Computationally Efficient Test for Gene Set Dysregulation

Adrien Dessy and Pierre Dupont

Université catholique de Louvain - ICTEAM/Machine Learning Group
Place Sainte Barbe 2 bte L5.02.01, B-1348 Louvain-la-Neuve, Belgium
adrien.dessy@uclouvain.be

Abstract. The identification of genetic regulatory pathways whose structure differs across biological conditions provides significant insights about organisms and diseases functioning at molecular level. In this paper, we propose a computationally efficient test to assess from gene expression data if a given group of genes is differentially regulated between two conditions. The method yields promising results in terms of precision and recall on real datasets.

1 Introduction

Differential analysis of Gene Regulatory Networks (GRNs) has been raising a growing interest lately. There is not yet a standard definition to this problem, but the high-level shared goal is to assess if the interactions or associations of genes differ between two or more biological conditions. This type of analysis can be performed at the network level, for a given group of genes (subnetwork) or for a specific interaction between two genes for instance.

Most of the proposed methods have in common that they compare networks infered for each condition from gene expression data [1,2]. Hence, they rely on network inference techniques or association measures between genes. The comparison is then based on a differentiation score whose significance is assessed by a permutation test. The first contribution of this work is to propose an alternative permutation test that is more computationally efficient.

The definition of differential network analysis slightly differs between studies. Either because they operate at different network levels, or because some studies test a given component for differentiation, while others try to discover differentiated parts of the network. Gill et al. [1] proposes three statistical tests to assess whether the modular structures of two networks are different, whether the connectivity of a group of genes or the connectivity of a single gene has changed. Liu et al. [3] describes a procedure to determine if genes of a pathway are differentially wired between two conditions. If so, a differential network is build by testing dysregulation of each interaction in the pathway using a t-test. The DINA procedure proposed by Gambardella et al. [2] also intends to assess whether co-regulation among a given set of genes depends on the condition, but accross multiple networks. Amar et al. [4] presents an algorithm to extract differential gene clusters. Our work is closely related to these techniques, especially [1,2]. In this paper, we describe a method to :

(a) test if a given group of genes (a module) is differentially regulated between two conditions;
(b) rank modules by observed dysregulation level.

In Section 2, we describe the details of the method. Section 3 discusses the results of our experiments on two real gene expression datasets. Finally, Section 4 presents the conclusions of this work and suggests some future works.

2 Method

The method that we propose is summarized in Figure 1. It consists of three main steps. Firstly, GRNs are infered from gene expression data for both conditions. Subsequently, a differentiation score is computed by comparing the GRNs. The significance of this score is finally estimated through a permutation test.

2.1 Inference of gene regulatory networks

For each biological condition, a GRN is infered using the MRNET approach [5]. MRNET inference consists in performing a sequence of mRMR gene selection procedures with each gene as output variable. mRMR algorithm selects iteratively variables depending on the previously selected variables (gene expression profiles in our case). At each iteration, it selects the gene that maximizes an objective function measuring a trade-off between the mutual information with the target gene (relevance) and the mean mutual information with the already selected genes (redundancy).

For the sake of computational efficiency, we made assumption of data normality. Under this assumption, mutual information can readily be computed as

$$\mathbf{MI}_{ij} = -\frac{1}{2}\ln(1-\rho_{ij}^2)$$

where ρ_{ij} is the pearson correlation between genes i and j.

This step produces two adjacency matrices $\mathbf{A}^1$ and $\mathbf{A}^2$ representing both GRNs. These matrices are symetric with null diagonal and their entries are in the range $[0, 1]$. Note that MRNET is quite an arbitrary choice. Another GRN inference method could have been chosen.

2.2 Differentiation score

The differentation score, denoted by s_Δ, represents the differentiation level observed between GRNs with respect to a module. A large score s_Δ indicates an important differentiation. In mathematical terms, it is computed as three-argument function :

$$s_\Delta = f_\Delta(\mathbf{A}^1, \mathbf{A}^2, \mathcal{M})$$

where $\mathcal{M}$ is the group of genes of interest.

The computation of s_Δ can be decomposed into two steps. The first part is applied separately to each network and aims to extract statistics that depend on the topology of the module. The second part compares these statistics to produce the differentiation score.

