OPEN ACCESS Freely available online

PLOS ONE

A Kalman-Filter Based Approach to Identification of Time-Varying Gene Regulatory Networks

Jie Xiong[1]*, Tong Zhou[2]

1 Department of Automation, Tsinghua University, Beijing, China, **2** Department of Automation and Tsinghua National Laboratory for Information Science and Technology(TNList), Tsinghua University, Beijing, China

Abstract

Motivation: Conventional identification methods for gene regulatory networks (GRNs) have overwhelmingly adopted static topology models, which remains unchanged over time to represent the underlying molecular interactions of a biological system. However, GRNs are dynamic in response to physiological and environmental changes. Although there is a rich literature in modeling static or temporally invariant networks, how to systematically recover these temporally changing networks remains a major and significant pressing challenge. The purpose of this study is to suggest a two-step strategy that recovers time-varying GRNs.

Results: It is suggested in this paper to utilize a switching auto-regressive model to describe the dynamics of time-varying GRNs, and a two-step strategy is proposed to recover the structure of time-varying GRNs. In the first step, the change points are detected by a Kalman-filter based method. The observed time series are divided into several segments using these detection results; and each time series segment belonging to two successive demarcating change points is associated with an individual static regulatory network. In the second step, conditional network structure identification methods are used to reconstruct the topology for each time interval. This two-step strategy efficiently decouples the change point detection problem and the topology inference problem. Simulation results show that the proposed strategy can detect the change points precisely and recover each individual topology structure effectively. Moreover, computation results with the developmental data of *Drosophila Melanogaster* show that the proposed change point detection procedure is also able to work effectively in real world applications and the change point estimation accuracy exceeds other existing approaches, which means the suggested strategy may also be helpful in solving actual GRN reconstruction problem.

Citation: Xiong J, Zhou T (2013) A Kalman-Filter Based Approach to Identification of Time-Varying Gene Regulatory Networks. PLoS ONE 8(10): e74571. doi:10.1371/journal.pone.0074571

Editor: Raya Khanin, Memorial Sloan Kettering Cancer Center, United States of America

Received January 16, 2013; **Accepted** August 4, 2013; **Published** October 7, 2013

Funding: This work was financially supported in part by the 973 Program under Grant 2012CB316504 and 2009CB320602, and by the National Natural Science Foundation of China under Grants 61174122, 61021063, 60721003 and 60625305. 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.

* E-mail: xiongj08@mails.tsinghua.edu.cn

Introduction

Identifying causal relationships of a gene regulatory network (GRN) is one of the fundamental problems in understanding cell behaviors. For most conventional identification methods, it is generally assumed the topological structure is constant over time. Based on this assumption, various models and methods have been proposed, such as Boolean networks [1], Bayesian networks [2], regression and correlation analyses based methods [3], ordinary differential equation (ODE) based methods [4], etc.

Recent research results, however, show that GRNs are dynamic in response to physiological and environmental changes. For instance, an example of such time-varying regulatory network can be provided by the development of the fruitfly *Drosophila Melanogaster*, which is segmented into different life stages: embryogenesis, larva, pupa and adult [5]. Moreover, some studies have also confirmed that the active regulatory paths in a gene expression network of *Saccharomyces cerevisiae* exhibit dramatic topological changes and hub transience during a temporal cellular process and in response to diverse stimuli [6]. Although there is a rich literature in modeling static or temporally invariant networks, how to systematically recover these temporally changing networks remains a major and significant pressing challenge.

To identify time-varying GRNs, some special methods have been proposed recently. A machine learning method called TESLA is presented in [7], which builds on a temporally smoothed l_1-regularized logistic regression formalism that can be cast as a standard convex-optimization problem and solved by using generic solvers scalable to large networks. However, the estimated topology by this method is undirected and suboptimal. In addition, there exist some methods that follow the Bayesian paradigm [8–11]. While these approaches also have their limitations. The method suggested in [8] assumes a fixed network structure and only allows the interaction parameters to vary with time, which is too rigid and idealistic in practice. The method proposed in [9] requires a discretization of the data, which incurs an ineluctable information loss. And, the limitation in [10,11] is that these methods need prior distributions on the network structure.

The purpose of this study is to suggest a two-step strategy that recovers time-varying GRNs. In this paper, the model for time-varying GRNs is adopted as the switching auto regressive model.

