Learning undirected graphical models from multiple datasets with the generalized non-rejection rate

Alberto Roverato
Università di Bologna, Italy
alberto.roverato@unibo.it

Robert Castelo
Universitat Pompeu Fabra, Barcelona, Spain
robert.castelo@upf.edu

Abstract

Learning graphical models from multiple datasets constitutes an appealing approach to learn transcriptional regulatory interactions from microarray data in the field of molecular biology. This has been approached both in a model based statistical approach and in an unsupervised machine learning approach where, in the latter, it is common practice to pool datasets produced under different experimental conditions. In this paper, we introduce a quantity called the *generalized non-rejection rate* which extends the non-rejection rate, introduced by Castelo and Roverato (2006), so as to explicitly keep into account the different graphical models representing distinct experimental conditions involved in the structure of the dataset produced in multiple experimental batches. We show that the generalized non-rejection rate allows one to learn the common edges occurring throughout all graphical models, making it specially suited to identify robust transcriptional interactions which are common to all the considered experiments. The generalized non-rejection rate is then applied to both synthetic and real data and shown to provide competitive performance with respect to other widely used methods.

1 Introduction

In the field of molecular biology, an important process that takes place within the cell is gene expression, where the DNA sequence of a gene is transcribed into a functional RNA molecule which, in the case of protein-coding genes, is translated into a protein. When, where and how often a gene is expressed is determined by the requirements imposed by the cell in order to fulfill the corresponding cellular functions. The control exerted on the expression of every gene is known as gene regulation and takes place through a wide range of mechanisms acting at different levels of the gene expression pathway in a coordinated manner. One such mechanisms is the initiation of the synthesis of the RNA molecule, or transcription initiation, which, among other things, requires the presence of a specific combination of proteins that belong to a class of genes called transcription factors. Transcription factor proteins play their role in the initiation of transcription by binding to the upstream genomic region of the regulated gene and then promoting the initiation of transcription (up-regulation) or repressing that step (down-regulation).

The interactions between transcription factors and the genes they target under specific cell environmental conditions, constitute one of the key pieces of information in the cellular program that governs gene expression. Therefore, identifying transcriptional regulatory interactions is a fundamental step towards reverse-engineering this cellular program which potentially contains clues to understanding biological processes like cell division, cell fate or disease.

Microarray technology in molecular biology enables measuring gene expression simultaneously for thousands of genes across a moderate number of samples corresponding to technical or biological replicates of one or more distinct experimental con-

ditions. The resulting gene expression data matrix conveys a snapshot of the expression level of genes under the essayed experimental conditions and efforts have been made in the last years in order to develop computational and statistical procedures that aid to infer, from these data, transcriptional regulatory interactions suitable of being followed-up by further functional experimental validation. Among those procedures, we shall distinguish in this paper between *model based statistical learning methods* and *unsupervised machine learning methods*. In unsupervised machine learning it is common practice to apply the learning algorithm to pooled datasets constructed by merging samples from smaller datasets generated under different experimental conditions (Wang et al., 2006; Steele and Tucker, 2008). The performance of different learning algorithms is then assessed with respect to benchmark problems for which the set of regulatory interactions between genes is (partially) known and available from the biological literature.

Castelo and Roverato (2006) introduced a quantity that they called the *non-rejection rate* and used it in the model based learning of transcriptional networks. Furthermore, Castelo and Roverato (2009) showed that the non-rejection rate provides satisfying results also in an unsupervised machine learning approach. In this paper, we introduce a generalized version of the non-rejection rate that can be naturally used as a meta-analysis approach when the available data are a compendium of microarray experiments. We show that it is suited for the unsupervised machine learning of robust transcriptional interactions, that is transcriptional interactions that are common to all the experimental conditions considered. We apply the proposed method to synthetic and experimental data from one of the best characterized organisms in terms of its transcriptional regulatory relationships, the bacterium Escherichia coli (E. coli), and compare it with some widely used unsupervised learning procedures.

The paper is organized as follows. In Section 2 we review the theory related to structural learning of biological networks. Section 3 addresses the issue of meta-analysis and introduces the generalized non-rejection rate. In Section 4 an analysis based on simulated data is carried out whereas in Section 5 the generalized non-rejection rate is compared with other unsupervised learning procedures on a microarray dataset from the E. coli system. Finally, Section 6 contains a brief discussion.