In the first step, network statistics are computed as $\mathbf{s}^1 = f(\mathbf{A}^1, \mathcal{M})$ and $\mathbf{s}^2 = f(\mathbf{A}^2, \mathcal{M})$, such that $\mathbf{s}^1$ and $\mathbf{s}^2$ are real vectors of same length ($\mathbb{R}^K$) for any instantiation of the scoring function f. The scores are then combined as

$$\begin{aligned} s_\Delta &= f_\Delta(\mathbf{A}^1, \mathbf{A}^2, \mathcal{M}) \\ &= \sum_{k=1}^{K} |\mathbf{s}^1_k - \mathbf{s}^2_k| = ||\mathbf{s}^1 - \mathbf{s}^2||_1 \end{aligned}$$

where $||.||_1$ denotes the $\ell 1$ norm.

Instead of this two-stage scheme, we could have adapted graph kernels to compute the differentiation scores. But, for the sake of computational efficiency, we chose this simple approach as a first step.

We now introduce in the rest of the section different instantiations of the scoring functions f. Several variants have been explored, but for the sake of brevity only simple graph statistics based on node degree will be reported here.

Degree The function f_{degree} returns the degree of each node in $\mathcal{M}$. More precisely, let's define $\mathbf{s} = f_{degree}(\mathbf{A}, \mathcal{M})$. Without loss of generality, we can reorder genes such that $\mathcal{M} = \{1, 2, \ldots, M\}$. We have that $\mathbf{s} \in \mathbb{R}^M$ and

$$\mathbf{s}_i = \sum_{j \in \mathcal{M}, i \neq j} \mathbf{A}_{ij}, \quad \forall i \in \mathcal{M}.$$

Notice that this definition uses only weights of edges between genes in $\mathcal{M}$. Furthermore, the score vector $\mathbf{s}$ can be normalised as $\mathbf{s}' = \frac{1}{\max_{i \in \mathcal{M}} \mathbf{s}_i} \mathbf{s}$.

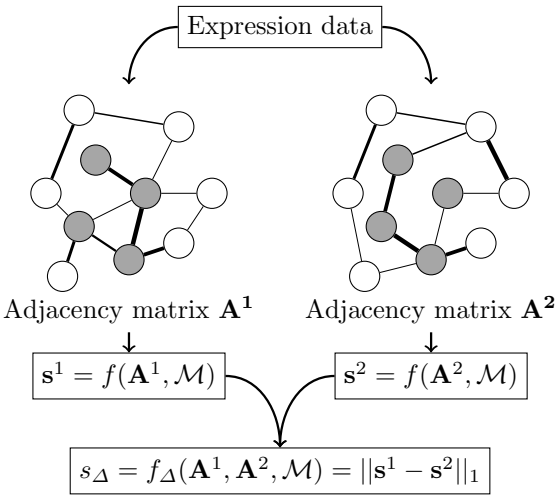


Fig. 1 – Method overview : A GRN is infered for both biological conditions from gene expression data. Given a module $\mathcal{M}$ (grey nodes), network statistics $\mathbf{s}^1$ and $\mathbf{s}^2$ are computed from each GRN and combined to produce the score of differentiation $\mathbf{s}_\Delta$.

Mean degree The function $f_{mean.deg}$ is simply defined as the mean degree of the module, that is

$$s = f_{mean.deg}(\mathbf{A}, \mathcal{M}) = \frac{1}{M} \sum_{m=1}^{M} \mathbf{s}_m^{degree}$$

where $\mathbf{s}^{degree} = f_{degree}(\mathbf{A}, \mathcal{M})$. In this case, the scoring function returns a scalar.

Degree variance The function $f_{deg.var}$ computes the degree variance of the module

$$s = f_{deg.var}(\mathbf{A}, \mathcal{M}) = \frac{1}{M-1} \sum_{m=1}^{M} (\mathbf{s}_m^{degree} - \overline{\mathbf{s}^{degree}})^2$$

where $\overline{\mathbf{s}^{degree}} = f_{mean.deg}(\mathbf{A}, \mathcal{M})$. As the previous scoring function, it returns a scalar.

2.3 Permutation test

A permutation test is performed to assess if the differentiation score s_Δ is significant. The standard approach consists in permuting the class labels N times [1,2,3,4]. This requires to reinfer a pair of GRNs for each permutation. This operation has a complexity of $\Omega(p^2)$ where p is the number of genes and becomes costful for real networks that involve thousands of genes.

