BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 18 2007, pages 2407–2414
doi:10.1093/bioinformatics/btm352

*Systems biology*

# LICORN: learning cooperative regulation networks from gene expression data

Mohamed Elati[1,2,*,†], Pierre Neuvial[3], Monique Bolotin-Fukuhara[4], Emmanuel Barillot[3], François Radvanyi[2] and Céline Rouveirol[1,†]

[1]LRI, CNRS UMR 8623, bât 490, Université Paris Sud, 91405 F-Orsay, [2]Institut Curie, CNRS UMR 144, 26 rue d'Ulm, 75248 F-Paris, [3]Institut Curie, Service de Bioinformatique, 26 rue d'Ulm, 75248 F-Paris and [4]IGM, CNRS UMR 8621, bât 400/409, Université Paris-Sud, 91405 F-Orsay, France

Received on April 27, 2007; revised on June 27, 2007; accepted on June 29, 2007

Associate Editor: Martin Bishop

**ABSTRACT**

**Motivation:** One of the most challenging tasks in the post-genomic era is the reconstruction of transcriptional regulation networks. The goal is to identify, for each gene expressed in a particular cellular context, the regulators affecting its transcription, and the co-ordination of several regulators in specific types of regulation. DNA microarrays can be used to investigate relationships between regulators and their target genes, through simultaneous observations of their RNA levels.

**Results:** We propose a *data mining* system for inferring transcriptional regulation relationships from RNA expression values. This system is particularly suitable for the detection of cooperative transcriptional regulation. We model regulatory relationships as labelled two-layer gene regulatory networks, and describe a method for the efficient learning of these bipartite networks from discretized expression data sets. We also evaluate the statistical significance of such inferred networks and validate our methods on two public yeast expression data sets.

**Availability:** http://www.lri.fr/~elati/licorn.html

**Contact:** mohamed.elati@curie.fr

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 INTRODUCTION

Gene regulation in eukaryotes involves many complex mechanisms, most of which are not well understood. With the advent of high-throughput microarray technologies (DeRisi *et al.*, 1997), the expression levels of thousands of genes can be measured simultaneously during various biological processes and for collections of related samples. Considerable effort has been devoted to the analysis of these data sets for the reconstruction of regulatory networks. A family of approaches based on mathematical models of the regulation process has been developed [e.g. Boolean (Liang *et al.*, 1998), Bayesian (Friedman *et al.*, 2000), piecewise-linear (de Jong *et al.*, 2004) and probabilistic Boolean (Bulashevska and Eils, 2005)]. Attempts to learn such models from expression data are hindered by the large number of potential solutions (Chu *et al.*, 2003), and the unrealistically large amount of data required to identify the best solution. In cases of complex formalism for the modelling of regulation, in particular, it has only been possible to reconstruct subnetworks with a few variables. Considerable effort is currently being dedicated to the charting of large-scale gene regulatory networks, relating the expression of a target gene to that of the genes encoding its regulators.

Recent integrative studies have aimed to derive complete yeast gene networks given additional information [e.g. protein–DNA binding from ChIP-chip experiments (Luscombe *et al.*, 2004) or computational analysis of transcription factor binding sites (Middendorf *et al.*, 2004)], with the computational advantage of restricting the number of possible regulators for a given target gene. However, these approaches are difficult to adapt to other organisms, for which the computational detection of *cis*-elements is more difficult, and the experimental detection of binding events is currently limited (e.g. *Homo sapiens*). In contrast, expression data sets are being collected rapidly, and methods based solely on the use of gene expression for network reconstruction are required.

Pe'er *et al.* (2002) have designed the *Minreg* system, a constrained Bayesian network for the reconstruction of large-scale regulatory networks from expression data. The maximal in-degree (i.e. the number of regulators) of target genes and the total number of regulators in the model are limited, so the model focuses on only a small set of global active regulators (AR). The authors made use of these constraints to devise an approximation algorithm for searching for high scoring networks among expression data. The system successfully and robustly identifies the key active regulators, but cannot learn the full detailed network, and may miss interesting regulation relationships: given a current set of active regulators AR, the greedy search of *Minreg* will ignore combinations of co-regulators AR $\cup\{r_1, r_2\}$ if the marginal score values of AR $\cup \{r_1\}$ and AR $\cup \{r_2\}$ are both low, although AR $\cup \{r_1, r_2\}$ may be significant. In such a case, $r_1$ and $r_2$ are said to *cooperate* (Nagamine *et al.*, 2005)—i.e. they act

*To whom correspondence should be addressed.
†Present address: LIPN, CNRS UMR 7030 Institut Galilée - Université Paris-Nord F-93430 Villetaneuse, France.

 

collectively to influence their target genes. Previous computational approaches, due to complexity reasons, have therefore only partly investigated the role of regulator cooperativity. However, such mechanisms have been identified in many organisms (e.g. *Saccharomyces cerevisiae*, *H.sapiens*).

We propose here an original, scalable technique called LICORN (Learning co-operative regulation networks) for deriving cooperative regulations, in which many co-regulators act together to activate or repress a target gene. Many forms of combinatorial logical control may theoretically occur in Boolean or Bayesian models, but we focus here on cooperative regulation patterns that (i) follow the biologically justified activator-repressor model (Woolf and Wang, 2000) (ii) operate on ternary expression level representation (iii) allow for efficient large-scale network computation. LICORN uses an original heuristic approach to accelerate the search for an appropriate structure for the regulation network. It first extracts a global, condensed representation of frequent co-regulator sets using constrained itemset mining techniques (Agrawal *et al*., 1993). From this representation, a limited subset of candidate co-regulator sets is then efficiently associated with each gene. As this candidate subset is modest in size, exhaustive search for the best gene regulatory network can be performed.

In section 2, we will introduce our model of regulation. Section 3 describes a three-step algorithm for inferring complex combinatorial regulation relationships and a procedure for selecting statistically significant relationships. Finally, in Section 4, we evaluate our system on two yeast data sets.