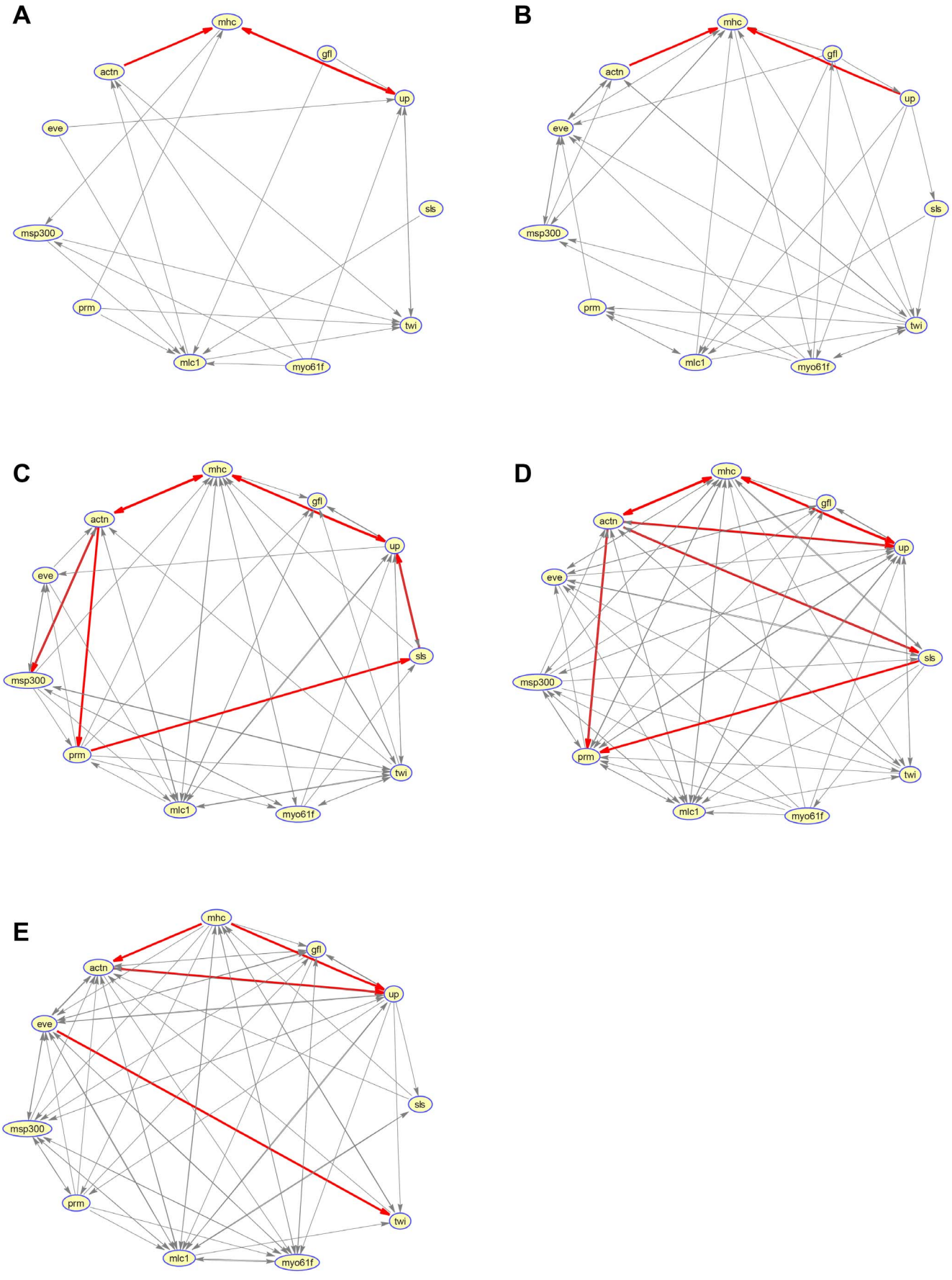
A
mhc
gfl
actn
up
eve
sls
msp300
prm
twi
mlc1
myo61f
B
mhc
gfl
actn
up
eve
sls
msp300
prm
twi
mlc1
myo61f
C
mhc
gfl
actn
up
eve
sls
msp300
prm
twi
mlc1
myo61f
D
mhc
gfl
actn
up
eve
sls
msp300
prm
twi
mlc1
myo61f
E
mhc
gfl
actn
up
eve
sls
msp300
prm
twi
mlc1
myo61f

Figure 1. Gene regulatory networks recovered from gene expression time series of *Drosophila Melanogaster*. For each physiological stage, a network has been reconstructed(A: Adult, B: Embryo1, C: Embryo2, D: Larva, E: Pupa). And, interactions that have been verified are marked in red.
doi:10.1371/journal.pone.0074571.g001

Consequently, in the first step, on the basis of a relation between the Kalman filter and recursive least squares (RLS) estimation, it is shown that a stochastic process can be constructed which is white if and only if the time series expression data are generated by the same sub-regulatory network. Based on this observation, a procedure is developed to detect change points of a time-varying GRN. The observed time series are divided into several segments using these detection results; and each time series segment belonging to two successive demarcating change points is associated with an individual static regulatory network. And then, in the second step, conditional network structure identification methods are used to reconstruct the topology for each time interval. In summary, in the suggested time-varying GRN identification strategy, the problem of identifying a time-varying regulatory network is transformed into that of identifying multiple single static regulatory networks. To solve the latter is much easier than to solve the former. For the performance evaluation, we use both time series data generated by a synthetic time-varying GRN and time series data provided by the DREAM3 challenge, and simulation results confirm that the proposed strategy can detect the change points precisely and recover each individual topological structure effectively. For a real data application, our proposed strategy is applied to time series data of *Drosophila Melanogaster* during a complete time-course of development, and computation results show that the proposed change point detection procedure has ability to work effectively in real world applications and the change point positioning accuracy exceeds other existing approaches, which means the suggested strategy may also be helpful in solving actual GRN reconstruction problem.

The rest of this paper is organized as follows. At first, the problem discussed in this paper and some mathematic preliminary results are given, then the change points estimation algorithm is illustrated and the two-step strategy is derived. Afterwards, the proposed estimation strategy is assessed using both *In Silico* data and the developmental data of *Drosophila Melanogaster*. Variations of estimation performances with respect to parameters of the suggested method will also be reported. Besides, some concluding remarks are given about the characteristics of the suggested method, as well as some future works worthy of further efforts. Finally, an appendix is included in Text S1 to give proof of some technical results.

The following notations are adopted in this paper. $\mathbf{vec}(X)$ denotes the operation of stacking the columns of matrix X from left to right, and $A\otimes B$ the Kronecker product of matrices A and B. $\mathbf{E}[x]$ and $\mathbf{E}[x|y]$ stand respectively for the expected value of a random variable x and the conditional expected value of a random variable x given an observation of the random variable y. Y_t is defined as $Y_t=\{y_t\ y_{t-1}\ \cdots\ y_1\}$, while $\hat{x}_{j|t}$ represents an estimate about x_j based on some observed data from $k=1$ to $k=t$. Both δ_{ij} and $\delta_{i,j}$ are used to denote the Kronecker delta function. To avoid an awkward presentation, no difference is made in this paper between a random variable and its realizations.

