_{\rm out}$ by using converse Jensen's inequality [33]:

$$R_{\rm out} = E[r_{\rm out}(k_{\rm out})] \leqslant p\, r_{\rm out}(1) + (1-p) r_{\rm out}(k_{\rm max}) = R_{\rm out}^{\rm opt}, \tag{26}$$

where $k_{\max}$ is the maximum number of output connections, and p is defined as

$$p = \frac{k_{\max} - \mu_{\text{out}}}{k_{\max} - 1}. \tag{27}$$

Although the boundary is not considered in general degree distributions, in our case, the upper bound $k_{\max}$ is necessary to obtain the optimal value. In the actual gene regulatory networks, the upper bound exists because the number of output connections does not exceed the number of genes N. Since it is assumed that μ_{out} is fixed, $R_{\text{out}}^{\text{opt}}$ is also fixed and can be regarded as the optimal value. By using Eq. (26), the optimal outdegree distribution is obtained as follows:

$$P(k_{\text{out}}) = \begin{cases} p & \text{if } k_{\text{out}} = 1 \\ 1-p & \text{if } k_{\text{out}} = k_{\max} \\ 0 & \text{otherwise} \end{cases}. \tag{28}$$

If $\mu_{\text{out}} \ll k_{\max}$, then the optimal outdegree distribution is obtained by the networks in which almost all genes have a single output and a small number of genes have a large number of outputs.

The proposed method cannot quantitatively explain the robustness of the actual gene regulatory networks because r_{in} and r_{out} are derived from the qualitative model. In addition, we assume equilibrium states as the dynamics, although oscillatory dynamics such as circadian rhythm can appear in the actual gene regulatory networks. Nevertheless, we consider that the analytic results are valid in the actual gene regulatory networks, because the optimal distributions of Eqs. (25) and (28) are derived only from the convexity of r_{in} and r_{out}. In other words, even though r_{in} and r_{out} do not quantitatively agree with the actual characteristics due to the model simplification or the oscillatory dynamics, the results are applicable if the convexity of r_{in} and r_{out} holds. Therefore, the optimal distributions proposed in this work are widely acceptable independently of our assumptions for the model.

C. Evaluation of network robustness by connection variance

The optimal distributions of Eqs. (25) and (28) are unrealizable in actual networks. Therefore, we require an indicator to evaluate the network robustness for realizable networks. We adopt Hölder's defect formula, which evaluates the difference between both sides of the Jensen's inequality [34].

For the concave function r_{in}, the Hölder's defect formula is represented as follows:

$$\frac{c_{\min}}{2}\sigma_{\text{in}}^2 \leqslant R_{\text{in}}^{\text{opt}} - R_{\text{in}} \leqslant \frac{c_{\max}}{2}\sigma_{\text{in}}^2, \tag{29}$$

where σ_{in}^2 is the variance of the indegree distribution, and $c_{\min}$ and $c_{\max}$ are the minimum and maximum values of negative of the second difference δ_{in}, respectively:

$$c_{\min} = \min_k[-\delta_{\text{in}}(k)], \quad c_{\max} = \max_k[-\delta_{\text{in}}(k)], \tag{30}$$

where δ_{in} is defined by

$$\delta_{\text{in}}(k) = r_{\text{in}}(k-1) + r_{\text{in}}(k+1) - 2r_{\text{in}}(k). \tag{31}$$

By the definition of the concave function [Eq. (22)], δ_{in} is nonpositive for all k. Therefore, $c_{\max}$ and $c_{\min}$ are nonnegative.

TABLE I. Examples of degree distributions. Here, ζ is the Riemann ζ function.

Degree distribution	Variance with mean $= \mu$
Power-law distribution	
$P(k) = \frac{k^{-\alpha}}{\zeta(\alpha)}$	$\sigma_{\text{pow}}^2 = \infty \quad (\text{if } \alpha \leqslant 3)$
Exponential (geometric) distribution	
$P(k) = \gamma(1-\gamma)^{k-1}$	$\sigma_{\text{exp}}^2 = \mu(\mu-1)$
Poisson distribution	
$P(k) = \frac{\lambda^k}{k!}e^{-\lambda}$	$\sigma_{\text{poi}}^2 = \mu$

Equation (29) shows that the deviation of the network robustness from the optimal value can be evaluated by the connection variance. In other words, a lower variance of the indegree distribution is indicative of higher robustness. If $\sigma_{\text{in}}^2 = 0$, then the equality of Eq. (29) holds and the network robustness is optimal [i.e., the optimal indegree distribution of Eq. (25)].

