Gustavo H. Esteves¹ / Luiz F. L. Reis²

A statistical method for measuring activation of gene regulatory networks

¹ Statistics Department, University of Paraíba State, Campina Grande, PB, Brazil, E-mail: gesteves@uepb.edu.br² Teaching and Research Institute, Sírío-Libânes Hospital, São Paulo-SP, Brazil

Abstract:

Motivation: Gene expression data analysis is of great importance for modern molecular biology, given our ability to measure the expression profiles of thousands of genes and enabling studies rooted in systems biology. In this work, we propose a simple statistical model for the activation measuring of gene regulatory networks, instead of the traditional gene co-expression networks. Results: We present the mathematical construction of a statistical procedure for testing hypothesis regarding gene regulatory network activation. The real probability distribution for the test statistic is evaluated by a permutation based study. To illustrate the functionality of the proposed methodology, we also present a simple example based on a small hypothetical network and the activation measuring of two KEGG networks, both based on gene expression data collected from gastric and esophageal samples. The two KEGG networks were also analyzed for a public database, available through NCBI-GEO, presented as Supplementary Material. Availability: This method was implemented in an R package that is available at the BioConductor project website under the name *maigesPack*.

Keywords: gene regulatory networks, hypothesis tests, systems biology

DOI: 10.1515/sagmb-2016-0059

1 Introduction

Several mathematical and statistical tools can be applied to gene expression datasets depending on the goal of the analysis. Methods for differential gene expression analysis (Dudoit et al., 2002; Yang et al., 2002; Draghici, 2003), discriminant analysis and clustering (Mardia, Kent & Bibby, 1979; Heyer, Kruglyak & Yoosheph, 1999; Zhu & Zhang, 2000; Johnson & Wichern, 2002), for example, are traditionally employed with this type of data. Some of these, however, do not provide meaningful results for gene sets or gene regulatory networks, such as Gene Ontology (GO) categories (Ashburner et al., 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa & Goto, 2000; Kanehisa et al., 2002).

In order to achieve a better understanding of higher level functioning in groups of genes, like GO categories, some methods were proposed to search for gene sets known as active modules, as proposed by Segal et al. (2004), or to perform gene set enrichment analysis (GSEA), as proposed by Subramanian et al. (2005).

Another type of modeling for gene sets is focused on gene regulatory networks or molecular pathways, as the construction of gene co-expression networks, as proposed by Butte et al. (2000) and Butte and Kohane (2000), that estimate linear correlation values (or mutual information) between all gene pairs inside one gene set and estimates sub-graphs, the co-expression networks, from zero correlation hypothesis tests. Langfelder and Horvath (2008) extended this type of analysis in a procedure named weighted gene co-expression network analysis (WGCNA), which presents additional methods to search inside a co-expression network for modules defined as clusters of genes with dense profile of interconnection between themselves. Also, Schäfer and Strimmer (2005) constructed a kind of gene co-expression networks using a shrinkage approach to estimate covariance and correlation matrices, that was proven to provide better statistical properties than traditional estimators.

More recently, Rahmatallah, Emmert-Streib, and Glazko (2014) proposed the gene sets net correlation analysis (GSNCA), which is a method that also constructs a correlation network, but introduces a weight vector for each gene of the network, whose values are proportional to a gene's cross-correlation with all the other genes of the network. These weight vectors are used in a statistic measure to test for differences between two biological conditions.

In 2002, Ideker et al. proposed a method that attempts to identify subnetworks inside a larger gene network, of genes that present coordinated and meaningful changes in their expression values. In general, given a network with m genes, the authors classify each gene as differentially expressed between the biological conditions of interest and define a statistic based on test p values. Thus, random groups of m genes are selected inside the

Gustavo H. Esteves is the corresponding author.

©2018 Walter de Gruyter GmbH, Berlin/Boston.

dataset several times, the statistic is recalculated, and the mean and standard deviation are estimated. Finally, subnetworks are searched based on another statistic calculated from both of these values.

Alternatively, Alberich et al. (2014) have proposed a method called MP-Align for the pairwise comparison of metabolic pathways that primarily aims to identify similar parts between networks. Other studies have proposed models for the reconstruction or estimation of molecular networks or pathways, but these mainly utilize time course datasets in conjunction with mathematical and statistical methods and models (Chang, Gray & Tomlin, 2014; Kiani & Kaderali, 2014).

