Learning gene regulatory networks with an Intelligent data analysis approach: an application to the yeast cell cycle

Riccardo Bellazzi, Roberta Amici, Fulvia Ferrazzi, Paolo Magni, Lucia Sacchi, Stefano Sotgiu
Dipartimento di Informatica e Sistemistica, Università di Pavia
Via Ferrata 1, 27100 Pavia, Italy
riccardo.bellazzi@unipv.it

Abstract

This paper presents a novel approach for the extraction of gene regulatory networks from DNA microarray data. The method is applied to the reconstruction of a network of interactions of genes involved into the cell cycle of Saccharomyces Cerevisiae. The approach is characterized by the integration of data coming from different experiments together with the knowledge available on the biological process under analysis and on the dynamics of the process itself. The method is capable to reconstruct known relationships among genes and to provide meaningful biological results.

1 Introduction

A noteworthy research effort in Biomedical informatics has been recently devoted to the development of methods for the automated extraction of gene regulatory networks from DNA microarrays data. Such interest is motivated by the capability of DNA microarrays to describe cell molecular processes at the whole genome level. The availability of experiments in which a certain cell condition is followed over time gives the chance to learn dynamic models of gene to gene interactions. Several algorithms have been implemented so far: a pioneering work is represented by the REVEAL approach, which extracts networks expressing Boolean relationships between genes through a heuristic search strategy based on mutual information [Liang et al., 1998]. More recently, other methods have been presented to derive regulatory networks from microarray data, including methods based on differential equations [De Jong, 2002] and dynamic probabilistic networks [Perrin et al., 2003]. All those methods have pros and cons; however, given the very nature of the data, none of the approaches may lead to reveal all the biochemical pathways underlying the observed processes. As a matter of fact, a certain mRNA stream does not always correspond to the same protein, due to potential modifications after transcription and after translation; even more importantly, the dynamics of biochemical reactions cannot be captured by the (low) sampling time available in DNA microarray experiments. For these reasons, it is of interest to integrate data coming from different sources, multiple experiments and the available background knowledge to derive models which should be able to describe as close as possible regulatory interactions occurring between genes. In this paper we present a novel approach to derive a network of potential interactions of genes involved in the yeast cell cycle. The approach integrates data coming from two different experiments and the knowledge available on the biological process and on the dynamics of cell cycle.

2 Modeling gene networks

Following the approach proposed by Schlitt and Brazma [Schlitt, 2005], it is possible to model gene networks at different levels of detail. As a consequence, four basic classes of models can be distinguished: a) Parts lists, referring to the collection and systematization of the network components; b) Topology models, describing the interactions between the parts; c) Control logic models, describing the effect of regulatory signals; d) Dynamic models, modeling the dynamics of gene interactions. The so-called part list is often directly extracted from knowledge available in Gene Ontology (Gene Ontology™ Consortium, http://www.geneontology.org). Such information allows to select only the genes which are known to be involved in the process which is under study. However, other secondary bioinformatics databases can be conveniently exploited, such as the Gene database, maintained at NCBI (http://www.ncbi.nlm.nih.gov). The gene-gene interaction network topology is learned from data. In this case, it is crucial to assign a meaning to the network connections. In the literature, a first interpretation is that, given two genes G1 and G2 represented in the network as nodes, G1 is directly linked to G2 only if G1 is a transcription factor for G2. In this case the link describes a physical interaction between the two genes. A second interpretation is that an edge between G1 and G2 means a generic “cause-effect” relationship, such that a change in the expression of G1 causes a change in the expression of G2. In this case we are describing a phenomenological event, regardless of the physical interactions between the two genes. Rather interestingly, in some model organisms, such as Saccharomyces Cerevisiae (baker’s yeast), it is now possible to learn from data both kind of networks.
An important data set on the interactions between the genes and their transcription factors has been collected by Lee et al [Lee et al 2002] in the so-called ChIP-on-chip experiments. Such data have been used to derive the topology of a network of physical interactions.

On the other side, Hughes et al. [Hughes et al., 2000] performed a complex experiment to detect the effects of a single gene mutation. Given a DNA microarray experiment on a mutant, corresponding to a single knocked-out gene, a significant change of the expression level in any of the non-mutated genes with respect to the wild-type case is supposed to highlight a relationship with the knocked out gene.

As mentioned in the introduction, a large number of control models have been studied in the literature, starting from Boolean relationships and moving towards probabilistic ones [Liang et al., 1998; De Jong, 2002, Perrin et al., 2003]. All those models can be considered also dynamic models, although the emphasis is not given to the description of the biochemical reactions, but rather to the phenomenological relationships between the problem variables, i.e. the genes. Such models are often derived from “dynamic data”, i.e. time series of gene expression profiles usually collected with experiments carried on in cell cultures [Spellman et al., 1998].

A consistent literature is also available on the quantitative modeling of the biochemical networks. For what concerns yeast, for example, several papers appeared on the cell cycle dynamics [Sveiczer et al., 2004]. It is important to notice that such models are designed for simulation purposes, and aim at describing at a “physical” level the gene product interactions. Since they must model also fast reactions, they are typically not identifiable from data, but they require knowledge on the stoichiometric coefficients of each single biochemical reaction.

