Identification of functional modules based on transcriptional regulation structure

Etienne Birmele, Mohamed Elati, Céline Rouveirol, Christophe Ambroise

BMC Proceedings

Proceedings

Identification of functional modules based on transcriptional regulation structure

Etienne Birmelé*[1], Mohamed Elati*[2], Céline Rouveirol[2] and Christophe Ambroise*[1]

Address: [1]Laboratoire Statistique et Génome, UMR CNRS 8071, INRA 1152, Tour Evry 2, F-91000 Evry, France and [2]LIPN – UMR 7030 CNRS – Université Paris 13, 99 Av. J.B. Clément, F-93430 Villetaneuse, France

Email: Etienne Birmelé* - etienne.birmele@genopole.cnrs.fr; Mohamed Elati* - Mohamed.Elati@lipn.univ-paris13.fr; Céline Rouveirol - rouveirol@lipn.univ-paris13.fr; Christophe Ambroise* - christophe.ambroise@genopole.cnrs.fr

* Corresponding authors

from Machine Learning in Systems Biology: MLSB 2007
Evry, France. 24–25 September 2007

Published: 17 December 2008

BMC Proceedings 2008, **2**(Suppl 4):S4

This article is available from: http://www.biomedcentral.com/1753-6561/2/S4/S4

Abstract

Background: Identifying gene functional modules is an important step towards elucidating gene functions at a global scale. Clustering algorithms mostly rely on co-expression of genes, that is group together genes having similar expression profiles.

Results: We propose to cluster genes by co-regulation rather than by co-expression. We therefore present an inference algorithm for detecting co-regulated groups from gene expression data and introduce a method to cluster genes given that inferred regulatory structure. Finally, we propose to validate the clustering through a score based on the GO enrichment of the obtained groups of genes.

Conclusion: We evaluate the methods on the stress response of S. Cerevisiae data and obtain better scores than clustering obtained directly from gene expression.

Background

An important step in analyzing gene functions is to cluster genes according to their expression patterns. Such clusters can then be analyzed in several ways, for example by assigning unannotated genes to the majority function of each cluster's genes (see [1] for a review).

However, this approach has several limitations. On the one hand, genes of similar expression patterns may not necessarily have the same or similar functions; on the other hand, genes with related functions may not show close correlation in their expression patterns. For example, a transcription factor can activate some genes and repress others in the same pathway.

The principal assumption of this paper is that unsupervised clustering of genes on the basis of similar regulators (activators/inhibitors) should assemble functional co-regulated groups of genes. To compute a similarity measure between genes as a function of inferred regulators of these

genes, we use the output of a data mining algorithm called LICORN [2], that infers cooperative regulation relations from expression data only. The resulting similarity matrix between genes is considered as the adjacency matrix of a weighted graph. Clustering is then performed to find functional modules of genes in the network.

To objectively evaluate clustering, we use Gene Ontology to determine if the obtained clusters can be associated with terms of the Biological Process ontology. The strength of such an association is given by a *p*-value from an Hypergeometric test. We compare different clusterings by calculating a score based on the *p*-values which becomes greater when the significant *p*-values are smaller and more numerous.

In section Methods, we introduce our model of gene regulation and briefly describe a data mining algorithm for inferring large-scale cooperative gene regulation. We then propose a similarity measure for the genes based on the inferred regulator sets and define the score of a clustering. Finally, in section Results and discussion, we evaluate our system on a yeast data set.