All the methods and models presented above are interesting and useful in a variety of situations, but none of them can handle with the topology of a gene regulatory network, especially aiming to verify its activation status in a specific time moment. Trying to mitigate this problem and given the importance of this type of data analysis for systems biology, we propose here a statistical method for the activation measuring of gene regulatory networks and metabolic pathways, i.e. KEGG pathways – <http://www.genome.ad.jp/kegg/>, (Kanehisa & Goto, 2000; Kanehisa et al., 2002). We use the p values obtained from the zero correlation statistical hypothesis test, as is performed for the gene co-expression networks, to construct a statistical measure that reflects the functional state of the network at a moment in time. This statistic has a known probabilistic distribution and can be used to easily classify regulatory or metabolic networks as active or inactive.

It is easy to see that an association measure is crucial to this type of data analysis, and the measure most commonly used is the traditional Pearson's linear correlation coefficient. However, this metric is known to be very susceptible to outliers, what is a common characteristic of gene expression datasets. So, some work has been presented that shows robust versions of correlation measure, specially focused on this type of data analysis (Hardin et al., 2007; Langfelder & Horvath, 2012). The shrinkage estimator for covariance and correlation matrices is also a good option (Schäfer & Strimmer, 2005). Other work shows a comparison between several correlation coefficients and other association measures, including mutual information (Song, Langfelder & Horvath, 2012).

Following, in the Methods section we present the statistic as defined in this work and a mathematical deduction of its probability distribution. In the Results section we present a numerical validation of the statistical distribution based on data permutation that makes it possible to measure the significance of a given result. Additionally, we present the results of the proposed measure for one small network and two KEGG pathways for a microarray dataset from gastric and esophageal samples. We have also analyzed both KEGG networks for a publicly available dataset, from Gene Expression Omnibus (GEO), available at <https://www.ncbi.nlm.nih.gov/geo/>, from the National Center for Biotechnology Information (NCBI), presenting the results in Section 1 from Supplementary Material. Finally, in the Discussion and conclusions section we present the main concluding remarks about the statistical methodology and the results obtained in this work.

2 Methods

In systems biology, a gene regulatory network is generally modeled by an oriented graph G in which nodes represent molecules (or in our case, genes) and edges represent interactions between molecules (i.e. nodes). Commonly, inducing associations are represented by a plus sign (+) or by +1 with an arrow, while repressing associations are denoted by a minus sign (−) or by −1 with a bar at the end.

Thus, given a graph G and a gene expression dataset, such as a microarray dataset, we can estimate the “weight” of each edge as the linear correlation value between each gene pair g_{i_1} and g_{i_2} in the network that is present in the dataset. Then, the significance of this correlation value must be evaluated by a specific hypothesis test or by permutation strategies. Because an edge can represent either an inductive or repressive association, we may test the null hypothesis (H_0) of $r(g_{i_1}, g_{i_2}) = 0$ against the alternative (H_A) of $r(g_{i_1}, g_{i_2}) > 0$ in the case of induction or against $r(g_{i_1}, g_{i_2}) < 0$ in the case of repression, where $r(g_{i_1}, g_{i_2})$ represents the linear correlation value between genes g_{i_1} and g_{i_2} .

Finally, denoting the total number of edges in G graph (the regulatory network) by S and the p value of the zero correlation test for the edge $s = 1, 2, \dots, S$ by p_s , the profile of activation for the G network can be quantified using the statistic

$$\mathcal{E} = -\frac{1}{S} \sum_{s=1}^S \log(p_s). \quad (1)$$

The statistic given in Eq. (1) above is a positive real number that is nearly one when few edges of the graph are active and increases with the number of active edges, thus the greater its value the greater the evidence

of network activation. Also, as the statistic is defined based on the p values of null correlation hypothesis tests for several edges from G graph, it is possible to argue about p values adjustment for multiple tests. However, these p values are used only to compose $\mathcal{G}$ statistic, and are not used to make any decision about the correlation values, so the adjustment must not be used here, because it changes the null hypothesis of uniform distribution of p values.

In this work, we used Pearson's linear correlation coefficient, but it has some limitations. The main one is that such a measure only estimates linear association between gene pairs, so non-linear relationships are not detected. Pearson's correlation is a good estimator for the covariance between two variables following normal distributions, but not for other families of probability distributions. This could be a problem with datasets where the gene expression values do not seem to follow the Gaussian density curve.

