

An application of the non-rejection rate in meta-analysis

Alberto Roverato

University of Bologna, e-mail: alberto.roverato@unibo.it

Robert Castelo

Pompeu Fabra University, e-mail: robert.castelo@upf.edu

Abstract: Learning of transcriptional regulatory interactions using microarray data is an important and challenging problem in bioinformatics. Several solutions to this problem have been proposed both in a model based statistical approach and in an unsupervised machine learning approach. In a model based approach a very popular choice are Gaussian graphical models where it is assumed that microarray data form an i.i.d. multivariate normal sample. In this framework, Castelo and Roverato (2006) introduced a quantity that they called the *non-rejection rate* that can be used to learn Gaussian graphical models when the sample size is smaller than the number of variables. Here we present an application of the non-rejection rate, in an unsupervised learning approach, to a compendium of data from different microarray experiments and shown that it provides competitive performance with respect to other widely used methods.

Keywords: graphical model, microarray data, non-rejection rate, partial correlation, small-sample inference, transcriptional regulatory interaction, unsupervised learning.

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 fulfil 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 regula-

⁰This work is supported by the Spanish Ministerio de Ciencia e Innovacion (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].

tory 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 experimental conditions. 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*, for inferring transcriptional regulatory relationships from microarray data. Model based approaches consider microarray data as a random sample from a multivariate distribution defined by a set of random variables associated to the genome functional elements under study (e.g., genes). The interactions among the genome functional elements are then usually described by using conditional independencies and, more concretely, graphical models (see Lauritzen, 1996). However, in a typical microarray data set the number of observations n (on the order of tens and sometimes hundreds) is substantially smaller than the number of variables p (on the order of thousands) and this prevents us from applying directly most of the existing multivariate methods for structural learning of graphical models. 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 networks. 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 edges of the graph (see Junker, 2008). Here, it is also important to underline 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. As a consequence, it makes sense to aim at learning transcriptional regulatory networks specific to particular environmental conditions of the cell.

In unsupervised machine learning the main goal is not to recover *all* transcriptional regulatory interactions but rather to recover *some* transcriptional interactions with high confidence (see d'Alché-Buc, 2006). Furthermore, it is not assumed that the data form a i.i.d sample from a random vector whose probability distribution belongs to a well-defined statistical model and it is common practice to apply these procedures to a compendium of data from different microarray experiments.

Castelo and Roverato (2006) introduced a quantity that they called the *non-rejection rate* and used it in the model based learning of transcriptional network. Furthermore, Castelo et al. (2009) showed that the non-rejection rate provides satisfying results also in an unsupervised machine learning approach. In this paper, we consider an application of the *generalized non-rejection rate*, introduced by Roverato and Castelo (2009) in a meta-analysis context, when the available data are a compendium of microarray experiments. We apply the proposed method to 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.

2. Application to 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-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

The *E. coli* microarray data set 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 ($\Delta arcA$, $\Delta appY$, Δfnr , $\Delta oxyR$, $\Delta soxS$ and the double knockout $\Delta arcA \Delta fnr$).

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 and used in Castelo and Roverato (2009) and are available as part of the *qppgraph* package from the Bioconductor project website (<http://www.bioconductor.org>). We have used here these same pre-processed data sets which 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 data set, 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 Figure 1 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 data sets 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).

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.

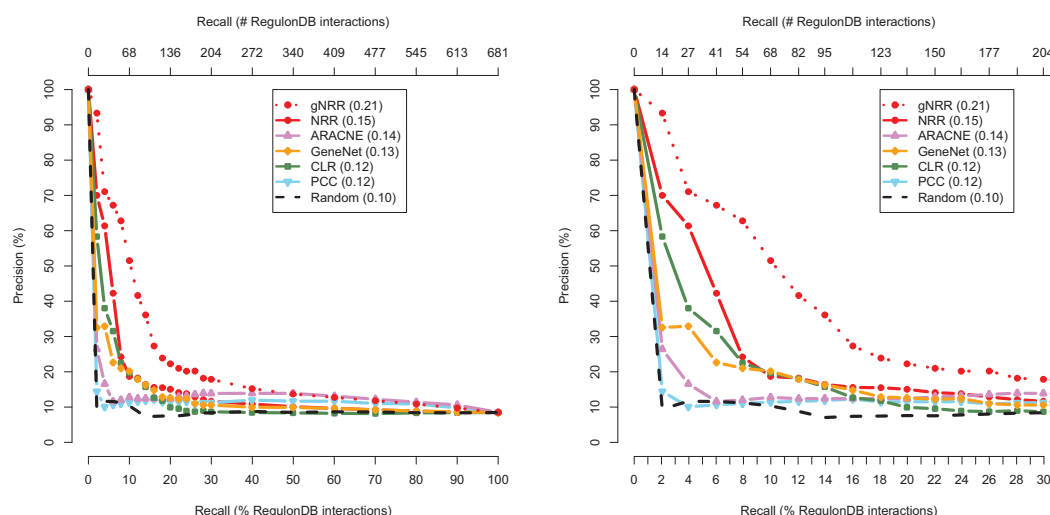


Figure 1: Comparison of Precision-Recall curves for different unsupervised machine learning procedures. The picture of the top shows the complete curves with recall ranging from 0 to 100%, whereas the picture on the bottom focuses on the recall from 0 to 30%. The AUC value is the number enclosed within brackets.

- 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*, to appear.
- 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.
- 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 Research*, vol 36, D120-D124.
- 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.
- Roverato, A. and Castelo, R., 2009. Learning transcriptional regulatory interactions from microarray data: a meta-analysis approach based on the non-rejection rate. Submitted.