Methods

Problem Statement and Preliminary Results

Generally speaking, a basic model for a time-varying GRN consisting of n genes can be expressed as [12]

$$y_t=A_t y_{t-1}+w_t \tag{1}$$

Here, $y_t\in\mathcal{R}^n$ is the time series experiment data of gene expression, and $w_t\in\mathcal{R}^n$ is the Guassian white noise as $w_t\sim N(0,R)$. The system matrix $A_t\in\mathcal{R}^{n\times n}$ captures the causal relationships of genes, that is, if $(A_t)_{ij}$ is positioned significantly far from zero, the j-th gene captures a large effect on the i-th gene from time point $t-1$ to t. On the other hand, it is generally assumed that the regulatory mechanism among genes is unlikely to change drastically over small time intervals [7,13], which presumes that the coefficient matrix A_t should vary smoothly with time. Based on this assumption, Equation (1) can be expressed as the switching auto regressive (SAR) model as follows:

$$y_t=A_{\lambda(t)}y_{t-1}+w_t,\lambda(t)\in\{1,2,\cdots,s\} \tag{2}$$

By means of the model described by Equation (2), some concepts are given. The so-called "change point" means that the time instant at which its discrete state $\lambda(t)$ changes value, and the time series segment belonging to two successive demarcating change points is associated with an individual static regulatory network which the causal relationship can be captured by $A_{\lambda(t)}$.

It has been pointed out that over the course of a cellular process, such as a cell cycle or an immune response, there may exist multiple underlying themes that determine the functionalities of each molecule and their relationships to each other, and such themes are dynamic and stochastic [7]. As a result, GRNs are dynamic in response to physiological and environmental changes.

In general, normal biological tissues will undergo morphologic changes when they are inflicted by some external stimuli, such as ionizing radiations. In fact, many literatures have studied the radiation tolerance [14–16], especially, the "Time-Dose" relationships [17]; that is to say, the time span that the normal biological networks remain unchanged when they are eroded by certain amount of ionization radiation is unknown. In other words, the change points are not known *a priori* in this type experiment. And by extension, the change points are not always known *a priori* in general. Therefore, it is assumed that the change point is unknown and its value is needed to estimate in some literatures on the identification of time-varying GRNs [10,11]. We also hold this assumption in the paper.

Based on the discussion above, the time-varying GRN identification problem discussed in this paper is as follows.

Problem. Given a series of gene expression vectors y_t generated by model (2), $t=1,2,\cdots,N$, estimate all the change points, the number of sub-networks s, and the model parameters $A_{\lambda(t)}$, $\lambda(t)\in\{1,2,\cdots,s\}$.

It is well known that the computation procedure of recursive least squares (RLS) parametric estimations for AR models possesses the same form as that of Kalman filtering [18,19]. Using these similarities, some system identification problems can be easily transformed into a state estimation one, and vice versa. To investigate the above change point estimation problem, some relations are introduced here between RLS estimations of an AR system and Kalman filtering.

Consider the following linear time invariant (LTI) AR system

$$y_t = Ay_{t-1} + w_t \tag{3}$$

in which w_t is a sequence of independent random vectors with zero mean and covariance matrix R. Rewrite Equation (3) into a state-space form as follows,

$$x_{t+1} = x_t, \quad y_t = h_t x_t + w_t \tag{4}$$

Here, $x_t = \mathbf{vec}(A^T)$, $h_t = I_n \otimes y_{t-1}^T$. Assume that x_0 is an *a priori* unbiased estimate about $\mathbf{vec}(A^T)$ and its covariance matrix is Π_0. Moreover, assume that this estimate is independent of w_t with $t \geq 1$. Through adopting the general results of [18,19] on relations between RLS parametric identification and state estimation to the above LTI AR system, the following results can be straightforwardly obtained.

Lemma 1. Set $\hat{x}_{0|0}$ and $P_{0|0}$ respectively as $\hat{x}_{0|0} = x_0$ and $P_{0|0} = \Pi_0$. Based on gene expression time series data y_t generated by model (3), $t = 1,2,\cdots,N$, the RLS estimate for its model parameters A, denote it by $\hat{x}_{t|t}$, can be recursively computed as follows,

$$\hat{x}_{t|t} = \hat{x}_{t-1|t-1} + K_t\left[y_t - h_t\hat{x}_{t-1|t-1}\right] \tag{5}$$

$$K_t = P_{t-1|t-1}h_t^T\left[R + h_t P_{t-1|t-1}h_t^T\right]^{-1} \tag{6}$$

$$P_{t|t} = P_{t-1|t-1} - P_{t-1|t-1}h_t^T\left[R + h_t P_{t-1|t-1}h_t^T\right]^{-1}h_t P_{t-1|t-1} \tag{7}$$

Moreover, if both x_0 and w_t are normally distributed, then, $\hat{x}_{t|t}$ is also normally distributed, and $\hat{x}_{t|t} = \mathbf{E}[x_t|Y_t]$.

Change point Detection Procedure

In the time-varying GRN identification, a common situation is that available knowledge about the actual network topology is nothing but its gene expression data. In order to develop the change points detection procedure, it appears appropriate to investigate at first whether or not there exist some detectable stochastic differences between gene expression data generated by the same sub-network and those generated by more than one sub-network. If the answer is positive, then a change of these stochastic properties reflects a switch between two sub-networks. In other words, a statistic can be constructed for estimating change points of a time-varying GRN. Based on these considerations, stochastic properties of an innovation process of the network described by Equation (2) are investigated with respect to the recursive estimation procedure given by Equations (5)–(7), in case that gene expression data are generated respectively by a single sub-network and multiple sub-networks.

