OPEN ACCESS Freely available online

PLOS ONE

Inference of Gene Regulatory Networks Incorporating Multi-Source Biological Knowledge via a State Space Model with *L*1 Regularization

Takanori Hasegawa[1]*, Rui Yamaguchi[2], Masao Nagasaki[3], Satoru Miyano[2], Seiya Imoto[2]

1 Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto, Japan, **2** Human Genome Center, The Institute of Medical Science, The University of Tokyo, Minato-ku, Tokyo, Japan, **3** Department of Integrative Genomics, Tohoku Medical Megabank Organization, Tohoku University, Sendai, Miyagi, Japan

Abstract

Comprehensive understanding of gene regulatory networks (GRNs) is a major challenge in the field of systems biology. Currently, there are two main approaches in GRN analysis using time-course observation data, namely an ordinary differential equation (ODE)-based approach and a statistical model-based approach. The ODE-based approach can generate complex dynamics of GRNs according to biologically validated nonlinear models. However, it cannot be applied to ten or more genes to simultaneously estimate system dynamics and regulatory relationships due to the computational difficulties. The statistical model-based approach uses highly abstract models to simply describe biological systems and to infer relationships among several hundreds of genes from the data. However, the high abstraction generates false regulations that are not permitted biologically. Thus, when dealing with several tens of genes of which the relationships are partially known, a method that can infer regulatory relationships based on a model with low abstraction and that can emulate the dynamics of ODE-based models while incorporating prior knowledge is urgently required. To accomplish this, we propose a method for inference of GRNs using a state space representation of a vector auto-regressive (VAR) model with *L*1 regularization. This method can estimate the dynamic behavior of genes based on linear time-series modeling constructed from an ODE-based model and can infer the regulatory structure among several tens of genes maximizing prediction ability for the observational data. Furthermore, the method is capable of incorporating various types of existing biological knowledge, *e.g.*, drug kinetics and literature-recorded pathways. The effectiveness of the proposed method is shown through a comparison of simulation studies with several previous methods. For an application example, we evaluated mRNA expression profiles over time upon corticosteroid stimulation in rats, thus incorporating corticosteroid kinetics/dynamics, literature-recorded pathways and transcription factor (TF) information.

Citation: Hasegawa T, Yamaguchi R, Nagasaki M, Miyano S, Imoto S (2014) Inference of Gene Regulatory Networks Incorporating Multi-Source Biological Knowledge via a State Space Model with *L*1 Regularization. PLoS ONE 9(8): e105942. doi:10.1371/journal.pone.0105942

Editor: Frank Emmert-Streib, Queen's University Belfast, United Kingdom

Received March 27, 2014; **Accepted** July 25, 2014; **Published** August 27, 2014

Copyright: © 2014 Hasegawa et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All synthetic data are within the paper and its Supporting Information files. All microarray files are available from the the NCBI Gene Expression Omnibus database (GSE490).

Funding: This work was supported by Grant-in-Aid for JSPS Fellows (24-9639) received by TH (http://www.jsps.go.jp/english/index.html). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: t-hasegw@kuicr.kyoto-u.ac.jp

Introduction

Transcriptional regulation, which is controlled by several factors, plays essential roles to sustain complex biological systems in cells. Thus, identifying the structure and dynamics of such regulation can facilitate recognition of and control over systems for many practical purposes, *e.g.*, treatment of diseases. To accomplish this, many mathematical methods have been developed for the analysis of high-throughput biological data, *e.g.*, time-course microarray data [1–3]. In addition, recent technological advances have facilitated experimental discoveries, *e.g.*, DNA-protein interactions and the pharmacogenomics of chemical compounds. These contributions have allowed the knowledge of GRNs to accumulate.

For elucidation of GRN dynamics, time-course observational data have been generally used. Currently, one strategy to elucidate transcriptional regulation using observational data is to apply an ordinary differential equation (ODE)-based approach, which can represent the dynamic behavior of biomolecular reactions based on biologically reliable models, *e.g.*, the Michaelis-Menten equation [4] or the S-system [5], which are described by differential equations. Thus, this approach can recapitulate the complex dynamic behavior of biological systems [6,7]. In this approach, several methods have been proposed to infer regulatory structures [8,9], to reproduce the dynamic behavior of biological systems recorded in the literature [10–13] and also to improve literature-recorded pathways so as to be consistent with the data [14]. However, nonlinearity of the system results in an analytically intractable problem of estimating the parameter values that minimize loss function with updating simulated results. Thus, under this statistically efficient paradigm [15], this approach cannot be applied to ten or more genes to infer regulatory structures if the missing information is extensive [10].

In contrast, a statistical model-based approach using highly abstracted models, *e.g.*, Bayesian networks [16–18] and the state space model [19–22], has been successfully applied to infer the

structure of transcriptional regulation from biological observational data. Because these methods simply describe biological systems, hundreds of genes can be handled computationally with ease. Whereas methods relying purely on data need to consider all possibilities of transcriptional regulation, some studies have further incorporated other information, *e.g.*, protein-protein interaction networks (PINs), literature-recorded pathways and transcription factor information [23–27]. Although these methods can infer relationships among hundreds of genes simultaneously, high levels of abstraction can also generate false regulations that are difficult to interpret biologically. Thus, when several tens of genes are handled with partially understood relationships, highly abstract models can be insufficient to represent biological systems. In this case, there is an urgent need for a method that can infer system dynamics and the structure of GRNs based on a model with a low abstraction that can emulate the dynamics of ODE-based gene regulatory models incorporating existing biological knowledge.

We propose a novel method for inference of GRNs based on a newly developed model that uses a state space representation of a vector auto-regressive model (VAR-SSM) [21,28,29]. The model is a type of state space models constructed from a typical gene regulatory system that is described by differential equations within a linear Gaussian model. The method can infer the dynamic behavior of gene expression profiles and the regulatory structure for several tens of genes by assimilating time-course observational data. Furthermore, the method is capable of integrating the existing biological knowledge, *e.g.*, literature-recorded pathways and intracellular kinetics/dynamics of chemical compounds, and can deal with even non-equally spaced time-course observational data. A regulatory structure is inferred by maximization of the $L1$ regularized likelihood. To this end, we developed a new algorithm to obtain active sets of parameters and estimate a maximizer of the $L1$ regularized likelihood using the EM algorithm.

To demonstrate its effectiveness, we compared this method to a state space model (SSM) [21], a general VAR model using LARS-LASSO algorithm [30], GeneNet [31,32] based on an empirical graphical Gaussian model (GGM), dynamic Bayesian networks using first order conditional dependencies [33], GLASSO [34] based on sparse GGM and the mutual information-based network inference algorithms: ARACNE [3], CLR [35] and MRNET [36] by implementing artificial simulation models. The first two observational datasets are generated by two simulation models representing pharmacogenomic pathways [37,38], including drug kinetics/dynamics, described by difference and differential equations, respectively. These pathways are initiated by the drug stimulation and observational data are obtained as non-equally spaced time-course data. The next observational dataset is generated by GeneNetWaver [39,40] using a yeast network that is a part of DREAM4 (Dialogue for Reverse Engineering Assessments and Methods) challenge. As an application example, we applied the proposed method to corticosteroid pharmacogenomics in rat skeletal muscle [37,38,41]. Because this system has been investigated previously through biological experiments, corticosteroid kinetics/dynamics and the related genes are already partly elucidated. Therefore, we incorporated time-course mRNA expression data (observational data), candidate genes/pathways related to corticosteroids, intracellular corticosteroid kinetics/dynamics and, additionally, TF information from ITFP (Integrated Transcription Factor Platform) [42]. As in the simulation experiment, the observational data were obtained as non-equally spaced time-course data (GSE490) after stimulating rat skeletal muscle with corticosteroid. Consequently, we propose candidate pathways for extensions of corticosteroid-related pathways and their simulation dynamics in the presence of corticosteroid.

Methods

Linear Description of Biological Systems from ODE-based Models

For gene regulatory systems, we postulate a general hill function-based model of transcriptional control, in which each gene has a synthesis process (regulated by other factors) and a degradation process, described by a differential equation [43,44]. Let $x_n(t)$ be a time-dependent function representing the abundance of the nth ($n=1,\ldots,N$) mRNA in a cell, where t means time. Further, we consider subsets of $\{1,\ldots,N\}$, $\mathcal{N}_1$ and $\mathcal{N}_2$ ($\mathcal{N}_1 \oplus \mathcal{N}_2 = \{1,\ldots,N\}$), whose regulatory functions are described by two different forms [38,45,46]. Then, the time-evolution of $x_n(t)$ is represented by

$$\frac{d}{dt}x_n(t) = \prod_{k=1}^{N}\{1+\phi_{n,k}(x_k(t))\}\cdot u_n - x_n(t)\cdot d_n, \quad n \in \mathcal{N}_1, \quad (1)$$

$$\frac{d}{dt}x_n(t) = \{1+\sum_{k=1}^{N}\phi_{n,k}(x_k(t))\}\cdot u_n - x_n(t)\cdot d_n, \quad n \in \mathcal{N}_2, \quad (2)$$

where $\phi_{n,k}$ represents the regulatory effect of the kth gene on the nth gene as a hill-function, $u_n > 0$ and $d_n > 0$ are the synthesis and degradation rates of the mRNA, respectively. For example, in a previous pharmacogenomic study [38], $\phi_{n,k}(x_k(t))$ was represented by

$$\phi_{n,k}(x_k(t)) = \frac{\alpha_{n,k}\cdot x_k(t)^{\gamma_{n,k}}}{\beta_{n,k}^{\gamma_{n,k}} + x_k(t)^{\gamma_{n,k}}}, \quad (3)$$

where $\alpha_{n,k}$, $\beta_{n,k}$ and $\gamma_{n,k}$ are tuning parameters.

In inferring the regulatory structure of GRNs consisting of several tens of genes, hill-function based differential equations, *e.g.*, Eqs. (1) and (2), become intractable. Therefore, we consider discretization and linearization of gene regulatory systems [8,19–22,27,29,44]. Here, linear functions are substituted for hill-functions and higher than quadratic terms are neglected. Furthermore, we assume that biological processes should include the effects by noise [47]. Let $\boldsymbol{x}_t = (x_{t,1}, \ldots, x_{t,n}, \ldots, x_{t,N})'$ be a series of N dimensional vectors containing expression levels of N genes at the tth time point. Then, we consider a gene regulatory system represented by

$$x_{t+\Delta t,n} - x_{t,n} = \{(1 + \boldsymbol{a}_n' \boldsymbol{x}_t)u_n - x_{t,n}\cdot d_n + v_{t,n}\}\Delta t, \quad (4)$$

where $a_n = (a_{n,1}, \ldots, a_{n,N})'$ is an N-dimensional vector including regulatory effects on the nth gene by other genes, $v_{t,n}$ is the effects by noise at the tth time point, and Δt indicates a minute displacement. Then, a VAR model for GRNs simulation can be constructed.

In constructing gene regulatory models, we make an assumption that observational data are measured with observational noise. Under this assumption, to separately handle a system model (*i.e.*, Eq. (4)) and biological observational data, we utilize a state space representation [13,21,24,29,48]. Here, a minimum observational time step and Δt are usually handled as 1 for reducing computational cost, however, we can set any value for Δt less than a minimum observational time step. Therefore, we evaluated the influence of changing Δt in the results section and describe the case of $\Delta t = 1$ in the following for simplicity. Consequently, we

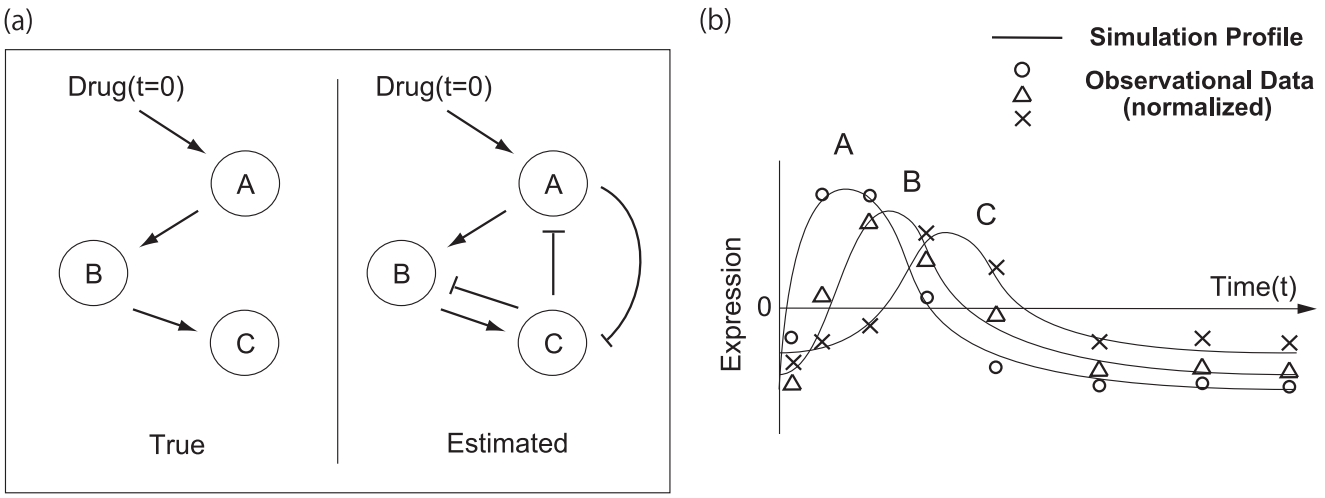

Figure 1. The problem of deleting a term representing a synthesis rate. A toy model indicating the problem of deleting a synthesis rate $\boldsymbol{u}$ by shifting an average of observed time-course data for each element to 0, *i.e.*, $\sum_{t\in\mathcal{T}_{obs}} y_{t,n}=0$ for $n=1,\ldots,N$ as a normalization procedure. The true network and the adjusted data are illustrated in the left panel in (a) and (b), respectively. As shown in the right panel in (a), some false positive edges are possibly estimated in comparison to the true relationships.
doi:10.1371/journal.pone.0105942.g001