## 2 REGULATION MODEL

We represent the regulatory network architecture as a bipartite graph: the top part contains a small number of regulators $\mathcal{R}$ (an estimated 10% of genes in many organisms); the bottom part contains target genes $\mathcal{G}$ (genes, without regulation activity); edges code for a regulatory interaction between regulators and target genes, each edge being labelled with a regulatory mode (i.e. *activator* or *inhibitor*). Like Pe'er *et al*. (2002) and Segal *et al*. (2003), we use a set of candidate regulatory proteins involved in various aspects of gene regulation, including transcription factors, but also signal transduction molecules, to obtain additional information about regulation by considering the levels of expression of signalling molecules with potential indirect effects on transcription.

As in most previous approaches, we chose to convert transcript levels into ternary expression values: $-1$ (under-expressed), 0 (no change) or 1 (over-expressed). This ternary discretization (see Supplementary Material, Section 1, for more details) is more accurate than a Boolean discretization: it allows for representing both over- and under-expression levels, without making the data representation too complex. Below, the matrix MR stores the expression of regulators in $\mathcal{R}$ and MG the expression of targets in $\mathcal{G}$ for samples from $\mathcal{S}$. For the sake of clarity, we assume that $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{S}$ are arbitrarily ordered and that each target, regulator or sample can be denoted, when it is clear from the context, by its index in $\mathcal{G}$, $\mathcal{R}$ or $\mathcal{S}$.

### 2.1 Local regulatory program

We model a gene regulatory network (GRN) associated with a target gene $g$ as 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 cooperative regulators in $A$ (or $I$), referred to below as *the co-regulator set*, operate collectively as activators or inhibitors of their target gene: for a given sample, they are aggregated in the model through the operator E_AND, which can be interpreted as a logical AND extended to a three-valued logic: E_AND$(X) = -1$ if for all $x_i \in X$, $x_i = 1$, E_AND$(X) = -1$ if for all $x_i \in X$, $x_i = -1$ and E_AND$(X) = 0$ otherwise.

In a simple activator-inhibitor model (Woolf and Wang, 2000), when the level of the activator is high and the level of inhibitor is low, the concentration of the target gene mRNA should be high. Conversely, when the inhibitor concentration is high, and the activator concentration is low, the concentration of the target gene mRNA is low. This qualitative heuristics models expert knowledge concerning regulation control, and was used as the basis for the development of 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 $s$ as described in Figure 1. The vector of $(\hat{g}_s(A, I))_{s\in S}$ is denoted $\hat{g}\ (A,\ I)$.

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 cooperative transcriptional regulation.

### 2.2 Formal problem definition

We can now formally define our inference problem. Given a set of target genes $\mathcal{G}$, a set of regulators $\mathcal{R}$, their discretized expression matrices (MG, MR) over the sample set and an evaluation score $h$, associating a real number with a candidate GRN, our goal is to find, for each target gene $g$, the set of regulators that best explains the level of expression

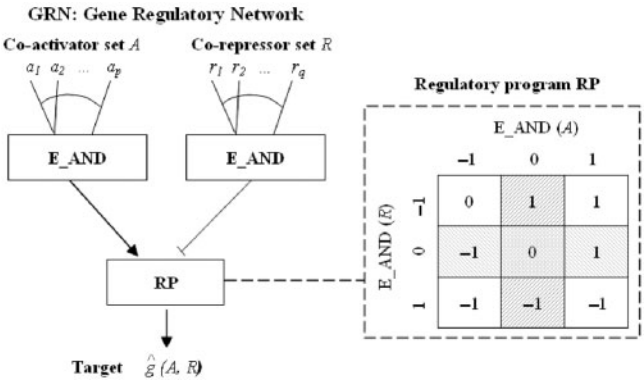


**Fig. 1.** Definition of the regulatory program RP, which can be interpreted as follows: (i) If GRN contains co-activators only, $\hat{g}\ (A,\ I)$ corresponds to the aggregated status of these co-activators. (ii) If GRN contains co-inhibitors only, $\hat{g}\ (A,\ I)$ is the inverse of the aggregated 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.

of $g$. Finding an optimal GRN—a network minimizing the discrepancy between predicted and observed states for a given gene $g$—is NP-hard (Pe'er *et al.*, 2002). We will therefore address the problem by adopting a three-step heuristic approach for the detection of cooperative transcriptional regulation.

## 3 LEARNING ALGORITHM

The first step generates a set of candidate co-regulator sets for all genes of $\mathcal{G}$, such that a candidate co-regulator set is a set of regulators frequently co-expressed in the data. During the second step, for each target gene of $\mathcal{G}$, LICORN efficiently computes a limited set of candidate GRNs and then exhaustively searches for the best one in this set—the activator and inhibitor sets best explaining the target gene status in the sample set. The last step of LICORN is a permutation-based method for the selection of statistically significant GRNs from the inferred GRNs for all target genes.

### 3.1 Mining global candidate co-regulator sets

*3.1.1 Frequent itemset mining* The main purpose of *data mining* (Agrawal *et al.*, 1993) is to reveal the relationships between the attributes or *items* of a sparse binary matrix. Sparseness implies very few co-occurrences of items, therefore, most of the counts in the pairwise marginal would be expected to be 0. It is therefore natural to assume that the frequently co-occurring itemsets contain most of the essential information about the data as a whole. A *frequent itemset* is a set of items that appear together in a set of samples (denoted *support*) with a size higher than a user-defined minimum support threshold.

A classical algorithm for mining frequent itemsets is the *Apriori* algorithm (Agrawal *et al.*, 1993). The algorithm relies upon a simple yet fundamental property of the minimum support constraint, namely anti-monotonicity.