Theorem 1. Suppose that gene expression time series data $y_t|_{t=1}^N$ are generated by the time-varying GRN described by Equation (2). On the basis of the recursive estimation procedure of Equations (5)–(7), define an innovation process $e_t|_{t=1}^N$ as follows,

$$e_t = y_t - h_t\hat{x}_{t-1|t-1} \tag{8}$$

Then, $e_t|_{t=1}^N$ is an independent random sequence if and only if the network described by Equation (2) collapses to a static network.

A proof of the above theorem is given in Text S1. Furthermore, from the above discussions, a direct result of Equation (2) is that y_t, $t = 1,2,\cdots,N$, is also normally distributed. On the other hand, from properties of Kalman filtering, it can be declared that $h_t\hat{x}_{t-1|t-1} = \mathbf{E}\{y_t|Y_{t-1}\}$. Therefore, e_t defined in Equation (8) is a normally distributed random vector. On the basis of Theorem 1 and its extensions, the following results can be obtained through straightforward algebraic manipulations. These results are very helpful in detecting the change point.

Corollary 1. Assume that for arbitrary $j,k = 1,2,\cdots,s$ and $j \neq k$, there exists at least one scenario that $A_j \neq A_k$. For gene expression time series data $y_t|_{t=1}^N$ generated by model (2), define $\hat{x}_{t|t}$ recursively using the procedure of Equations (5)–(7). Moreover, define a time series $\bar{e}_t|_{t=1}^N$ as follows

$$\bar{e}_t = \left(R + h_t P_{t-1|t-1}h_t^T\right)^{-1/2}(y_t - h_t\hat{x}_{t-1|t-1}) \tag{9}$$

Then, $\bar{e}_t|_{t=1}^N$ is a sequence of independently distributed random variables with zero mean and unit covariance matrix, if and only if there is no sub-network switch during the time period $1 \leq t \leq N$.

Corollary 1 makes it clear that change point estimation for a time-varying GRN described by Equation (2) can be transformed to independence validation of a Gaussian random sequence. The latter can be checked by chi-square test. Based on the definition of the χ^2-distribution, we have the following result.

Corollary 2. Based on the conditions of Corollary 1, define Q as

$$Q = \sum_{i=t-1}^{t} \bar{e}_i^2 \tag{10}$$

Then, Q obeys the χ^2-distribution with degrees of freedom $2n$, if and only if there is no sub-network switch during the time period $t-1 \leq i \leq t$.

The results of Corollary 2 are helpful in detecting the change point. As a matter of fact, in actual applications, if $Q \leq \chi^2_{1-\alpha}(2n)$, then, the hypothesis that the collected gene expression data are generated by the same sub-network can not be rejected with a confidence level $1-\alpha$. Based on the results of Corollary 1 and Corollary 2, a procedure can be developed for detecting the change point. Details of this procedure are given in Table 1.

An attractive property of the change point detection procedure is that its computational complexity does not depend on the number of change points. Moreover, it is also worthwhile to point out that in this detection procedure, neither prior distribution on the number of change points nor knowledge about the change time instant is required, i.e., our change points detection procedure do not require the structure prior distribution of a GRN, which is the major difference from the method proposed in [10,11].

Two-Step Strategy

In the above subsection, the change point estimates have been obtained, which are denoted by $t_s^i, i = 1,2,\cdots,h$. Base on these change point estimates, the observed gene expression time series data are divided into $h+1$ segments, which are $L_1 : y_t|_{t=1}^{t_s^1-1}, L_2 : y_t|_{t_s^1}^{t_s^2-1}, \ldots, L_{h+1} : y_t|_{t_s^h}^{N}$; and each time series segment belonging to two successive demarcating change points is supposed to associate with an individual static GRN. Consequently, for each time interval, the causal relationships inference problem can resort to conditional network structure identification

Table 1. Change point Detection Procedure.

S1:	Initialization: Select a positive number $\alpha\in(0,1)$. Set $\hat{x}_{0\|0}=0$, $P_{0\|0}=I$.
S2:	Calculate recursively $\bar{e}_t$ using Equation (9) and the procedure of Equations (5)–(7).
	If the number of the computed $\bar{e}_t$ is greater than 2, compute the statistic Q.
S3:	If $Q\le\chi^2_{1-\alpha}(2n)$, the current time instant isn't a change point. Let $t\rightarrow t+1$ and return
	to S2. Otherwise, record the current time instant as a change point, assign it to be the
	initial time for detecting the next change point, and return to S1.
S4:	Stop the procedure in case that every gene expression data has been utilized.

doi:10.1371/journal.pone.0074571.t001

methods. The suggested two-step strategy inference method for time-varying GRNs is summarized as follows.

1. Estimate change points using the change point detection procedure in Table 1.
2. For each time series segment, infer the causal relationships by conditional network structure identification methods, such as IOTA [20], and LASSO [21], etc.

In summary, the suggested two-step strategy can decouple the change point detection problem and the topology inference problem, that is, the problem of identifying a time-varying regulatory network is transformed into that of identifying multiple single static regulatory networks. To solve the latter is much easier than to solve the former.

Remark. It should be noted that the number of the biological experimental time series data is very limited. If there exist some *a priori* information about the network topology, it is also desirable to jointly learn the static networks across time segments. A feasible method is as follows.

