other hand, BAK1 is not considered as a hub, but is as an accumulation node of the network, and is selected as a reporter. Moreover, it seems that some of the selected proteins have significant biological meanings as follows. p53, a tumour suppressor gene that responds to DNA-damage, is influential on TRAIL-induced apoptosis by up-regulating TRAIL receptor (Wu et al., 1997). Bcl-2 superfamily regulates cell death that is amplified via the mitochondrial pathway (Sprick and Walczak, 2004). BAX may be related with possible amplification of apoptosis via the intrinsic pathway in response to JNK. The caspase-9 (CASP9) may be essential for border-cell migration in the Drosophila ovary (Geisbrecht and Montell, 2004), and the regulation of cell migration may also point to a roll in the cleavage of several adhesion- and cell motility- related proteins during mammalian apoptosis (Fischer et al., 2003).

Figure 3 Apoptosis pathway maps in a HeLa cell, which contain 132 proteins and 337 binomial relations

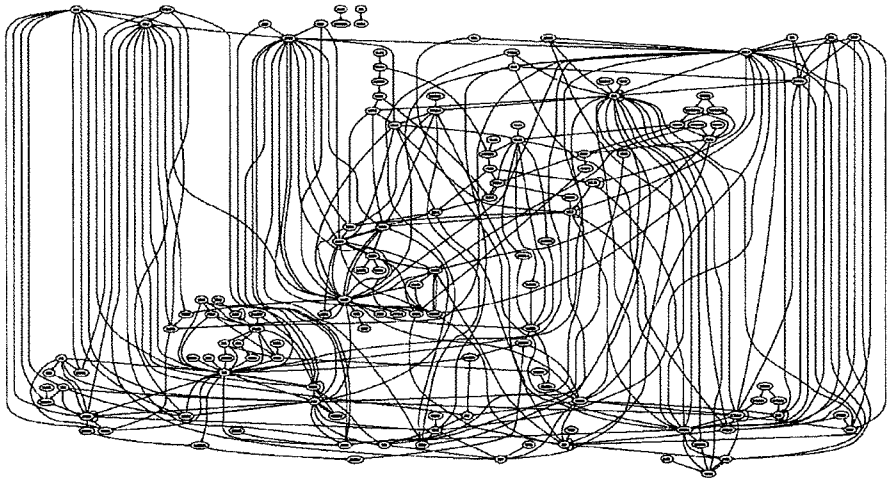


Table 1 The optimal solution of IP1 and IP2 for each L and K in apoptosis pathway maps, where the numbers of covered nodes and the numbers of the selected reporters are shown for IP1 and IP2, respectively

	IP1 for each K							
L	*1*	*2*	*3*	*4*	*5*	*6*	*IP2*	*Reporter in K = 1 (indegree/outdegree)*
1	20	36	47	56	62	68	42	TP53 (19/5)
2	60	76	85	92	98	104	22	BCL2 (17/4)
3	88	103	110	116	118	120	15	BAX (16/6)
4	109	116	120	122	124	126	12	BAX (16/6)
5	118	121	123	125	127	128	10	BAK1 (6/1)
6	121	123	125	127	128	129	9	BAK1 (6/1)
132	121	123	125	127	128	129	9	BAK1 (6/1)

Table 2 shows the selected proteins as reporters for each L and K. The protein selected as a reporter for smaller K was not always selected for larger K. For example, for $L = 2$, BCL2 was selected as a reporter in the case of $K = 1$, but was not in the cases of $K = 2, \ldots, 4$. If we use a simple greedy algorithm for solving P1, we may not be able

to find CASP9 and BAX for $K = 2$, or CASP9, BAX and IKBKG for $K = 3$ since the greedy algorithm often tends to add a new node to the solution for $K - 1$. Actually, for $L = 1$ the greedy algorithm selected 44 reporters to cover all nodes of V although only 42 reporters are required as we see from the result of IP2. On the other hand, our integer programming-based methods can always find optimal solutions if any. For each case, the elapsed time of optimising IP1 or IP2 was at most 0.023 seconds. These results suggest that our methods are practical.

Table 2 Selected proteins as reporters for each L and K in apoptosis pathway maps

L	*K*	*IP1*	*Reporters*
1	1	20	TP53
1	2	36	TP53, BCL2
1	3	47	TP53, BCL2, BAX
1	4	56	TP53, BCL2, BAX, CASP9
1	5	62	TP53, BCL2, BAX, CASP9, FADD
1	6	68	TP53, BCL2, BAX, CASP9, FADD, MAP3K1
1	7	73	TP53, BCL2, BAX, CASP9, FADD, MAP3K1, BIRC4
2	1	60	BCL2
2	2	76	CASP9, BAX
2	3	85	CASP9, BAX, IKBKG
2	4	92	CASP9, BAX, IKBKG, MAP2K7
2	5	98	CASP9, IKBKG, MAP2K7, BCL2, VDAC2
2	6	104	CASP9, IKBKG, MAP2K7, BCL2, VDAC2, TP53
3	1	88	BAX
3	2	103	BAX, IKBKG
3	3	110	IKBKG, BCL2, VDAC2
3	4	116	IKBKG, BCL2, BAK1, MAP2K7
3	5	118	IKBKG, BAK1, MAP2K7, CASP9, TP53
4	1	109	BAX
4	2	116	BCL2, BAK1
4	3	120	BAX, VDAC2, IKBKG
4	4	122	BAX, VDAC2, IKBKG, FASLG
5	1	118	BAK1
5	2	121	BAK1, BCL2
5	3	123	BCL2, VDAC2, TNFRSF1A
5	4	125	BCL2, VDAC2, TNFRSF1A, DFFB
6	1	121	BAK1
6	2	123	BAK1, FASLG
6	3	125	BAK1, FASLG, TNFRSF1A
132	1	121	BAK1
132	2	123	BAK1, TNFRSF1A
132	3	125	BAK1, TNFRSF1A, FASLG

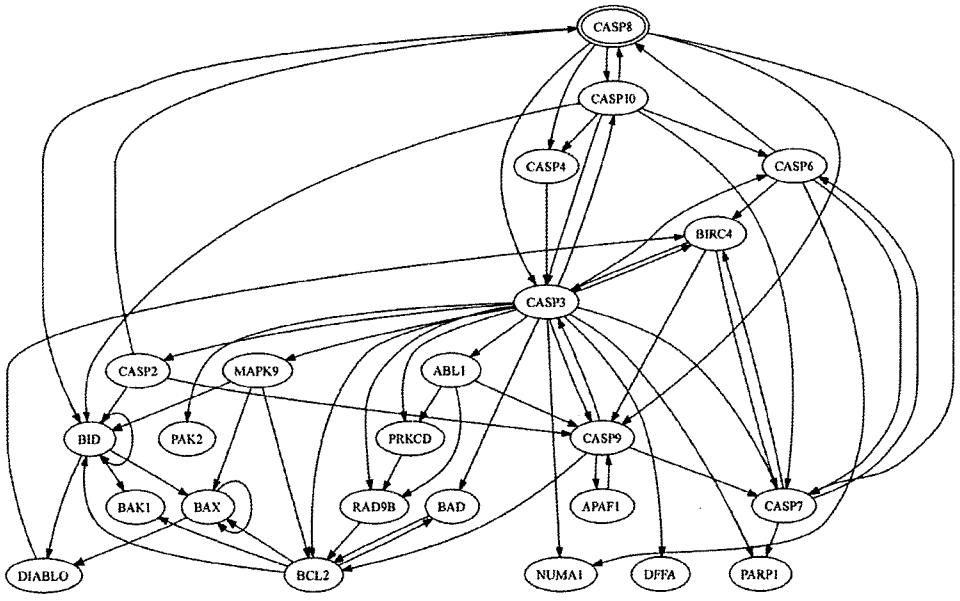
5.1.1 Effects of specific nodes

It is also important to observe the effects of signals on specific proteins or genes using cell arrays. In this section, we used CASP8, which is a protease located at the upstream of the caspase cascade that is a main pathway of the apoptosis initiated by

TRAIL (Lamkan et al., 2007), as a specific protein among the apoptosis pathway maps. Then, we extracted the downstream proteins within the distance 2 from CASP8 (See Figure 4). We excluded CASP8 from this downstream subnetwork not to select it as a reporter. Thus, we obtained the subnetwork with 23 proteins and 58 binomial relations excluding CASP8.

Table 3 shows selected proteins as reporters for each L and K as Table 2. In both the whole network and the subnetwork, the same proteins such as BCL2, BAK1 and CASP9 were selected as reporters. It is reasonable because they have similar connections in both networks. For $L = 4, \ldots, n(= 23)$, five proteins without outward edges were selected as the optimal reporter nodes in IP2.

Figure 4 Downstream proteins of CASP8 within the distance 2 in apoptosis pathway maps. CASP8 is highlighted with the double circles. We excluded CASP8 from this subnetwork not to select it as a reporter



5.2 *Artificial scale-free networks*

It is known that many real biological networks have the scale-free property (Barabási and Albert, 1999). The degree distribution, $P(k)$, of a scale-free network follows a power-law relationship ($P(k) \propto k^{-\gamma}$). In the network, most nodes have one connection, and a few nodes have many connections. In particular, it is observed that gene regulatory networks have the power-law outdegree distribution and the Poisson indegree distribution (Guelzim et al., 2002). Thus, we generated scale-free networks with power-law outdegree distributions and Poisson indegree distribution as follows. We first choose the outdegree for each node from a power-law distribution. That is, the outdegree d_i of node v_i is drawn from a power-law distribution. Then, we choose d_i output nodes randomly with uniform probability from n nodes. Thus, the indegree distribution should follow a Poisson distribution.

Table 4 shows the average CPU time over 100 networks for each case. Since it is known that the exponent $\gamma = 2 \sim 3$ for many real biological networks, we performed experiments for $\gamma = 2.0, 2.2, 2.5$, and 3.0. For large $n(= 1000, 5000, 10000)$, the elapsed

time was sufficiently short (even in almost cases of $L = 3$ and $K = 5$). However, the elapsed time was slightly long for small γ. It is reasonable because such a scale-free network often has more reporter candidates than that with large γ. This result suggests that the proposed methods are scalable to realistic size instances. The elapsed time of IP2 was shorter than that of IP1 for almost all cases. It is reasonable because IP1 has twice as many integer variables as IP2, and the number of constraints in IP1 is larger than that in IP2.