Here, we propose an alternative approach that is less computationally expensive. The idea is to sample N random modules $\mathcal{M}_n$ ($\forall n, 1 \leq n \leq N$) of size $|\mathcal{M}|$ from all the available genes. This alternative test postulates that the probability of these random modules being differentiated is very low. Hence, a background distribution of s_Δ can be estimated by computing permutation scores

Kegg ID	Name	Size
hsa04010	MAPK signaling pathway	220
hsa04060	Cytokine-cyt. receptor interaction	217
hsa04110	Cell cycle	117
hsa04115	p53 signaling pathway	37
hsa04151	PI3K-Akt signaling pathway	292
hsa04210	Apoptosis	46
hsa05215	Prostate cancer	59

Table 1 – Kegg pathways used as differentiated modules in prostate cancer compared to healthy condition.

$s^n_\Delta = f_\Delta(\mathbf{A}^1, \mathbf{A}^2, \mathcal{M}_n)$ from these random modules. The test p-value is then defined as $p\text{-}value = \#\{s^n_\Delta \geq s_\Delta\}/N$ where $\#\{s^n_\Delta \geq s_\Delta\}$ is the number of permutation scores greater or equal to the original differentiation score. These p-values can used to rank a set of modules according to their dysregulation level.

3 Experiments

We investigate the performance of our approach on real datasets and compare it with a baseline approach. The experiments show promising results in terms of recall and precision.

3.1 Baseline

In order to validate our approach, we compare it with the following baseline inspired from gene set enrichment analysis [6]. Firstly, we select genes with differentiated expression using a Welch's t-test with Benjamini-Hochberg correction. A hypergeometric test is used to test the significance of the overlap between the selected genes and the genes of a given module. The modules can then be ranked by p-value in ascending order.

3.2 Datasets

We tested our approach on two real gene expression datasets : GSE6919[1] and GSE13159[1], retrieved from InSilico DB [7]. In order to be able to measure precision and recall, a set of differentiated modules as well as a set of undifferentiated modules must be known for each dataset. This information has been retrieved from Kegg database of annotated pathways. A Kegg pathway can be readily converted into a module by considering its set of genes.

GSE6919 : prostate cancer. This dataset is composed of gene expression data from normal and prostate cancer tumor tissues. It consists of 171 samples (18 healthy and 153 cancer) and 8801 genes. For computational reasons, it has been reduced to 2000 genes. The set of differentiated modules was defined from the prostate cancer pathway and its related pathways reported in Table 1. And the set of undifferentiated modules was then formed by selecting randomly 50 other Kegg pathways.

GSE13159 : leukemia. This dataset is part of the MILE Study (Microarray Innovations In LEukemia) program and encompasses 2096 gene expression data from different kinds of leukemia. We restricted the dataset to two conditions : chronic myeloid leukemia (74 samples) and healthy (76 samples). Furthermore, the dataset has also been reduced to 2500 genes. In the same way as for GSE6919, the set of undifferentiated modules is formed from random pathways while the set of differentiated modules is composed of pathways related to chronic myeloid leukemia.

3.3 Results and discussion

The precision-recall curves for both datasets are shown in Figure 2 and AUPR measures are reported in Table 2. We can observe from the GSE6919 curve that the degree scoring function performs best, followed by the mean-degree statistics. However, turning now our attention to the GSE13159 dataset, we can see that the baseline outperforms our approach. It is followed this time by the degree-variance scoring function. Hence, no single technique appears superior to others in all cases.

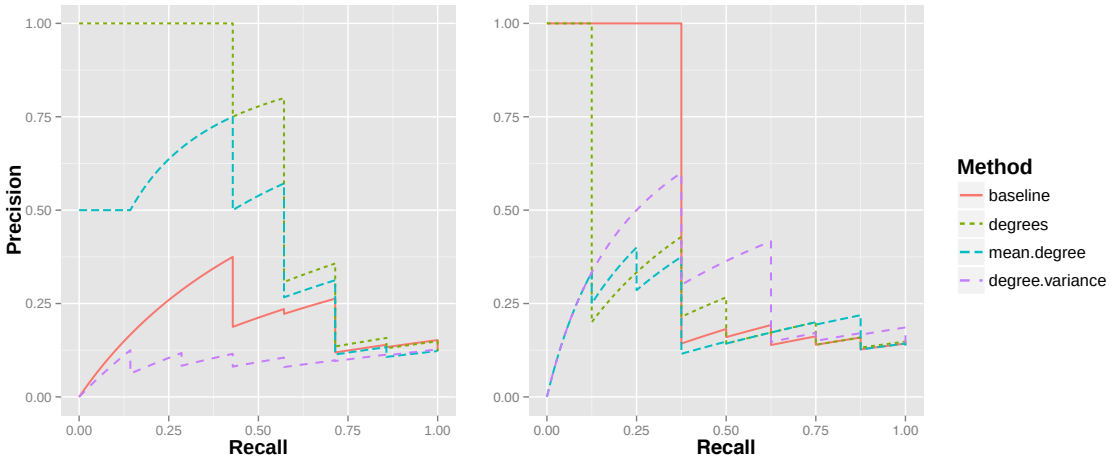
However, these results seem to underestimate the actual performances of our method. Indeed, if we consider for instance the most dysregulated pathways in prostate cancer according to the f_{degree} scoring function (as reported in Table 3), we can see that meaningful results are penalized by our initial definition of the differentiated modules. According to multiple studies [8,9], steroid hormones play a major role in human prostatic carcinogenesis. Besides, studies have shown associations between prostate cancer and alpha-linolenic acid [10]. Eventually, Brockhausen et al. [11] reports links between some kinds of *O*-glycans and adenocarcinomas from the prostate. Hence, pathways 4, 6 and 7 (in Table 3) are actually relevant to the disease, but are considered as false positives by the evaluation protocol.