Suppose that the time series $y_t|_{t=1}^{N}$ is cut into two segments $P\overset{\Delta}{=}\{y_t|_{t=1}^{k}\}$, and $Q\overset{\Delta}{=}\{y_t|_{t=k+1}^{N}\}$ by the suggested change point detection procedure. If gene j regulate gene i throughout this time period based on other *a priori* knowledge, then we can set the initial value $(A_Q)_{ij}$ as $\left(\hat{A}_P\right)_{ij}$ that is obtained from the data segment P, when learn the network topology based on the data segment Q. Therefore, jointly learning the static networks across time segments can be done by this way.

Results and Discussion

Simulation Study

In order to evaluate the properties of the suggested two-step strategy, gene expression time series data are generated by an academic dynamic network. This simulated dynamic network include two sub-networks denoted by A_1 and A_2. The simulation time span is 60, and at time instant 31, the active sub-network is changed from A_1 to A_2. The simulated dynamic network include 10 genes; and the nonzero elements for A_1 are $(A_1)_{1,1}=0.02656$,$(A_1)_{2,2}=-0.0324$,$(A_1)_{7,2}=0.0767$,$(A_1)_{3,3}=0.1900$, $(A_1)_{8,3}=0.2004$,$(A_1)_{9,3}=0.0803$,$(A_1)_{5,4}=0.7089$, $(A_1)_{2,5}=0.2441$, $(A_1)_{4,5}=0.0265$,$(A_1)_{5,6}=-0.0183$, $(A_1)_{6,6}=-0.0215$,$(A_1)_{1,7}=0.5605$, $(A_1)_{4,7}=0.1922$,$(A_1)_{7,7}=0.2841$, $(A_1)_{3,8}=0.6424$,$(A_1)_{10,8}=0.0982$, $(A_1)_{8,9}=0.2512$,$(A_1)_{9,9}=0.2605$, $(A_1)_{2,10}=0.5514$, $(A_1)_{6,10}=0.3873$, $(A_1)_{10,10}=0.5897$ respectively; while the nonzero elements for A_2 are $(A_2)_{1,1}=0.03778$,$(A_2)_{8,1}=0.3624$, $(A_2)_{2,2}=0.3975$, $(A_2)_{9,2}=-0.0009$, $(A_2)_{7,3}=-0.0228$, $(A_2)_{3,4}=0.6232$, $(A_2)_{10,4}=-0.0157$, $(A_2)_{4,5}=0.5368$, $(A_2)_{5,6}=-0.0222$, $(A_2)_{8,6}=0.2090$, $(A_2)_{6,7}=0.4607$, $(A_2)_{7,7}=0.3293$, $(A_2)_{5,8}=0.4301$, $(A_2)_{8,8}=0.3317$, $(A_2)_{2,9}=0.2580$, $(A_2)_{4,9}=0.0402$, $(A_2)_{9,9}=0.2110$, $(A_2)_{1,10}=0.2886$, $(A_2)_{10,10}=0.0578$, respectively. The noise w_t is a sequence of independent Gaussian random variable with mean 2 and variance 0.5.

In the second step, we apply a recent identification method, named the inner composition alignment (IOTA) [20]. In IOTA, a measurement τ is defined to characterize the causality of two time series. For the given short time series $y^{(l)}$ and $y^{(k)}$, sort $y^{(l)}$ with the order $\psi^{(l)}$, such that $\forall i,\left[y^{(l)}\left(\psi^{(l)}\right)\right]_i\le\left[y^{(l)}\left(\psi^{(l)}\right)\right]_{i+1}$, and reorder the time series $y^{(k)}$ with respect to $\psi^{(l)}$ as $g^{(k,l)}=y^{(k)}\left(\psi^{(l)}\right)$. Define τ as follows,

$$\tau_{kl}=1-\frac{\sum_{i=1}^{m}\sum_{j=i+1}^{m-1}\omega_{ij}\Theta\left[\left(g_{j+1}^{(k,l)}-g_i^{(k,l)}\right)\left(g_i^{(k,l)}-g_j^{(k,l)}\right)\right]}{\Delta}$$

Here, m is the length of the time series, $\Delta=(m-1)(m-2)/2$ is a normalization constant which corresponds to the maximum number of crossings, ω_{ij} denotes a weight, and $\Theta[x]$ is the Heaviside step function,

$$\Theta[x]\begin{cases}1, x>0\\ 0, x\le 0\end{cases}$$

In the last ranking list of τ_{kl} , if the magnitude is larger, the corresponding transcription regulation will be established in a larger probability from gene l to gene k.

In systems biology, predictions are compared with the actual network structure using the following two different metrics in topology prediction accuracy evaluations.

- AUPR: The area under the precision-recall curve;
- AUROC: The area under the receiver operating characteristic curve.

Choose α as 0.05, independently simulate this dynamic network 500 times, and the results are summarized in Table 2.

The change point estimated mean in Table 2 are very close to the actual change time instant, and the variance of these estimates

Table 2. Performances with simulation data.

	changetime estimate	AUROC1	AUPR1	AUROC2	AUPR2
mean	31.0900	0.6208	0.4043	0.6454	0.3414
standard deviation	0.2865	0.0410	0.0588	0.0417	0.0476

doi:10.1371/journal.pone.0074571.t002

is very small, which shows our change point detection procedure is effectively. And, the estimate variances of AUROC and AUPR are quite small. It is concluded that the main difficulty of the switched autoregressive exogenous model identification is that the identification problem includes a classification problem in which each data point must be associated to the most suitable sub-model; and, the more precise the data classification is, the better the identification results are [22]. This argument is also suitable for the scenario of identifying a time-varying GRN. Due to the high accuracy of the change point estimates, the estimate variances of AUROC and AUPR are quite small. Therefore, these simulation results show that our two-step strategy is appropriate for reconstructing time-varying GRNs.