For the converse Jensen's inequality in Eq. (26), there exists an equation similar to the Hölder's defect formula [33], which is given as follows:

$$\frac{C_{\min}}{2}\left(\sigma_{\max}^2 - \sigma_{\text{out}}^2\right) \leqslant R_{\text{out}}^{\text{opt}} - R_{\text{out}} \leqslant \frac{C_{\max}}{2}\left(\sigma_{\max}^2 - \sigma_{\text{out}}^2\right), \tag{32}$$

where $\sigma_{\max}^2$ is the maximum variance of bounded distributions and σ_{out}^2 is the variance of the outdegree distribution. $C_{\min}$ and $C_{\max}$ are the minimum and maximum values of the second difference δ_{out}, respectively, and are given as follows:

$$C_{\min} = \min_k[\delta_{\text{out}}(k)], \quad C_{\max} = \max_k[\delta_{\text{out}}(k)], \tag{33}$$

where δ_{out} is defined by

$$\delta_{\text{out}}(k) = r_{\text{out}}(k-1) + r_{\text{out}}(k+1) - 2r_{\text{out}}(k). \tag{34}$$

The maximum variance $\sigma_{\max}^2$ is the variance of the optimal outdegree distribution of Eq. (28):

$$\sigma_{\max}^2 = (k_{\max} - \mu_{\text{out}})(\mu_{\text{out}} - 1). \tag{35}$$

Therefore, if $\sigma_{\text{out}}^2 = \sigma_{\max}^2$, the equality of Eq. (32) holds and the network robustness is optimal. In contrast to the indegree distribution, Eq. (32) shows that a higher variance of the outdegree distribution is indicative of higher robustness.

Table I shows the examples of the degree distributions that appear in actual complex networks, and their variances if the mean value is fixed at μ. Let σ_{pow}^2, σ_{exp}^2, and σ_{poi}^2 be the variances of power-law, exponential, and Poisson distributions, respectively. As mentioned earlier, an outdegree distribution with high variance is indicative of high robustness. Within a normal parameter region ($\alpha \leqslant 3$), σ_{pow}^2 can be an infinite value. Therefore, the power-law distribution of outdegree is most robust in the three distributions.

An indegree distribution with low variance is indicative of high robustness. Therefore, the exponential and Poisson distributions can be candidates of the indegree distribution because σ_{exp}^2 and σ_{poi}^2 are finite. As shown in Table I, σ_{exp}^2 is a quadratic function of μ and σ_{poi}^2 is equivalent to μ. If $\mu > 2$, then the order of these variances is $\sigma_{\text{exp}}^2 > \sigma_{\text{poi}}^2$, and

hence the Poisson distribution of indegree can be regarded as more robust. However, if $\mu < 2$, the order of these variances is $\sigma^2_{\rm exp} < \sigma^2_{\rm poi}$. In organisms such as bacteria, which have a small-size genome, the mean value of connections can be less than two. Therefore, in a small-size genome, the exponential distribution of indegree can be regarded as more robust.

IV. RESULTS AND DISCUSSION

A. Estimation of distributions and statistical tests

We evaluate actual indegree and outdegree distributions of the gene regulatory networks in terms of their fit with power-law, exponential, and Poisson distributions. However, in actual indegree distributions, it is observed that the variance is greater than the mean as will be shown in the next subsection. The variance of the Poisson distribution is equivalent to its mean. Therefore, we use the normal distribution instead of the Poisson distribution for the evaluation (we discuss the variance of the indegree distribution in Sec. IV C).

For each distribution, we assume that there is a maximum degree, $k_{\rm max}$. Then, the distribution of the degree k is represented as follows:

$$p_i(k) = \frac{q_i(k)}{\sum_{l=1}^{k_{\rm max}} q_i(l)}, \tag{36}$$

where the function $q_i(k)$ is represented for the power-law ($i = 1$), exponential ($i = 2$), and normal ($i = 3$) distributions as follows, respectively:

$$q_1(k) = k^{-\alpha}, \quad q_2(k) = (1-\gamma)^k, \quad \text{and } q_3(k) = e^{\frac{-(k-\mu)^2}{2\sigma^2}}. \tag{37}$$

The value of each parameter is estimated by the maximum likelihood estimation using the Newton method (we use the quasi-Newton method for normal distributions).

