Discrete Applied Mathematics 161 (2013) 580–592

Contents lists available at SciVerse ScienceDirect

Discrete Applied Mathematics

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

An integer optimization approach for reverse engineering of gene regulatory networks

Roberto Cordone [a], Guglielmo Lulli [b,*]

[a] University of Milano, Department of Computer Science, Via Comelico 39, 20135 Milano, Italy

[b] University of Milano "Bicocca", Department of Informatics, Systems and Communication, Viale Sarca 336, 20122 Milano, Italy

ARTICLE INFO

Article history:
Received 29 November 2010
Received in revised form 9 November 2011
Accepted 8 February 2012
Available online 5 March 2012

Keywords:
Gene regulatory networks
Lagrangian relaxation
Tabu search
Variable neighborhood search

ABSTRACT

Gene regulatory networks are a common tool to describe the chemical interactions between genes in a living cell. This paper considers the Weighted Gene Regulatory Network (WGRN) problem, which consists in identifying a reduced set of interesting candidate regulatory elements which can explain the expression of all other genes. We provide an integer programming formulation based on a graph model and derive from it a branch-and-bound algorithm which exploits the Lagrangian relaxation of suitable constraints. This allows to determine lower bounds tighter than CPLEX on most benchmark instances, with the exception of the sparser ones. In order to determine feasible solutions for the problem, which appears to be a hard task for general-purpose solvers, we also develop and compare two metaheuristic approaches, namely a Tabu Search and a Variable Neighborhood Search algorithm. The experiments performed on both of them suggest that diversification is a key feature to solve the problem.

1. Introduction

Gene regulatory networks are the on–off switches and rheostats of a cell operating at the gene level. They dynamically orchestrate the level of expression for each gene in the genome by controlling whether and how vigorously that gene will be transcribed into RNA [21]. The network structure is an abstraction of the system's chemical dynamics, describing the manifold ways in which one substance affects all the others to which it is connected. Schematically, these dynamics can be formalized using graph theoretical models, where nodes represents genes and arcs represent direct regulatory interactions (i.e., influences of genes upon the expression of other genes). These interactions can be either activating, i.e., an increase in the level of expression of one gene leads to an increase in the other, or inhibiting, i.e., an increase in the level of expression of one gene prevents the transcription of the other.

Unveiling such networks is essential for understanding how genomic expression programs unfold during developmental processes, how the molecular machinery of cells works to respond adequately to environmental clues and to maintain homeostasis, and, consequently, how to manipulate these processes to human advantage. Hence, gaining an understanding of the emergence of complex patterns of behavior from the interactions between genes in a regulatory network poses a huge scientific challenge with potentially high industrial pay-offs.

In view of recent technological advances, which enable biomedical investigators to observe the genome of entire organisms in action by simultaneously measuring the level of expression of thousands of genes under the same experimental conditions, several methods have been proposed to reconstruct gene regulatory networks. The goal of these methods is

[*] Corresponding author. Fax: +39 0250316412.
E-mail addresses: roberto.cordone@unimi.it (R. Cordone), lulli@disco.unimib.it (G. Lulli).

0166-218X/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2012.02.010

to produce a high-fidelity representation of the cellular network topology as a graph, thus explaining gene expression data [16]. There is a wide spectrum of techniques and criteria to define an arc, such as Bayesian networks (e.g., [2,10]), Boolean networks (e.g., [19]), systems of Ordinary Differential Equations (e.g., [4,5]) and statistical methods (e.g., [8,20]). Among others, de Jong [6], Filkov [7], Friedman [9], Lee [16], van Someren [22] surveyed most of the existing methods to date. Although all these approaches may seem unrelated, they might not be as far apart as it appears and could serve complementary roles. Indeed, some of these approaches provide a complete description of gene regulatory networks while others focus on either specific influences or small parts of a gene regulatory network. All these methods – as any reverse engineering approach – show a trade-off between the level of computational difficulty and the level of approximation of the real network. In fact, due to the reduced amount of data and to measurement noise, the inferred gene regulatory networks might include both spurious correlations and indirect relations (effect of a sequence of direct influences).

