Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 48-28 (2015) 251–256

Reconstructing Large-scale Cellular Networks Using a Sparse Approximation Approach ⋆

Wanhong Zhang * **Yan Zhang** ** **Tong Zhou** ***

* *Department of Automation, Tsinghua University, Beijing, 100084 &,School of Chemical Machinery, Qinghai University, Qinghai, 810016 China (e-mail: zhangwh11@mails.tsinghua.edu.cn).*
** *School of Chemical Machinery, Qinghai University, Qinghai, 810016 China(e-mail: zy210825@163.com)*
*** *Department of Automation & TNList, Tsinghua University, Beijing, 100084 China(e-mail: tzhou@mail.tsinghua.edu.cn)*

Abstract: In this paper, a sparse reconstruction framework is proposed on the basis of steady-state experiment data to identify Gene Regulatory Networks (GRNs) structure. Different from traditional methods, this approach is adopted which is well suitable for a large-scale underdetermined problem in inferring a sparse vector. We investigate how to combine the noisy steady-state experiment data and a sparse reconstruction algorithm to identify causal relationships. Efficiency of this method is tested by an artificial linear network and the DREAM networks. The performance of the suggested approach is compared with two state-of-the-art algorithms, the widely adopted total least-squares (TLS) method and those available results on the DREAM project website. Actual results show that with a lower computational cost, the proposed method can significantly enhance estimation accuracy.

Keywords: Sparse approximation, GRNs, systems biology, causal identification, DREAM.

1. INTRODUCTION

In biological sciences, a significant task is to reconstruct GRNs from experiment data and other a priori information, which is a fundamental problem in understanding cellular functions and behaviors, see Hecker et al. (2009), and Feala et al. (2010). Spurred by advances in experimental technology, it is considerably interesting to develop a systematic method to provide new insights into the evolution of some target genes both in normal physiology and in human diseases. The present challenges in biological research are that GRNs are generally large-scaled and there are many restrictions on probing signals in biochemical experiments. These challenges make the problem of identifying a GRN much more difficult than other reverse engineering problems.

At present, numerous classical methods have been developed to unravel the interactions of GRNs, including Boolean network approaches in Shmulevich and Dougherty (2010), Bayesian network inference in Li et al. (2011), partial or conditional correlation analysis in Penfold et al. (2012), differential equation analysis in Karlebach and Shamir (2008), and others. However, while their absolute and comparative performances remain poorly understood, some of results are associated with heavy computational burdens. Recently, an approach based on the total differential formula and total least-squares is proposed to infer a GRN from measured expression data in Sontag (2008). Although this method can weaken the effect of experimental uncertainty, there exist significant false positive and negative errors. To overcome these difficulties, researchers have obtained some positive and constructive results and improvements in inferring a GRN, including incorporating power law in Xiong and Zhou (2012), distinguishing direct and indirect regulations in Wang and Zhou (2012), penalizing the regulation strength in Xiong and Zhou (2013), etc. However, these methods either have higher computing complexities or have lower estimation accuracies. Moreover, many methods may not be suited to large-scale network identifications. Then, how is it possible to accurately identify the causal relationships based on certain observable quantities extracted from partial measurements?

Note that great similarities exist between the network identification of a single gene (also called a node) and a sparse vector reconstruction, which often relates to the determination of the number, location, and magnitude of the nonzero entries by solving the problem of underdetermined system of linear equations $y = \Phi x$. In sparse reconstruction, the aim is to find the sparse solution x from the compressed measurement y and measurement matrix Φ. The classical algorithms find the solution to a sparse problem with minimal ℓ_1 norm. Since these algorithms, based on convex optimization, can guarantee global optimum and have strong theoretical assurance, the problem can be solved via linear programming in Donoho (2006); Candes (2008). However, the complexity is

⋆ This work was supported in part by the 973 Program (2009CB320602, 2012CB316504), the NNSFC (61174122 and 51361135705), and the SRFDPHE (20110002110045).

10.1016/j.ifacol.2015.12.134

burdensome and unacceptable for the application in large-scale systems. Recently, greedy algorithms have received considerable attention as cost effective alternatives of the ℓ_1-*minimization* in Wang et al. (2012). In the greedy algorithm family, stagewise orthogonal matching pursuit (StOMP) algorithm with the property either Φ that is random or that the nonzeros in x are randomly located, or both, is well suited to large-scale underdetermined applications for sparse vector estimations in Donoho et al. (2012). It can reduce computational complexity and has some attractive asymptotical statistical properties. However, the estimation speed is at the cost of accuracy violation. In our paper, an improvement algorithm on the StOMP which is called stagewise modified orthogonal matching pursuit (SmOMP), is suggested. This algorithm is more efficient at finding a sparse solution of large-scale underdetermined problems. Moreover, compared with StOMP, this modified algorithm can not only more accurately estimate parameters for the distribution of matched filter coefficients, but also improve estimation accuracy for the sparse vector in Zhang et al. (2013).

In this paper, a linear description of the causal interacting relationships for a GRN is firstly established from steady-state experiment data based on nonlinear differential equations. Then, we adopt a sparse reconstruction algorithm to find the sparse solution of a large-scale underdetermined problem. Finally, some applications, on an artificially generated linear network with 100 nodes and networks of size 100 in DREAM3 and DREAM4 subchallenges, are employed to demonstrate efficiency of this proposed algorithm. Moreover, we compare the performance of suggested approach with two state-of-the-art methods which are called subspace likelihood maximization (SubLM1 and SubLM2) methods in Zhou and Wang (2010), the widely adopted TLS method in Berman et al. (2007) and those available results on the DREAM project website. Computation results show that with a lower computational cost, the proposed method can significantly improve estimation accuracy.