consider a model described by

$$\boldsymbol{x}_t = A\,\boldsymbol{x}_{t-1} + \boldsymbol{u} + \boldsymbol{d}_{t-1} + \boldsymbol{v}_t, \tag{5}$$

$$\boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{w}_t, \tag{6}$$

$$\boldsymbol{d}_t = ((1-d_1)\cdot x_{t,1},\ldots,(1-d_N)\cdot x_{t,N})', \tag{7}$$

where $A=(\boldsymbol{a}_1,\ldots,\boldsymbol{a}_N)'$ is an $N\times N$ matrix representing regulation among genes, $\boldsymbol{x}_t$ is an N-dimensional hidden state variable, $\boldsymbol{u}=(u_1,\ldots,u_N)'$ is an N-dimensional vector including synthesis rates, $\boldsymbol{y}_t\in R^N$ is a series of vectors containing observed expression levels of N genes at the tth time point and $\boldsymbol{w}_t\in R^N$ is observational noise. Here, we define a set of all points of time $\mathcal{T}$ ($t\in\mathcal{T}$), consisting of the observed time set $\mathcal{T}_{obs}$ ($\mathcal{T}_{obs}\subset\mathcal{T}$). We set system noise $v_t\sim N_N(\boldsymbol{0}_N,Q)$ and observation noise $\boldsymbol{w}_t\sim N_N(\boldsymbol{0}_N,R)$, where Q and R are $N\times N$ diagonal matrices. The initial state vector $\boldsymbol{x}_0$ is assumed to be a Gaussian random vector with mean vector $\boldsymbol{\mu}_0$ and covariance matrix Σ_0, *i.e.*, $\boldsymbol{x}_0\sim N_N(\boldsymbol{\mu}_0,\Sigma_0)$. Note that $\boldsymbol{u}$ and $\boldsymbol{d}$ must be dense vectors; nevertheless, A should be a sparse matrix, and activation and repression correspond to positive and negative values of $a_{n,k}$, respectively, because hill-functions are monotonic.

Contrary to the derivation of Eq. (5), in previous linear state space models for GRN analysis [21,29], a simulation model was constructed as

$$\boldsymbol{x}_t = F\boldsymbol{x}_{t-1} + \boldsymbol{v}_t, \tag{8}$$

where F is an $N\times N$ matrix in which the nth row and kth column element is represented by

$$f_{n,k}=\begin{cases} 1-d_n+a_{n,k} & (n=k) \\ a_{n,k} & (n\neq k)\end{cases}. \tag{9}$$

In this model, $\boldsymbol{u}$ is removed by shifting the average of the observed time-course data for each element to 0, *i.e.*, $\sum_{t\in\mathcal{T}_{obs}} y_{t,n}=0$ for $n=1,\ldots,N$, where $y_{t,n}$ is the nth row element of $\boldsymbol{y}_t$, as a normalization procedure. However, this model may cause marked difficulty in estimating gene regulatory relationships if the observed time-course includes a steady state. Fig. 1 exemplifies such a situation.

Fig. 1 shows a small pathway consisting of three genes (left panel in Fig. 1 (a)) and the averages of the observed time-course data for each element are shifted to 0 (Fig. 1(b)). By applying Eq. (8) to the observed data, we expect to obtain three false edges added to the true pathway (right panel in Fig. 1(a)) because nodes must retain a constant steady state regardless of their negative steady state values and positive regulation from negative nodes. In some cases, such additional false regulation possibly hide true regulation. The above result encourages us to use a model explicitly implementing terms to represent a steady state of gene expressions to estimate gene regulatory relationships precisely. Furthermore, in using Eq. (8), when elements of F are regularized to be selected non-zero elements, even $1-d_n$ is regularized and $f_{n,n}$ can be zero. To penalize the regulatory effect $a_{n,k}$ only, A and $\boldsymbol{d}$ are separately described in our proposed model.

Incorporation of Biomolecules Affecting Biological Systems

When simulating the dynamic behavior of GRNs including biomolecules that cannot be represented by $\boldsymbol{x}_t$ and can affect biological systems, *e.g.*, corticosteroids in corticosteroid-stimulated GRNs, we should consider the concentration of such biomolecules. For these cases, we remodel Eq. (5) to add a term representing the concentration of such biomolecules as

$$\boldsymbol{x}_t = A\,\boldsymbol{x}_{t-1} + \boldsymbol{u} + \boldsymbol{d}_{t-1} + G\boldsymbol{z}_{t-1} + \boldsymbol{v}_t, \tag{10}$$

where $\boldsymbol{z}_t$ is an M-dimensional vector containing the concentration of the biomolecules at the tth time point, $G=(\mathcal{g}_1,\ldots,\mathcal{g}_N)'$ is an $N\times M$ matrix and $\mathcal{g}_n=(\mathcal{g}_{n,1},\ldots,\mathcal{g}_{n,M})'$ is an M-dimensional vector representing their regulatory effects on the nth gene. We consider the case that the concentration is known or can be

simulated. In the results section, for an application example, we deal with corticosteroid drug pathways that have been well studied previously [37,38,41]; $\boldsymbol{z}_t$ is given the concentration of the intranuclear corticosteroid-receptor complex employed in Yao *et al.* [38].

State Space Model and Kalman Filter for Estimating the Hidden State

Recently, many types of state space models have been proposed and applied in the context of systems biology [13,19–21,23,48,49]. They are roughly divided into two major classes, *i.e.*, linear and nonlinear models. In using linear state space models, posterior probability densities of the hidden state can be obtained as Gaussian distributions and the optimal mean and covariance matrices can be analytically calculated by the Kalman filter algorithm [50,51]. In contrast, for nonlinear state space models, because the analytical form can be intractable, several extensions of the Kalman filter algorithm, *e.g.*, extended Kalman filter [52], unscented Kalman filter [53,54] and particle filter [55], which utilize approximation techniques, have been applied to obtain posterior probability densities of hidden state and parameters [13,24,48,49,56]. In using linear state space models [19–22], the main concern is to infer causal relationships among genes, for which regulatory structure is assumed to be sparse, *i.e.*, genes are regulated by only a few specific regulators. Imposing such a sparse constraint to regression approaches is a general problem, but for state space models to simultaneously estimate optimal hidden state and parameter values (including penalization parameters), it is not a trivial problem [27–30,57]. Then, for example, a sparse regulatory structure was extracted by statistical tests after estimating parameter values [21]. In this article, under the framework of a state space representation of a VAR model, we intend to infer the parameter values and the hidden state maximizing prediction ability for observational data with a sparse regulatory structure. For this purpose, we apply the EM algorithm [58] in the next subsection and the conditional expectations of hidden state are given by using the Kalman filter algorithm.

Kalman Filter Algorithm for VAR-SSM

Let U_t be the sum of $\boldsymbol{u}$ and $G\boldsymbol{z}_t$. For simplicity, we here use F in Eq. (8) rather than A. The prediction, filtering, and smoothing of the Kalman filter are calculated by the following formulas:

- Prediction:

$$\boldsymbol{x}_{t|t-1} = F\boldsymbol{x}_{t-1|t-1} + \boldsymbol{U}_{t-1}, \tag{11}$$

$$\Sigma_{t|t-1} = F\Sigma_{t-1|t-1}F' + Q, \tag{12}$$

- Filtering:

$$\boldsymbol{x}_{t|t} = \boldsymbol{x}_{t|t-1} + \Sigma_{t|t}R^{-1}(\boldsymbol{y}_t - \boldsymbol{x}_{t|t-1}), \tag{13}$$

$$\Sigma_{t|t} = (R^{-1} + \Sigma_{t|t-1}^{-1})^{-1}, \tag{14}$$

- Smoothing

$$\boldsymbol{x}_{t|T} = \boldsymbol{x}_{t|t} + J_t(\boldsymbol{x}_{t+1|T} - \boldsymbol{x}_{t+1|t}), \tag{15}$$

$$\Sigma_{t|T} = \Sigma_{t|t} + J_t(\Sigma_{t+1|T} - \Sigma_{t+1|t})J_t', \tag{16}$$

$$\Sigma_{t,t-1|T} = \Sigma_{t|t}J_{t-1}' + J_t(\Sigma_{t+1,t|T} - F\Sigma_{t|t})J_{t-1}', \tag{17}$$

$$J_t = \Sigma_{t|t}F'\Sigma_{t+1|t}^{-1}, \tag{18}$$

$$\Sigma_{T,T-1|T} = (I - \Sigma_{T|T}R^{-1})F\Sigma_{T-1|T-1}, \tag{19}$$

where $\mathrm{E}[\boldsymbol{x}_t]$ given $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s$ is represented by $\boldsymbol{x}_{t|s}$ and $\mathrm{Var}[\boldsymbol{x}_t]$ given $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s$ is represented by $\boldsymbol{\Sigma}_{t|s}$. To calculate an inverse of the $N \times N$ matrix, we use a matrix inversion theorem [29].

Maximum Likelihood Estimation Using the EM Algorithm with L1 Regularization

In biological systems, most genes are regulated by a few specific genes, *i.e.*, A and G can be sparse matrices. Thus, we applied $L1$ regularization to select effective sets of elements for A and G. Let $\{Y_T, X_T\}$ be the complete data set, where $Y_T = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_T\}$ is the set of observed data and $X_T = \{\boldsymbol{x}_0, \cdots, \boldsymbol{x}_T\}$ is the set of state variables. Furthermore, let the probability densities $P(\boldsymbol{x}_0)$, $P(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ and $P(\boldsymbol{y}_t|\boldsymbol{x}_t)$ be the N-dimensional Gaussian distributions $N(\boldsymbol{\mu}_0, \Sigma_0)$, $N(F_{t-1}\boldsymbol{x}_{t-1} + U_{t-1}, Q)$ and $N(\boldsymbol{x}_t, R)$, respectively. Then joint likelihood for the complete data set is given by

$$P(Y_T, X_T; \boldsymbol{\theta}) = P(\boldsymbol{x}_0)\prod_{t\epsilon T} P(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})\prod_{t\epsilon T_{obs}} P(\boldsymbol{y}_t|\boldsymbol{x}_t), \tag{20}$$

where $\boldsymbol{\theta} = \{A, \boldsymbol{u}, \boldsymbol{d}, G, Q, R, \boldsymbol{\mu}_0\}$. In this study, we used the EM algorithm [58] to search for the parameter vector $\boldsymbol{\theta}$ that maximizes $P(Y_T; \boldsymbol{\theta})$ under $L1$ regularization. The $L1$ regularized log-likelihood is given by

$$\begin{aligned}&\log\int P(\boldsymbol{x}_0)\prod_{t\epsilon T} P(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})\prod_{t\epsilon T_{obs}} P(\boldsymbol{y}_t|\boldsymbol{x}_t)d\boldsymbol{x}_0 \ldots d\boldsymbol{x}_T \\ &- \sum_{n=1}^{N}\sum_{k=1}^{N}\lambda_n|A_{n,k}| - \sum_{n=1}^{N}\sum_{k=1}^{M}\lambda_n|G_{n,k}|,\end{aligned} \tag{21}$$

where λ_n is the $L1$ regularization term for the nth row. In the EM algorithm, the conditional expectation of the joint log-likelihood of the complete data set

$$q(\boldsymbol{\theta}|\boldsymbol{\theta}_i) = E[\log P(Y_T, X_T|\boldsymbol{\theta})|Y_T, \boldsymbol{\theta}_i], \tag{22}$$

is iteratively maximized with respect to $\boldsymbol{\theta}$ until convergence, where $\boldsymbol{\theta}_i$ is the parameter vector obtained at the ith (previous) iteration.

The detailed solution for estimating parameter values using the EM algorithm for VAR-SSM with $L1$ regularization can be found in Method S1.