In this way, it is important to note that the Pearson's correlation coefficient should be used with caution, and to avoid some of these eventual problems it is possible to use other correlation values, such as Spearman, Kendall or mutual information (MI) values, bearing in mind that it is important to adjust the hypothesis test according to the association measure used. Special attention must be paid when using mutual information, because it is not possible to perform hypothesis tests for negative values because MI is only a positive measure.

2.1 $\mathcal{G}$ statistic distribution

This section addresses the deduction of the $\mathcal{G}$ statistic, given by Eq. (1), distribution. Let P_s be a random variable that gives the p value from the hypothesis test for the edge s in the graph. Assuming the null hypothesis that the gene regulatory network represented by graph G is inactive, P_s may assume any value within the $[0, 1]$ interval independently of the other edges in the graph. In other words, we can assume $P_s \sim U(0, 1)$, meaning that P_s follows a uniform distribution over $[0, 1]$ for all $s = 1, 2, \dots, S$.

Taking $Y = -\log(P_s)$, we find that

$$P(Y \leq y) = P[-\log(P_s) \leq y] = P(P_s \geq e^{-y}) = 1 - e^{-y},$$

indicating that $f_Y(y) = e^{-y}$, implying that $Y = -\log(P_s)$ follows an exponential distribution for the parameter 1 (one) for all $s = 1, \dots, S$. Then, defining a new random variable $Z = \sum_{s=1}^S -\log(P_s)$ and using the fact that the moment generating function (m.g.f.) of an exponential distribution with the parameter α is given by

$$\frac{\alpha}{\alpha - t}, \quad t < \alpha,$$

we can find the distribution of Z by taking its m.g.f.

$$\begin{aligned} g_Z(t) &= E(e^{tZ}) \\ &= E\left(e^{t \sum_{s=1}^S -\log(P_s)}\right) \\ &= E\left(e^{t(-\log P_1)}\right) \dots E\left(e^{t(-\log P_S)}\right) \\ &= \left(\frac{1}{1-t}\right)^S. \end{aligned}$$

However, $g_Z(t) = (1/(1-t))^S$ is the m.g.f. of a gamma (Γ) distribution with parameters S and 1 (one). Thus, $\sum_{s=1}^S -\log(P_s) \sim \Gamma(S, 1)$ and has a density probability function given by

$$f_Z(z) = \frac{1}{\Gamma(S)} z^{S-1} e^{-z}, \quad z \geq 0. \quad (2)$$

Finally, we can see that $\mathcal{G} = Z/S$, yielding

$$P(\mathcal{G} \leq x) = P\left(\frac{Z}{S} \leq x\right) = P(Z \leq xS) = F_Z(xS),$$

and

$$f_{\mathcal{G}}(x) = S f_Z(xS) = \frac{S^S}{\Gamma(S)} x^{S-1} e^{-xS},$$

which is the probability density function of a gamma distribution with parameters S and $1/S$. Thus, we have shown that $\mathcal{G} \sim \Gamma(S, 1/S)$. Figure 1 illustrates this density function for certain values of S .

Figure 1: Probability density functions from $\Gamma(S, 1/S)$ for $S = 1, 2, 7, 15$.

Note that a gamma distribution with parameters α and β has a mean and variance given by $\alpha\beta$ and $\alpha\beta^2$, respectively. We can see that $\mathcal{G}$ has a mean of one and a variance of $1/S$ for any value of S . Note that the variance decreases as the number of edges increases, which can be observed in Figure 1. Even with the fact that $\mathcal{G}$ has a mean of one for all values of S , the values obtained for this statistic may not be comparable between different networks due to differences in the variance relative to the number of edges.

To validate this theoretical probability distribution, we used a gene co-expression network generated from a real dataset previously analyzed by Gomes et al. (2005) and random permutations as described in the Results section. We next present the construction of hypothesis tests to verify the profile of activation and inactivation in any network G .

