

Reverse engineering molecular regulatory networks from microarray data with qp-graphs

Robert Castelo^{1,2,*} Alberto Roverato³

(this is a preprint of the publication in the Journal of Computational Biology, 16(2):213-227, 2009)

1. Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain. Email: robert.castelo@upf.edu

2. Research Program on Biomedical Informatics, Institut Municipal d'Investigació Mèdica, Barcelona, Spain

3. Department of Statistical Science, Università di Bologna, Bologna, Italy. Email: alberto.roverato@unibo.it

* Corresponding author. Dept. of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona Biomedical Research Park, Dr. Aiguader 88, E-08003 Barcelona, Spain. Tel: +34 933 160 514. Fax: +34 933 160 550. Email: robert.castelo@upf.edu

Running title: Reverse engineering regulatory networks with qp-graphs

Key words: molecular regulatory network, microarray data, reverse engineering, qp-graph

Abstract

Reverse engineering bioinformatic procedures applied to high-throughput experimental data have become instrumental to generate new hypotheses about molecular regulatory mechanisms. This has been particularly the case for gene expression microarray data where a large number of statistical and computational methodologies have been developed in order to assist in building network models of transcriptional regulation.

A major challenge faced by every different procedure is that the number of available samples n for estimating the network model is much smaller than the number of genes p forming the system under study. This compromises many of the assumptions on which the statistics of the methods rely, often leading to unstable performance figures. In this work we apply a recently developed novel methodology based in the so-called q-order limited partial correlation graphs, qp-graphs, which is specifically tailored towards molecular network discovery from microarray expression data with $p \gg n$.

Using experimental and functional annotation data from *Escherichia coli* here we show how qp-graphs yield more stable performance figures than other state-of-the-art methods when the ratio of genes to experiments exceeds one order of magnitude. More importantly, we also show that the better performance of the qp-graph method on such a gene-to-sample ratio has a decisive impact on the functional coherence of the reverse-engineered transcriptional regulatory modules and becomes crucial in such a challenging situation in order to enable the discovery of a network of reasonable confidence that includes a substantial number of genes relevant to the essayed conditions.

An R package called `qpgraph` implementing this method is part of the Bioconductor project and can be downloaded from <http://www.bioconductor.org>. A parallel standalone version for the most computationally expensive calculations is available from <http://functionalgenomics.upf.edu/qpgraph>.

1 Introduction

Building a network model of a molecular regulatory layer of the cell has become a routine task in order to scratch at the surface of the complexity of the cellular program under study. The ever increasing availability of structured high-throughput data about the cell's molecular phenotype under all sorts of experimental conditions has enabled the development of statistical and computational procedures that aid in building such models. Gene expression microarray data has been, and still is, paradigmatic for developing and assessing bioinformatic approaches for reverse engineering network models of molecular regulatory mechanisms and this can be observed just from the number of reviews about the subject that have been published during the last year (see, for instance, Bansal et al., 2007; Markowetz and Spang, 2007; Tegner and Björkegren, 2007). Gene expression profiling provides us with a matrix of n expression values measured for p genes where p is much larger than n , typically somewhere between one and three orders of magnitude.

A very popular approach to infer a transcriptional regulatory network from gene expression data consists of considering some pairwise measure of association between two expression profiles (e.g., Pearson correlation coefficient or mutual information), compute it for every pair of genes of interest (e.g., transcription-factor gene vs. target gene) and output those gene pairs with an association strength above a given threshold. One of the first applications of this approach to gene network inference was by Butte et al. (2000). However, marginal pairwise associations cannot distinguish direct from indirect (that is, spurious) relationships and specific enhancements to this pairwise approach have been made in order to address this problem (see, for instance, Basso et al., 2005; Faith et al., 2007). The simplicity of these *off the shelf* pairwise methods enables their direct application with very few samples and also allows them to scale up easily to very large gene sets.

A sensible approach is to try to apply multivariate statistical methods like undirected Gaussian graphical modeling (see Whittaker, 1990, Ch. 6) and compute partial correlations which are a measure of association between two variables while controlling for the remaining ones. However, these methods require inverting the sample covariance matrix of the gene expression profiles and this is only possible when $n > p$ (Dykstra, 1970). This problem has been addressed using the so-called moderation and regularization techniques that provide a shrinkage estimate of the inverse of the sample covariance matrix (see, for instance, Schäfer and Strimmer, 2005) enabling the estimation of partial correlation coefficients. Another choice within graphical models are Bayesian networks which are associated to acyclic directed graphs but present similar requirements that preclude their direct application on microarray data. Several strategies have been devised to approach this problem as, for instance, in the seminal work by Friedman et al. (2000), where the number of transcription factor genes that are allowed to target other genes in the network is bounded in order to enable the necessary calculations.

An analogous idea with Gaussian graphical models associated to undirected graphs is to employ limited-order partial correlations, concretely, q -order partial correlations with $q < (n - 2)$. In this way, we can carry out a test for the hypothesis of zero q -order partial correlation between two genes with the hope that if they are not directly associated, their indirect association is mediated by less or equal than q other genes. This idea has been applied in several works (de la Fuente et al., 2004; Magwene and Kim, 2004; Wille and Bühlmann, 2006) but only considering when $q = 1$ or $q = 2$. Besides the fact that $q < 3$ may be insufficiently small, it is unclear what sort of mathematical object one obtains by exclusively inferring interactions from

the outcome of pairwise-single tests of first or second-order partial correlations and how close the resulting networks may be to the full-order partial correlation network. Recently, we have addressed these two questions providing a precise definition of a q -order inferred network (the qp-graph) and a principled methodology to derive it (Castelo and Roverato, 2006). In this paper we show for the first time its application to gene expression microarray data from one of the best characterized systems, *Escherichia coli*. We compare its performance with some of the state-of-the-art methods in terms of network accuracy with respect to the *Escherichia coli* transcriptional network described in RegulonDB and in terms of functional coherence of the reverse-engineered regulatory modules. Finally, we show how in a very challenging situation with $p = 4\,205$ genes and $n = 43$ experiments a network of a reasonable nominal accuracy obtained with our method reflects a sensible functional organization of the genes under the essayed experimental conditions.

