A Novel Graph Optimisation Algorithm for the Extraction of Gene Regulatory Networks from Temporal Microarray Data

Judit Kumuthini[1], Lionel Jouffe[2], and Conrad Bessant[1]

[1] Cranfield University, Barton Road, Silsoe, Beds, UK
[2] Bayesia – 6, rue Léonard de Vinci – BP0102 – 53001 Laval Cedex - France
s.j.kumuthini.s02@cranfield.ac.uk

Abstract. A gene regulatory network (GRN) extracted from microarray data has the potential to give us a concise and precise way of understanding how genes interact under the experimental conditions studied [1, 2]. Learning such networks, and unravelling the knowledge hidden within them is important for drug targets and to understand the basis of disease. In this paper, we analyse microarray gene expression data from *Saccharomyces cerevisiae,* to extract Bayesian belief networks (BBNs) which mirror the cell cycle GRN. This is achieved through the use of a novel structure learning algorithm of Taboo search and a novel knowledge extraction technique, target node (TN) analysis. We also show how quantitative and qualitative information captured within the BBN can be used to simulate the nature of interaction between genes. The GRN extracted was validated against literature and genomic databases, and found to be in excellent agreement.

Keywords: Gene regulatory networks, Bayesian belief networks, Taboo search and knowledge extraction.

1 Introduction

The entire network system of mechanisms that governs gene expression is a very complex process, regulated at several stages of protein synthesis [3]. However, the activity of genes is a result of protein and metabolite regulation, and proteins themselves are gene products. Thus, how genes are regulated in a biological system, i.e. which genes are expressed, when and where in the organism, and to what extent, is crucial for determination of the regulatory network.

A GRN is not only the representation of the biochemical activity at the system level, it also provides a large-scale, coarse-grained view of the physiological state of an organism at the mRNA level. It not only provides an overall view of the system at the genetic level, it also provides the precise way of gene interaction under the experimental conditions studied and the dynamic properties of those underlying interactions [1, 2]. The knowledge within a gene regulatory network might provide valuable clues and lead to new ideas for treating complex diseases. Such knowledge can aid pharmaceutical research to fulfil nonexistent drug and gene therapy, providing

S. Hochreiter and R. Wagner (Eds.): BIRD 2007, LNBI 4414, pp. 28–37, 2007.

possibilities to tailor target drugs for individual patients, to unravel the causal factor of some serious genetic diseases, and to better understand side effects.

In recent years, several groups have developed methodologies for the extraction of regulatory networks from gene expression data. Spellman *et al* [4] used DNA microarrays to analyse mRNA levels in yeast cell cultures that had been subjected to synchronised cell cycle arrest using three independent methods. This protocol allowed the inclusion of data from previously published studies such as Cho *et al.*, 1998 [5], and identified 800 genes that met an objective minimum criterion for cell cycle regulation. In contrast, traditional methods identified 104 genes as cell cycle regulated, while the study by Cho *et al.*, 1998, using a manual decision process identified 421 genes. In fact, the Spellman study included 304 of the genes identified by Cho *et al*, perhaps indicating that the technical differences in the way the two studies were carried out may have contributed to the differences between the results. Spellman *et al* in particular, employed a statistical approach for identification of genes associated in cell cycle process, using a larger and diverse number of experiments.

In this paper we describe the use of the novel Bayesian network methodology of Taboo search to extract GRNs from temporal microarray data of the yeast cell cycle. From a statistical viewpoint, a Bayesian network (BN) efficiently encodes the joint probability distribution of the variables describing an application domain. BNs are represented in a graphical annotated form that allows us to quickly understand the qualitative part of the encoded knowledge. The nodes of a BN correspond to random variables describing the domain and the arcs that connect the nodes correspond to direct probabilistic relations between these variables.

We go on to interrogate the networks learnt using TN analysis. This shows how all the genes in the GRN are associated with a specific gene of interest, and allow us to interactively evaluate how changes in the expression of selected genes impact on the behaviour of other genes.

2 Methods

2.1 Gene Expression Data

The data was obtained from a cell cycle analysis project, published by [4], whose aim was to identify all genes regulated by the cell cycle. This time series data has been used by [1, 6, 7], Yihui *et al.*, 2002, Anshul *et al.*, and others for genome wide gene expression and cell cycle regulation studies. The data is freely available to the public from the yeast cell cycle project site, http://genome-www.stanford.edu/cellcycle/. The Spellman dataset contained temporal gene expression measurements of the mRNA levels of 6177 *S. cerevisiae* ORFs. This project considers the cell cycle regulation of yeast, which traditional laboratory analysis indicates consists of six separable phases. These are M/G1, late G1 (SCB regulated), late G1 (MCB regulated), S, S/G2, and G2/M. In our analysis, late G1 (SCB regulated) is denoted as G1, and late G1 (MCB regulated) is denoted as Late G1. At the time of writing, the KEGG database for the yeast cell cycle pathway only contains 88 genes, and these are grouped according to cell cycle phases, G1, S, G2 and M.