BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Inferring gene regression networks with model trees

Isabel A Nepomuceno-Chamorro [1*], Jesus S Aguilar-Ruiz [2*], Jose C Riquelme[1]

Abstract

Background: Novel strategies are required in order to handle the huge amount of data produced by microarray technologies. To infer gene regulatory networks, the first step is to find direct regulatory relationships between genes building the so-called gene co-expression networks. They are typically generated using correlation statistics as pairwise similarity measures. Correlation-based methods are very useful in order to determine whether two genes have a strong global similarity but do not detect local similarities.

Results: We propose model trees as a method to identify gene interaction networks. While correlation-based methods analyze each pair of genes, in our approach we generate a single regression tree for each gene from the remaining genes. Finally, a graph from all the relationships among output and input genes is built taking into account whether the pair of genes is statistically significant. For this reason we apply a statistical procedure to control the false discovery rate. The performance of our approach, named REGNET, is experimentally tested on two well-known data sets: *Saccharomyces Cerevisiae* and E.coli data set. First, the biological coherence of the results are tested. Second the E.coli transcriptional network (in the Regulon database) is used as control to compare the results to that of a correlation-based method. This experiment shows that REGNET performs more accurately at detecting true gene associations than the Pearson and Spearman zeroth and first-order correlation-based methods.

Conclusions: REGNET generates gene association networks from gene expression data, and differs from correlation-based methods in that the relationship between one gene and others is calculated simultaneously. Model trees are very useful techniques to estimate the numerical values for the target genes by linear regression functions. They are very often more precise than linear regression models because they can add just different linear regressions to separate areas of the search space favoring to infer localized similarities over a more global similarity. Furthermore, experimental results show the good performance of REGNET.

Background

In the area of microarray data analysis, inferring gene-gene interactions involved in biological function is a relevant task. Over the past few years several statistical and machine learning techniques have been proposed to carry out the inferring task of gene-gene interactions or gene regulatory networks. Clustering algorithm represents one of the first approaches to support the identification of regulatory modules [1,2]. These approaches are motivated by a simple idea which is still widely used in functional genomic. It is called the guilt-by-association heuristic: co-expression means co-regulation, i.e. if two genes show similar expression profiles, they are supposed to follow the same regulatory regime.

In order to formalize the idea of similar expression, several statistical measures have been proposed as solution. In correlation methods, interactions are inferred using correlation statistics as pairwise similarity measures between gene expression profiles over multiple conditions, as for example in [3]. In this kind of methods, if the correlation between gene pairs is higher than a threshold value, then it is considered that these gene pairs interact directly in a signaling pathway and are relevant in a biological way [4-6]. These methods build gene co-expression networks, also known as gene association, gene interaction or gene relevance networks. These networks provide a framework for assigning biological function to group of genes as it was argued in

* Correspondence: inepomuceno@us.es; aguilar@upo.es
[1]Dpt Lenguajes y Sistemas Informaticos, Universidad de Sevilla, Seville, Spain
[2]School of Engineering, Pablo de Olavide University, Seville, Spain
Full list of author information is available at the end of the article

[7]. Correlation coefficient is widely used as a way of obtaining an association measure between two random variables but does not provide a causal measure between them. However, correlation is still informative about the underlying structure [8]. The causal properties that can be inferred from correlations have been investigated in [9,10].

Correlation-based methods are very useful to determine whether two genes have a strong global similarity over all conditions from the data set. This is an important constrain as there might exist a strong local similarity over a subset of conditions, which could not be detected with global similarity measures. In addition, many pairs of genes show similar behavior in gene expression profiles by chance even though they are not biologically related [11], i.e. the significance of the results should be assessed in interaction networks.

On the other hand, Gaussian graphical models (GGM) are a full conditional independence model. These models try to explain the correlation between two genes by the rest of the genes and they are a popular tool to represent gene association network [8,12,13]. Recently, [14] has proposed estimating partial correlations to attach lengths to the edges of the GGM, where the length of an edge is inversely related to the partial correlation between a gene pair. As a drawback, these models are hard to estimate if the number of samples is small compared to the number of variables. In contrast to GGMs, other models try to explain the correlation between two genes not by the rest of the genes, but only by single third genes. This idea can also be implemented using sparse Gaussian graphical model based on partial correlation [15] or conditional mutual information to test for first-order independence [16-18].

Bayesian networks try to explain the dependence between genes if there are no subset of other genes that explain the dependency [19]. An example of Bayesian networks can be found in [20] where a *stochastic expectation and maximization* algorithm is used to learn a probabilistic model, and regression trees are used to learn graph topologies that maximize Bayesian scores. Recently, [21] has revised the approach before using an ensemble method, and [22] has incorporated prior knowledge from literature on Bayesian networks. Also, several approaches have been developed to build Boolean networks [23], or to infer regulatory rules [24,25] using machine learning principles.

In this paper, we present a novel method inspired by model trees as a way to detect linear dependencies between genes and to set a group of gene-gene dependencies. From that set, our method provides as gene-gene interactions all those significant dependencies in a statistical sense. Then, it builds undirected dependency graphs (UDGs) from these gene-gene interactions. Furthermore, our method analyzes which dependencies between genes are considered as a discovery by means of the Benjamini and Yekutieli procedure [26]. This statistical procedure enables the control of the expected proportion of false discoveries among all the discoveries made. One of the main contributions of our approach is that it addresses the issue of searching for local similarities arising from conditional regulatory relationships -instead of global similarities.

The remainder of this paper is organized as follows. In Section *Method*, a detailed explanation of the methodology and the algorithm are presented. In Section *Results and Discussion*, experimental results tested on an in silico benchmark suite of datasets, yeast and E.coli data are provided. Finally, Section *Conclusions* summarizes the most relevant conclusions and future research directions.

Method

Correlation methods are focused on the global match of two gene expression profiles, analyzing each possible pair of genes. Instead, our approach analyzes each gene in an iterative way. At each iteration a gene is taken as target gene and the remaining genes as input for splitting the search space. In each subspace generated by that division, a linear model is built to identify a linear dependency between the target gene and a subgroup of genes, i.e. the target gene expression values are estimated by this subgroup of genes involved in that linear model. As a consequence, the dependency between the genes is not calculated for the complete gene expression profile, but for a localized subspace of the profiles using M5' model tree algorithm.

Our method consists of three steps as it is depicted in Figure 1. The first step involves building M5' trees. M5' is a model tree algorithm, an extension of regression tree algorithms [27], which has several linear models, each one of them built in a leaf of the tree. The aim of this step is to obtain a set of genes associated to other genes from their prediction ability by means of linear regression functions. We use model trees because these representations work like several linear regression functions at the same time, each of them identified by a leaf in the tree. The main advantage of this methodology is that each regression is specialized in a specific area of the search space, i.e. in a local subspace of gene expression profiles, hence the model tree is generally more accurate than a global linear regression.

The second step implies the extraction of the set of gene-gene dependencies from the forest of trees obtained by the previous step. Specifically, our approach considers which hypothetical evidences of gene-gene dependency exist between the target gene and every gene participating in the linear regression functions of the target gene.

