An Integrative Analysis of Time-varying Regulatory Networks From High-dimensional Data

Zi Wang *, Yun Guo *, Haijun Gong *†
* Department of Mathematics and Statistics
Saint Louis University, St. Louis, MO 63103, USA
† Research School of Finance, Actuarial Studies and Statistics
Australian National University, Acton, ACT, 2601 Australia
Email: haijun.gong@{anu.edu.au, slu.edu}

***Abstract*—Directed networks have been widely used to describe many biological processes and functions. Understanding the structure of biological networks, especially regulatory networks, could help discover the mechanisms underlying important biological processes and pathogenesis of diseases. Most network inference methods assume the network structure is time-invariant or stationary. However, in some processes, the network structure is non-stationary or time-varying. The stationary network inference methods might not be able to directly used to reconstruct time-varying networks. Some non-stationary network learning methods have been proposed to infer the networks, but, the inferred networks are not regulatory networks which require activation and inhibition information. This work proposes an integrative approach, which combines the changepoint estimation, weighted network learning and searching, and model checking technique, to reconstruct time varying regulatory networks from high-dimensional time series data. We illustrate this approach to study the structure changes of Drosophila's regulatory networks in its life cycle.**

***Index Terms*—Regulatory network, time series data, changepoint estimation, dynamic Bayesian network, model checking.**

I. Introduction

Biological networks, e.g., gene regulatory networks and neural connectivity networks, have been widely used to describe many biological processes and help study the dynamic behaviors of the system. Modern genomic technologies have generated a vast amount of omics data. One of the most challenging tasks for biologists and data scientists is how to correctly reconstruct the networks from the high-dimensional data. Different computational techniques, including the deterministic [1], [2] and probabilistic methods [3]–[6], have been developed to reconstruct networks. Deterministic methods can not handle the noise in the data (though stochastic differential equations (SDE) can simulate the stochastic process by adding a noise term to the ODEs, but they can not handle high-dimensional data); Bayesian network model can not learn the feedback loops which exist in many biological systems; the networks inferred by the graphical LASSO is undirected. So, most directed network inference methods, including our previous work [7], [8], are based on the dynamic Bayesian networks (DBN) model. These network inference methods assume the network structure does not change over time: it is stationary.

Several studies have revealed that the structures of some cellular networks are non-stationary or time-varying, that is, the network undergoes systematic rewiring at different stages or under some circumstances. For example, the neural information flow networks [9] of brains are changing during learning; the structure of Drosophila's regulatory network is evolving from the embryonic, larval, pupal, to adult stages in its life cycle [10]. Moreover, recent studies found the naive/effector T cells would be converted into senescent cells [11] due to the structure changes of genetic network [12] during tumorigenesis. The stationary network inference methods can not be directly used to reconstruct time-varying networks and investigate the structure changes of genetic networks.

Some researchers have attempted to develop new computational methods based on different assumptions for the time-varying networks inference from high-dimensional data generated by modern genomic technology. For example, the dynamic vector autoregressive model [13] and and heterogeneous DBN model [14] assume some parameters are changing over time, but the network structure is still time-invariant. A temporally smoothed L1-regularized logistic regression method [15] can infer un-directed, rather than directed, time-evolving networks. A dynamic linear model with Markov switching were developed [16] to estimate the time-dependent network structure, but this approach assumes the number of stages is known and fixed. Though these methods contributed to our understanding of network structure, no existing methods can correctly reconstruct and efficiently verify time-varying regulatory networks due to some severe drawbacks.

The time-varying network inferred by all existing machine learning methods [13]–[17] are either undirected, or correlation or causality graphs (whose arrow represents statistical dependence, i.e., one gene's expression is statistically dependent on others), not regulatory networks (which need activation and inhibition information). Regulatory networks can provide more information than the correlation and causality graphs. Without verification or validation, the inferred network cannot be used to predict the dynamic behaviors of the cellular system. Another challenge is how to correctly and efficiently verify the inferred time-varying networks. Other researchers'

This work was partially supported by the USA NIH-NIGMS grant 1R15GM129696-01A1 and Australian National University.

work just manually assess the accuracy of the inferred optimal network by comparing with existing database or known models; this naive verificaiton procedure works for small network. But, most biological networks are highly complex, it is not realistic and efficient to manually check multiple time-varying networks, which are composed of hundreds or thousands of nodes and edges. No methods can overcome all these challenges. Our previous work had applied weighted dynamic Bayesian network [8] and symbolic model checking [18]–[21] technique to infer and verify stationary networks from time series microarray data.

The study of time-varying network involves the estimation of changepoints describing the stage transition, network structure learning at different stages, and validation of inferred networks. This work proposes to integrate the changepoint estimation method, weighted network learning and searching algorithm, and model checking technique together to study the time-varying networks from high-dimensional data. The outline of the paper is as follows. First, we will briefly review a changepoint estimation algorithm; then we will discuss a weighted dynamic Bayesian network inference technique which can learn weighted and directed networks at a given stage; furthermore, we will introduce a symbolic model checker and integrate all these technique together for the time-varying network inference and verification. In Section III, we apply this integrative approach to study the Drosophila's regulatory network. Finally, we discuss the advantages and limitations of our technique and future work.

II. Methods

Most learning algorithms assume the network structure to be time-invariant, that is, the data generation process is stationary; and the time-varying networks inferred by all existing methods are not regulatory networks. In this work, the regulatory network describes the signed information flow, which is an analogy of the road traffic of cities. The cities (nodes) are connected by roads (directed edges), the traffic of the road could change (blocked or smooth) due to some traffic control or accidents; that is, some connections (roads or edges) might change over time.

In the biological system, under some circumstances (e.g., genetic mutation, stress or drug treatment, etc.), the network of signal transduction or information flow will be changing with time (connected or disconnected, activation or inhibition). For example, in the normal cell, the oncoprotein/oncogene MDM2 is regulated by the tumor suppressor protein/gene P53 [22]. But, in the cancer cell, due to the mutation of P53, the inhibition connection between P53 and MDM2 might not exist any more during tumorigenesis, that is, regulatory network structure is changing over time.

