An Information Theoretic Exploratory Method for Learning Patterns of Conditional Co-Expression in Gene Microarray Data

Riccardo Boscolo, James C. Liao, Vwani P. Roychowdhury

Abstract

In this article, we introduce an exploratory framework for the detection of patterns of conditional co-expression in gene expression data. The main idea behind the proposed approach is that of modeling gene expression levels as random variables whose statistical dependence appears only conditionally on the values assumed by a separate set of indicator variables. The method is non-parametric and it is based on the concept of statistical *co-information*, which, unlike conventional correlation based techniques, is not restricted in scope to linear conditional dependency patterns. A moment based approximation of the co-information measure is derived that efficiently gets around the problem of estimating high-dimensional multi-variate probability density functions from the data, a task usually not viable due to the intrinsic sample size limitations that characterize expression data measurements. By applying the proposed exploratory method, we analyzed a whole genome microarray assay of the eukaryote *Saccharomices cerevisiae* and were able to detect statistically significant patterns of conditional co-expression. A selection of such interactions that carry a meaningful biological interpretation are discussed.

I. Introduction

DNA Microarray technology has revolutionized the field of life science with the introduction of an experimental technique allowing the simultaneous monitoring of the expression levels of all genes in a particular organism. Although DNA microarray technology has resulted in

Riccardo Boscolo and Vwani P. Roychowdhury are with the Electrical Engineering Department, University of California Los Angeles (e-mail: {riccardo,vwani}@ee.ucla.edu).

James C. Liao is with the Department of Chemical Engineering, University of California Los Angeles (e-mail: liaoj@ucla.edu).

the breakthrough capability of obtaining high-throughput gene expression measurements, the statistical analysis of such expression data presents several challenges. The noise component in the data can be significantly large: a substantial variability in the experimental data can result from assaying cell cultures grown at different times, from using different types of microarray chips and optical readers, or simply from different laboratory personnel performing the experiments. Only a few of these sources of variability can be tightly controlled. Moreover, the sampling characteristics of a time-course experiment are generally poor and characterized by frequent missing values.

The ultimate goal of any statistical analysis tool for gene expression data is to detect patterns of interactions (also known as *co-expressions*) between genes as well as to assess their statistical significance. Several attempts aiming at adapting well-known statistical learning frameworks to gene expression data, such as Bayesian Networks [5], Support Vector Machines (SVM) [6], K-means clustering or Self-Organizing Maps (SOM) [9], have led to results that are generally difficult to interpret from a biological standpoint. For this reason, members of the biology community tend to resort to simple analysis tools, which are widely accepted mainly because of their straightforward biological interpretation: an example is given by correlation analysis and its extension to gene clustering by hierarchical agglomeration [3].

The motivation behind our work stems from the following considerations:

- One of the primary goals of any exploratory method suited for the analysis of biological signals should be that of estimating the information content in the data. In other terms, the scope of the learning framework should be tailored to the actual amount of information that the data can provide, when experimental parameters such as noise level or sampling characteristics are taken into consideration.
- The analysis should retain a certain degree of biological significance, if not being directly biologically inspired. Although the derivation of causal relationships purely from data analysis results is not generally feasible, one should still pursue the discovery of patterns of interaction that can be associated to or explained by known biological mechanisms.
- The existence of patterns of co-dependency in gene expression data that can be systematically extracted by an unsupervised exploratory technique has not been clearly established. A recently proposed approach [10], suggests that patterns of linear pairwise correlation between functionally related genes often appear only conditionally on the value of a third

scouting gene. The extension of this and similar approaches to non-linear dependency structures is still an open problem.

- The computational tractability of the analysis is constrained by two key factors: the first is related to the sampling characteristics of the data. Such issue intrinsically limits, for example, the ability of estimating reliably high-dimensional probability density functions of the variables in the model. Secondly, because of the combinatorial nature of any unsupervised exploratory approach that aims at unveiling patterns of conditional dependency, constraints on the model size are necessary in order to limit the computational cost associated with the procedure.

The proposed gene expression analysis framework is based on the concept of *co-information*, a measure of statistical conditional dependence that is *non-parametric* and it is not restricted to linear models. In section II, the conditional co-expression model is introduced and certain properties of the proposed cost function are elucidated. In particular, we demonstrate that for the special case involving networks of three nodes (where each node represents a random variable), the co-information measure is equivalent to the residual mutual information, computed as the difference in mutual information between any two nodes with and without conditioning on the third node. The application of the resulting exploratory method to gene expression data is detailed in Section III, where certain issues related to the practical implementation of the algorithm are also addressed. A moment based approximation of the co-information measure is derived that efficiently solves the problem of estimating high-dimensional multi-variate probability density functions from the data. In Section IV, the results we obtained by analyzing a whole genome microarray assay of *Saccharomyces cerevisiae* [4] are presented.

II. Conditional Co-expression Model

The idea behind the proposed approach consists of identifying groups of genes that are co-expressed *only conditionally on the expression level of other genes*.

Figure 1 shows the time-courses of three genes whose expression levels were generated from synthetic data in order to exemplify a case of conditional co-expression. The scatter plots of the three genes (shown in Figure 2) does not present any significant pattern of dependency. When the points in the scatter plot of gene_x versus gene_y are labeled according to the value of

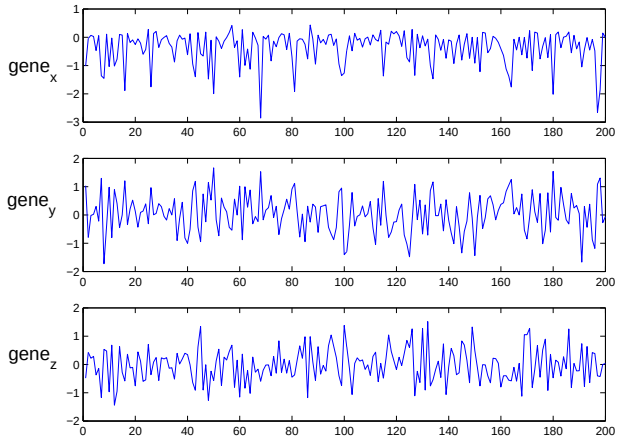

Fig. 1. The plot shows the expression time-courses of three hypothetical genes, generated from synthetic data.

gene$_z$[1], however, a clear dependency structure appears. This is shown in Figure 3: when gene$_z$ is significantly up-regulated, the expression level of the remaining two genes appear to positively correlated. Clearly, gene$_x$ and gene$_y$ follow a specific pattern of co-expression when gene$_z$ is down-regulated or near the reference level, but such pattern changes significantly when gene$_z$ is up-regulated. Although gene$_z$ might not be directly controlling the mechanism responsible for the change in behavior, it plays the role of indicator of the underlying biological process determining the observed co-expression pattern. The problem can be thus posed as follows: how can one choose a measure of statistical dependency that is capable of detecting conditional co-expressed genes, and how can one identify a set of genes whose expression levels are indicator functions of a significant change in the transcriptional regulation mechanisms within the cell.