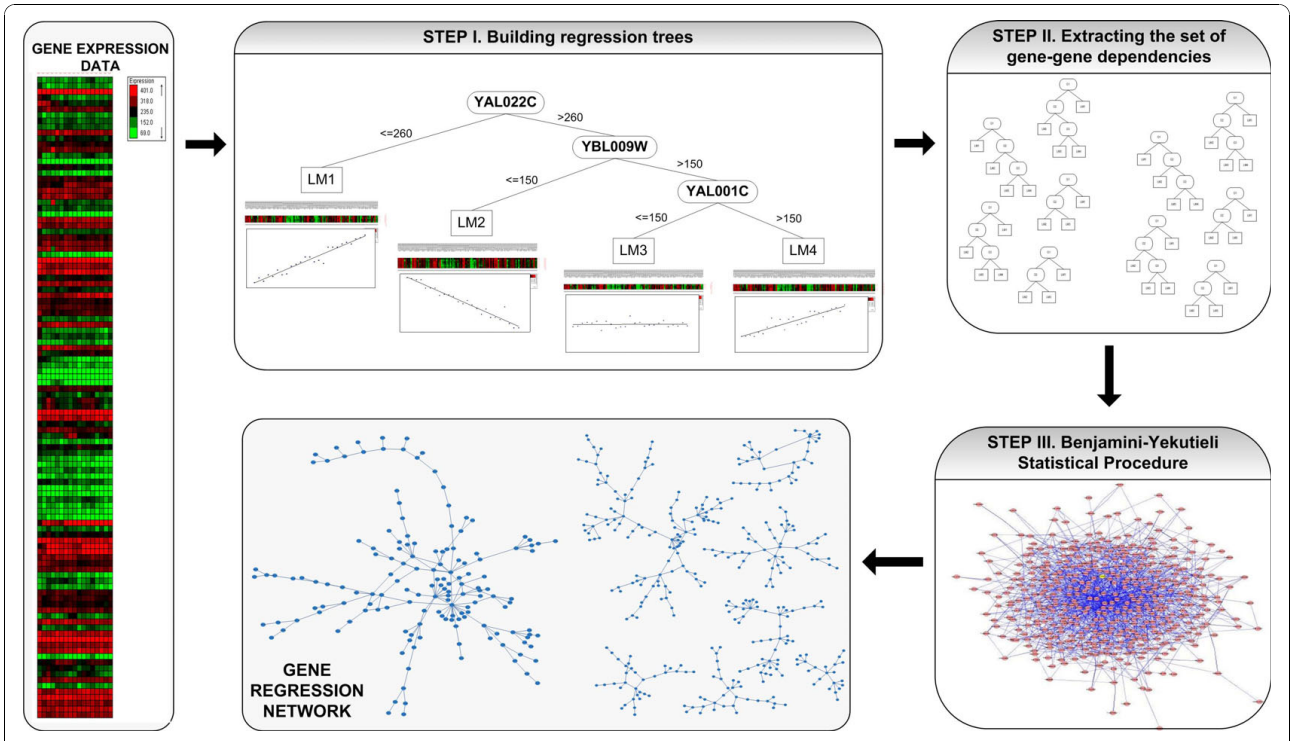


Figure 1 Schematic view of the proposed method. In the first step, for each gene (target gene) a model tree is generated, which provides a partition of the space. Linear regression functions are built in the leaves of the tree. These linear regression functions can be seen as prototypes to estimate the value of the target gene. Genes involved in the linear regression functions might be identified as potential dependencies. This is an iterative process that is made for each gene taking the remaining genes as input to build the model tree, and it provides a set of hypothetical gene-gene interactions. Only the model trees with low prediction error will be conserved. Next, the Benjamini-Yekutieli statistical procedure is applied in order to assess the significance of the dependencies.

Finally, the third step involves learning a graph model of gene co-expression network by assessing the significance of the set of hypothetical evidences. Many sets of genes show similar behavior in expression profile by chance even though they do not share the same biological function. Therefore, the aim of this step is to minimize the number of false discoveries among all those discoveries made in the previous step. For that reason, we apply a statistical procedure to control the false discovery rate instead of the increase of type I error when a family of hypotheses is being tested simultaneously. The reliability of our method is strengthened by applying the Benjamini-Yekutieli statistical procedure to assess the significance of the results.

Building model trees

The first work on regression trees dates from [28], although the most popular reference is the seminal work of [29]. Later on, [30] introduced the system M5. It builds multivariate trees using linear regression functions at the leaves. M5' is introduced in [31], a rational reconstruction of Quinlan's M5 algorithm. Throughout the description of model tree, we will refer to gene as attribute, and sample as instance space.

The algorithm M5' is divided into two phases. First, a tree is built by a decision-tree induction algorithm, and second, a pruning procedure is applied. Given a gene as a target, M5' constructs a tree by recursively splitting the instance space. In this decision-tree induction algorithm the splitting criterion is based on treating the standard deviation, i.e. the attribute which maximizes the expected error reduction is chosen. After the tree has been built, a linear regression function is obtained for every internal node of the tree and the regression models are reduced by dropping attributes to minimize the estimated error on future data. The number of attributes in the linear regression functions decreases and the average error will offset over the training example. After this has been done, every subtree is considered for pruning. Pruning takes place if the estimated error for the linear regression function at the root of a subtree is smaller than or equal to the expected error for this subtree. After pruning is done, M5' applies a smoothing process to compensate sharp discontinuities that occur between adjacent regression models at the leaves of the tree. Finally, M5' has an associated relative error ε that will be used to reject some of the trees, those with low

INPUT M: a microarray (gene expression data)
θ: threshold value
α: significance level
OUTPUT $Q^{\star}$: graph of gene–gene interactions
begin
$FT \leftarrow \emptyset$ {STEP 1 – Building model trees}
for all gene $g_j \in G$ **do**
$FT \leftarrow FT \cup MT_j$
end for
$FT^{\theta} \leftarrow \emptyset$ {STEP 2 – Extracting gene-gene interactions}
for all model tree $MT_j \in FT$ **do**
if ε of $MT_j < \theta$ **then**
$FT^{\theta} \leftarrow FT^{\theta} \cup MT_j$
end if
end for
$TG^{\theta} \leftarrow \emptyset$; $LG^{\theta} \leftarrow \emptyset$;$D^{\theta} \leftarrow \emptyset$
for all model tree $MT_j \in FT^{\theta}$ **do**
$TG^{\theta} \leftarrow TG^{\theta} \cup g_j$
for all $g \in \Delta(g_j)$ **do**
$LG^{\theta} \leftarrow LG^{\theta} \cup g$
$D^{\theta} \leftarrow D^{\theta} \cup (g_j, g)$
end for
end for
$Q \leftarrow (TG^{\theta} \cup LG^{\theta}, D^{\theta})$
$Q^{\star} \leftarrow$ BY(Q, α) {STEP 3 – Controlling the false discovery rate}
end

Figure 2 Pseudocode. Pseudocode of REGNET.

precision. The result is a forest of trees (FT^{θ} in Figure 2). This algorithm is described in [30,31].