Fig. 1 illustrates the time-invariant regulatory network inference using the stationary dynamic Bayesian network method to model a simple regulatory network, which is composed of 3 nodes with activation (arrow) and inhibition (circlehead arrow) connections, and the network structure is invariant (always stays in the stage G1), which assumes the first-order Markov

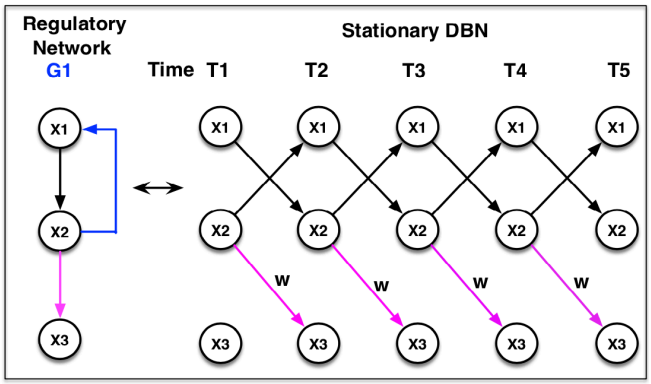


Fig. (1) Illustration of time-invariant regulatory network inference (3 nodes) using the weighted stationary dynamic Bayesian network model based on the first-order Markov property.

property: the state of each node measured at time t+1 is dependent on the states of its parents measured at time t only, so the feedback loop is allowed. The signed weight (w) is also assumed time-invariant in previous studies [7], [8]. The traditional stationary DBN method can not be directly used to reconstruct the time-varying networks.

Reconstruction of time-varying networks from high-dimensional dataset **D** faces several challenging questions. One of the difficult questions is to identify the number and locations of changepoints, which describe when the system makes transitions from one stage to another one. In many high-dimensional observation data, both the number and locations of changepoints are unknown. Different supervised and unsupervised learning methods have been proposed to estimate the changepoints from high-dimensional data. In the supervised learning methods, the changepoint detection problem can be trained as a multi-class classification problem (e.g., logistic regression [23] and hidden Markov model [24]). Most of changepoint detection methods are based on the unsupervised learning, e.g., the likelihood or density-ratio test [25], Bayesian inference approaches [26], [27], product partition models [28], reversible jump MCMC sampling method [29]. Most of them are difficult to handle very high dimensional data, so some dimensionality reduction techniques are applied before the change-points estimation is implemented. Some of these methods have been used to learn the time-varying biological networks [17], [30]. These methods have different advantages and disadvantages in the high-dimensional data analysis.

A. Changepoint Estimation

In this section, we will briefly review a changepoint estimation method called INSPECT (informative sparse projection for estimation of changepoints) proposed in [31], which can be used for the single and multiple changepoints estimation from high-dimensional time-series data.

Let the matrix $X = (X_1, ..., X_n)^T \in R^{n \times p}$ describe the observed time series data, which consist of p features

(genes) measured at n different time points, that is, X_t is a p-dimensional random vector, which is sampled from some unknown distribution, for example $N_p(\mu_t, \sigma^2 I_p)$ if they are normally and independently distributed, where $1 \leq t \leq n$; and the random variable X_{ti} denotes the ith feature/gene whose value is measured at time t. Let's assume there are ν changepoints denoted by $C = (c_1, ..., c_\nu)$, where $1 \leq c_1 < c_2 < ... < c_\nu \leq n-1$, in the high-dimensional time series data. The interval between two adjacent changepoints (c_i, c_{i+1}) corresponds to a segment or stage s_i, totally there are $\nu+1$ stages or segments $S = (s_1, ..., s_{\nu+1})$. In each stage s_i, we assume the mean vectors are piecewise-constants, that is, $\mu_{c_i+1} = ... = \mu_{c_{i+1}} = \mu^{(i)}$ for $0 \leq i \leq \nu$ considering the network structure is stationary in any stage.

The INSPECT method [31] estimates the number and locations of multiple changepoints from high-dimensional time series data by finding the optimal projection directions which are closely aligned with the vector of mean structure changes between two consecutive stationary stages. The changes of mean vectors are defined as $\theta^{(i)} = \mu^{(i)} - \mu^{(i-1)}$ for $i = 1, 2, ..., \nu$ in any stage, which are sparse since we assume very few edges make changes during stage transitions, that is, $||\theta^{(i)}||_0 = \sum_j 1_{\theta_j^{(i)} \neq 0} \leq k$, where $k \in \{1, ..., p\}$. The optimal projection direction can be obtained by finding the k-sparse leading left singular vectors $\hat{v}_k$ of the CUSUM (cumulative sum) transformation [31], [32] matrix $T_{n,p} : R^{n \times p} \rightarrow R^{(n-1) \times p}$ which is defined as

$$[T_{n,p}(X)]_{t,i} = \sqrt{\frac{n}{t(n-t)}} \Big(\frac{t}{n} \sum_{j=1}^{n} X_{j,i} - \sum_{j=1}^{t} X_{j,i} \Big). \quad (1)$$

To find the k-sparse leading left singular vectors $\hat{v}_k$ [31], [33] of the CUSUM is equivalent to solving the convex optimization problem:

$$\hat{v}_k = argmax ||T^T v||_2. \quad (2)$$

More details with proof are in Wang *et al.*'s work [31], which proposed both single and multiple changepoint estimation algorithms based on the ADMM (alternating direction method of multipliers) algorithm by performing sparse singular value decomposition on the CUSUM transformation matrix T given in Eq. 1, and find k-sparse leading left singular vectors given in Eq. 2.

B. Time-varying Directed Network Learning

Fig. 2 illustrates a weighted time-varying DBN model which is composed of 3 nodes: the network structure is changing from one stage (G1) to another one (G2) at a specific timepoint (e.g., T3 in Fig. 2), which is called the 1st transition time or changepoint, from which the network structure starts to change and the system enters a new stage (e.g., senescence, apoptosis). In each stage, the network structure remains unchanged or stationary. The changes of network structure involve some newborn edges (e.g., red one), death or removal of old edges (e.g., blue one), interaction changes (e.g., pink one) of old edges. The time-varying connections are described by the stage-dependent conditional probability dependences between nodes of the network. The signed weight (w and w'), whose sign denotes the activation or inhibition relationship, should also be stage-dependent. Next, we will discuss how to learn the structures of networks and their changes after the changepoints are identified.

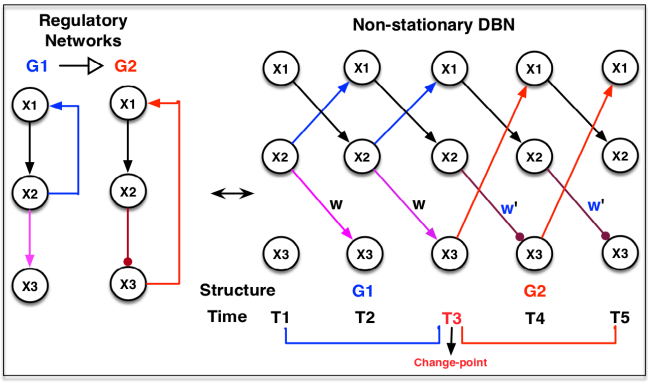


Fig. (2) Illustration of time-varying regulatory networks (2 stages) and weighted non-stationary dynamic Bayesian network model. Different colors of edges represent the structure changes.