To determine the distribution with the best fit for the given data, we use the likelihood ratio test [35,36]. Let k_n ($n = 1,2,\ldots,M$) be the observed degree data. The log-likelihood for the ith distribution is represented as follows:

$$L_i = \sum_{n=1}^{M} l_i(n), \tag{38}$$

where $l_i(n)$ is the log-likelihood for each degree:

$$l_i(n) = \log p_i(k_n). \tag{39}$$

Then, the log-likelihood ratio between the ith and jth distributions is represented as follows:

$$\mathcal{R}_{ij} = L_i - L_j = \sum_{n=1}^{M} [l_i(n) - l_j(n)]. \tag{40}$$

If $\mathcal{R}_{ij} > 0$ (or $\mathcal{R}_{ij} < 0$), then the ith (or jth) distribution has a better fit for the data. To evaluate the significance of the fit, we calculate the p value with the null hypothesis in which $\mathcal{R}_{ij} = 0$. In the null hypothesis, $\mathcal{R}_{ij}$ is approximately represented as the normal distribution by using the central limit theorem. The variance of $\mathcal{R}_{ij}$ is approximately represented by the empirical variance of the data [35,36]:

$$\sigma^2_{ij} = \frac{1}{M}\sum_{n=1}^{M}\left\{[l_i(n) - l_j(n)] - \frac{L_i - L_j}{M}\right\}^2. \tag{41}$$

Therefore, we calculate the p value as follows:

$$p \text{ value} = \left|\text{erfc}\left(\frac{\mathcal{R}_{ij}}{\sqrt{2M\sigma^2_{ij}}}\right)\right|, \tag{42}$$

where erfc is the complementary Gaussian error function. If the p value is smaller than the significance level, then the distribution indicated by the sign of $\mathcal{R}_{ij}$ can be regarded as a better fit for the given data.

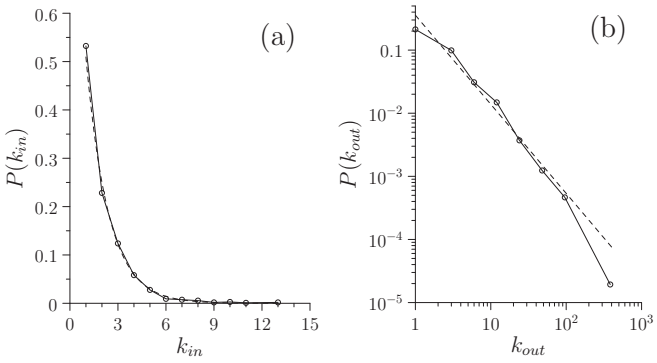

FIG. 4. Degree distributions of *E. coli* (RegulonDB). The solid lines indicate the degree distributions and the dashed lines indicate the theoretical distributions with the best fit. (a) Indegree distribution. The exponential distribution has the best fit ($\gamma = 0.509$, $k_{\rm max} = 14$). (b) Outdegree distribution. The power-law distribution has the best fit ($\alpha = 1.41$, $k_{\rm max} = 281$).

B. Degree distributions of gene regulatory networks

We show the goodness of fit for the three distributions mentioned above for the degree data of *E. coli*, fruit fly, and human. We adopt RegulonDB 9.4 [37] for *E. coli*, DroID v2015_12 [38] for fruit fly, and ENCODE 3 [39] for human. The data preprocessing for each data set is described in the Appendix.

We show the degree distributions of *E. coli* in Fig. 4. We use a log-log plot to illustrate the outdegree distribution [Fig. 4(b)]. The data size of the outdegree distribution is smaller than that of the indegree distribution because the number of transcription factors is small in the data set. Thus, in the outdegree distribution, we adopt the logarithmic binning, wherein the bin width is constant in the logarithmic scale [40].

In the statistical tests shown in Table II, we choose the significance level at 0.05. In the indegree distribution, $\mathcal{R}_{12} < 0$, $\mathcal{R}_{23} > 0$, and each p value is significant. Therefore, the exponential distribution corresponding to index 2 is the best distribution [as indicated by the dashed line in Fig. 4(a)]. Although $\mathcal{R}_{13}$ is not significant, it does not affect this conclusion. In the outdegree distribution, $\mathcal{R}_{12} > 0$, $\mathcal{R}_{13} > 0$, and each p value is significant. Therefore, the power-law distribution corresponding to index 1 is the best distribution [as indicated by the dashed line in Fig. 4(b)].