Our approach takes each gene as a target gene and builds a model tree to predict the target gene expression values. By construction of model tree, linear regression functions are built to infer localized similarities over a more global similarity. Figure 3 presents a hypothetical example, the correlation between the target gene and two other genes is weak, however we can observe two strong local dependencies between them.

Extracting gene-gene dependencies

This step extracts a set of dependencies between the target gene and the genes involved in the linear regression functions from each tree. Correlation-based methods extract gene-gene dependencies by computing a similarity score for each pair of genes. These methods are based on the assumption that two genes show similar expression profiles if they follow the same regulatory regime, i.e. coexpression hints at coregulation [11]. Our approach analyzes each gene as a target by taking into account the remaining genes as inputs to obtain linear models that estimate the expression value of that target gene. We assume that the genes involved in these linear models control or influence the target expression value and they follow the same regulatory regime. This influence can be explained when several genes fit a specific area of the space, which leads to an evidence for dependency.

Let LM be a multivariate linear model of a M5' tree defined by $LM: g_x = \sum_i \lambda_i g_{\gamma_i}$, where g_x belongs to the set of target genes, g_{γ_i}, is a gene involved in the linear

regression that belongs to the set of genes, and λ_i is a coefficient of the linear model. Our approach considers that an hypothetical evidence of dependency or expression pattern exists between g_x and every g_{y_i}, which will be statistically tested in the next step.

The output of this step is a set of gene-gene dependencies (Q in Figure 2) that are potential interactions for the problem under study.

Building the gene regression network

After obtaining the set of gene-gene interactions, the significance of these results must be assessed. The authors in [32] have shown that for microarrays studies, the expected proportion of false discoveries among all the discoveries made (so-called *false discovery rate*, FDR) is more important than the low number of false discoveries or the small probability of making at least one false positive (calculated by means of adjustments of p-values). For this reason we apply a statistical procedure in order to control the number of type I errors (connections inferred which do not correspond to a connection in the real network, also called *false positives*) among the number of discoveries when a family of hypotheses is tested simultaneously.

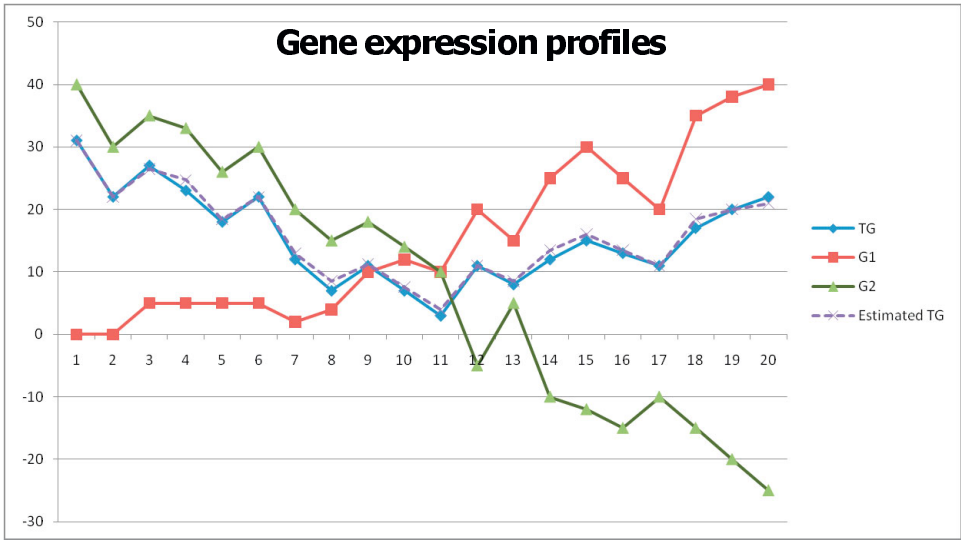
Once the set of gene-gene dependencies (D) has been provided, our approach builds a graph Q of interactions defined as a tuple (N, E) of $|N|$ nodes and $|E|$ edges. We will denote by $g_x \sim g_y$ an hypothetical gene-gene dependency. Our approach takes several $g_x \sim g_y$ from D and the genes g_x and g_y are mapped as two nodes in the set of nodes N, and the dependency is mapped as an edge of the set E. This step, to decide which $g_x \sim g_y$ is mapped onto an edge, i.e. which dependency is considered as a discovery, is carried out by means of the Benjamini-Yekutieli (BY) procedure.

The BY procedure is applied in order to test m null hypotheses $H_0^1, H_0^2, \ldots, H_0^m$. Let $p_1, \ldots, p_m$ be the corresponding p-values to m null hypotheses. Let $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$ be the ordered p-values. This procedure defines k as detailed in Eq. 1 and rejects all hypothesis.

$$k = max\left\{ i : p_{(i)} \frac{m}{i} \sum_{k=1}^{m} \frac{1}{k} \leq \alpha \right\} \quad (1)$$

If no such i exists, none of the hypotheses will be rejected. This procedure controls the proportion of false discoveries (FDR) among all the discoveries.

TG	G_1	G_2
31	0	40
22	0	30
27	5	35
23	5	33
18	5	26
22	5	30
12	2	20
7	4	15
11	10	18
7	12	14
3	10	10
11	20	-5
8	15	5
12	25	-10
15	30	-12
13	25	-15
11	20	-10
17	35	-15
20	38	-20
22	40	-25



Corr(TG,G1) = -0.09

Corr(TG,G2) = 0.35

IF G1<=10 AND G2>10 then Estimated-TG = 0.9 * G2 -5

IF G1>10 then Estimated-TG = 0.5 * G1 + 1

Figure 3 Hypothetical example of localized similarities. The table represents the gene expression values from 20 samples. The correlation coefficients between the target gene *TG* and the two other genes are weak ($\rho(TG, G_1)$ = -0.09 and $\rho(TG, G_2)$ = 0.35). However we can observe in this hypothetical example two strong localized similarities detected by construction of this hypothetical model tree: IF $G_1 \leq 10$ AND $G_2 > 10$ THEN $TG = 0.9 * G_2 - 5$. IF $G_1 > 10$ THEN $TG = 0.5 * G_1 + 1$ The dot line is the results of apply the linear regression functions that estimate the target gene expression value.

In this context, we will say that $g_x \sim g_y$ is not an interaction in Q^* if and only if there is not any significant monotonic relationship between the two variables, i.e. $H_0 : \rho_{xy} \approx 0$ (where ρ is a correlation measure), taking into account the subspace of the input data identified by the leaf of the linear model in the M5' tree. If this null hypothesis is rejected at the significance level represented by α, this dependency is mapped into the graph. To test whether a significant monotonic relationship exists, we use the Kendall's τ (under the subspace or subset of gene expression samples) as non-parametric measure of association [33].

Algorithm

In order to formalize the algorithm, named REGNET, several definitions are required.

