Neurocomputing 160 (2015) 213–227

Contents lists available at ScienceDirect

Neurocomputing

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

Network decomposition based large-scale reverse engineering of gene regulatory network

Ahsan Raja Chowdhury [a,b,*], Madhu Chetty [a]

[a] School of Engineering and Information Technology, Federation University, Churchill, Vic 3842, Australia
[b] Department of Computer Science and Engineering, University of Dhaka, Dhaka, Bangladesh

ARTICLE INFO

Article history:
Received 15 August 2014
Received in revised form
10 February 2015
Accepted 13 February 2015
Communicated by L. Kurgan
Available online 25 February 2015

Keywords:
Gene regulatory network
Reverse engineering
Decomposition

ABSTRACT

A Gene Regulatory Network (GRN) is the functional circuitry of a living organism that exhibits the regulatory relationships among genes of a cellular system at the gene level. In real-life biological networks, the number of genes present are very large exhibiting both, the instantaneous and time-delayed regulations. While our recent technique [1] addresses the modeling of time-delays occurring in genetic interactions, the issue of large-scale GRN modeling still remains. In this paper, we propose a novel methodology for large-scale modeling of GRNs by decomposing the GRN into two independent sub-networks utilizing its biological traits. Using the time-delayed S-system model [1], these two sub-networks are learnt separately and then combined to get the entire GRN. To speed up the inference mechanism, a cardinality-based fitness function, especially developed for inferring large-scale GRNs is proposed to allow incorporation of knowledge of maximum in-degree. A novel local-search method is also proposed to further facilitate the incorporation of biological knowledge by gene clustering and gene ranking. Experimental studies demonstrate that the proposed approach is successful in learning large genetic networks, currently not achievable with existing S-system based modeling approaches.

1. Introduction

Reverse engineering Gene Regulatory Network is the process of representing the genetic interactions given by time-series data with an appropriate mathematical model. It can reveal the underlying biological processes of living organisms, and provide new insights into, e.g., the causes of complex diseases [2]. Accurate prediction of the behaviour of regulatory networks can also speed up biotechnological research since predictions are quicker, consistent and cheaper than the wet lab experiments. With the advent of cutting-edge microarray technologies, it has become possible to generate time series data embedded with biological knowledge which can help unravel the underlying genetic interactions using any of the model-based system identification methods. Computational methods, both for supporting the development of network models and for the analysis of their functionality, have already proved to be a valuable research tool.

The models for reverse-engineering GRNs can be broadly categorized into three major groups, namely co-expression network, Bayesian network and differential equation. Co-expression networks [3] are coarse-scale, simplistic models that employ pairwise association measures for inferring interaction between genes. Although these models require low computational time and can be scaled up to very large networks of thousands of genes, they lack the precision necessary for accurate modeling of system dynamics. On the other hand, Bayesian network (BN) models, based on the strong foundation of probability and statistics, are sophisticated and accurate but unable to implement feedbacks that are common in genetic network. The Dynamic Bayesian Network (DBN), a temporal form of BN, overcomes this limitation and allows feedback [4–6]. The third group, differential equation based models, belongs to a sophisticated and well established class of methods for modeling GRNs [7,1]. A salient feature of all differential equation based approaches is their ability to accurately model system dynamics in *continuous* time. While most approaches in this group apply Ordinary Differential Equation (ODE) to model interactions, some Delay Differential Equations (DDE) based approaches have also been reported [8,9].

The S-system, with a set of tightly coupled differential equations, is amongst the best non-linear differential models for modeling biochemical interactions. It is a rich model for capturing system dynamics and has been considered to provide an excellent balance between model complexity and mathematical tractability. These advantages of S-system model have led to large number of recent applications and improvements [10,11]. To reduce the computational burden, Maki et al. [12] proposed decoupled S-system model that divides the problem into N sub-problems, which improves the time required for

* Corresponding author at: School of Engineering and Information Technology, Federation University, Churchill, Vic 3842, Australia.
E-mail addresses: ahsan.chowdhury@federation.edu.au, farhan717@gmail.com (A. Raja Chowdhury), madhu.chetty@federation.edu.au (M. Chetty).

http://dx.doi.org/10.1016/j.neucom.2015.02.020
0925-2312/

reverse engineering. Despite application of decoupled system, the ability to infer GRN on large-scale, with thousands of genes, remains a major challenge. The S-System approach is known to be intrinsically compute-intensive [11,10,13] mainly because of the large number of parameters to be estimated and also due to the associated numerical integration to be carried out for each gene at every time step. The number of parameters (including time delays) to be learnt for a N-gene network is quadratic in the number of genes, i.e., $2N(N+1)+2\times N\times N=2N(2N+1)$. Hence, even for a network with, say 100 genes, the number of parameters would be 402 for a decoupled system (40200 if decoupling was not applied), which is a significantly large number of parameters to be learnt. To empirically provide an understanding of the time complexity, we executed our time-delayed S-system (TDSS) model [1] to infer a 100-gene synthetic network from a time-series having 10 datasets, each with 11 samples. On average, it took $\sim$1 min to complete a single iteration, requiring $\sim$25 h to infer the regulations for single gene (implemented in C++ using a 2.16 GHz Dual CPU with 3 GB of RAM with the same set of parameters mentioned in [1]). Thus, to reverse engineer the entire GRN of 100 genes, it would require 3.5 months for TDSS on a single PC. Using clusters of computers could ameliorate the situation, albeit the problem would still remain. For example, in 2003 Kikuchi et al. [11] used 1024 CPUs to solve the 5-gene network, which still took over 10 h to learn the parameters.