TABLE II. Results of likelihood ratio tests. The log-likelihood ratios $\mathcal{R}_{ij}$ and the corresponding p values (in parentheses) are shown. The indices 1, 2, and 3 correspond to the power-law, exponential, and normal distributions, respectively.

	$\mathcal{R}_{12}$	$\mathcal{R}_{13}$	$\mathcal{R}_{23}$	Best distribution
E. coli indegree	-62.5 (1.07×10^{-5})	27.3 (0.375)	89.8 (3.69×10^{-7})	Exponential
E. coli outdegree	61.8 (0.0196)	212 (9.29×10^{-6})	151 (4.53×10^{-9})	Power-law
Fruit fly indegree	-51.1 (0.0)	-518 (0.0)	-467 (0.0)	Normal
Fruit fly outdegree	57.7 (2.43×10^{-5})	74.9 (5.30×10^{-6})	17.3 (6.95×10^{-6})	Power-law
Human indegree	-2.20×10^3 (0.0)	-2.82×10^3 (0.0)	-615 (0.0)	Normal
Human outdegree	75.1 (5.23×10^{-4})	174 (1.05×10^{-9})	98.6 (0.0)	Power-law

Figures 5 and 6 show the degree distributions for fruit fly and human, respectively. As shown in Table II, since all p values are smaller than 0.05 for the results of fruit fly and human, all likelihood ratios are statistically significant. The results indicate that the best degree distributions of fruit fly and human are equivalent; the best indegree distribution is the normal distribution and the best outdegree distribution is the power-law distribution.

Table III shows the mean and variance of the indegree distribution (μ_{in} and σ^2_{in}) and the outdegree distribution (μ_{out} and σ^2_{out}). To compare the variance with the mean, the Fano factors of the indegree and outdegree distributions are also shown:

$$F_{\text{in}} = \frac{\sigma^2_{\text{in}}}{\mu_{\text{in}}}, \quad F_{\text{out}} = \frac{\sigma^2_{\text{out}}}{\mu_{\text{out}}}. \tag{43}$$

The normal distribution is the best indegree distribution for fruit fly and human, while the exponential distribution is the best indegree distribution for *E. coli*. As mentioned in the previous section, this discrepancy can be attributed to the mean μ_{in}. If $\mu_{\text{in}} < 2$, then the variance of the exponential distribution is lower than that of the Poisson distribution. The proposed model indicates that the networks are robust if the variance of the indegree distribution is low. Since the mean of *E. coli* is $\mu_{\text{in}} = 1.96 < 2$, the exponential distribution is advantageous. Therefore, this discrepancy of the indegree distributions among species further points to the validity of the proposed model.

The power-law distribution has the best fit for the outdegree distribution for all species (Table II). As shown in Table III, the Fano factor of the outdegree distribution is two to three orders of magnitude greater than that of the indegree distribution. These results suggest that the variance in the outdegree distribution is maintained at a higher value than that in the indegree distribution, as indicated in the proposed model.

If the indegree distribution is Poisson distributed, then the Fano factor is $F_{\text{in}} = 1$. However, the Fano factors of fruit fly and human are greater than one. Therefore, these results indicate that their indegree distributions are not Poisson distributed. Hence, in the following section, we discuss why the variance of the indegree distribution is higher than expected.

C. Variance of indegree distribution

The proposed model suggests that the networks are optimally robust if the variance of the indegree distribution is zero; i.e., all genes have an identical number of input connections. However, in actual cases, zero variance is virtually impossible because the number of input connections is determined locally in each gene. In actual systems, the random graph can be the second-best structure because the number of input connections are determined based on an identical probability for each gene [41]. In the random graph, the indegree distribution is Poisson distributed.