2 Background

2.1 Unsupervised machine learning of biological networks

There is a substantial amount of recent work on high-dimensional and computationally tractable procedures for learning of biological networks. These can be grouped into two main families: *model based* and *unsupervised machine learning* procedures.

Model based procedures mainly rely on graphical models (see, for instance, Friedman, 2004). Graphical models are well-defined statistical models and the associated network has a precise probabilistic interpretation in terms of conditional independencies. Several procedures are available for learning an independence graph from data (Edwards, 2000), however, structural learning in this context poses new challenges because in microarray data the sample size n is smaller than the number of variables p. This has led to the development of specific structural learning procedures which try to overcome the *small n and large p problem* by exploiting specific biological background knowledge on the structure of the network. From this viewpoint, the most relevant feature of biological networks is that they are *sparse*, that is the direct regulatory interactions between genes represent a small proportion of all possible interactions in the network (see Junker, 2008).

With the term *unsupervised machine learning approach*, shortly *unsupervised learning*, we mean a set of methods that distinguish themselves from the model based statistical learning procedures both for the assumptions underlying the analysis and for the interpretation of the results (see d'Alché-Buc, 2006). More specifically, unsupervised learning procedures aim at identifying some "direct", to be read as "non-spurious", associations with high confidence (see, for instance, Faith et al., 2007) and, usually, no underlying statistical model is assumed. In this framework, the most popular procedures belong to the family of *relevance networks*. One of the first applications of relevance networks was by Butte and Kohane (2003) who proposed to consider

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. Widely used enhancements to this pairwise approach aimed at reducing the number of identified spurious associations are the ARACNE procedure (Margolin et al., 2006) and the CLR procedure (Faith et al., 2007).

In order to validate and compare different unsupervised learning procedures it may be useful to make use of a *benchmark set of transcriptional interactions*, hereafter *benchmark set* for short, that can be obtained by mining existing literature on functional experiments that essay the actual activation or inhibition of a gene by a transcription factor. The benchmark set can be used to construct several measures of performance of the procedure. In particular, specificity and sensitivity can be used to produce ROC curves; however, because of the sparsity of biological networks, and of the consequent large skew in the class distribution, ROC curves present an overly optimistic view of an algorithm's performance (see Fawcett, 2006). In this case, ROC analyses are usually replaced by a *precision-recall curve* where the fraction of the interactions in the benchmark set that the procedure successfully identifies (*recall*) is plotted against the fraction of identified interactions that are true positives (*precision*). We also provide a summary value associated to a precision-recall curve obtained by computing its Area Under the Curve (AUC).

The connection between the model based and the unsupervised learning approaches comes from the fact that model based learning procedures which provide a measure of association of the edges of the complete graph can also be applied in unsupervised learning; see, for instance, Soranzo et al. (2007). This is the case, for instance, of the shrinkage estimator (Schäfer and Strimmer, 2005) implemented in the R package GeneNet. Furthermore, Castelo and Roverato (2006) introduced a quantity that they called the non-rejection rate to be applied in a model based approach but they then showed (Castelo and Roverato, 2009) that it provides satisfying results also in an unsupervised learning approach.

2.2 The non-rejection rate

For a finite set $V = \{1, \ldots, p\}$ let $G = (V, E)$ be an undirected graph with vertex set V and edge set E. Furthermore, let X_V be a multivariate normal random vector, indexed by V, with mean vector μ and covariance matrix Σ. The probability distribution of X_V is said to belong to a Gaussian graphical model with graph G if every missing edge in the graph, $(i, j) \notin E$, implies that X_i is conditionally independent of X_j given the remaining variables $X_{V\setminus\{i,j\}}$. We refer to Lauritzen (1996) for a comprehensive account on Gaussian graphical models, but it is worth recalling here that such conditional independence relationship holds if and only if the partial correlation coefficient $\rho_{ij.V\setminus\{i,j\}}$ is equal to zero.