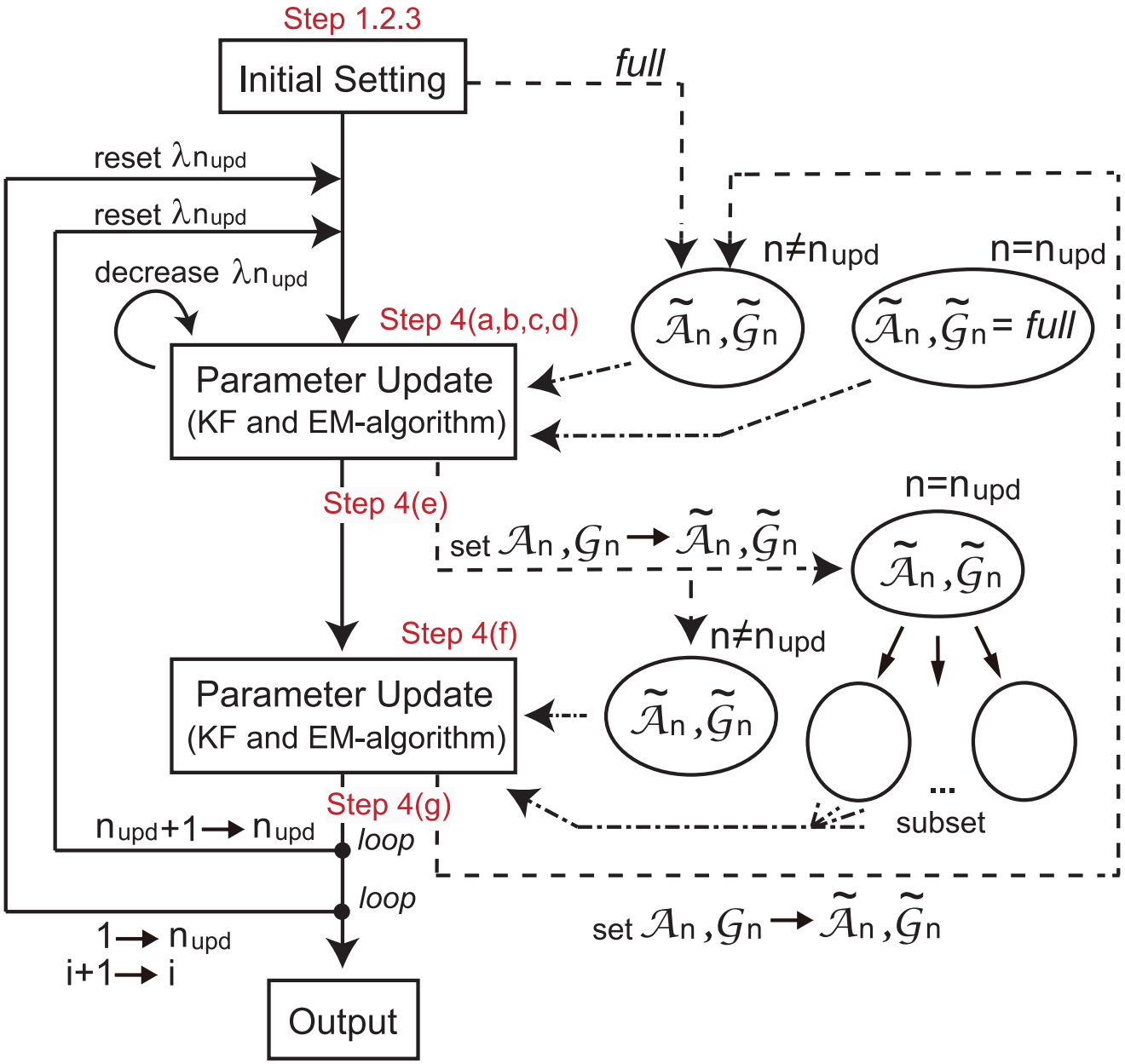

Figure 2. The conceptual view of the proposed algorithm. This figure illustrates a conceptual view of the proposed algorithm. The notations 'Step' correspond to those of the proposed algorithm. Solid, dashed and chain lines represent flowchart of the algorithm, setting the parameter values and active sets, and setting candidates of active sets used for selecting active sets.
doi:10.1371/journal.pone.0105942.g002

Parameter Optimization Algorithm with *L*1 Regularization

Because of the combination of the regularization terms and a state space representation, updating an element of $\boldsymbol{\lambda}=(\lambda_1,\ldots,\lambda_N)'$ influences the other active sets. Thus, it is difficult to select the optimal active sets $\mathcal{A}$ and $\mathcal{G}$, the values of $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ at the same time. Therefore, we proposed a novel algorithm to separately update $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ in each row as follows. In this algorithm, we consider candidates of active sets for $\mathcal{A}_n$ and $\mathcal{G}_n$ as $\tilde{\mathcal{A}}_n$ and $\tilde{\mathcal{G}}_n$, respectively. In the EM algorithm in Method S1, we constraint that the active sets $\mathcal{A}_n$ and $\mathcal{G}_n$ can be selected from $\tilde{\mathcal{A}}_n$ and $\tilde{\mathcal{G}}_n$, respectively, *i.e.*, $\mathcal{A}_n \subseteqq \tilde{\mathcal{A}}_n$ and $\mathcal{G}_n \subseteqq \tilde{\mathcal{G}}_n$

Algorithm

-Initial Settings

1. Set $\boldsymbol{\lambda}=\mathbf{0}$ and recursively update $\boldsymbol{\theta}$ to obtain $\boldsymbol{x}_{t|T}(t=1,\ldots,T)$ using the EM algorithm until convergence is attained. In this step, active sets $\mathcal{A}_n$ and $\mathcal{G}_n$ $(n=1,\ldots,N)$ consist of all elements, *i.e.*, A and G become dense matrices, since the regularization terms can be neglected. Thus, the solution of the EM algorithm is directly obtained from Eqs. (SI–11)–(SI–17) in Method S1.

2. Set the maximum number of iterations to be i_{max}, the maximum number of regulatory edges for each gene to be k_{max}

Table 1. Algorithm 1: A pseudo code of Main Routine (step 4 and 5) in the proposed algorithm.

1: $BIC_{min} \leftarrow +\infty$;
2: **for** $i=1$ to i_{max} **do**
3: **for** $n_{upd}=1$ to N **do**
4: $\tilde{\mathcal{A}}_{n_{upd}} \leftarrow full; \tilde{\mathcal{G}}_{n_{upd}} \leftarrow full; \lambda_{n_{upd}} \leftarrow$ a sufficiently high value;
5: **while** $\lvert\mathcal{A}_{n_{upd}}\rvert+\lvert\mathcal{G}_{n_{upd}}\rvert \leq k_{max}$ **do**
6: **while** convergence criterion is not satisfied **do**
7: Update X_T and parameter values using the Kalman filter and the EM-algorithm;
8: **end while**
9: **if** $BIC_{min} > BIC_{current}$; **then**
10: $BIC_{min} \leftarrow BIC_{current}$; Store the current parameter values;
11: where $BIC_{current}$ is the BIC score of the current parameter values
12: **end if**
13: Decrease $\lambda_{n_{upd}}$;
14: **end while**
15: Set the stored parameter values as the current parameter values;
16: $sub_A \leftarrow$ the set of all subsets of the current $\mathcal{A}_{n_{upd}}$;
17: $sub_G \leftarrow$ the set of all subsets of the current $\mathcal{G}_{n_{upd}}$;
18: **for all** $s_A \in sub_A$ **do**
19: $\tilde{\mathcal{A}}_{n_{upd}} \leftarrow s_A$;
20: **for all** $s_G \in sub_G$ **do**
21: $\tilde{\mathcal{G}}_{n_{upd}} \leftarrow s_G$;
22: **while** convergence criterion is not satisfied **do**
23: Update X_T and parameter values using the Kalman filter and the EM-algorithm;
24: **end while**
25: **if** $BIC_{min} > BIC_{current}$ **then**
26: $BIC_{min} \leftarrow BIC_{current}$; Store the current parameter values;
27: **end if**
28: **end for**
29: **end for**
30: Set the stored parameter values as the current parameter values;
31: **end for**
32: **end for**

doi:10.1371/journal.pone.0105942.t001

and $\boldsymbol{\lambda}$ to be sufficiently high to allow all elements of A and G to become 0, and $\tilde{\mathcal{A}}_n$ and $\tilde{\mathcal{G}}_n$ to be full. Alternatively, i_{max} can be set as a value when the Bayesian information criterion (BIC) [59–61], which are used to select the best model in this algorithms, is not updated through iterations and k_{max} can be set a sufficiently high value, *e.g.*, $\frac{N}{2}$. The BIC score in this algorithm is defined as

$$BIC_{VARSSM} = -2\log L(Y_N|\,\boldsymbol{\theta}) + \mathrm{df}(\boldsymbol{\lambda},\,\boldsymbol{\theta})\log v, \qquad (23)$$

$$L(Y_N|\,\boldsymbol{\theta}) = \int P(\boldsymbol{x}_0)\prod_{t\in T} P(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})\prod_{t\in T_{obs}} P(\boldsymbol{y}_t|\boldsymbol{x}_t)d\boldsymbol{x}_0\ldots d\boldsymbol{x}_T, \qquad (24)$$

where $\mathrm{df}(\boldsymbol{\lambda},\,\boldsymbol{\theta})$ is the degree of freedom, *i.e.*, the number of active parameters [61], and v is the number of samples.

3. Set $i=1$ and recursively update $\{\boldsymbol{\lambda},\mathcal{A},\mathcal{G},\tilde{\mathcal{A}}=\{\tilde{\mathcal{A}}_1,\ldots,\tilde{\mathcal{A}}_N\}, \tilde{\mathcal{G}}=\{\tilde{\mathcal{G}}_1,\ldots,\tilde{\mathcal{G}}_N\},\ \boldsymbol{\theta}\}$ as follows. Note that, at $i=1$, we fix $\boldsymbol{x}_{t|T}$ as the values obtained at Step 1, except for the updating elements indicated as n_{upd} in the next step. Thus, we only update the values of the parameters for the n_{upd}th row at $i=1$.

-Main Routine

4. For $n_{upd}=1,\ldots,N$

(a) Set $\tilde{\mathcal{A}}_{n_{upd}}$ and $\tilde{\mathcal{G}}_{n_{upd}}$ full and $\lambda_{n_{upd}}$ sufficiently high to allow all elements of $\boldsymbol{a}_{n_{upd}}$ and $\boldsymbol{g}_{n_{upd}}$ become $\mathbf{0}$. Through the following steps, fixing λ_n $(n\neq n_{upd})$, $\lambda_{n_{upd}}$ is gradually

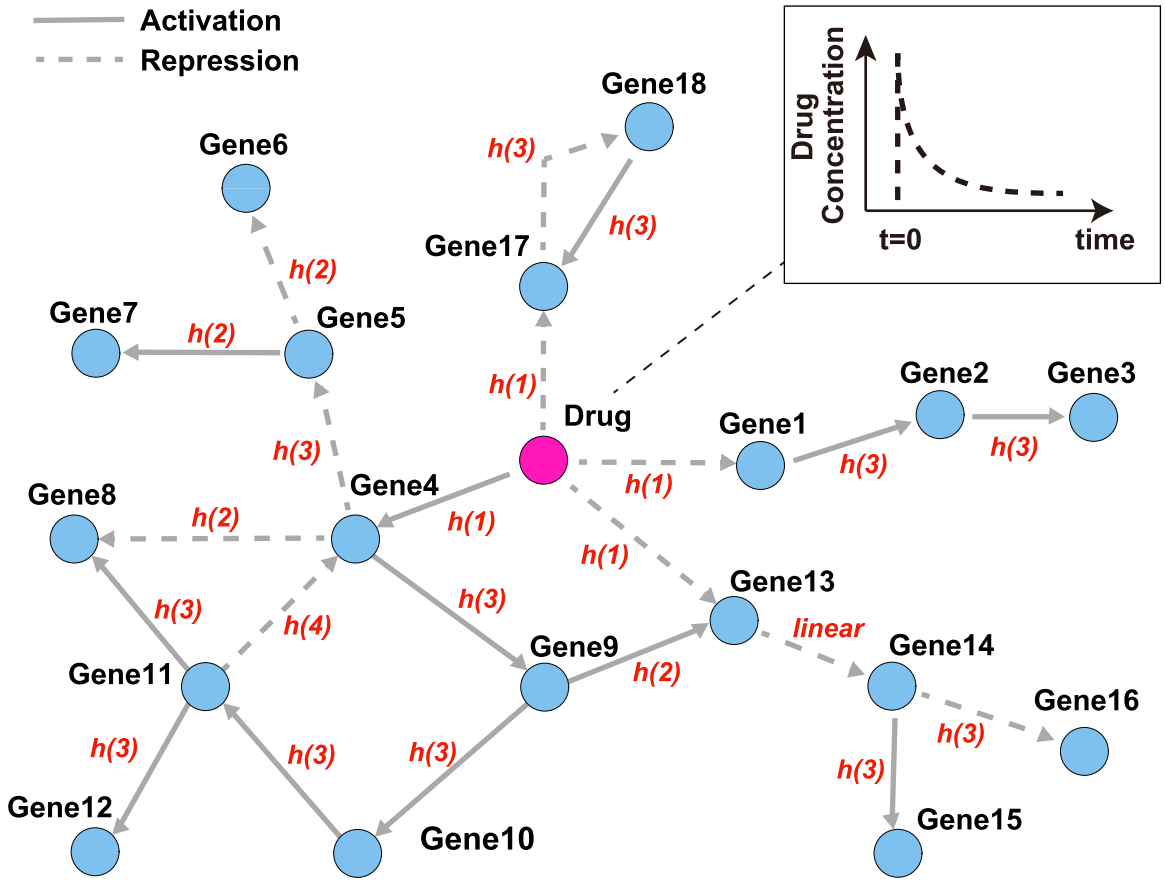

Figure 3. A pharmacogenomic pathway of the artificial simulation model. The figure illustrates the pathway of the artificial simulation model used for datasets (i) and (ii). Each regulation is represented by (i) a linear and (ii) a nonlinear function, such as Eq. (3). For dataset (ii), descriptions on edges as *linear* or $h(c=1,2,3 \text{ or } 4)$ means a linear function and a hill function, described in Eq. (3) when $\gamma_{n,k}=c$, respectively. The system is stable at first ($t<0$) and undergoes stimulation by a drug at $t=0$. The concentration of the drug is gradually decreased according to the drug kinetics. A solid arrow and a dotted arrow mean activation and repression, respectively.
doi:10.1371/journal.pone.0105942.g003

decreased to find an optimum $\lambda_{n_{upd}}$ for which the BIC score is minimized.

(b) Calculate conditional expectations using the Kalman filter.

(c) Update $\mathcal{A}$, $\mathcal{G}$, and $\boldsymbol{\theta}$ by Eqs. (SI–11)–(SI–19) in Method S1. Here, $\mathcal{A}_n$ and $\mathcal{G}_n$ of Eqs. (SI–18)–(SI–19) in Method S1 can be constructed from $\tilde{\mathcal{A}}_n$ and $\tilde{\mathcal{G}}_n$, respectively.

(d) Calculate the BIC score and decrease $\lambda_{n_{upd}}$ if the regularized log-likelihood of Eq. (21) is converged. Then, repeat from step (b) until the sum total of $\mathcal{A}_{n_{upd}}$ and $\mathcal{G}_{n_{upd}}$ becomes k_{max}.