Herein, we present a mathematical model for making sense of large, multiple time-series data sets arising in expression analysis. Analyzing the expression profiles for all pairs of genes it is possible to identify the set of all putative activation/inhibition influences. These influences can be even scored by mean of the activation–inhibition index described in [8]. This index ($s \in [-1, 1]$) measures how strongly one gene activates or inhibits another: absolute values close to 1 denote a correlation between the two genes ($+1$ if activating, -1 if inhibiting); the value zero denotes the absence of a relation between the two genes. The mathematical model herein presented is purposely designed to generate networks in which a relatively small number of regulators explain the expression of all other genes, i.e., it aims to identify a small set of interesting candidate regulatory elements. Those genes that are not identified as candidate regulatory elements are considered neutral, i.e., do not have any activation/inhibition influence upon other genes of the network, though they play a role in biochemical intracellular and intercellular processes. We do not assert that our approach identifies the regulatory network, but we believe that it quickly enables biologists to identify and visualize interesting features from raw expression array data sets. The mathematical model we propose is similar in spirit to the maximum gene regulation problem proposed by Chen et al. [3]. However, with respect to that model, we remove some of their simplifying hypotheses thus obtaining a more biologically accurate model. In fact, we include in our model the presence of neutral genes, which are in general the great majority. For instance, the number of known regulators (activator/inhibitor) in the *Saccharomyces cerevisiae* is approximately 8% of the total number of genes (http://www.yeastgenome.org). Moreover, in the Chen et al. model each gene can only exert either activation or inhibition influences, while in our model, coherently with biological evidence [12,18], we do allow some genes to exert both activation and inhibition influences.

The unweighted version of the model herein studied has been already investigated in [17] and validated on a real world test derived from the *Saccharomyces cerevisiae* genome.

The paper is organized as follows. In Section 2 we present the mathematical programming formulation of the model. Some simple properties of the problem are described in Section 2.1. To solve instances of the problem, we present both an exact method based on branch-and-bound and two metaheuristic algorithms to compute good quality solutions in a reasonable amount of time. Finally, Section 6 contains the conclusions.

2. The weighted gene regulatory formulation

In this section, we first define the Weighted Gene Regulatory Network (WGRN) problem, which consists in designing a network in which a relatively small number of regulators explain the expression of all the genes.

In mathematical terms, we suppose that a given putative gene network is represented by a graph $\mathcal{G}(N, A \cup I, w)$. N, the set of nodes, represents the genes, $A \subset N \times N$ and $I \subset N \times N$ are the sets of arcs representing the putative activations (A) and inhibitions (I) respectively. The map w associates to each arc of the set $A \cup I$ a weight derived from the activation–inhibition index [8]. More specifically, $w = 1 - |s|$, so that a strong correlation between two genes corresponds to a small weight and a weak correlation to a large weight. The decision problem is to identify a subset of genes which explain the expression of all the genes by acting as either activators or inhibitors. This means that each gene has at least one activator and one inhibitor arc incoming from the identified subset of genes. In general the genes identified as activators (resp. inhibitors) only exert activation (resp. inhibition) influences. However, in order to promote or to repress the expression of all the genes of the network, few exceptions to the cornerstone mentioned above are allowed, i.e., we allow some genes to exert both activation and inhibition influences. The presence of such genes is consistent with biological evidence, but their number should be rather small [12,18]. Therefore, in our model, we minimize the total weight of activation (resp. inhibition) influences exerted by a node (gene) labeled as inhibitor (resp. activator). This type of influences are here named *incoherent*.

To formulate the problem we define the following decision variables. To each node of the network, we assign two binary variables:

$$z_i^{(A)} = \begin{cases} 1 & \text{if node } i \text{ is labeled as activator,} \\ 0 & \text{otherwise.} \end{cases}$$

$$z_i^{(I)} = \begin{cases} 1 & \text{if node } i \text{ is labeled as inhibitor,} \\ 0 & \text{otherwise.} \end{cases}$$

We also use a binary decision variable for each arc of the putative network to discern if it is used as incoherent influence.

$$x_{ij} = \begin{cases} 1 & \text{if arc } (i, j) \in A \cup I \text{ is an incoherent influence,} \\ 0 & \text{otherwise.} \end{cases}$$