Any time-series dataset D can be represented by an $n \times p$ matrix consisting of p feasures measured at n different time points. A stationary DBN model is a pair (G, Θ), where $G = (V, E)$ represents a directed graph, in which V represents a set of random variables or nodes, E is set of edges, and $\Theta = P(X|Pa(X))$ is a set of conditional probability distributions of the nodes $X \in V$ given its parents $Pa(X)$. DBN assumes first-order Markov property: the state of each node measured at time t+1 is dependent on the states of its parents measured at time t only, which was illustrated in the Fig. 1. So, the stationary DBN can be encoded by a joint distribution over V: $P(V) = \prod_{X \in V} P(X|Pa(X))$. In the stationary DBN, learning the structure of an optimal network (G^*, Θ) from a dataset D is equivalent to maximizing the posterior distribution of the network conditional on the data, which is written as the following optimization question:

$$G^* = \underset{G}{\text{argmax}}\, P(G|D) = \underset{G}{\text{argmax}}\, P(G, D)$$
$$= \underset{G}{\text{argmax}}\, P(D|G)P(G),$$

where, $P(G)$ is the prior of the network G, e.g, the minimal description length (MDL) was used in [4]. $P(D|G)$ is the likelihood function, if the parameter vector Θ is continuous, then it can be rewritten as

$$P(D|G) = \int P(D|G, \Theta)P(\Theta|G)d\Theta.$$

Bayesian Dirichlet equivalence (BDe) metric [34] is one of the most widely used methods to learn the stationary DBN network structures. The BDe metric assumes the structure parameters Θ are globally and locally independent and its prior follows Dirichlet distribution with a hyperparameter vector

α given a network G, that is, $\Theta|G \sim Dirichlet(\alpha)$; while the data D conditional on the Θ follows the multinomial distribution, that is $D|\Theta \sim Multinomial(\Theta)$. Later, some heuristic searching algorithms (e.g., simulated annealing) are applied to find the optimal network.

In the time-varying DBN model, as illustrated in Fig. 2, assuming there exist ν changepoints $(c_1, ..., c_\nu)$, that is there are $\nu + 1$ stages, $S = (s_1, ..., s_{\nu+1})$, which are depicted by a sequence of different network structures $G = (G_1, G_2, ..., G_{\nu+1})$. The goal is to find a sequence of optimal networks $G^* = (G_1^*, ..., G_{\nu+1}^*)$ at different stages. Learning the time-varying DBN network structures is equivalent to maximizing the stage (S)-dependent joint posterior distribution

$$\begin{aligned}(G_1, ..., G_{\nu+1})^* &= \underset{(G_1, G_2, ..., G_{\nu+1})}{\operatorname{argmax}} P((G_1, ..., G_{\nu+1})|D, S) \\ &= \underset{G=(G_1, G_2, ..., G_{\nu+1})}{\operatorname{argmax}} P(D|S)P(G|S)\end{aligned}$$

This time-varying DBN optimization problem is non-trivial compared with the stationary DBN optimization problem if the number and locations of changepoints are unknown. Now the first-order Markov property for the time-varying DBN model should be updated as: the state of each node measured at time t+1 should be dependent on the states of its parents measured at time t and its current network structure or stage. So the time-varying conditional probability distribution describing the dependencies can be expressed as $\Theta = P(X|Pa(X), s_t)$. In this preliminary work, we adopt a "divide-and-conquer" strategy. That is, a changepoint estimation algorithm is applied first to identify the stages, then, we can assume the network structures are independent at each stage. This assumption allows us to use the stationary DBN technique to learn the network structures at different stages.

C. Time-varying Weight Estimation

Our aim is to reconstruct regulatory network which contains both activation and inhibition information to describe the relationship between any two nodes. The activation/inhibition on each edge is very important information in the gene regulatory network or signaling pathway, which can provide more information to help us comprehensively understand the mechanism underlying the biological networks. Most non-stationary network inference methods could not learn the activation/inhibition information and its changes over time, so the inferred network is only a time-varying correlation or causality graph.

The quantity w (shown in Fig. 2) describes the signed interaction strength between any two nodes, so the sign of w can be used to describe the activation/inhibition relationship. In previous studies [8], weight was assumed to be time-invariant as shown in Fig. 1. In the non-stationary DBN model, weight is time-varying because the changes of network structure involve the birth or death or change of some edges, which will influence the sign and magnitude of interaction strength.

How to estimate the time-varying signed weight at different stages? A function $G_{ijk}(t, S_t)$, which is an extension of the work in [6], is defined by Eq. 4 and used to measures the probability that, at time t and stage S_t, a node X_i takes a (discrete) value ($m \in \{0, 1, ..., k\}$) no larger than k given its parent nodes $Par(X_i)$ taking a value of j at its current stage. If there is a high probability for the node X_i taking a larger value given its parent's value increases, that is, If w_{ijm} is an increasing function of its parent node's value, then, this interaction will be voted as an activation event at a specific stage S_t by a predefined voting machine [6] based on the values of $G_{ijk}(t, S_t)$; else, it will be voted as an inhibition event.

$$G_{ijk}(t, S_t) = \sum_{m=0}^{k} \omega_{ijm}(t, S_t) \tag{3}$$

$$= \sum_{m=0}^{k} P(X_{ti} = m|Par(X_{ti}) = j, S_t). \tag{4}$$

The signed "weight" will be converted into an integer by a renormalization procedure, which is required by the model checker (discussed later), to describe the time-varying interaction strengths in the regulatory network at different stages. Sometimes, the sign can not be identified if the weight is close to zero. The time-varying weight estimation formula is dependent on the stage, so its estimation is challenging compared with the stationary weight estimation formula. However, if the changepoints are known or estimated in the beginning, then, the above formula is simplied to the stationary weight estimation formula for a given stage if we assume the network structures are independent at different stages. Inference of optimal time-varying regulatory networks need the integration of the changepoints estimation, network structure learning and searching, and weight estimation methods together.

Due to the limited number of experimental replications and time points, almost all the network reconstruction tools face the same inference uncertainty problem, that is, the outcome is very sensitive to the values of some parameters and choice of structure learning and searching algorithms. So, a statistically "optimal" network might not be correct in biology. The proposed integrative approach can generate multiple optimal time-varying networks (high BDe scores) at different stages. Without validation, the inferred networks cannot be used to study the dynamic properties of the system and make predictions. Next we will discuss how to verify which network is correct or most consistent with the experiment.