Table 3 Selected proteins as reporters for each L and K in the downstream proteins of CASP8

L	*K*	*IP1*	*Reporters*
1	1	6	BCL2
1	2	10	BID, CASP7
1	3	13	BCL2, BID, BIRC4
1	10 (IP2)	23	CASP9, RAD9B, BCL2, BAK1, DIABLO, CASP3, DFFA, NUMA1, PAK2, PARP1
2	1	13	BCL2
2	2	18	BCL2, BIRC4
2	3	19	BCL2, DIABLO, NUMA1
2	7 (IP2)	23	BCL2, BAK1, DIABLO, DFFA, NUMA1, PAK2, PARP1
3	1	16	BAD
3	6 (IP2)	23	CASP9, BAK1, DFFA, NUMA1, PAK2, PARP1
4	5 (IP2)	23	BAK1, DFFA, NUMA1, PAK2, PARP1
23	1	19	BAK1
23	5 (IP2)	23	BAK1, DFFA, NUMA1, PAK2, PARP1

Table 4 Elapsed time (sec) of solving IP1 and IP2 for each n, γ, L and K

n	*γ*	*L*	*K*	*IP1*	*IP2*
1000	2.0	1	1	0.0284183	0.0222008
1000	2.5	1	1	0.0147972	0.00932519
1000	3.0	1	1	0.0127831	0.00622164
1000	2.0	3	5	7.87762	0.168065
1000	2.5	3	5	0.904964	0.0526494
1000	3.0	3	5	0.114845	0.0205262
5000	2.0	1	1	0.15739	0.133003
5000	2.5	1	1	0.102972	0.0485728
5000	3.0	1	1	0.0936872	0.0322236
5000	2.2	3	5	272.207	7.84515
5000	2.5	3	5	2.90922	0.841976
5000	3.0	3	5	0.179411	0.153181
10000	2.0	1	1	0.423074	0.301631
10000	2.5	1	1	0.276991	0.101553
10000	3.0	1	1	0.259448	0.0655794
10000	2.2	3	5	604.545	430.068
10000	2.5	3	5	5.62986	4.01971
10000	3.0	3	5	0.374687	0.392894

6 Concluding remarks

We have studied the problem of allocating a set of reporter genes that are most effective for analysing genetic networks and signaling pathways. We proposed two formalisations P1 and P2 of this problem. P1 selects a set of nodes that covers as many nodes (genes or proteins) as possible whereas P2 selects a minimal set of nodes that covers all the nodes in a network. We showed hardness results on approximation of these problems. On the other hand, by means of reduction to the set cover problem, we showed that P1 and P2 can be approximated within a factor of $e/(e-1)$ and $O(\log n)$, respectively.

We proposed integer programming-based methods IP1 and IP2 so as to find optimal solutions for practical instances of P1 and P2, respectively. We applied them to apoptosis pathway maps, and found that such proteins as TP53, BCL2 and BAX selected by our methods often correspond to hubs in the network. These proteins are also considered to play important biological roles. Furthermore, we applied our methods to artificial scale-free networks with up to 10,000 nodes, and we showed that our methods can compute optimal solutions for these networks in practical time.

We did not consider specific mathematical models such as Boolean networks and Bayesian networks since there is no consensus on mathematical models of biological networks. Instead, biological networks are treated simply as directed and unweighted networks. However, IP1 and IP2 can be modified for undirected and/or weighted networks. If more detailed information (e.g., rates of reactions and flux distribution) should be taken into account, it may be embedded in weights of edges. Furthermore, we can add various kinds of constraints to IP1 and IP2 because these are based on integer programming. Such a flexibility would be useful for modifying the proposed methods according to requirements from experimental biologists.

Acknowledgments

We would like to thank Prof. Yuichi Sugiyama in University of Tokyo for valuable suggestions. This work is partially supported by the Cell Array Project from NEDO, Japan and by a Grant-in-Aid 'Systems Genomics' from MEXT, Japan.

References

Akutsu, T. and Bao, F. (1996) 'Approximating minimum keys and optimal substructure screens', *Lecture Notes in Computer Science* 1090 (Proc. COCOON 96), pp.290–299.

Bailey, S.N., Wu, R.Z. and Sabatini, D.M. (2002) 'Applications of transfected cell microarrays in high-throughput drug discovery', *Drug Discovery Today*, Vol. 7, pp.S113–S118.

Barabási, A-L. and Albert, R. (1999) 'Emergence of scaling in random networks', *Science*, Vol. 286, pp.509–512.

Chabrier-Rivier, N., Chiaverini, M., Danos, V., Fages, F. and Schächter, V. (2004) 'Modeling and querying biomolecular interaction networks', *Theoretical Computer Science*, Vol. 325, pp.25–44.

Eker, S., Knapp, M., Laderoute, K., Lincoln, P. and Talcott, C.L. (2002) 'Pathway logic: executable models of biological networks', *Electric Notes in Theoretical Computer Science*, Vol. 71, pp.144–161.

Fischer, U., Janicke, R.U. and Schulze-Ostho, K. (2003) 'Many cuts to ruin: a comprehensive update of caspase substrates', *Cell Death Differentiation*, Vol. 10, pp.76–100.

Geisbrecht, E.R. and Montell, D.J. (2004) 'A role for Drosophila IAP1-mediated caspase inhibition in Rac-dependent cell migration', *Cell*, Vol. 118, pp.111–125.

Golzio, M., Mazzolini, L., Ledoux, A., Paganin, A., Izard, M., Hellaudais, L., Bieth, A., Pillaire, M.J., Cazaux, C., Hoffmann, J.S., Couderc, B. and Teissié, J. (2007) 'In vivo gene silencing in solid tumors by targeted electrically mediated siRNA delivery', *Gene Therapy*, Vol. 14, pp.752–759.

Guelzim, N., Bottani, S., Bourgine, P. and Képès, F. (2002) 'Topological and causal structure of the yeast transcriptional regulatory network', *Nature Genetics*, Vol. 31, pp.60–63.

Hadjantonakis, A.K., Dickinson, M.E., Fraser, S.E. and Papaioannou, V.E. (2003) 'Technicolour transgenics: imaging tools for functional genomics in the mouse', *Nature Reviews Genetics*, Vol. 4, pp.613–625.

Hochbaum, D.S. (1982) 'Approximation algorithms for the set covering and vertex cover problems', *SIAM Journal on Computing*, Vol. 11, pp.555, 556.

Kato, K., Umezawa, K., Miyake, M., Miyake, J. and Nagamune, T. (2004) 'Transfection microarrays of nonadherent cells on an oleyl poly (ethylene glycol) ether-modified glass slide', *Biotechniques*, Vol. 37, pp.444–452.

Kimberley, F.C. and Screaton, G.R. (2004) 'Following a TRAIL: update on a ligand and its five receptors', *Cell Research*, Vol. 14, pp.359–372.

Lamkan, M., Festjens, N., Declercq, W., Vanden Berghe, T. and Vandenabeele, P. (2007) 'Caspases in cell survival, proliferation and differentiation', *Cell Death and Differentiation*, Vol. 14, pp.44–55.

Ruths, D.A., Nakhleh, L., Iyengar, M.S., Reddy, S.A.G. and Ram, P.T. (2006) 'Hypothesis generation in signaling networks', *Journal of Computational Biology*, Vol. 9, pp.1546–1557.

Sprick, M.R. and Walczak, H. (2004) 'The interplay between the Bcl-2 family and death receptor-mediated apoptosis', *Biochim. Biophys. Acta*, Vol. 1644, pp.125–132.

Stearman, R.S., Grady, M.C., Nana-Sinkam, P., Varella-Garcia, M. and Geraci, M.W. (2007) 'Genetic and epigenetic regulation of the human prostacyclin synthase promoter in lung cancer cell lines', *Molecular Cancer Research*, Vol. 5, pp.295–308.

Tran, N., Baral, C., Nagaraj, V.J. and Joshi, L. (2005) 'Knowledge-based framework for hypothesis formation in biochemical networks', *Bioinformatics*, Vol. 21, pp.ii213-ii219.

Vazirani, V.V. (2001) *Approximation Algorithms*, Springer, Berlin.

Watts, D.J. and Strogatz, S.H. (1998) 'Collective dynamics of small-world networks', *Nature*, Vol. 393, pp.440–442.

Wu, G.S., Burns, T.F., McDonald III., E.R. Jiang, W., Meng, R., Krantz, I.D., Kao, G., Gan, D-D., Zhou, J-Y., Muschel, R., Hamilton, S.R., Spinner, N.B., Markowitz, S., Wu, G. and El-Deiry, W.S. (1997) 'KILLER/DR5 is a DNA damage-inducible p53-regulated death receptor gene', *Nature Genetics*, Vol. 17, pp.141–143.

Yoshikawa, T., Uchimura, E., Kishi, M., Funeriu, D.P., Miyake, M. and Miyake, J. (2004) 'Transfection microarray of human mesenchymal stem cells and on-chip siRNA gene knockdown', *Journal of Controlled Release*, Vol. 96, pp.227–232.

Ziauddin, J. and Sabatini, D.M. (2001) 'Microarray of cells expressing defined cDNAs', *Nature*, Vol. 411, pp.107–110.

BMC Systems Biology

BioMed Central

Methodology article

Open Access

Network evaluation from the consistency of the graph structure with the measured data

Shigeru Saito[†1,2], Sachiyo Aburatani[†1] and Katsuhisa Horimoto[*1]

Address: [1]Biological Network Team, Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan and [2]Chem & Bio Informatics Department, INFOCOM CORPORATION, Mitsui Sumitomo Insurance Surugadai Annex Building, 3-11, Kanda-surugadai, Chiyoda-ku, Tokyo 101-0062, Japan

Email: Shigeru Saito - sh.saito@infocom.co.jp; Sachiyo Aburatani - s.aburatani@aist.go.jp; Katsuhisa Horimoto* - k.horimoto@aist.go.jp

* Corresponding author †Equal contributors

Published: 1 October 2008

BMC Systems Biology 2008, **2**:84 doi:10.1186/1752-0509-2-84

Received: 28 May 2008
Accepted: 1 October 2008

This article is available from: http://www.biomedcentral.com/1752-0509/2/84

Abstract

Background: A knowledge-based network, which is constructed by extracting as many relationships identified by experimental studies as possible and then superimposing them, is one of the promising approaches to investigate the associations between biological molecules. However, the molecular relationships change dynamically, depending on the conditions in a living cell, which suggests implicitly that all of the relationships in the knowledge-based network do not always exist. Here, we propose a novel method to estimate the consistency of a given network with the measured data: i) the network is quantified into a log-likelihood from the measured data, based on the Gaussian network, and ii) the probability of the likelihood corresponding to the measured data, named the graph consistency probability (*GCP*), is estimated based on the generalized extreme value distribution.

Results: The plausibility and the performance of the present procedure are illustrated by various graphs with simulated data, and with two types of actual gene regulatory networks in *Escherichia coli*: the SOS DNA repair system with the corresponding data measured by fluorescence, and a set of 29 networks with data measured under anaerobic conditions by microarray. In the simulation study, the procedure for estimating *GCP* is illustrated by a simple network, and the robustness of the method is scrutinized in terms of various aspects: dimensions of sampling data, parameters in the simulation study, magnitudes of data noise, and variations of network structures.

