Automatica 50 (2014) 2765–2776

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Structure identification for gene regulatory networks via linearization and robust state estimation☆

Jie Xiong [a,1], Tong Zhou [b]

[a] Department of Automation, Tsinghua University, Beijing, China

[b] Department of Automation and Tsinghua National Laboratory for Information Science and Technology(TNList), Tsinghua University, Beijing, China

ARTICLE INFO

Article history:
Received 18 May 2013
Received in revised form
22 May 2014
Accepted 5 June 2014
Available online 2 October 2014

Keywords:
Causal relationships
Extended Kalman filter
Gene regulatory networks
Robust state estimator
Modelling error

ABSTRACT

Inferring causal relationships among cellular components is one of the fundamental problems in understanding biological behaviours. The well known extended Kalman filter (EKF) has been proved to be a useful tool in simultaneously estimating both structure and actual gene expression levels of a gene regulatory network (GRN). First-order approximations, however, unavoidably result in modelling errors, but the EKF based method does not take either unmodelled dynamics or parametric uncertainties into account, which makes its estimation performances not very satisfactory. To overcome these problems, a sensitivity penalization based robust state estimator is adopted in this paper for revealing the structure of a GRN. Based on the specific structure of the estimation problem, it has been proved that under some weak conditions, both the EKF based method and the method suggested in this paper provide a consistent estimate, but the suggested method has a faster convergence speed. Compared with both the EKF and the unscented Kalman filter (UKF) based methods, simulation results and real data based estimations consistently show that both convergence speed and parametric estimation accuracy can be appreciably improved. These lead to significant reductions in both false positive errors and false negative errors, and may imply helpfulness of the suggested method in better understanding the structure and dynamics of actual GRNs.

1. Introduction

Most functions in a cell are a result of mutual regulation effects among a large number of molecular components underlying a biochemical reaction network. One fundamental problem in systems biology is to describe the causal relationships among these components. This could help to more deeply understand cell functions, gain further insights into the regulation processes, and find new target genes of complex diseases, etc. With the development of high-throughput technologies, such as DNA microarrays, possibilities become significantly higher for revealing the structure of a gene regulatory network (GRN) and developing its mathematical model, as thousands of gene expression data become available.

Several methods for modelling GRNs have been reported in literature. The simplest model of GRNs is the Boolean network model (Akutsu, Miyano, & Kuhara, 1999; Kauffman, 1993), for which some modelling techniques for inferring interactions among genes have already been successfully developed. Earlier methods typically employed correlation or partial correlation coefficients between expression patterns of all gene pairs to infer "coexpression networks" (Eisen, Spellman, Brown, & Botstein, 1998). Due to the nonlinear nature of GRNs, these coefficients usually fail to capture more complicated statistical dependencies between expression patterns. To overcome these difficulties, a mutual information (MI) based method has been proposed, which computes MIs between all gene pairs and obtains a "relevance network" through selecting gene pairs whose MI is larger than a given threshold (Butte & Kohane, 2000). In addition, various refinements have been proposed to discriminate between direct and indirect interactions in relevance networks, such as the CLR algorithm (Faith et al., 2007), the ARACNE algorithm (Margolin et al., 2006), and

☆ This work was financially supported in part by the 973 Program under Grants 2012CB316504 and 2009CB320602, and by the National Natural Science Foundation of China under Grants 61174122, 61021063, and 51361135705, and by the Specialized Research Fund for the Doctoral Program of Higher Education, P.R.C., under Grant 20110002110045. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Er-Wei Bai under the direction of Editor Torsten Söderström.

E-mail addresses: xiongj08@mails.tsinghua.edu.cn (J. Xiong), tzhou@mail.tsinghua.edu.cn (T. Zhou).

[1] Tel.: +86 10 6279 7430; fax: +86 10 6278 6911.

http://dx.doi.org/10.1016/j.automatica.2014.08.003
0005-1098/

the MRNET algorithm (Meyer, Kontos, Lafitte, & Bontempi, 2007). Moreover, to speed up estimations, some regression analysis based methods have also been developed to GRN identification (Irrthum, Wehenkel, & Geurts, 2010; Xiong & Zhou, 2012). All of these algorithms, however, can model only static relations. A more precise and insightful construction method is needed, which can effectively incorporate random effects caused by perturbations and temporal evolutions of gene interactions. To facilitate information extraction from time series expression profiles, various dynamical models have been developed, such as dynamic Boolean networks (Martin, Zhang, Martino, & Faulon, 2007), neural networks (Tian & Burrage, 2003), and Bayesian networks (Friedman, Linial, Nachman, & Pe'er, 2000; Liu, Sung, & Mittal, 2006), etc.

Among the statistical techniques currently adopted in modelling GRNs, Bayesian inferences have received the most widespread attention (Kim, Imoto, & Miyano, 2003; Murphy et al., 1999; Perrin et al., 2003). Under the dynamic Bayesian regime, the model of GRNs is extensively considered as a state-space model, which consists of gene expression measurement equations and gene regulation equations (Bansal, Della Gatta, & Di Bernardo, 2006; Perrin et al., 2003). In this state-space model, gene expression values are assumed to depend not only on the current cellular states but also on external inputs or disturbances, which reflects the nature of a dynamic network. In the early works, it is generally assumed that gene regulations can be described by linear differential/difference equations, and the well known Kalman filter is used to recover the structure of a GRN (Perrin et al., 2003). However, due to the inherent nonlinear nature of GRNs, there exist some restrictions when a linear model is applied to describe gene behaviours (Qian, Wang, & Dougherty, 2008). In short, linear approximation is valid only when a GRN has slow dynamics around its steady-state. In order to capture complex gene interactions more efficiently, it is crucial to alleviate this linearity assumption. One way to make the GRN model more appropriate is to include nonlinear terms, such as the so called S-system (Wang, Qian, & Dougherty, 2010), sigmoid function (De Jong, 2002; Huang, Tienda-Luna, & Wang, 2009; Noor, Serpedin, Nounou, & Nounou, 2012; Wang, Liu, Liu, Liang, & Vinciotti, 2009), etc.