D. Time-varying Network Verification

Most networks inference methods validate the inferred networks through manually comparing with the online database or known models. To overcome the drawback of naive verification procedure, we will introduce a powerful model checking technique, a popular hardware system verification technique, to intelligently verify the inferred time-varying regulatory networks.

Model Checking is a Turing Award winning technique [35], which is written as $M \models \psi$: a given model M is checked

whether or not it satisfies some temporal logic formula ψ desired by the system. Many model checking tools (e.g., BLAST, FDR2, Prism, SPIN, Java Pathfinder, NuSMV ...) have been developed and successfully applied to verify the design of hardware and software systems in the past thirty years, though all these model checking tools face a same unsolved issue: state explosion problem. This technique was recently applied to the biomedical and cyber-physical research [36]–[39].

Our recent studies [18], [20], [21], [40]–[42] have proposed different model checking technique and apply stand-alone to formally analyze several models of signaling pathways. In particular, SMV model checker was integrated with the weighted stationary dynamic Bayesian network method together to find the best candidate from the inferred optimal regulatory networks in our recent work [8]. Different model checking technique, including the statistical model checker, symbolic model checking and probabilistic model checking PRISM, have been discussed in our recent work [8], [18], [20], [21], [40]–[42], for completeness, we will review some fundamentals of symbolic model checking to help those who are not familiar with SMV model checker to understand this technique.

In the formal language, a model is described by a Kripke structure [35] $M = (S, s_0, R, L)$, where, $s_0 \in S$ represents the initial state, R and L represent the states transition relation and a labeling function respectively. During model checking, the proposed model M is converted into a state transition system, then, start from the initial state s_0, model checker will automatically and exhaustively search the state transition system to verify or falsify the desired property ψ which is expressed as a linear temporal logic (LTL) or computation tree logic (CTL) formula. SMV [43] is one popular symbolic model checker that is encoded by ordered binary decision diagram (OBDD) [44], and it has been applied to verify inferred stationary DBN models [8].

TABLE (I) Some important CTL operators.

Operators	Meanings in English
A	For All paths
E	For some paths (there Exist some paths)
AF ϕ	ϕ will necessarily become true
EF ϕ	ϕ will possibly become true
AG ϕ	ϕ is an invariant (globally true on all paths)
EG ϕ	ϕ is a potential invariant (always true on some path)

During model checking, one key step is how to express the temporal logic formula. Linear temporal logic (LTL) formula is made by Boolean logic connectives (e.g., ! | & $\rightarrow$), temporal operators (e.g., $\mathbf{X}, \mathbf{F}, \mathbf{G}, \mathbf{U}$) describing some property on a path. There are also two path quantifiers (**A**, **E**) describing the branching structure in the computation tree. The CTL formulas are constructed using both path quantifiers and LTL operators, while the path quantifiers must be put in front of LTL operators. Table (I) lists some important CTL operators and their meanings. More details of the temporal logic operators, formula, syntax and semantics of CTL logic are discussed in [35]. SMV can exhaustively search the state transition system and check whether $M \models \psi$ is true or not. If the the model M satisfies the property ψ, SMV will output "True", else, "False" with a counterexample will be given. More details of the temporal logic operators, formula, syntax and semantics of CTL, and the original SMV algorithm are discussed in [35].

E. Integrative Analysis of Time-varying Network

The proposed time-varying regulatory networks reconstruction from high-dimensional time series data involves three major steps: changepoints estimation, optimal network structure learning, searching, and signed weight estimation, and final networks verification, which have been discussed in the above sections. Now, we will combine these methods together to develop an integrative approach to study the time-varying regularoty networks.

The Algorithm 1 shows the integrative procedure for the time-varying regulatory networks reconstruction. In the first step, INSPECT method [45] is applied to estimate the number and locations of changepoints from high-dimensional time series data. Then, the data are divided into different segments or stages according to the changepoints estimated in the first step. We assume the networks structure at different stages are independent, then the time-varying entwork inference problem is reduced to the inference of several stationary networks at different stages. In the 2nd step, the weighted DBN method will be implemented to infer and generate any desired number of high-BDe-scoring regulatory networks and estimate the signed integer weights for all edges at different stages. Finally, any inferred optimal network will be encoded by SMV for model verification. The generation of SMV code follows the following requirement: the SMV program starts with "MODULE MAIN", and the keywords "VAR" and "init" are used to define and initialize variables. The state transition for any node is updated by a state transfer function which is dependent on the integer weights and corresponding parental nodes' values. After the network is automatically defined and initialized, the next step is to encode some desired CTL formulas using the keyword "SPEC". The model checker SMV will intelligently check which one best describes the biological system by checking some desired properties abstracted from experiment or known database.

III. Application

In this section, we will apply the proposed integrative approach to analyze time series microarray data. Arbeitman *et al.*'s work [10] has measured and analyzed the RNA expression levels of 4028 genes of Drosophila melanogaster during its life cycle from the embryonic, larval, and pupal periods, to the first 30 days of adulthood, which includes 67 time points observations during 4 stages. The collected data have been analyzed by many researchers to identify genetic signatures and networks that are specific to the development of different tissues and organs. Biologists take the Drosophila as a good model to study muscle development because it has

Algorithm 1: Integrative Approach for Time-Varying Network Inference and Verification

Part 1: INSPECT for Changepoints Estimation

Input 1: High-dimensional time-series data D;
Parameters (regularization, thresholding...)

1) Perform CUSUM transformation T;
2) Find k-sparse leading left singular vector of CUSUM using ADMM algorithm;
3) Locate changepoints by wild Binary segmentation.

Output 1: Number and locations of changepoints

Part 2: Network Structure Learning and Searching

Input 2 : High-dimensional time-series data D;
Number and Locations of Changepoints

for *data in each stage* S_t **do**
  for *each network* G **do**
    Optimal network structure learning;
      Dirichlet prior distribution for $\Theta|G$;
      Calculate Bayesian Dirichlet equivalence (BDe) scores;
    Network searching;
      Sort network by BDe scores;
      Search network by simulated annealing;
    Integer signed weight estimation;
      for *each edge* $(X_i, Par(X_i))$ **do**
        Compute $\omega_{ijk}(t, S_t)$ and $G_{ijk}(t, S_t)$;
        Calculate integer weights;
      end
  end
end

Output 2: Optimal time-varying regulatory networks;
Signed integer time-varying weights.