Besides measuring AUPR performances, we also checked that the test behaves properly by testing the uniformity of the empirical distribution of p-values for undifferentiated modules. This has

Method	GSE6919	GSE13159
Baseline	0.20	0.40
Degrees	0.57	0.22
Mean degree	0.40	0.21
Degree variance	0.09	0.30

Table 2 – AUPR measures.

[1] Gene Expression Omnibus identifiers.



(a) GSE6919 **(b)** GSE13159

Fig. 2 – Precision-Recall curves

been done for the different scoring functions with a χ^2 test. None of these tests has shown enough evidence to reject the null hypothesis of uniformity. This result and a complementary visual inspection of the distribution indicate in particular a good control of type-I error.

4 Conclusions and perspectives

In this paper, we proposed a statistical framework to test the differential regulation of sets of genes. The primary contribution is the introduction of a computationally efficient permutation test. Indeed, this test does not require to reinfer GRNs (or recompute association measures for each pair of genes) for each permutation.

Promising results in terms of precision and recall has been obtained using very simple scoring functions. Besides testing the approach on additional datasets, there are plenty of opportunities for future works. Hitherto, we only considered scoring functions that rely on local properties of GRN topology. One might implement network statistics that takes long range dependencies into account. Furthermore, the data and procedure of evaluation are certainly a point to refine.

Rank	Pathway name
1	**MAPK signaling pathway**
2	**Prostate cancer**
3	**Apoptosis**
4	Steroid hormone biosynthesis
5	**Cytokine-cyt. receptor interaction**
6	Linoleic acid metabolism
7	Other types of O-glycan biosynthesis

Table 3 – Top-ranked pathways for GSE6919 dataset (prostate cancer) using f_{degree} scoring function. Pathways labeled as differentiated for the evaluation are in boldface.

References

1. Gill R. et al.: A statistical framework for differential network analysis from microarray data. BMC Bioinformatics 11(95) (2010)
2. Gambardella G. et al.: Differential network analysis for the identification of condition-specific pathway activity and regulation. Bioinformatics 29(14) (2013)
3. Liu, Y. et al.: Gene interaction enrichment and network analysis to identify dysregulated pathways and their interactions in complex diseases. BMC systems biology 6 (2012)
4. Amar, D. et al.: Dissection of regulatory networks that are altered in disease via differential co-expression. PLoS Comput. Biol. 9(3) (2013)
5. Meyer, P. E. et al.: minet: AR/Bioconductor package for inferring large transcriptional networks using mutual information. BMC bioinformatics 9(1) (2008)
6. Falcon, S. et al.: Hypergeometric testing used for gene set enrichment analysis. In Bioconductor case studies. Springer New York. (2008)
7. Coletta et al.: InSilico DB genomic datasets hub: an efficient starting point for analyzing genome-wide studies in GenePattern, Integrative Genomics Viewer, and R/Bioconductor. Genome Biology, 13(11) (2012)
8. Bosland, M. C.: The role of steroid hormones in prostate carcinogenesis. JNCI Monographs 2000(27) (2000)
9. Wilding, G.: The importance of steroid hormones in prostate cancer. Cancer surveys 14 (1991)
10. Azrad, M. et al.: Prostatic alpha-linolenic acid (ALA) is positively associated with aggressive prostate cancer: a relationship which may depend on genetic variation in ALA metabolism. PloS one, 7(12) (2012)
11. Brockhausen, I.: Pathways of *O*-glycan biosynthesis in cancer cells. Biochimica et Biophysica Acta (BBA)-General Subjects, 1473(1) (1999)