The above simulation is an academic case, in which the experimental environment is very close to the fundamental assumption in the section of Methods. In the rest of this subsection, we will give another simulation, in which gene expression data are from the DREAM3 *in silico* size10 challenge [23,24]. Although gene expression data in the DREAM challenges are emulational, the simulation model in the DREAM challenges is nonlinear and the noise is not exactly Gaussian. Therefore, the source networks are closer to the real situation. Due to the static nature of the network in the DREAM challenges, we concatenate gene expression data generated by 5 different networks to simulate a time-varying GRN. Applying our strategy to these data, the simulation results are shown in Table 3.

From Table 3, we know the actual change time instants are 22, 43, 64, 85, respectively, and our change point detection procedure estimate these change points accurately. That is, gene expression time series are successfully divided into five segments, and the problem of identifying a time-varying regulatory model is effectively transformed into that of identifying five single regulatory models. Consequently, conditional identification methods for the static GRN are used to reconstruct the topology for each segment, which means that the proposed two-step strategy can ease the difficulty level in recovering a time-varying GRN.

Real Data Application

The gene expression data of *Drosophila Melanogaster* have been well-studied in different aspects [5,10,11,25]. Here, we apply our two-step strategy to the developmental data provide by [5]. In this study, the researchers have reported gene expression data for nearly one third of all *Drosophila Melanogaster* genes during a complete time-course of development. And, cDNA microarrays were used to analyze the RNA expression levels during 66 sequential time periods, including the embryonic period (30 samples), the larval period (10 samples), the pupal period (18 samples) and the first 30 days of adulthood (8 samples). In addition, a major morphological change relates to a modification of transcriptional regulations during the first 0 to 6.5 hours of embryonic development, which consists of 12 samples. Therefore, the actual change instants ought to be 13, 31, 41, and 59. Here, we use a sub-dateset of this developmental data, containing the following 11 genes: 'actn','eve','gfl','mhc','mlc1','msp300','myo61f','prm','sls','twi', and 'up'.

To apply our change point detection procedure, we first estimate noise covariance matrix $\boldsymbol{R}$. Reformulize the basic model (1) as

$$y_t = A_t y_{t-1} + w_t = \phi_t \theta_t + w_t$$

Here, $\theta_t = \mathbf{vec}(A_t^T)$, $\phi_t = I_n \otimes y_{t-1}^T$. Based on this formulation, we utilize the weighted recursive least square algorithm to estimate $\boldsymbol{R}$ [18,19]. Concretely, given the initial condition P_0 and $\hat{\theta}_0$, the recursive expression equations to calculate θ_t are given as follows.

$$\hat{\theta}_t = \hat{\theta}_{t-1} + K_t\left[y_t - \phi_t \hat{\theta}_{t-1}\right] \tag{11}$$

$$K_t = P_{t-1}\phi_t^T\left[\phi_t P_{t-1}\phi_t^T + \mu I\right]^{-1} \tag{12}$$

$$P_t = \mu^{-1}[I - K_t\phi_t]P_{t-1} \tag{13}$$

On the basis of the weighted recursive least square procedure of Equations (11)–(13), define a residual sequence $\varepsilon_t|_{t=1}^N$ as follows,

$$\varepsilon_t = y_t - \phi_t \hat{\theta}_t. \tag{14}$$

Then, the noise covariance matrix $\boldsymbol{R}$ can be estimated as

$$\hat{R} = \frac{1}{N}\sum_{t=1}^{N} \varepsilon_t \varepsilon_t^T. \tag{15}$$

Table 3. Performances with DREAM3 *in silico* size10 challenge.

segment number	starting point (actual/estimate)	ending point (actual/estimate)	AUROC	AUPR
1	1/1	21/21	0.7344	0.2007
2	22/22	42/42	0.6933	0.2893
3	43/43	63/63	0.6011	0.1354
4	64/64	84/84	0.6789	0.3522
5	85/85	105/105	0.7681	0.4408

doi:10.1371/journal.pone.0074571.t003

The typical initial value can be selected as $P_0=\delta I$, where δ is a large number, $\hat{\theta}_0=0$, and the weighting coefficient μ is typically chosen as $0.98<\mu<1$ [18,19].

By setting $\delta=10^3,\mu=0.985,\alpha=0.05$, the change point estimates by the suggested change point detection procedure are 13, 31, 46, 59. Although the third change point estimate is a little away from the actual value, the other changetime estimates are equal to the actual change instants. On the other hand, using the same dataset, a report that time intervals {18 to 19}, {31 to 33}, {41 to 43} and {59 to 61} contain more than 40% of the change points can be found in [10]; and in [11], the authors have given only the last three change points. These results show that the proposed change point detection procedure appears to exceed the the two alternative methods, and our approach has ability to work effectively in real world applications.

Based on these change point estimates, the developmental data of *Drosophila Melanogaster* have divided into five segments. For each segment, we use the well-studied LASSO model in statistics to infer the causal relationship [21]. Specifically, let $\mathrm{Y}=[y_1,y_2,\cdots,y_N]$, and $\mathrm{Y}_1=\mathrm{Y}(:,1:N-1)$, $\mathrm{Y}_2=\mathrm{Y}(:,2:N)$, then reconstructing a static GRN can be formulized as follows,

$$\begin{aligned}&\min\|A\mathrm{Y}_1-\mathrm{Y}_2\|_2+\lambda\|A\|_1\\&s.t.\ \sum_{j=1}^{n}|a_{ij}|\le 1,\text{for all } i=1,\cdots,n\end{aligned}\tag{16}$$

The elements on each row of matrix A are independent and constraints in (16) are independent of each row in matrix A. Therefore, optimization problem (16) can be reduced to n optimization problems, each having n variables which are elements on a row of matrix A. That is, for each row in matrix A, i.e., for each gene i ($i=1,2,\cdots,n$), we have

$$\begin{aligned}&\min\left\|\mathrm{Y}_1^T a_i-\mathrm{Y}_{2j}\right\|_2+\lambda\|a_i\|_1\\&s.t.\ \|a_i\|_1\le 1\end{aligned}\tag{17}$$

in which, $\mathrm{Y}_{2j}=\left[(y_2)_j,\cdots,(y_N)_j\right]^T$. By solving optimization problem (17), the topological structure of a GRN can be recovered.