Part 3: SMV Code Generation and Verification

Input 3 : Inferred regulatory networks M;
Signed integer weights;
Abstracted temporal logic formula ψ

for *each network, SMV code generator will* **do**
  MODULE MAIN: Starter of SMV code;
  VAR: Declare variables;
  ASSIGN & init: Initialize variables and assign values;
  next: Update state by weighted transfer functions;
end

Output 3: SMV Code for each network

Input 4 : SMV Code;
Experimental results

1) Design desired CTL formula from experiment;
2) **SPEC**: Specify the CTL formula ψ;
3) Attach CTL formula to SMV code;
4) Run SMV Model Checker to verify $M \models \psi$.

Output 4: Network is True or False

fewer muscle types and shorter life span compared with other species.

Recently, some work [17], [26], [30] proposed different network learning algorithms to investigate how the network structures are changing during the muscle development, and these studies tried to reconstruct networks composed of eleven wing muscle development-related genes that were identified by Zhao and Dondelinger *et al.*'s work [46], [47] to study the network evolution. As we mentioned in the previous sections, these networks are either undirected or causality graph. Next, we will apply the proposed integrative approach to study how the regulatory network structures regulating the Drosophila's muscle development are changing in its life cycle.

Since there are only 67 observations (time points) with 4028 genes ($p >> n$), during the changepoints estimation using the INSPECT method, we consider two cases: one is a small dataset which contains only 11 genes that were previously identified to be involved in the wing muscle development; another case is, each time, we randomly sample 500 genes from the pool (4028) and repeat 500 times and count the frequencies for each time-point to be identified as a changepoint. Our studies found the number of changepoints is senstive to the value of threhold, a parameter for testing whether an identified changepoint is a true changepoint in INSPECT. Fig. 3(A) shows 3 possible changepoints around time point (18, 40, 52) with a threshold value 40, and the Fig. 3(B) is a histogram of estimated changepoint locations through randomly sampling 500 genes (repeated 500 times), three most likely changepoints are (23, 39, 56). Arbeitman *et al.*'s work [10] indicates that, it takes around 20 hours for more than 80% of developmentally modulated genes to turn on (the 28th time-point in the data) before the end of embryogenesis (24 hours). Considering the inference uncertainty, the changepoints identified by INSPECT are very close to the real changepoints (31, 41, 59) in the Drosophila data.

Then, we applied a weighted dynamic Bayesian network inference method to reconstruct the time-varying regulatory networks of Drosphila's muscle development. Since the chagepoints are already known before [10], we divide the data into 4 different stages for individual analysis and assume the network structure in any stage is stationary and independent of that in other stages. The goal is to find out how the structures of regulatory networks are changing over time, and what structure changes and how these changes influence the cellular development and functions. Our method can infer multiple optimal regulatory networks (high BDe scores) and estimate the interaction strength on each edge. Fig. 4 illustrates four selected networks with high BDe scores during the Drosophila's life cycle from the embryonic, larval, pupal to adulthood. In these networks, the solid lines with arrows represent an activation event, while the circle-head arrows represent inhibition processes. The integers on the directed edges represent the signed weights or interaction strength, which will be used for network verification in the next step. This method is very sensitive to the choise of data discretization, searching algorithm and some parameters.

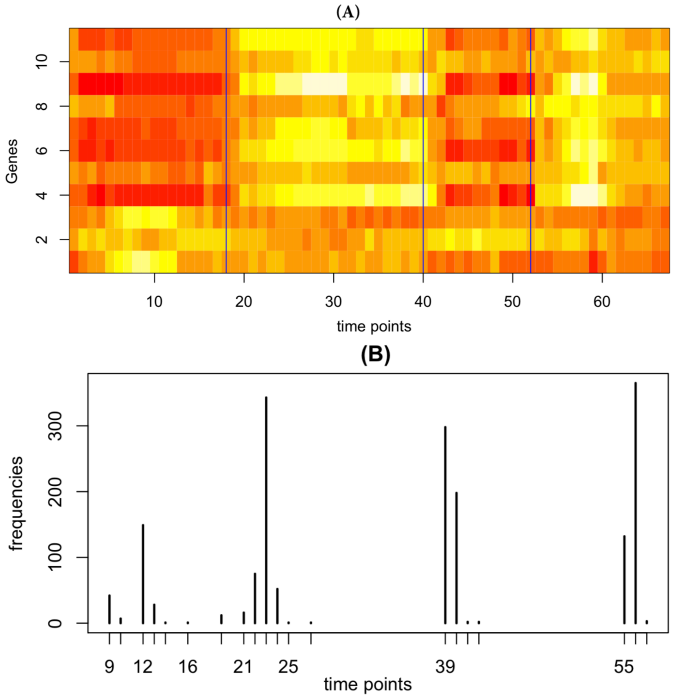


Fig. (3) (A) Estimation of changepoints from 11 genes, where the blue verical lines represent the changepoint lines; (B) Histogram of estimated changepoint locations using 500 genes randomly sampled from 4028 genes each time.

Comparing these inferred optimal regulatory networks at four different stages, it is clearly shown that in the Drosophila's life cycle, the networks regulating the wing muscular development are non-stationary, and it undergoes systematic rewiring. For example, previous experimental study [48] has found that, msp300 plays an important role in actin-dependent nuclear anchorage during cytoplasmic transport. Our results revealed that, the msp300 is a hub gene, which is highly connected with other genes. It can activate several genes including the mlc1, up, eve and myo61f in the enbryoic stage to promote the cellular development; however, in the later stages, its activities will be inhibited by mlc1 and other genes, in particular after the Drosophila enters the pupa and adult stage. Although these inferred network are "optimal" in statistics, according to our studies, they are very sensitive to the changepoints, data discretization choice and values of some parameters, due to the small number of measurements at different stages.

The weighted DBN method can automatically generate SMV code for each inferred network candidate for future model checking. SMV model checker is then implemented to formally verify the inferred networks through checking some temporal logic formulas abstracted from experiment or known database. Using the model checker, it is easy to design and verify different complicated temporal properties associated with some events that can not be expressed or analyzed by traditional simulation methods. The network that satisfies all properties can be used for further analysis, or be re-checked by more future experiments, then, a "statistically optimal" network will be verified to be a biologically-consistent network in the model checking step. The computer code and data are available at Gong's homepage: http://stat.slu.edu/∼gong/Research/BigData.zip.