As shown in Table III, the Fano factors of fruit fly and human are sufficiently greater than one. Therefore, the indegree

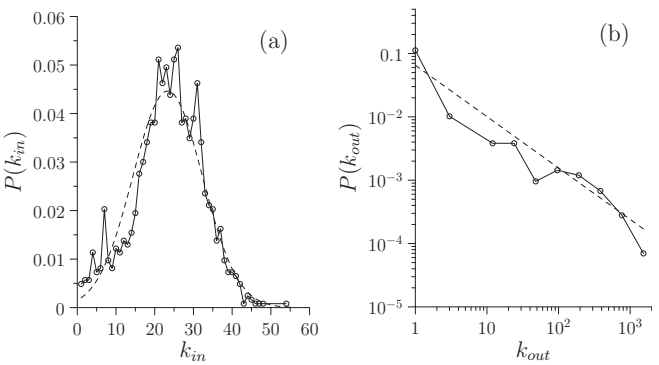

FIG. 5. Degree distributions of fruit fly (DroID). The solid lines indicate the degree distributions and the dashed lines indicate the theoretical distributions with the best fit. (a) Indegree distributions. The normal distribution has the best fit ($\mu = 23.3, \sigma^2 = 80.8, k_{\text{max}} = 55$). (b) Outdegree distributions. The power-law distribution has the best fit ($\alpha = 0.811, k_{\text{max}} = 1{,}150$).

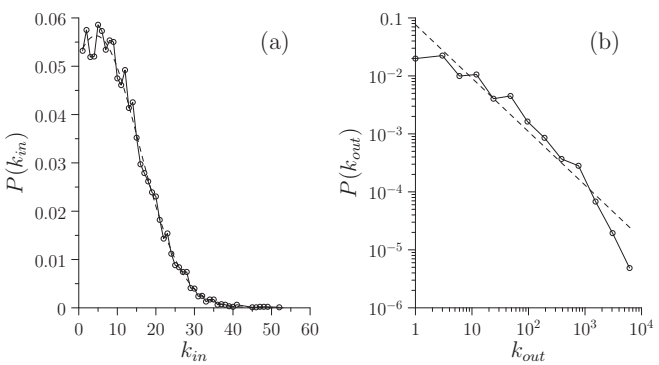

FIG. 6. Degree distributions of human (ENCODE). The solid lines indicate the degree distributions and the dashed lines indicate the theoretical distributions with the best fit. (a) Indegree distributions. The normal distribution has the best fit ($\mu = 4.45, \sigma^2 = 123, k_{\text{max}} = 53$). (b) Outdegree distributions. The power-law distribution has the best fit ($\alpha = 0.923, k_{\text{max}} = 6{,}246$).

TABLE III. Means, variances, and Fano factors of degree distributions.

	μ_{in}	σ^2_{in}	F_{in}	μ_{out}	σ^2_{out}	F_{out}
E. coli	1.96	2.24	1.14	10.5	682	65.0
Fruit fly	23.5	77.3	3.29	295	1.23×10^5	417
Human	10.9	55.0	5.05	505	9.23×10^5	1.83×10^3

distribution is not Poisson distributed. These results appear to be inconsistent with the proposed model, where a low variance implies high robustness. However, the structure of gene regulatory networks is determined not only by the robustness but also by the biological functions.

Paralogs of a transcription factor play a biologically important role and are expected to significantly influence the variance of the indegree distribution. Since the rearrangement of gene regulatory networks is elicited by paralogs, they are considered to be important for biological functions [42]. Paralogs frequently have a similar DNA-binding activity because they originate from the same ancestral gene. Therefore, multiple paralogs of a transcription factor can bind a single binding motif. Consequently, the existence of the paralogs increases the variance of the indegree distribution because the appearances of the paralogs in a promoter are positively correlated with each other.

To evaluate the effect of paralogs on the indegree distribution, we adopt TRANSFAC database [43]. Unlike the experimentally derived databases used in the previous section, TRANSFAC estimates the binding sites of transcription factors with high accuracy by using position specific matrices, which are the model of binding motifs. In addition, TRANSFAC has a data set in which redundancy of binding motifs is eliminated. Since the paralogous transcription factors are bounded to similar motifs, the elimination of motif redundancy implies that paralogs are eliminated. We evaluate the effect of the paralogs by comparing the redundant and nonredundant data (see the Appendix for the data).

As in the case of the ENCODE database results, in both the redundant and nonredundant data of TRANSFAC, the normal and power-law distributions have the best fit for the indegree and outdegree distributions, respectively (data not shown). However, the result of the outdegree distribution for the nonredundant data is not statistically significant, while that of the redundant data is significant. This may be attributed to the reduction in the number of data (transcription factors). However, the results of the indegree distributions for both the redundant and nonredundant data are statistically significant. Therefore, the results of TRANSFAC are sufficiently consistent with those of ENCODE.