2 Results

2.1 Basic principles of qp-graphs

Let's assume we represent the underlying molecular regulatory network we want to reverse-engineer as an undirected graph denoted by $G = (V, E)$ where V is its set of vertices (each associated to a gene in our context) and E is its set of undirected edges (each associated to a molecular regulatory relationship). If we would like to interpret this graph as, for instance, a transcriptional regulatory network, we may label then those vertices that are transcription factors, put directions to the edges that connect them with non-transcription factor genes and disregard the edges between genes that are not transcription factors. However, for the time being let's assume our representation G of the molecular regulatory network is undirected and delay its biological interpretation till we infer a network from data. For some organisms, like *Escherichia coli*, we may know G partially but in general G is unknown and we want to estimate it using microarray data. By making the further assumption that microarray data form a multivariate normal sample of size n drawn from p genes in one-to-one correspondence with the vertex set V (i.e., $|V| = p$), the estimation of G could be done by calculating between every pair of vertices (i, j) the so-called partial correlation coefficient denoted by $\rho_{ij.Q}$, which is a measure of correlation between variables i and j that takes into account the variables in the set Q of size q . One can be more precise and refer to them as q -order partial correlations and if $Q = V \setminus \{i, j\}$ (i.e., $q = p - 2$) then $\rho_{ij.Q}$ are full-order partial correlations. Whenever $q = p - 2$ and $\rho_{ij.Q} = 0$ then $(i, j) \notin E$ (i.e., they are disconnected in G), otherwise vertices i and j form an edge in G (i.e., $(i, j) \in E$). Graphical model theory (see Whittaker, 1990) shows that there is a direct relationship between the statistical assertion $\rho_{ij.Q} = 0$ and the lack of an edge between vertices (i, j) in G in terms of the so-called graph separation. A pair of disconnected vertices (i, j) are separated in G by a subset Q if and only if all paths connecting i and j intersect Q and therefore $\rho_{ij.Q} = 0$ whenever i and j are separated by Q in G .

The decision of whether $\rho_{ij.Q} = 0$ in a particular microarray data set can be made on the basis of a hypothesis test for zero partial correlation, but if $q = p - 2$ then this will be most of the times impossible to calculate because normally in microarray data $p \gg n$ (Dykstra, 1970). However, if we employ a subset Q of size q such that $q < (n - 2)$ then we can test whether $\rho_{ij.Q} = 0$ with some first and second type error probabilities α and $\beta_{ij.Q}$, respectively. Note that α can be fixed by setting a desired significance level to our tests while $\beta_{ij.Q}$ is unknown

and depends on the particular pair of genes (i, j) , the subset Q and on the available effective sample size $n - q$.

At this point we can finally introduce the main concept of our methodology, the qp-graph. A q -order partial correlation graph, or qp-graph, is an undirected graph denoted by $G^{(q)} = (V, E^{(q)})$ with the same vertex set V as G and with an edge set $E^{(q)}$ such that a pair of vertices (i, j) are disconnected in $G^{(q)}$, i.e., $(i, j) \notin E^{(q)}$, if and only if there is a subset U with $|U| \leq q$ and $(i, j) \notin U$ such that U separates i from j in G (Castelo and Roverato, 2006, Def. 1). It is relatively straightforward to see that $G^{(q)}$ is always equal or larger than G , i.e., each edge in G is in $G^{(q)}$ and therefore we can interpret a qp-graph $G^{(q)}$ as an approximation to G (Castelo and Roverato, 2006, Cor. 3). In Figure 1a we may see this illustrated with an example. However, the usefulness of $G^{(q)}$ as a surrogate of G depends on how close $G^{(q)}$ is to G . Full details on these questions can be found in (Castelo and Roverato, 2006).

2.2 A new measure of association: the average non-rejection rate

In (Castelo and Roverato, 2006) we introduced a new measure of association between two variables, called the non-rejection rate, that ranges from 0 to 1 and helps in deciding what edges are present or missing from a qp-graph $G^{(q)}$. Therefore this measure depends on the particular value of q being used and can be succinctly described as follows. Let \mathcal{Q}_{ij} be the set made up of all subsets $Q \in V \setminus \{i, j\}$ such that $|Q| = q$. Let T_{ij}^q be a binary random variable associated to the pair of vertices (i, j) that takes values from the following three-step procedure: 1. an element Q is sampled from \mathcal{Q}_{ij} according to a (discrete) uniform distribution; 2. using the available data, test the null hypothesis of zero partial correlation controlling for the subset Q of size q (i.e., $H_0 : \rho_{ij,Q} = 0$); and 3. if the null hypothesis H_0 is rejected then T_{ij}^q takes value 0, otherwise takes value 1. It follows that T_{ij}^q has a Bernoulli distribution and the non-rejection rate is defined as its expectancy $E[T_{ij}^q] = \Pr(T_{ij}^q = 1)$. In order to estimate the non-rejection rate we can use the three-step procedure previously described and take the average of non-rejections, however, since there are $\binom{p-2}{q}$ elements in \mathcal{Q}_{ij} we will sample uniformly only a limited number of subsets Q from \mathcal{Q}_{ij} , for instance one-hundred.

Interestingly, it can be shown (Castelo and Roverato, 2006, Eq. 5) that the theoretical non-rejection rate equals the sum of two terms:

$$\Pr(T_{ij}^q = 1) = \beta_{ij}^q (1 - \pi_{ij}^q) + (1 - \alpha) \pi_{ij}^q, \quad (1)$$

where α is the probability of the first type error of the tests for zero partial correlation, π_{ij}^q is the proportion of subsets Q of size q in \mathcal{Q}_{ij} that separate i and j in G and β_{ij}^q is the mean value of the second type errors $\beta_{ij,Q}$ for all $Q \in \mathcal{Q}_{ij}$.

Notice that if a pair of vertices (i, j) are directly connected in G (in our context this is as much as saying that a transcription factor directly binds to the promoter region of a target gene) then they will never be separated in any qp-graph so that $\pi_{ij}^q = 0$ and therefore the theoretical non-rejection rate equals exactly β_{ij}^q which will be only very high when all the second-type errors $\beta_{ij,Q}$ are uniformly very high over \mathcal{Q}_{ij} . Since the statistical power of the tests for zero partial correlation with null hypothesis $H_0 : \rho_{ij,Q} = 0$ equals $1 - \beta_{ij,Q}$, the non-rejection rate of an edge (i, j) in G corresponds to the one minus the average statistical power to detect that association. Therefore, in this case, it follows that a high value of the non-rejection rate implies a uniformly low statistical power throughout all $Q \in \mathcal{Q}_{ij}$ and thus that such an edge is very difficult to detect with the available sample size. In (Castelo and Roverato, 2006, Sec. 5.1) it

is shown that for connected pairs of vertices (i, j) in G the non-rejection rate converges to 0 as $n - q$ increases while for disconnected vertices (i, j) in G the term π_{ij}^q increases as q grows large and concretely when $q = p - 2$ then $\Pr(T_{ij}^q = 1) = 1 - \alpha$. We can see this effect in Figure 1b where we have plotted the distribution of the non-rejection rate calculated from two sets of synthetic data sampled from a network with $p = 150$ genes each of them connected on average to 5 other genes. In these simulations we have used $q = 20$ and the boxplots from the left correspond to one data set where we sampled $n = 200$ observations and the ones on the right correspond to $n = 50$.