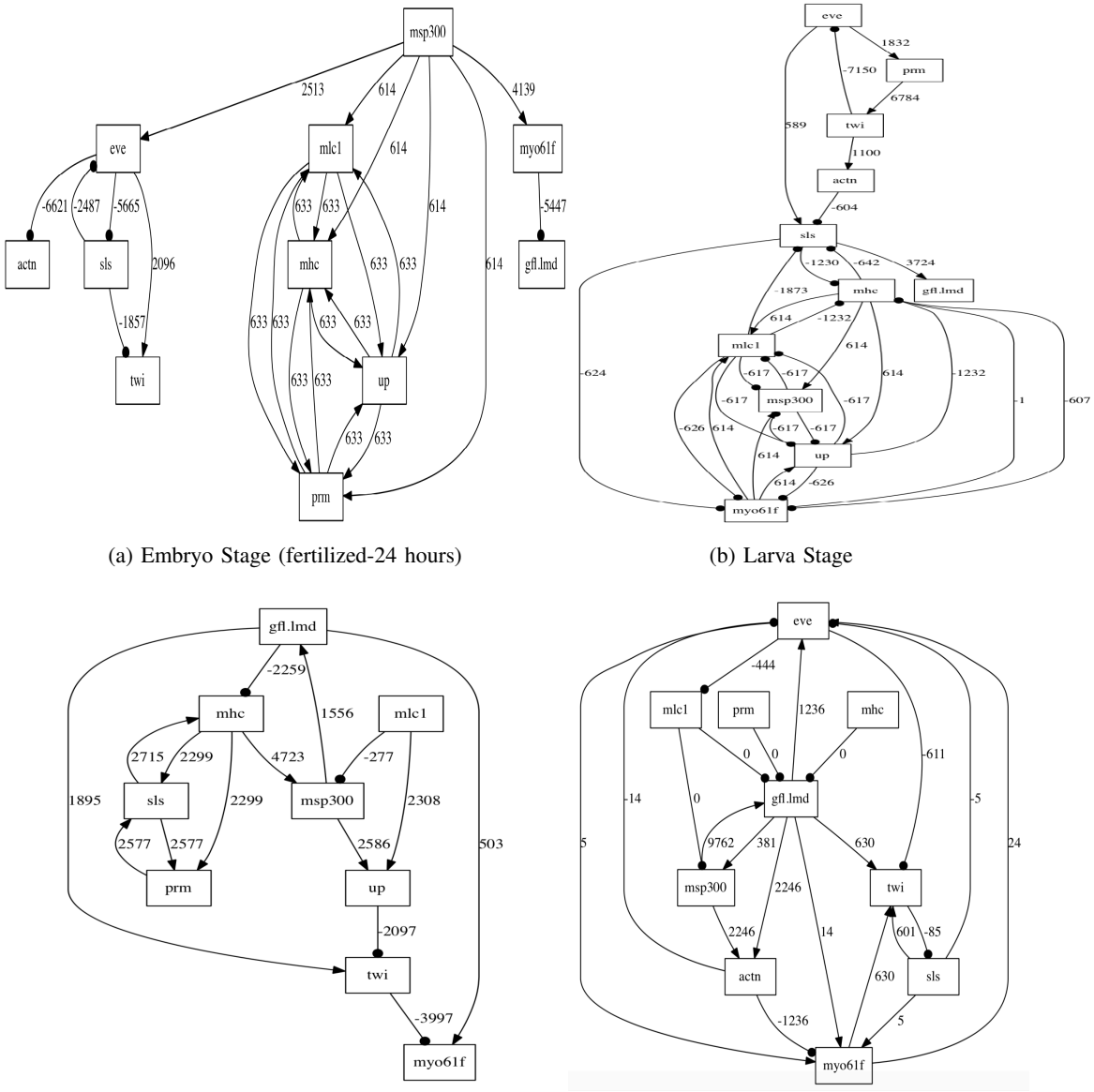
For illustration of network verification, and due to page limits, we only designed and summarized some putative CTL formulas in Table (II) to show how to verify the inferred time-varying networks at different stages. In the SMV code, all the variables can only take discrete values. In this work, we assume that each gene or variable can take three possible values: $-1, 0, 1$, which denote down-regulated, normal and up-regulated, and the initial state is randomly set to be either 0 or -1.

TABLE (II) Putative CTL formulas for the inferred time-varying networks verification

Property	CTL Formula
P1	A(msp300 = 1 $\rightarrow$ AX(mlc1 = 1))
P2	AG(msp300 = 1 $\rightarrow$ AF(twi = -1))
P3	AG(msp300 = 1 $\rightarrow$ AF(twi = +1))
P4	AG(mhc=1 $\rightarrow$ AF(prm = 1 & mlc1 = 1 & up = 1))
P5	EG((mlc1 = 1 $\mid$ msp300 = 1) $\rightarrow$ EF(twi $\leq$ 0 & up $\geq$ 0))
P6	AG((mlc1 = 1 $\mid$ msp300 = 1) $\rightarrow$ AF(twi $\leq$ 0 & up $\geq$ 0))
P7	AG((mhc = 1 $\rightarrow$ msp300 = 1) $\rightarrow$ EF(msp300 = 1 & mhc $\leq$ 0))

In the network verification, one question is, for a given gene, what's its downstream genes? This is easy to manually check in a small network, not easy for the complicated networks, or if there are many downstream genes. Formula P1 checks whether the activation of msp300 can activae mlc1 immediately using the operators **AX**. This formula can automatically and intelligently identify more downstream genes in the complex network. The formulas P2-P3 indicate that, there exists a path on which msp300's activation will finally inhibit (P2) or activate (P3) twi's expression. P4 checks whether mhc's activation will finally activate the genes prm, mlc and up's activity. Formula 5 or 6 shows, for some (P5) or all (P6) paths, it is globally true that either mlc1 or msp300's overexpression will finally inhibit twi's activity but promote up's expression. The formula P5 is weaker than P6 in verification. Property 7 describes a negative feedback loop between mhc and msp300. Table (III) summarizes the verification results of these CTL formulas at different stages. Our results show some properties are true in all stages, but some are false due to the network structure changes in the cell cycle progression.

SMV model checker can be applied to identify some key components (e.g., biomarkers) or processes that give rise to a specific event or reach a specified state (e.g., apoptosis or senescence) from the biological network by checking some pre-designed temporal logic formula, which are difficult to be expressed or analyzed by traditional computational technique. For example, the below property P8, which describes a sequence of reaction events, can be easily and concisely expressed using the CTL formula and verified or falsified by



(a) Embryo Stage (fertilized-24 hours)

(b) Larva Stage

(c) Pupa Stage

(d) Adult Stage

Fig. (4) Structure changes of the regulatory networks at different stages during the muscle development.