TABLE IV. Means, variances, and Fano factors of indegree distributions estimated by TRANSFAC.

	μ_{in}	σ^2_{in}	F_{in}
Redundant data	461	4.42×10^3	9.59
Nonredundant data	31.2	25.2	0.808

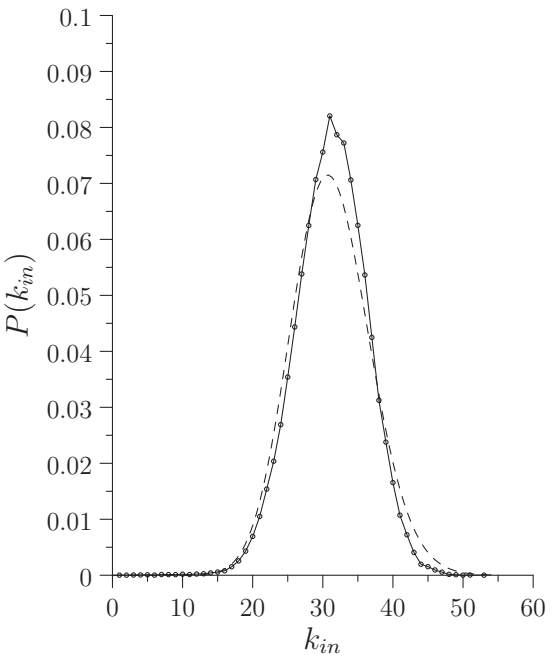

FIG. 7. Indegree distribution estimated by nonredundant data of TRANSFAC. The solid line indicates the indegree distribution and the dashed line indicates the Poisson distribution ($\lambda = 31.2$).

Table IV shows the means, variances, and Fano factors of the indegree distributions. As can be seen in Table IV, the Fano factor of the redundant data is greater than one, which is similar to the results of ENCODE. Therefore, the indegree distribution is not Poisson distributed. However, the Fano factor of the nonredundant data is approximately equivalent to one. As shown in Fig. 7, the indegree distribution of the nonredundant data is approximated by the Poisson distribution. These results indicate that the indegree distribution is Poisson distributed if the effect of paralogs is eliminated. Therefore, although the gene regulatory networks have many paralogs and cause an increase in the variance to realize the biological functions, the networks are essentially Poisson distributed, as indicated by the proposed model.

V. CONCLUSIONS

In this work, we defined the dynamical robustness of gene regulatory networks and showed that the discrepancy between the indegree and outdegree distributions can be attributed to the dynamical robustness. The robustness characteristics as a function of the number of input and output connections are concave and convex, respectively. Applying these characteristics to Jensen's inequality, we showed that the networks are robust if the variance of the indegree distribution is low and the variance of the outdegree distribution is high. By analyzing the actual gene regulatory networks of *E. coli*, fruit fly, and human, we showed that the indegree distribution has a goodness of fit for normal or exponential distributions with low variance, and the outdegree distribution has a goodness of fit for power-law distributions with high variance. Although the variance of the indegree distribution is higher than expected, we elucidated

that this is due to the effect of the paralogs of the transcription factors.

As discussed in the results, the variance of degree distributions is a suitable indicator for the evaluation of dynamical robustness. This is confirmed by the observations that the type of the indegree distribution is dependent on the size of the genome; the indegree distribution has a goodness of fit for exponential distributions if the size of the genome is small and for Poisson distributions if the size of the genome is sufficiently large. We can theoretically explain these observations because the variance of exponential distributions is lower than that of Poisson distributions if the mean value of the number of connections is less than two and higher if not. The outdegree distribution has a goodness of fit for power-law distributions because of high variance. Power-law distributions can have infinite variance. However, the proposed model cannot explain why power-law distributions are observed as outdegree distributions among various distributions with infinite variance. This is because magnitudes of infinite variance cannot be compared unlike the case of the indegree distribution with finite variance. Therefore, in addition to dynamical robustness, there may be other factors that explain the selection of power-law distributions. Future work will focus on the origin of power-law distributions.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant No. 15KT0074.