A. Conditional Mutual Information as a Measure of Conditional Co-expression

In order to detect patterns of conditional co-expression in the data, one must choose a measure of statistical dependence. Although conditional correlation is probably the simplest such measure, it is only suitable for detecting patterns of linear dependency [10]. A natural measure of statistical dependence that does not make any assumptions on the linearity of the model is given by the mutual information [2]. Let us start by considering the definition of mutual information between two random variables x_1 and x_2 conditioned on a third random variable y:

[1]In this case the expression levels of gene$_z$ are quantized according to three discrete levels.

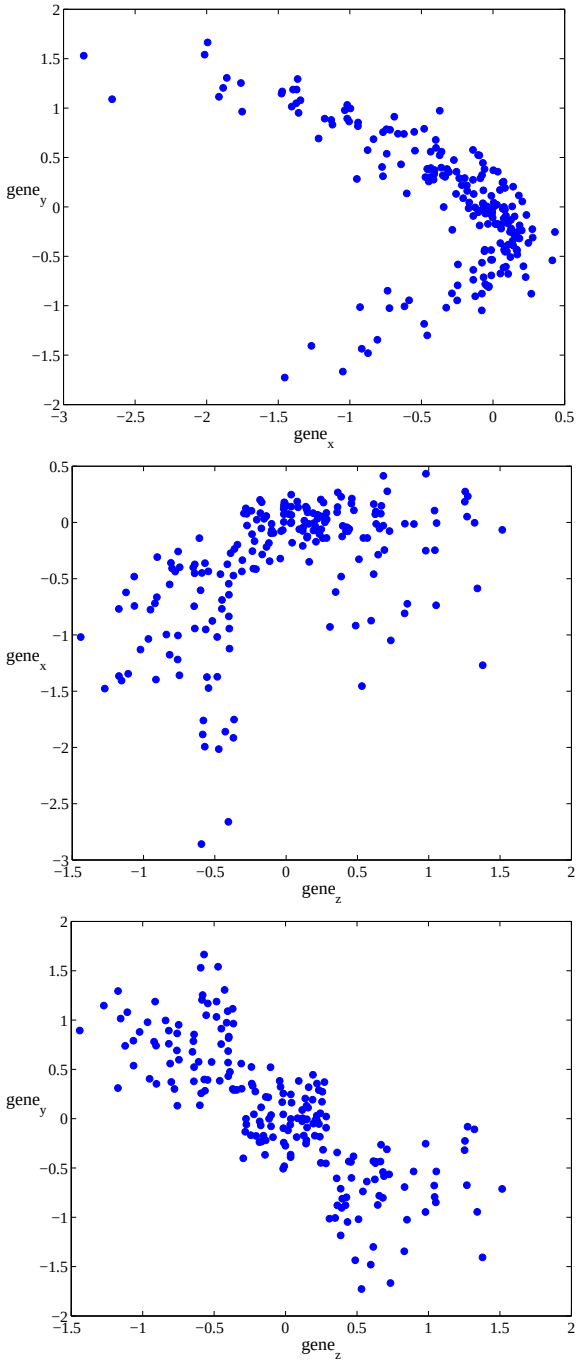

Fig. 2. Scatter plots of the synthetically generated expression levels for the three genes.

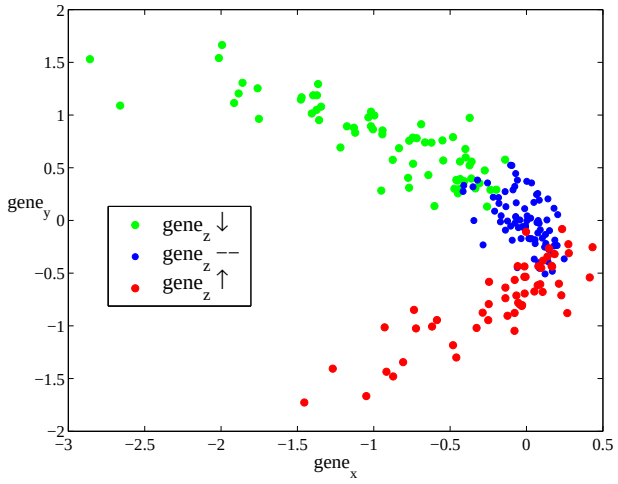

Fig. 3. The co-expression pattern between gene$_x$ and gene$_y$ appears when conditioning on gene$_z$.

$$I(x_1, x_2|y) \triangleq D\left(p(x_1, x_2|y) \,\middle\|\, p(x_1|y)p(x_2|y)\right), \tag{1}$$

where D is the Kullback-Leibler distance or relative entropy, defined as:

$$D(q||r) \triangleq E_q\left[\log \frac{q(u)}{r(u)}\right]. \tag{2}$$

It can be shown [2] that the relative entropy is always non-negative and is zero if and only if $q = r$ almost everywhere.