TABLE (III) Verification results of CTL formulas at 4 stages

Property	Embryo	Larval stage	Pupal stage	Adulthood
P1	True	True	True	True
P2	True	True	True	True
P3	True	False	False	True
P4	True	False	False	True
P5	True	False	False	False
P6	True	False	False	False
P7	True	False	False	True

the SMV model checker, but it is not easy for researchers to apply the traditional simulation methods (e.g., stochastic simulation or ODE) to model or analyze this kind of phenomenon. So, the model checker provides a new and convenient way to investigate the temporal behaviors of complex biological systems.

P8: AG((mlc1 = 1 $\rightarrow$ AF(msp300 $\leq$ 0)) & (twi $\leq$ 0 $\rightarrow$ AF(myo61f $\leq$ 0)) & ((actn$\geq$ 0| sls $\geq$ 0) $\rightarrow$ AF(eve $\leq$ 0))).

IV. Discussion

In this paper, we showed how to apply an integrative approach, which combines the changepoint estimation, weighted dynamic Bayesian network inference and searching algorithm, and symbolic model checking technique, to study the time-varying regulatory networks from high-dimensional data. Most network learning algorithms assume the network structure is stationary, and some non-stationary network inference methods could only reconstruct time-varying undirected graph, or correlation/causality graph, instead of regulatory network. The proposed approach first applied the INSPECT algorithm [31]

to identify the number and locations of changepoints from high-dimensional time series data in order to divide the data into different segments or stages. When the changepoints are known, we can apply the weighted stationary dynamic Bayesian network technique to learn the optimal structures of directed networks, and also estimate the interaction strength using an integer weight estimation method, whose sign describes the activation or inhibition information of each edge, which is very important information in the regulatory network. We applied this technique to study the structure changes of Drosophila's regulatory networks in its life cycle, which has been analyzed in many researchers' work. Compared with other researchers' work, our approach could provide more information about the network structure, including the birth or death of edges, interaction changes (activation or inhibition) of edges. Another novelty is, a SMV model checker is integrated to formally verify the inferred time-varying networks by checking some putative temporal logic formulas, this technique has some advantages compared with traditional simulation or modeling methods, it could design many complicated properties, which can not be analyzed by traditional methods, but easy for the model checker to verify or falsify, these results could also be experimentally validated in the wet lab.

The inference of time-varying networks from high-dimensional omics data is still a challenging task in systems biology. This work's major aim is to develop an integrative approach and procedure, which combines different existing methods, for the time-varying regulatory networks reconstruction. Though different changepoints estimation methods have been developed, it is still difficult to handle the high-dimensional genetic data, including the INSPECT algorithm used in this work, and its changepoints estimation results are very sensitive to some parameters. For example, the number and locations of changepoints estimated by INSPECT is dependent on a predefined threshold value. Another problem in this proposed approach is, we assume the network structures at different stages are independent, so we can apply the stationary DBN method to learn the structures of regulatory network and estimate the weight changes on each edge at different stages. This assumption can simplify our calculations in this work, but it is not always true. Our studies found that the network structure might undergo major changes under this assumption, but this is not consistent with the experiment which observes a small number of changes. Though our technique can automatically generate SMV code, however, the abstraction of CTL formula is not automatic translation from experiment or database. Our future work will propose new methods to resolve these problems; develop an integrative platform which can correctly estimate the changepoints, infer optimal time-varying regulatory networks from high-dimensional omics data, automatically abstract CTL formulas from database or experiment, intelligently verify or falsify the inferred time-varying networks, and investigate the temporal and dynamic behaviors of the biological systems.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHOR'S CONTRIBUTIONS

HG proposed the study, ZW, YG, HG wrote the code and analyzed the results, ZW and HG wrote the manuscript. All authors approved the final manuscript.

ACKNOWLEDGMENT

This work was partially supported by the NIH-NIGMS grant 1R15GM129696-01A1 (HG) and Australian National University during Dr. Gong's sabbatial leave.

REFERENCES

[1] T. Chen, H. He, and G. Church, "Moeling gene expression with differential equations," in *Pacific symposium on biocomputing*, 1999, pp. 29–40.

[2] T. Akutsu, S. Miyano, and S. Kuhara, "Inferring qualitative relations in genetic networks and metabolic pathways," *Bioinformatics*, vol. 16, pp. 727–734, 2000.

[3] S. Kim, S. Imoto, and S. Miyano, "Inferring gene networks from time series microarray data using dynamic bayesian networks," *Briefings in Bioinformatics*, vol. 4, pp. 228–235, 2003.

[4] N. Friedman, K. Murphy, and S. Russell, "Learning the structure of dynamic probabilistic networks," *Prpceedings of the 14th conference on the uncertainty in artificial intelligence*, 1998.

[5] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using bn to analyze expression data," *J. Comp. Biol.*, vol. 7, pp. 601–620, 2000.

[6] J. Yu, V. Smith, P. Wang, A. Hartemink, and E. Jarvis, "Advances to bayesian network inference for generating causal networks from observational biological data," *Bioinformatics*, vol. 20, pp. 3594–3603, 2004.

[7] H. Gong, J. Klinger, K. Damazyn, X. Li, and S. Huang, "A novel procedure for statistical inference and verification of gene regulatory subnetwork," *BMC Bioinformatics*, vol. V16:S7, 2015.

[8] Y. Ma, K. Damazyn, J. Klinger, and H. Gong, "Inference and verification of probabilistic graphical models from high-dimensional data," *Lecture Notes in Bioinformatics*, vol. 9162, 2015.

[9] B. Horwitz, "The elusive concept of brain connectivity," *NeuroImage*, vol. 19, pp. 466–470, 2003.

[10] M. Arbeitman, E. Furlong, F. Imam, E. J. E *et al.*, "Gene expression during the life cycle of drosophila melanogaster," *Science*, vol. 297, pp. 2270–5, 2002.

[11] T. Doering *et al.*, "Network analysis reveals centrally connected genes and pathways involved in cd8+ t cell exhaustion versus memory," *Immunity*, vol. 37, 2012.

[12] N. Luscombe *et al.*, "Genomic analysis of regulatory network dynamics reveals large topological changes," *Nature*, vol. 431, pp. 308–312, 2004.

[13] A. Fujita, J. Sato *et al.*, "Time varying modeling of gene expression regulatory networks using the wavelet dynamic vector autoregressive method," *Bioinformatics*, vol. 23(13), pp. 1623–1630, 2007.

[14] M. Grzegorczyk and D. Husmeier, "Nonstationary continuous dynamic bayesian networks." *Advances in Neural Information Processing Systems (NIPS)*, vol. 22, pp. 682–690, 2009.