Obviously, a key question when using the non-rejection rate with microarray data is what value of q should we use. We know that a large value of q should provide a qp-graph $G^{(q)}$ closer to G but this may be compromised by the available statistical power which depends on $n - q$. In the context of microarray data where $p \gg n$ we propose to average (taking the arithmetic mean), for each pair of genes, the estimates of the non-rejection rates for different values of q spanning its entire range from 1 to somewhere close $n - 3$. Averaging the non-rejection rate $E[T_{ij}^q]$ over q is a sensible way of accounting for the uncertainty we have about a proper q value because when i and j are connected in G then π_{ij}^q is 0 for every q value and therefore the resulting quantity equals the average probability of the second type error. In this case, the average non-rejection rate has the same interpretation as the non-rejection rate defined for a particular q value and converges to 0 as $n - q$ increases. Conversely, when a pair of vertices (i, j) are not connected in G averaging the non-rejection rate has the following consequences. Let q_{ij}^* be the size of the smallest subset(s) separating the pair of vertices (i, j) . When we average through values $q < q_{ij}^*$ then we have a similar situation as if (i, j) were connected in G because $\pi_{ij}^q = 0$ and the non-rejection rate equals the average probability of the second type error β_{ij}^q preventing us from seeing that (i, j) is actually not present in G . On the other hand, when we average through values $q \geq q_{ij}^*$ it holds that $\pi_{ij}^q > 0$ and moreover $\pi_{ij}^{q_1} \geq \pi_{ij}^{q_2}$ for every two $q_1 > q_2$. As shown in (Castelo and Roverato, 2006, Sec. 5.1), in this case, the non-rejection rate belongs to the interval $(\beta_{ij}^q, (1 - \alpha))$ and, although it can take any value in such interval, it is important to notice that it will be closer to the boundary $(1 - \alpha)$ for larger values of π_{ij}^q . In particular, when i and j belong to different connected components of G (i.e., when there is no path in G connecting i and j) then $\pi_{ij}^q = 1$ and the theoretical average non-rejection rate equals $(1 - \alpha)$.

We can conclude that the average non-rejection rate is more stable than the non-rejection rate, avoids having to specify a particular value of q and it behaves similarly to the non-rejection rate for connected pairs of vertices in the true underlying graph G (i.e., for interacting genes in the underlying molecular regulatory network). However, *there is no free lunch* and the drawback of averaging is that a disconnected pair of vertices (i, j) in a graph G with a large value of q_{ij}^* will be easier to identify with the non-rejection rate using a particular value $q \geq q_{ij}^*$ than with the average non-rejection rate which will have as a consequence an increase in the number of false positive predicted interactions. Fortunately, there is a growing body of literature (see, for instance, Barabasi and Oltvai, 2004) providing evidence that molecular regulatory networks show high degrees of modularity and sparseness and for this reason we can expect that q_{ij}^* should not be very large.

2.3 Performance comparison with RegulonDB

Escherichia coli (*E. coli*) is the free-living organism for which a largest part of its transcriptional regulatory network is supported by some sort of experimental evidence. As a result of an effort in combining all this evidence the database RegulonDB (Gama-Castro et al., 2008) provides a curated set of transcription factor and target gene relationships that we have used as a gold-standard to assess the performance of the qp-graph method in comparison with some of the current state-of-the-art approaches, concretely, CLR (Faith et al., 2007), ARACNE (Basso et al., 2005) and GeneNet (Schäfer and Strimmer, 2005). Additionally, in a similar vein as the relevance network approach from Butte et al. (2000), we have used as a method the calculation of the Pearson correlation coefficient between all pairs of genes which we shall call the Pairwise-PCC method and, finally, as a baseline comparison we have used the assignment of a random uniform number between -1 and 1, mimicking a random correlation, to every pair of genes and labeled it as the Random method. For each of these methods we produced a ranking of all possible transcription-factor and target gene pairs according to the score that the method assigns to each interaction. In order to compare the accuracy of the methods with respect to RegulonDB we built precision-recall curves (see Fawcett, 2006) by going through each ranking from the top to the bottom calculating the precision of the method as a function of the recall in fixed recall steps of 0.5% (see Methods). We should, however, keep in mind that both estimates of precision and recall might be biased due to the incompleteness of RegulonDB which, as reported in its latest release (Gama-Castro et al., 2008), contains some information on transcriptional regulation for about one third of the genes, and therefore we can only expect and assume that these two figures are still representative enough.

We have assessed the performance of the methods in two different situations. A first one, more “optimistic”, where we have many available microarray experiments and a number of genes that does not exceed them in more than one order of magnitude, and a second one, more challenging and closer to the reality of microarray data availability for most living organisms, where we have a number of genes that exceeds the number of experiments in more than one order of magnitude.

We have arranged the first situation by using one of the largest available sets of microarray data for *E. coli*, a compendium of $n = 380$ experiments from the latest release of the M3d database (Faith et al., 2008). In order to work with a gene set whose size would not exceed these $n = 380$ experiments in more than one order of magnitude we have restricted the initial *E. coli* gene set from the microarray chip to those genes that participate in at least one transcriptional regulatory relationship from RegulonDB. This leaves us with a gene set of $p = 1428$ genes participating in 3283 transcriptional regulatory relationships (see Methods for further details on the filtering process). When using the qp-graph method we have averaged the non-rejection rates across 16 different q values that go from 1 to 350 and in Figure 2a we may see the resulting precision-recall curves for each of these individual q values jointly with the curve resulting from using the average non-rejection rate and, as a baseline comparison, we have added the Random method. Observe that for $q = 1$ its precision-recall curve is nearly as bad as the Random method while the average non-rejection seems to provide a nice trade-off between all the best-performing q values. In Figure 2b we see the precision-recall curve for the average non-rejection rate (labeled qp-graph) in comparison with the other methods. All the methods perform similarly with ARACNE and CLR providing up to 15% better precisions at 1%-1.5% and 4%-4.5% recall intervals.

To set up the second situation we have used a microarray data set from the NCBI Gene Expression Omnibus (Barrett et al., 2007) with accession GDS680 corresponding to 43 experiments of various mutants under oxygen deprivation (Covert et al., 2004) and used the full gene set of *E. coli* with $p = 4\,205$ genes (see Methods for details on filtering steps). In this case, precision-recall curves are calculated on the subset of 1 428 genes forming the 3 283 RegulonDB interactions. In Figure 2c we have again the precision-recall curves for several q values and we can observe again how the average non-rejection rate provides a smoother precision-recall curve. It is interesting also to observe in this case what the outcome is when using the largest possible q value ($q = n - 3 = 40$) which provides the worst precision-recall curve probably due to an absolute lack of statistical power which depends on $n - q$. Note, however, that the average non-rejection rate is robust to the inclusion of values from $q = 40$ as its precision-recall curve still displays a good trade-off between all the best-performing q values. We have, in fact, calculated the precision-recall for the average non-rejection rate excluding values from $q = 40$ and the precision improvement in the resulting curve is limited to less than 5% between 0.5% and 1.5% recalls (data not shown). In Figure 2d we have the comparison of the average non-rejection rate (labeled qp-graph) with the other methods and, in this case, we see a dramatic drop in performance for GeneNet, ARACNE and Pairwise-PCC ($\sim 50\%$ precision loss at 1% and 3% recall), an important drop for CLR (24% and 43% precision losses at 1% and 3% recalls resp.) and more moderate drops for qp-graph (14% and 37% precision losses at 1% and 3% recalls resp.). As a consequence of the less pronounced drops in performance from qp-graph we have that for precision levels between 40% and 80% the qp-graph method doubles the recall with respect to the other methods. We will see later that this has an important impact when targeting a network of a reasonable nominal precision in such a data set with $p = 4\,205$ and $n = 43$.

