Integration of Known Transcription Factor Binding Site Information and Gene Expression Data to Advance from Co-Expression to Co-Regulation

Maarten Clements*, Eugene P. van Someren, Theo A. Knijnenburg, and Marcel J.T. Reinders

Information and Communication Theory Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2600 GA Delft, the Netherlands.

The common approach to find co-regulated genes is to cluster genes based on gene expression. However, due to the limited information present in any dataset, genes in the same cluster might be co-expressed but not necessarily co-regulated. In this paper, we propose to integrate known transcription factor binding site information and gene expression data into a single clustering scheme. This scheme will find clusters of co-regulated genes that are not only expressed similarly under the measured conditions, but also share a regulatory structure that may explain their common regulation. We demonstrate the utility of this approach on a microarray dataset of yeast grown under different nutrient and oxygen limitations. Our integrated clustering method not only unravels many regulatory modules that are consistent with current biological knowledge, but also provides a more profound understanding of the underlying process. The added value of our approach, compared with the clustering solely based on gene expression, is its ability to uncover clusters of genes that are involved in more specific biological processes and are evidently regulated by a set of transcription factors.

Key words: gene clustering, gene regulation, binding motifs

Introduction

Current technologies have enabled scientists access to complete sequence information as well as to genome-wide gene activity measurements for an ever-growing number of organisms. However, unraveling gene regulation by means of promotor analysis and/or cluster analysis remains a challenging task. In the last few years, many new computational methods have been developed to automatically detect regulatory motifs. These tools can be divided into two main categories: scanning methods and *de novo* methods. The scanning methods use a motif representation resulting from experimentally determined binding sites to scan the genome sequence to find additional matches (*1*). The *de novo* methods attempt to find novel motifs that are enriched in a set of upstream sequences (*2*–*6*). In order to identify regulatory programs, those *de novo* motif detection methods can be applied to the promotor regions of gene clusters to detect frequently occurring sequence patterns, which may be related to certain transcription factors (TFs) (*7*, *8*). However, in these methods, the identified regulation program of a gene cluster is considered as the final result; whether the regulatory program sufficiently explains the observed expression of all members of the gene cluster is not evaluated.

Segal *et al* (*9*) used a more advanced method that attempts to construct complex regulatory mechanisms from the expression profiles of known TFs. They assume that the expression level of the TFs is directly related to the expression of the genes that are regulated by them. There exists, however, clear biological evidence that this simple model is not always valid (*10*). Beer *et al* (*11*) circumvented the need to use the TF profiles as input by using sequence data instead. Utilizing AND, OR, and NOT logic and placing severe constraints on motif strength, orientation, and relative position, a large number of complex rules can be derived. However, these hypotheses need to be biologically validated before they would be useful to be incorporated in a clustering scheme.

***Corresponding author.**
E-mail: m.clements@tudelft.nl

We propose to incorporate TF binding potential data into the clustering scheme, such that for each newly discovered cluster, a *single* common regulatory structure sufficiently explains the behavior of *all* the genes in the cluster. Recently, different methods have been proposed that also let the regulation program adapt the grouping of genes. Segal *et al* (*12*) employed the expectation maximization algorithm that iteratively partitions the gene set and applied this gene partition to detect new motif candidates. In this way transcriptional modules are built that are both coherent in expression profiles and have common binding sites. Middendorf *et al* (*13*) used both gene regulators and putative binding sites to build a decision tree that tries to explain the gene expression profiles in terms of regulators and motifs. A similar method from Ruan *et al* (*14*) applies a multivariate regression tree to discover a model for gene expression patterns.

The above methods generally aim to find new motifs that are assumed to be involved in the regulation of the uncovered clusters of genes. In other words, both the clusters and the motifs are free parameters that have to be optimized. However, the rather poor performance of *de novo* motif discovery methods (*15*), combined with the uncertainty that remains in gene clustering (*16*), make it often difficult to link the regulatory programs with existing biological knowledge. As both the motifs and the gene clusters can be unknown, the biological interpretation of such results is, therefore, severely limited.

In this work, we propose to integrate the occurrence of known regulating elements in the upstream region of genes together with their expression levels as a combined input to the clustering system. The fact that our method only inputs validated TF motifs allows for an easier biological interpretation of the clusters and their discovered regulation structure. This increases the usefulness of the results and facilitates biologists in their studies to decipher the function of the genes regulated under given experimental conditions. More specifically, we identify three different scenarios where the integration of known TF binding site information and gene expression data leads to clusters of co-regulated genes that are not only expressed similarly under the measured conditions, but also share a regulatory structure that may explain their common regulation (Figure 1).

Results

Combining gene expression and gene regulation

Our proposed methodology is depicted in Figure 2. Complete details can be found in Materials and Methods. Here we give a short description of each step.

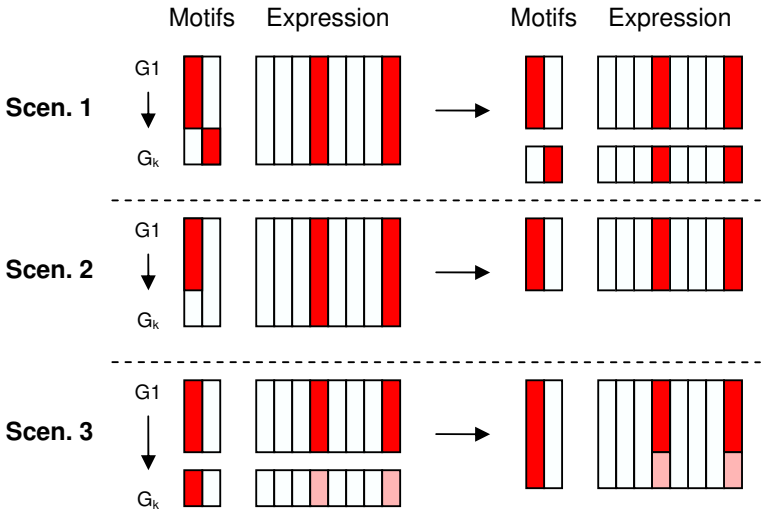

Fig. 1 The goal of the proposed method is to find co-regulated gene clusters that have similar expression profiles and share a similar set of motifs. The reason why the integration of motif enrichment results in a more functionally related module is threefold. Scenario 1: A cluster that is actually regulated by two different motifs is split up into separate clusters. Scenario 2: A cluster showing homogeneous expression is shrunk to a smaller cluster in which all genes contain the same motif. Scenario 3: Genes that show weak co-expression are integrated in one cluster because they share the same motif.