Inferring *E. coli* SOS Response Pathway from Gene Expression Data Using IST-DBN with Time Lag Estimation

Lian En Chai, Mohd Saberi Mohamad*, Safaai Deris,
Chuii Khim Chong, and Yee Wen Choon

Artificial Intelligence and Bioinformatics Research Group, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Skudai 81310, Johor, Malaysia
{lechai2,ckchong2,ywchoon2}@live.utm.my, {saberi,safaai}@utm.my

Abstract. Driven to discover the vast information and comprehend the fundamental mechanism of gene regulations, gene regulatory networks (GRNs) inference from gene expression data has gathered the interests of many researchers which is otherwise unfeasible in the past due to technology constraint. The dynamic Bayesian network (DBN) has been widely used to infer GRNs as it is capable of handling time-series gene expression data and feedback loops. However, the frequently occurred missing values in gene expression data, the incapability to deal with transcriptional time lag, and the excessive computation time triggered by the large search space, are attributed to restraint the effectiveness of DBN in inferring GRNs from gene expression data. This paper proposes a DBN-based model (IST-DBN) with missing values imputation, potential regulators selection, and time lag estimation to address these problems. To assess the performance of IST-DBN, we applied the model on the *E. coli* SOS response pathway time-series expression data. The experimental results showed IST-DBN has higher accuracy and faster computation time in recognising gene-gene relationships when compared with existing DBN-based model and conventional DBN. We also believe that the ensuing networks from IST-DBN are applicable as a common framework for prospective gene intervention study.

Keywords: Dynamic Bayesian network, missing values imputation, time-series gene expression data, gene regulatory networks, network inference.

1 Introduction

In the post-genomic era, aided by the breakthroughs in technology, researchers have begun to shift the research paradigm from the classical reductionism to the modern holism, wherein biological systems and experimental design are viewed as a whole instead as collections of parts [1]. One of the innovations conceived in such era, the

* Corresponding author.

A.S. Sidhu & S.K. Dhillon (Eds.): *Adv. in Biomedical Infrastructure 2013*, SCI 477, pp. 5–14.
DOI: 10.1007/978-3-642-37137-0_3

DNA microarray technology, which is capable of representing the expression of thousands of genes under various circumstances (otherwise known as gene expression profiling), has allowed the development of numerous new experiments for exploring into the complex system of gene expression and regulation [2]. Since its conception, various organisms and mammalian cells have been profiled, such as *S. cerevisiae* [3], human cancerous tissue [4], and *E. coli* [5]. The consequent output, commonly known as gene expression data, comprises immense information such as the robustness and behaviours denoted by the cellular system under diverse situations [6], assists us in understanding the underlying mechanism of gene expression and regulation.

From a computational perspective, a GRN can be represented as a directed graph containing nodes (genes) and edges (interaction/relationship). In recent years, various computational methods have been developed to infer GRNs from gene expression data. Among them, Bayesian network (BN) [7], which uses probabilistic correlation to distinguish relationships between a set of variables, was popular in GRNs inference. This is mainly due to several factors: BN is capable of working on local elements, assimilating other mathematical models to avert data overfitting, and merging prior knowledge to fortify the causal relationships. Nonetheless, BN also has two disadvantages: it is unable to deal with time-series gene expression data and construct feedback loops.

From a biological perception, feedback loops actually embody the homeostasis procedure in living organisms. Hence, to take account of the feedback loops, researchers have developed the dynamic Bayesian network (DBN) [8] as a replacement to tackle BN's weaknesses. However, the scattering missing values commonly found in gene expression data could affect more than 90% of the genes and subsequently negatively influencing downstream analysis and inferring approaches [9]. Furthermore, in identifying gene-gene relationships, conventional DBN generally comprises all genes into the subsets of potential regulators for each target gene, and thus instigated the large search space and the excessive computational time [10]. To address the two problems, Chai *et al.* [11] suggested a three-step DBN-based model (ISDBN) with missing values imputation and potential regulators selection, and the proposed model showed better performance than conventional DBN in GRNs inference.

Yet, ISDBN and conventional DBN is still not adept enough to effectively take account of the transcriptional time lag, in which a time delay exists before the target genes are being expressed into the system. This shortcoming hampers the accuracy of DBN-based approaches in GRNs inference. To solve this problem, we proposed to further improve the aforesaid DBN-based model with time lag estimation (IST-DBN) which would take account of the transcriptional time lag based on the time difference between the initial changes of expression level of potential regulators and their target genes.

2 Methods

Essentially, IST-DBN involves four main steps: missing values imputation, potential regulators selection, time lag estimation and DBN inference. Fig. 1 illustrates the schematic overview of IST-DBN.