DEFINITION 1. (Anti-monotonic property). *A constraint Const is anti-monotonic (with respect to itemset inclusion) if and only if whenever Const is satisfied by an itemset X, Const is also satisfied by all subsets of X.*

*Apriori* proceeds iteratively, first identifying itemsets of length 1 (1-itemsets). Then, candidate frequent $k$-itemsets are generated by extending the frequent $(k-1)$-itemsets obtained in the previous iteration. This process is repeated until no more candidate itemsets are found. Considering only candidates obtained by extending existing frequent itemsets allows for an optimized search space exploration. Anti-monotonicity of minimum support guarantees that *Apriori* does not miss any frequent itemset when using this optimized candidate generation.

*3.1.2 Candidate co-regulator sets* Global candidate co-regulator sets are mined to compute a condensed representation of the discretized expression matrix MR, by looking for all combinations of co-regulators co-occurring frequently in MR. As our input data is three-valued rather than Boolean, each co-regulator set does not have a single support (implicitly a support for value 1 in binary data), but has a support for each value of interest: 1 (denoting over-expression) and −1 (under-expression).

| | $r_1$ | $r_2$ | $r_3$ | $r_4$ | $r_5$ |
|---|---|---|---|---|---|
| $s_1$ | 1 | 1 | 0 | 0 | −1 |
| $s_2$ | 1 | 1 | 0 | 0 | 0 |
| $s_3$ | 1 | 0 | 0 | 0 | 1 |
| $s_4$ | −1 | 1 | 1 | 0 | 0 |
| $s_5$ | −1 | 1 | 0 | 1 | −1 |
| $s_6$ | −1 | 1 | −1 | 0 | −1 |

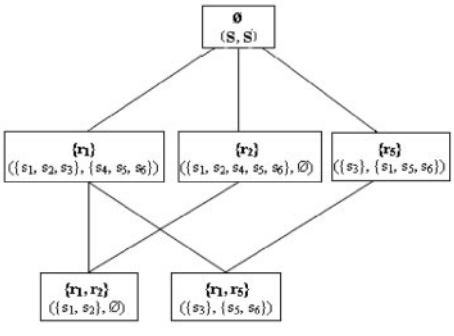


**Fig. 2.** Given the three-valued expression matrix MR on the left, the right-hand part of the figure shows the sub-lattice of frequent co-regulator sets, with a minimum support of 2 (20% of $|\mathcal{S}|$). Each node of the sub-lattice consists of a co-regulator and its 1- and −1-supports.

DEFINITION 2. (Frequent co-regulator set). *Given the three-valued expression matrix MR, a co-regulator set $C \subseteq \mathcal{R}$ and its 1- and −1-supports, denoted $\mathcal{S}^1(C)$, $\mathcal{S}^{-1}(C) \subseteq \mathcal{S}$ C is frequent if and only if max $(|\mathcal{S}^1(C)|, |\mathcal{S}^{-1}(C)|) \geq T_s$, a user-defined minimum support threshold.*

We have implemented an extension of the *Apriori* algorithm that handles in parallel 1 and −1-supports for building the lattice of frequent itemsets, as shown in Figure 2. At this stage, we opt for a relatively small $T_s$ (20% or less), as the aim is to select candidate co-regulator sets with a low level of stringency, as relevant observed regulations may have medium to low frequency in the data set. This step, the most complex in LICORN, is performed only once in the algorithm.

### 3.2 Searching for gene regulatory networks

The sub-lattice CL of global frequent co-regulator sets obtained is now used to generate all possible co-regulator sets for each target gene. The criterion for the involvement of a frequent co-regulator set in the regulatory program of a given target gene is hereafter referred to as the *overlap constraint*. Like the co-regulator sets, each gene $g$ has a 1-support $\mathcal{S}^1(g)$ and a −1-support $\mathcal{S}^{-1}(g)$. The overlap constraint (**cov**) checks the size of the intersection between supports of the target gene and a given candidate co-regulator set.

DEFINITION 3. (Overlap constraint). *Given a co-regulator set C, a gene g, and their respective supports $\mathcal{S}^x(C)$ and $\mathcal{S}^y(C)$ for the states $x, y \in \{-1, 1\}$. C in state x co-varies with g in state y, denoted* **cov**$(\mathcal{S}^x(C), \mathcal{S}^y(g))$ *if and only if* $\frac{|\mathcal{S}^y(g) \cap \mathcal{S}^x(C)|}{|\mathcal{S}^y(g)|} \geq T_o$, *a user-defined minimum overlap threshold.*

$T_o$ is the lower limit of the proportion of samples in which the target $g$ is over- or under-expressed while the co-regulator set $C$ is over- or under-expressed. In other words, it is the conditional probability $\mathbb{P}$ (E_AND $(C) = x \mid g = y$), with $x, y \in \{-1, 1\}$. Note that $T_o$ should exceed 50%, as a small overlap size makes the definition of the regulatory program meaningless. We distinguish co-regulator sets satisfying **cov** for a given target gene according to their roles: a candidate

*co-activator set A* for $g$, is a co-regulator set that positively co-varies with $g$, and a candidate *co-inhibitor set I* negatively co-varies with $g$. $\mathcal{A}(g)$ and $\mathcal{I}(g)$ denote, respectively, all candidate co-activator and co-inhibitor sets for $g$.

$$\mathcal{A}(g) = \left\{A \in \mathrm{CL} | \mathbf{cov}(\mathcal{S}^x(A), \mathcal{S}^x(g)); x \in \{-1, 1\}\right\}$$
$$\mathcal{I}(g) = \left\{I \in \mathrm{CL} | \mathbf{cov}(\mathcal{S}^x(I), \mathcal{S}^{-x}(g)); x \in \{-1, 1\}\right\}$$

CL may be too large and it may therefore be too expensive to generate candidate co-regulator sets of each target gene blindly. As CL is a sublattice partially ordered by $\subseteq$ and given that **cov** is anti-monotonic (see Definition 1) with respect to $\subseteq$, efficient pruning during search is possible: when a coregulator set $C$ does not satisfy **cov**, no superset of $C$ can ever satisfy **cov**. Therefore, large parts of the sublattice need not to be explored.