In the case of continuous random variables (1) can be expressed as:

$$I(x_1, x_2|y) = E_{p(x_1,x_2,y)}\left[\log\left(\frac{p(x_1, x_2|y)}{p(x_1|y)p(x_2|y)}\right)\right] \tag{3}$$
$$= \int_{-\infty}^{\infty} p(x_1, x_2, y) \log\left(\frac{p(x_1, x_2|y)}{p(x_1|y)p(x_2|y)}\right) dx_1 dx_2 dy.$$

This definition can be extended to the mutual information of M random variables $\mathbf{x} = [x_1, \ldots, x_M]^T$, conditioned on a separate set of L variables $\mathbf{y} = [y_1, \ldots, y_L]^T$ as follows:

$$I(\mathbf{x}|\mathbf{y}) = E_{p(\mathbf{x},\mathbf{y})}\left[\log \frac{p(\mathbf{x}|\mathbf{y})}{\prod_{i=1}^{M} p(x_i|\mathbf{y})}\right] \quad (4)$$

$$= \int_{-\infty}^{\infty} p(\mathbf{x},\mathbf{y}) \log \frac{p(\mathbf{x}|\mathbf{y})}{\prod_{i=1}^{M} p(x_i|\mathbf{y})} d\mathbf{x}d\mathbf{y}, \quad (5)$$

This expression provides us with a measure of the expected mutual information of $\mathbf{x}$ conditionally on the value of $\mathbf{y}$. Evidently, when $\mathbf{x}$ and $\mathbf{y}$ are statistically independent, we have trivially that:

$$I(\mathbf{x}|\mathbf{y}) = I(\mathbf{x}) \int_{-\infty}^{\infty} p(\mathbf{y}) d\mathbf{y} = I(\mathbf{x}). \quad (6)$$

Recalling that we are after certain structure in the data that appears only under conditioning, this result prompts us with the idea of adopting the following cost function:

$$\boxed{\mathcal{L}(\mathbf{x}|\mathbf{y}) \stackrel{\triangle}{=} I(\mathbf{x}|\mathbf{y}) - I(\mathbf{x})} \quad (7)$$

Clearly, we have that $\mathcal{L}(\mathbf{x}|\mathbf{y}) = 0$ when $\mathbf{x}$ and $\mathbf{y}$ are independent. In this case, even if a cluster of genes possesses a high information content, *i.e.* $I(\mathbf{x})$ is large, such structure appears regardless of the set of conditioning variables. On the other hand, $\mathcal{L}(\mathbf{x}|\mathbf{y})$ is a large positive number when the information content is significantly increased under conditioning. This is the case of interest in our framework. Notice that the quantity in (7) might assume negative values and it is not lower-bounded in general.

B. Certain Theoretical Properties of the Cost Function

In this section, we derive certain properties of the cost function (7). In particular, we show that the proposed cost function is equivalent to the measure of *co-information* [1] when restricted to networks involving three nodes. Consider a simple network with a single parent node x_0 and two children nodes x_1 and x_2. The conditional information content of this network, according to (7), given by:

$$\mathcal{L}(x_1, x_2|x_0) = I(x_1, x_2|x_0) - I(x_1, x_2). \quad (8)$$

Let us consider the first term on the right hand side of (8):

$$
\begin{aligned}
I(x_1, x_2|x_0) &= E_{p(x_0,x_1,x_2)}\left[\log \frac{p(x_1, x_2|x_0)}{p(x_1|x_0)p(x_2|x_0)}\right] && (9)\\
&= \int_{-\infty}^{\infty} p(x_0, x_1, x_2) \log \frac{p(x_1, x_2|x_0)p(x_0)^2}{p(x_1|x_0)p(x_2|x_0)p(x_0)^2}\, dx_0 dx_1 dx_2\\
&= \int_{-\infty}^{\infty} p(x_0, x_1, x_2) \log \frac{p(x_0, x_1, x_2)p(x_0)}{p(x_0, x_1)p(x_0, x_2)}\, dx_0 dx_1 dx_2\\
&= -H(x_0, x_1, x_2) - H(x_0) + H(x_0, x_1) + H(x_0, x_2), && (10)
\end{aligned}
$$

where:

$$H(x) = -\int p(x) \log p(x) dx, \tag{11}$$

is the differential entropy of the random variable x. Therefore, recalling that $I(x_1, x_2) = H(x_1) + H(x_2) - H(x_1, x_2)$, we have that (8) is equal to:

$$
\begin{aligned}
\mathcal{L}(x_1, x_2|x_0) = &-H(x_0) - H(x_1) - H(x_2) + H(x_0, x_1)\\
&+H(x_0, x_2) + +H(x_1, x_2) - H(x_0, x_1, x_2). && (12)
\end{aligned}
$$

From this expression we conclude that when considering a simple network with one conditioning node and two children nodes, the cost function (7) is indeed equal to the negative *co-information* between the three random variables. The general definition of co-information of N random variables is given in [1]:

$$\mathcal{C}(\mathbf{x}) = \sum_{E_j \subseteq E_N} q_j H(\mathbf{x}_{E_j}), \tag{13}$$

where E_j is the power set of j and q_j is the Möbius inversion function, defined as:

$$q_j = -(-1)^{|E_j|} = \begin{cases} 1 & \text{if } |E_j| \text{ is odd} \\ -1 & \text{if } |E_j| \text{ is even} \end{cases}, \tag{14}$$

and $|E_j|$ is the cardinality of E_j. The co-information provides a measure of the total information content shared by all the random variables, unlike the conventional mutual information which

includes all the information shared by the variables two at the time. Therefore, maximizing (7) is equivalent to seeking clusters whose representatives simultaneously share the least amount of information between each other. Equivalently:

$$\max_{x_0,x_1,x_2} \mathcal{L}(x_1, x_2|x_0) = \min_{x_0,x_1,x_2} \mathcal{C}(x_0, x_1, x_2). \tag{15}$$

Notice that expression (12) is not altered if we exchange the variables x_0, x_1, or x_2. Hence, it holds that:

$$\mathcal{L}(x_1, x_2|x_0) = \mathcal{L}(x_0, x_2|x_1) = \mathcal{L}(x_0, x_1|x_2) \tag{16}$$

Thus, the information content of the sub-network does not change if we exchange one of the children nodes with the parent node.