Definition 1 (Microarray)

Let M *be the microarray data, defined as* $M = (\mathcal{C}, \mathcal{G}, \mathcal{L})$, *where* $\mathcal{C} = \{c_1, c_2, ..., c_n\}$ *is a finite set of experimental conditions,* $\mathcal{G} = \{g_1, g_2, ..., g_m\}$ *is a finite set of genes, and* $\mathcal{L} = (v_{ij})$ *is a* $n \times m$ *gene expression matrix, where* $v_{ij} = \ell(c_i, g_j)$ *given by the level function* $\ell : \mathcal{C} \times \mathcal{G} \to \mathbb{R}$.

Definition 2 (Partition)

A partition Π *of a set* S *is a non-empty collection of non-empty subsets of* S, $\Pi = \{\pi_i\}_{i = 1,..., p}$ *such that* $\cup\pi_i = S$ *and* $\pi_i \cap \pi_j = \varnothing$ *when* $i \neq j$ *for* $i, j = 1,..., p$. *The set of partitions of* S *is denoted by* PART(S).

Definition 3 (Model Tree)

A model tree MT_j *is aimed at estimating the values of the level function* ℓ *for the column j, i.e. for the target gene* g_j, $MT_j = \{(\psi_i, \varphi_i)\}_{i=1,..., q}$, *where* $\bigcup\psi_i \in \text{PART}(\mathcal{C})$, *and* φ_i *is a linear function defined on a subset of genes* $\Omega_i \subset \mathcal{G} - \{g_j\}$, *i.e.,* $\varphi_i : \Omega_i \to \mathbb{R}$. *Therefore, each function* φ_i *will be applied in a subspace of conditions* ψ_i *to locally estimate the level function of the gene* g_j.

Given a relative error threshold for the model tree θ, *then* MT_j^θ *defines a non-empty model tree when its relative error* ε *is smaller than* θ.

$$MT_j^\theta = \begin{cases} MT & \text{if } \varepsilon < \theta \\ \varnothing & \text{if } \varepsilon \geq \theta \end{cases}$$

Definition 4 (Forest)

The forest of model trees FT is the collection of every model tree MT generated from each gene g_j, $1 \leq j \leq m$, $FT = \{MT_1, MT_2, ..., MT_m\}$, *where each* MT_j *is built by minimizing the error* ε *at estimating the level function for gene* g_j *and the conditions within* ψ_i *by means of the functions* φ_i.

Definition 5 (Association)

A gene g is potentially associated with the gene g_j ($g \sim g_j$) *if g appears in any of the* Ω_i *of the corresponding functions* $\varphi_1, \varphi_2, ..., \varphi_q$ *defined at the leaves of the model tree* MT_j, *whose target gene is* g_j. *Each function* φ_i *involves a set of genes* Ω_i *related to* g_j, *and therefore, all the genes associated with* g_j, *represented as* $\Delta(g_j) = \bigcup_{i=1}^{q} \Omega_i$, *constitute potential associations.*

Given a threshold θ *there is an association between two genes,* $g_x \overset{\theta}{\sim} g_y$, *if and only if* g_x *belongs to the set of genes that form the regression of* g_y.

$$g_x \overset{\theta}{\sim} g_y \Leftrightarrow g_y \in TG^\theta \wedge g_x \in \Delta(g_y)$$

where TG^θ *is the set of target genes*

$$TG^\theta = \{g_j \in \mathcal{G} \mid MT_j^\theta \neq \varnothing\}$$

Definition 6 (Gene Regression Network)

A gene regression network is a graph Q defined for a given θ *as:*

$$Q = (TG^\theta \bigcup LG^\theta, D^\theta)$$

where LG is the set of associated genes

$$LG^\theta = \{g \in \mathcal{G} \mid g \in \Delta(g_i), g_j \in TG^\theta$$

and D is the set of dependencies

$$D^\theta = \{(g_x, g_y) \mid g_x \overset{\theta}{\sim} g_y\}$$

The input is the gene expression matrix M, a threshold value θ to prune the model trees generated, and the significance level α for the Benjamini-Yekutieli procedure. The output is a graph of interactions Q^* among the genes in $\mathcal{G}$.

Regarding the computational complexity of REGNET, the cost of building the forest of trees is m times the cost of building a M5' tree, i.e. $O(m^2 n \log(n))$, where m is the number of genes and n the experimental conditions; extracting the hypothetical dependencies is an iterative process which has a linear complexity $O(m)$; and finally, the BY procedure involves sorting the p-values calculated before, i.e., $O(m \log(m))$. Consequently, the overall cost of the algorithm is $O(m^2 n \log(n))$.

Results and Discussion

The robustness of the methodology is shown by means of the analysis on an *in silico* benchmark suite of

datasets, the *Saccharomyces Cerevisiae* cell cycle and the *E. coli* data set.

In silico benchmark suite of datasets

We tested our approach on a published in silico benchmark suite of datasets [34]. The goal is the prediction of network structure from the given in silico gene expression dataset. We use this suite as a blind performance test to compare our approach REGNET against several benchmark methods.

We used the simulated steady-state gene expression datasets reported in DREAM4 (In silico Network Challenge) [35]. The challenge is to infer 5 networks of size 100 hidden in 15 different experiments of microarray. For each network, the GNW tool [36] is used to simulate three different experiments of microarray: the steady-state levels of single-gene knockouts (deletions); knockdowns experiments by reducing the transcription rate of the corresponding gene by half; multifactorial experiment where each expression profile could be extracted from a patient.

For network inference, we applied several benchmark methods:

- A heuristic algorithm for learning high-dimensional dependency networks from genomic data. We used the *GeneNet* R package to infer causal networks based on partial correlations. *GeneNet* implements the methods of [37] and [38] for learning large-scale gene dependency networks.
- Weighted-LASSO for structured network inference implemented in the *Simone* R package [39] and [40]. This algorithm uses the GLasso procedure to estimate a sparse inverse covariance matrix using a lasso (L1) penalty.
- For learning Bayesian networks (BN) we used the R package named *Deal* [41] and the R package named *G1DBN* http://cran.r-project.org/web/packages/G1DBN.

Results reported here were obtained from *GeneNet*, *Simone* and *G1DBN*. The task of learning Bayesian Networks (BN) from data is NP-hard with respect to the number of network vertices, i.e. Bayesian methods are computationally intractable for a huge number of genes. The *Deal* algorithm for learning BN was unsuitable to obtained networks because of the number of genes in the input microarray (100 genes). The *G1DBN* was suitable to obtain networks because this algorithm performs Dynamic BN inference using first order conditional dependencies as heuristic.

Results reported by REGNET and the benchmark methods are shown in Figure 4. In this graphic, the accuracy is represented for each of the fifteen synthetic data sets. M, O and D represent the microarray data set obtained from a multifactorial, knockout and knockdown experiment, respectively. Results reported here by REGNET were obtained with α = 0.001.

Our approach outperformed the results reported by *G1DBN* and *SIMONE* in all the data set (knockout, knockdown and multifactorial experiments of microarray). In general, our approach showed higher accuracy. Only in five out of fifteen data sets, out approach did not outperform the results obtained by *GeneNet*.