2.4 Assessment of functional coherence with Gene Ontology

A critical question when reverse engineering a molecular regulatory network is to know the extent to which the inferred regulatory relationships reflect the functional organization of the system under the experimental conditions employed to generate the microarray data. We have addressed this question using the output of the methods previously applied to the genome-wide oxygen deprivation data from Covert et al. (2004) and the Gene Ontology database (<http://www.geneontology.org>) which provides structured functional annotations on genes for a large number of organisms including *E. coli*. The approach we have followed consists of assessing the functional coherence of each regulatory module within a given network. We define here a regulatory module as a transcription factor and its set of regulated genes and assess its functional coherence by relying on the observation that for many transcription factor genes, their biological function, beyond regulating transcription, is related to the genes they regulate (Zhou et al., 2005). Note that different regulatory modules may form part of a common pathway and thus share some more general functional annotations which can lead to some degree of functional coherence between target genes and transcription factors of different modules. However, we expect functional coherence to still be tighter within a regulatory module than between modules in a pathway and this makes it also an appealing measure for assessing the discriminative power between direct and indirect interactions independently from the network accuracy assessment with RegulonDB. We shall see below that our data leading to Figure 3a verifies empirically this conjecture.

Using Gene Ontology (GO) annotations, concretely those that refer to the biological process ontology, we build two GO graphs where in each of them vertices are GO terms and (directed) links are GO relationships. One GO graph is induced (i.e., grown towards vertices representing more generic GO terms) from GO terms annotated on the transcription factor gene discarding those terms related to transcription factor regulation. The other GO graph is induced from GO terms over-represented among the regulated genes in the regulatory module. These over-represented GO terms are found by using the conditional hyper-geometric test implemented in the `GOstats` package from Bioconductor (Falcon and Gentleman, 2007). We calculate the level of functional coherence of the regulatory module as the degree of similarity between the two GO graphs which, in this case, amounts to a comparison of the two corresponding subsets of vertices (see Methods). The level of functional coherence of the entire network is determined by the distribution of the functional coherence values of all the regulatory modules for which this measure was calculated¹.

Similarly to what happens with RegulonDB, current functional genome-wide annotations, as those from the GO consortium, exist only for a subset of the genes. In the case of GO, most available annotations are computationally derived without manual curation (Rhee et al., 2008) and hence potentially including some fraction of false positives. Therefore, we cannot expect that the current functional annotation of a transcription factor follows exactly that of its regulated genes and vice versa and thus it becomes necessary to have an estimate of what functional coherence levels can we expect from a top-performing reverse engineering procedure. We have addressed this by calculating functional coherence values for the regulatory modules in the transcriptional network from RegulonDB. Moreover, in order to assess whether functional coherence is higher within a regulatory module than when the module includes genes indirectly related we have repeated this calculation introducing increasing levels of noise in the following way. We define here noise as the fraction of target genes in a regulatory module of RegulonDB that are replaced by other genes outside the regulatory module. We considered five levels of noise ranging from 0% (the original RegulonDB network) to 100% (the entire target set being replaced in every regulatory module) in steps of 25%. Given a noise level, for each regulatory module we randomly sample a subset of the corresponding size defined by the noise level from the target gene set and let this size be noted by s . This subset is then replaced by the first s genes outside the regulatory module with largest absolute Pearson correlation coefficients with respect to the transcription factor, trying to reproduce what a simple network inference algorithm would do when picking a wrong target for a given transcription factor. This is nevertheless an optimistic simulation as, in general, we do not know *a priori* what the size of the regulatory module is.

We can see the results in Figure 3a and observe that indeed functional coherence values are far from being distributed close to the perfect coherence (i.e., value 1) and that they decrease as the noise level increases. This decrease is about 30% between the original RegulonDB network (0% noise) and the completely noisy network (100% noise) where every regulatory module has its entire target gene set replaced by genes outside the module. One might have expected that in this latter network functional coherence levels would be distributed close to zero but we hypothesize that this is not the case because highly-correlated genes (with respect to a transcription factor, see caption of Figure 3), although not in the same regulatory module, are likely to be part of the same pathway and thus sharing some functional annotation with the

¹This depends on the availability of GO functional annotations for genes (see Methods).

transcription factor. This phenomenon is likely to be enhanced by the fact that the sizes of the original regulatory modules are preserved in this exercise leading to an underestimation of the negative effects of false positives in functional coherence. We can, therefore, make the observation that wrongly predicted transcription regulatory relationships, even when displaying high pairwise correlations, have a negative impact in the functional coherence of the corresponding regulatory module and thus we can expect that the currently available GO annotations for *E. coli* should provide reasonably good estimates of functional coherence for the purpose of comparing different reverse engineering approaches.

In order to compare functional coherence between methods, we have considered three different strategies to reverse-engineer a transcriptional regulatory network: 1. using a nominal RegulonDB-precision of 50%; 2. using a nominal RegulonDB-recall of 3%; and 3. using the top-scored 1 000 interactions. In Figure 3 (b,c,d) we may see boxplots for the functional coherence levels of the networks obtained from each method and selection strategy. Through the three different strategies, the qp-graph method is the one displaying larger values of functional coherence, particularly the mean and median values of functional coherence are the largest among all the methods. Functional coherence values of the qp-graph method are also most similar to the ones from RegulonDB, specially for the 50%-precision network. In this case, the distribution of functional coherence is made out of just 3 values due to the limited recall at this precision level and the lack of GO annotations (see Figure 3b). Concretely, these values are 0.54, 0.33 and 0.13 corresponding to the regulatory modules of *mhpR*, *appY* and *glcC*, respectively. While three data points are not necessarily representative of the distribution they come from, it is, however, remarkable that most of the interactions in two of these three modules (*mhpR* and *glcC*) are novel (see Figure 4).

Note from Figure 2d that at 3% recall the precision for qp-graph is about 30%, for CLR about 20% and for the rest about 10%, while in the third strategy the top-scored 1000 interactions yield a precision figure between 10% to 15% for all the methods. This decrease in precision is clearly correlated with a decrease in functional coherence with the qp-graph method while this trend does not exist with any of the other methods, in particular with the Random method whose distribution largely overlaps with all the methods except with qp-graph at 50% precision where this overlap is the smallest among all networks. Note also that at 50% precision GeneNet could not provide any predicted regulatory module with more than 4 target genes while CLR, ARACNE and Pairwise-PCC just a few of them (respectively, 18, 18 and 13 modules where 5, 4 and 3 had 5 or more target genes) and only qp-graph has provided a network of a reasonable size (46 modules with 9 having 5 or more target genes) for further analysis. Of course, the Random method scatters uniformly all the predictions along the ranking and this leads to the largest networks at all precision levels.

2.5 Analysis of the 50%-precision qp-graph regulatory network

Finally, we take a closer look to the 50%-precision qp-graph transcriptional regulatory network. This network consists of 147 genes involved in 125 transcriptional regulatory relationships. Since the dimension of the undirected graph (its maximum clique size) formed by these 125 edges is now below the number of experiments $n = 43$ we have been able to carry out, using standard procedures (Lauritzen, 1996), a maximum likelihood estimation of the partial correlation coefficients which we shall use to provide information on the direction of the inferred correlations. The 125 regulatory relationships forming the network are organized into 38 con-

nected components containing 46 transcription factor genes where most of them (25) regulate a single target gene, 18 regulate between 2 and 11 targets and three of them (*gadE*, *glcC*, *mhpR*) regulate, respectively, 12, 13 and 15 targets. In Figure 4 we can see the connected components from this network with at least 5 genes and the *soxS* regulatory module with one single target.