Led D be a random sample of n observations i.i.d. from X_V. The non-rejection rate is a quantity, introduced by Castelo and Roverato (2006), that can be used in structural learning of Gaussian graphical models when $p > n$ and the network has a sparse structure. It is based on *q-order partial correlations*, that is on partial correlations $\rho_{ij.Q}$ where $Q \subset V\setminus\{i, j\}$ is such that $|Q| = q$.

In the rest of this section we review the theory of the non-rejection rate required for this paper and refer to Castelo and Roverato (2006) for a more complete discussion. The non-rejection rate is a measure associated with every pair of vertices $i, j \in V$ and depends on the particular value of $q < (n - 2)$ being used. It can be 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} be a binary random variable associated to the pair of vertices i and j that takes values from the following two-stage experiment:

1. An element Q is sampled from $\mathcal{Q}_{ij}$ according to a (discrete) uniform distribution;

2. using the available data D the null hypothesis $H_0 : \rho_{ij.Q} = 0$ is verified on the basis of a statistical test of level α. If H_0 is rejected then T_{ij} takes value 0, otherwise it takes value 1.

Thus T_{ij} is a Bernoulli random variable and the non-rejection rate is defined as its expectancy; formally

$$\mathrm{NRR}(i, j \mid q, D) := E[T_{ij}] = \Pr(T_{ij} = 1).$$

It is shown in Castelo and Roverato (2006) that the non-rejection rate can be written as

$$\begin{aligned} \text{NRR}(i,j\,|q,D) &= \beta_{ij}(1-\pi_{ij}) + (1-\alpha)\pi_{ij} \\ &= \beta_{ij} + (1-\alpha-\beta_{ij})\pi_{ij} \quad (1) \end{aligned}$$

where α is the probability of the first type error of the used statistical test, π_{ij} is the proportion of subsets Q in $\mathcal{Q}_{ij}$ that separate i and j in G and β_{ij} is the mean value of the second type errors $\beta_{ij.Q}$ for all the subsets $Q \in \mathcal{Q}_{ij}$ such that Q does not separate i and j in G; see Lauritzen (1996, p. 6) for a formal definition of separator. It follows that the non-rejection rate is a probability that takes a value between 0 and $(1-\alpha)$. If a pair of vertices i and j are adjacent, that is $(i,j) \in G$, (in our context this is as much as saying that a transcription factor directly binds to the promoter region of a target gene) then $\pi_{ij} = 0$ and equation (1) makes clear that the theoretical non-rejection rate equals exactly β_{ij}. Hence, since the statistical power of the tests for zero partial correlation with null hypothesis $H_0 : \rho_{ij.Q} = 0$ equals $1-\beta_{ij.Q}$, then the non-rejection rate of an edge $(i,j) \in G$ corresponds to the one minus the average statistical power to detect that association. It follows that for $(i,j) \in G$ the non-rejection rate $\text{NRR}(i,j\,|q,D)$ converges to 0 as $n-q$ goes to infinity whereas for finite sample size it is a summary measure of the strength of the linear association represented by the edge (i,j) over all the marginal distributions of size $q+2$, with 0 representing maximum strength.

3 Meta-analysis

In unsupervised learning it is common practice to overcome the difficulties related to the small sample size by applying the procedures in a meta-analysis approach. More specifically, a pooled dataset is obtained by merging smaller datasets generated under different experimental conditions. We remark that this practice would make little sense in a model based approach, whereas in a unsupervised learning approach it is justified whenever it leads to an improvement of the precision-recall performance of the procedure.

In this section we formally approach this issue and assume that the data involve a common set of genes V but are obtained from m batches of microarray experiments, possibly under different experimental conditions. Formally, set $M = \{1,\ldots,m\}$ and for every microarray experiment $s \in M$ let $X_V^{(s)}$ be a random vector, indexed by V, corresponding to the expression level of genes in the experimental condition $s \in M$. Furthermore, let $D^{(s)}$ be a random sample of $n^{(s)}$ i.i.d. observations from $X_V^{(s)}$ so that $D^* = \{D^{(1)},\ldots,D^{(m)}\}$ is the pooled dataset made up of $n = \sum_{s=1}^m n^{(s)}$ independent, but not identically distributed, observations.

Our standpoint is that biological networks are dynamic objects which modify their interaction structure to allow the cell to respond effectively to changes of its internal and external environments. This is formalized by assuming that a different graph is associated with every experimental condition and, specifically, by assuming that for every $s \in M$ the probability distribution of $X_V^{(s)}$ belongs to an undirected Gaussian graphical model with graph $G^{(s)} = (V, E^{(s)})$.