Saccharomyces Cerevisiae dataset

We use *Saccharomyces Cerevisiae* cell cycle expression data set [42], which contains 2884 genes and 17 experimental conditions. In the first experiment, the effect of pruning and non-pruning the forest of model trees is compared. Simplifying the forest involves rejecting all the M5' trees that have a relative error greater than a threshold. For both experiments a level α = 0.05 is fixed for the statistical BY procedure. To analyze the biological coherence of the results we use Gene Ontology attributes to characterize the resulted genes derived from our algorithm. We use FuncAssociate [43] to provide a measure (p-value) that determines whether the set of genes obtained is due to chance, or instead, to common biological behavior. Furthermore, this tool calculates appropriate corrections for multiple hypothesis testing, such as Westfall-Young [44].

Figure 5 depicts the experimental results, which consist of a network with eight main subgraphs or connected components. The algorithm also obtains other minor subgraphs (not depicted in the Figure) that are not considered because they are composed only by three or four edges. From these eight subgraphs, we calculated the correlation between pair of genes to obtain the number of weak correlated genes detected by our approach focused on localized similarities (see

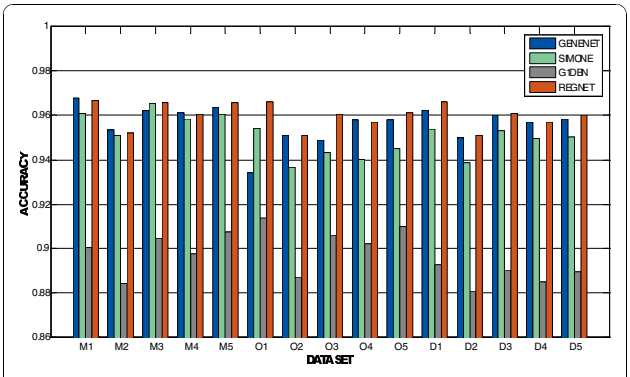


Figure 4 Benchmark analysis. Results reported by REGNET and the benchmark methods using the in silico benchmark suite of datasets [34]. The bars represent the accuracy of the prediction of network structure from the given in silico gene expression dataset.

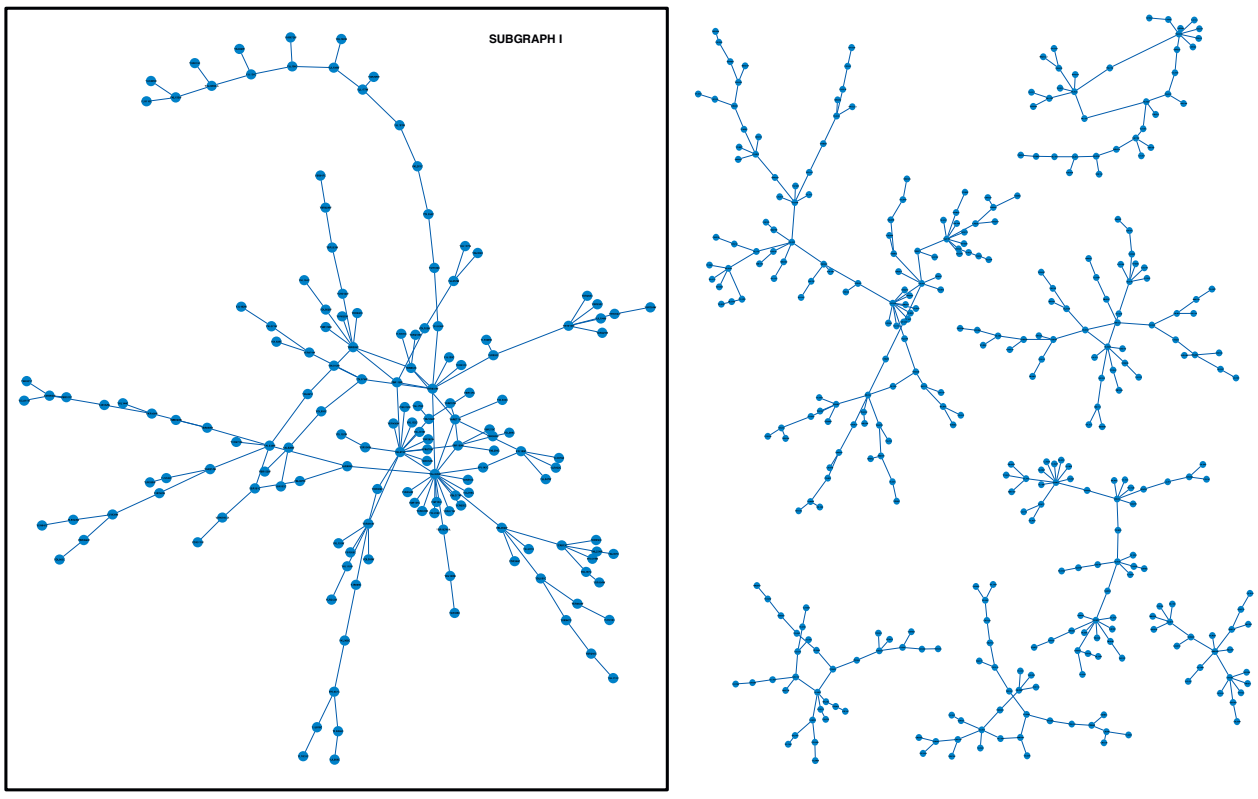


Figure 5 Saccharomyces Cerevisiae dataset. Experiment I. Gene regression network resulting from our approach at level α = 0.05 for the BY procedure. The image was created with Cytoscape [47]. This graph has 502 nodes and 504 edges. Subgraph I is the biggest subgraph obtained and it has 154 nodes and 162 edges.

Additional file 1). We use the biggest subgraph in Figure 5, which has 154 genes, to analyze the result.

The resulted genes are functionally enriched for GO attributes and the great majority of these GO attributes are related with ribosome cellular component, as we can see in Table 1. This table reports these GO attributes, the number of genes in the subgraph with this attribute and the adjusted p-value less than α = 0.05 provided by the FuncAssociate tool [43]. In the first subgraph, there can be seen several genes related with the small subunit of the ribosome that is found in the cytosol (part of the cytoplasm that does not contain membranous or particulate subcellular components) of the cell. There are several genes that contribute to the structural integrity of these small ribosomal subunits which are involved in translation. Specifically, our approach has found genes related with the biological process of aggregation, arrangement and bonding together of constituent RNAs and proteins to form and maintain those small ribosomal subunits. In addition, there are several genes that are involved in the process of assembly and maintenance of the large subunit of the ribosome.

Table 1 Saccharomyces Cerevisiae data. Experiment I

N	P-adj	GO Attribute
38	< 0.001	0005830: cytosolic ribosome
42	< 0.001	0005840: ribosome
37	< 0.001	0003735: structural constituent of ribosomal protein
46	< 0.001	0030529: ribonucleoprotein complex
20	< 0.001	0005843: cytosolic small ribosomal subunit
20	< 0.001	0015935: small ribosomal subunit
17	< 0.001	0015934: large ribosomal subunit

Gene Ontology attributes are used to characterize the genes obtained by our method from the yeast data set. It shows the biological analysis of the biggest subgraph I (see Figure 5). The first column represents the number of genes in the subgraph with this GO attribute, the second column is the adjusted p-value by Westfall and Young corrections and the third column is the name of the GO attribute.