One problem associated with the distribution of the $\mathcal{G}$ statistic is the fact that it is deduced with the assumption of independence between the edges on the graph G . While this could be a reasonable assumption for edges that do not have a common node, this is not true for edges in this situation. In Section 2 from Supplementary Material, we present a simulation study to evaluate these assumption. In summary, we show that breaking this assumption do not significantly affects the distribution of $\mathcal{G}$, but can result in a little increase in the probability of false positives occurrence (Sup. Figures 3 and 4).

2.2 Hypothesis test construction

From the $\mathcal{G}$ statistic, given by Eq. (1), as well as its probability distribution under the null hypothesis of network inactivation, it is possible to perform two types of statistical hypothesis tests aimed at evaluating the status of the network. The more important of these is testing whether the network G is active under the proposed biological model (H_A) versus its inactivation (H_0). In other words, we can test the following hypothesis:

$$\begin{cases} H_0 : & \text{inactivation of } G \text{ (no statistical evidence of activation)} \\ H_A : & G \text{ network is active under the proposed model.} \end{cases}$$

Supposing that H_0 is true, $\mathcal{G}$ follows a gamma distribution as described above, and its value should remain relatively close to one. If, however, H_A is true, the $\mathcal{G}$ statistic should return large positive values [greater than one, as expected from the $\Gamma(S, 1/S)$ distribution], providing evidence for H_A and indicating statistical significance that the network appears do be “turned on” or active at that particular moment in time. This is therefore a right-tailed test.

For the second case, we can test whether the network is in a state of activation (in H_A), but in an opposite manner relative to the biological knowledge represented by the network. It is therefore possible to test the two hypothesis:

$$\begin{cases} H_0 : & \text{inactivation of } G \text{ (no statistical evidence of activation)} \\ H_A : & G \text{ network is inversely active relative to the proposed model.} \end{cases}$$

In this case, if H_A is true, $\mathcal{G}$ results in small positive values (near zero); thus, this test is left-tailed. However, in both situations, the null distribution of the $\mathcal{G}$ statistic is given by a gamma distribution parameterized by S and $1/S$, where S is the total number of edges in the biological network under consideration. In this manner, the p values for both tests can be easily computed from this distribution.

The first test presented above is paramount, in the sense that it evaluates network activation according to the proposed biological model. However, it is nevertheless possible to test a kind of opposite activation of the

network in cases where the resulting $\mathcal{S}$ values are very close to zero. This can be done using the second test presented above.

For both tests, if several gene regulatory networks are to be tested simultaneously it is important to do the p values adjustment for multiple tests. There are several methods for this procedure, and here we used the false discovery rate (FDR) to this end (Benjamini & Hochberg, 1995).

This statistical method has been implemented in the `activeNet` function of the `maigesPack` R package, which integrates other gene expression data analysis methods and is part of the BioConductor project (available at <http://www.bioconductor.org>).

2.3 Microarray dataset

Gomes et al. (2005) used microarrays to analyze the gene expression profile of gastric-esophageal cells. The expression of 4800 cDNA clones representing 4565 full-length human genes were observed in the RNA samples of 71 patients: 39 esophagus and gastric-esophageal junction (GEJ) samples [9 normal esophageal mucosa, 6 esophagitis mucosa, 10 Barrett's mucosa (4 of the long type and 6 of the short type), 5 adenocarcinomas of the esophagus, and 9 adenocarcinomas of the GEJ] and 32 stomach samples (6 normal body and antrum mucosa, 5 normal cardiac mucosa, 9 intestinal metaplasia mucosa, 7 samples of intestinal-type adenocarcinoma, and 5 samples of diffuse-type carcinoma). More details can be obtained in the original work.

Normalized data, as described in the original article, were used to validate the $\mathcal{S}$ statistic distribution and to exemplify the hypothesis tests using a small gene network and two KEGG pathways, as described below.

The KEGG pathways investigated here were cytokine-cytokine receptor interaction (id 04060) and glycerolipid metabolism (id 00561), which were obtained from KEGG database on 08/17/2015 using `retrieveKGML` function and parsed to graphs using the `parseKGML2Graph` function, both from the `KEGGgraph` package available for the R software for statistical computing (Ihaka & Gentleman, 1996). These two pathways were used because they appeared to be activated by Segal's method in the original work (Gomes et al., 2005).