In the following section we introduce the generalized non-rejection rate which keeps into explicit account the pooled structure of the dataset.

3.1 The generalized non-rejection rate

The non-rejection rate is defined in Section 2.2 with respect to a set D of i.i.d. observations as the expected value of a Bernoulli random variable generated by means of a two stage experiment. Similarly, the *generalized non-rejection rate* between two variables, X_i and X_j is defined with respect to the dataset D^* as the expected value of a Bernoulli random variable T_{ij}^*,

$$\text{gNRR}(i,j\,|q,D^*) := E[T_{ij}^*] = \Pr(T_{ij}^* = 1),$$

where T_{ij}^* is defined by adding a third step to the two stage experiment as follows:

1. A random value s is generated from S, where S is a discrete random variable that takes values in M with probability $\Pr(S = s) = n^{(s)}/n$;

2. an element Q is sampled from $\mathcal{Q}_{ij}$ according to a (discrete) uniform distribution;

3. using the dataset $D^{(s)}$ the null hypothesis $H_0 : \rho_{ij.Q} = 0$ is verified on the basis of a statistical

test of level α. If H_0 is rejected then T^*_{ij} takes value 0, otherwise it takes value 1.

The value q is chosen so that $q < \min\{(n^{(s)} - 2); s \in M\}$.

We use the generalized non-rejection rate values to produce a ranking of all possible pair of genes, i.e., of all possible edges of the graph, and its usefulness will be assessed by means of a precision-recall analysis; nevertheless, in the following we shortly discuss the interpretation of this quantity.

As well as the non-rejection rate, also the generalized non-rejection rate is a probability. Furthermore, it can be written as weighted average of the non-rejection rates for the single datasets with weights proportional to sample sizes,

$$\begin{aligned}\text{gNRR}(i,j\ |q,D^*) &= \sum_{s=1}^{m} \text{NRR}(i,j\ |q,D^{(s)}) P(S=s) \\ &= \frac{1}{n}\sum_{s=1}^{m} \text{NRR}(i,j\ |q,D^{(s)})\, n^{(s)}.\end{aligned}$$

It follows that if $(i,j) \in G^{(s)}$ for every $s \in M$, then the generalized non-rejection rate is the weighted average of the mean second type errors $\beta^{(s)}_{ij}$ for all the datasets $D^{(s)} \in D^*$, formally $\text{gNRR}(i,j\ |q,D^*) = \bar{\beta}_{ij}$ where $\bar{\beta}_{ij} := \sum_{s=1}^{m} \beta^{(s)}_{ij} P(S=s)$. Since for all $s \in M$ the quantity $\beta^{(s)}_{ij}$ converges to zero as $n^{(s)}$ increases, it follows that the generalized non-rejection rate is specially useful to identify edges that belong to all graphs. These are robust transcriptional interactions that are common to all the experiments considered.

4 Assessment with simulated data

In order to empirically show the behavior of the generalized non-rejection rate we considered $m = 2$ and two different graphs $G^{(1)}$ and $G^{(2)}$ and repeat the following simulation 100 times.

1. Generate randomly $G^{(1)}$ and $G^{(2)}$ with $p = 50$ vertices each of them with an average number of adjacent vertices equal to 3.

   For each of these two graphs build a random precision matrix whose zero structure reflects the conditional independencies encoded in the graph (i.e., the pattern of missing edges) and sample $n = 30$ observations from the Gaussian distribution with that precision matrix and zero mean vector. This step provides us with two datasets, $D^{(1)}$ and $D^{(2)}$, with $p = 50$ and $n = 30$ belonging to two different joint multivariate Gaussian distributions.

2. Estimate the non-rejection rate with $q = 4$ and the Pearson correlation coefficient for each pair of variables, separately from $D^{(1)}$ and $D^{(2)}$. Apply methods GeneNet, ARACNE and CLR separately to $D^{(1)}$ and $D^{(2)}$.