We run our algorithm again but we introduce a variation that involves rejecting all the M5' that has a relative error greater than 50%. This variation restricts the number of linear models taken into account in the learning process of gene-gene interactions. Figure 6 shows the biggest subgraph obtained, which has 62 nodes and all of them belong to the first subgraph mentioned in Experiment I.

The main contribution of this variation is that the size of the subgraph is reduced more than 50% with respect to Experiment I, but the biological information is the

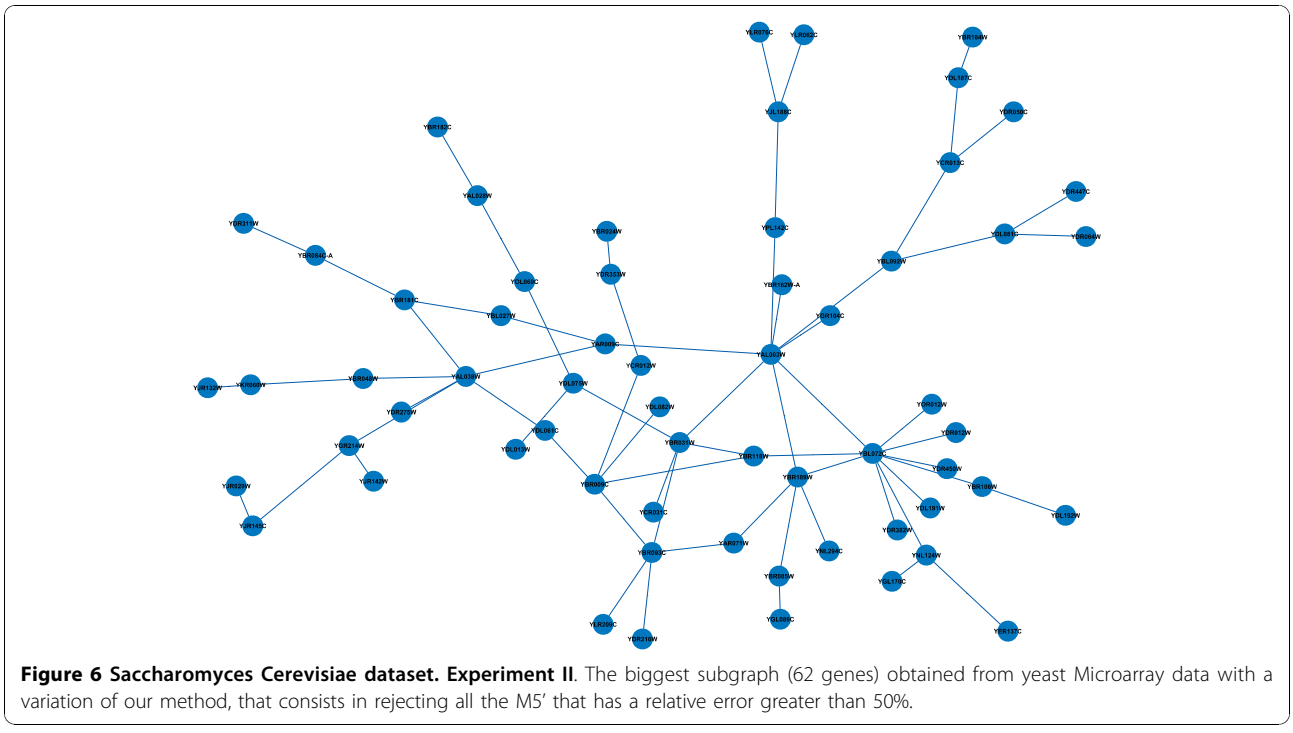


Figure 6 Saccharomyces Cerevisiae dataset. Experiment II. The biggest subgraph (62 genes) obtained from yeast Microarray data with a variation of our method, that consists in rejecting all the M5' that has a relative error greater than 50%.

same, as it can be noticed in Table 2. This table reports the biological study provided by GO database, that relates most of genes to ribosome cellular component (c.f. Table 1). In fact, all the GO attributes in Experiment I have remained in Experiment II, and they are obtained from the simplified forest (all the M5' trees have a relative error smaller than 50%).

In summary, the use of constrains to provide more accurate model trees does not have negative influence on the quality of results. Selecting the best M5' trees from the forest reduces the size of the gene network without decreasing the quality of the results from a biological perspective.

Escherichia coli dataset

The predictive performance of our approach was tested using Escherichia coli (E.coli) gene expression database from [45]. The E.coli gene expression database M^{3D}

Table 2 Saccharomyces Cerevisiae data. Experiment II

N	p-adj	GO Attribute
21	< 0.001	0005830: cytosolic ribosome
23	< 0.001	0005840: ribosome
21	< 0.001	0003735: structural constituent of ribosomal protein
22	< 0.001	0005198: structural molecule activity
23	< 0.001	0030529: ribonucleoprotein complex
11	< 0.001	0005843: cytosolic small ribosomal subunit
11	< 0.001	0016283: eukaryotic 48S initiation complex
11	< 0.001	0016282: eukaryotic 43S preinitiation complex/eukaryotic 43S pre-initiation complex
25	< 0.001	0005829: cytosol
11	< 0.001	0015935: small ribosomal subunit
10	< 0.001	0005842: cytosolic large ribosomal subunit
24	< 0.001	0009059: macromolecule biosynthesis
23	< 0.001	0006412: protein biosynthesis
10	< 0.001	0015934: large ribosomal subunit
4	< 0.001	0000028: ribosomal small subunit assembly and maintenance

Biological analysis of the biggest subgraph from Experiment II (see Figure 6).

(Many Microbe Microarrays Database) is used and *E_coli_v3_Build_3* from T. Gardner Lab is built. This dataset consists of 524 arrays from 13 different collections corresponding to various conditions. The experiments were carried out on Affymetrix GeneChip E.coli Antisense Genome arrays, containing 4292 gene probes. A RMA normalization procedure was performed on the data prior to the application of our approach and the benchmark method.

Our approach REGNET and a gene relevance network method based on Partial Correlation were applied. Firstly, REGNET was applied several times with different values as a threshold of pruning phase: 25%, 50% and 100%. Second, the method proposed in [8] is used to provide partial Pearson and Spearman correlations (zeroth and first order correlations, with level α = 0.001, are calculated). Partial correlation coefficients quantify the correlation between two variables when conditioning on one or several other variables, which seems closer to causal relationships.

We chose the E.coli K12 transcriptional network in the Regulon database, version 6.3 [46] as true gene interaction network. From this transcriptional network we derived a gene association graph of 3288 interactions.