By setting $\lambda=0.5$, the topological structure for each segment is shown in Figure 1. An objective assessment of the reconstruction accuracy is not feasible due to the limited existing biological knowledge and the absence of a gold standard. However, we can mark some interactions that have been verified in red. More specifically, the interactions 'actn↔mhc', 'actn→up', and 'up↔mhc' have been verified in [26,27]; the interaction 'eve→twi' twi' has been verified in [28]; and the interactions 'actn→msp300', 'actn→prm', 'prm↔sls' and 'sls→up' have been verified in [29]; and the interaction 'actn→sls' has been verified in [30]. These computation results using the developmental data of *Drosophila Melanogaster* show that the suggested two-step strategy may also be helpful in solving actual GRN reconstruction problem.

Apart from the developmental data of *Drosophila Melanogaster*, there exist some other real microarray compendia. Especially, the more recent DREAM5 network inference challenge offer some alternative real microarray compendia, which can be found in the web site at http://wiki.c2b2.columbia.edu/dream/index.php/D5c4 or in [31]. Whereas, the time series data in a single experiment are quite short. Thus, using them to reconstruct a simulative time-varying GRN that is obtained by concatenating gene expression data of different networks is very tricky. However, there exists an especial long time series, i.e., time series data of No. 49 experiment, Network4. This long time series has 48 samples; and from Sample 1 to Sample 11 there is no external interference to the network, while from Sample 12 to Sample 48 there is an uninterrupted external interference (P19) to the network. As mentioned before, biological networks change in response to environmental cues, and the change point is not always known *a priori* in general. Therefore, we use the suggested change point detecting algorithm to check whether there is a topological change in response to P19.

Based on the gold standard of Network4 suggested by the Dream project organizers, we select three sub-networks. In this way, although it is not clear that whether there is a biological significance for these sub-networks, it can be guaranteed that the system matrix A for each sub-network is not a zero matrix. The first one include 7 genes, which are G20, G61, G76, G111, G224, G273, G319. The second one include 8 genes, which are G15, G21, G45, G95, G101, G111, G212, G213. And, the third one include 8 genes, which are G15, G45, G47, G87, G101, G112, G152, G273. By setting $\delta=10^3,\mu=0.98$ in Equations (11)–(15), we can obtain $\hat{\boldsymbol{R}}$ for the dataset. Then, setting $\alpha=0.001$ and using the change point detection procedure in Table 1, we find that each sub-network changes its network topology at the time interval 16 to 17. This result also verifies the general conclusion that biological networks are dynamic in response to environmental changes [6,7], and the rewiring processes may be time-delayed [11].

Finally, some information about the dataset can be found in [31]. More specifically, Network 4 is *S. cerevisiae*; the external interference P19 is phenelzine treatment; and the de-anonymized gene names are listed as follows: G15: YKL043W, G20: YJR147W, G21: YER045C, G45: YMR016C, G47: YNL167C, G61: YLR131C, G76: YJR060W, G87: YHR206W, G95: YNL314W, G101: YGL162W, G111: YOR028C, G112: YER111C, G152: YLR182W, G212: YDL106C, G213: YIL130W, G224: YEL009C, G273: YDR259C, and G319: YPR104C.

Concluding Remarks

In this paper, we consider the time-varying GRN identification problem. The switching auto-regressive model is used to approximate the regulatory model for time-varying GRNs. And, a two-step strategy is proposed to recover the topological structure. In the first step, on the basis of a relation between the Kalman filter and recursive least squares estimation, it is shown that the innovation process is white if and only if the time series expression data are generated by the same sub-regulatory network. Based on this observation, a procedure is developed to detect change points of a time-varying GRN. The observed time series are divided into several segments based on these detection results; and each time series segment belonging to two successive demarcating change points is associated with an individual static regulatory network. Therefore, in the second step, for each time interval, the causal relationships inference problem can resort to conditional network structure identification methods, such as IOTA, and LASSO, etc.

The main difficulty of the time-varying GRN identification problem is that the identification problem includes a classification problem in which each data must be associated to the most suitable sub-network. The more precise the data classification is, the better the identification results are. The proposed two-step strategy efficiently estimates the change point, which results in the decoupling of the change point detection problem and the topology inference problem. Hence, the problem of identifying a time-varying regulatory model is transformed into that of

identifying multiple single static regulatory models, which means that the proposed two-step strategy can ease the difficulty level in recovering a time-varying GRN. Simulation results show that the proposed strategy can detect the change point precisely and recover each individual topology structure effectively. Moreover, computation results with the developmental data of *Drosophila Melanogaster* show that the suggested strategy may also be helpful in solving actual GRN reconstruction problem.

Under our two-step strategy architecture, recovering a static GRN from time series is the most basic problem. However, this problem is not solved completely and efficaciously! Therefore, the most urgent problem is how to utilize gene expression time series data to obtain a static network structure with high accuracy.