3. Build the collection of datasets $D^* = \{D^{(1)}, D^{(2)}\}$. By using D^* estimate the generalized non-rejection rate, the non-rejection rate and the Pearson correlation for each pair of variables. Apply methods GeneNet, ARACNE and CLR to D^*.

4. For each method and dataset, calculate a precision-recall curve with respect to the union graph $G^* = (V, E^{(1)} \cup E^{(2)})$. For each precision-recall curve calculate the area under this curve (AUC).

5. In order to have a baseline comparison method we have generated three sets of random correlations sampling values uniformly from (-1,+1) for each pair of the $p = 50$ variables. This will be referred to as the Random method.

In Figure 1 we show the AUC values across the 100 simulations for each of the methodologies followed where the larger the value, the better the performance. The two panels on top correspond to using $D^{(1)}$ and $D^{(2)}$ separately in order to try to infer all the edges in the union of the two graphs. The panel at the bottom shows the performance when using a meta-analysis approach in which all datasets in D^* are used together. We can observe that only the generalized non-rejection rate provides a clear improvement over the use of any of the methods on one single dataset. GeneNet also achieves a small increase in its median AUC value.

We have previously pointed out that the generalized non-rejection rate is specially useful to identify edges that belong to all graphs. In the previous simulations those correspond to edges in both $G^{(1)}$ and $G^{(2)}$ in each simulation. We have used the previous

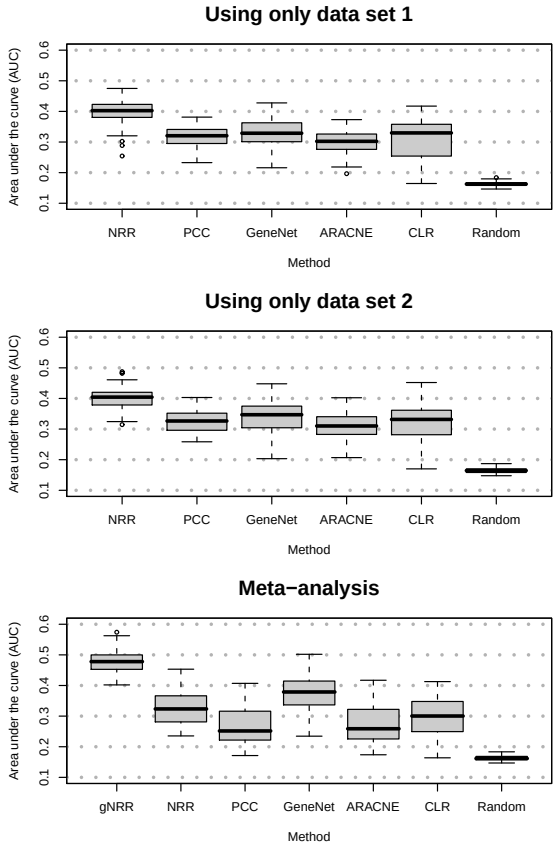

Figure 1: AUC value comparison (the larger the value the better the performance).

results to empirically verify this fact by grouping together the generalized non-rejection rate values in the following four subsets: edges in $G^{(1)}$ and $G^{(2)}$, edges in $G^{(1)}$ but not in $G^{(2)}$, edges in $G^{(2)}$ but not in $G^{(1)}$ and missing edges in both $G^{(1)}$ and $G^{(2)}$. Then, we have examined the distribution of generalized non-rejection rates separately for each of these subsets and we may see those values from the 100 simulations in Figure 2. Clearly, generalized non-rejection rates are most of the time smaller for edges that belong to the two graphs than for edges that belong to one of the two graphs only or edges that are missing in both graphs.

5 Assessment with microarray data

In order to assess with real microarray experimental data whether the generalized non-rejection rate increases our accuracy when trying to identify transcriptional regulatory relationships through a meta-