III. METHOD

When designing a practical implementation of the algorithm seeking clusters that maximize the cost function (7), certain issues must be taken into account:

- A direct evaluation of the cost function (7) requires an estimate of the multi-variate probability density function of all the N variables included in the cluster. Although for small dimensional problems methods for estimating directly the joint probability density function (pdf) have been developed [11], for higher dimensional problems the direct estimation of the pdf is usually not feasible.
- The noise level in the data might significantly limit the number of parameters that can be learned with a certain degree of accuracy.
- Current microarray based experimental procedures are affected by inherent limits in the number of expression samples that can be measured in a given interval of time. Hence, the sampling characteristics of any microarray assay are generally poor in the time-domain. This issue imposes a further limitation on the capability of estimating joint probability density functions.
- For a sub-network of a given size, the optimization of the cost function (7) requires a search through all possible combinations of nodes chosen among the set of genes included

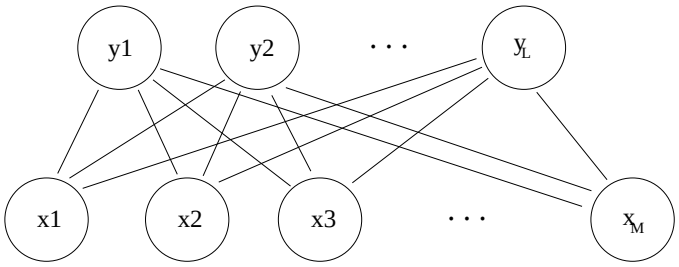

Fig. 4. Cluster of genes composed of L conditioning genes and M children nodes. This cluster represents the generic sub-network explored to identify conditional structure.

in the experiment. We will show that such number of combinations can be quite large if the parameters of the search algorithm are not chosen properly, quickly yielding to an intractable computational cost.

A. Combinatorial Optimization Approach

The goal is to identify a list of sub-networks that yield large values of the cost function (7). Such goal can be achieved by exhaustively selecting sub-networks (see Figure 4) consisting of all possible combinations of L genes as conditioning variables (which we will refer to as parent nodes), and all possible combinations of M genes among the remaining ones as conditioned variables (also known as children nodes), and by evaluating the corresponding value of the cost function (7). Clearly, the cost of this combinatorial approach increases quite rapidly with the total number of genes assayed in the experiment, as a non-linear function of M and L. When a total number of N gene expression profiles are measured in the experiment, the total number of possible sub-networks with L parent nodes and M children nodes is given by the following expression:

$$\mathcal{K}(N, M, L) = \begin{pmatrix} N \\ L \end{pmatrix} \begin{pmatrix} N-L \\ M \end{pmatrix} \tag{17}$$

$$= \frac{N!}{M!L!(N-L-M)!}. \tag{18}$$

For example, when dealing with $N = 2,000$ genes, a choice of $L = 3$ and $M = 5$ will result in $3.5 \cdot 10^{23}$ possible combinations! In general, for small values of M and L, we have that:

$$\mathcal{K}(N, M, L) \approx \mathcal{O}(N^{M+L}). \tag{19}$$

Hence, unless a technique is devised that allows for efficient pruning of non-informative clusters, the problem will result computationally tractable only for very small values of M and L. In addition, as it will be discussed more in details in the next section, for large values of M and L we will unavoidably incur in the problem of having to estimate high-dimensional multivariate statistics of the data, thus requiring a significantly large number of samples in order to get a robust estimate.

These constraints clearly suggest that a simple framework in which a sub-network involving only three genes (one parent node and two children nodes) should be the subject of an initial investigation and validation of the proposed approach. From the symmetric expression of the cost function given in (12), it is possible to show that the computational complexity associated with evaluating the co-information content of each possible sub-network, when $L = 1$ and $M = 2$ simplifies as:

$$\begin{aligned} \mathcal{K}(N, 2, 1) &= N(N-1)(N-2) + N(N-1) + N \\ &= N^3 - 2N^2 + N \qquad (20) \\ &= \mathcal{O}(N^3), \qquad (21) \end{aligned}$$

for a total number of N genes assayed. As an example, when $N = 2,000$, approximately $8 \cdot 10^9$ possible combinations need to be considered, and the corresponding cost function evaluated. This kind of task can be completed in a reasonable amount of time by any modern off-the-shelf single-processor machine. It is also clear that the algorithm could be easily parallelized to run on clusters of processors, since the evaluation of the cost function for a given sub-network is an independent task.

B. A Moment Based Approximation of the Mutual Information

The expression of the cost function given in (12) suggests that some kind of estimate of the multivariate joint probability density function of the three variables in the cluster is required

in order to evaluate the corresponding entropies. However, considering that the typical experimental setting in DNA microarray assays results in a limited number of samples per gene, such poor sampling properties generally discourage the use of standard probability density function estimators such as parametric models or kernel methods. Therefore, in the design of a practical implementation of the principle (15), we opted for the use of a moment based approximation of the information theoretical quantities involved in the calculation of the cost function.

Let us first examine equation (12), which is used as a starting point to define our working approximation:

$$\begin{aligned} \mathcal{L}(x_0, x_1, x_2) &= I(x_1, x_2|x_0) - I(x_1, x_2) \\ &= H(x_1|x_0) + H(x_2|x_0) - H(x_1, x_2|x_0) \\ &\quad -H(x_1) - H(x_2) + H(x_1, x_2). \end{aligned} \tag{22}$$

This expression suggests that a method to compute univariate and bivariate entropies must be devised. A moment based approximation of the univariate entropy is obtained by approximating the marginal probability density function of each variable with its Gram-Charlier expansion [12]. Recalling that the entropy is shift invariant, we can assume that all the sample data has been centered. Moreover, since it holds that $H(ax) = H(x) + \log(|a|)$, where a is a deterministic constant parameter, we can re-scale each variable to be unit-variance and compute the entropy estimate as follows:

$$H(x_i) = H(\tilde{x}_i) + \log(\sigma_i), \qquad i = 1, 2 \tag{23}$$

where σ_i^2 is the variance of x_i, and $\tilde{x}_i \triangleq x_i/\sigma_i$ is unit-variance. A Gram-Charlier approximation of $p_{\tilde{x}_i}(u)$, including moments up to the fourth order is given by the following expression:

$$p_{\tilde{x}_i}(u) = g(u)\left(1 + \kappa_{3,i}H_3(u)/6 + \kappa_{4,i}H_4(u)/24\right) \tag{24}$$

where $H_3(u)$ and $H_4(u)$ are the $3rd$ and $4th$ order Chebyshev-Hermite polynomial [8], respectively, $g(u)$ is the zero-mean, unit-variance, normal probability density function, and $\kappa_{3,i}$ and $\kappa_{4,i}$ are the third and fourth order cumulants of $\tilde{x}_i$. For a zero mean, unit-variance random variable, these can be computed as follows:

$$\kappa_{3,i} = E[\tilde{x}_i^3] \tag{25}$$

$$\kappa_{4,i} = E[\tilde{x}_i^4] - 3. \tag{26}$$

By substituting the approximation of $p_{\tilde{x}_i}(u)$ considered in (24) in the definition of entropy, we can compute the following approximation:

$$H(\tilde{x}_i) \approx \frac{1}{2}\log(2\pi e) - (\kappa_{3,i}^2 + \kappa_{4,i}^2/4)/12. \tag{27}$$

which is consistent with the fact that the entropy of a random variable with a given variance is maximized when the variable is normally distributed. The maximum of (27) is indeed attained when $\kappa_3 = \kappa_4 = 0$ and is equal to the entropy of a unit variance gaussian random variable. A similar approximation can be obtained for the bivariate entropy. The derivation is simplified if the data is pre-whitened by principal component analysis, so that the resulting variables are uncorrelated. We denote the whitening matrix as $S^{-1/2}$, where S is the sample covariance matrix of x_1 and x_2, and $S^{-1/2}$ is an inverse square root factor of S. Hence, if we define:

$$\begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} \triangleq S^{-1/2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \tag{28}$$

we have that:

$$H(x_1, x_2) = H(\hat{x}_1, \hat{x}_2) + \frac{1}{2}\log|\det(S)|, \tag{29}$$

where S is always full rank unless x_1 and x_2 are linearly dependent. A detailed derivation of an approximation of $H(\hat{x}_1, \hat{x}_2)$ can be found for example in [7], and is based on a bivariate Gram-Schmidt expansion of the corresponding probability density function. The resulting expression for the approximated entropy is given by:

$$\begin{aligned} H(\hat{x}_1, \hat{x}_2) \approx \log(2\pi e) - \frac{1}{12}\Big[&\kappa_{30}^2 + 3\kappa_{21}^2 + \\ &+ 3\kappa_{12}^2 + \kappa_{03}^2 + \frac{1}{4}(\kappa_{40}^2 + 4\kappa_{31}^2 + 6\kappa_{22}^2 + 4\kappa_{13}^2 + \kappa_{04}^2)\Big], \end{aligned} \tag{30}$$

where the bivariate cross-cumulants can be computed as follows from the sample data:

$$\kappa_{30} = E[\hat{x}_1^3] \tag{31}$$

$$\kappa_{03} = E[\hat{x}_2^3] \tag{32}$$

$$\kappa_{21} = E[\hat{x}_1^2 \hat{x}_2] \tag{33}$$

$$\kappa_{12} = E[\hat{x}_1 \hat{x}_2^2] \tag{34}$$

$$\kappa_{40} = E[\hat{x}_1^4] - 3 \tag{35}$$

$$\kappa_{04} = E[\hat{x}_2^4] - 3 \tag{36}$$

$$\kappa_{31} = E[\hat{x}_1^3 \hat{x}_2] \tag{37}$$

$$\kappa_{13} = E[\hat{x}_1 \hat{x}_2^3] \tag{38}$$

$$\kappa_{22} = E[\hat{x}_1^2 \hat{x}_2^2] - 1 \tag{39}$$

By combining (23) and (29), the following approximation of $I(x_1, x_2)$ is obtained, which involves only cross cumulants up to the fourth order:

$$I(x_1, x_2) = H(x_1) + H(x_2) - H(x_1, x_2) \tag{40}$$

$$\begin{aligned} &= H(\tilde{x}_1) + \log(\sigma_1) + H(\tilde{x}_2) + \log(\sigma_2) \\ &\quad - H(\hat{x}_1, \hat{x}_2) - \frac{1}{2} \log |\det(S)| \end{aligned} \tag{41}$$

$$\begin{aligned} &= \frac{1}{12} \left\{ - \left[\kappa_{3,1}^2 + \kappa_{3,2}^2 + \frac{1}{4}(\kappa_{4,1}^2 + \kappa_{4,2}^2) \right] + \right. \\ &\quad + \left[\kappa_{30}^2 + 3\kappa_{21}^2 + +3\kappa_{12}^2 + \kappa_{03}^2 + \frac{1}{4}(\kappa_{40}^2 + \right. \\ &\quad \left. \left. + 4\kappa_{31}^2 + 6\kappa_{22}^2 + 4\kappa_{13}^2 + \kappa_{04}^2) \right] \right\} + \\ &\quad + \log(\sigma_1) + \log(\sigma_2) - \frac{1}{2} \log |\det(S)|. \end{aligned} \tag{42}$$

An analogous expression involving *conditional* cross-cumulants of the variables can be used to estimate the conditional mutual information $I(x_1, x_2|x_0)$.

IV. Results and Discussion

For a given microarray experiment, according to the framework outlined in the previous section, all possible unique combinations of three genes are considered and the co-information is evaluated to assign a score to each such combination. The highest scoring clusters are recorded in order to be further evaluated. The actual software implementation includes a set of tools devoted to pre-processing the expression data, performing a series of tasks which include pruning the set of genes according to a user defined criterion (*e.g.* their sample variance), correcting for outliers or accounting for missing values.

DNA microarray data is conventionally expressed as the logarithm (usually in base 10) of the ratio between the estimated expression level and a reference value. Therefore, a log-ratio value of zero indicates that the gene is expressed at levels close to the reference. A reading of 0.3 or above is equivalent to a 2-fold increase in the transcription level. We will conventionally refer to gene levels that show at least a 2-fold increase as *up-regulated* or *over-expressed*, compared to the reference level. Equivalently, when the log-ratio level is -0.3 or less, the gene shows at least a 2-fold decrease in the expression level and will be referred to as *down-regulated*, or *under-expressed*.

Due to the small sample size available, the estimation of the set of conditional entropies is most efficiently accomplished by discretizing the expression levels of the parent node into three levels, according to whether the gene is down-regulated, close to the reference level (baseline), or up-regulated. The choice of the discretization levels is arbitrary and will, in general, affect the outcome of the exploratory analysis. Throughout our analysis, the default thresholds of -0.3 for down-regulation and 0.3 for up-regulation were adopted. Such choice is dictated by considerations that are both biological and statistical. The goal is clearly to select a level at which the up- or down-regulation can be robustly established. Due to the large measurement error affecting the data, it is widely accepted that at least a 2-fold increase or decrease in the measured expression level is necessary in order to establish significant up or down regulations.