In absolute terms, there is a huge number of edges which does not correspond to any true edge from the Ecoli K12 transcriptional network. This situation shows the complexity of the gene expression regulation system. However, if we focus only on relative terms, i.e. the number of true positives divided by the size of the inferred network, we can observe that REGNET produces better results than the partial correlation-based methods. Figure 7 depicts the low proportion of true positives for each method. However, REGNET is much more selective, and builds smaller networks. For example, while 61 true edges are found in the REGNET network with 15908 interactions (0.0038), the smaller network obtained by a partial correlation-based method had 123 true edges in the network with 79372 interactions (0.0015), when using the first-order Pearson partial correlation. For zeroth-order partial correlations, the number of edges surpasses four millions of interactions.

Conclusions

Inferring any type of relationship from data is a difficult task, particularly when non-linearity is present. Gene networks provide a framework to analyze regulation and causality.

Our approach, named REGNET, generates new hypothesis of interactions among genes from gene expression data, and differs from correlation-based methods in that the relationship between one gene and others is calculated simultaneously, and statistically validated when all these genes show linear dependency only in a region of the space. Our method is based on the idea that, given some control genes which define

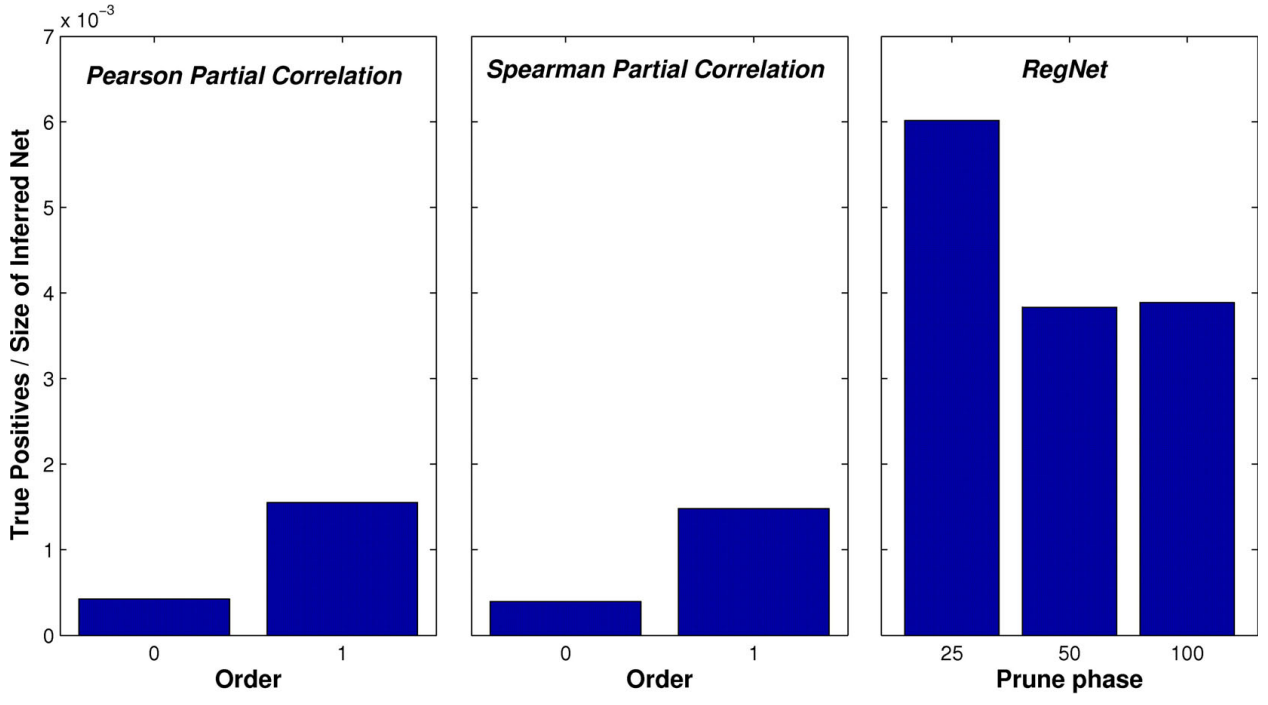


Figure 7 Escherichia coli dataset. Regulon Database is used to test the performance of REGNET and zeroth and first-order Pearson/Spearman correlation-based methods to build gene networks. The bars represent the proportion of true edges found in the gene network.

subspaces of the input data, multivariate linear models can be estimated for the target gene. REGNET strongly favours localized similarities over more global similarity, which it is one of the major drawbacks of correlation-based methods.

Experimental results show the good performance of REGNET. The first experiment, with yeast cell cycle data, is consistent with Gene Ontology. The aim of the second experiment is to check the ability of finding true gene associations from gene expression data in comparison with E.coli transcriptional network from Regulon database.

In general, REGNET is a powerful method to hypothesize on unknown relationships, and therefore, on genes potentially related to biological functions.

Additional material

Additional file 1: yeastSubNET1-8.xls. Gene-gene associations resulting from our approach using *Saccharomyces Cerevisiae* data as input. The correlation measure between pair of genes from the network is reported, together with the number of weak correlated genes detected by our approach focus on localized similarities.

Acknowledgements
This research work is partially supported by the Ministry of Science and Innovation, projects TIN2007-68084-C02-00, PCI2006-A7-0575, and by the Junta de Andalucia, projects P07-TIC-02611 and TIC-200. We are grateful to the anonymous reviewers who provided valuable feedback on our manuscript.

Author details
[1]Dpt Lenguajes y Sistemas Informaticos, Universidad de Sevilla, Seville, Spain. [2]School of Engineering, Pablo de Olavide University, Seville, Spain.

Authors' contributions
IN refined the method and designed the experiments for testing the performance of REGNET. JAR conceived the method and leaded the project. JRS critically revised the computational and statistical steps of the method. All authors read, edited and approved the final manuscript.