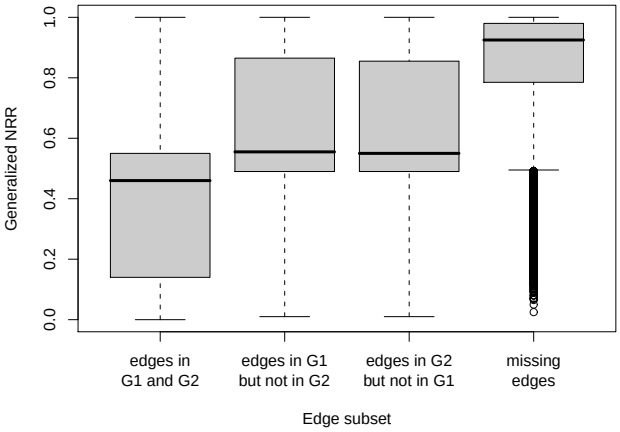

Figure 2: Behaviour of the generalized non-rejection rate for different kind of edges in graphs $G^{(1)}$ and $G^{(2)}$.

analysis approach, we have used experimental and functional annotation data from the E. coli system. These bacteria are the free-living organism for which the largest part of its transcriptional regulatory network has undergone some sort of experimental validation. The database RegulonDB (Gamma-Castro et al., 2008) contains a set of transcription factor and target gene relationships curated from the literature, and we have used those as benchmark set of regulatory interactions. The E. coli microarray dataset that we have used to assess our meta-analysis approach corresponds to the oxygen deprivation data from Covert, et al. (2004), available from the Gene Expression Omnibus (GEO) database (Barrett, 2007) with accession number GDS680, which monitors the response of E. coli during the transition from aerobic to anaerobic conditions which are essayed in two groups of 21 and 22 experiments, respectively. In order to obtain the necessary variability on the expression levels of genes that form part of the transcriptional network relevant to these experiments, Covert, et al. (2004) used six E. coli strains with knockouts of key transcriptional regulators in the oxygen response (Δ*arcA*, Δ*appY*, Δ*fnr*, Δ*oxyR*, Δ*soxS* and the double knockout Δ*arcA*Δ*fnr*). Both, the microarray oxygen deprivation data where we are going to assess our approach and the RegulonDB interactions which we are going to use as

benchmark set, were pre-processed by Castelo and Roverato (2009) and are available as part of the `qpgraph` package from the Bioconductor project website (`http://www.bioconductor.org`). These pre-processed datasets consist, in one hand, of 3283 transcriptional regulatory interactions in RegulonDB involving 1428 genes and, on the other hand, $p = 4205$ genes and $n = 43$ experiments in the oxygen deprivation microarray expression data matrix. From these latter dataset, we have discarded one of the aerobic experiments (GSM18237) which showed a very dissimilar profile to the rest of the aerobic experiments. More concretely, we have used only those genes involved in the regulatory modules of the five transcription factors knocked-out in the experiments of Covert, et al. (2004) according to the RegulonDB database. This subnetwork is formed by 378 genes out of which 22 are transcription factors involved in 681 transcriptional interactions. In the bottom panel of Figure 3 we may see the resulting precision-recall curves where we compare the generalized non-rejection rate with $q = 10$ (gNRR) with other methods applied to the union of the aerobic and anaerobic datasets including the non-rejection rate with $q = 15$ (NRR), the absolute Pearson correlation coefficient (PCC), ARACNE, CLR, GeneNet and the assignment of a uniformly random correlation between every transcription factor and target gene (Random). The gNRR improves the rest of the meta-analysis approaches up to a 40% larger AUC with respect to the second best-performing approach (NRR).

An issue in the computation of the non-rejection rate concerns the choice of the parameter q. In this application, q may take any value between 1 and 18 and there is a trade-off between larger values of q, which increase the probabilities π_{ij}, and smaller values of q, which improve the power of statistical tests. In the bottom panel of Figure 3 we set $q = 10$ for gNRR which is the median of the possible values of q within a sensible range bounded by the number of available samples but, in fact, our procedure is not very sensitive to the choice of q as shown in the top panel of Figure 3 where the precision-recall curves of gNRR for different values of q are drawn.