In the actual networks, the former example revealed that our method operates well for an actual network with a size similar to those of the simulated networks, and the latter example illustrated that our method can select the activated network candidates consistent with the actual data measured under specific conditions, among the many network candidates.

Conclusion: The present method shows the possibility of bridging between the static network from the literature and the corresponding measurements, and thus will shed light on the network structure variations in terms of the changes in molecular interaction mechanisms that occur in response to the environment in a living cell.

Background

The knowledge-based approach to construct biological network models is recognized as one of the most promising advances in computational biology [1]. In this approach, the causal relations between biological molecules are described as a directed graph, based on the interaction information extracted from a large number of previous reports, in a manual or automatic manner [2,3]. Since each relation has been identified by experimental studies, the existence of edges in the network model is supported by strong evidence. Due to the high reliability of each relation, many network models, even those with large, complex structures, have been constructed for various biological phenomena by a knowledge-based approach [4-6]. Note that a network generated by a knowledge-based approach is a mixture of molecular relationships identified by experimental studies under different conditions. Indeed, it is well known that the relationships between the molecules in a living cell change dynamically, depending on the cellular environment. Fortunately, an abundance of such information about molecular interactions under different conditions has been obtained by measuring them on a genomic scale, due to recent advances in experimental techniques, and the information about the interactions is available at various web sites [7]. Thus, we can evaluate the consistency of the knowledge-based network structure by the available information about the data measured under the different conditions. Although the inference of static network structures from the data has been intensively studied by various approaches, such as the Bayesian network [8], the dynamic Bayesian network [9], the Boolean network [10], and the graphical Gaussian model [11], the consistency evaluation will be useful to trace the dynamic network structure variations reflecting the molecular relationships that change coordinately in response to the cellular environment.

The consistency evaluation between the network structure and the measured data is well known in statistics as the test for causal hypotheses by using the measured data. The origin of the test for causal hypotheses is attributed to path analysis [12]. Unfortunately, the importance of this cornerstone research was not recognized for a long time, but the natural extension of path analysis has been established as the well-known structural equation model (SEM) [13]. Indeed, the SEM has been utilized recently in various fields, in accordance with increased computer performance. However, the SEM without any latent variables, which is a natural assumption for its application to biological networks, sometimes has difficulties in the numerical calculation of the maximum likelihood for the observed data. To overcome the problem with this calculation, the d-sep test [14] has been developed, based on the concept of d-separation in a directed acyclic graph (DAG) [15]. Note that the graph consistency with the data in the d-sep test is considered by focusing on the absence of edges in the graph [16,17].

Recently, linear regression was applied to reconcile the gene regulatory network with the corresponding data [18]. This application is based on the concept that the entire network of gene regulation can be divided into a few network motifs, with a two-layer relationship between the transcription factors and their regulated genes [19]. Indeed, the division of the entire network into a small and simple network enables us to utilize the standard statistical tests in linear regression for the consistency of the gene relationships with the measured data. Unfortunately, the linear regression is limited to the two-layer relationships, and subsequently, its application is constrained to the simple structures of gene regulatory networks.

In this study, we propose a new method for estimating the consistency of a causal graph with the measured data, in combination with the Gaussian network (GN) [20] and the generalized extreme value distribution (GEV) [21]. The present study partly exploits the previous study [18] about the consistency between the network motif with two-layer gene relationships and the measured data. However, instead of the network motifs with simple structures, here we consider rationally complicated network structures based on the graphical model, and its consistency with the data is expressed as a probability, referred to as the graph consistency probability (*GCP*). The performance of the present method is examined by artificial networks with various structures and actual data measured in *Escherichia coli*. Furthermore, the merits and pitfalls of our method are discussed in terms of its possible utility with various actual issues and methodologies, in comparison with previous methods.

Results and discussion

Calculation of Graph Consistency Probability (GCP)

We will illustrate the procedure for calculating the graph consistency probability (*GCP*) with a simple graph, G_0, which is a directed acyclic graph with ten nodes and nine edges, and with the corresponding data that are artificially generated on the assumption that the data noise follows the normal distribution. The procedure for calculating the *GCP* is composed of five steps, as schematically shown in Fig. 1 (see details of the mathematical description in the Materials and Methods and the additional file 1: Details of the schematic description of the procedure).

At the first step, the given graph, G_0, is recursively factorized into the subgraphs, according to the parent-descent relationships in DAG [15]. By recursive factorization, G_0 is

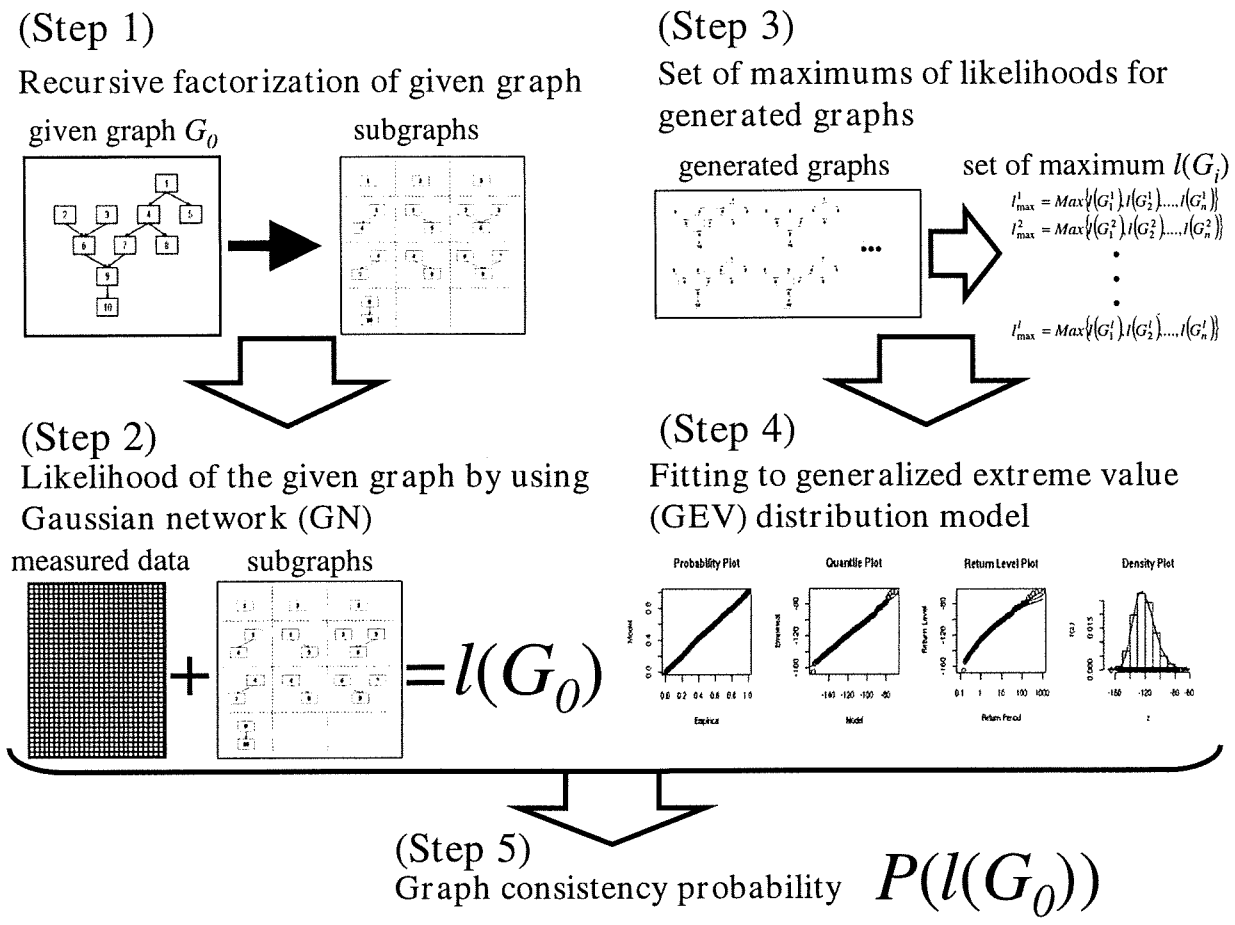


Figure 1
Flow of the calculation of graph consistency probability. The calculation is composed of five steps (see details in the text).

rationally divided into 10 subgraphs, based on the parent-descent relationships in the directed graph.

At the second step, we calculated the log-likelihood of the given graph, $l(G_0)$, with the corresponding data by a Gaussian network model [20]. To calculate $l(G_0)$, here, we generated the data $\{X_{kl}\}$ for each node with 50 sample dimensions, i.e., for k = 1,2,...,10 and l = 1,2,...,50, instead of the actual data, by the structural equations (see details in Methods). The $l(G_0)$ of the given graph was then calculated to -31.14. We will estimate the probability of $l(G_0)$, *GCP*, by the following three steps.

At the third step, we generated the graphs based on the given graph, and then calculated the log-likelihoods of the generated graphs according to the two preceding steps.

(1) We generated 50 random graph sets, $\{G_i\}$, to form a data set, in which each graph has the same number of nodes and edges, but with different connections from those of G_0.

(2) 50 corresponding log-likelihoods of $\{G_i\}$ were calculated according to the first and second steps. Among the 50 log-likelihoods, the maximum of the log-likelihood, $l(G_{max})$, is selected.

(3) The above procedure is iterated 1000 times to finally obtain 1000 values of $l(G_{max})$. In this step, the dimensions of the sampling data, the number of graphs in one set, and that of the iterations to select $l(G_{max})$ are changeable parameters, and the robustness of our method with them will be evaluated in the following sections.

At the fourth step, we fit the log-likelihoods calculated in the third step to the GEV model. The maximization of the GEV log-likelihood leads to the following estimate:

$$\left(\hat{\mu}, \hat{\sigma}, \hat{\xi}\right) = (-126.58, 13.15, -0.148),$$

for which the GEV log-likelihood is 4063.59. Although the maximum likelihood estimate for $\hat{\xi}$ is negative, corresponding to a bounded distribution, the ξ value larger than -0.5 indicates that the maximum likelihood functioned well for the estimation [21,22]. Furthermore, the goodness of fit can be visually diagnosed, using the three diagnostic plots for assessing the accuracy of the GEV model, fitted to the 1000 log-likelihoods data by the three parameterizations. Neither the probability plot nor the quantile plot gave any cause to doubt the validity of the fitted model: each set of plotted points was nearly linear. The return-level curve asymptotes to a finite level as a consequence of the negative estimate $\hat{\xi}$, and also provides a satisfactory representation of the empirical estimates. In addition, the corresponding density estimate seems consistent with the histogram of the data. Consequently, the four diagnostic plots lend support to the fitted GEV model.