The cytokine-cytokine receptor interaction network originally had 265 nodes and 254 edges. However, after matching the network with genes available in the dataset, we identified 45 nodes and 11 edges for this network, yielding coverages of approximately 17% and 4% for nodes and edges, respectively. For glycerolipid metabolism, we began with 59 nodes and 763 edges. However, a comparison with the gene expression data resulted in 17 nodes and 57 edges, approximately 29% and 7% coverage, respectively.

Thus, we used these two subnetworks obtained from the KEGG database mainly for the purpose of illustrating the operation of the method proposed here. Biologically, it is not reliable to discuss in depth the results obtained due to the poor coverage of both networks for this particular database. We also used a public database obtained through GEO repository to replicate the analysis of these two KEGG networks using an alternative gene expression data, as presented in the Section 1 from Supplementary Material.

3 Results

In this section, we present the results of the validation of the $\mathcal{S}$ statistic distribution and the values obtained for a small network and two KEGG pathways, illustrating the functionality of the method. All results were obtained for a real microarray dataset published previously, as described above.

3.1 Validation of $\mathcal{S}$ distribution

One way to validate the theoretical probabilistic distribution of $\mathcal{S}$ is data permutation, where we shuffled gene observations and recalculate the statistic for a network known to be active. In the process, we break any associations present into the network. Thus, for each data permutation, the $\mathcal{S}$ statistic given by Eq. (1) is recalculated, and the process is repeated a number of times to estimate the empirical distribution for the $\mathcal{S}$ statistic and to compare it with the theoretical distribution given in the Methods section.

For this, we require an activated network, so we constructed a co-expression network using 30 random genes with pairwise absolute correlation values above 0.75 from microarray dataset cited above (Gomes et al., 2005). This number of 30 random genes were defined empirically, as it is near the mean nodes for the two KEGG subnetworks used to illustrate the method. This generated a network with 30 nodes and 16 edges, with ten nodes without any in or out edges, as seen in Figure 2 (nodes without edges are not represented in the figure). Because the network was obtained from the dataset, all correlation values for each edge are real. $\mathcal{S}$ statistic calculation

resulted in an approximate value of 4.4, which is a rather unlikely value from a $\Gamma(16, 1/16)$ distribution, leading to a p value very close to zero (on the order of 10^{-15}) in the activation test for this co-expression network, as would be expected since it was constructed from the gene expression dataset.

Figure 2: Gene co-expression network constructed for 30 random genes. Ten nodes without edges are not represented in the figure. Edge labels represent estimated Pearson correlation values.

Then, gene expression values for this network were randomly permuted 2000 times, recalculating the $\mathcal{G}$ value each time. Figure 3 presents a histogram of these 2000 values, which gives the empirical distribution for $\mathcal{G}$ superimposed with the density probability function of $\Gamma(16, 1/16)$ in dashed line. As we can see in the figure, there is a good fit between the histogram and the theoretical density, indicating that the $\mathcal{G}$ statistic really seems to follow the $\Gamma(16, 1/16)$ distribution.

Figure 3: Histogram from the $\mathcal{G}$ statistic applied over a co-expression network with 30 nodes and 16 edges after 2 thousand permutations from gene expression values. Indicated by the dashed line is the $\Gamma(16, 1/16)$ theoretical distribution.

Another way to verify the theoretical model is to use `qqplot`. This type of graphic allows a comparison between quantiles from two groups of observations to verify whether they follow the same pattern. Thus, we can compare the sample quantile from the permuted $\mathcal{G}$ statistic with that of $\Gamma(16, 1/16)$. Figure 4 shows the good fit of the `qqplot`.

The main focus of this section was to show the good fit between the theoretical probability distribution deduced previously and the empirical one estimated by permutation methods and none of the results presented here are meant to be biologically meaningful or relevant.

3.2 Activation measuring of a small gene network

This subsection presents the results of the classification of a small and simple hypothetical gene regulatory network. According to the previous results of Gomes et al. (2005), it is possible to suppose that an increase in

the expression of the pro-inflammatory genes CCL20, CCL18 and IFNAR2 in normal tissues must be accompanied by an increase in expression of genes ADH1B, AKR1B10, ALDH3A2 and IL1R2 in order to promote the detoxification of products generated during the inflammatory reaction. When these interactions are broken, the accumulation of toxic products synthesized via the inflammatory process – in particular, the production of aldehyde – can favor the accumulation of mutations in DNA and an eventual malignant transformation.