[15] A. Ahmed and E. Xing, "Recovering time varying networks of dependencies in social and biological studies." *Proceedings of the National Academy of Sciences*, vol. 106, p. 11878?11883, 2009.

[16] R. Yoshida, S. Imoto, and T. Higuchi, "Estimating time-dependent gene networks form time series microarray data by dynamic linear models with markov switching." *CSB05, IEEE CSBC*, 2005.

[17] J. Robinson and R. Hartemink, "Learning non-stationary dynamic bayesian networks," *Journal of Machine Learning Research*, vol. 11, p. 3647?3680, 2010.

[18] H. Gong, P. Zuliani, A. Komuravelli, J. R. Faeder, and E. M. Clarke, "Analysis and verification of the HMGB1 signaling pathway," *BMC Bioinformatics*, vol. 11, no. 7, 2010.

[19] H. Gong and L. Feng, "Probabilistic verification of er stress-induced signaling pathways," *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine*, 2014.

[20] H. Gong, P. Zuliani, A. Komuravelli, J. Faeder, and E. Clarke, "Computational modeling and verification of signaling pathways in cancer," *Proceedings of Algebraic and Numeric Biology, LNCS*, vol. 6479, 2012.

[21] H. Gong and L. Feng, "Computational analysis of the roles of er-golgi network in the cell cycle," *BMC Systems Biology*, vol. 8, p. S4, 2014.

[22] B. Vogelstein, D. Lane, and A. J. Levine, "Surfing the p53 network," *Nature*, vol. 408, pp. 307–310, 2000.

[23] K. Feuz, D. Cook, C. Rosasco, K. Robertson, and M. Schmitter-Edgecombe, "Automated detection of activity transitions for prompting," *IEEE Trans Human-Machine Syst*, vol. 45(5), pp. 1–11, 2014.

[24] S. Reddy, M. Fun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava, "Using mobile phones to determine transportation modes," *ACM Trans Sens Networks*, vol. 6(2), pp. 1–27, 2010.

[25] Y. Kawahara and M. Sugiyama, "Sequential change-point detection based on direct density-ratio estimation," *SIAM International Conference on Data Mining*, pp. 389–400, 2009.

[26] L. Schwaller and S. Robin, "Exact bayesian inference for off-line change-point detection in tree-structured graphical models," *Statistics and Computing*, vol. 27(5), 2016.

[27] P. Fearnhead, "Exact and efficient bayesian inference for multiple change point problems," *Statistics and Computing*, vol. 16(2), pp. 203–213, 2006.

[28] D. Barry and J. Hartigan, "Product partition models for change point problems," *The Annals of Statistics*, vol. 20(1), pp. 260–279, 1992.

[29] P. Green, "Reversible jump markov chain monte carlo computation and bayesian model determination," *Biometrika*, vol. 82, pp. 711–732, 1995.

[30] S. Lebre, J. Becq, F. Devaux, M. Stumpf, and G. Lelandais, "Statistical inference of the time-varying structure of gene-regulation networks," *BMC Systems Biology*, vol. 4:130, 2010.

[31] T. Wang and R. Samworth, "High dimensional change point estimation via sparse projection," *Statistical Methodology*, 2017.

[32] D. Darling and P. Erdos, "A limit theorem for the maximum of normalized sums of independent random variables," *Duke Math. J.*, vol. 23, pp. 143–155, 1956.

[33] Y. You, T. Wang, and R. Samworth, "A useful variant of the davis-kahan theorem for statisticians," *Biometrika*, vol. 102, pp. 315–323, 2015.

[34] D. Heckerman, D. Geiger, and D. Chickering, "Learning bayesian networks: The combination of knowledge and statistical data." *Machine Learning*, vol. 20(3), 1995.

[35] E. M. Clarke, O. Grumberg, and D. A. Peled, *Model Checking*. MIT Press, 1999.

[36] S. K. Jha, E. M. Clarke, C. J. Langmead, A. Legay, A. Platzer, and P. Zuliani, "A bayesian approach to model checking biological system," in *CMSB*, ser. LNCS, vol. 5688, 2009, pp. 218–234.

[37] M. Ceccarelli, L. Cerulo, and A. Santone, "De novo reconstruction of gene regulatory networks from time series data, an approach based on formal methods," *Methods*, vol. 69(3), pp. 298–305, 2014.

[38] E. Bartocci, L. Bortolussi, and G. Sanguinetti, "Learning temporal logical properties discriminating ecg models of cardiac arrhytmias," *arXiv:1312.7523*, 2013.

[39] O. Parvu and D. Gilbert, "A novel method to verify multilevel computational models of biological systems using multiscale spatio-temporal meta model checking," *PLOS One*, 2016.

[40] H. Gong, Q. Wang, P. Zuliani, M. T. Lotze, J. R. Faeder, and E. M. Clarke, "Symbolic model checking of the signaling pathway in pancreatic cancer," *Proceedings of the International Conference on Bioinformatics and Computational Biology (BICoB)*, 2011.

[41] H. Gong, P. Zuliani, and E. Clarke, "Model checking of a diabetes-cancer model," *3rd International Symposium on Computational Models for Life Sciences*, 2011.

[42] H. Gong, Q. Wang, P. Zuliani, and E. Clarke, "Formal analysis for logical models of pancreatic cancer," *50th IEEE Conference on Decision and Control and European Control Conference*, 2011.

[43] K. L. McMillan, *PhD thesis: Symbolic model checking - an approach to the state explosion problem*. Carnegie Mellon University, 1992.

[44] R. Bryant, "Graph-based algorithms for boolean function manipulation," *IEEE Tran. on Computers*, vol. 35, no. 8, pp. 677–691, 1986.

[45] T. Wang and R. Samworth, "Inspect package: High-dimensional change-point estimation via sparse projection," *CRAN*, 2016.

[46] W. Zhao, E. Serpedin, and E. Dougherty, "Inferring gene regulatory networks from time series data using the minimum description length principle," *Bioinformatics*, vol. 22, 2006.

[47] F. Dondelinger, S. Lebre, and D. Husmeier, "Non-homogeneous dynamic bayesian networks with bayesian regularization for inferring gene regulatory networks with gradually time-varying structure," *Machine Learning*, vol. 90, 2013.

[48] J. You, D. Starr, X. Wu, S. Parkhurst, Y. Zhuang, T. Xu, R. Xu, and M. Han, "The kash domain protein msp-300 plays an essential role in nuclear anchoring during drosophila oogenesis." *Deve Biol*, vol. 289(2), pp. 336–45, 2006.