Due to computational complexity, the current application of S-system model for reverse engineering GRN is restricted to inferring either small-scale GRNs (i.e., 5–10 genes) or medium-scale GRNs (i.e., upto 50 genes) [14,15,1]. However, in reality, the genetic networks are large-scale networks consisting of thousands of genes and its reconstruction is referred to interchangeably as either "large-scale reverse engineering" or "reverse engineering large-scale GRNs". In this paper, we propose a modeling approach which enables the inference of such large-scale GRNs using S-system.

In this paper, to perform S-system based large-scale GRN modeling, we propose improvements in both, the modeling paradigm and the model parameter learning. To develop a new S-system model, we bifurcate the entire set of genes into two groups. The first group comprise of all genes which are responsible for controller action, i.e., Transcription Factor (TF), Enzyme catalyst, antibody, etc. We will call all these genes collectively as Regulatory Genes (RGs). The second group will have genes, other than RGs, called Target Genes (TGs), which can be controlled by RG but do not have any controlling action of their own. Separating genes into two categories of RG and TG and incorporating this knowledge into the modeling process allows us to decompose the overall GRN into two sub-networks, which can be reconstructed independently thereby significantly reducing the enormous computational complexity if a complete GRN network reconstruction were to be undertaken in traditional manner. The RGs, TGs and their interactions are illustrated by a synthetic GRN of 20 genes [7] shown in Fig. 1(a). Of these 20 genes, 4 genes $(G_9, G_{16}, G_{19}, G_{20})$ do not regulate any other genes but merely perform self-degradation. On the other hand, the remaining genes regulate each other and also these 4 genes (i.e., TGs). Fig. 1(b), redrawn from Fig. 1(a), clearly illustrates the difference between the roles of RGs $(G_1-G_8, G_{10}-G_{15}, G_{17}-G_{18})$ and TGs $(G_9, G_{16}, G_{19}, G_{20})$. For improving evolutionary optimization technique, suitable for large-scale modeling of GRN, we propose to learn the model parameters using a novel cardinality based fitness function inspired by power law distribution of genes' in-degree along with a novel local search method based on knowledge-based gene clustering and gene ranking.

The modeling approach, briefly highlighted above, allows modeling large-scale GRNs (typically thousands of genes) with S-system model. This is made possible due to the bifurcation of the entire GRNs into two sub-networks and adapting the existing S-system modeling approach to these two sub-networks. Without decomposition, optimization of model parameters would not have been possible due to very high computational complexity due to large number of genes involved. Since the real-life GRNs consist of thousands of genes, the proposed method could thus find application in any large-scale GRN modeling, e.g., understanding the p53-MDM2 genes feedback cycle in cancerous cells. Further, the approach can be applied for drug design after accurate inference of the interactions amongst disease genes.

The rest of the paper is organized as follows. Section 2 presents the preliminaries relevant to the proposed method. The details of the proposed model called TDSS$^+$ are elaborated in Section 3. Section 4 is devoted to the evaluation of TDSS$^+$ for various networks, while the discussion of the results are presented in Section 5. Finally, Section 6 concludes the paper.

2. Preliminaries

2.1. The S-system model

The S-system model, proposed by Savageau [16], is a well-known system for biochemical networks and is found to be both promising and challenging for GRN modeling. For an N gene network, the

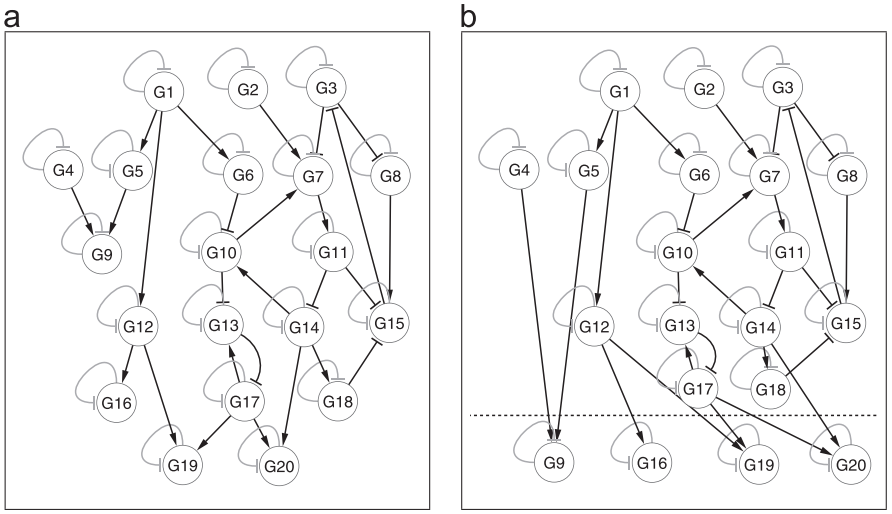

Fig. 1. (a) 20-gene network of [7], (b) 20-gene network after re-arrangement where 4 TGs $(G_9, G_{16}, G_{19}, G_{20})$ and 16 RGs $(G_1-G_8, G_{10}-G_{15}, G_{17}-G_{18})$ are shown separately. Arrow and block ended edges represent activations and suppressions, respectively.