At the final step, we estimated the *GCP* of the log-likelihood of the given graph based on the fitted GEV distribution. According to the GEV distribution, the *GCP* corresponding to $l(G_0)$ (= -31.14) was calculated to be less than 10^{-10}. As a result, it is natural that the examined given graph was highly consistent with the data generated according to the graph structure.

Robustness of the Present Method

The high performance of the present method described in the preceding subsection depends on a few parameters. By using the same network structure as in Fig. 1, we tested the robustness of the present method in terms of the dimensions of the analyzed data, the two parameters in generating artificial graphs for GEV, and the degree of noise in the data. Furthermore, the robustness of the network structure variation is tested by using the typical network structure in biological interactions in the following four subsections.

Robustness in Terms of the Dimensions of the Analyzed Data

We test the robustness of our method in terms of the number of data samples for one variable (data dimension) that is smaller than the data dimension ($\{X_i\}$ for i = 1,2,...,50) in Fig. 1. This is because the experimental conditions are frequently limited, due to the technical difficulty of performing experiments for different growth conditions. Thus, small data dimensions are expected in the actual data.

We performed the same estimation of *GCP* as that in Fig. 1, by using the data with 15 and 30 dimensions, and in both cases, the present method operated well. The GEV fit well to the data: the estimated $\hat{\xi}$ was larger than -0.5: the estimated $\hat{\xi}$ values are -0.1132 for the 15-dimension data and -0.1332 for the 30-dimension data. In addition, the four GEV-diagnostic plots for assessing the accuracy of the GEV model show the validity of the fitted model in each case (see additional file 2: Robustness in terms of data dimensions). By the estimated GEV distributions, the *GCP*s in the two cases were less than 10^{-4} and 10^{-8}, respectively. The probability for the 30-dimension data was smaller than that for the 15-dimension data. Considering that the probability was 10^{-10} in the 50-dimension data in the preceding section, this indicates that the resolution degree about the consistency is higher with larger dimensions.

Robustness in Terms of Parameters in Generating the GEV Model

The GCP depends on two parameters in the graph generation for GEV: the number of graphs for selecting the maximum of the likelihood in one set of the generated graphs, l, and the number of iterations for sampling the maximum values from each set of generated graphs, n. In Fig. 1, l and n were set to 50 and 1000, and a total of 50,000 graphs were generated for GEV. Here, we examined the fitness of the log-likelihoods to GEV based on the graph shown in Fig. 1, with nine pairs of l and n: l was set to 25, 50 and 100, and n was set to 100, 500, and 1000. The total numbers of graphs for GEV ranged from 2500 to 100000, and all of the examinations with the above parameter pairs are provided in an additional file (see additional file 3: Robustness in terms of the parameters). Here, we focused on the case when fewer graphs are generated than the number in the default case. This is because a small number of generated graphs in each set and iterations may tend to violate the distribution of GEV, due to some biases in the graph generation.

In the comparison of (l, n) = (50, 100) and (25, 1000) with (50, 1000) in Fig. 1, the log-likelihoods calculated in the two cases were fitted to the GEV model. Indeed, the two $\hat{\xi}$ values were larger than -0.5: the estimated $\hat{\xi}$ values were -0.1545 in (l, n) = (50, 100) and -0.1670 in (l, n) = (25, 1000), respectively. The two sets of diagnostic plots for assessing the accuracy also showed the validity of the fitted model in each case (see additional file 3). A closer

inspection revealed that the $\hat{\xi}$ value in the (50, 100) case was slightly less than that in the (25, 1000) case. This indicated that the number of graphs in each set, l, is more sensitive to the goodness of fitness than the number of iterations, n, regardless of the total number of generated graphs. At any rate, the present method operates well, even in the case of a relatively small number of generated graphs.

In general, the optimized values of l and n depend on the size of the examined graph, and may be expressed as the fraction to the total number of possible graphs with the same numbers of nodes and edges as those of the examined graph. Although the number of possible graphs composed of arbitrary numbers of labeled nodes and edges can be estimated asymptotically under some constraints on the edge connectivity [23], unfortunately, the total number of possible graphs, in which all of the nodes are connected to form one graph, is not still obtained. In the present stage, we should heuristically define l and n by diagnosing the goodness of fit to the GEV model.

Robustness in Terms of the Magnitude of Noise in the Analyzed Data

We estimated the *GCP* in various noise ranges. For this purpose, the value of the standard deviation in the structural equations for data generation (σ = 0.1 in Fig. 1) was changed to three values (σ = 0.5, 1.0, and 2.0). By the same procedure as that in Fig. 1, we calculated 100 *GCPs* for the three ranges of standard deviations. Finally, the probabilities of the generated graphs were calculated.

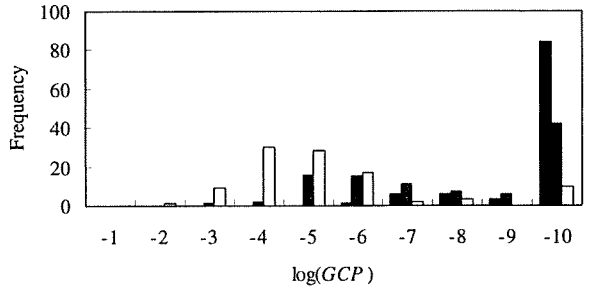


Figure 2
Robustness in terms of the noise in measured data. *GCP*(=$P(I(G_0))$) for the graph in Fig. 1 was calculated with simulated data with distinct standard deviations, and the frequencies of *GCPs* are plotted against the probability degree. The horizontal axis indicates the log(*GCP*) value, and the vertical axis is its frequency: black-colored bar, σ = 0.5; gray-colored bar, σ = 1.0; and boxed bar, σ = 2.0.

The histograms of the *GCPs* in the three ranges of standard deviations are shown in Fig. 2. In this figure, 100 *GCPs* were plotted against the number of connections in the generated graphs that were different from those in the examined graph in the respective cases of standard deviations. In the cases of the two small standard deviations (σ = 0.5 and 1.0), less than 10^{-10} of the *GCPs* emerged most frequently, but the most frequent *GCP* was found at 10^{-4} in the case of the largest standard deviation (σ = 2.0). In the former two cases, the largest *GCP* was 10^{-6} in σ = 0.5 and 10^{-3} in σ = 1.0. Although some exceptional *GCPs* were also found, the present method operates well within the range of the two noise levels. In contrast, the last case shows the limitation of our method, in terms of the noise of the measured data. Careful preprocessing of the measured data may be required to apply our method to actual data. Note that the noise is amplified as the number of parents grows in the present simulation. For example, the standard deviation is $(\alpha_1^2 + \alpha_2^2 + 1)\sigma$, when a descent has two parents and α_i is the path coefficient between the descent and the i-th parent. At any rate, the limitation of the present method in terms of the data noise can be examined by describing the histogram of the *GCP*, and was estimated between 1.0 and 2.0 for the graph in Fig. 1. In addition, we assumed that the distribution of the data noise also follows uniform and gamma distributions, and obtained similar results in terms of the robustness about the data noise (see additional file 4: Robustness in terms of the noise according to the gamma and uniform distributions).

Robustness Regarding the Variation of the Network Structure

We applied the present method to the three network structures shown in Fig. 3. The three networks are analogous to the typical structures of biological networks; the first is analogous to part of a chain reaction in a metabolic pathway, the second represents the simple structure of a gene regulatory network, and the third depicts a cascade in a signal transduction pathway. According to the connectivity in the network, the data were generated with the corresponding structural equations, and the present method was applied to estimate the graph consistency with the generated data.

The present method operated well in all of the network structures. Indeed, the log-likelihoods in the three networks fit well to the GEV (see statistics in the legend of Fig. 3, and additional file 5: Robustness regarding the network structure variation). In addition, the *GCPs* were very small: The *GCPs* of the three networks were less than 10^{-11}, 10^{-4}, and 10^{-7}, respectively. Interestingly, the magnitudes of the *GCPs* may be related to the network structures. The *GCP* in Fig. 3B is relatively larger than the *GCPs* in Figs. 3A and 3C. This is because the present path coeffi-