Methods

Cooperative regulation networks

Let us denote by $\mathcal{R}$ the set of genes with a known or putative regulation activity and $\mathcal{G}$ as the set of genes without such an activity. The input of the mining method is a discretised expression matrix for genes of $\mathcal{R} \cup \mathcal{G}$. Each expression value can take the value -1 (under-expressed), 0 (normal), or 1 (over-expressed). A gene regulatory network (GRN) associated with a target gene g is a pair (A, I), where $A \subseteq \mathcal{R}$ is a co-activator set, and $I \subseteq \mathcal{R}$ is a co-inhibitor set. The set of GRNs for all target genes can also be seen as a bipartite graph where the top layer contains regulators, the bottom layer contains target genes, and edges code for a regulatory interaction between regulators and target genes, each edge being labelled with a regulatory mode (i.e., *activator* or *inhibitor*). The regulation relations we are interested in are combinatorial: each target gene has a number of activators and/or inhibitors. Activators on one side and inhibitors on the other side are aggregated in our model through an extended logical AND, i.e., a regulator set S (activator or repressor) is over-expressed (resp. under-expressed) if and only if all the regulators in S are over-expressed (resp. under-expressed). Finally, we describe in Figure 1 a discrete function called *Regulatory Program* RP, which, given the combined states of activators A and inhibitors I of g in a sample s computes $\hat{g}_s(A, I)$, the estimated state of g in s. The main features of our regulation model are therefore the explicit representation of activation and repression relationships for a given target gene, and the representation of co-operative transcriptional regulation.

Learning algorithm

We have recently proposed [2] an original, scalable technique called LICORN for deriving co-operative regulations, in which many co-regulators act together to activate or repress a target gene. LICORN uses an original heuristic approach to accelerate the search for an appropriate structure for the regulation network. It first computes frequent co-regulator sets, i.e., regulator sets that frequently occur together as over (1) or under (-1)-expressed in the discretised expression matrix. This is done by using an extension of the Apriori algorithm [3] to handle both 1 and -1 supports (The x-support of a co-regulator C in the three-valued expression matrix is the set of samples that include all the regulators of C with the state x).

From this representation, a limited subset of candidate co-regulator sets is then associated with each gene. The learning algorithm looks for each gene for regulator sets which have a high "overlap" with the target gene. Intuitively, the overlap constraint checks the size of the intersection between supports of the target gene and a given candidate co-regulator set. A candidate activator set for a target gene

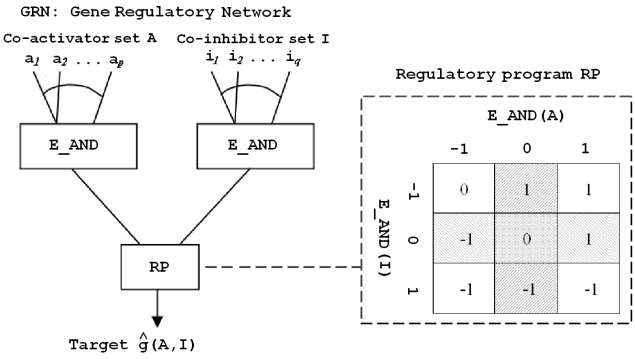

Figure 1

The regulatory program. Definition of the regulatory program RP, which can be interpreted as follows: i) If GRN contains co-activators only, $\hat{g}(A, I)$ corresponds to the aggregated status of these co-activators. ii) If GRN contains co-inhibitors only, $\hat{g}(A, I)$ is the inverse of the aggregated status of these co-inhibitors. iii) Otherwise, $\hat{g}(A, I)$ depends on a combination of the statuses of co-activators and co-inhibitors, as described by the matrix on the right. For example, $\hat{g}(A, I) = 1$ when the co-activators are over-expressed and the co-inhibitors are not.

g is frequently over-expressed when g is over-expressed or frequently under-expressed when g is under-expressed. On the opposite, a candidate repressor set for a target gene g is frequently over-expressed when g is under-expressed and vice-versa. This search can be efficiently performed because of the property of anti-monotonicity of the overlap constraint with respect to set inclusion. Then, once a limited number of candidate activator and inhibitor sets have been obtained, exhaustive search for the best gene regulatory network can be performed. Finally, a permutation-based procedure is used for selecting statistically significant regulation relations. We have shown in [2] that the co-operative regulation patterns inferred by LICORN cannot be identified by clustering or pairwise methods, and are only partly revealed by constrained Bayesian or decision tree-based techniques, such as those used in previous studies [4,5].

Identification of functional co-regulation modules

Partial overlap of the regulator sets for a set of target genes can be used as an alternative measurement of the distance between genes.