A. Analysis of Saccharomyces cerevisiae Expression Data

In order to evaluate the effectiveness of the proposed approach in unveiling hidden dependencies between gene transcription levels, we considered a dataset composed of several experiments

Sample index	Condition	Number of samples
1–15	Heat shock from 25^{o}C to 37^{o}C	15
16–20	Temperature shift from 37^{o}C to 25^{o}C	5
21–25	Heat shock from various temp. to 37^{o}C	5
26–35	Mild heat shock at variable osmolarity	10
36–45	Hydrogen peroxide treatment	10
46–54	Menadione exposure	9
55–69	DTT exposure	15
70–77	Diamide treatment	8
78–84	Hyper-osmotic shock	7
85–90	Hypo-osmotic shock	6
91–95	Amino-acid starvation	5
96–105	Nitrogen source depletion	10
106–112	Glucose depletion (diauxic shift)	7
113–134	Stationary phase growth	22
135–139	Response of mutant cells to heat shock	5
140–144	Mutant cells exposed to H_2O_2	5
145–147	Over-expression studies	3
148–160	Steady-state growth on alternate carbon source	13
161–173	Steady-state growth at constant temperatures	13

TABLE I

EXPERIMENTAL CONDITIONS AND CORRESPONDING NUMBER OF MEASUREMENTS FOR THE *S.cerevisiae* DATASET USED IN THE SIMULATIONS.

involving whole-genome assays of the gene expression levels of the yeast *S.cerevisiae*. The data[2] consists of a total of 6,152 genes with 173 sample points per gene. Table I provides a basic listing of the experimental conditions. A detailed description of the actual experimental conditions can be found in [4].

The dataset comprises a variety of experimental conditions, including temperature induced shock, exposure to various chemicals, aminoacid starvation, nitrogen depletion and so on. The resulting large oscillations in the expression levels of several genes ensure that the dataset

[2]All datasets were downloaded from the publicly available Stanford Microarray Database (http://genome-www5.stanford.edu/MicroArray/SMD).

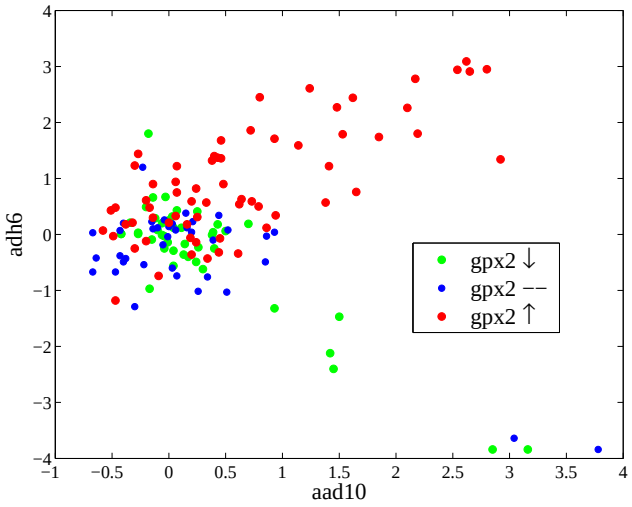

Fig. 5. Co-expression pattern between the genes *aad10* and *adh6*, when gene *gpx2* is the conditioning node. The plot shows that when *gpx2* is up-regulated, *adh6* and *aad10* are in general positively correlated and above the reference level. On the other hand, when *gpx2* is down-regulated, a negative pattern of correlation appears between *adh6* and *aad10*. Such conditional expression program could be explained by considering that *gpx2* was found to be considerably under-expressed in those experiments involving a depletion in sources of nitrogen. Therefore, in such conditions the enzyme translated from *adh6* lacks its primary activation mechanism, and its enzymatic role appears to be replaced by *aad10*.

provides enough variability to allow the consistent detection of specific patterns, in a statistically meaningful way. A preliminary analysis of the dataset, suggested the removal of the sample points 113–134 (*cfr.* Table I), which were collected over a considerably larger span of time (few days vs. few hours for the other experiments). We observed that in such experiments several genes were characterized by a specific expression pattern, most likely associated with the fact that the yeast cells had reached a steady state and stopped replicating. Therefore, such points were treated as outliers and were not included in our exploratory procedure.

Among all the possible triplets of genes whose score was evaluated, 3,124 resulted in a value of the co-information measure that was above a pre-selected significance threshold. A selection of the conditional interaction patterns that were identified by the algorithm are described in detail below. For each cluster of genes whose conditional co-expression pattern appeared to be relevant, we evaluated separately the statistical significance of the co-information score by using the following procedure. For each conditioning node, we randomly permuted its sample points several times (at least 100 million permutations), and re-evaluated the score of the selected triplet

of gene by using the permuted version of the conditioning variable. A *p*-value expressing the statistical significance of the interaction is obtained by counting the number of times that the score obtained with the scrambled values is larger than the score obtained when using the actual data.

The cluster of genes resulting in the highest value of the co-information cost function included *aad10* and *adh6*, which were found to be co-expressed conditionally on the expression levels of the genes *rot2*, *alg7*, and *gpx2* (Figures 6, 7, and 5). *aad10* is a putative alcohol dehydrogenase, *i.e.* an enzyme involved the alcohol metabolism. The product of *adh6* is also an alcohol dehydrogenase, whose activity is NADPH dependent. Figure 5 shows that when *gpx2* (a glutathione peroxidase induced during glucose starvation) is up-regulated, *adh6* and *aad10* are in general positively correlated and above the reference level. On the other hand, when *gpx2* is down-regulated, a negative pattern of correlation appears between *adh6* and *aad10*. Such conditional expression program could be explained by considering that *gpx2* was found to be considerably under-expressed in those experiments involving a depletion in sources of nitrogen. Therefore, in such conditions the enzyme translated from *adh6* lacks its primary activation mechanism, and its enzymatic role appears to be replaced by *aad10*. Figures 6 and 7 show that the genes *rot2* (involved in normal cell wall synthesis) and *alg7* (responsible for protein glycosylation) also act as indicators of such alternative gene expression program, with their expression being repressed under the same conditions.

The conditional co-expression pattern involving *gpx2*, *aad10* and *adh6* was found to be significant at a p-value of less than 10^{-8}. A histogram of the scores obtained by scrambling the values of the conditioning variable is shown in Figure 8, where the score obtained by using the actual sample points is also shown for comparison.