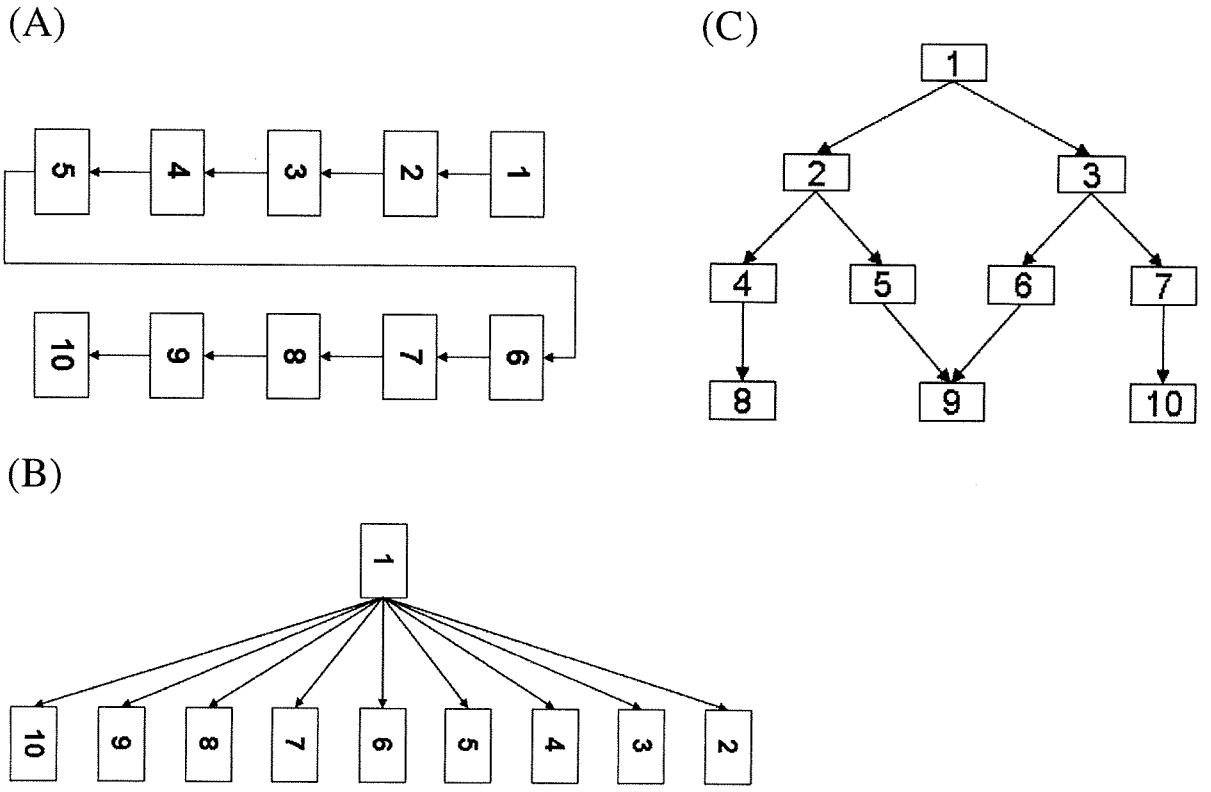


Figure 3

Robustness regarding graph structure variation. The calculation is composed of five steps (see details in the text). Three networks with typical structures in biology are examined in (A), (B), and (C). To generate the simulation data by structural equations, we set the standard deviation to 0.1 in all three graphs, and the path coefficients between the variables are as follows: (A) $\alpha_{1,2} = 0.6$, $\alpha_{2,3} = 0.3$, $\alpha_{3,4} = 0.1$, $\alpha_{4,5} = 0.7$, $\alpha_{5,6} = 0.8$, $\alpha_{6,7} = 0.9$, $\alpha_{7,8} = 0.2$, $\alpha_{8,9} = 0.5$, and $\alpha_{9,10} = 0.4$; (B) $\alpha_{1,2} = 0.1$, $\alpha_{1,3} = 0.2$, $\alpha_{1,4} = 0.3$, $\alpha_{1,5} = 0.4$, $\alpha_{1,6} = 0.5$, $\alpha_{1,7} = 0.6$, $\alpha_{1,8} = 0.7$, $\alpha_{1,9} = 0.8$, and $\alpha_{1,10} = 0.9$; and (C) $\alpha_{1,2} = 0.5$, $\alpha_{1,3} = 0.7$, $\alpha_{2,4} = 0.4$, $\alpha_{2,5} = 0.8$, $\alpha_{3,6} = 0.6$, $\alpha_{3,7} = 0.3$, $\alpha_{4,8} = 0.2$, $\alpha_{5,9} = 0.1$, $\alpha_{6,9} = 1.0$, and $\alpha_{7,10} = 0.9$. The value of log-likelihood and the parameters of GEV distribution in the respective networks are as follows: (A) $l(G_0) = 163.4805$, $\mu = 89.8375$, $\sigma = 12.9694$, and $\xi = -0.1743$; (B) $l(G_0) = 61.6096$, $\mu = 3.0217$, $\sigma = 12.5220$, and $\xi = -0.1314$; and (C) $l(G_0) = 124.8894$, $\mu = 46.9002$, $\sigma = 12.1395$, and $\xi = -0.1406$. See also the corresponding GEV plots at additional file 5: Robustness regarding the network structure variation.

cients between the 10 nodes were set at different values, but in the same order of digits. This indicates that the most similar data for respective variables were generated in Fig. 3B, and caused pseudo-correlations between the variables with no edges in the network in Fig. 3B. Although the performance for estimating the graph consistency may slightly decrease, depending on the number of two -layer relationships in the examined data, this simulation shows that the present method can be applied to various structures of networks.

Examinations of Actual Graphs

We examined the performance of the present method with two sets of actual networks in *Escherichia coli* and the corresponding actual measured data. One set is a regulatory network for the SOS response system with the expression degrees of the constituent genes measured by fluorescence [24], and the other is 29 networks classified by gene functions, with the expression degrees under anaerobic conditions measured by microarray [25]. The former examination is a verification of the present method for an actual network with a size similar to the networks shown in Fig. 1, and the latter is a demonstration of a high-throughput search of network candidates, consistent with the data measured under particular conditions.

Verification for a Simple Network

The gene network in the SOS system is schematically shown in Fig. 4A. The SOS DNA repair system in *Escherichia coli* is a well-characterized transcriptional network [26,27]. One of the SOS proteins, *RecA*, acts as a sensor of DNA damage, and a master repressor (*LexA*) binds sites in the promoter regions of these operons. The corresponding data to the constituent molecules in the network are the transcriptional activity of genes measured with real-time monitoring by means of low-copy reporter plasmids, in which a promoter controls green fluorescent protein [24].

The GEV plots with the likelihood values and the statistics are shown in Fig. 4B. The value of $\hat{\xi}$ was larger than -0.5, and the GEV plots were quite similar to those in Fig. 1. Indeed, each set of plotted points was nearly linear, and

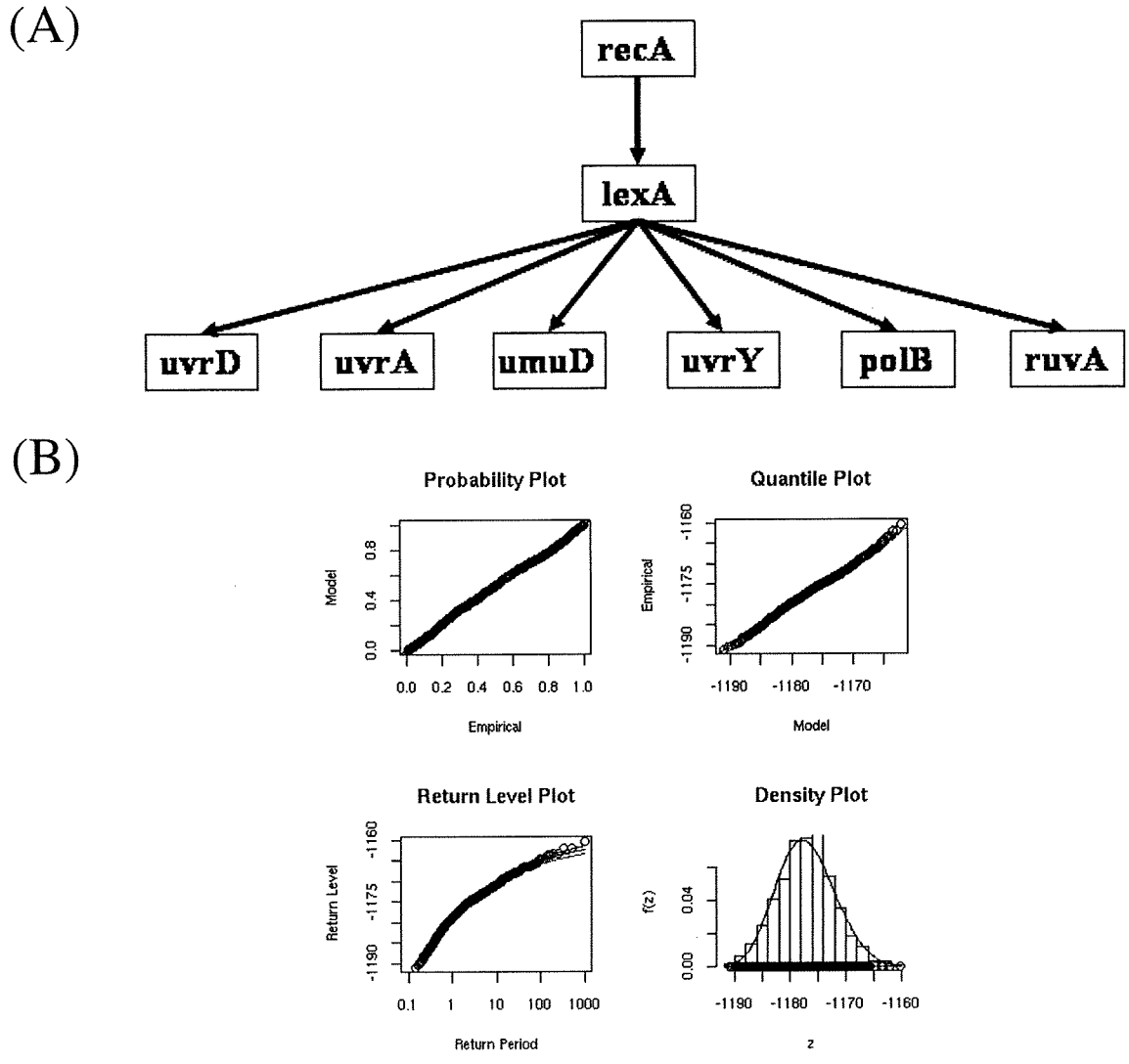


Figure 4
Evaluation of the transcriptional network of the SOS DNA repair system in *Escherichia coli*. The network is schematically shown in (A), and the corresponding GEV plots and the box-plot are also shown in (B). The value of log-likelihood between the examined network and the measured data is -1168.453, and the parameters of GEV distribution are as follows: μ, -1179.079, σ, 4.957; ξ, -0.236. The data for the promoter activities of eight genes in the SOS system are cited from [24].

the return-level curve asymptotes to a finite level. In addition, the corresponding density estimate seems consistent with the histogram of the data. Consequently, the goodness of fitness in the actually measured data lends support to the GEV model.

The *GCP* of the SOS network with the corresponding measured data was estimated as 0.049, and the network structure was estimated to be consistent with the data measured from the examined network. However, the *GCP* was large in comparison with the *GCP*s in the simulation studies in the preceding sections. This is partly because the cyclic relationship of *RecA* is neglected in the examined network, and partly because most of the relationships in the examined network are composed of 2-layer relationships, due to the production of similar degrees of expression data, as in the situation in Fig. 3B. At any rate, the performance of the present method was verified by a well-known network, with a size similar to that in the simulation, and with the corresponding data measured by an experimental study.

Demonstration for an Actual Network Set

We further tested the performance of our method for selecting the networks consistent with the data measured under specific conditions from many network candidates. Here, we arranged 29 regulatory networks in *Escherichia coli* and the corresponding gene expression profiles measured under anaerobic conditions (for details about the examined network reconstruction and the profile data, see Methods).

Table 1 shows the analyzed networks and the corresponding graph consistency probabilities of the 29 networks (see additional file 6: the 29 network structures analyzed in the present study). When we set the significance probability to 5%, only two networks (Nos. 14 and 28) in Fig.