(e) Set $\{\boldsymbol{\lambda},\mathcal{A},\mathcal{G},\boldsymbol{\theta}\}$ as the value with the lowest BIC score obtained through the above described steps. Furthermore, set $\tilde{\mathcal{A}}\leftarrow\mathcal{A}$ and $\tilde{\mathcal{G}}\leftarrow\mathcal{G}$.

(f) Consider the set of all subsets of $\mathcal{A}_{n_{upd}}$ and $\mathcal{G}_{n_{upd}}$ as sub_A and sub_G, respectively. For all $s_A \in sub_A$ and $s_G \in sub_G$, setting $\tilde{\mathcal{A}}_{n_{upd}}\leftarrow s_A$ and $\tilde{\mathcal{G}}_{n_{upd}}\leftarrow s_G$, repeat steps 4(b) and (c), and then obtain the BIC scores of converged log-likelihood.

(g) Set $\{\mathcal{A},\mathcal{G},\boldsymbol{\theta}\}$ as the value with the lowest BIC score. Furthermore, $\tilde{\mathcal{A}}\leftarrow\mathcal{A}$ and $\tilde{\mathcal{G}}\leftarrow\mathcal{G}$.

5. Set $i\rightarrow i+1$ and repeat from step 4 until i becomes i_{max}.

A conceptual view and a pseudo code of the algorithm are shown in Figure 2 and Algorithm 1 in Table 1, respectively. We should note that, since the active sets $\mathcal{A}$ and $\mathcal{G}$ obtained at step 4(e) may not be the optimal ones for the selected $\boldsymbol{\lambda}$, *i.e.*, there can exist better ones having lower BIC scores for the selected $\boldsymbol{\lambda}$, the proposed algorithm further explores such better ones by evaluating subsets of the obtained active sets at step 4(e) through steps 4(f) and (g).

Weighting Known Regulations

To weight parameters of known regulations, *e.g.*, as recorded in the literature, we derive the weighted regularization [62]. For the nth row, we define the weight vectors $\boldsymbol{\omega}_n^a=(\omega_{n,1}^a,\ \ldots,\ \omega_{n,N}^a)'$ and $\boldsymbol{\omega}_n^g=(\omega_{n,1}^g,\ \ldots,\ \omega_{n,M}^g)'$. The elements of these vectors for known regulations are set to less than 1 or, otherwise set to 1. Then, in the

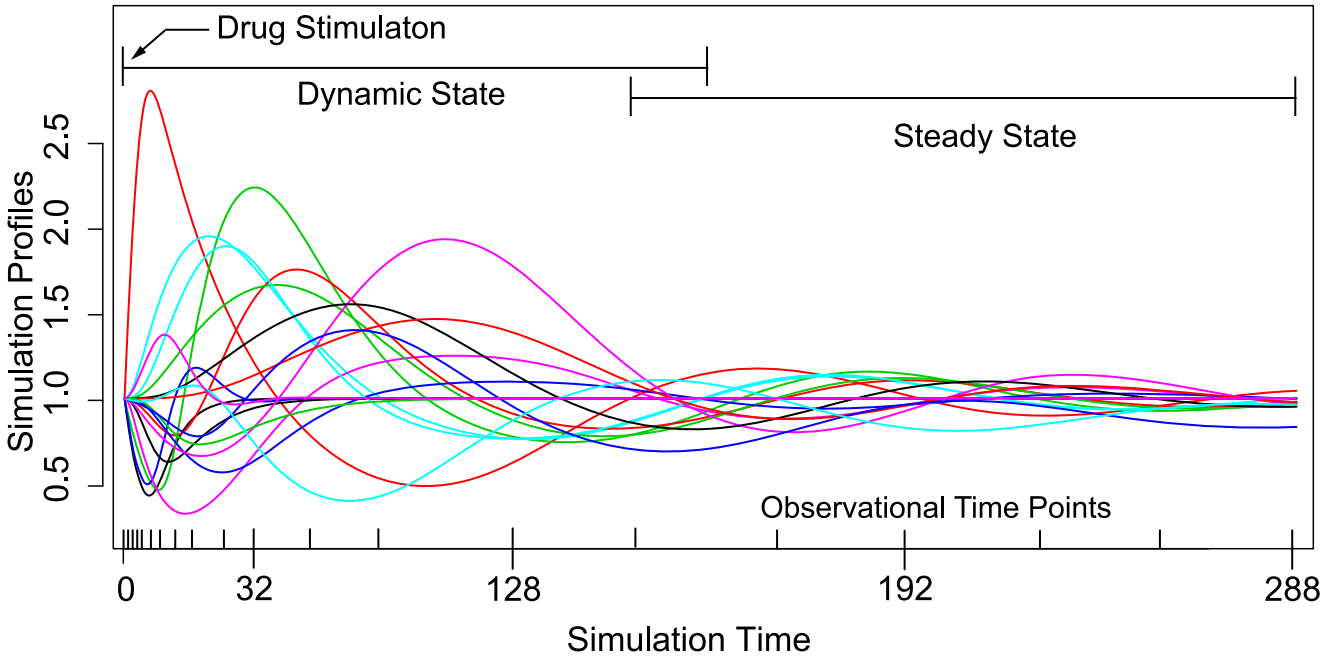

Figure 4. The simulation expression profiles of genes of the artificial simulation model. This illustrates the simulation expression profiles of genes of the artificial simulation model used for dataset (ii). The simulated data for datasets (i) and (ii) have both dynamic and steady state, and stimulated by the drug at $t=0$. Observational time-course data is obtained with Gaussian noise from the simulation expression at the time points that are indicated on the bottom axis. The observational data, parameter values and simulation models are available at Models S1 and S2.
doi:10.1371/journal.pone.0105942.g004

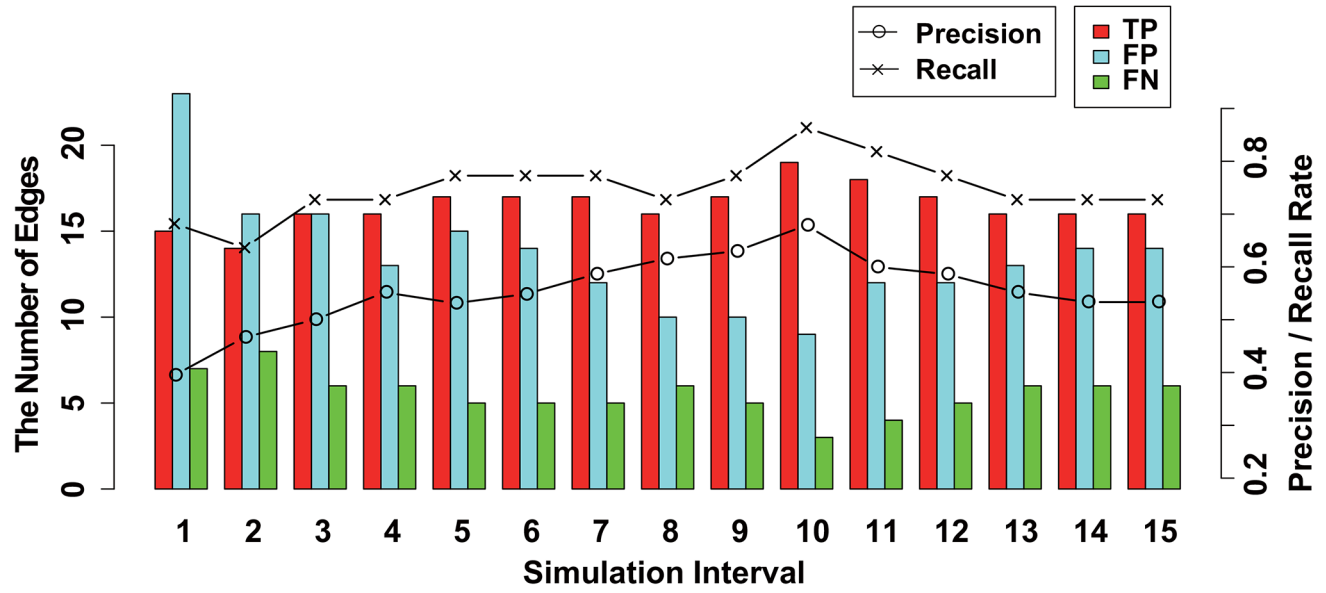

Figure 5. The results of the structure inference using dataset (i) of a pharmacogenomic pathway by the proposed method. This figure illustrates the results of the structure inference after applying the proposed method to dataset (i) for each simulation time interval Δt. The histogram represents the number of true positive (TP), false positive (FP), and false negative (FN) findings for each $\frac{1}{\Delta t}=(1,2,\cdots,15)$ as red, blue, and green bars, respectively. Black lines with circles and crosses represent 'precision rate (PR $=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$)' and 'recall rate (RR $=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$)', respectively. The values of the histogram and lines correspond to the left and right axes, respectively.
doi:10.1371/journal.pone.0105942.g005

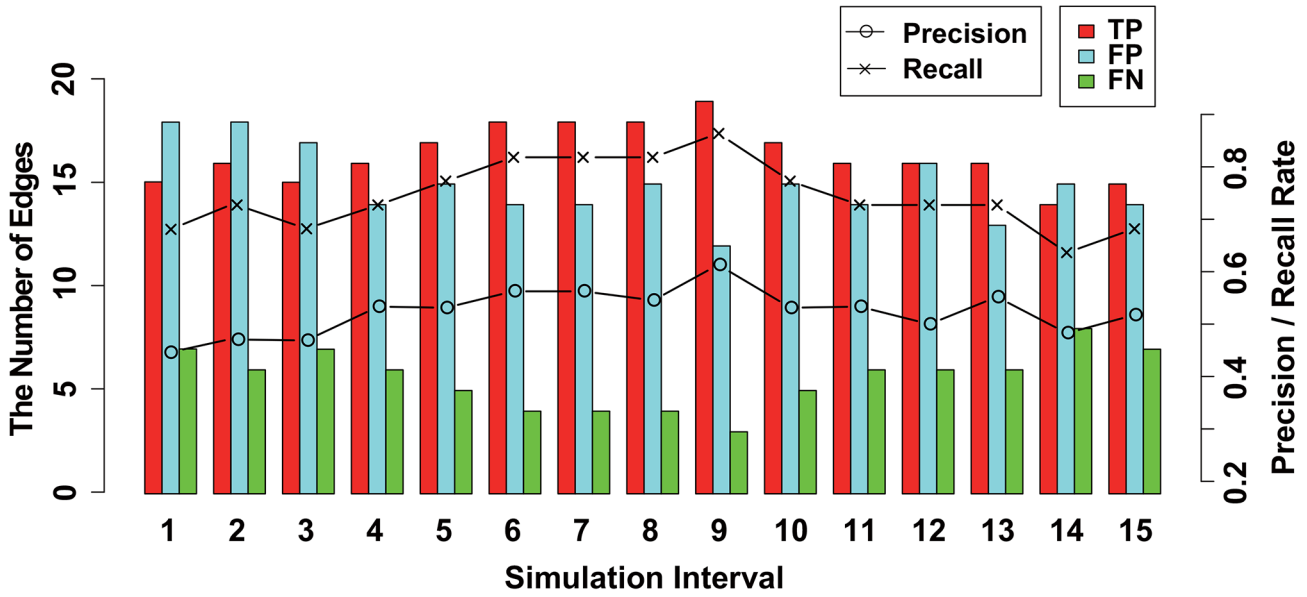

Figure 6. The results of the structure inference using dataset (ii) of a pharmacogenomic pathway by the proposed method. This figure illustrates the results of the structure inference after applying the proposed method to dataset (ii) for each simulation time interval Δt. The histogram represents the number of true positive (TP), false positive (FP), and false negative (FN) findings for each $\frac{1}{\Delta t}=(1,2,\cdots,15)$ as red, blue, and green bars, respectively. Black lines with circles and crosses represent 'precision rate ($PR=\frac{TP}{TP+FP}$)' and 'recall rate ($RR=\frac{TP}{TP+FN}$)', respectively. The values of the histogram and lines correspond to the left and right axes, respectively.
doi:10.1371/journal.pone.0105942.g006

M step of the EM algorithm and the regularized log-likelihood, regularization terms are handled as

$$\sum_{k=1}^{N}\lambda_n|a_{n,k}|\rightarrow\sum_{k=1}^{N}\omega_{n,k}^{a}\lambda_n|a_{n,k}|, \qquad (25)$$

$$\sum_{k=1}^{M}\lambda_n|g_{n,k}|\rightarrow\sum_{k=1}^{M}\omega_{n,k}^{g}\lambda_n|g_{n,k}|. \qquad (26)$$

In practice, the purpose of the weight is to select known regulation in the instance where multiple candidates are highly