When a nonlinear state-space model is adopted, the extended Kalman filter (EKF) is one efficient method for GRN structure recovering (Noor et al., 2012; Wang et al., 2009). The EKF based approach works well with both steady state data and slow dynamical data. On the other hand, there may occur considerable performance deteriorations in this approach if either the initial state estimate is incorrect or there are appreciable deficiencies in the system model caused by first-order approximations (Noor et al., 2012). More specifically, as the EKF based approach does not take either unmodelled dynamics or parametric uncertainties into account, its estimation performances may not be satisfactory due to its slow convergence speed which usually leads to low estimation accuracy. Mistakes are often caused by the low estimation accuracy of an estimation algorithm. For example, in inferring the structure of a GRN, an estimated parameter, say $\hat{g}^{[ij]}$, is often used to decide whether gene j directly regulates gene i. A false positive error is made when the actual value of $g^{[ij]}$ equals to zero, but its estimate $\hat{g}^{[ij]}$ has a large magnitude. This means that unmodelled dynamics and parametric uncertainties should not be ignored in identifications.

To enhance estimation performances, GRN structure recovering is resorted in this paper to the robust state estimator suggested in Zhou (2010), after the first-order approximation of GRNs. As a result, the suggested method is robust against model errors due to GRN linearizations and state estimate inaccuracies. Moreover, based on the specific structure of the estimation problem, it has been proved that under some weak requirements, the estimated network topology by both the EKF based method and the suggested method converges to the actual structure in the mean square sense, but the convergence speed of the suggested method is faster than that of the EKF based method. The suggested method has been used to identify an artificially constructed nonlinear GRN. Compared with the EKF based method, computation results show that the convergence speed is distinctly improved, and parametric estimation accuracy is significantly increased, which greatly reduces both false positive errors and false negative errors. Consistent results have also been obtained when these methods are applied to a benchmark problem proposed in the DREAM series project. Moreover, computation results with a real GRN of *yeast* show that the proposed method can identify causal relationships effectively.

In these computations, the suggested method has also been compared to another well known state estimation method for nonlinear dynamic systems, that is, the unscented Kalman filter (UKF), and similar observations have been obtained. This means that the suggested method may be helpful in solving actual GRN reconstruction problems.

The rest of this paper is organized as follows. In the next section, a robust structure identification algorithm is derived. Afterwards, convergence conditions are investigated in Section 3. In Section 4, some calculated results are reported. The paper is concluded by Section 5, in which some important characteristics of the suggested method are summarized, as well as some important works worthy of further efforts. An Appendix is included to give proofs of some technical results.

The following notation and symbols are adopted. $\mathrm{vec}(X)$ denotes the operation of stacking the columns of matrix X from left to right, while $\mathrm{diag}\left\{X_i|_{i=1}^N\right\}$ a block diagonal matrix with its ith block diagonal element being X_i, and $\mathrm{col}\left\{X_i|_{i=1}^N\right\}$ the vector/matrix stacked by $X_i|_{i=1}^N$ with its ith row block vector/matrix being X_i. $\left[X_{ij}|_{i=1,j=1}^{i=M,j=N}\right]$ represents a matrix with $M \times N$ blocks and its ith row jth column block matrix being X_{ij}, while $\mathrm{reshape}(A, M, N)$ an M row N column matrix whose elements are taken column-wise from an MN dimensional column vector A. $\mathbf{E}[x]$ stands for the expected value of a random variable x. δ_{ij} is the Kronecker delta function which equals to 1 when $i = j$ and to zero whenever $i \neq j$. $\mathrm{tr}(A)$ stands for the trace of the matrix A. Given two symmetric matrices P and Q with compatible dimensions, the inequality $P \geq Q$ means that $P - Q$ is positive semi-definite. Moreover, $\lambda_{\min}(A)$ represents the minimum eigenvalue of a symmetric matrix A.

2. Robust structure identification algorithm for GRNs

According to chemical principles, such as the Michaelis–Menten kinetics, etc., dynamic reactions occurred in a practical biochemical networks are inherently nonlinear, which means that GRNs must be treated in general as a nonlinear dynamic system (Akutsu et al., 1999; De Jong, 2002; Kauffman, 1993; Wang et al., 2010). An extensively adopted way in dealing with dynamic systems is the so-called state-space approach. In particular, a nonlinear state evolution equation for GRNs consisting of n genes can be described by[2]

$$x_{k+1} = f(x_k, \theta) + w_k, \tag{1}$$

in which, k stands for the temporal variable, $x_k = \mathrm{col}\{x_{k,i}|_{i=1}^n\}$ is the vector consisting of expression levels of all the genes, $f(x_k, \theta)$ is a vector of nonlinear functions, $\theta \in \mathbb{R}^p$ is a vector consisting of

[2] It is worthwhile to maintain here that in actual GRNs, rather than directly act on another gene, a gene exerts its influence through its mRNAs, proteins, etc. However, when relations among genes are discussed, models like Eqs. (1) and (3) are usually adopted (Huang et al., 2009; Kauffman, 1993; Marbach et al., 2012; Prill et al., 2010).