We can thus compute the set of all candidate GRNs for each target gene $g$ as follows:

$$\mathcal{C}(g) = \{(A,\ I) | A \in \mathcal{A}(g), I \in \mathcal{I}(g) \text{ and } A \cap I = \emptyset\}$$

A candidate GRN for $g$, or a GRN for short, is an element of $\mathcal{C}(g)$.

### 3.3 Scoring gene regulatory networks

In the preceding steps, we have built, for each gene $g$, a relatively small number of candidate regulatory networks, based on the recurrent positive and negative co-variation of candidate co-regulator sets with $g$. We now define a scoring function to compare the different GRNs inferred for a given gene, and to choose the best one. We propose a resampling approach for estimating the statistical significance of each *best candidate* GRN for each target, and a method for determining which candidates are significant enough to be retained.

*3.3.1 Best GRN for each gene* We propose a heuristic measurement for comparing discretized expression profiles, in which each candidate GRN associated with a given gene is scored. As discretized expression values are *ordinal variables*, mean absolute error (MAE) is used to measure distance between gene expression profiles: ideally, over-expressed genes should be closer to genes with no change in expression than to under-expressed genes.

$$h_g(A,\ I) = \mathrm{MAE}(g, \hat{g}(A,\ I)) = \sum_{s \in \mathcal{S}} |g_s - \hat{g}_s(A,\ I)|$$

where $g_s = \mathrm{MG}_{sg}$. Note that $0 \leq \mathrm{MAE} \leq 2$. The best candidate GRN for gene $g$ is then defined as

$$\mathrm{GRN}^*(g) = \underset{(A,I) \in \mathcal{C}(g)}{\mathrm{Argmin}}\ \ h_g(A,\ I)$$

*3.3.2 Significance estimation* Our scoring function $h$ allows us to define a best GRN for each gene, but the scores of the best GRN associated with two different genes may not be directly comparable, as different genes have different probabilities of being under- or over-expressed in the study. Moreover, a GRN is selected because the expression of activator and inhibitor sets co-varies in a recurrent fashion with expression of the gene of interest. Most distances are therefore necessarily small, and a small distance for a given gene does not guarantee that the best GRN is statistically significant. We use statistical hypothesis testing to evaluate how unusually low the score of the best GRN is with respect to the scores that would have been observed if there was no biological relationship between regulators and target gene expression.

The absence of a biological relationship between the target and candidate regulators in the GRN is checked, using random permutations of the samples in the gene expression matrix MG. $B = 1000$ randomized matrices $\mathrm{MG}^{(b)}$ are generated, each corresponding to a particular permutation of the samples. For each permutation $b$, we infer for each gene $g$ a set $\mathcal{C}_b(g)$ of candidate GRNs, and select the best candidate $\mathrm{GRN}_b^*$ from this set, as described above. The statistical significance ($P$-value) of gene $g$ is estimated as the proportion of permutations for which the best score is lower than that obtained with real data:

$$P(g) = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}_{\{h_g(\mathrm{GRN}_b^*) \leq h_g(\mathrm{GRN}^*)\}}$$

*3.3.3 Correction for multiple hypothesis testing* Selecting the genes for which the best candidate GRN is significant based on these $P$-values consists of a multiple hypothesis testing problem, which can be addressed using the false discovery rate paradigm (FDR) introduced by Benjamini and Hochberg (1995). The idea is to control the expected fraction of false positives (i.e. the FDR) among those GRNs selected. We used the FDR control procedure proposed by Benjamini and Yekutieli (2001), which provides strong FDR control for any kind of dependence between test statistics.

## 4 RESULTS AND DISCUSSION

As a proof of concept, we used LICORN for the mining of gene regulatory networks separately on two different gene expression data sets for *S.cerevisiae*. The Gasch data set (Gasch *et al.*, 2000) measures the response of yeast to 173 stress conditions for 6152 genes. The Spellman data set (Spellman *et al.*, 1998) consists of a series of 73 microarray experiments measuring gene expression during the cell cycle for 6178 genes. These two expression matrices were discretized into three states $-1$, 0 and 1: for the Gasch data set, discretized values reflect the expression levels of each gene in each experimental condition; for the Spellman data set, discretized values reflect *expression changes* between consecutive time points. Discretization thresholds, as described in the Supplementary Material (Section 1), were chosen so as to yield balanced frequencies of 1, $-1$ and 0 in the data set. No gene selection was performed at this step: the discretized matrices still contain 6152 and 6178 genes, respectively.

We used a set of 475 regulators compiled by Middendorf *et al.* (2004), consisting of 237 known and putative transcription factors and 250 known and putative signalling molecules, with an overlap of 12 genes of unknown function. A large amount of biological knowledge on yeast is available: function information, contained in the *Saccharomyces Genome Database (SGD)* (Cherry *et al.*, 1998), documented regulations in the *YEASTRACT* database (Teixeira *et al.*, 2006), protein–protein interactions in the *BioGRID* database

(Stark *et al.*, 2006) and data about DNA-binding to transcriptional regulators in ChIP-chip experiments (Harbison *et al.*, 2004; Lee *et al.*, 2002). Thus, the transcriptional networks identified for these two data sets can be checked by comparison with various sources of information.

## 4.1 Performance evaluation

*4.1.1 Objective measurement of prediction performance* Above, we used MAE as a measure of the discrepancy between actual gene expression and the gene expression inferred from the activity of regulators through the GRN. The prediction error of a particular gene *on a given sample set* $\mathcal{T}$ is defined as the MAE within $\mathcal{T}$:

$$e(g) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} |g_t - \hat{g}_t|$$

Averaging individual prediction errors across all selected genes leads to the following *global measure of prediction error* of the model:

$$e = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} e(g)$$

For a prediction measure to be objective, it must be evaluated on a validation set that has not been used to build the predictor. Cross-validation involves partitioning the observed population $\mathcal{S}$ into $K$ subgroups $\mathcal{S}_1, \ldots, \mathcal{S}_K$. For $k = 1, \ldots, K$, the predictor is built on the *training population* $\mathcal{S} \setminus \mathcal{S}_k$, and its performance is evaluated on the *test population* $\mathcal{S}_k$. In practice, 10-*fold cross-validation* ($K = 10$) is often considered, as this method provides a fair estimate of the prediction error at a reasonable computational cost (10 training runs with $\frac{|\mathcal{S}|}{10}$ observations each).