Received: 30 April 2010 Accepted: 15 October 2010
Published: 15 October 2010

References

1. Spellman P, Sherlock G, Zhang M, Iyer V, Anders K, Eisen M, Brown P, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.** *Molecular biology of the cell* 1998, **9(12)**:3273-3297.
2. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95**:14863-14868.
3. 'Haeseleer P, Wen X, Fuhrman S: **Mining the gene expression matrix: inferring gene relationships from large scale gene expression data.** *Proceedings of the second international workshop on Information processing in cell and tissues* 1998, 203-212.
4. hou X, Kao M, Wong W: **From the Cover: Transitive functional annotation by shortest-path analysis of gene expression data.** *Proceedings of the National Academy of Sciences* 2002, **99(20)**:12783-12788.
5. Stuart J, Segal E, Koller D, Kim S: **A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules.** *Science* 2003, **302(5643)**:249-255.
6. Lee H, Hsu A, Sajdak J, Qin J, Pavlidis P: **Coexpression analysis of human genes across many microarray data sets.** *Genome Research* 2004, **14(6)**:1085-1094.
7. Wolfe C, Kohane I, Butte A: **Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks.** *BMC Bioinformatics* 2005, **6**:227.
8. de la Fuente A, Bing N, Hoeschele I, Mendes P: **Discovery of meaningful associations in genomic data using partial correlation coefficients.** *Bioinformatics* 2004, **20(18)**:3565-3574.
9. Pearl J: *Causality: Models, Reasoning, and Inference* Cambridge, UK: Cambridge University Press 2000.
10. Shipley B: *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference* Cambridge, UK: Cambridge University Press 2002.
11. Florian M, Rainer S: **Inferring cellular networks-a review.** *BMC Bioinformatics* 2007, **8**:S5.
12. Matsuno T, Tominaga N, Arizono K, Iguchi T, Kohara Y: **Graphical Gaussian modeling for gene association structures based on expression deviation patterns induced by various chemical stimuli.** *IEICE Transactions on Information and Systems* 2006, **E89-D(4)**:1563-1574.
13. Banerjee O, El Ghaoui L, d'Aspremont A: **Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data.** *The Journal of Machine Learning Research* 2008, **9**:485-516.
14. Fitch A, Jones M: **Shortest path analysis using partial correlations for classifying gene functions from gene expression data.** *Bioinformatics* 2009, **25**:42-47.
15. Chiquet J, Smith A, Grasseau G, Matias C, Ambroise C: **SIMoNe: Statistical Inference for MOdular NEtworks.** *Bioinformatics* 2009, **25(3)**:417-418.
16. Margolin A, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera R, Califano A: **ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context.** *BMC Bioinformatics* 2006, **7(Suppl 1)**:S7.
17. Zhao W, Serpedin E, Dougherty ER: **Inferring Connectivity of Genetic Regulatory Networks Using Information-Theoretic Criteria.** *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2008, **5(2)**:262-274.
18. Qiu P, Gentles A, Plevritis S: **Fast calculation of pairwise mutual information for gene regulatory network reconstruction.** *Comput Methods Programs Biomed* 2009, **94(2)**:177-180.
19. Wilczynski B, Dojer N: **BNFinder: exact and efficient method for learning Bayesian networks.** *Bioinformatics* 2009, **25(2)**:286-287.
20. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.** *Nature Genet* 2003, **34**:166-176.
21. Joshi A, De Smet R, Marchal K, Van de Peer Y, Michoel T: **Module networks revisited: computational assessment and prioritization of model predictions.** *Bioinformatics* 2009, **25(4)**:490-496.
22. Steele E, Tucker A, 't Hoen PAC, Schuemie MJ: **Literature-based priors for gene regulatory networks.** *Bioinformatics (Oxford, England)* 2009, **25(14)**:1768-1774.
23. Mehra S, Hu W, Karypis G: **A Boolean algorithm for reconstructing the structure of regulatory networks.** *Metabolic Engineering* 2004, **6(4)**:326-339.
24. Soinov L, Krestyaninova M, Brazma A: **Towards reconstruction of gene networks from expression data by supervised learning.** *Genome Biol* 2003, **4**:R6.
25. Ponzoni I, Azuaje F, Augusto J, Glass D: **Inferring Adaptive Regulation Thresholds and Association Rules from Gene Expression Data through Combinatorial Optimization Learning.** *IEEE/ACM Trans Comput Biol Bioinformatics* 2007, **4(4)**:624-634.
26. Benjamini Y, Yekutieli D: **The control of the false discovery rate in multiple testing under dependency.** *Ann. Statist* 2001, **29(4)**:1165-1188.
27. Malerba D, Esposito F, Ceci M: **Top-down induction of model trees with regression and splitting nodes.** *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2004, **26**:1-14.
28. Morgan J, Sonquist J: **Problems in the analysis of survey data, and a proposal.** *Journal of American Statistics Society* 1963, **58**:415-434.
29. Breiman L, Friedman J, Stone C, Olshen R: *Classification and Regression Trees* Chapman & Hall/CRC 1984, **67.**
30. Quinlan J: **Learning with continuous classes.** *5th Australian Joint Conference on Articial Intelligence* 1992, 343-348.

31. Witten I, Frank E: *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* San Francisco: Morgan Kaufmann 2000.
32. Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A: **False discovery rate, sensitivity and sample size for microarray studies.** *Bioinformatics* 2005, **21(13)**:3017-3024.
33. Sheskin D: *Handbook of Parametric and Nonparametric Statistical Procedures* Boca Raton: CRC Press 2004.
34. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G: **Revealing strengths and weaknesses of methods for gene network inference.** *Proceedings of the National Academy of Sciences* 2010, **107(14)**:6286-6291.
35. Marbach D, Schaffter T, Floreano D, Prill R, Stolovitzky G: **The DREAM4 in-silico network challenge.** *Tech rep* Computer Science and Artificial Intelligence Laboratory Massachusetts Institute of Technology, Cambridge MA, USA 2009 [http://gnw.sourceforge.net/resources/DREAM4%20in%20silico%20challenge.pdf].
36. Marbach D, Schaffter T, Mattiussi C, Floreano D: **Generating Realistic In Silico Gene Networks for Performance Assessment of Reverse Engineering Methods.** *Journal of Computational Biology* 2009, **16(2)**:229-239.
37. Schafer J, Strimmer K: **A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics.** *Statistical applications in genetics and molecular biology* 2005, **4**:Article32.
38. Opgen-Rhein R, Strimmer K: **From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data.** *BMC Systems Biology* 2007, **1**:37.
39. Charbonnier C, Chiquet J, Ambroise C: **Weighted-LASSO for structured network inference from time course data.** *Statistical applications in genetics and molecular biology* 2010, **9, Article 15.**
40. Ambroise C, Chiquet J, Matias C: **Inferring sparse Gaussian graphical models with latent structure.** *Electronic Journal of Statistics* 2009, **3**:205-238.
41. Boettcher SG, Dethlefsen C: **deal: A Package for Learning Bayesian Networks.** *Journal of Statistical Software* 2003, **8(20)**:1-40.
42. Cho R, Campbell M, Winzeler E, L S, Conway A, Wodicka L, Wolfsberg T, Gabrielian A, Landsman D, Lockhart D: **A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle.** *Molecular Cell* 1998, **2**:65-73.
43. Berriz G, King O, Bryant B, Sander C, Roth F: **Characterizing gene sets with FuncAssociate.** *Bioinformatics* 2003, **19(18)**:2502-2504.
44. Westfall P, Young S: *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment* North Carolina: Wiley-Interscience 1993.
45. Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, Juhn FS, Schneider SJ, Gardner TS: **Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata.** *Nucleic acids research* 2008, , **36 Database:** D866-70.
46. Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Penaloza Spindola M, Contreras-Moreira B, Segura-Salazar J, Muniz Rascado L, Martinez-Flores I, Salgado H, Bonavides-Martinez C, Abreu-Goodger C, Rodriguez-Penagos C, Miranda-Rios J, Morett E, Merino E, Huerta A, Collado-Vides J: **RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation.** *Nucleic Acids Research* 2008, , **36 Database:** D120-4.
47. Shannon P, Markiel A, Ozier O, Baliga N, Wang J, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.** *Genome Research* 2003, **13(11)**:2498-2504.

doi:10.1186/1471-2105-11-517
Cite this article as: Nepomuceno-Chamorro *et al.*: Inferring gene regression networks with model trees. *BMC Bioinformatics* 2010 **11**:517.