2. MATERIALS AND METHODS

2.1 A description of the GRN model

In a GRN with n genes, we assume that the dynamics of the i-th gene concentration x_i can be described by the following nonlinear differential equation

$$\frac{dx_i}{dt} = f(x_1, x_2, \cdots, x_n; \theta_i), \tag{1}$$

in which θ_i stands for a kinetic parameter that can be changed through external perturbations. While each gene in the GRN reaches an equilibrium, there exist $dx_i/dt = 0, i = 1, 2, \cdots, n$, i.e. $f(x_1, x_2, \cdots, x_n; \theta_i) = 0$. In order to quantitatively measure the direct effect among genes, we quantify the causal interaction between two genes in terms of the fractional changes $\Delta x_i/\Delta x_j$ of the i-th gene caused by a change of another gene j. As argued in Kholodenko et al. (2002), at a stable equilibrium, the direct effect of the j-th gene on the i-th gene ($i \neq j$) can be measured by u_{ij} which results in log-to-log derivatives

$$u_{ij} = \lim_{\Delta x_i, \Delta x_j \to 0} \left(\frac{\Delta x_i/x_i}{\Delta x_j/x_j}\right) = \frac{\partial\, lnx_i}{\partial\, lnx_j} = -\frac{\partial f_i/\partial lnx_j}{\partial f_i/\partial lnx_i}. \tag{2}$$

If $u_{ij} = 0$, it means that gene j has no direct effect on gene i, whereas $u_{ij} > 0$ and $u_{ij} < 0$ mean activation and inhibition respectively. Let $\Delta_{x_j}^{[s]}$ denote the variation of the steady state ${x_j}^{[s]}$ when a kinetic parameter changes by Δ_{θ_j}. Then, taking the first-order Taylor expansions and normalization of each component at an equilibrium in the GRN, the following equation is obtained

$$\sum_{j=1}^{n} \frac{\partial f_i/\partial lnx_j}{\partial f_i/\partial lnx_i} \times \frac{\Delta_{x_j}^{[s]}}{x_j^{[s]}} \approx 0. \tag{3}$$

Suppose that m experiments have been performed, and the relative variation quantity of the j-th gene in the ℓ-th experiment is denoted by $\phi_{j\ell} = \Delta_{x_j}^{[s]} \big/ x_j^{[s]}$. Then, from the definition of u_{ij} and the above equation, we can easily obtain the causal relationship model of the i-th gene associated with the interaction among others as $\sum_{k=1,k\neq i}^{n} u_{ik}\phi_{k\ell} \approx \phi_{i\ell},\ \ell = 1, 2, \cdots, m$. Then, this causal regulation model can be compactly expressed as a linear equation (4), if a vector $[u_{i1}, \cdots u_{i(i-1)}, u_{i(i+1)} \cdots u_{in}]^{\mathrm{T}}$ is denoted by α_i, and an $m \times (n-1)$ measurement matrix Φ and the observation vector $b \in R^m$ are defined respectively as $b = \phi_i = [\phi_{i1}, \phi_{i2}, \cdots, \phi_{im}]^{\mathrm{T}}$, $\Phi = [\phi_1, \cdots, \phi_{i-1}, \phi_{i+1}, \cdots, \phi_n]$, in which T denotes the operation of transposing.

$$\Phi\alpha_i = b. \tag{4}$$

The problem of inferring a GRN requires the precise estimation α_i using steady-state experiment data. Since the distribution of the degree of nodes in most GRNs obeys approximately the so-called power law, the unknown vector α_i to be reconstructed is a sparse vector. Therefore, under the condition that both Φ and b are known, the problem discussed in this paper is to reconstruct a sparse vector. A distinctive characteristic of this problem is that both matrix Φ and vector b are corrupted by measurement noise. In the following section, the use of SmOMP for inferring GRNs is described.

2.2 The SmOMP algorithm

SmOMP aims to estimate the distribution parameters for matched filter coefficients more accurately and improve the estimate accuracy of the sparse solution based on the true positive rate (TPR). Suppose that the undetermined linear system equation is $y = \Phi x$ in which x is the original sparse vector. SmOMP operates in $s \leq S$ stages, building up a sequence of approximations x_0, x_1, $\cdots$ by removing detected structure from a sequence of residual vectors r_0, r_1, $\cdots$. Starting from $x_0 = 0$ and initial residue $r_0 = y$, it iteratively constructs approximations by maintaining a sequence of estimates for the locations of the nonzeros in x as $I_1, \ldots, I_s$.

At the s-th stage, we apply matched filtering to the current residual, obtaining a vector of residual correlations $c_s = \Phi^{\mathrm{T}} r_s$. In StOMP, authors demonstrate that $\langle \phi_j, r_s \rangle, j = 1, 2, \cdots, n$, are subject to the Gaussian distribution with zero or nonzero mean, which are corresponding to the first distribution and the second distribution.

We consider an m_s-dimensional subspace, using k_s nonzeros out of n_s possible terms. Note that the coefficients of this subspace are obtained by matched filtering as follows