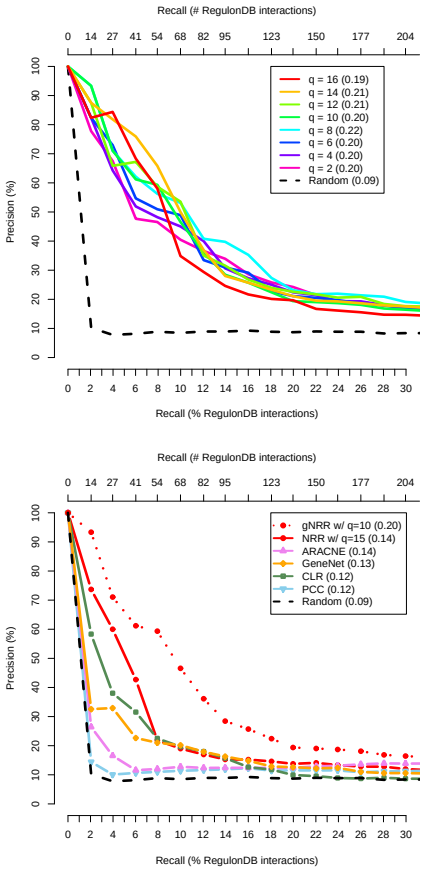

Figure 3: Comparison of precision-recall curves through the first 30% fraction of the recall. The top panel shows curves from the generalized non-rejection rate for a wide range of possible q values. The bottom panel shows curves for different unsupervised machine learning methods. The AUC value is the number enclosed within brackets.

6 Discussion

The idea at the basis of the generalized non-rejection rate is that of exploiting the information provided by the different microarray experiments on the common part of the underlying regulatory networks in a way that does not require pooling the datasets. We believe that this is the crucial aspect that makes the generalized non-rejection rate outperform other procedures, but it is worth pointing out that it also has the drawback that the power and efficiency of the applied statistical procedures depend on the samples sizes of the single experiments, $n^{(s)}$, rather than the overall sample size n.

Acknowledgments

This work is supported by the Spanish Ministerio de Ciencia e Innovación (MICINN) [TIN2008-00556/TIN] and the ISCIII COMBIOMED Network [RD07/0067/0001]. The first author acknowledges support from the Ministero dell'Università e della Ricerca [PRIN-2007AYHZWC, FISR MITICA]. Robert Castelo is a research fellow of the "Ramon y Cajal" program from the Spanish MICINN [RYC-2006-000932].

References

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 Research, 35 D760-D765.

Butte, A.S. and Kohane,I.S., 2003. Relevance networks: a first step toward finding genetic regulatory networks within microarray data. In G. Parmigiani, E.S. Garett, R.A. Irizarry and S.L. Zeger (Ed.s), The Analysis of Gene Expression Data, Springer, New York, 428-446.

Castelo, R. and Roverato, A., 2006. Gaussian graphical model search from microarray data with p larger than n. Journal of Machine Learning Research, 7 2621-2650.

Castelo, R. and Roverato, A., 2009. Reverse engineering molecular regulatory networks from microarray data with qp-graphs. Journal of Computational Biology, 16(2):213-227.

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 92-96.

d'Alché-Buc, F., 2006. Inference of biological regulatory networks: machine learning approaches. In: F. Képès, (Ed.). Biological networks, World Scientific, 41-82.

Edwards, D.E., 2000. Introduction to graphical modelling. Springer-Verlag, New York.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern recognition letters, 27 861-874.

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.

Friedman. N., 2004. Inferring cellular network using probabilistic graphical models. Science, 33 799-805.

Junker, B.H., 2008. Networks in biology. In: B.H. Junker and F. Schreiber (Ed.s). Analysis of Biological Networks, Wiley, 3-12.

Lauritzen, S.L., 1996. Graphical models. Oxford University Press, Oxford.

Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C. Stolovitzky, G., Favera, R.D. and Califano, A., 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks ina mammalian cellular context. BMC Bioinformatics, 7(Suppl 1) S7.

Schäfer, J. and Strimmer, K., 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology, 4(1) article 32.

Gama-Castro, S. et al. 2008. RegulonDB (Version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and textpresso navigation. Nucleic Acids Research, vol 36, D120-D124.

Soranzo, N., Bianconi, G. and Altafini, C., 2007. Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data. Bioinformatics, 23(13) 1640-1647.

Steele, E. and Tucker, A., 2008. Consensus and meta-analysis regulatory networks for combining multiple microarray gene expression datasets. Journal of Biomedical Informatics, 41(6), 914-926.

Wang, Y., Joshi, T., Zhang, X., Xu, D. and Chen, L. 2006. Inferring gene regulatory networks from multiple microarray datasets. Bioinformatics, 22(19), 2413-2420.