Table 1: Consistency of the twenty-nine networks with expression profiles measured under anaerobic conditions in *Escherichia coli*

No.	ID	Description	node	edge	*GCP*
1	C9333	detoxification	6	8	1.000
2	C9448	amino acids	6	9	1.000
3	C9449	carbon compounds	6	9	1.000
4	C9426	colanic acid (M antigen)	6(7)	9(11)	1.000
5	C9509	operon	6(7)	9(11)	1.000
6	C9448, C9462	amino acids, formyl-THF biosynthesi	7	10	1.000
7	C9449	carbon compounds	8(9)	7(8)	1.000
8	C9331	motility, chemotaxis, energytaxis	9	8	0.998
9	C9340	flagella	9	8	0.647
10	C9362	nucleoproteins, basic proteins	9	8	0.925
11	C9401	tryptophan	9	8	1.000
12	C9449	carbon compounds	9	8	1.000
13	C9376	cytoplasm	10	9	1.000
14	C9449	**carbon compounds**	10	9	**0.006**
15	C9449	carbon compounds	10	11	0.976
16	C9337	SOS response	11	10	0.127
17	C9354	DNA repair	11	10	0.068
18	C9383	arginine	11	10	1.000
19	C9474	nucleotide and nucleoside conversion	11	15	0.378
20	C9493	fermentation	11	10	1.000
21	C9376	cytoplasm	12	11	0.302
22	C9393	isoleucine/valine	13	12	1.000
23	C9420	purine biosynthesis	13	12	1.000
24	C9394	leucine	14	17	1.000
25	C9504	phosphorous metabolism	23	22	1.000
26	C9528	repressor	52(53)	77(79)	1.000
27	C9523	activator	58(59)	92(93)	1.000
28	C9490	**anaerobic respiration**	89(91)	161(162)	**0.016**
29	C9372	Transcription related	91(93)	143(146)	0.772

GCP values with less than 5% significance probability are indicated in bold type. The ID in the classification scheme by EcoCyc [44] and the corresponding gene function are denoted in the second and third columns, respectively. Two networks in the functions C9448 and C9462 are composed of the same constituent genes with the same connectivity. In the following columns, the numbers of nodes and edges of the analyzed networks are denoted: the original network was constructed based on the information about the relationship between the transcription factor and its regulated genes in EcoCyc, and the analyzed network was constructed from the original network by excluding the genes that were not found in the expression profile data from NCBI GEO (accession number: GSE1107) [25]. The numbers of nodes and edges of the original networks are denoted in parentheses. The graph consistency probability (*GCP*) is denoted in the last column.

5, which are composed of regulatory gene pairs related to carbon compounds and anaerobic respiration, were selected among the 29 networks. As seen in the figure, the network structures are quite different. The network related to the carbon compounds in Fig. 5A shows a relatively simple structure that is a two-layer relationship between one TF and its nine regulated genes. In contrast, the other network related to anaerobic respiration in Fig. 5B has a highly complicated form, with 89 nodes and 161 edges. The selection of the two networks can be interpreted in terms of biological functions, as described below.

The first network (No. 14) is composed of *malT* and its regulated genes. *malT* is involved in maltose transport [28]. Besides the *malT*-regulating network, four networks related to carbon compounds (Nos. 3, 7, 12 and 15) were also included in the examined networks, but they showed no significant probability: the TFs in each network are *araC*, *galR*, *gutM*, and *exuR*, and they regulate products related to the transport of arabinose, galactose, glucitol, and hexuronate, respectively [29-32]. Among the four networks, the *galR* and *exuR*-regulating networks (Nos. 12 and 15) are coordinated in terms of their products: the *exuR* regulatory gene product controls the expression of the galacturonate pathway operons (*exuT*, *uxaC*, *uxuA*, and *uxaB*) [33]. Interestingly, galactose was the least efficiently utilized under anaerobic conditions, among glucose, lactose, galactose, maltose, maltotriose, and maltohexaose [34]. This fact may be one of the reasons why our method revealed the consistency of network No. 14 with the data, and the lack of consistency of two of the networks, Nos. 12 and 15. In the remaining two networks, Nos. 3 and 7, there are no reasons for their lack of consistency with the present data. Thus, the detection of the network related to maltose metabolism is reasonable, at least in comparison with the galactose- and hexuronate-related networks.

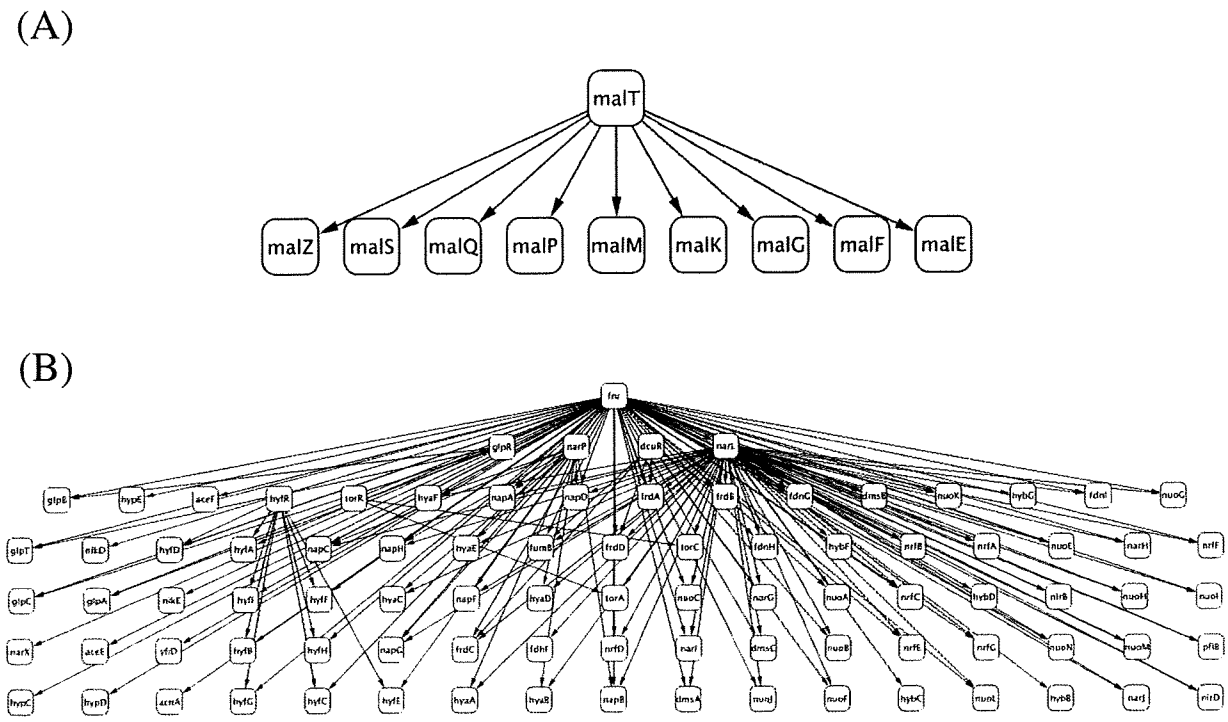


Figure 5
Networks with 5% significance probability in graph consistency search. By corresponding between the regulatory relationships and the gene functions in EcoCyc [44], 29 regulatory networks were reconstructed, and their consistency with the expression profiles measured under anaerobic conditions (accession number GSE1107 in NCBI Gene Expression Omnibus (GEO); http://www.ncbi.nlm.nih.gov/geo/) [25] was examined. Among the 29 regulatory networks, two networks showed 5% significance probability: the network related with carbon compounds (EcoCyc ID: C9449_11) (A) and that with anaerobic respiration (EcoCyc ID: C9490_1) (B). The details of the network structures of the 29 regulatory networks are shown in the additional file 6: the 29 network structures analyzed in the present study.

As for the second network (No. 28), the biological function is defined as anaerobic respiration, and its detection is clearly reasonable. The gene encoding the transcription factor in the network is *fnr*, one of the seven global regulators in *E. coli* [35], and the modular controlled by its product, Fnr, encodes proteins involved in cellular adaptation to growth in anoxic environments [36-38]. Since the network is related to adaptations to environmental changes, many genes are comprehensively associated with each other, and the network structure is complex, as seen in Fig 5B. Thus, the consistency of the *fnr*-regulating network with the present data demonstrates the validity of the present method for searching a large-size network consistent with data measured under particular cellular conditions.

It may seem overly strict to estimate the network consistency by the present method. Some other networks besides the two detected networks might be operating under anaerobic conditions. However, the strictness of the consistency estimation is one of the prerequisites for exploring unknown networks. The falsely detected networks should be excluded as much as possible, and the detection of a few definite network candidates may serve as the initial step for investigating the unknown networks that are unexpected, in terms of biological knowledge. In addition, the strictness for consistency estimation is easily modified by setting the selection degree with a significance probability. As a result, the present method reveals the strictly consistent networks with the expression profiles measured under specific conditions, and will be useful to find the activated network candidates among many given networks.

Merits and Pitfalls of the Present Method

The present method successfully evaluates the consistency of a network with the artificial and actual data, which is expressed as a probability, *GCP*. The *GCP* of each known network is estimated from one set of data in which the constituent molecules of the network were measured under one particular condition. Although a large amount of noise prevents a confident estimation of the *GCP*, the present method is robust in terms of the data sampling dimensions, the parameters in the method, and the network structure variation. The plausibility of the structure variation and scale is illustrated by the detection of actual networks for the simple network of the SOS response and the large and complicated network for anaerobic respiration. Thus, the present method is feasible to evaluate the consistency of the networks with a set of data measured under particular conditions.

The present method may be further applied to various analyses of biological issues. One example is a simple extension of the demonstration shown in the preceding section, as follows. Assume that we know more than two distinctive cell stages, and that we can measure the data of the constituent molecules in different stages. Then, we evaluate the consistency of a set of known networks with the respective data. By this evaluation, we may detect the activated networks, among the known networks that are specific to the respective cell stages. For example, the present method may address the problem of which known networks are activated in progressive diseases and in cell differentiation processes. Thus, the present method will be useful to investigate the network variation in various cell stages responding to different environments. Another example is a utilization of the graphs generated in GEV modeling. Assume that we know a network model for a biological phenomenon, and that a few molecules have been newly detected, and are responsible for the phenomenon. Then, we face the issue of how the newly detected molecules should be connected to the previous network. In this situation, our method may present a solution. A new network is tentatively constructed, by connecting the newly detected molecules into the previous network with the full use of biological knowledge, and then the consistency of the tentative network is estimated with the data measured under the conditions where the relationship of the new molecules with the phenomenon was found. If the *GCP* shows the significance probability, then the network is a promising model for the phenomenon. If not, then we can list some network candidates with the significance probability that commonly share the structure of the previous network, among the generated networks for the GEV distribution. Note that the present method aims to evaluate the consistency between the known network structures and the measured data. Thus, the network inference without any given network structures is beyond the present study. At any rate, these two examples will be demonstrated by appropriate networks and data in the near future.