*4.1.2 Results* Using 10-fold cross-validation, we compared four methods: (i) a majority vote in which the predicted gene expression value in the test set is simply the most frequent expression value for this gene in the training set; (ii) a re-implementation of the *Minreg* system, as previously described (Pe'er *et al.*, 2002). We limited running time by filtering out the least informative genes—those remaining almost unchanged in more than 65% of samples—and we have set the maximal in-degree of target genes in the networks to 2 (iii) LICORN algorithm without selection of significant GRNs and (iv) LICORN algorithm with selection of significant GRNs at the 0.05 FDR level.

We used the same 10 cross-validation subgroups to evaluate each of the methods, to facilitate comparisons of performance. The significance of the difference between the prediction rates of two methods on these subgroups was assessed using a paired *t*-test. Cross-validation results are given in Figure 3 for the Gasch data set. Similar results were obtained for the Spellman data set (Supplementary Material, Section 2). It should be noted that the ranking of the methods was the same for all folds, for both data sets. LICORN significantly outperformed *Minreg*, with a *P*-value in paired *t*-tests of $1.6 \times 10^{-8}$ for the Spellman data set, and $6.7 \times 10^{-9}$ for the Gasch data set. Focusing only on those GRNs selected at a given FDR threshold resulted in significant further decrease in MAE: LICORN with FDR < 5% outperformed LICORN, with $P = 1.3 \times 10^{-10}$ for the Spellman data set, and $5.2 \times 10^{-13}$ for the Gasch data set.

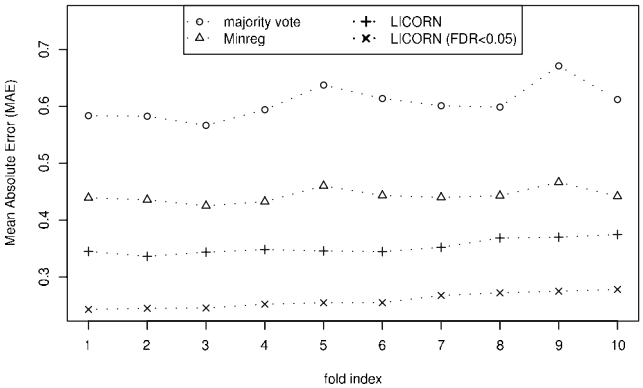


**Fig. 3.** Results of the 10-fold cross-validation on the Gasch data set: comparison of the MAE for all GRNs for the test set for each fold. We recall that $0 \leq \text{MAE} \leq 2$. Folds were sorted in increasing order of MAE for the method 'LICORN (FDR<0.05)'.

## 4.2 Biological analysis

We applied LICORN as described in the Section 3, and retained only those GRNs (gene regulatory networks) identified as significant with a 5% FDR level by the Benjamini and Yekutieli (2001) procedure. We chose the 5% level empirically: it is stringent enough to guarantee that the overwhelming majority of selected GRNs are true discoveries, but relaxed enough for almost half the genes to be retained: for the Gasch data set, 2795 GRNs (of 5703 GRNs) were identified as significant, whereas for the Spellman data set, 2792 GRNs (of 5677 GRNs) were identified as significant. We show some examples of learned GRNs in the Gasch and Spellman data sets in the Supplementary Material, Section 3. We discuss below the structural organization of the learned GRNs. We then provide two kinds of biological evidence to support the inferred GRNs: (i) documented regulation and high-throughput ChIP-chip data sets for confirming transcription factor-target interactions; (ii) protein–protein interactions and functional evaluation for confirming co-regulator cooperativity.

*4.2.1 Overall network structure* Analysis of the structure and organisation of the inferred networks revealed several notable features. In both stress response GRNs (Gasch data set) and cell cycle GRNs (Spellman data set), we found about 10 000 interactions between regulators and target genes. On average, each target is regulated by three regulators in both data sets. Regulators in stress response conditions have a greater influence than cell cycle regulators, as they target more genes simultaneously (on average 30 targets versus 23 targets). We have shown that the distribution of the outgoing connectivity is best approximated by a power-law equation (Supplementary Material, Section 4.2). This allowed us to detect regulator hubs (Lee *et al.*, 2002) with high out-going connectivity (e.g. the heat shock and osmolarity stress regulator *PPT1* regulates 300 target genes). For most regulators in both data sets, a linear dependence was observed between

the number of target genes regulated by a given regulator and the number of co-regulators of that regulator (Supplementary Material, Section 4.3). However, regulators from the Spellman data set have a higher number of co-regulators than do regulators from the Gasch data set (on average 16 co-regulators versus 12 co-regulators), indicating that there are more cooperative associations between regulators in cell cycle GRNs.

All these results, fully detailed in the Supplementary Material (Section 4), are consistent with recent advances (Balaji *et al*., 2006; Guelzim *et al*., 2002; Luscombe *et al*., 2004) concerning the characterization of topological transcriptional network features in yeast and provide the first evidence of the relevance of inferred GRNs.

*4.2.2 Evaluating transcription factor-target interactions* Firstly, as expected, we found that the transcription factors frequently occurring in GRNs inferred from the Gasch data set (e.g. MSN4, XBP1, YAP1, CAD1) played a major role in the response to stress and that many frequent transcription factors in the Spellman data set (e.g. MBP1, FKH1, XBP1, SWI4, ACE2) were involved in the cell cycle. In addition, the *SGD* annotations, concerning the role (activator/inhibitor) of transcription factors, when available, corresponded to the role most frequently assigned within the GRNs inferred, for both data sets (Supplementary Material, Section 5.1). We also showed that LICORN-inferred TF-target interactions have a significant overlap with condition-specific TF-target interactions obtained by Luscombe *et al*. (2004) with their recent integrative method when applied on the same data sets (Supplementary Material, Section 5.2).