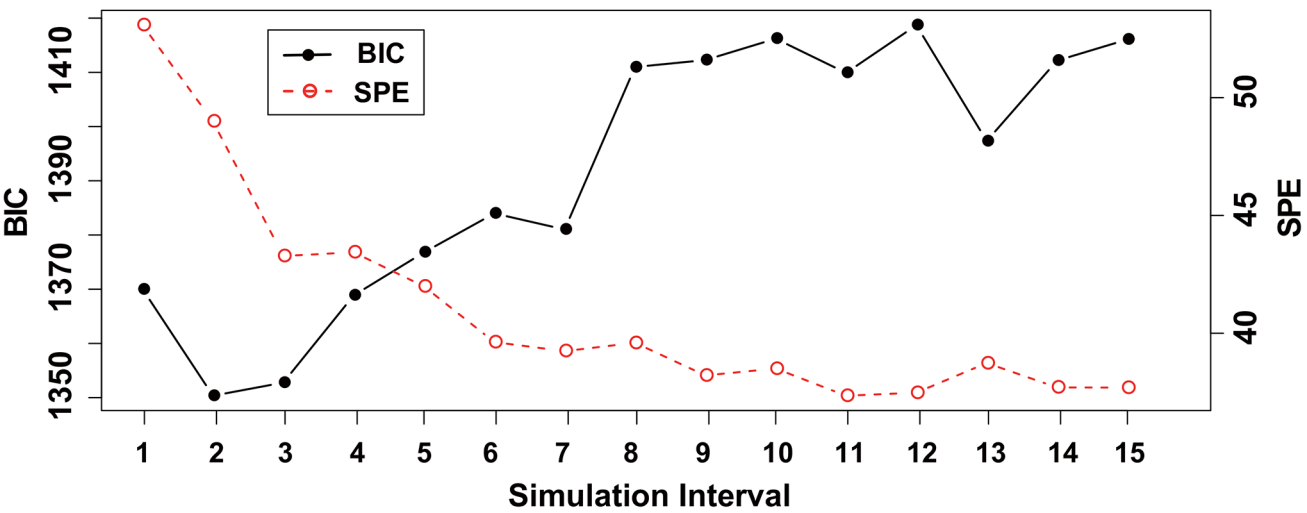

Figure 7. The result of the BIC scores and SPE for each simulation time interval using dataset (i). This illustrates the BIC scores and SPE ($t=6,8,12$) for $\frac{1}{\Delta t}=(1,2,\cdots,15)$ for dataset (i). The values of the BIC scores and SPE correspond to the left and right axes, respectively.
doi: 10.1371/journal.pone.0105942.g007

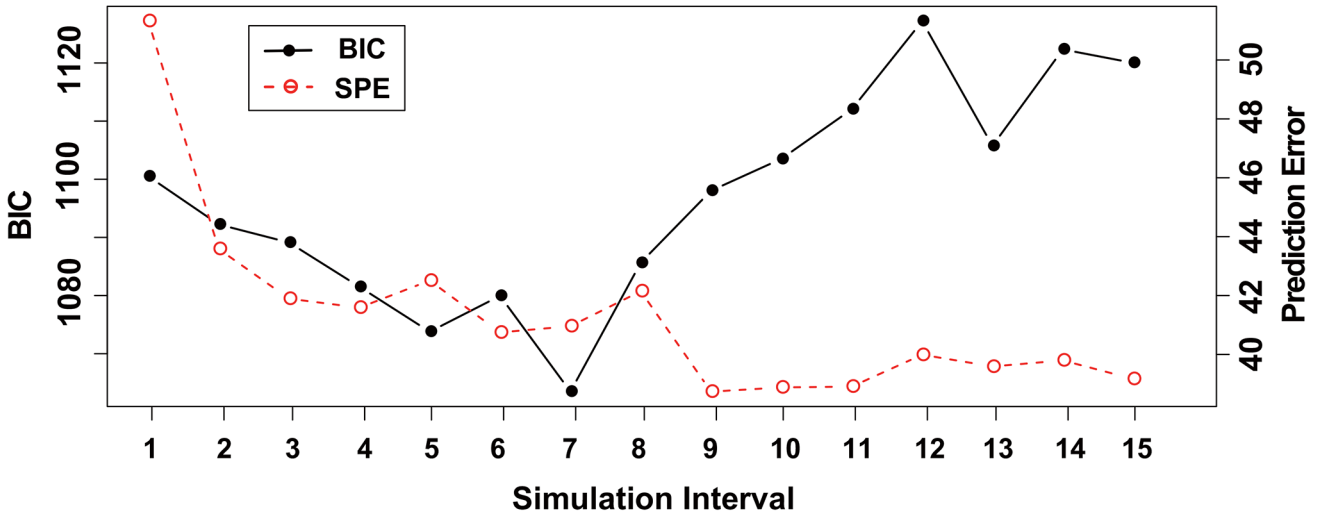

Figure 8. The result of the BIC scores and SPE for each simulation time interval using dataset (ii). This illustrates the BIC scores and SPE ($t=6,8,12$) for $\frac{1}{\Delta t}=(1,2,\cdots,15)$ for dataset (ii). The values of the BIC scores and SPE correspond to the left and right axes, respectively.
doi: 10.1371/journal.pone.0105942.g008

correlated with the same gene. Thus, when the correlation of a known regulation is still a low value, the regulation should not be selected as an active regulation. For example, weights for literature-recorded pathways and regulations by TFs are set as $\frac{1}{20}$ and $\frac{1}{10}$ in the real data experiment, respectively. The effectiveness of the weighted regularization is demonstrated in the results section.

Results

Comparison Results

To show the effectiveness of the proposed method, we compared it with other GRN inference methods, *i.e.*, a state space model (SSM) [21,63], a general VAR model using the LARS-LASSO algorithm [30,61], GeneNet [31,32] based on an empirical graphical Gaussian model (GGM), dynamic Bayesian networks using first order conditional dependencies (G1DBN) [33], GLASSO [34] based on sparse GGM and the mutual information-based network inference algorithms: ARACNE [3], CLR [35] and MRNET [36]. We applied these inference methods by using R-package ('GeneNet', 'G1DBN', 'glasso' and 'parmigene') and implementing the others. The comparison analysis was performed using three artificial data, which were generated based on pharmacogenomic pathways that we assumed and a yeast network that was produced as a part of the DREAM4 (Dialogue for Reverse Engineering Assessments and Methods) challenge. We should note that, because ARACNE, CLR and MRNET are intended to infer static relationships between genes, we considered time-course observational data as static data utilizing a time-lag matrix, in which the tth row vector consists of $\boldsymbol{y}_{t+1}-\boldsymbol{y}_t$, according to Shimamura *et al.* [57]. Note that the Jar file of the proposed method is available at: http://sunflower.kuicr.kyoto-u.ac.jp.

Comparison Using Pharmacogenomic Pathways

For the comparison, we first generated two time-courses from (i) linear difference equations as Eq. (4) and (ii) nonlinear differential equations as Eqs. (1)–(3) representing pharmacogenomic pathways (*e.g.*, Yao *et al.* [38]) using Cell Illustrator 5.0 (http://www.cellillustrator.com/home). The details of the artificial simulation models are as follows.

Table 2. Comparison of the proposed method and the existing methods using dataset (i).

	PR	RR	TP	FP	TN	FN
(a)VARSSM(BIC)	0.467	0.634	14	16	286	8
(b)VARSSM(SPE)	0.600	0.818	18	12	290	4
(c)SSM	0.308	0.182	4	9	293	18
(d)VAR	0.150	0.773	17	97	205	5
(e)Genenet	0.280	0.667	14	36	114	7
(f)G1DBN	0.314	0.500	11	24	278	11
(g)GLASSO	0.094	0.286	6	58	92	15
(h)ARACNE	0.131	0.524	11	71	79	10
(i)CLR	0.135	0.619	13	83	67	8
(j)MRNET	0.121	0.571	12	87	63	9

doi:10.1371/journal.pone.0105942.t002

Table 3. Comparison of the proposed method and the existing methods using dataset (ii).

	PR	RR	TP	FP	TN	FN
(a)VARSSM(BIC)	0.563	0.818	18	14	288	4
(b)VARSSM(SPE)	0.613	0.864	19	12	290	3
(c)SSM	0.234	0.318	7	23	279	15
(d)VAR	0.206	1.000	22	84	236	0
(e)Genenet	0.278	0.714	15	39	111	6
(f)G1DBN	0.647	0.500	11	6	296	11
(g)GLASSO	0.052	0.143	3	55	95	18
(h)ARACNE	0.191	0.429	9	38	112	12
(i)CLR	0.156	0.667	14	76	74	7
(j)MRNET	0.156	0.667	14	76	74	7

doi:10.1371/journal.pone.0105942.t003

-*Dataset*(*i*)

1. The number of genes is 18.
2. Each gene undergoes synthesis and degradation processes, and genes are mutually regulated as shown in Fig 3 (The details of the figure are explained below).
3. A drug is added at $t=0$ and its concentration gradually decreases according to one compartment model, *i.e.*, $\frac{d}{dt}z(t)=\zeta z(t)$, where $z(t)$ is the concentration of the drug as a function of time t and ζ is the degradation rate. The simulated expression profiles of the genes are initiated by the drug at $t=0$ and gradually converge to their steady states as illustrated in Fig. 4.
4. The expression data is observed at $\mathcal{T}_{obs}=(0, 1, 2, 3, 4, 6, 8, 12, 16, 24, 32, 48, 52, 96, 128, 160, 192, 224, 256 and 288)$ with Gaussian observation noise of mean 0 and a variance that is proportional to the intensity.
5. The number of replicated observations with different observational noise for each time point is three.
6. The simulated expression is updated according to the linear difference equations represented by Eq. (4) at $\Delta t=\frac{1}{5}$.

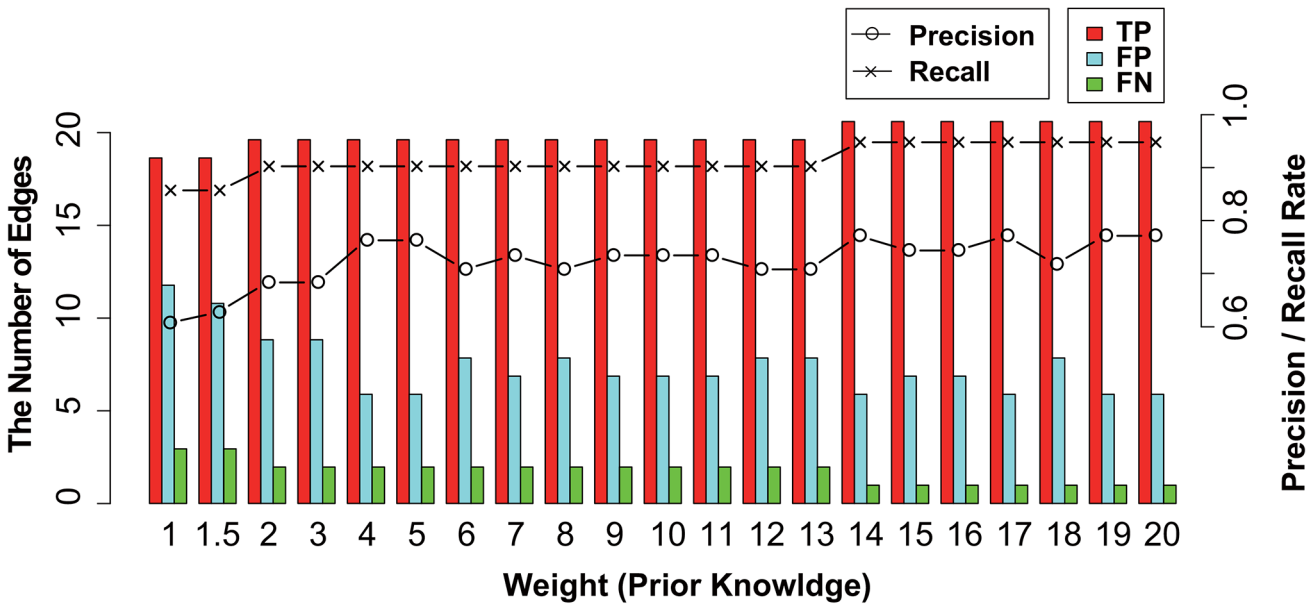

Figure 9. The performance of using prior knowledge as the weighted regularization. This figure illustrates the effectiveness of the weighted regularization (prior knowledge) at simulation time interval $\frac{1}{\Delta t}=9$ using dataset (ii). The histogram represents the number of true positive (TP), false positive (FP), and false negative (FN) findings for each $\frac{1}{w_{n,k}}=(1,1.5,2,3,, \cdots ,20)$ as red, blue, and green bars, respectively. Black lines with circles and crosses represent 'precision rate (PR$=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$)' and 'recall rate (RR$=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$)', respectively. The values of the histogram and lines correspond to the left and right axes, respectively.
doi:10.1371/journal.pone.0105942.g009

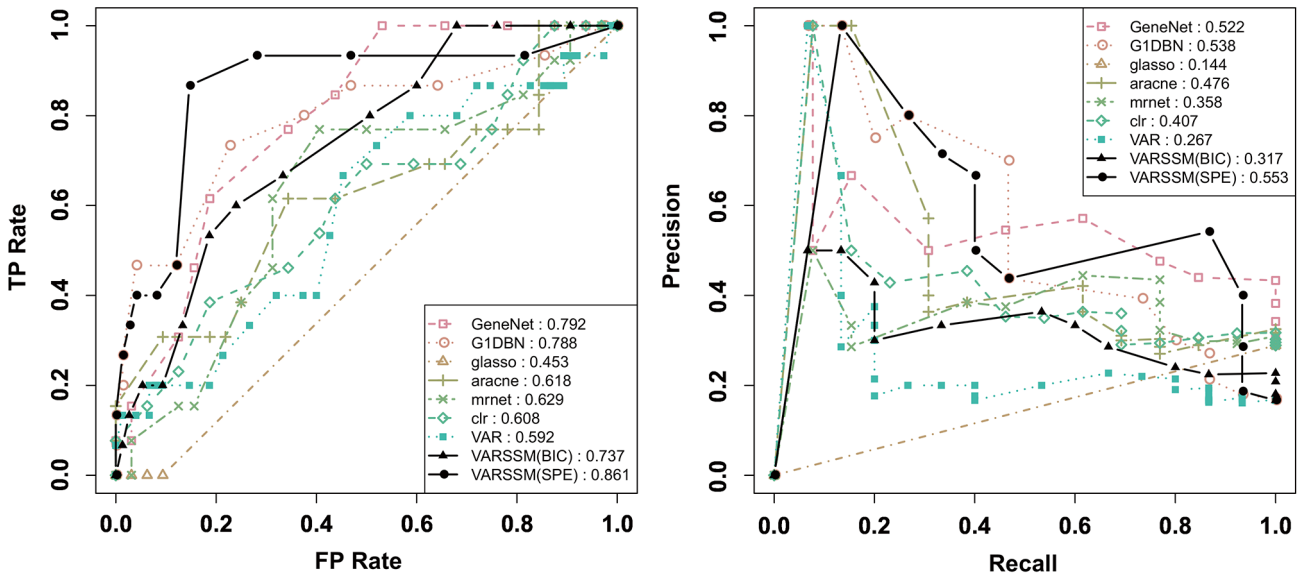

Figure 10. The ROC and PR curves using dataset (iii). The left and the right figures illustrate the ROC and PR curves for dataset (iii), respectively. In the left figure, the vertical axis and horizontal axis correspond to TP rate and FP rate, respectively. In the right figure, the vertical axis and horizontal axis correspond to PR and RR, respectively. AUROC and AUPR are represented at the right side of the inference methods.
doi:10.1371/journal.pone.0105942.g010

7. The observational data and the values of the parameters are available at Model S1.

-Dataset(ii)

1 to 5 of dataset (i) are also satisfied in dataset (ii).