In our case, we are interested in providing a description of the interactions of the genes involved in the cell cycle of Saccharomyces Cerevisiae, taking into account all the four levels mentioned above: we will propose a network model based on different data sources and on the knowledge available in the knowledge repositories (parts lists), which relies on a network topology derived from data (topology modeling), and which models the dynamics of control interactions between genes (control logic and dynamic models).

3 The proposed approach

In this paper we propose a method to infer gene to gene interaction networks in Saccharomyces Cerevisiae cell cycle. The basic steps of the method are described in Figure 1; they can be summarized as follows: 1) learning of an initial network topology from mutant data; 2) selection of the genes involved in the cell cycle; 3) filtering of the selected genes on the basis of the available data on the cell cycle dynamics; 4) learning the final interaction network and a dynamic model of control with a genetic algorithm search.

![Figure 1. The proposed method](image)

3.1 Learning the initial network topology from mutant data

This step is based on the analysis of the data made available by Hughes et al. [Hughes et al., 2000], already introduced in Section 2. They collected the data of about 300 experiments in which a single gene has been knocked-out and the RNA abundance of all the other genes (about 6800) has been measured through c-DNA microarrays. The goal of this study was the detection of the functional modules of each mutated gene. Starting from the mutants experiments, it is possible to derive a first network of gene interactions: this network can be easily represented with a connection matrix \( D \) with elements \( D_{ij} \) which express the relationships between gene \( i \) and gene \( j \); if \( D_{ij}=1 \) the connection is present, if \( D_{ij}=0 \) the connection is absent.

After the analysis of the Huges data, we obtained a matrix of 6800 x 276 elements, where each column corresponds to an experiment with a single mutated gene, while each row corresponds to a certain gene. The semantic of the network can be augmented with the sign of the relationship (enhancement or inhibition).

3.2 Gene ontology and dynamic networks filtering

The dimension of the matrix \( D \) can be conveniently reduced by resorting to the knowledge available in Gene Ontology. In our case we selected only the genes involved in the cell cycle biological process, thus reducing the matrix \( D \) to 502 x 34.

Since our main goal was to learn a dynamical model of the control of genes involved in the cell cycle, we then resorted to the “dynamic” data sets available in the literature. In the case of yeast cell cycle, the reference data are the ones coming from a well-known experiment from Spellman [Spellman et al., 1998]. In this case the mRNA data have been collected in 18 different time points (one each 7 minutes). Since the cell cycle for the yeast under the experimental conditions lasts 66 minutes, it is possible to observe almost two complete mitotic cycles.
3 Results

Interesting results have been obtained in all phases of the learning process. To evaluate such results, we considered 22 genes whose role in the cell cycle is well characterized and we investigated the capability of our method of reconstructing the known relationships on the basis of the available data.

We first exploited the data coming from the Hughes disruption experiment, in which only 6 of those 22 genes have been mutated. We thus inferred a network (shown in Figure 2a) in which some connections appear to be supported by the information available in the literature (e.g. some links involve a gene and its transcription factor). This network was extended following the strategy proposed in this paper: in the final graph obtained (shown in Figure 2b) a significant number of the inferred connections between the 22 cell cycle genes reflects the knowledge available in the literature about the gene to gene interactions. In particular, the network shows the following interesting relationships:

a) Mcm1 interacts with Clb1: the genes that normally exhibit a G2-to-M-phase-specific expression pattern, such as Clb1, are not induced in the absence of functional Mcm1; moreover, it was demonstrated that Clb1 transcript levels are substantially reduced when functional Mcm1 is absent. b) the Clb5-Clb1 and Clb2-Clb1 links express complex (indirect) interactions between cyclins, the proteins which regulate the overall cell cycle (see http://mips.gsf.de/genre/proj/yeast/). c) Far1 is a cyclin-dependent kinase inhibitor, and it is therefore activated by the cyclin levels, such as Clb1.

Figure 2. Graph connectivity of some of the 22 well-characterized cell cycle genes: a) initial disruption network, b) final network obtained exploiting background knowledge and dynamic data

Examining the overall derived network, we observed a scale-free connectivity: about 170 genes out of 226 are linked with no more than 5 genes, while only 10 genes are connected with more than 40 other genes. Such latter genes are the hubs of the final gene interaction network. Some of the hubs are: Swi4, the DNA binding component of SBF transcription factor; the two B-type cyclins Clb1 and Clb2, activators of Cdc28 at G2/M phase of the cell cycle; Cdc46, that encodes a member of the Mcm2-7 family of proteins involved in the initiation of DNA replication; Cdc27, subunit of the Anaphase-Promoting
Acknowledgments

This work is part of the PRIN Project “Modelli dinamici dell’espressione genica da dati di microarray: tecniche di clustering e reti di regolazione” funded by the Italian Ministry of Education. We gratefully thank Sonia Rinaldi for her help in running the Genetic algorithm tests.

References