Figure 4: *Qqplot* comparing the sample quantiles from the permuted $\mathcal{G}$ statistic with the $\Gamma(16, 1/16)$ theoretical ones.

From the idea behind the above explanation, it is possible to suppose that the expression values observed for the genes CCL20, CCL18 and IFNAR2 must be positively correlated with that ones for the genes ADH1B, AKR1B10, ALDH3A2 and IL1R2 in normal tissues and not associated for tumor ones; which can be represented by the graph in Figure 5.

Figure 5: Graph representing a simple gene regulatory network, where the expression levels for genes CCL20, CCL18 and IFNAR2 must be positively associated with the the ones for genes ADH1B, AKR1B10, ALDH3A2 and IL1R2 in normal tissues.

The graph at Figure 5 represents a gene regulatory network that could be active in normal tissues, both for stomach and esophagus and inactive for tumors. Obviously, some additional biological experimentation is needed to validate this hypothesis, but even so, this graph can be used to test the functionality of the statistical methodology proposed here.

In this way, we applied the $\mathcal{G}$ statistic and hypothesis test proposed above to verify the activation profile of this network using the original dataset, comparing tissue types classified as Normal (normal tissues from esophagus, GEJ and stomach), Metaplasia (Barrett’s esophagus and intestinal metaplasia from the stomach) and Adenocarcinoma (all adenocarcinomas from the esophagus and stomach, both intestinal and diffuse types). We found that for a p value less than 0.001 the $\mathcal{G}$ statistic of approximately 3.4 in normal tissues is significant, and it is possible to say that there is a strong statistical evidence for network activation for this biological condition, while it is lacking for metaplasia and adenocarcinomas (results in Table 1), as hypothesized above.

Table 1: Results of activation measuring for the gene network of Figure 5, with $\mathcal{G}$ values and respective p values.

Tissue type	$\mathcal{G}$ statistic	p value
Normal	3.403837	<0.001
Metaplasia	1.555719	0.061
Adenocarcinoma	1.208244	0.220

From these results, we could conclude that our findings seems to be in accordance with the biological explanation underlying the construction of the proposed network and demonstrate the great potential of the

statistical method proposed in this work. Biologically, as mentioned above, some additional experimentation is needed to validate these results.

3.3 Activation measuring of two KEGG networks

Thus far, we have presented a simulation-based result for the validation of a probabilistic model of the $\mathcal{S}$ statistic using a co-expression network and the resulting classification of a simple hypothetical network based on the results of original work by Gomes et al. (2005), whose behavior could be considered known among normal and tumor tissues.

Now we present the results for the activation measuring of two networks obtained from KEGG (cytokine-cytokine receptor interaction and glycerolipid metabolism) that appeared to be significantly altered in the original work using the gene group classification method proposed originally by Segal et al. (2004). Cytokine-cytokine receptor interaction was considered inactive in normal esophagus samples, whereas glycerolipid metabolism was inactive in adenocarcinomas and active in intestinal metaplasia of the stomach.

Using these two KEGG pathways presents a great opportunity for testing the statistical methodology proposed here under a more realistic scenario by calculating the $\mathcal{S}$ statistic and applying the respective activation test to both networks. To this end, we used the two subnetworks whose genes were present in the database. Figure 6 and Figure 7 show graphs representing these two subnetworks. In Figure 6, some nodes without edges are not represented in the figure.

Figure 6: Graph representing the subnetwork for cytokine-cytokine receptor interaction, after the match with genes from the dataset. Some nodes without edges are not represented in the figure.

For the cytokine-cytokine receptor interaction pathway, Table 2 shows that, at 5% significance level, the $\mathcal{S}$ statistic presented significant values only for normal tissue type ($\mathcal{S}$ approximately equal to 1.76, 1.27 and 1.10, for normal, metaplasia and adenocarcinomas, respectively), indicating that the subnetwork for this pathway seems to be activated in normal tissue and inactivated for the other two tissue types. In a similar manner, the glycerolipid metabolism subnetwork also presented similar results for the three tissue type ($\mathcal{S}$ approximately equal to 1.96, 0.92 and 0.97, for normal, metaplasia and adenocarcinomas, respectively). These results can be found in Table 3.

