Targeting Differentially Co-regulated Genes by Multiobjective and Multimodal Optimization

Oscar Harari[1], Cristina Rubio-Escudero[1], and Igor Zwir[1,2]

[1] Dept. Computer Science and Artificial Intelligence, University of Granada, E-18071, Spain
[2] Howard Hughes Medical Institute, Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO 63110-1093, USA
oharari@decsai.ugr.es, crubio@decsai.ugr.es, zwir@borcim.wustl.edu

Abstract. A critical challenge of the postgenomic era is to understand how genes are differentially regulated in and between genetic networks. The fact that such co-regulated genes may be differentially regulated suggests that subtle differences in the shared *cis*-acting regulatory elements are likely significant, however it is unknown which of these features increase or reduce expression of genes. In principle, this expression can be measured by microarray experiments, though they incorporate systematic errors, and moreover produce a limited classification (e.g. up/down regulated genes). In this work, we present an unsupervised machine learning method to tackle the complexities governing gene expression, which considers gene expression data as one feature among many. It analyzes features concurrently, recognizes dynamic relations and generates profiles, which are groups of promoters sharing common features. The method makes use of multiobjective techniques to evaluate the performance of profiles, and has a multimodal approach to produce alternative descriptions of same expression target. We apply this method to probe the regulatory networks governed by the PhoP/PhoQ two-component system in the enteric bacteria *Escherichia coli* and *Salmonella enterica*. Our analysis uncovered profiles that were experimentally validated, suggesting correlations between promoter regulatory features and gene expression kinetics measured by green fluorescent protein (GFP) assays.

1 Introduction

Genetic and genomic approaches have been successfully used to assign genes to distinct regulatory networks. However, little is known about the differential expression of genes within a regulon. At its simplest, genes within a regulon are controlled by a common transcriptional regulator in response to the same inducing signal. Moreover it is suggested that subtle differences in the shared *cis*-acting regulatory elements are probably significant in the genes expression. However, it is not known which of these features, independently or collectively, can set expression patterns apart. Indeed, similar expression patterns can be generated from different or a mixture of multiple underlying features, thus, making it more difficult to discern the causes of analogous regulatory effects.

E. Marchiori, J.H. Moore, and J.C. Rajapakse (Eds.): EvoBIO 2007, LNCS 4447, pp. 68 – 77, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The material required for analyzing the promoter features governing bacterial gene expression is widely available. It consists of genome sequences, transcription data, and biological databases containing examples of preciously explored cases. In principle, genes could be differentiated by incorporating into the analysis quantitative and kinetic measurements of gene expression [1] and/or considering the participation of other transcription factors [2-4]. However, there are constraints in such analyses due to systematic errors in microarray experiments, the extra work required to obtain kinetic data and the missing information about additional signals impacting on gene expression. These constraints hitherto allow a relatively crude classification of gene expression patterns into a limited number of classes (e.g., up- and down-regulated genes [5, 6]), thus concealing distinctions among expression features, such as those that characterize the temporal order of genes or their levels of intensity

Here we describe an unsupervised machine learning method that discriminates among co-regulated promoters by simultaneously considering both cis-acting regulatory features and gene expression. By virtue of being an unsupervised method, it is neither constrained by a dependent variable [2, 7], such as expression data, which would restrict the classification to the dual expression classes reported by microarray experiments; nor it requires pre-existing kinetic data. Our method treats each of the promoter features with equal weight, because it is not known beforehand which features are important. Thus, it explores all of the possible aggregations of features; and applies multiobjective and multimodal techniques [8, 9] to identify alternative optimal solutions that describe target sets of genes from different perspectives.

We applied our methodology to the investigation of genes regulated by the PhoP protein of *Escherichia coli* and *Salmonella enterica* serovar Typhimurium. We recovered several profiles that were experimentally validated [10] to establish that PhoP uses different configurations of promoter to regulate genes. We finally correlated these groups with more accurate independent experiments that measure gene expression over time by using GFP assays.

2 Methods

The purpose of this method is to identify all of the possible substructures, here termed profiles (i.e., groups of promoters sharing a common set of features), that characterize sets of genes. These common attributes can ultimately clarify the key *cis*-features that produce distinct kinetic patterns, shedding light in the transcriptional mechanisms that the cell employs to differentially regulate genes belonging to a regulon.

The identification of the promoter features that determine the distinct expression behavior of co-regulated genes is a challenging task because (i) the difficulty in ascertaining the role of the differences in the shared *cis*-acting regulatory elements of co-regulated promoters; (ii) detailed kinetic data that would help the classification of expression patterns is not always available, or it is available for a limited subset of genes; and (iii) the limited extent of genes regulated by a transcriptional factor. To circumvent these constrains, our method explores all of the possible *cis*-feature aggregations, looking for those that better characterize different subset of genes; uses an unsupervised approach, where pre-existing classes are not required; and allows a