Figure 9, shows the conditional co-expression pattern involving genes *sul1*, *sam4*, and *cwp1*. The gene *sul1* is one of two major mediators (*sul2* is the other one) of the sulfate transport pathway, being responsible for controlling the concentration of endogenous activated sulfate intermediates. Its activity is closely related to the one of *sam4*, the latter being involved in the metabolism of sulfur-containing aminoacids. The gene *cwp1*, is mainly involved in cell wall organization and biogenesis. *sam4* and *cwp1* appear to be in general negatively correlated, possibly due to the fact that their activity peaks in completely different stages of the yeast cell cycle. However, when *sul1* is over-expressed (signaling an increase in the concentration

of activated sulfate compounds), the two genes appear to be positively correlated as well as generally under-expressed.

The last example involves the genes *gln3*, *vap1*, and *pph3*. This case is of particular interest since it demonstrates the method's capability of detecting certain types of regulatory interactions that could not be directly derived from simple correlation patterns. *gln3* is a transcription factor responsible for the regulation of nitrogen utilization. Its product is generally inactive unless activated by the protein phosphatase translated from *pph3*. Figure 10 shows a scatter plot of *gln3* vs. *pph3*: no pattern of co-expression appears when examining the expression profiles of these two genes. On the other hand, as it is shown in Figure 11, these genes are conditionally co-expressed. The plot shows that *vap1* (whose product is an amino-acid transport protein) and *pph3* are positively correlated, when *gln3* is under-expressed or near the reference level. An up-regulation of *gln3* results in the opposite correlation pattern for *vap1* and *pph3*. This mechanism can be explained considering that most of the expression levels responsible for such pattern are relative to conditions of either nitrogen depletion or amino-acid starvation. In such conditions, the expression level of *gln3* rapidly increases (concurrently with the level of its activator *pph3*), in order to repress the transcription of nitrogen demanding gene products. At the same time, *vap1* is strongly down-regulated due to the fact the amino-acid transport mechanisms are significantly slowed down during this stage.

In general, not all significant patterns of conditional interaction detected by the algorithm carry a straightforward biological explanation. This is most likely due to the fact that such patterns of conditional co-expression are often mediated by several factors that are not directly measurable. Moreover, it is often the case that conditionally informative clusters include one or more genes whose biological role is only partially known or completely unknown. Despite such limitations, the proposed framework provides a valuable tool to biologists, being capable of highlighting patterns of interaction whose biological significance can be elucidated through further experimental analysis.

V. Conclusions

We introduced a novel method capable of detecting linear as well as non-linear patterns of conditional co-expression in gene expression measurements. Due to the significant computational cost associated with the proposed exploratory method, the derivation of an efficient technique

for evaluating the co-information score played a key role in order to make the method computationally tractable. We applied the method to a whole genome micro-array dataset of the yeast *S.cerevisiae* and were able to detect several statistically significant patterns of conditional interaction between genes. This result proves unquestionably that such patterns of conditional co-expression appear indeed very frequently in the data, and raises the very important question of whether a general biological model capable of explaining such interactions can be devised.

REFERENCES

[1] A. J. Bell. The co-information lattice. In *Proc. 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 921–926, Nara, Japan, April 2003.

[2] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

[3] M.B. Eisen, P.T. Spellman, P.O. Brown, , and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, December 1998.

[4] A.P. Gash *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell.*, 11:4241–4257, 2000.

[5] N. Friedman, I. Nachman, and D. Peér. Learning bayesian network structure from massive datasets: The "sparse candidate" algorithm. In Kathryn Blackmond Laskey and Henri Prade, editors, *Proc. Fifteenth Conf. on Uncertainty in Artificial Intelligence (UAI)*, pages 206–215, San Francisco, 1999. Morgan Kaufmann, Inc.

[6] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning*. Springer Verlag, New York, 2001.

[7] M.C. Jones. *The Projection Pursuit Algorithm for Exploratory Data Analysis*. PhD thesis, University of Bath, School of Mathematics, 1983.

[8] M.G. Kendall and A. Stuart. *The Advanced Theory of Statistics. Volume I: Distribution Theory (4*th *ed.)*. Griffin, London, 1977.

[9] T. Kohonen. Self-organizing formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982.

[10] Ker-Chau Li. Genome-wide coexpression dynamics: Theory and application. *Proc. Natl. Acad. Sci. (PNAS) USA*, 99(26):16875–16880, December 2002.

[11] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York, 1985.

[12] D. L. Wallace. Asymptotic approximations to distributions. *Ann. Math. Stat.*, 29:635–654, 1958.

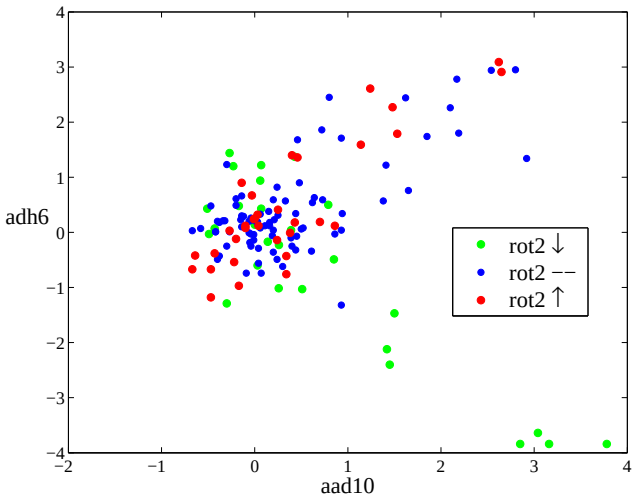

Fig. 6. Co-expression pattern between the genes *aad10* and *adh6*, when gene *rot2* is the conditioning node. The expression levels of *aad10* and *adh6* switch from a positive to a negative correlation pattern when *rot2* is under-expressed.

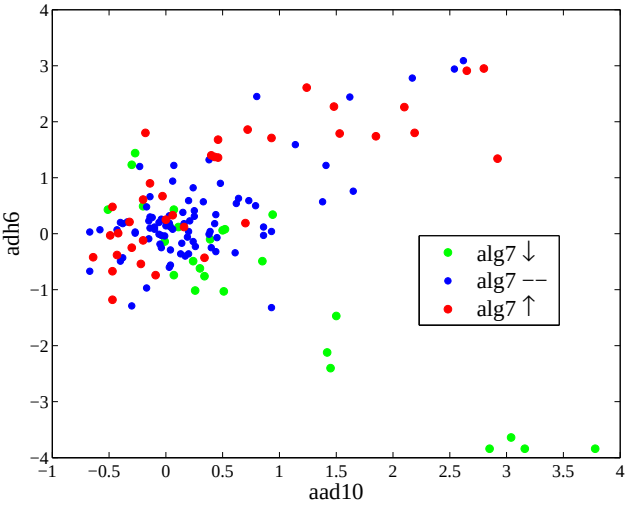

Fig. 7. Co-expression pattern between the genes *aad10* and *adh6*, when gene *alg7* is the conditioning node. The expression levels of *aad10* and *adh6* switch from a positive to a negative correlation pattern when *alg7* is under-expressed.