APPENDIX: DATA SETS

We adopt RegulonDB 9.4 to construct the gene regulatory networks of *E. coli* [37]. As the target of output connections, we use the transcription unit, which is the set of genes regulated by a single promoter. Consequently, we obtained 1082 transcription units and 202 transcription factors.

We adopt DroID (the Drosophila interaction database) v2015_12 to construct the gene regulatory networks of fruit fly [38]. DroID includes REDFly, which is a high quality, curated database of gene interactions [44], and modENCODE, which is an interaction database experimentally determined using ChIP-seq and ChIP-chip [45]. ChIP-seq and ChIP-chip are methods to experimentally determine whether transcription factors bind promoters. However, since there are inactive promoters in actual cells, the absence of binding signals by these methods does not necessarily imply that there is no promoter binding. Thus, we select only the active promoters by using the data of DNase I sensitivity (Kc167 cell line) [46]. Since DNase I sensitivity identifies low-density regions of nucleosomes, such regions are positively correlated with active promoters. We take an average of DNase I read density in the promoter region, which is obtained from the eukaryotic promoter database [47], and we adopt the top 10% of the promoters. Consequently, we obtained 1232 promoters and 98 transcription factors.

We adopt ENCODE (Encyclopedia of DNA Elements) version 3 to construct the gene regulatory networks of human [39]. Similar to modENCODE, ENCODE is the interaction database that is experimentally determined by ChIP-seq and ChIP-chip. Given that ENCODE is a large database, we adopt only the peak data whose scores are greater than 200 to reduce noise. Even if there are multiple peaks for an identical transcription factor in an identical promoter, we treat them as a single peak. We use DBTSS R8 as the promoter data (hg19, upstream 500 bp from transcription start sites) [48]. Similar to the data of fruit fly, we adopt only the top 10% of the promoters evaluated by DNase I sensitivity (ENCODE, K562 cell line). Consequently, we obtained 9286 promoters and 201 transcription factors.

To construct the gene regulatory networks of human, we also use the data estimated by TRANSFAC, which is a database of binding motifs [43]. We search the binding motifs of TRANSFAC by using the condition of "minimize false positive" for the promoters of DBTSS R8 (upstream 1000 bp and downstream 200 bp from transcription start sites) [48]. Similar to ENCODE data, even if there are multiple hits for a motif in a promoter, we treat them as a single hit. We use the two types of redundant and nonredundant data sets of vertebrates. Consequently, we obtained 96 807 promoters for both of the redundant and nonredundant data. We also obtained 4638 and 163 transcription factors for the redundant and nonredundant data, respectively.

[1] A.-L. Barabasi and Z. N. Oltvai, Nat. Rev. Genet. **5**, 101 (2004).
[2] R. Albert, J. Cell Sci. **118**, 4947 (2005).
[3] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabasi, Nature **407**, 651 (2000).
[4] A. Wagner and D. A. Fell, Proc. Roy. Soc. London B: Biol. Sci. **268**, 1803 (2001).
[5] H. Jeong, S. P. Mason, A.-L. Barabasi, and Z. N. Oltvai, Nature **411**, 41 (2001).
[6] S.-H. Yook, Z. N. Oltvai, and A.-L. Barabasi, Proteomics **4**, 928 (2004).
[7] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, Science **302**, 249 (2003).
[8] M. Arita, Proc. Natl. Acad. Sci. USA **101**, 1543 (2004).
[9] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, Nat. Genet. **31**, 64 (2002).
[10] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J.-B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young, Science **298**, 799 (2002).
[11] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley, Proc. Natl. Acad. Sci. USA **97**, 11149 (2000).
[12] A. L. Barabási, *Network Science* (Cambridge University Press, Cambridge, 2016).
[13] R. Albert, H. Jeong, and A.-L. Barabasi, Nature **406**, 378 (2000).
[14] M. E. Tsuda and M. Kawata, PLoS Comput. Biol. **6**, 1 (2010).