Computation of the co-regulation matrix

We design the co-regulation matrix by using a similarity measure defined as follows: let $\lambda \in [0, 1]$ and (g_1, g_2) be two genes. The similarity between g_1 and g_2 is defined by

$$\phi(g_1, g_2) = \frac{|A| + |I| + \lambda |AI|}{|TF|},$$

where $|A|$ and $|I|$ are respectively the number of activators and inhibitors of both g_1 and g_2, $|AI|$ is the number of regulators which activate one gene and inhibit the other and $|TF|$ is the number of transcription factors regulating at least one of the genes. This similarity considers two genes as being far appart ($\varphi(g_1, g_2) = 0$) if they do not share any regulators. Two genes are considered most similar if their set of activators and inhibitors are exactly the same ($\varphi(g_1, g_2) = 1$). In intermediate situations, λ represents the weight given to common regulators which have opposite effects.

Clustering

To cluster genes from the similarity matrix, we use the MCL algorithm [6,7]. That algorithm, based on the fluxes

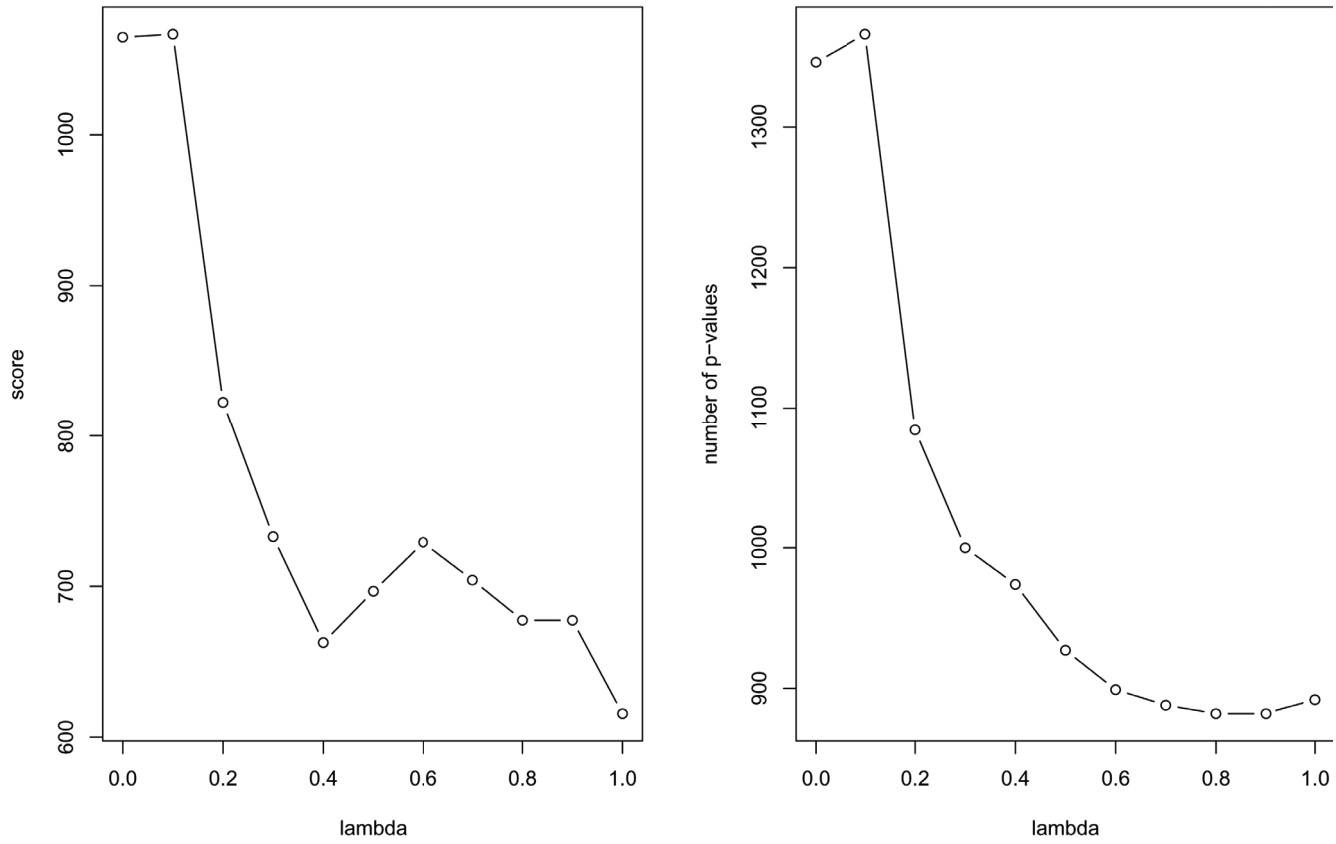

Figure 2
Score and number of p-values for λ varying from 0 to 1.

