INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 76S (2007) S462–S475

journal homepage: www.intl.elsevierhealth.com/journals/ijmi

Inferring gene regulatory networks by integrating static and dynamic data

Fulvia Ferrazzi[a], *Paolo Magni*[a], *Lucia Sacchi*[a], *Angelo Nuzzo*[a], *Uroš Petrovič*[b], *Riccardo Bellazzi*[a,*]

[a] *Dipartimento di Informatica e Sistemistica, Università degli Studi di Pavia, via Ferrata 1, 27100 Pavia, Italy*
[b] *J. Stefan Institute, Ljubljana, Slovenia*

ARTICLE INFO

Article history:
Received 5 January 2007
Received in revised form
13 June 2007
Accepted 26 July 2007

Keywords:
Gene regulatory networks
Machine learning
Data interpretation
Microarray analysis of gene expression
Genetic algorithm

ABSTRACT

Objectives: The purpose of the paper is to propose a methodology for learning gene regulatory networks from DNA microarray data based on the integration of different data and knowledge sources. We applied our method to *Saccharomyces cerevisiae* experiments, focusing our attention on cell cycle regulatory mechanisms. We exploited data from deletion mutant experiments (static data), gene expression time series (dynamic data) and the knowledge encoded in the Gene Ontology.

Methods: The proposed method is based on four phases. An initial gene network was derived from static data by means of a simple statistical approach. Then, the genes classified in the Gene Ontology as being involved in the cell cycle were selected. As a third step, the network structure was used to initialize a linear dynamic model of gene expression profiles. Finally, a genetic algorithm was applied to update the gene network exploiting data coming from an experiment on the yeast cell cycle.

Results: We compared the network models provided by our approach with those obtained with a fully data-driven approach, by looking at their AIC scores and at the percentage of preserved connections in the best solutions. The results show that several nearly equivalent solutions, in terms of AIC scores, can be found. This problem is greatly mitigated by following our approach, which is able to find more robust models by fixing a portion of the network structure on the basis of prior knowledge. The best network structure was biologically evaluated on a set of 22 known cell cycle genes against independent knowledge sources.

Conclusions: An approach able to integrate several sources of information is needed to infer gene regulatory networks, as a fully data-driven search is in general prone to overfitting and to unidentifiability problems. The learned networks encode hypotheses on regulatory relationships that need to be verified by means of wet-lab experiments.

1. Introduction

In recent years, the reverse engineering of gene networks from gene-expression microarray data has received increasing interest [1,2]. Microarray data can be static, i.e. snapshots of gene expression in different experimental conditions, such as in mutants, or dynamic, i.e. time series of gene expression data collected during the evolution of processes of particular

* *Corresponding author.*
E-mail address: riccardo.bellazzi@unipv.it (R. Bellazzi).
1386-5056/$ – see front matter © 2007 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2007.07.005

interest, for example the cell cycle. In particular, thanks to the availability of dynamic data, it is possible to learn models that describe how gene interactions evolve over time.

Several models, which differ on the way the interactions between the involved genes are represented, were proposed in the literature to describe gene networks. The first models to be introduced were Boolean models, where the interactions between genes are represented as logical relationships between binary variables that describe the behavior of genes in terms of activation and deactivation [3–5]. To overcome the potential limitations introduced by the strong assumptions underlying Boolean models, which basically treat genes as being either in an 'on' or 'off' status, other methodological approaches were developed; among these we recall differential equations models and Bayesian networks. In differential equation models, the rate of production of a cellular component is described as a function of the concentrations of other components [6–8]. Bayesian networks and dynamic Bayesian networks (an extension of Bayesian networks to explicitly manage the system dynamics) appear particularly suited to represent probabilistic relationships between stochastic variables [9–12]. Other interesting approaches proposed in the literature are information-theoretic methods and module networks. An example of the former type of methods was proposed by Margolin et al. [13]: mutual information is used to select correlated gene expression profiles; a filtering procedure is then applied to remove correlations resulting from indirect relationships between genes and thus identify a set of potentially interacting genes. Module networks highlight the temporal relationships between groups of synchronized co-regulated genes, the so-called *modules*, which share a common function. The modules are reconstructed by integrating gene expression data and biological knowledge on the interactions between the involved genes; such a network structure can be conveniently used to infer the temporal sequence of biological sub-processes [14].

These models can be very useful to generate hypotheses about gene interactions. Yet, due to the complexity of learning true regulatory networks, they usually extract phenomenological models, i.e. models able to describe the available data but with no guarantee that the discovered interactions have a "true" biological meaning. For example, it is typically not possible to distinguish between direct or indirect interactions between genes. Moreover, as DNA microarrays give information only at transcriptional level, none of the approaches may lead to reveal all the biochemical pathways underlying the observed processes. Indeed, a certain mRNA molecule 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 rate available in DNA microarray experiments. Finally, in a pure data-driven approach, a large number of models might describe the available data equivalently well. This is caused by the conjunction of three main factors: the large number of available genes, the low number of available samples and the presence of measurement noise. This problem, well-known in the mathematical modeling community, is known as a posteriori unidentifiability [15,16].

A potential solution to the above-mentioned problems is the integration of data coming from different sources with the background knowledge available on the process under study. This integrative approach is expected to provide models that are more likely to describe the true regulatory interactions occurring between genes. In particular, gene deletion experiments are considered an extremely effective way to infer gene interactions. However, such experiments are typically static, since they are expensive and the number of potential deletions is very high. Their joint use with dynamic data may be a powerful way to infer functional relationships between genes.

In this paper, we concentrate on the issue of knowledge and data integration and we present a novel approach to derive a network of potential interactions between genes involved in the *Saccharomyces cerevisiae* cell cycle. The approach integrates data coming from static deletion experiments [17] and from wild-type time series measurements [18] as well as prior knowledge available in the literature on genes involved in the cell cycle and on the dynamics of the cell cycle.

The method exploits a linear regression model to describe gene expression dynamics and uses a genetic algorithm to search through the space of gene network structures. A set of possible interaction networks, sampled from the space of solutions, is inferred. This allows an analysis on a posteriori identifiability. In particular, we will show the advantage of exploiting prior knowledge and multiple data sources with respect to a fully data-driven approach.

2. Method

In the literature several interpretations of the meaning of a gene regulatory network are given. Here, a gene regulatory network is defined as a directed graph where an arc between two genes is drawn when it is assumed that the first gene directly or indirectly participates in regulating (activating or repressing) the expression of the second. In particular we will work on networks aimed at describing co-expression relationships between genes [19], which can be interpreted as potential control relationships by analyzing the delays between the gene expression profiles. Because of this, in our study regulators are not necessarily transcription factors.

Learning gene regulatory networks requires a number of modeling choices. Typically, such choices involve the dynamic model describing gene relationships (difference or differential equations, regression, logic-based models) and the search strategy for model selection. In our case, the choices also involve the inclusion of prior knowledge and the use of heterogeneous data sets. In particular, we defined a four-step procedure, which can be summarized as follows (Fig. 1):

1. learning of an initial gene network topology using deletion mutant data;
2. selection of the genes involved in the cell cycle on the basis of the knowledge reported in the Gene Ontology;
3. filtering of the genes on the basis of the available knowledge on cell cycle dynamics;
4. learning of the final interaction network and of a model of gene expression dynamics through a genetic algorithm search coupled with regression models.