The chromatin immunoprecipitation (ChIP) method profiles the binding sites for each transcription factor throughout the entire genome. We compared our results for the Gasch and the Spell-man data sets respectively with those for a stress-response (Harbison *et al*., 2004) and a cell cycle (Lee *et al*., 2002) ChIP-chip data sets. For each condition, we then checked the overlap of both sets of prediction with more than 12 000 demonstrated TF-target relationships described in diverse studies, organized in the *YEASTRACT* knowledge base (Teixeira *et al*., 2006).

In Figure 4, we show the relative overlap between the three sets of identified interactions. For the Gasch data set, 47% of the relationships predicted by LICORN were confirmed by *YEASTRACT*, and 29% of these relationships were also confirmed by the ChIP-chip predictions. Overall, 25% of LICORN predictions were confirmed by ChIP-chip predictions, and 17% of ChIP-chip predictions were confirmed by LICORN. Similar proportions were obtained for predictions based on the Spellman data set: 50% of the relationships predicted by LICORN were confirmed by *YEASTRACT*, and 32% of these relationships were also confirmed by the ChIP-chip predictions. Overall, 26% of the LICORN predictions were confirmed by ChIP-chip predictions and 20% of the ChIP-chip predictions were confirmed by LICORN. The agreement between LICORN results and regulation documented in YEASTRACT is consistent with the agreement between ChIP-chip predictions and this database. For both data sets, ~40% of the regulations learned by LICORN were not supported

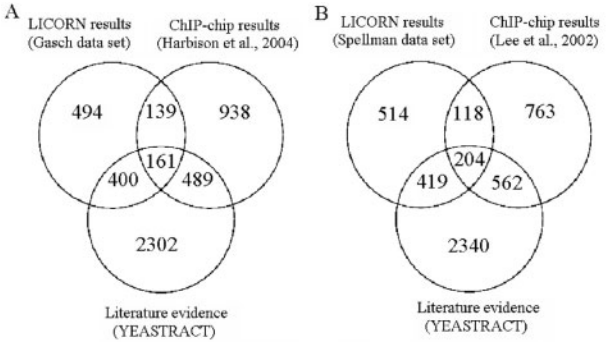


**Fig. 4.** LICORN interaction predictions and ChIP-chip interaction results (with *P*-values <0.001), compared with experimental evidence concerning regulation collected from *YEASTRACT*: (**A**) The number of TF-target interactions, for the 82 TFs shared by LICORN-inferred GRNs from the Gasch data set, the stress-response ChIP data set (Harbison *et al*., 2004) and *YEASTRACT*. (**B**) Number of TF-target interactions for the 69 transcription factors shared by LICORN-inferred GRNs from the Spellman data set, normal growth ChIP data set (Lee *et al*., 2002) and *YEASTRACT*.

by YEASTRACT or by ChIP-chip experiments. There are several possible explanations for this: (i) the usual noise in expression data and the 5% FDR yield a number of false discoveries (ii) large portions of the underlying true network remain unknown and some of these interactions, currently unsupported experimentally, may enable researchers to propose new hypotheses potentially corresponding to new regulation relations.

Evaluation of the non-documented candidate genes under control of a specific TF in GRNs learned from the Gasch data set revealed some biological connections. These connections were obvious for the GAT1 TF which is know to be a transcriptional activator of genes involved in nitrogen catabolite repression (Coffman *et al*., 1995) and associated in our results to several non-documented genes among which DAL3 and YLR164W. Both genes are associated directly or indirectly with nitrogen utilization (Scherens *et al*., 2006; Yoo and Cooper, 1991). More interestingly, YAP1, a transcription factor required for oxidative stress tolerance (Schnell *et al*., 1992) was found to be associated with the EAF3 and TPP1 genes. Both these genes have functions classified as *DNA repair biological process*. EAF3, a chromatin acetylase component (Eisen *et al*., 2000), is probably involved in transcription-coupled repair, a DNA repair mechanism associated with chromatin modifications (Teng *et al*., 2005). TPP1 repairs endogenous damage to double-stranded DNA (Vance and Wilson, 2001). As oxidative stress is known to induce damages in proteins, lipids and DNA, it seems logical that YAP1, in addition to controlling genes necessary to cope with oxygen reactive species, also induces the transcription of genes involved in DNA repair. Finally, this method can be used to identify less direct connections that are nonetheless biologically sound. An example is provided by the BAS1 TF, which is involved in regulating the basal and induced expression of genes of the purine and histidine biosynthesis pathways (Daignan-Fornier and Fink, 1992). Among the non-documented genes predicted to be controlled by BAS1 we found, DPH5, encoding a

**Table 1.** List of the 20 most frequent co-regulator pairs involved in the GRNs learned from the Gasch data set

| Co-regulators | NT | BG | Shared *GO-Slim* terms |
|---|---|---|---|
| LSG1 PPT1 | 93 | N* | Ribosome biogenesis and assembly protein biosynthesis |
| TPK1 TPK2 | 74 | Y | Response to stress, RNA metabolism |
| RAP1 PPT1 | 72 | N | RNA metabolism |
| PDE1 GLC8 | 49 | N | Protein biosynthesis organelle organization and biogenesis |
| LSG1 YVH1 | 44 | Y | Organelle organization and biogenesis ribosome biogenesis and assembly |
| MSN4 TPK1 | 42 | Y | Response to stress |
| XBP1 TOS8 | 39 | N* | Translation |
| GIS1 TPK1 | 35 | Y* | Response to stress |
| PHO2 BAS1 | 32 | Y | RNA metabolism |
| MSN4 USV1 | 30 | N* | Cell wall organization and biogenesis |
| BCY1 TPK1 | 29 | Y | Morphogenesis, response to stress |
| PPT1 YVH1 | 29 | N | Ribosome biogenesis and assembly |
| BMH2 TPK1 | 27 | Y | Protein catabolism, response to stress |
| CLB6 CLB5 | 23 | Y | Cell cycle, DNA metabolism |
| MSN4 TPK2 | 23 | Y | RNA metabolism |
| GCN20 GCN1 | 22 | Y | Sporulation |
| XBP1 USV1 | 20 | N* | Organelle organization and biogenesis |
| FAR1 CLN2 | 19 | Y | Cell wall organization and biogenesis protein modification |
| PPT1 RAS1 | 19 | N | Response to stress |
| BMH2 BMH1 | 19 | Y | Carbohydrate metabolism |