In terms of the methodology, the present method is a rational extension of the previous study based on linear regression [18], by the combination of the Gaussian network and the extreme value distribution. Indeed, the application range on the network structure is expanded, from simple networks with two-layer relationships to more complex networks with multiple-layer relationships. In addition, the present method is complementary with the d-sep test; the graph consistency is estimated for the associations between variables (existence of edges in the graph) in our method, and in contrast, no associations between variables (no edges in the graph) are considered in the d-sep test [14]. However, the d-sep test failed to select the activated networks: when we set the significance probability to 5%, 27 networks among the 29 networks were consistent with the data measured under specific conditions, and only two networks were not (see addi-

tional file 7: d-sep test and SEM for 29 network structures). Interestingly, SEM also failed (see also the additional file 7): 27 networks among the 29 networks were consistent with the data, and one of the two remaining networks could not be evaluated, due to a numerical calculation violation. Thus, our method may be appropriate for tightly estimating the graph consistency in comparison with the d-sep test and SEM. Furthermore, our method differs from the d-sep test in a strict sense. The present method is based on the generation of artificial graphs in the estimation of graph consistency with the measured data, while the d-sep test is based on the direct hypothesis of a population distribution [14,16]. Thus, our method is an asymptotic approach, and is similar to various methods for model selection in network inference, such as various Bayesian network models [8,9]. Note that the present *GCP* is an occurrence probability, and definitely differs from the model selection procedure by using the scores that show a relative difference.

The consistency of a model with the observed data also reminds us of the identifiability problem in the compartmental models for tracer kinetics [39]. The identifiability problem addresses the issue of whether the unknown parameters can be determined uniquely or non-uniquely from the tracer data. Although a systematic algorithm for the identifiability problem was proposed regardless of the model structure [40], its application is limited to the ideal context of noise-free data. Recently, we have partially exploited the identifiability problem algorithm to treat data including noise [41]. Indeed, a network including a cyclic relationship has been examined to estimate the consistency with noisy data. Although this method has a limitation of the network size to smaller than 10 nodes and 15 edges, another method with a symbolic approach may partly compensate for the statistical approach presented here for the limitation of the network structure.

Conclusion

We have proposed a novel method to estimate the consistency of a given network with the measured data as a probability (*GCP*: graph consistency probability), based on the Gaussian network and the generalized extreme value distribution. The performance of the present method was validated by application to artificial graphs with simulated data and actual graphs with measured data from *Escherichia coli*. The plausible evaluation of the consistency between the network structures and the corresponding measured data promises to help reveal the network structure variations depending on the environments in a living cell, as well as to form a bridge between the static network from the literature and the corresponding measurements.

Methods

Data Generation for Simulation

We generates the numerical data according to a standard statistical procedure [16]. The data for 10 nodes with 50 sampling dimensions, $\{X_{kl}$, for k = 1,2,...10, and l = 1,2,...50$\}$, are generated by using the following structural equations that correspond to the parent-descent relationships in Fig. 1:

$$\begin{cases} X_{1l} = N(0,\sigma) \\ X_{2l} = N(0,\sigma) \\ X_{3l} = N(0,\sigma) \\ X_{4l} = \alpha_{1,4}X_{1l} + N(0,\sigma) \\ X_{5l} = \alpha_{1,5}X_{1l} + N(0,\sigma) \\ X_{6l} = \alpha_{2,6}X_{2l} + \alpha_{3,6}X_{3l} + N(0,\sigma) \\ X_{7l} = \alpha_{4,7}X_{4l} + N(0,\sigma) \\ X_{8l} = \alpha_{4,8}X_{4l} + N(0,\sigma) \\ X_{9l} = \alpha_{6,9}X_{6l} + \alpha_{7,9}X_{7l} + N(0,\sigma) \\ X_{10l}10 = \alpha_{9,10}X_{9l} + N(0,\sigma) \end{cases} \quad (1)$$

where $N(0, \sigma)$ means a value that follows a normal distribution with a zero mean and a standard deviation of σ, and $\alpha_{i,j}$ is a path coefficient relating variables i and j. Here, we set σ to 0.1, and the following parameterization was used: $\alpha_{i,j}$ = 0.5. Thus, we obtain a graph and examine the corresponding data to estimate their consistency with the graph. Note that the above data generated by linear equations may not precisely reflect the measured data underlying various non-linear relationships. Here, we adopted the linear relationships as the first approximation to test the performance of the present method. The performance for the complex relationships will be tested by actually measured data.

Recursive Factorization of Causal Graph

Suppose a causal graph is a directed acyclic graph (DAG), $G(V_i, E_j)$, where V_i is a vertex (i = 1, 2, ..., n_v) and E_j is an edge (j = 1, 2, ..., n_e) in the graph. The DAG can be factorized into subgraphs according to the parent-descent relationships [15]. Then, the joint density function $f(X_i)$, corresponding to V_i for the graph G, can be factorized into the conditional density functions according to the graph, as follows:

$$f(X_1, X_2, \cdots, X_{n_v}) = \prod_{i=1}^{n_v} f(X_i \mid pa\{X_i\}), \quad (2)$$

where $pa\{X_i\}$ is the set of variables corresponding to the parents of V_i in the graph.

Gaussian Network (GN)

The causal graph meets the measured data based on the Gaussian network model [20]. On the assumption that the probability variable X_i is subjected to a multiple normal distribution, each conditional function in equation (2) is obtained by linear regression for the measured data of the constituent nodes (molecules) measured at m points, i.e.,

$$f(X_i \mid pa\{X_i\}) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[\frac{-1}{2\sigma_i^2} \sum_{k=1}^{m} \left(x_{ik} - \sum_{j=1}^{n_i} \beta_{ij} x_{jk} \right)^2 \right], \quad (3)$$

where x_{ik} is the measured value of X_i, at the k-th point, and n_i is the number of variables corresponding to the parents of V_i. Thus, the joint density function in equation (2) is expressed by the regression for the measured data in equation (3). Finally, the logarithm of the likelihood of the equation (3) is calculated for the measured data as

$$l(G_0) = \ln \prod_{i=1}^{n_v} f(X_i \mid pa\{X_i\}) = -\frac{1}{2} \sum_{i=1}^{n_v} \sum_{j=1}^{n_i} \left\{ \frac{1}{\sigma_i^2} \sum_{k=1}^{m} \left(x_{ik} - \sum_{j=1}^{n_i} \beta_{ij} x_{kj} \right)^2 + \ln\left(2\pi\sigma_i^2\right) \right\}. \quad (4)$$

Thus, the GN allows us to quantify a given network into the corresponding numerical value from the measured data, according to the network form. Note that the calculation of likelihood itself requires no assumptions on the relationships between variables. Indeed, the likelihood can be calculated in the case of non-linear regressions, such as spline regression.

Generalized Extreme Value Distribution (GEV)

Next, we estimate the probability of $l(G_0)$ by using the generalized extreme value distribution [21]. First, the log-likelihoods of an ensemble of n networks generated according to the *GN* are calculated, and then the maximum log-likelihood is selected from them. The above procedure is iterated l times, i.e.,

$$\begin{aligned} l^1_{\max} &= Max\left\{ l\left(G_1^1\right), l\left(G_2^1\right), \ldots, l\left(G_n^1\right) \right\} \\ l^2_{\max} &= Max\left\{ l\left(G_1^2\right), l\left(G_2^2\right), \ldots, l\left(G_n^2\right) \right\} \\ &\vdots \\ l^l_{\max} &= Max\left\{ l\left(G_1^l\right), l\left(G_2^l\right), \ldots, l\left(G_n^l\right) \right\} \end{aligned} \quad (5)$$

The distribution of the maximum values by l iterations is expected asymptotically to be a generalized extreme value distribution, i.e.,

$$G\left(l_{\max}\right) = \exp\left\{ -\left[1 + \xi\left(\frac{l_{\max} - \mu}{\sigma} \right) \right]^{-1/\xi} \right\} \quad (6)$$

defined on the set,

$$\left\{ l_{\max} : 1 + \xi\left(\frac{l_{\max} - \mu}{\sigma} \right) > 0 \right\}$$

where the parameters satisfy $-\infty < \mu < \infty$, $\sigma > 0$, and $-\infty < \xi < \infty$. The model has three parameters: μ, σ, and ξ are a location parameter, a scale parameter, and a shape parameter, respectively. Maximization of the log-likelihood of equation (6) with respect to the parameter vector (μ, σ, ξ) leads to the maximum likelihood estimate for any given dataset, using standard numerical optimization. In the present study, the *R extRemes* package [42] was used to fit the data to the GEV distribution.

Note that the standard likelihood ratio test [43] cannot be applied straightforwardly to a Gaussian network in the present case. This is because the density function of the population and the degrees of freedom in the likelihood ratio test are unclear when maximizing the likelihoods of the generated graphs. In the present method, the GEV distribution of the maximum values of likelihoods in the blocks of generated graphs is adopted analogically, instead of the maximum likelihood in the likelihood ratio test. The utilization of the GEV distribution requires the model fitting to the data, but allows us to set the significance probability arbitrarily, as usual in statistical tests.

Graph Consistency Probability (GCP)

If the goodness of fitness of the maximum values from the generated graphs is ascertained, then the occurrence probability of a given graph (*GCP: graph consistency probability*) can be directly estimated by corresponding the $l(G_0)$ in equation (1) to the probability density function of GEV obtained in (6), i.e.

$$P\left(l\left(G_0\right)\right) = \int_{(G_0)}^{+\infty} G\left(l_{\max}\right) dl_{\max}. \quad (7)$$

Thus, the present method expresses the consistency in the form of a probability. The probability examines the possibility of whether the tested known networks are activated in the environment where the data were measured. If the probability is small, which corresponds to a large likelihood value, then the data are generated, according to the molecular relationships in the network.

Actual Networks and Data for High-Throughput Consistency Search

We first classified a transcription factor (TF) and its regulated genes compiled in EcoCyc [44], according to the classification scheme of gene functions http://biocyc.org/ECOLI/class-tree?object=Genes. Using the gene sets of the TF and the regulated genes in each function, we next reconstructed the networks: respective networks were reconstructed, so as to form the network structure with as many connections between the genes as possible. Thus, we obtained 130 regulatory networks that are characterized by biological functions. Since some networks were characterized by more than two functions, the 130 regulatory networks were redundant in terms of the connectivity and the constituent genes. Then, 29 networks were kept, after excluding the redundancy and the small networks with less than 8 edges (see Table 1 and additional file 6: 29 network structures analyzed in the present study).

The consistency of each of the 29 networks was estimated with one set of expression profiles measured under 22 different anaerobic conditions (GSE1107) [25] cited from NCBI GEO [45]. The expression profiles were standardized by the average and the standard deviation in each condition, as preprocessing of the measured data. In a few nodes (genes) in the original network constructed from the information in EcoCyc, the corresponding expression profiles were not found in the analyzed data (GSE1107), and the corresponding parts in the network were excluded.

Authors' contributions

SS carried out the implementation and the calculations, and participated in the design of the study. SA participated in the design of the study, and helped to draft the manuscript. KH conceived of the study, participated in its design and coordination, and drafted the manuscript. All authors read and approved of the final manuscript.