The microarray experiments from Covert et al. (2004) monitor the response from *E. coli* during an oxygen shift targeting the *a priori* most relevant part of the network by using six strains with knockouts of key transcriptional regulators in the oxygen response ($\Delta arcA$, $\Delta appY$, Δfnr , $\Delta oxyR$, $\Delta soxS$ and the double knockout $\Delta arcA\Delta fnr$). Many of the metabolic transitions between aerobic and anaerobic growth states are controlled at transcriptional level by the activity of the transcription factors *fnr* and *arcA* which are presumed to be inactive under aerobic conditions (Salmon et al., 2003). The five knocked-out transcription factors form part of the 50%-precision qp-graph network except for *fnr* (see Figure 4). This might be due to the observation made by Covert et al. (2004) by which for *arcA*, the expression of the regulatory protein correlates positively with its activity while for *fnr* when the protein is activated under anaerobic conditions, which is when *fnr* is known to be active, then its mRNA level is significantly reduced. This underscores the limitations of using bioinformatic procedures based only in co-expression, like the ones described in this article, to decipher transcriptional regulatory networks.

We have searched for enriched GO terms among the 147 genes of this network which we provide in Table 1. They reflect three broad functional categories one being transcription which is the most enriched but it is also probably a byproduct of the network models themselves that are anchored on transcription factor genes. The other two are metabolism and response to an external stimulus, which are central among the biological processes that are triggered by an oxygen shift. Particularly related to this, is the fatty acid oxidation process (fifth most enriched category) since fatty acid metabolism is crucial to allow the cell to adapt quickly to environmental changes and allows *E. coli* to grow under anaerobic conditions (Cho et al., 2006).

3 Discussion

We have applied a novel methodology for reverse engineering molecular regulatory networks from microarray data where the number of genes p is much larger than the number of experiments n , based on limited-order partial correlations. In a previous work (Castelo and Roverato, 2006) we had elaborated the necessary theory and concepts that lead to the so-called q -order partial correlation graphs, qp-graphs, and a new measure of association to learn them from data, the non-rejection rate. This measure depends on the order q of the correlations that is going to be employed and in this article we show that averaging this measure through different q orders is a simple but sensible way of avoiding to choose a particular one and thus eases its application to microarray data. In fact, the non-rejection rate itself is based on a simple uniform subset sampling procedure that makes its implementation straightforward and we are investigating the development of more effective quantities for learning qp-graphs, based on different, and possibly more sophisticated, sampling schemes.

In our performance assessment with *E. coli* and RegulonDB we have seen that when the ratio of genes to sample size rises from one to two orders of magnitude the qp-graph method provides a more stable performance due to its specific design towards the modeling of the

small n large p situation. This method allows us to see the problem of $n - p < 0$ as $n - p = (n - q) - (p - q)$ so that we can explore the tradeoff between a small value of q that would provide a $(n - q)$ large enough to guarantee the necessary statistical power, and a large value of q that would provide a $(p - q)$ small enough so that the resulting network is as accurate as possible.

We have also seen that the networks obtained with the qp-graph method display larger values of functional coherence between each transcription factor and its set of regulated genes and that these values approach closer the estimates derived from the RegulonDB network. In a challenging situation as the one we have approached with $n = 43$ experiments on oxygen deprivation in *E. coli* and $p = 4\,205$ genes, many regulatory modules in our networks may show functional coherence levels similar to those seen in networks built at random. More dramatically, we have seen with this particular data set that when targeting a network of a reasonable precision most of the methods could not output a minimum-sized network for further analysis and only the qp-graph method provided a network of 147 genes and 125 transcriptional regulatory relationships. The further analysis of this network confirmed that its functional organization was capturing some of the most important processes under the essayed conditions.

A key element in the calculation of functional coherence is the enrichment of GO terms within the target gene set which we have done using a conditional hyper-geometric test. This test assumes, in our context, that genes occur independently in the regulatory module which is clearly not true. Goeman and Bühlmann (2007) show, in the context of analysis of differential expression, that dependent genes lead to anti-conservative P-values in the hyper-geometric test and thus eventually to falsely enriched GO terms. Note that a GO term that is erroneously identified as over-represented because of the correlation between target genes, but that it is not part of the transcription factor functional annotation, will lead to smaller values of functional coherence on that regulatory module. Therefore, the particular use made here of the hyper-geometric test, and the resulting enriched GO terms, penalize those methods that associate genes to a transcription factor because of the correlation between genes, rather than because they are directly regulated by the transcription factor. Thus, although further research is required to develop a proper test for GO enrichment in the context of functional coherence, the violation of the independence assumption in the hyper-geometric test should, in principle, reward those methods that better discriminate between direct and indirect associations. In this respect we know, and we have seen here, that methods based on pairwise associations have more difficulties to distinguish direct from indirect associations, leading to higher rates of false positive predictions. Thus, the better performance of the qp-graph method in both network accuracy and functional coherence strengthens the idea that q -order partial correlations are a very useful tool to distinguish direct from indirect associations among genes in microarray data.

All together, the qp-graph and the average non-rejection rate constitute a principled and effective methodology for reverse engineering molecular regulatory networks from microarray data allowing one to approach this challenging task in an intuitive way through the concept of q -order partial correlations which link sparseness of the underlying network with statistical power.

4 Materials and methods

4.1 Performance assessment on different reverse engineering methods

GeneNet, ARACNE and CLR have been applied with default parameters and particularly with CLR, following the advice on its documentation, we have used the Rayleigh approximation with the data set of $p = 1\,428$ genes and the normal approximation with $p = 4\,205$ genes. Before calculating precision-recall curves the scores for all non-transcription factor and target gene pairs were set to the minimum score of the corresponding method.

Following the conventions from Fawcett (2006), when using RegulonDB interactions for comparison the recall (also known as sensitivity) is defined as the fraction of RegulonDB interactions that are present above a particular score threshold and the precision (also known as positive predictive value) is defined as the number of predicted interactions that form part of RegulonDB over the number of predicted interactions whose genes belong to at least one transcription factor and target gene relationship in RegulonDB.

4.2 E. coli functional and microarray data processing

A goal in our methodology has been to be able to perform all the analyses, including those from our own procedures, within the R and Bioconductor platforms (Gentleman et al., 2004). For this reason we have worked at all times with Entrez Gene Identifiers (Entrez IDs) and built an annotation package for the EcoliASv2 Affymetrix array platform using the AnnBuilder tool from Bioconductor.