For each co-regulator pair, the number of targets (NT) and the existence or otherwise of known protein–protein interactions in the *BioGRID* database (BG) are indicated, together with the list of GO-Slim terms significantly shared by more than 40% of their target genes. (*) indicates that the co-regulators found were identified in the results of Segal *et al.* (2003).

methyltransferase required for the synthesis of a modified histidine residue (Mattheakis *et al.*, 1992) and TRM1, encoding an N2,N2-dimethylguanosine-specific tRNA methyltransferase (Ellis *et al.*, 1986). Although not directly involved in the purine or histidine biosynthetic pathways, their functions depend on these pathways. It is therefore reasonable to assume that they may be co-regulated with the well-identified BAS1 target genes, consistent with the coupling of main pathways with secondary ones. These interesting new possibilities require experimental testing.

*4.2.3 Evaluating candidate co-regulators* We evaluated cooperativity between the co-regulators inferred by LICORN, based on two assumptions: (i) the existence of protein–protein interactions between co-regulators implies participation in the same regulatory mechanism, and (ii) targets contributing to similar biological process are regulated by the same control mechanism. In Table 1, we have listed the 20 most frequent regulator pairs in co-activator or co-inhibitor sets in GRNs learned from the Gasch data set. For each co-regulator pair, we have checked whether the two co-regulators are known to interact (protein or genetic interaction), based on information in the *BioGRID* database. In total, 60% of the co-regulator pairs in the list were reported to interact in *BioGRID*. This proportion is high and confirms the validity of LICORN predictions, with *P*-value close to 0 ($<10^{-15}$). For all co-regulator pairs, we found GO-Slim terms (high-level GO terms that represent the major biological processes in *S.cerevisiae*) significantly shared by at least 40% of the target genes, using the *GO-Slim* mapping tool of the *SGD* (Cherry *et al.*, 1998), demonstrating the functional robustness of the co-regulators inferred by LICORN.

The pairs of co-regulators in Table 1 include the known heat shock and osmolarity stress regulators TPK1, PPT1 and USV1, which occur at high frequency. This observation correlates well with the results obtained by Segal *et al.* (2003) and Middendorf *et al.* (2004) for the Gasch data set. Segal *et al.* (2003) identified these proteins as the master regulators for this data set, as they occurred in more than 5 of the 50 inferred modules of co-regulated genes and their regulators. Four of the eight co-regulator pairs not found in *BioGRID* were identified by Segal *et al.* (2003). Moreover, Segal *et al.* (2003) did not identify some of our confirmed co-regulators (e.g. TPK1-TPK2, GCN20-GCN1 and CLB6-CLB5), as in cases in which several regulators are involved in the same regulatory event, this method typically identifies only one representative of the group.

Finally, we obtained similar results for the list of the 20 most frequent co-regulator pairs involved in the GRNs learned from the Spellman data set (see Supplementary Material, Section 6.1). We also found significant agreement between the extent of cooperative associations between regulators and physical interactions between regulatory proteins during the yeast cell cycle, as reported by de Lichtenberg *et al.* (2005). More details are given in the Supplementary Material (Section 6.2). These results confirm those of recent studies (Balaji *et al.*, 2006; Nagamine *et al.*, 2005) connecting regulator cooperativity and protein–protein interactions.

## 5 CONCLUSION

We provide here a model for cooperative regulation and an algorithm, LICORN, for the inference of cooperative regulation from gene expression data. We used a permutation-based procedure selecting the most statistically significant regulation networks and have shown that this selection step improves prediction performance in a 10-fold cross-validation framework. Moreover, validation on two yeast data sets showed that LICORN was a powerful *data mining* tool for the analysis of gene expression. The results obtained with this algorithm were consistent with published experimental results. The labelled relationships (activation/inhibition) found with our method do not require post-treatment analysis for interpretation, unlike the combinatorial interactions learned with Bayesian network algorithms (Friedman *et al.*, 2000; Pe'er *et al.*, 2002).

Cooperative regulation patterns cannot be identified by clustering or pairwise methods (Woolf and Wang, 2000), and are only partly revealed by constrained Bayesian or decision tree-based techniques, such as those used in previous studies (Middendorf *et al.*, 2004; Pe'er *et al.*, 2002; Segal *et al.*, 2003). Rather than selecting regulators independently, LICORN efficiently reduces the search space for the candidate regulators of the targets to the sub lattice of frequent co-regulators. This decreases the number of regulator combinations to be evaluated, and LICORN does not require strong a priori selection criteria based on uncertain or incomplete information, such as

DNA-binding data (Middendorf *et al*., 2004). LICORN thus speeds up the inference of gene regulatory networks including co-activator and co-inhibitor sets. LICORN also avoids the use of 'gene modules' (Segal *et al*., 2003) for factorizing the search for the best regulation network. Modularity may be an organizing principle of regulatory networks, but it may be too coarse for the learning of specific regulatory programs (LICORN learns a regulation network for each gene). Instead, partial overlap of the regulator sets for a set of target genes, once inferred, can be used as an alternative measurement of the distance between genes.

Future work should focus on extending the LICORN model, to increase accuracy and generalization. For instance, LICORN can be extended to the learning of other classes of combinatorial regulation, in which several co-activator or co-inhibitor sets may function independently, or in which regulatory relationships may link the regulators themselves. This requires care, to avoid problems of over-fitting, given the small size of the training sets available. Finally, the gene regulatory networks learned by LICORN from expression data can be enriched by integrating various gene networks from diverse data sources (motif networks, ChIP-chip data, protein–protein interactions, functional category, etc.). This suggests the use of a logical representation for gene networks, and the use of adapted integrative algorithms, such as those developed in *Inductive Logic Programming* (Fröhler and Kramer, 2006).