Additional material

Additional file 1

Details of the schematic description of the procedure. The graph factorization at Step 1 and the four GEV-diagnostic plots of the probability plot, the quantile plot, the return-level curve, and the density plot at Step 4 (PDF file) are shown.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-84-S1.pdf]

Additional file 2

Robustness in terms of data dimensions. Four GEV-diagnostic plots of the probability plot, the quantile plot, the return-level curve, and the density plot (PDF file) are shown for the 15- and 30-dimension data, respectively.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-84-S2.pdf]

Additional file 3

Robustness in terms of the parameters. Four GEV plots (PDF file) are shown when two parameters were set as follows: l *was set to 25, 50 and 100, and* n *was set to 100, 500, and 1000.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-84-S3.pdf]

Additional file 4

Robustness in terms of the noise according to the gamma and uniform distributions. GCP(=$P(l(G_0))$) *for the graph in Fig. 1 was calculated with simulated data according to the gamma and uniform distributions, and the frequencies of* GCPs *are plotted against the probability degree. The horizontal axis indicates the log(*GCP*) value, and the vertical axis is its frequency: black-colored bar,* $\lambda = 1$ *in gamma distribution; gray-colored bar,* $\lambda = 3$*; striped bar,* $\lambda = 5$*; and boxed bar, between 0 and 1 in uniform distribution.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-84-S4.pdf]

Additional file 5

Robustness regarding the network structure variation. GEV plots (PDF file) are shown for the three types of network structures in Fig. 3.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-84-S5.pdf]

Additional file 6

The 29 network structures analyzed in the present study. The 29 regulatory networks of Escherichia coli *with more than 8 edges (PDF file) are shown, as constructed from the information on the regulatory relationships between two genes in EcoCyc [44].*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-84-S6.pdf]

Additional file 7

SEM and d-sep test for 29 network structures. The 29 regulatory networks of Escherichia coli *were also tested by SEM and the d-sep test.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-84-S7.pdf]

Acknowledgements

S.A. was supported by a Grant-in-Aid for Scientific Research (grant 18681031), and K.H. was partly supported by a Grant-in-Aid for Scientific Research on Priority Areas "Systems Genomics" (grants 18016008 and 20016028) and by a Grant-in-Aid for Scientific Research (A) (grant 19201039), from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. Bonetta L: **Bioinformatics-from genes to pathways.** *Nature Methods* 2004, **1**:169-176.
2. Yuryev A, Mulyukov Z, Kotelnikova E, Maslov S, Egorov S, Nikitin A, Daraselia N, Mazo I: **Automatic pathway building in biological association networks.** *BMC Bioinformatics* 2006, **7**:171.
3. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, Walk TC, Zhang P, Karp PD: **The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.** *Nucl Acids Res* 2008, **36**:D623-D631.
4. Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, Cho RJ, Chen RO, Brownstein BH, Cobb JP, Tschoeke SK, Miller-Graziano C, Moldawer LL, Mindrinos MN, Davis RW, Tompkins RG, Lowry SF: **Inflammation and Host Response to Injury Large Scale Collaborative Research Program. A Network-Based Analysis of Systemic Inflammation in Humans.** *Nature* 2005, **437**:1032-1037.
5. Rudd MF, Webb EL, Matakidou A, Sellick GS, Williams RD, Bridle H, Eisen T, Houlston RS, GELCAPS Consortium: **Variants in the GH-IGF axis confer susceptibility to lung cancer.** *Genome Res* 2006, **16**:693-701.
6. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z, Croshaw R, Willis J, Dawson D, Shipitsin M, Willson JKV, Sukumar S, Polyak K, Park BH, Pethiyagoda CL, Krishna Pant PV, Ballinger DG, Sparks AB, Hartigan J, Smith DR, Suh E, Papadopoulos N, Buckhaults P, Markowitz SD, Parmigiani G, Kinzler KW, Velculescu VE, Vogelstein B: **The Genomic Landscapes of Human Breast and Colorectal Cancers.** *Science* 2007, **318**:1108-1113.
7. Bateman A: **Editorial.** *Nucl Acids Res* 2008, **36**:D1.
8. Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data.** *J Comp Biol* 2000, **7**:601-620.
9. Ghahramani Z: **Learning Dynamic Bayesian Networks.** *Adaptive Processing of Sequences and Data Structures* 1998:168-197.
10. Akutsu T, Miyano S, Kuhara S: **Algorithms for inferring qualitative models of biological networks.** *Proc Pacific Symp Biocomput* 2000:290-301.
11. Toh H, Horimoto K: **Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling.** *Bioinformatics* 2002, **18**:287-297.
12. Wright S, Adhya S: **The method of path coefficients.** *Ann Math Statist* 1934, **5**:161-215.
13. Joreskog KG: **A general method for analysis of covariance structures.** *J Biometrika* 1970, **57**:239-251.
14. Shipley B: **A new inferential test for path models based on directed acyclic graphs.** *Structural Equation Modeling* 2000, **7**:206-218.
15. Pearl J: *Probabilistic Reasoning in Intelligent Systems* California, Kaufmann Morgan Publishers; 1988.
16. Shipley B: *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations, and Causal Inference* Oxford, Oxford University Press; 2000.
17. Bisits AM, Smith R, Mesiano S, Yeo G, Kwek K, MacIntyre D, Chan EC: **Inflammatory aetiology of human myometrial activation tested using directed graphs.** *PLoS Comput Biol* 2005, **1**:132-136.
18. Herrgard MJ, Covert MW, Palsson BO: **Reconciling gene expression data with known genome-scale regulatory network structures.** *Genome Research* 2003, **13**:2423-2434.
19. Shen-Orr SS, Milo R, Mangan S, Alon U: **Network motifs in the transcriptional regulation network of Escherichia coli.** *Nat Genet* 2002, **31**:64-68.
20. Whittaker J: *Graphical Models in Applied Multivariate Statistics* New York, John Wiley and Sons; 1990.
21. Coles S: *An Introduction to Statistical Modeling of Extreme Values* London, Springer-Verlag; 2001.
22. Smith RL: **Maximum likelihood estimation in a class of non-regular cases.** *Biometrika* 1985, **72**:67-90.
23. Bender EA, Canfield ER, McKay BD: **The asymptotic number of labeled graphs with n vertices, q edges, and no isolated vertices.** *J Combinatorial Theory, Series A* 1997, **80**:124-150.
24. Ronen M, Rosenberg R, Shraiman BI, Alon U: **Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics.** *Proc Natl Acad Sci* 2002, **99**:10555-10560.
25. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO: **Integrating high-throughput and computational data elucidates bacterial networks.** *Nature* 2004, **429**:92-96.
26. Kenyon CJ, Walker GC: **DNA-damaging agents stimulate gene expression at specific loci in Escherichia coli.** *Proc Natl Acad Sci* 1980, **77**:2819-2823.
27. Little JW, Mounta DW: **The SOS regulatory system of Escherichia coli.** *Cell* 1982, **29**:11-22.
28. Chapon C: **Expression of malT, the regulator gene of the maltose region in Escherichia coli, is limited both at transcription and translation.** *EMBO J* 1982, **1**:369-374.
29. Lee NL, Gielow WO, Wallace RG: **Mechanism of araC autoregulation and the domains of two overlapping promoters, Pc and PBAD, in the L-arabinose regulatory region of Escherichia coli.** *Proc Natl Acad Sci* 1981, **78**:752-756.
30. Hugovieux-Cotte-Pattat N, Robert-Baudouy J: **Regulation and transcription direction of exuR, a self-regulated repressor in Escherichia coli K-12.** *J Mol Biol* 1982, **156**:221-228.
31. Yamada M, Saier MH: **Positive and negative regulators for glucitol (gut) operon expression in Escherichia coli.** *J Mol Biol* 1988, **203**:569-583.
32. Weickert MJ, Adhya S: **Control of transcription of gal repressor and isorepressor genes in Escherichia coli.** *J Bacteriol* 1993, **175**:251-258.
33. Portalier RC, Robert-Baudouy J, Stoeber F: **Regulation of Echerichia coli K-12 hexuronate system genes: exu regulon.** *J Bacteriol* 1980, **143**:1095-1107.
34. Muir M, Williams L, Ferenci T: **Influence of Transport Energization on the Growth Yield of Escherichia coli.** *J Bacteriol* 1985, **163**:1237-1242.
35. Martinez-Antonio A, Collado-Vides J: **Identifying global regulators in transcriptional regulatory networks in bacteria.** *Curr Opin Microbiol* 2003, **6**:482-489.
36. Lynch AS, Lin EC: **Responses to molecular oxygen. In *Escherichia coli* and *Salmonella typhimurium*.** In *Cellular and Molecular Biology* 2nd edition. Washington DC; 1996:1526-1539.
37. Unden G, Schirawski J: **The oxygen-responsive transcriptional regulator FNR of Escherichia coli: the search for signals and reactions.** *Mol Microbiol* 1997, **4**:205-210.
38. Unden G, Achebach S, Holighaus G, Tran HG, Wackwitz B, Zeuner Y: **Control of FNR function of Escherichia coli by O_2 and reducing conditions.** *J Mol Microbiol Biotechnol* 2002, **4**:263-268.
39. Cobelli C, Foster D, Toffolo G: *Tracer Kinetics in Biomedical Research: From Data to Model* New York, Kluwer Academic/Plenum Publishers; 2000.
40. Buchberger B: **An Algorithmic Criterion for the Solvability of a System of Algebraic Equations.** In *Peer review in Gröbner Bases and Applications Volume 251*. Edited by: Buchberger B, Winkler F. London, Mathematical Society Lecture Notes Series; 1998:535-545.
41. Yoshida H, Nakagawa K, Anai H, Horimoto K: **An algebraic-numeric algorithm for the model selection in kinetic networks.** *Proceedings of 10th CASC. LNCS 4770* 2007:433-447.
42. Gilleland E, Katz RW: **Analyzing seasonal to interannual extreme weather and climate variability with the extremes toolkit (extRemes).** *18th Conference on Climate Variability and Change, 86th American Meteorological Society (AMS) Annual Meeting* 2006:2-15.
43. Lehmann EL: *Testing Statistical Hypotheses* 2nd edition. New York, John Wiley and Sons; 1986.
44. Karp PD, Keseler IM, Shearer A, Latendresse M, Krummenacker M, Paley SM, Paulsen I, Collado-Vides J, Gama-Castro S, Peralta-Gil M, Santos-Zavaleta A, Penaloza-Spinola MI, Bonavides-Martinez C, Ingraham J: **Multidimensional annotation of the Escherichia coli K-12 genome.** *Nucl Acids Res* 2007, **35**:7577-7590.
45. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R: **NCBI GEO: Mining tens of millions of expression profiles – database and tools update.** *Nucleic Acid Res* 2007, **35**:D760-D765.