We downloaded the Release 6.1 from RegulonDB formed by an initial set of 3 472 transcriptional regulatory relationships. We translated the Blattner IDs into Entrez IDs, discarded those interactions for which an Entrez ID was missing in any of the two genes and did the rest of the filtering using Entrez IDs. We filtered out those interactions corresponding to self-regulation and among those conforming to feedback-loop interactions we discarded arbitrarily one of the two interactions. Some interactions were duplicated due to a multiple mapping of some Blattner IDs to Entrez IDs, in that case we removed the duplicated interactions arbitrarily. We finally discarded interactions that did not map to genes in the array and were left with 3 283 interactions involving a total of 1 428 genes.

The M3d data (Faith et al., 2008, Version 4, Build 5) was readily pre-processed using RMA and we have obtained RMA expression values for the (Covert et al., 2004) data using the `rma()` function from the `affy` package in Bioconductor. We filtered out those genes for which there was no Entrez ID and when two or more probesets were annotated under the same Entrez ID we kept the probeset with highest median expression level. This left a total of $p = 4\,205$ probesets mapped one-to-one with Entrez genes.

4.3 Functional analysis with Gene Ontology

We searched for Gene Ontology (GO) terms enriched in gene sets using E. coli GO annotations from the `org.EcK12.eg.db` Bioconductor package and the conditional hyper-geometric test from the `GOSTATS` Bioconductor package (Falcon and Gentleman, 2007). All tests were done on the Biological Process (BP) ontology and we filtered out from the genome-wide set of $p = 4\,205$ genes those for which there was no BP annotation leaving a gene universe of 1 955

genes. Each set of genes to which the enrichment test was applied was also filtered to consider only those with BP annotations.

For the calculation of functional coherence, only those regulatory modules with at least 5 target genes with BP annotations were considered. The functional coherence was calculated as the similarity between the GO graphs induced, in one hand, by the transcription factor GO annotations (excluding those containing the word “transcription”) and, on the other hand, by the enriched GO terms ($P\text{-value} \leq 0.05$) among the target genes. This similarity was calculated as the size of the intersection of the node sets divided by the size of the union of the node sets as implemented in the function `simUI` from the `GOstats` Bioconductor package. In <http://functionalgenomics.upf.edu/supplements/qpgraph> can be found the R-scripts to reproduce these calculations.

Acknowledgements

This work is supported by the Spanish Ministerio de Ciencia e Innovación (MICINN) [TIN2008-00556 / TIN] and the ISCIII COMBIOMED Network [RD07/0067/0001]. Robert Castelo is a research fellow of the “Ramon y Cajal” program from the Spanish MICINN [RYC-2006-000932]. The second author acknowledges support from the Ministero dell’Università e della Ricerca [PRIN-2007AYHZWC, FISR MITICA]. The authors thankfully acknowledge the computer resources, technical expertise and assistance provided by the Barcelona Supercomputing Center - Centro Nacional de Supercomputación.

Disclosure statement

No competing financial interests exist.

References

- Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Mol Syst Biol*, 3:78.
- Barabasi, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell’s functional organization. *Nat Rev Genet*, 5(2):101–113.
- Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M., and Edgar, R. (2007). NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res*, 35(Database issue):D760–5.
- Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nat Genet*, 37(4):382–390.
- Butte, A. J., Tamayo, P., Slonim, D., Golub, T. R., and Kohane, I. S. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci U S A*, 97(22):12182–12186.

- Castelo, R. and Roverato, A. (2006). A robust procedure for Gaussian graphical model search from microarray data with p larger than n . *J Mach Learn Res*, 7:2621–2650.
- Cho, B.-K., Knight, E. M., and Palsson, B. O. (2006). Transcriptional regulation of the *fad* regulon genes of *Escherichia coli* by arca. *Microbiology*, 152(Pt 8):2207–2219.
- Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J., and Palsson, B. O. (2004). Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429(6987):92–96.
- de la Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574.
- Dykstra, R. L. (1970). Establishing positive definiteness of sample covariance matrix. *Ann Math Statist*, 41:2153–2154.
- Faith, J. J., Driscoll, M. E., Fusaro, V. A., Cosgrove, E. J., Hayete, B., Juhn, F. S., Schneider, S. J., and Gardner, T. S. (2008). Many microbe microarrays database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res*, 36(Database issue):D866–70.
- Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., and Gardner, T. S. (2007). Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol*, 5(1):e8.
- Falcon, S. and Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–258.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recogn Lett*, 27:861–874.
- Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *J Comput Biol*, 7(3-4):601–620.
- Gama-Castro, S., Jimenez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Penaloza-Spinola, M. I., 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. M., Trevino-Quintanilla, L., and Collado-Vides, J. (2008). RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res*, 36(Database issue):D120–4.
- Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80.
- Goeman, J. J. and Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987.

- Lauritzen, S. (1996). *Graphical models*. Oxford University Press.
- Magwene, P. M. and Kim, J. (2004). Estimating genomic coexpression networks using first-order conditional independence. *Genome Biol*, 5(12):R100.
- Markowetz, F. and Spang, R. (2007). Inferring cellular networks—a review. *BMC Bioinformatics*, 8 Suppl 6:S5.
- Rhee, S. Y., Wood, V., Dolinski, K., and Draghici, S. (2008). Use and misuse of the gene ontology annotations. *Nat Rev Genet*, 9(7):509–515.
- Salmon, K., Hung, S.-p., Mekjian, K., Baldi, P., Hatfield, G. W., and Gunsalus, R. P. (2003). Global gene expression profiling in Escherichia coli K12. the effects of oxygen availability and *fnr*. *J Biol Chem*, 278(32):29837–29855.
- Schäfer, J. and Strimmer, K. (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764.
- Tegner, J. and Björkegren, J. (2007). Perturbations to uncover gene networks. *Trends Genet*, 23(1):34–41.
- Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. John Wiley & Sons.
- Wille, A. and Bühlmann, P. (2006). Low-order conditional independence graphs for inferring genetic networks. *Stat Appl Genet Mol Biol*, 5:Article1.
- Zhou, X. J., Kao, M.-C. J., Huang, H., Wong, A., Nunez-Iglesias, J., Primig, M., Aparicio, O. M., Finch, C. E., Morgan, T. E., and Wong, W. H. (2005). Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat Biotechnol*, 23(2):238–243.

Tables

Table 1. Gene Ontology biological process terms enriched ($P\text{-value} \leq 0.05$) among the 147 genes forming the 50%-precision qp-graph network inferred from the (Covert et al., 2004) oxygen deprivation data.

GOBPID	P-value	Odds Ratio	Exp. Count	Count	Size	Term
GO:0006350	< 0.0001	4.32	14.39	39	293	transcription
GO:0009063	0.0003	4.24	3.24	11	66	amino acid catabolic process
GO:0009059	0.0013	1.96	27.15	41	553	macromolecule biosynthetic process
GO:0043283	0.0019	1.89	32.16	46	655	biopolymer metabolic process
GO:0019395	0.0027	5.10	1.47	6	30	fatty acid oxidation
GO:0030258	0.0027	5.10	1.47	6	30	lipid modification
GO:0044238	0.0055	2.08	71.35	82	1453	primary metabolic process
GO:0044270	0.0072	2.51	5.50	12	112	nitrogen compound catabolic process
GO:0006542	0.0107	8.53	0.49	3	10	glutamine biosynthetic process
GO:0009268	0.0134	19.76	0.20	2	4	response to pH
GO:0019752	0.0281	1.67	16.40	24	334	carboxylic acid metabolic process
GO:0042594	0.0473	4.25	0.83	3	17	response to starvation

Figure legends

Figure 1. The qp-graph and the non-rejection rate. **(A)** A graph G and its corresponding qp-graphs $G^{(q)}$ for $q = \{0, 1, 2\}$. Note that while the maximum order q in this example is 3, an order of $q = 2$ is sufficient to obtain a qp-graph that coincides with G . **(B)** The left pair of boxplots show the non-rejection rate for present and missing edges from a given graph from which we sampled $n = 200$ observations while the right pair of boxplots show the non-rejection rate for synthetic data from this same graph but this time sampling only $n = 50$ observations.