in a graph, is well suited to weighted graphs and does not require any prior knowledge about the number of clusters. Moreover, it does not require any initial conditions and is therefore reproducible. The inflation parameter of the algorithm is fixed to 1.8, as suggested in [8].

Mapping to GO-terms

To assess the functional significance of obtained clusters, and suggest putative functions for genes with unknown functions, we calculate the enrichment of gene ontology (GO) [9].

To determine the over-represented GO terms in each cluster, we apply the R package GOstats [10] with a p-value cut-off of 5% and the *biological process ontology*. For each cluster C, we obtain a set T_C of GO terms over-represented in C with a rate of 5% and a set of associated p-values $\{p_t, t \in T_C\}$.

We define the score of the clustering by

$$S(\lambda) = \sum_{C, c_{min} \leq |C| \leq c_{max}} \sum_{t \in T_C} -\log(p_t) \frac{|Ct|}{|C|}$$

where C_t is the set of genes of C associated with the GO term t. Parameters c_{min} and c_{max} allow us to avoid clusters that are too small, which don't have a biological meaning, as well as too big ones, which don't have any functional unity.

Results and discussion

As a proof of concept, we used gene expression data sets for *S. Cerevisiae*. The Gasch data set [11] measures the response of yeast to 173 stress conditions for 6152 genes. We used a set of 237 known and putative transcription factors.

We applied LICORN and retained only those GRNs (gene regulatory networks) identified as significant with a 5% FDR level (see [2] for details). 2041 GRNs (of 5703 GRNs) were identified as significant. The structural organization of the learned GRNs has been shown to be consistent with recent advances [12] concerning the characterization of topological transcriptional network features in yeast and provide the first evidence of the relevance of inferred GRNs.

In order to choose the parameter λ for the similarity matrix, we computed the matrices and the associated clusterings for several values of λ and compared their scores with parameters $c_{min} = 5$ and $c_{max} = 200$. Figure 2 shows that the best one is obtained for $\lambda = 0.1$.

For $\lambda = 0.1$, the clustering gives 30 clusters among which one is too big to be considered (407 genes) and 3 have less than 5 genes. Table 1 gives the best GO term association

Table 1: GO-enrichment of the clusters obtained for *S. Cerevisiae*.

Cluster Id	GO BP Id	*p*-value	Cluster size	Biological process
6	0022613	1.28e – 23	80	ribonucleoprotein complex biogenesis and assembly
15	0006119	3.26e – 13	18	oxidative phosphorylation
9	0042254	3.15e – 11	55	ribosome biogenesis and assembly
4	0006081	4.89e – 07	142	aldehyde metabolic process
2	0000746	5.89e – 07	155	conjugation
7	0007001	9.48e – 07	68	chromosome organization and biogenesis (sensu Eukaryota)
13	0006974	4.10e – 05	30	response to DNA damage stimulus
27	0008652	2.02e – 04	6	amino acid biosynthetic process
10	0046907	3.50e – 04	52	intracellular transport
8	0019754	7.84e – 04	66	one-carbon compound catabolic process

Table of the ten best clusters among the 59 ones obtained for $\lambda = 0.3$ ranked by their best association with a GO term. For each of them, the best associated GO term and the corresponding *p*-value are given, as well as the size of the cluster and the biological process associated to the GO term.

for the 10 best of them when ranked according to their best *p*-value. The biological evaluation of these clusters is ongoing.

The cluster number 15 that appears on the second line of Table 1 is in fact associated to 32 GO terms with a *p*-value lower than $1e-07$, most of those terms being related to phosphorylation or triphosphate metabolic process. Moreover, five genes of that cluster belong to the 167 genes having no associated GO term, namely the genes YLR296W, YDR215C, YBL044W, YIR040C and YPR027C. All of them appear in the Entrez gene database but without known functions.