Supporting Information

Text S1 Appendix: Proof of Theorem 1.
(PDF)

Author Contributions

Conceived and designed the experiments: JX. Performed the experiments: JX. Analyzed the data: JX. Contributed reagents/materials/analysis tools: JX. Wrote the paper: JX TZ.

References

1. Martin S, Zhang Z, Martino A, Faulon J (2007) Boolean dynamics of genetic regulatory networks inferred from microarray time series data. Bioinformatics 23: 866–874.
2. Ferrazzi F, Sebastiani P, Ramoni M, Bellazzi R (2007) Bayesian approaches to reverse engineer cellular systems: a simulation study on nonlinear gaussian networks. BMC bioinformatics 8: S2.
3. Xiong J, Zhou T (2012) Gene regulatory network inference from multifactorial perturbation data using both regression and correlation analyses. PLOS ONE 7: e43819.
4. Zhou T, Wang Y (2010) Causal relationship inference for a large-scale cellular network. Bioinformatics 26: 2020–2028.
5. Arbeitman M, Furlong E, Imam F, Johnson E, Null B, et al. (2002) Gene expression during the life cycle of drosophila melanogaster. Science 297: 2270–2275.
6. Luscombe N, Babu M, Yu H, Snyder M, Teichmann S, et al. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. Nature 431: 308–312.
7. Ahmed A, Xing E (2009) Recovering time-varying networks of dependencies in social and biological studies. Proceedings of the National Academy of Sciences 106: 11878–11883.
8. Grzegorczyk M, Husmeier D (2011) Non-homogeneous dynamic bayesian networks for continuous data. Machine Learning 83: 355–419.
9. Robinson J, Hartemink A (2010) Learning non-stationary dynamic bayesian networks. The Journal of Machine Learning Research 11: 3647–3680.
10. Lèbre S, Becq J, Devaux F, Stumpf M, Lelandais G (2010) Statistical inference of the time-varying structure of gene-regulation networks. BMC systems biology 4: 130.
11. Dondelinger F, Lèbre S, Husmeier D (2013) Non-homogeneous dynamic bayesian networks with bayesian regularization for inferring gene regulatory networks with gradually time-varying structure. Machine Learning 90: 191–230.
12. Yoshida R, Imoto S, Higuchi T (2005) Estimating time-dependent gene networks from time series microarray data by dynamic linear models with markov switching. In: Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference. IEEE Computer Society, pp. 289–298.
13. Zhang K, Han J, Groesser T, Fontenay G, Parvin B (2012) Inference of causal networks from time-varying transcriptome data via sparse coding. PloS one 7: e42306.
14. von Essen CF (1969) Radiation tolerance of the skin. Acta Oncologica 8: 311–330.
15. Wara WM, Phillips TL, Sheline GE, Schwade JG (1975) Radiation tolerance of the spinal cord. Cancer 35: 1558–1562.
16. Stafford SL, Pollock BE, Leavitt JA, Foote RL, Brown PD, et al. (2003) A study on the radiation tolerance of the optic nerves and chiasm after stereotactic radiosurgery. International journal of radiation oncology, biology, physics 55: 1177.
17. Van der Kogel A (1977) Radiation tolerance of the rat spinal cord: Time-Dose relationships. Radiology 122: 505.
18. Ljung L (1999) System identification. PTR Prentice Hall, Upper Saddle River, NJ.
19. Ljung L, Söderström T (1983) Theory and practice of recursive identification. MIT Press, Cambridge, MA.
20. Hempel S, Koseska A, Kurths J, Nikoloski Z (2011) Inner composition alignment for inferring directed networks from short time series. Physical review letters 107: 54101.
21. Wu F, Liu L, Xia Z (2010) Identification of gene regulatory networks from time course gene expression data. In: Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE. IEEE, pp. 795–798.
22. Paoletti S, Juloski A, Ferrari-Trecate G, Vidal R (2007) Identification of hybrid systems: a tutorial. European Journal of Control 13: 242–260.
23. Prill RJ, Marbach D, Saez-Rodriguez J, Sorger P, Alexopoulos L, et al. (2010) Towards a rigorous assessment of systems biology models: the dream3 challenges. PloS one 5: e9202.
24. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, et al. (2010) Revealing strengths and weaknesses of methods for gene network inference. Proceedings of the National Academy of Sciences 107: 6286–6291.
25. Marbach D, Roy S, Ay F, Meyer PE, Candeias R, et al. (2012) Predictive regulatory models in drosophila melanogaster by integrative inference of transcriptional networks. Genome research 22: 1334–1349.
26. Homyk-Jr T, Emerson-Jr CP (1988) Functional interactions between unlinked muscle genes within haploinsufficient regions of the drosophila genome. Genetics 119: 105.
27. Nongthomba U, Cummins M, Clark S, Vigoreaux J, Sparrow J (2003) Suppression of muscle hypercontraction by mutations in the myosin heavy chain gene of drosophila melanogaster. Genetics 164: 209–222.
28. Parkhurst S, Ish-Horowicz D (1991) wimp, a dominant maternal-effect mutation, reduces transcription of a specific subset of segmentation genes in drosophila. Genes & development 5: 341–357.
29. Formstecher E, Aresta S, Collura V, Hamburger A, Meil A, et al. (2005) Protein interaction mapping: a drosophila case study. Genome research 15: 376–384.
30. Sanchez C, Lachaize C, Janody F, Bellon B, Röder L, et al. (1999) Grasping at molecular interactions and genetic networks in drosophila melanogaster using flynets, an internet database. Nucleic acids research 27: 89–94.
31. Marbach D, Costello J, Küffner R, Vega N, Prill R, et al. (2012) Wisdom of crowds for robust gene network inference. Nature methods 9: 796–804.