6. The simulated expression is updated according to the differential equations. Regulatory relationships are the same as in (i) but the regulatory effects are represented by hill functions, such as Eqs. (1)–(3), or linear functions, as illustrated in Fig 3. In this figure, $h(c)$ indicates that the regulation is described by Eq. (3) when $\gamma_{n,k}=c$.
7. The observational data and the csml (cell system markup language) file are available at Model S2.

A true positive (TP), false positive (FP), false negative (FN), precision rate ($\mathrm{PR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$), and recall rate ($\mathrm{RR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$) were used to measure the performance. At first, in applying the proposed method to the data, we changed the simulation time interval of Eq. (4) to $\frac{1}{\Delta t}=(1,2,\ldots,15)$, and estimated active sets of regulation ($\mathcal{A}$ and $\mathcal{G}$) and the values of the parameters for each $\frac{1}{\Delta t}$ for each dataset. The results for datasets (i) and (ii) are illustrated in Figs. 5 and 6, respectively. The precision and recall rates in Figs. 5 and 6 show that the performance of the structure inference gradually increases from $\frac{1}{\Delta t}=1$ and is optimal at $\frac{1}{\Delta t}=10$ for (i) and $\frac{1}{\Delta t}=9$ for (ii). This indicates that the simulation time interval Δt can influence the performance of structure inference and we should carefully design Δt for biological simulations. In order to determine Δt, we measured the BIC scores and the sum of squared prediction errors (SPE) at three time points ($t=6$, 8 and 12) for each $\frac{1}{\Delta t}$ using (i) and (ii), as represented in Fig. 7 and Fig. 8, respectively. Here, we measured the prediction errors for each time point by optimizing the values of the estimated parameters without using the observational data at the corresponding time point ($t=6$, 8, 12).

For dataset (i), although the PR and RR values peak at $\frac{1}{\Delta t}=10$, the BIC scores become lowest at $\frac{1}{\Delta t}=2$. Similarly, the BIC score becomes lowest at $\frac{1}{\Delta t}=7$ but peaks at $\frac{1}{\Delta t}=9$ for dataset (ii). SPE gradually converges when $\frac{1}{\Delta t}$ becomes large and has the lowest value at $\frac{1}{\Delta t}=11$ and $\frac{1}{\Delta t}=9$ for datasets (i) and (ii), respectively. Therefore, SPE can be an indicator for determining the best time interval for this hill function-based system of pharmacogenomics. Note that the measured time points for the prediction errors should be the points that are not steady state values.

Next, we compared the results of (a) the proposed VAR-SSM with the lowest BIC and (b) the proposed VAR-SSM with the lowest SPE to (c) SSM [21,63] (permutation tests were utilized to select regulations), (d) VAR model with $L1$ regularization using the LARS-LASSO algorithm [30,61] (the BIC score is used to determine the value of the regularization parameters), GeneNet [31,32], G1DBN [33], GLASSO [34], ARACNE [3], CLR [35] and MRNET [36]. The comparison results for datasets (i) and (ii) are listed in Tables 2 and 3, respectively. In these comparisons, we added the drug profiles to the observational data and did not count regulations in response to drugs and self-regulation. For the methods inferring undirected regulations, *i.e.*, GeneNet, GLASSO, ARACNE, CLR and MRNET, we considered the true network (directed network) as an undirected network and then measured the performance by comparing this undirected network to the inferred networks. Additionally, for GeneNet, G1DBN and mutual information-based methods (ARACNE, CLR and MRNET), which are required to set a threshold value to determine the existence of regulation, we checked the results of setting the threshold q-value (GeneNet) and posterior probability (G1DBN) to (0.01,0.05,0.1,0.2,..., and 0.5) and a cut-off value

Table 4. The number of selected simulation time intervals for dataset (iii).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
BIC	99	0	0	0	0	1	0	0	0	0	0	0	0	0	0
SPE	3	13	9	1	4	0	2	6	5	6	3	11	6	12	19

doi:10.1371/journal.pone.0105942.t004

(ARACNE, CLR and MRNET) to (0,0.01,0.05,0.1,0.2, . . . , and 0.8), and adopted the best thresholds with respect to *F-measure*$=\frac{2\cdot PR\cdot RR}{PR+RR}$. We should note that the simulation time interval of SSM is set $\Delta t=1$ (no other choice is available) due to the implementation of Tamada *et al.* [63]. It is hard to make the simulation time interval short; hence, the simulated expression profiles often oscillated in such situations.

Consequently, the proposed method achieved a low false positive rate while maintaining a high true positive rate. These results may be acceptable because the system model of the proposed method is the same as or similar to the artificial simulation models. Thus, it is conceivable that the proposed method is highly capable of inferring the regulatory structure of the assumed hill-function based model. Furthermore, we demonstrated the effectiveness of the weighted regularization for known prior information using dataset (ii). To evaluate the performance, we adapted a simulation time interval of $\frac{1}{\Delta t}=9$. Setting weights for true regulations as $\frac{1}{w_{n,k}}=(1.5,2,3,\cdots,20)$, PR and RR were evaluated as illustrated in Fig. 9. The correct weights reduced the FP and FN edges, and the performance was gradually improved according to the increase in the weight coefficient. In contrast, several FP edges still exist even when the weight coefficients take on high values. It can be considered that the simplification of the true regulatory system using the proposed model generates these false edges to effectively predict the data.

Comparison Using Yeast Network of a Part of the DREAM4 Challenge

In contrast to the previous comparisons, for which the data were based on the assumed models as Eqs. (1)–(4), we next prepared data generated by GeneNetWaver [39,40] using a 10-node yeast network (*yeast 1*) of a part of the DREAM4 challenge (in silico network challenge). To measure the performance of the proposed method, in this comparison, we generated dataset (iii), which was a set of 100 time-course observational data, in which the measured time points were $t=(0,1,\ldots,30)$.

According to the original setting, three genes, which were randomly selected for each time-course, were perturbed among $t=0$ to 15. Here, since we intended to consider the case that observational data have a steady state, the number of time points was to be set larger than those of the original setting $t=(0,1,\ldots,20)$. The dataset (iii) is available at Model S3.

We applied the methods (a)–(j) to dataset (iii); however, since SSM [21,63] requires large computational costs to perform permutation tests for each time-course, we neglected SSM for this comparison. The time points to calculate SPE for the proposed method are $t=(16,17,18)$, which are the time points shortly after removal of perturbations. For each method, we summed the existence of the estimated regulation on the ith gene by jth gene as $est_{i,j}$ and considered the values $\frac{est_{i,j}}{100}$ as the confidence level for the regulation. Then, TP rate (TPR$=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$), FP rate (FPR$=\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$), precision rate (PR$=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$) and recall rate (RR$=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$) were calculated to draw ROC and PR curves. Using these curves, we measured the performance with respect to the AUROC (area under the ROC curve) and AUPR (area under the PR curve). These comparison results are illustrated in Fig. 10. Note that, similarly to

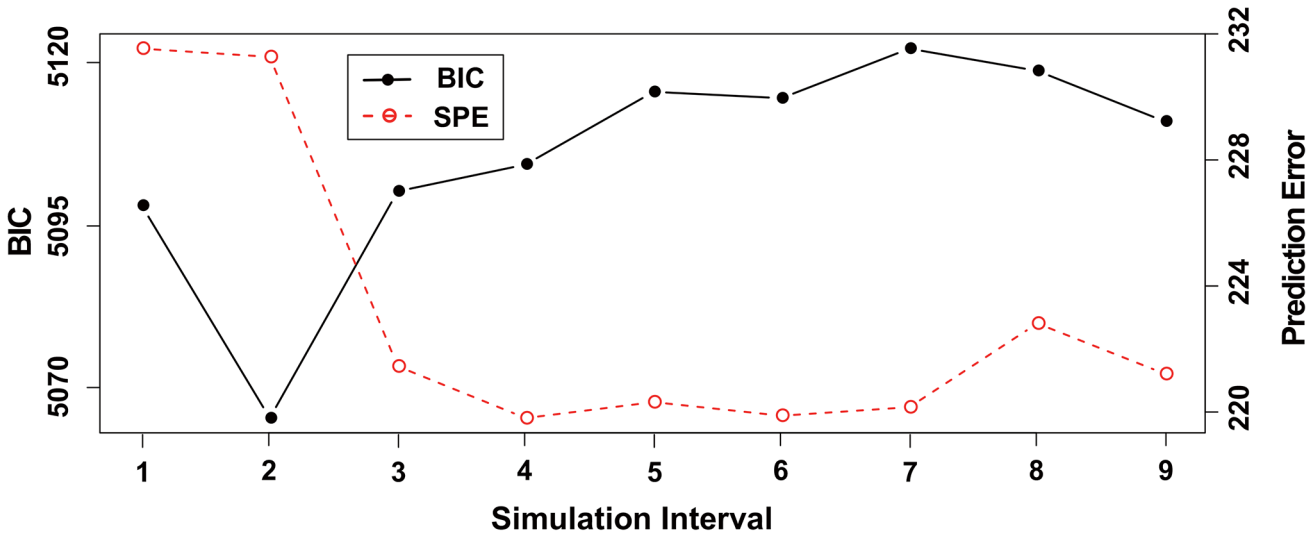

Figure 11. The result of the BIC scores and SPE for each simulation time interval using the real data. This illustrates the BIC scores and SPE ($t=1,2,4$) for each time interval $\frac{1}{\Delta t}=\{1,2,\ldots,9\}$. The values of these indicators corresponds to the left and right axes, respectively.
doi: 10.1371/journal.pone.0105942.g011

the previous experiments, we selected the best threshold values with respect to AUROC for the methods (e), (f) and (h)–(j).

As a result, although the simulation model for dataset (iii) is different from the models that we assumed, the proposed method using SPE outperformed the other methods in terms of both AUROC and AUPR. The number of selected simulation time intervals Δt is shown in Table 4. These results indicate that the proposed method has good ability for inferring the regulatory relationships using time-course observational data for which regulations are not based on the model that we assumed. Furthermore, we can consider the SPE as a good indicator for determining the simulation time interval.

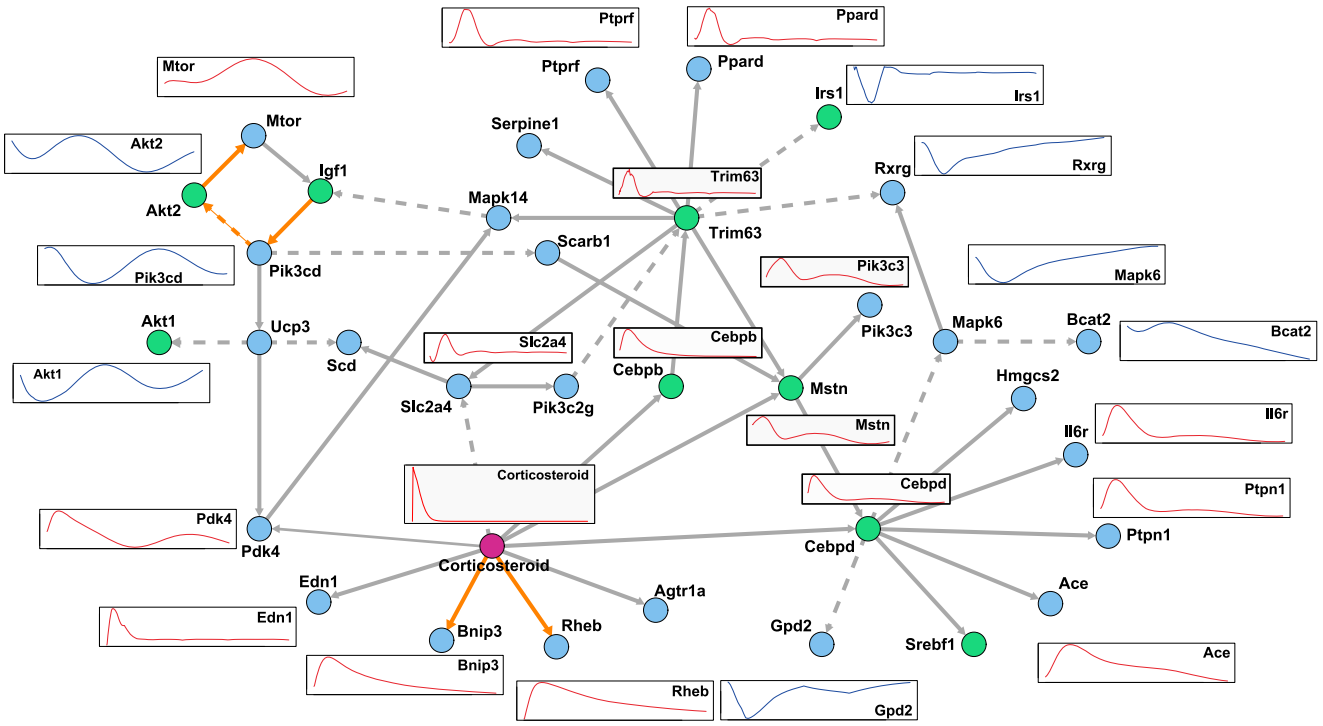

Figure 12. The estimated network with weighting literature-recorded pathways. This figure illustrates the inferred gene regulatory network with weights for literature-recorded pathways. *Corticosteroid* and genes of TFs are drawn as a red circle and green circles, respectively. Estimated edges with weights are illustrated as orange. Further, on some genes, simulation expression profiles are attached as examples. Red and blue profiles are roughly distinguished to up-regulated and down-regulated genes, respectively.
doi:10.1371/journal.pone.0105942.g012