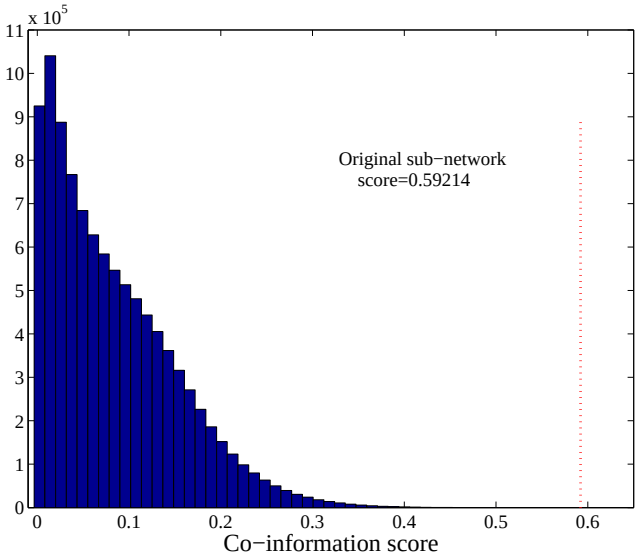

Fig. 8. Statistical significance of the conditional co-expression pattern. The plot shows a histogram of the co-information values obtained by scoring the triplet *aad10*, *adh6*, and *gpx2*, when the samples odf the latter are randomly permuted. The score of the actual sub-network is also shown for reference.

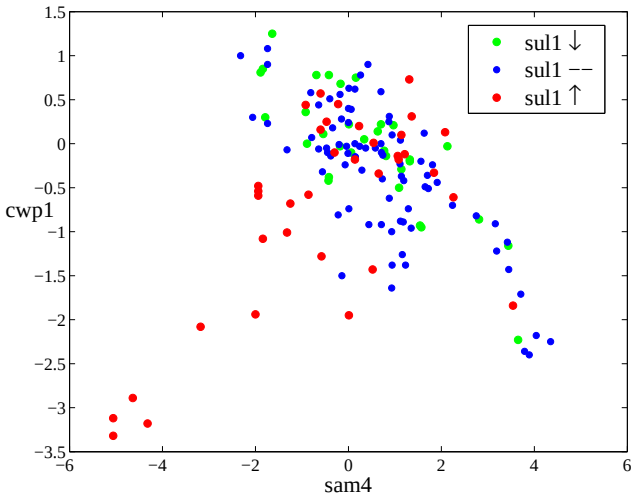

Fig. 9. Co-expression pattern between the genes *sam4* and *cwp1*, when gene *sul1* is the conditioning node. *sam4* and *cwp1* appear to be in general negatively correlated, possibly due to the fact that their activity peaks in completely different stages of the yeast cell cycle. However, when *sul1* is over-expressed (signaling an increase in the concentration of activated sulfate compounds), the two genes appear to be positively correlated as well as generally under-expressed.

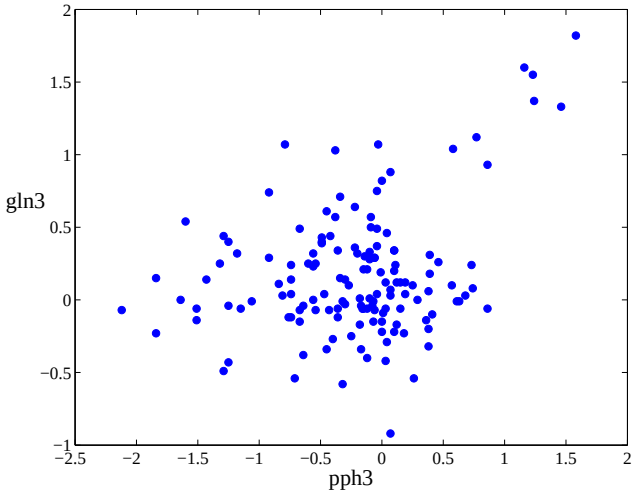

Fig. 10. Scatter plot of the expression levels of *pph3*. Although *pph3* is involved in the activation of the transcriptional regulator *gln3*, no significant co-expression between the expression levels of the two genes can be observed from the scatter plot.

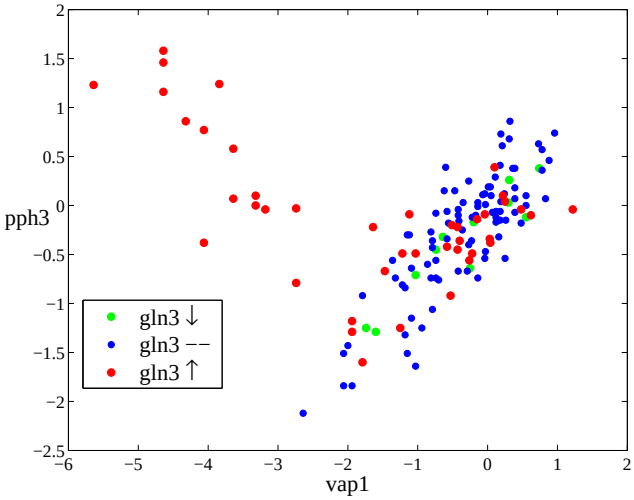

Fig. 11. Co-expression pattern between the genes *vap1* and *pph3*, when gene *gln3* is the conditioning node. The plot shows that *vap1* (whose product is an amino-acid transport protein) and *pph3* are positively correlated, when *gln3* is under-expressed or near the reference level. An up-regulation of *gln3* results in the opposite correlation pattern for *vap1* and *pph3*. This mechanism can be explained considering that most of the expression levels responsible for such pattern are relative to conditions of either nitrogen depletion or amino-acid starvation. In such conditions, the expression level of *gln3* rapidly increases (concurrently with the level of its activator *pph3*), in order to repress the transcription of nitrogen demanding gene products.