We have finally validated our method by comparing clustering performances based on other similarity matrices. We therefore have computed from the original expression data matrices of euclidian distance, partial correlation [13] and mutual information [14]. To compare clustering results with the same number of clusters, we used the hierarchical clustering method AGNES [15] to cluster the genes in 20, 30, 40, and 50 groups. Figure 3 shows the scores for those three methods as well as for ours with $\lambda = 0.1$. It clearly shows that inferring the regulatory network from LICORN preprocessing improves the score of the clustering and provide more biologically relevant clusters.

Conclusion

The problem of discovering functional modules from expression data is both biologically important and computationally challenging. From a biological perspective, identifying members of functional modules is the first step toward understanding the regulatory network of the cell. We provide here an alternative way for constructing gene modules: genes are clustered in the same module if they share many regulators, as they have been inferred by LICORN from gene expression data. We expect this way of clustering will discover modules that are out the scope of classical co-expression clustering techniques. From a computational perspective, one of the key challenges is dealing with over-fitting as the number of data samples is so small.

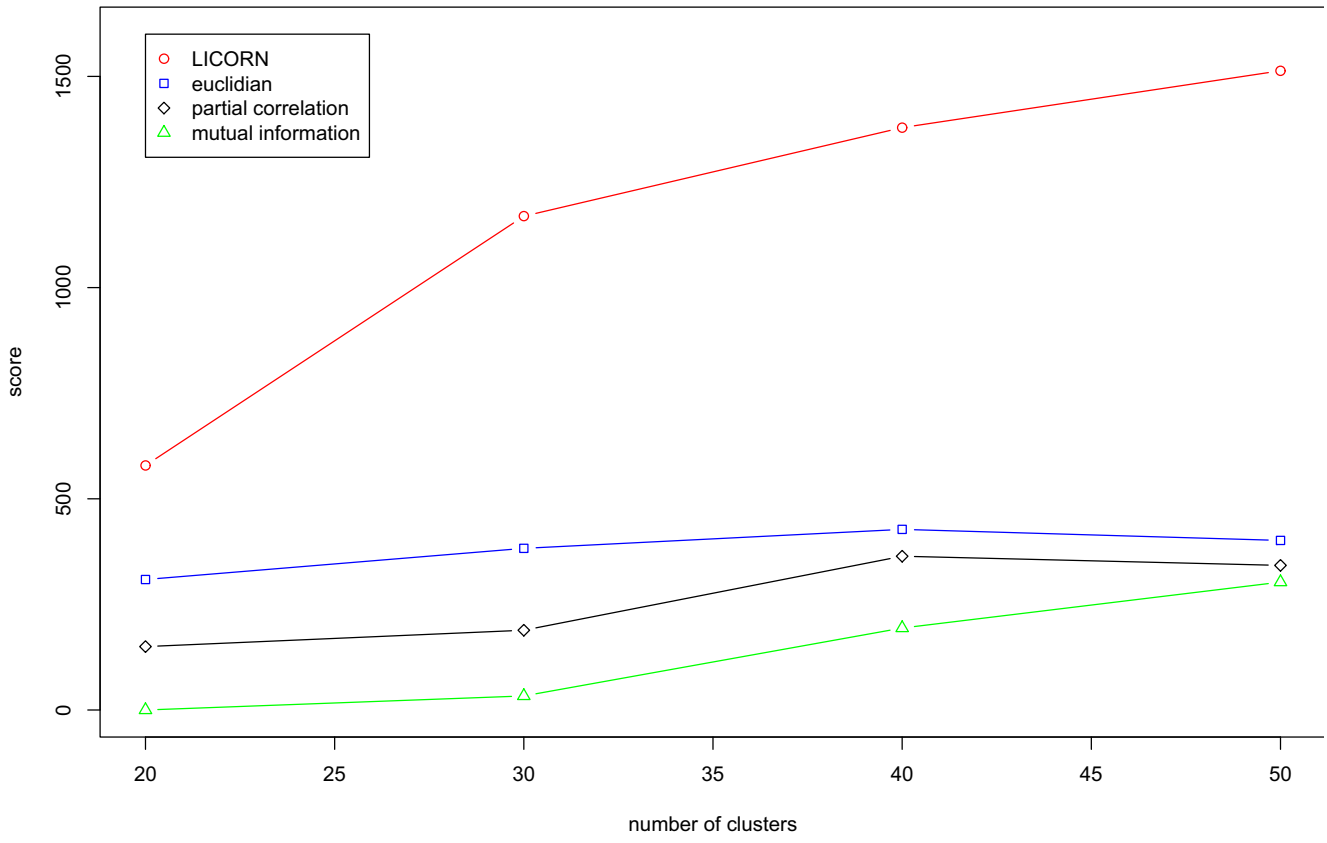

Figure 3
Comparison of the clustering based on LICORN with existing methods. Figure of the scores obtained for hierarchical clustering into 20, 30, 40 and 50 clusters. The red circles are the scores obtained for the similarity matrix given by LICORN and $\lambda = 0.1$. The similarity measures which are compared to are euclidian distance, partial correlation and mutual information.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
ME and CR designed the inference and participated in experimentations and drafting the manuscript. EB and CA proposed the clustering and evaluation methods.

Acknowledgements
Te authors would like to thank Monique Bolotin-Fukuhara for her helpful comments.

This article has been published as part of *BMC Proceedings* Volume 2 Supplement 4, 2008: Selected Proceedings of Machine Learning in Systems Biology: MLSB 2007. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/2?issue=S4.