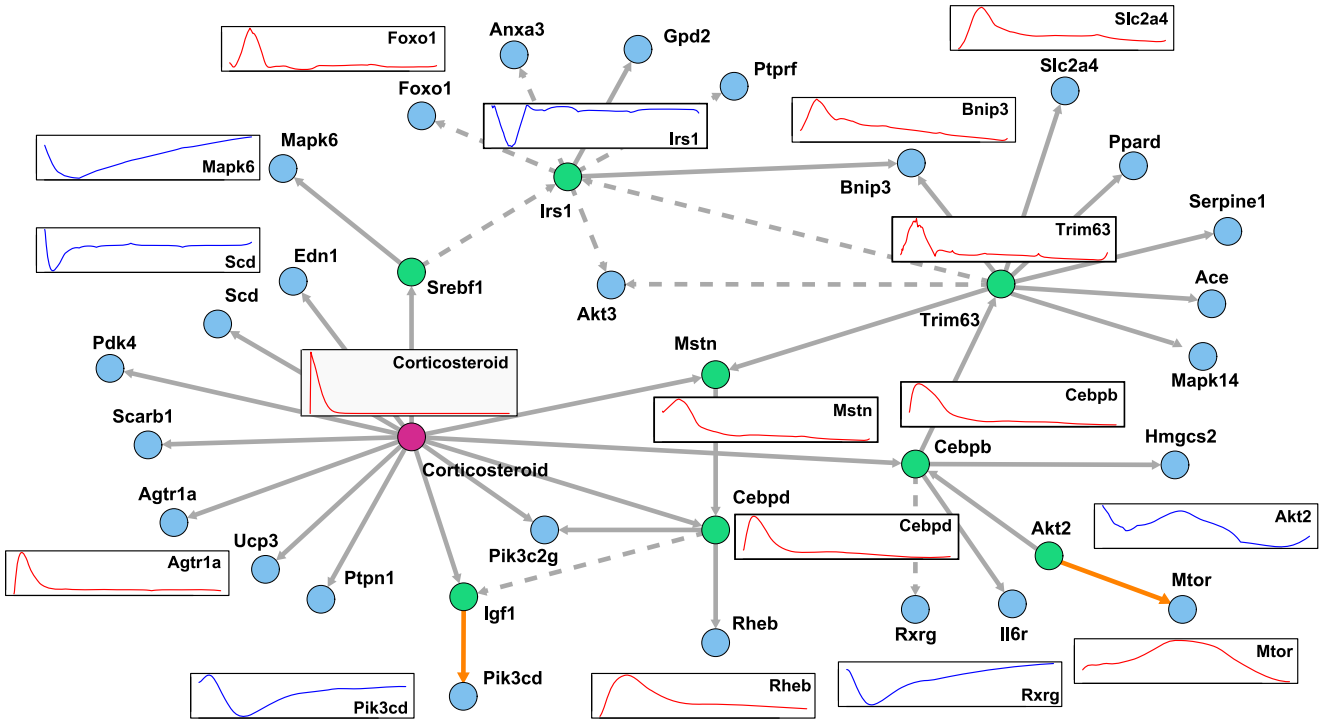

Figure 13. The estimated network with weighting literature-recorded pathways and regulations by TFs. This figure illustrates the inferred gene regulatory network with weights for literature-recorded pathways and regulations by TFs. *Corticosteroid* and genes of TFs are drawn as a red circle and green circles, respectively. Estimated edges with weights for literature derived regulations are illustrated as orange. Red and blue simulation profiles are roughly distinguished to up-regulated and down-regulated genes, respectively.
doi:10.1371/journal.pone.0105942.g013

Application to Corticosteroid Pathways in Rats

As an application example, we analyzed microarray time-course gene expression data from rat skeletal muscle [37,38], which is assumed to have the same system used in simulation studies. The microarray data were downloaded from the GEO database (GSE490). The time-course gene expression was measured at 0, 0.25, 0.5, 0.75, 1, 2, 4, 5, 5.5, 7, 8, 12, 18, 30, 48, and 72 [h] (16 time points) after the glucocorticoid was applied. The data at time 0 represent controls (untreated). There were two, three, or four replicated observations for each time point.

Because corticosteroid pharmacokinetics/dynamics in skeletal muscle have been modeled based on differential equations [38] as shown in Model S4, the time-dependent concentration of corticosteroid in rat skeletal muscle can be obtained as $\boldsymbol{z}_t$. Furthermore, corticosteroid catabolic/anabolic processes in rat skeletal muscle have been partly established [41]; thus, these regulatory relationships can also be used. Given this information, we included *Mtor*, *Anxa3*, *Bnip3*, *Bcat2*, *Foxo1*, *Trim63*, *Akt1*, *Akt2*, *Akt3*, *Rheb*, *Igf1*, *Igf1r*, *Pik3c3*, *Pik3cd*, *Pik3cb*, *Pik3c2g*, *Slc2a4*, and *Mstn*. Note that the microarray (GSE490) does not include three genes in the original pathway [41], *Redd1*, *Bcaa* and *Klf15*. In addition, we employed the genes, *Irs1*, *Srebf1*, *Rxrg*, *Scarb1*, *Gpam*, *Scd*, *Gpd2*, *Mapk6*, *Ace*, *Ptpn1*, *Ptprf*, *Edn1*, *Agtr1a*, *Ppard*, *Hmgcs2*, *Serpine1*, *Cebpb*, *Cebpd*, *Il6r*, *Mapk14*, *Ucp3*, and *Pdk4*, which have been suggested to be corticosteroid-induced genes [37]. In summary, we applied the method to these 40 genes with weights for the established pathway and the concentration of corticosteroid.

First, to determine the simulation time interval from $\frac{1}{\Delta t}=\{1,2,\ldots,9\}$, we evaluated the BIC scores and SPE ($t=1,2,4$). The results are shown in Fig. 11. Interestingly, even for the observational data, we obtained the same tendency for both indicators. Therefore, we obtained $\frac{1}{\Delta t}=4$ for the lowest SPE. Next, we analyzed the result of $\frac{1}{\Delta t}=4$. The inferred structure with some simulated expression profiles are illustrated in Fig. 12. From the figure, we can capture the propagation of gene expression stimulated by *corticosteroid* and hub genes regulating other genes. However, these results may be difficult to biologically interpret because some mRNAs are not considered to regulate other genes. Therefore, to exploit biological meaning correctly and demonstrate the effectiveness of incorporating prior information in the case of real biological data, we finally performed an experiment using TF information from ITFP (Integrated Transcription Factor Platform) [42]. Then, weights for regulations by TFs, *Trim63*, *Akt1*, *Akt2*, *Mstn*, *Irs1*, *Srebf1*, *Gpam*, *Cebpb*, and *Cebpd*, were set $\frac{1}{w_{n,k}}=10$. The inferred structure at $\frac{1}{\Delta t}=4$ using the TF information is illustrated in Fig. 13.

In Figs. 12 and 13, there are some interesting observations. At first, some genes are directly regulated by corticosteroids, which are included in the model as $\boldsymbol{z}$. Thus, other models that do not include the drug terms cannot estimate such regulation. Second, only weighted regulations, *i.e.*, literature-recorded pathways and regulation by TFs, were inferred in contrast to the non-weighted network in Fig. 12. Thus, we could successfully incorporate prior knowledge, and further candidates may extend our understanding of regulation not yet reported in literature. Additionally, some weighted genes, *Cebpb*, *Mstn*, *Cebpd*, and *Trim63*, were also selected as hub genes with no weight in Fig. 12. Third, *Cebpb*,

Table 5. The confidence levels of estimated pharmacogenomic regulations using GeneNet and G1DBN.

Regulator	Target	*q*–val	post–prob.
Corticosteroid	Srebf1	0.101	0.000
Corticosteroid	Agtr1a	0.864	0.002
Corticosteroid	Cebpd	0.021	0.003
Corticosteroid	Cebpb	0.747	0.003
Trim63	Serpine1	0.375	0.005
Corticosteroid	Mstn	0.198	0.012
Trim63	Irs1	0.385	0.065
Corticosteroid	Scd	0.905	0.068
Akt2	Mtor	0.881	0.069
Cebpb	Il6r	0.836	0.102
Trim63	Ppard	0.395	0.105
Trim63	Slc2a4	0.915	0.189
Corticosteroid	Ucp3	0.663	0.195
Trim63	Bnip3	0.629	0.217
Trim63	Mstn	0.935	0.273
Mstn	Cebpd	0.413	0.280
Irs1	Ptprf	0.928	0.452
Igf1	Pik3cd	0.897	0.457
Trim63	Mapk14	0.909	0.503
Irs1	Anxa3	0.107	0.632
Irs1	Gpd2	0.853	0.749
Corticosteroid	Edn1	0.833	0.799
Corticosteroid	Pik3c2g	0.929	0.821
Cebpb	Trim63	0.864	0.991
Irs1	Akt3	0.396	1.000
Srebf1	Mapk6	0.453	1.000
Corticosteroid	Scarb1	0.651	1.000
Cebpb	Rxrg	0.734	1.000
Corticosteroid	Ptpn1	0.827	1.000
Srebf1	Irs1	0.832	1.000
Akt2	Cebpb	0.863	1.000
Corticosteroid	Pdk4	0.871	1.000
Cebpd	Pik3c2g	0.888	1.000
Irs1	Foxo1	0.894	1.000
Cebpb	Hmgcs2	0.897	1.000
Corticosteroid	Igf1	0.908	1.000
Trim63	Akt3	0.913	1.000
Cebpd	Igf1	0.924	1.000
Irs1	Bnip3	0.925	1.000
Cebpd	Rheb	0.935	1.000
Trim63	Ace	0.936	1.000

doi:10.1371/journal.pone.0105942.t005

which is known as a transcription factor related to immune and inflammatory responses, is indicated as a hub gene (illustrated as a green circle). *Cebpd* and *Cebpb* are assumed to be candidate genes for insulin-related transcription factors [64]. This finding may confirm the findings of previous studies [37,38] indicating that corticosteroid stimulation of skeletal muscle can induce the expression of insulin.

Finally, we applied the other methods, *i.e.*, GeneNet and G1DBN, to the pharmacogenomic data and attached significance levels (*q*-val and *posterior probability* for GeneNet and G1DBN, respectively) for the regulations inferred by the proposed method. The results are presented in Table 5. Interestingly, some regulations have very high significance levels but others do not. For example, regulations of *Srebf1*, *Agtr1a*, *Cebpb* and *Cepbd* by a

corticosteroid are quite probable. In contrast, some regulations were not significant when using these methods. We can suppose, for example, that differences between the models, the prior weights for TF candidates and literature derived pathways, steady state gene expression profiles and corticosteroid drug dynamics in the proposed model may have caused the results. Although some inferred regulations had low significance levels in other approaches, we believe that these regulations can be candidates for true regulation in corticosteroid pharmacogenomic pathways because the proposed method outperformed the other methods through the comparison using synthetic pharmacogenomic pathways.

Although we actually used 40 genes, only 35 genes were found to be regulated because the expression of residual genes did not vary through the time-course. Hence the expression of these genes can represent only synthesis and degradation processes, for which regulation was not estimated.

Discussion

In this study, we proposed a novel method for inference of gene regulatory networks incorporating existing biological knowledge and time-course observation data. The properties of the method are as follows; (i) the dynamics of the gene expression profiles can be estimated based on the proposed linear model with a hidden state, (ii) $L1$ regularized log-likelihood is maximized to infer the active sets of regulation, (iii) the dynamics of other biomolecules can be included in the model, (iv) existing biological knowledge, *e.g.*, literature-recorded pathways and TF information, can be integrated. Furthermore, we proposed an indicator for selecting a simulation time interval for the inference.

To show the effectiveness of the proposed method, we compared it to the previously reported GRNs inference methods using hill function-based pharmacogenomic pathways [38] and a yeast network that is a part of the DREAM4 challenge [39,40]. Since the artificial simulation models were described by differential equations or difference equations, in which the time intervals were smaller than the measurement interval, to reproduce a realistic biological system, the simulated expressions was updated in detail. In this situation, we assumed that the simulation time interval for the method is crucial for inference. As we expected, the results demonstrated that inference of the regulatory structure depends greatly on the simulation time interval. This indicates that we should carefully design the simulation time interval even for analysis of real observational data. For this purpose, we introduced indicators to determine the simulation time interval and measured their validity. Here, since the tendency of the indicator for the simulation time interval depends on the analyzed biological system, it is recommended to check the tendency by using simulation models. Upon comparison of the inferred structures, the proposed method using the indicator showed the highest performance in terms of precision and recall rates for all three data types. The fact that the proposed method outperformed the other methods in using synthetic datasets, which includes the model we do not assume, indicates the adaptability of our proposed method.