The method proposed by Segal et al. (2004) and used in the work of Gomes et al. (2005) classifies each sample type as active or inactive based on the number of genes over- or under-expressed, respectively, using a hypergeometric distribution to calculate the significance of the result. Thus, it may be classified as a kind of GSEA method, which does not consider the interrelation between genes that constitute the group. So the results obtained by Gomes et al. (2005) are not comparable with those ones presented in this section.

Figure 7: Graph representing the subnetwork for glycerolipid metabolism, after the match with genes from the dataset.

It is also important to note that any interpretations regarding the biology of these two networks based on the results obtained here should be made with caution, because only a small portion of the original edges (and thus, number of genes) are represented by the genes available in this database. Furthermore our results need biological validation to be discussed in more depth.

Table 2: Results of activation measuring for cytokine-cytokine receptor interaction subnetwork, with $\mathcal{E}$ values and respective p values.

Tissue type	$\mathcal{E}$ statistic	p value
Normal	1.757011	0.046
Metaplasia	1.274981	0.261
Adenocarcinoma	1.099118	0.338

The same type of analysis for both KEGG networks were done using a public database, available through GEO (access code GSE33651). The subnetworks identified in this case are shown in Sup. Figures 1 and 2, and the results obtained comparing normal and tumor gastric tissues are seen at Sup. Tables 1 and 2 (Supplementary Material).

Table 3: Results of activation measuring for glycerolipid metabolism subnetwork, with $\mathcal{E}$ values and respective p values.

Tissue type	$\mathcal{E}$ statistic	p value
Normal	1.961607	<0.001
Metaplasia	0.916536	0.727
Adenocarcinoma	0.967062	0.727

4 Discussion and conclusions

In this work, a simple but useful statistical method was presented for the activation measuring of genetic regulatory networks, represented by undirected graphs. This method is based on a simple statistic whose probability distribution, when H_0 is true, was deduced and verified through a study involving data permutation.

The statistic used here is not new to the statistical theory, but it is equivalent to one originally proposed by Fisher (1934) for use in applied statistics. Chang et al. (2013), for example, have evaluated various statistical

methods for the meta-analysis of microarray data, and one of the methods compared was based on Fisher's statistic. However, to our knowledge, no previous work has used this type of statistic or methodology for the activation measuring of gene regulatory networks or gene pathways, taking into account the network topology, as presented here.

Previous work have presented methodologies somewhat similar to what we proposed here. Segal et al. (2003) presented a probabilistic method, based in graphical models that searches for gene groups regulated by common regulatory genes or transcription factors. The method is iterative and is based into the maximization of a Bayesian score, using the EM algorithm. On the other hand, Ulitsky and Shamir (2007) presented an algorithmic framework, named MATISSE, that search for jointly active connected subnetworks (abbreviated by the authors as JACS). The method is based on a mixture of two gaussian distributions to search for highly interconnected subnetworks inside a greater gene network. However, both methods are still rather different from the statistical technique presented here.

In a context of application to gene expression data, we used a small network hypothesized from the results obtained in a previous study (Gomes et al., 2005). According to prior biological knowledge, this network should be activated in normal tissues but not in tumors, which was verified by applying our method to the original dataset. This gives a clear indication that the hypothesis test proposed here can be used for the activation measuring of gene networks from real gene expression data.

To verify its functionality in a more realistic situation, graphs of two pathways available in KEGG (interaction of cytokine-cytokine receptors and glycerolipid metabolism) were downloaded and indexed with genes available in the database. Both of them were classified as significantly activated in normal tissues and not activated for metaplasia and tumor ones. These results were similar to the measuring of the subnetworks of the same KEGG pathways identified for the GSE33651 dataset presented significant $\mathcal{Z}$ values for normal gastric tissues.

These results shows that the method could be used without major problems with these KEGG networks, and it is important to remember that only subnetworks were used for the classification presented here. Thus, any interpretation of biological processes related to entire pathways would be pure speculation.

It is important to reinforce that the main focus of this work was to present the statistical methodology proposed here. In this way, the activation measuring of these three gene networks was done just as didactic examples, that is, to illustrate the functionality of the methodology. So, no significant biological results are meant to be extrapolated from these results.

Also, we are currently trying to identify other gene expression datasets having considerably more genes where our methodology could be applied for a more detailed network level analysis, using KEGG networks which have more overlapping genes between the networks and the dataset, in order to evaluate with more biological confidence the efficacy of the proposed model.