Figure 2. Performance comparison of different reverse engineering methods with respect to RegulonDB. **(A, C)** Precision-recall curves resulting from the use of different values of q in the calculation of the non-rejection rate for qp-graphs. **(B, D)** Precision-recall curves comparing different methods. Plots (A) and (B) are derived from a microarray data set with $p = 1428$ genes and $n = 380$ experiments from the M3d database while plots (C) and (D) are derived from a microarray data set with $p = 4205$ and $n = 43$ experiments from NCBI GEO GDS680.

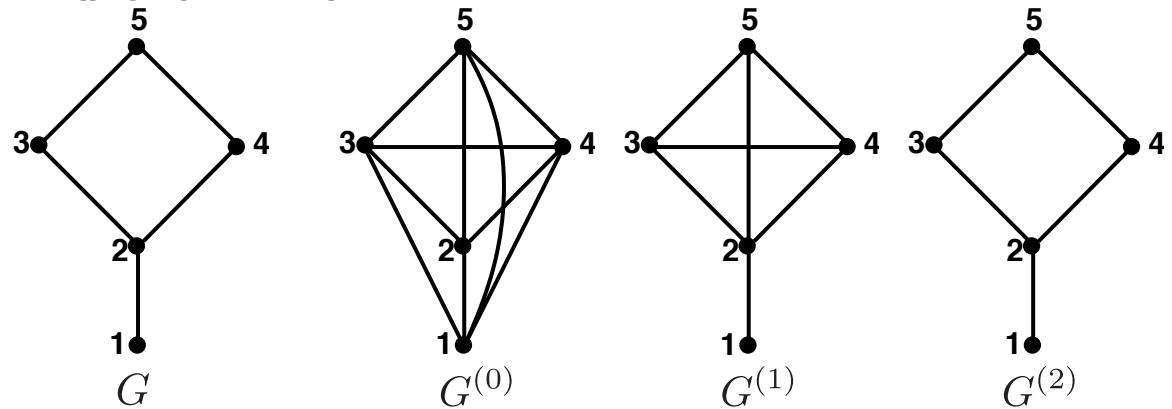
Figure 3. Functional coherence. **(A)** Distribution of functional coherence values in the RegulonDB transcriptional network with different levels of noise. Solid lines within boxes indicate medians and diamonds indicate means. On the x-axis and between square brackets, under positive noise levels, are indicated minimum, maximum and mean values, respectively, of the absolute Pearson correlation coefficients between each of the genes introducing noise and the transcription factor of the corresponding regulatory module. **(B, C, D)** Distribution of values of functional coherence for reverse-engineered networks from different methods and applying three different strategies: (B) a nominal RegulonDB-precision of 50%, (C) a nominal RegulonDB-recall of 3%, and (D) using the top ranked 1000 interactions. On the x-axis and between square brackets, under each method, are indicated, respectively, the total number of regulatory modules of the network, the number of them with at least 5 genes and the number of them with at least 5 genes with GO-BP annotations. Among this latter number of modules, the number of them where the transcription factor had GO annotations beyond transcription regulation is noted above between parentheses by n and corresponds to the number of modules on which functional coherence could be calculated.

Figure 4. Connected components with 5 or more genes from the 50%-precision qp-graph network, including the single-target *soxS* regulatory module. This network was obtained from the (Covert et al., 2004) oxygen deprivation data set of $n = 43$ experiments and $p = 4205$ genes. Arrows indicate transcriptional regulatory relationships between a transcription factor gene and a non-transcription factor target gene, and bidirected arrows indicate these relationships between two transcription factor genes. Positive (+) and negative (-) signs are labeling the actual annotated direction of the interaction in RegulonDB whenever the estimated partial correlation did not agree with it.

Figures

Figure 1

A qp-graph examples



B Non-rejection rate

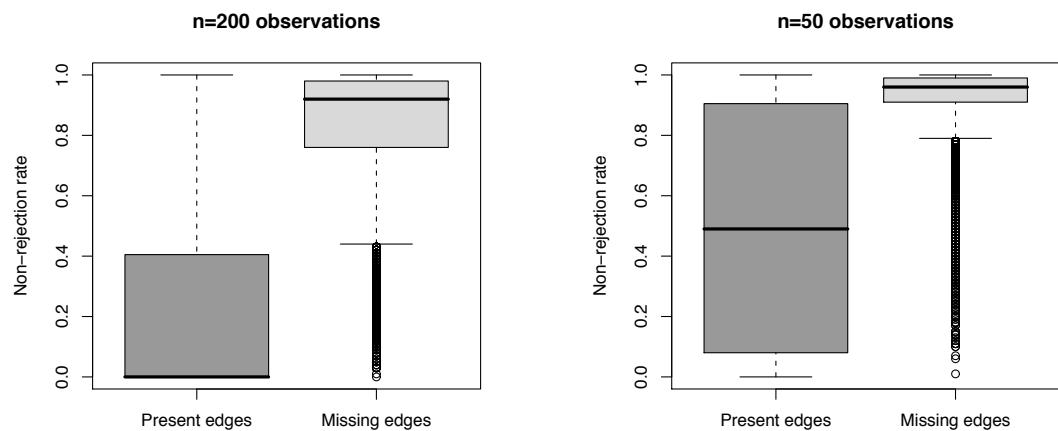
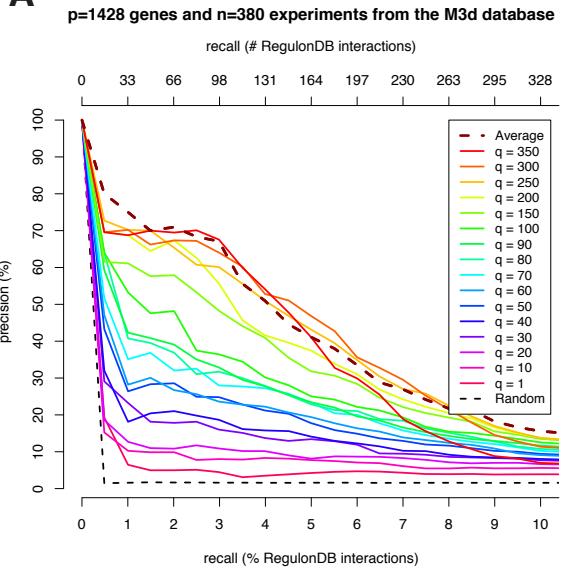
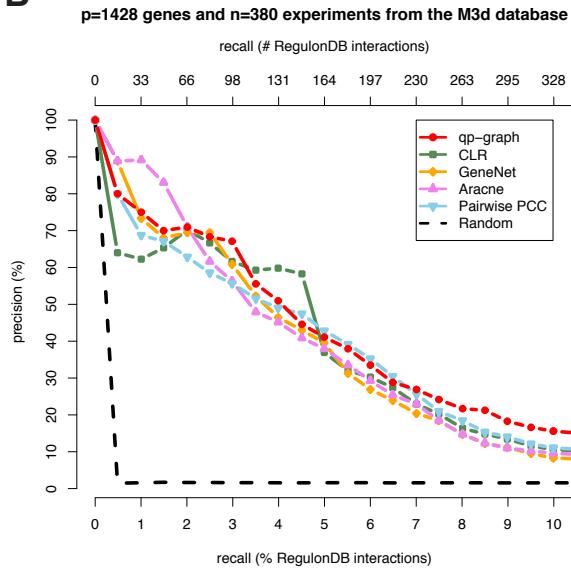