For an application example, we applied the proposed method to a corticosteroid-stimulated pathway in rat skeletal muscle. Because pathways and genes related to corticosteroids have been widely investigated, we were able to obtain the concentration of the drug as a function of time from the corticosteroid kinetics/dynamics and the literature-recorded pathways. By incorporating time-course mRNA expression data, corticosteroid kinetics/dynamics, literature-recorded pathways and TF information, we inferred the regulatory relationships among 40 genes that are candidate or known corticosteroid-related genes. The tendency of the BIC scores and the SPE for the simulated time intervals were the same as in the simulation studies, in which the regulatory systems were based on the previous corticosteroid pharmacogenomic studies, and interesting findings for corticosteroid regulation were obtained. For example, genes that are suggested to be significant factors in corticosteroid pharmacogenomics were predicted to be hub genes regulating other genes in the results both with and without prior information. Furthermore, we found that the properties of the proposed method, *i.e.*, the weighted regularization and inclusion of a term for other biomolecules, influenced the results of selecting potential regulators and introducing drug effects to genes, respectively. Finally, these inferred regulations were evaluated by GeneNet and G1DBN, and some of the regulations had high significance. Since our approach imposed prior weights for reliable regulations and included drug terms to explicitly represent their dynamics, not only these regulations but also regulations that are evaluated as non-significant could be candidate regulations for corticosteroid pharmacogenomics. These results indicate that the proposed method can help to elucidate candidates that will allow extension of GRNs in which the regulation among genes is partly understood by incorporating multi-source biological knowledge.

Supporting Information

Method S1 A solution for estimating parameter values and active sets. The detailed solution of estimating parameter values using the EM-algorithm for VAR-SSM with $L1$ regularization.
(PDF)

Model S1 Artificial data and parameter values for dataset (i). The artificial observational data and parameter values for dataset (i).
(ZIP)

Model S2 Artificial data and simulation files for dataset (ii). The artificial observational data and a csml file for dataset (ii).
(ZIP)

Model S3 Artificial data for dataset (iii). The artificial observational data (100 time-courses) for dataset (iii).
(ZIP)

Model S4 Corticosteroid pharmacokinetics/dynamics in rat muscle. A corticosteroid pharmacokinetics/dynamics described in differentia equations in rat muscle [38].
(PDF)

Acknowledgments

The super-computing resource was provided by Human Genome Center, the Institute of Medical Science, the University of Tokyo (http://sc.hgc.jp/shirokane.html).

Author Contributions

Conceived and designed the experiments: TH RY MN. Performed the experiments: TH. Analyzed the data: TH. Contributed reagents/materials/analysis tools: TH RY. Contributed to the writing of the manuscript: TH RY MN SI SM.

References

1. Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. Journal of Computational Biology 7: 601–620.
2. Imoto S, Goto T, Miyano S (2002) Estimation of genetic networks and functional structures between genes by using bayesian networks and nonparametric regression. In: Pacific Symposium on Biocomputing. pp.175–186.
3. Margolin A, Nemenman I, Basso K, Wiggins C, Stolovitzky G, et al. (2006) ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics 7: S7+.
4. Savageau MA (1969) Biochemical systems analysis: II. The steady-state solutions for an n-pool system using a power-law approximation. Journal of Theoretical Biology 25: 370–379.
5. Savageau MA, Voit EO (1987) Recasting nonlinear differential equations as s-systems: a canonical nonlinear form. Mathematical Biosciences 87: 83–115.
6. Lawrence ND, Sanguinetti G, Rattray M (2006) Modelling transcriptional regulation using gaussian processes. In: NIPS. MIT Press, pp.785–792.
7. Rogers S, Khanin R, Girolami M (2007) Bayesian model-based inference of transcription factor activity. BMC Bioinformatics 8.
8. Opper M, Sanguinetti G (2010) Learning combinatorial transcriptional dynamics from gene expression data. Bioinformatics 26: 1623–1629.
9. Henderson J, Michailidis G (2014) Network reconstruction using nonparametric additive ode models. PLoS ONE 9: e94003.
10. Nakamura K, Yoshida R, Nagasaki M, Miyano S, Higuchi T (2009) Parameter estimation of *in silico* biological pathways with particle filtering toward a petascale computing. In: Pacific Symposium on Biocomputing 2009. volume 14, pp.227–238.
11. Nagasaki M, Yamaguchi R, Yoshida R, Imoto S, Doi A, et al. (2006) Genomic data assimilation for estimating hybrid functional petri net from time-course gene expression data. Genome Informatics 17(1): 46–61.
12. Ramsay JO, Hooker G, Campbell D, Cao J (2007) Parameter estimation for differential equations: a generalized smoothing approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69: 741–796.
13. Quach M, Brunel N, d'Alche Buc F (2007) Estimating parameters and hidden variables in non-linear state-space models based on odes for biological networks inference. Bioinformatics 23: 3209–3216.
14. Hasegawa T, Yamaguchi R, Nagasaki M, Imoto S, Miyano S (2011) Comprehensive pharmacogenomic pathway screening by data assimilation. In: Proceedings of the 7th international conference on Bioinformatics research and applications. Berlin, Heidelberg: Springer-Verlag, ISBRA'11, pp.160–171.
15. Bard Y (1974) Nonlinear parameter estimation. New York: Academic Press.
16. Friedman J, Hastie T, Tibshirani R (2007) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9: 432–441.
17. Kim S, Imoto S, Miyano S (2004) Dynamic bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. Biosystems 75(1–3): 57–65.
18. Zacher B, Abnaof K, Gade S, Younesi E, Tresch A, et al. (2012) Joint Bayesian inference of condition-specific miRNA and transcription factor activities from combined gene and microRNA expression data. Bioinformatics 28: 1714–1720.
19. Barenco M, Tomescu D, Brewer D, Callard R, Stark J, et al. (2006) Ranked prediction of p53 targets using hidden variable dynamic modeling. Genome Biology 7: R25+.
20. Beal MJ, Falciani F, Ghahramani Z, Rangel C, Wild DL (2005) A bayesian approach to reconstructing genetic regulatory networks with hidden factors. Bioinformatics 21: 349–356.
21. Hirose O, Yoshida R, Imoto S, Yamaguchi R, Higuchi T, et al. (2008) Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. Bioinformatics 24: 932–942.
22. Rangel C, Angus J, Ghahramani Z, Lioumi M, Sotheran E, et al. (2004) Modeling t-cell activation using gene expression profiling and state-space models. Bioinformatics 20: 1361–1372.
23. Sabatti C, James GM (2006) Bayesian sparse hidden components analysis for transcription regulation networks. Bioinformatics 22: 739–746.
24. Asif HMS, Sanguinetti G (2011) Large-scale learning of combinatorial transcriptional dynamics from gene expression. Bioinformatics 27: 1277–1283.
25. Eduati F, De Las Rivas J, Di Camillo B, Toffolo G, Saez-Rodriguez J (2012) Integrating literatureconstrained and data-driven inference of signalling networks. Bioinformatics 28: 2311–2317.
26. do Rego TG, Roider HG, de Carvalho FAT, Costa IG (2012) Inferring epigenetic and transcriptional regulation during blood cell development with a mixture of sparse linear models. Bioinformatics 28: 2297–2303.
27. Greenfield A, Hafemeister C, Bonneau R (2013) Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. Bioinformatics 29: 1060–1067.
28. Dong CY, Shin D, Joo S, Nam Y, Cho KH (2012) Identification of feedback loops in neural networks based on multi-step granger causality. Bioinformatics 28: 2146–2153.
29. Kojima K, Yamaguchi R, Imoto S, Yamauchi M, Nagasaki M, et al. (2010) A state space representation of var models with sparse learning for dynamic gene networks. Genome informatics International Conference on Genome Informatics 22: 56–68.
30. Efron B, Hastie T, Johnstone L, Tibshirani R (2004) Least angle regression. Annals of Statistics 32: 407–499.
31. Schäfer J, Strimmer K (2005) An empirical bayes approach to inferring large-scale gene association networks. Bioinformatics 21: 754–764.
32. Opgen-Rhein R, Strimmer K (2007) From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC Systems Biology 1: 37.
33. Lébre S (2009) Inferring dynamic genetic networks with low order independencies. Statistical Applications in Genetics and Molecular Biology 8: 1–38.
34. Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9: 432–441.
35. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, et al. (2007) Large-scale mapping and validation of *escherichia coli* transcriptional regulation from a compendium of expression profiles. PLoS Biol 5: e8.
36. Meyer P, Kontos K, Lafitte F, Bontempi G (2007) Information-theoretic inference of large transcriptional regulatory networks. EURASIP Journal on Bioinformatics and Systems Biology 2007: 79879.
37. Almon RR, DuBois DC, Jin JY, Jusko WJ (2005) Temporal profiling of the transcriptional basis for the development of corticosteroid-induced insulin resistance in rat muscle. Journal of Endocrinology 184: 219–232.
38. Yao Z, Hoffman EP, Ghimbovschi S, DuBois DC, Almon RR, et al. (2008) Mathematical modeling of corticosteroid pharmacogenomics in rat muscle following acute and chronic methylprednisolone dosing. Molecular Pharmaceutics 5: 328–339.
39. Marbach D, Schaffter T, Mattiussi C, Floreano D (2009) Generating Realistic In Silico Gene Networks for Performance Assessment of Reverse Engineering Methods. Journal of Computational Biology 16: 229–239.
40. Schaffter T, Marbach D, Floreano D (2011) Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. Bioinformatics 27: 2263–2270.
41. Shimizu N, Yoshikawa N, Ito N, Maruyama T, Suzuki Y, et al. (2011) Crosstalk between Glucocorticoid Receptor and Nutritional Sensor mTOR in Skeletal Muscle. Cell metabolism 13: 170–182.
42. Zheng G, Tu K, Yang Q, Xiong Y, Wei C, et al. (2008) Itfp: an integrated platform of mammalian transcription factors. Bioinformatics 24: 2416–2417.
43. Elowitz MB, Leibler S (2000) A synthetic oscillatory network of transcriptional regulators. Nature 403: 335–338.
44. de Jong H (2002) Modeling and simulation of genetic regulatory systems: A literature review. Journal of Computational Biology 9: 67–103.
45. Hazra A, DuBois DC, Almon RR, Snyder GH, Jusko WJ (2008) Pharmacodynamic modeling of acute and chronic effects of methylprednisolone on hepatic urea cycle genes in rats. Gene Regulation and Systems Biology 2: 1–19.
46. Jin JY, Almon RR, DuBois DC, Jusko WJ (2003) Modeling of corticosteroid pharmacogenomics in rat liver using gene microarrays. Journal of Pharmacology and Experimental Therapeutics 307: 93–109.
47. Chen KC, Wang TY, Tseng HH, Huang CYF, Kao CY (2005) A stochastic differential equation model for quantifying transcriptional regulatory network in saccharomyces cerevisiae. Bioinformatics 21: 2883–2890.
48. Lillacci G, Khammash M (2010) Parameter estimation and model selection in computational biology. PLoS Comput Biol 6: e1000696.
49. Sun X, Jin L, Xiong M (2008) Extended Kalman Filter for Estimation of Parameters in Nonlinear State-Space Models of Biochemical Networks. PLoS ONE 3: e3758+.
50. Kalman RE (1960) A New Approach to Linear Filtering and Prediction Problems. Transactions of the ASME - Journal of Basic Engineering: 35–45.
51. Shumway RH, Stoffer DS (1982) An approach to time series smoothing and forecasting using the em algorithm. Journal of Time Series Analysis 3: 253–264.
52. Maybeck PS (1979) Stochastic models, estimation and control. Volume I. Academic Press.
53. Julier SJ, Uhlmann JK (1997) A new extension of the kalman filter to nonlinear systems. In: Proc. of AeroSense: The 11th Int. Symp. on Aerospace/Defense Sensing, Simulations and Controls. pp.182–193.
54. Julier S, Uhlmann J (2004) Unscented filtering and nonlinear estimation. Proceedings of the IEEE 92: 401–422.
55. Kitagawa G (1996) Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models. Journal of Computational and Graphical Statistics 5: 1–25.
56. Liu X, Niranjan M (2012) State and parameter estimation of the heat shock response system using kalman and particle filters. Bioinformatics 28: 1501–1507.
57. Shimamura T, Imoto S, Yamaguchi R, Nagasaki M, Miyano S (2010) Inferring dynamic gene networks under varying conditions for transcriptomic network comparison. Bioinformatics 26: 1064–1072.
58. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B 39: 1–38.
59. Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6: 461–464.
60. Yamaguchi R, Higuchi T (2006) State space approach with the maximum likelihood principle to identify the system generating time-course gene expression data of yeast. Int J Data Min Bioinformatics 1: 77–87.
61. Zou H, Hastie T, Tibshirani R (2007) On the "degrees of freedom" of the lasso. The Annals of Statistics 35: 2173–2192.

62. Shimamura T, Imoto S, Yamaguchi R, Fujita A, Nagasaki M, et al. (2009) Recursive regularization for inferring gene networks from time-course gene expression profiles. BMC systems biology 3: 41.
63. Tamada Y, Yamaguchi R, Imoto S, Hirose O, Yoshida R, et al. (2011) Sign-ssm: open source parallel software for estimating gene networks with state space models. Bioinformatics 27: 1172–1173.
64. Foti D, Iuliano R, Chiefari E, Brunetti A (2003) A nucleoprotein complex containing sp1, c/ebpb, and hmgi-y controls human insulin receptor gene transcription. Molecular and Cellular Biology 23: 2720–2732.