## ACKNOWLEDGEMENTS

We thank Ch. Froidevaux for her constant support, Ch. Battail for fruitful discussions and the anonymous referees for their pertinent suggestions. We also thank J. Sappa from Alex Edelman and Associates for careful reading of the manuscript. This work was supported by the CNRS, the Institut Curie, the Plan Pluri-Formation Bioinformatique et Génomique and the IFR Génome. M. Elati and F. Radvanyi are members of the Equipe Oncologie Moléculaire, labellisée par La Ligue Nationale Contre le Cancer. M.E. was supported by a fellowship from the French Ministry of Foreign Affairs. P.N. was supported by a fellowship from the association Courir pour la vie, courir pour Curie.

*Conflict of Interest*: none declared.

## REFERENCES

Agrawal,R. *et al*. (1993) Mining association rules between sets of items in large databases. In *Proceedings of the International Conference on Management of Data,* pp.207–216.

Balaji,S. *et al*. (2006) Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J. Mol. Biol.*, **360**, 204–212.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, **57**, 289–300.

Benjamini,Y. and Yekutieli,Y. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1198.

Bulashevska,S. and Eils,R. (2005) Inferring genetic regulatory logic from expression data. *Bioinformatics*, **21**, 2706–2713.

Cherry,J. *et al*. (1998) SGD: Saccharomyces Genome Database. *Nucleic. Acids Res.*, **26**, 73–79.

Chu,T. *et al*. (2003) A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays. *Bioinformatics*, **19**, 1147–1152.

Coffman,J. *et al*. (1995) Genetic evidence for Gln3p-independent, nitrogen catabolite repression-sensitive gene expression in Saccharomyces cerevisiae. *J. Bacteriol.*, **177**, 6910–6918.

Daignan-Fornier,B. and Fink,G. (1992) Coregulation of purine and histidine biosynthesis by the transcriptional activators BAS1 and BAS2. *Proc. Natl Acad. Sci.*, **89**, 6746–6750.

de Jong,H. *et al*. (2004) Qualitative simulation of genetic regulatory networks using piecewise-linear models. *Bull. Math. Biol.*, **66**, 301–340.

de Lichtenberg,U. *et al*. (2005) Dynamic complex formation during the yeast cell cycle. *Science*, **307**, 724–727.

DeRisi,J. *et al*. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Eisen,A. *et al*. (2000) The yeast NuA4 and Drosophila MSL complexes contain homologous subunits important for transcriptional regulation. *J. Biol. Chem.*, **276**, 3484–3491.

Ellis,S. *et al*. (1986) Isolation and characterization of the TRM1 locus, a gene essential for the N2,N2-dimethylguanosine modification of both mitochondrial and cytoplasmic tRNA in Saccharomyces cerevisiae. *J. Biol. Chem.*, **261**, 9703–9709.

Friedman,N. *et al*.Using bayesian network to analyze expression data. *Comput. Biol.*, **7**, 601–620.

Fröhler,S and Kramer,S. (2006) Logic-based information integration and machine learning for gene regulation prediction. In *Proceedings of the 9th International Conference on Molecular Systems Biology*.

Gasch,A. *et al*. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.

Guelzim,N. *et al*. (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet.*, **31**, 60–63.

Harbison,C. *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Lee,T.I. *et al*. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.

Liang,S. *et al*. (1998) Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Pacific Symposium in Biocomputing*, 18–29.

Luscombe,N.M. *et al*. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**, 308–312.

Mattheakis,L. *et al*. (1992) DPH5, a methyltransferase gene required for diphthamide biosynthesis in Saccharomyces cerevisiae. *Mol. Cell. Biol.*, **12**, 4026–4037.

Middendorf,M. *et al*. (2004) Predicting genetic regulatory response using classification. *Bioinformatics*, **20**, 232–240.

Nagamine,N. *et al*. (2005) Identifying cooperative transcriptional regulations using protein-protein interactions. *Nucleic. Acids Res.*, **33**, 4828–4837.

Pe'er,D. *et al*. (2002) Minreg: inferring an active regulator set. *Bioinformatics*, **18**, 258–267.

Scherens,B. *et al*. (2006) Identification of direct and indirect targets of the gln3 and gat1 activators by transcriptional profiling in response to nitrogen availability in the short and long term. *FEMS Yeast Res.*, **6**, 777–791.

Schnell,N. *et al*. (1992) The par1 (yap1/snq3) gene of saccharomyces cerevisiae, a c-jun homologue, is involved in oxygen metabolism. *Curr. Genet.*, **21**, 269–273.

Segal,E. *et al*. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34**, 166–176.

Spellman,P. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Stark,C. *et al*. (2006) BioGRID: a general repository for interaction datasets. *Nucleic. Acids Res.*, **34**, 535–539.

Teixeira,M. C. *et al*. (2006) The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic. Acids Res.*, **34**, 446–451.

Teng,Y. *et al*. (2005) Histone acetylation, chromatin remodelling, transcription and nucleotide excision repair in s. cerevisiae: studies with two model genes. *DNA Repair*, **4**, 870–883.

Vance,J. and Wilson,T. (2001) Uncoupling of 3' phosphatase and 5' kinase functions in budding yeast: characterization of S. cerevisiae DNA 3' phosphatase (TPP1). *J. Biol. Chem.*, **276**, 15073–15081.

Woolf,P. and Wang,Y. (2000) A fuzzy logic approach to analyzing gene expression data. *Physiol. Genomics*, **3**, 9–15.

Yoo,H. and Cooper,T. (1991) The ureidoglycollate hydrolase (dal3) gene in saccharomyces cerevisiae. *Yeast*, **7**, 693–698.