Figure 2

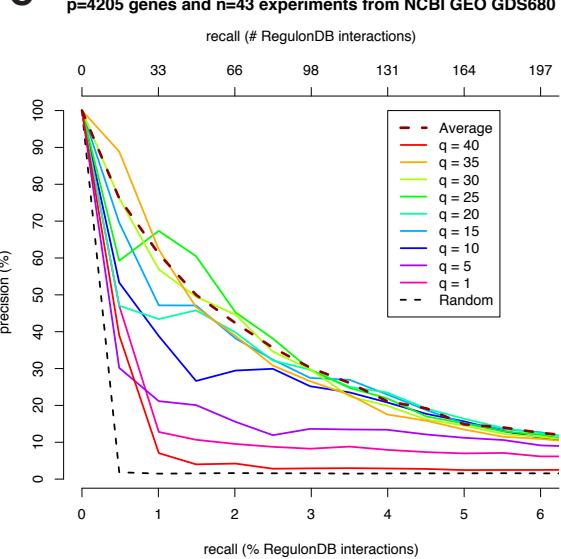
A



B



C



D

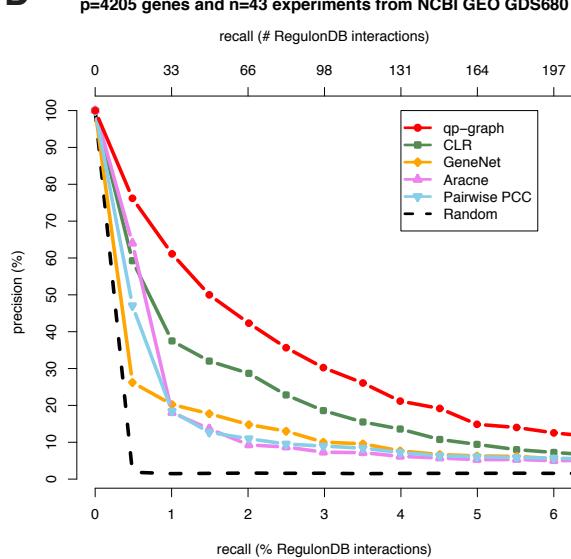


Figure 3

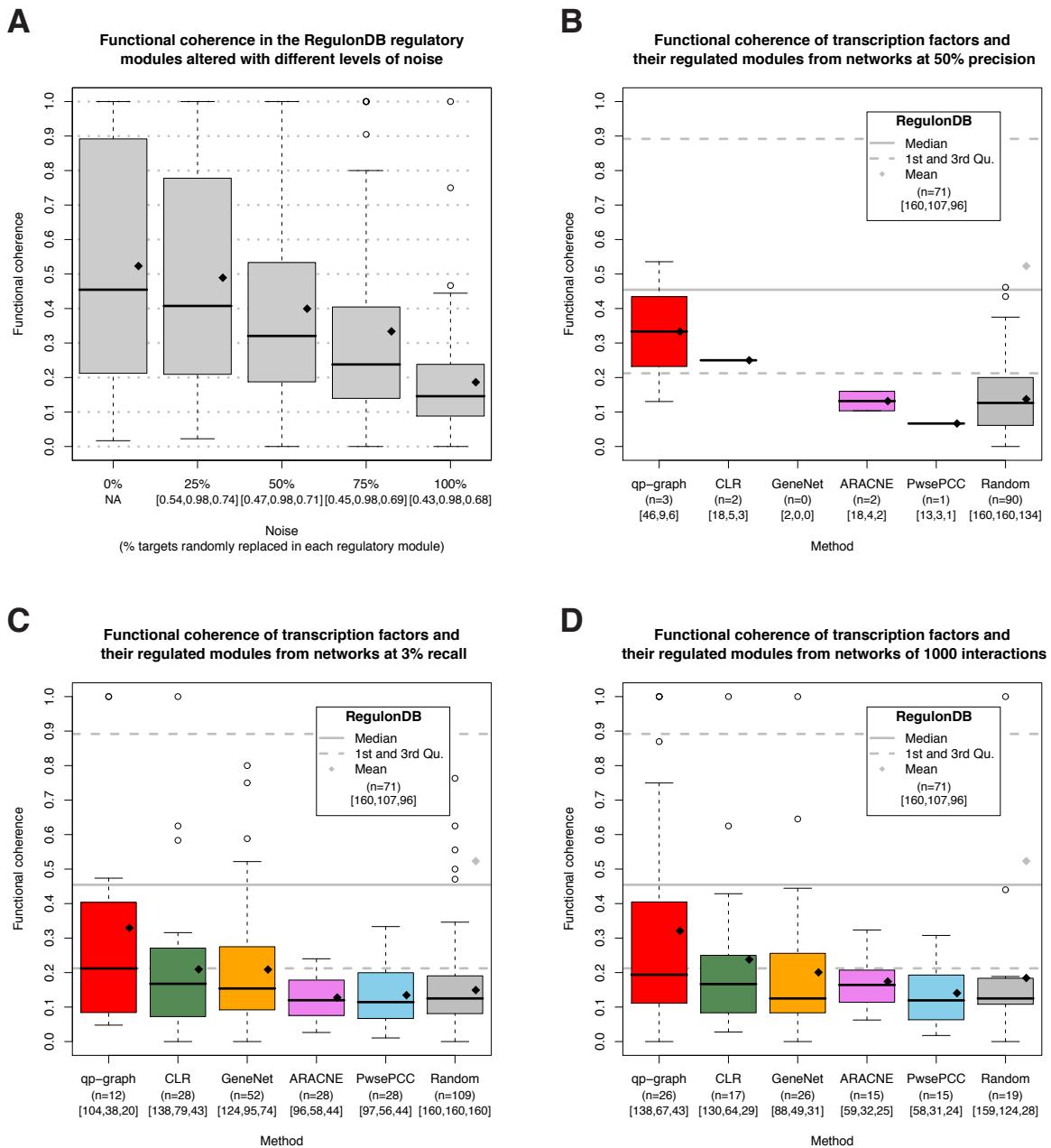


Figure 4 50%-precision qp-graph regulatory network