References
1. Armstrong N, Wiel M van de: **Microarray data analysis: from hypotheses to conclusions using gene expression data.** *Cellular Oncology* 2004, **26:**.
2. Elati M, Neuvial P, Bolotin-Fukuhara M, Barillot E, Radvanyi F, Rouveirol C: **LICORN: learning co-operative regulation networks from gene expression data.** *Bioinformatics* 2007, **23:**2407-2414.
3. Agrawal R, Imielinski T, Swami A: **Mining Association Rules between sets of items in large databases.** *Proceedings of the International Conference on Management of Data* 1993:207-216.
4. Pe'er D, Regev A, Tanay A: **Minreg: inferring an active regulator set.** *Bioinformatics* 2002, **18(Suppl 1):**S258-S267.
5. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.** *Nature Genetics* 2003, **34:**166-176.
6. van Dongen S: **Graph Clustering by Flow Simulation.** In *PhD thesis University of Utrecht*; 2000.
7. Enright A, Ouzounis C, van Dongen S: **An efficient algorithm for large-scale detection of protein families.** *Nucleic Acids Research* 2002, **30(7):**1575-1584.
8. Brohée S, van Helden J: **Evaluation of clustering algorithms for protein-protein interaction networks.** *BMC Bioinformatics* 2006, **7:**488.
9. Cherry J, Adler C, Ball C, Chervitz S, Dwight S, Hester E, Jia Y, Juvik G, Roe T, Schroeder M, Weng S, Botstein D: **SGD: Saccharomyces Genome Database.** *Nucleic Acids Res* 1998, **26:**73-79.
10. Falcon S, Gentleman R: **Using GOstats to test gene lists for GO term association.** *Bioinformatics* 2007, **23(2):**257-258.
11. Gasch A, Spellman P, Kao C, Carmel-Harel O, Eisen M, Storz G, Botstein D, Brown P: **Genomic expression programs in the response of yeast cells to environmental changes.** *Mol Biol Cell* 2000, **11(12):**4241-57.
12. Guelzim N, Bottani S, Bourgine P, Képès F: **Topological and causal structure of the yeast transcriptional regulatory network.** *Nature Genetics* 2002, **31:**60-63.
13. Opgen-Rhein R, Strimmer K: **Inferring gene dependancy networks from genomic longitudinal data: a functional data approach.** *REVSTAT* 2006, **4:**53-65.
14. Meyer P, Kontos K, Lafitte F, Bontempi G: **Information-theoretic inference of large transcriptional regulatory networks.** *EURASIP Journal of Bioinformatics and Systems Biology* 2007.
15. Kaufman L, Rousseeuw P: *Finding Groups in Data: an Introduction to Cluster Analysis Wiley, New-York*; 1990.