[15] J. L. Faulon, S. Martin, and R. D. Carr, in *Proceedings of the IEEE Computational Systems Bioinformatics Conference* (CSB'04), Stanford, CA, USA (IEEE, 2004), pp. 595–596.
[16] C. H. Waddington, Nature **150**, 563 (1942).
[17] A. Wagner, Evolution **50**, 1008 (1996).
[18] M. L. Siegal and A. Bergman, Proc. Natl. Acad. Sci. USA **99**, 10528 (2002).
[19] R. D. Leclerc, Mol. Syst. Biol. **4**, 213 (2008).
[20] M. L. Siegal, D. E. L. Promislow, and A. Bergman, Genetica **129**, 83 (2007).
[21] S. Ciliberti, O. C. Martin, and A. Wagner, PLoS Comput. Biol. **3**, e15 (2007).
[22] E. Rünneburger and A. Le Rouzic, BMC Evol. Biol. **16**, 239 (2016).
[23] S. Roy and T. K. Kundu, IUBMB Life **66**, 100 (2014).
[24] M. Laurent and N. Kellershohn, Trends Biochem. Sci. **24**, 418 (1999).
[25] M. Santillán, M. C. Mackey, and E. S. Zeron, Biophys. J. **92**, 3830 (2007).
[26] K. Nakano, IEEE Trans. Syst. Man Cybernet. **SMC-2**, 380 (1972).
[27] J. A. Anderson, Math. Biosci. **14**, 197 (1972).
[28] T. Kohonen, IEEE Trans. Comput. **C-21**, 353 (1972).
[29] S.-I. Amari, IEEE Trans. Comput. **C-21**, 1197 (1972).
[30] D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory* (Psychology Press, UK, 2005).
[31] H. Zhao, Phys. Rev. E **70**, 066137 (2004).
[32] J. L. W. V. Jensen, Acta Mathematica **30**, 175 (1906).
[33] M. K. Bakula and K. Nikodem, J. Math. Anal. Appl. **434**, 516 (2016).
[34] J. M. Steele, *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities* (Cambridge University Press, New York, NY, 2004).
[35] A. Clauset, C. R. Shalizi, and M. E. J. Newman, SIAM Rev. **51**, 661 (2009).
[36] Q. H. Vuong, Econometrica **57**, 307 (1989).
[37] S. Gama-Castro, H. Salgado, A. Santos-Zavaleta, D. Ledezma-Tejeida, L. Muñiz-Rascado, J. S. García-Sotelo, K. Alquicira-Hernández, I. Martínez-Flores, L. Pannier, J. A. Castro-Mondragón, A. Medina-Rivera, H. Solano-Lira, C. Bonavides-Martínez, E. Pérez-Rueda, S. Alquicira-Hernández, L. Porrón-Sotelo, A. López-Fuentes, A. Hernández-Koutoucheva, V. D. Moral-Chávez, F. Rinaldi, and J. Collado-Vides, Nucleic Acids Res. **44**, D133 (2016).
[38] T. Murali, S. Pacifico, J. Yu, S. Guest, G. G. Roberts, and R. L. Finley, Nucleic Acids Res. **39**, D736 (2010).
[39] ENCODE Project Consortium and others, Nature **489**, 57 (2012).
[40] E. P. White, B. J. Enquist, and J. L. Green, Ecology **89**, 905 (2008).
[41] P. Erdős and A. Rényi, Publicationes Mathematicae (Debrecen) **6**, 290 (1959).
[42] J. C. Pérez, P. M. Fordyce, M. B. Lohse, V. Hanson-Smith, J. L. DeRisi, and A. D. Johnson, Genes Dev. **28**, 1272 (2014).
[43] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender, Nucleic Acids Res. **34**, D108 (2006).
[44] M. S. Halfon, S. M. Gallo, and C. M. Bergman, Nucleic Acids Res. **36**, D594 (2007).
[45] S. Roy, J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre, M. L. Eaton, J. M. Landolin, C. A. Bristow, L. Ma, M. F. Lin *et al.*, Science **330**, 1787 (2010).
[46] P. V. Kharchenko, A. A. Alekseyenko, Y. B. Schwartz, A. Minoda, N. C. Riddle, J. Ernst, P. J. Sabo, E. Larschan, A. A. Gorchakov, T. Gu *et al.*, Nature **471**, 480 (2011).
[47] R. Dreos, G. Ambrosini, R. Groux, R. Cavin Périer, and P. Bucher, Nucleic Acids Res. **45**, D51 (2017).
[48] A. Suzuki, H. Wakaguri, R. Yamashita, S. Kawano, K. Tsuchihara, S. Sugano, Y. Suzuki, and K. Nakai, Nucleic Acids Res. **43**, D87 (2014).