In summary, our results show that the statistical methodology proposed in this paper can be used to analyze real data and can yield new insights into the biological processes associated with various tissue types or diseases. Notably, because the method is based on significance tests for linear correlation values, other association measures, like mutual information, Kendal and Spearman coefficients or robust correlations (Hardin et al., 2007) can be used in addition to the calculation of p values by bootstrap methods.

Finally, because all calculations are based on numerical data, this method is not only limited to microarray data and can be easily adapted for other types of gene expression databases, such as those for the relatively new next generation sequencing technologies (Shendure & Ji, 2008), like RNAseq or miRNA, for example. The method could also be adapted even for datasets of protein (and metabolites) levels (Boersema, Kahraman & Picotti, 2015), since this could be measured in large scale.

Acknowledgement

We would like to thank Raydonal Ospina Martinez and Diana Maia for additional support writing the manuscript and E. Jordão Neves and Roberto Hirata Jr. for help and suggestions regarding this work.

Funding

GHE has also been supported by CNPq (grant 478184/2012-3).

References

- Alberich, R., M. Llabrés, D. Sánchez, M. Simeoni, and M. Tuduri (2014): "MP-Align: alignment of metabolic pathways," *BMC Syst. Biol.*, 8, 58.
- Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock (2000): "Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat. Genet.*, 25, 25–29.
- Benjamini, Y. and Y. Hochberg (1995): "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. R. Stat. Soc. B*, 57, 289–300.
- Boersema, P. J., A. Kahraman, and P. Picotti (2015): "Proteomics beyond large-scale protein expression analysis," *Curr. Opin. Biotechnol.*, 34, 162–170.
- Butte, A. J. and I. S. Kohane (2000): "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," *Pac. Symp. Biocomput.*, 5, 415–426.
- Butte, A. J., P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane (2000): "Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks," *Proc. Natl. Acad. Sci. USA*, 97, 12182–12186.
- Chang, L.-C., H.-M. Lin, E. Sibille, and C. C. Tseng (2013): "Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline," *BMC Bioinformatics*, 14, 368.
- Chang, Y., J. W. Gray, and C. J. Tomlin (2014): "Exact reconstruction of gene regulatory networks using compressive sensing," *BMC Bioinformatics*, 15, 400.
- Draghici, S. (2003): *Data Analysis tools for DNA microarrays*, London: Chapman & Hall.
- Dudoit, S., Y. H. Yang, M. J. Callow, and T. P. Speed (2002): "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments," *Stat. Sin.*, 12, 111–139.
- Fisher, R. A. (1934): "Statistical methods for research workers," in *Biological monographs and manuals*, V, Edinburgh: Oliver and Boyd.
- Gomes, L. I., G. H. Esteves, A. F. Carvalho, E. B. Cristo, R. Hirata, W. K. Martins, S. M. Marques, L. P. Camargo, H. Brentani, A. Pelosof, C. Zitron, R. a. Sallum, A. Montagnini, F. a. Soares, E. J. Neves, and L. F. L. Reis (2005): "Expression profile of malignant and nonmalignant lesions of esophagus and stomach: differential activity of functional modules related to inflammation and lipid metabolism," *Cancer Res.*, 65, 7127–7136.
- Hardin, J., A. Mitani, L. Hicks, and B. VanKoten (2007): "A robust measure of correlation between two genes on a microarray," *BMC Bioinformatics*, 8, 220.
- Heyer, L. J., S. Kruglyak, and S. Yooseph (1999): "Exploring expression data: identification and analysis of coexpressed genes," *Genome Res.*, 9, 1106–1115.
- Ideker, T., O. Ozier, B. Schwikowski, and A. F. Siegel (2002): "Discovering regulatory and signaling circuits in molecular interaction networks," *Bioinformatics*, 18, S233–S240.
- Ihaka, R. and R. Gentleman (1996): "R: A language for data analysis and graphics," *J. Comput. Graph. Stat.*, 5, 299–314.
- Johnson, R. and D. Wichern (2002): *Applied multivariate statistical analysis*, 5th edition. New Jersey: Prentice Hall.
- Kanehisa, M. and S. Goto (2000): "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Res.*, 28, 27–30.
- Kanehisa, M., S. Goto, S. Kawashima, and A. Nakaya (2002): "The KEGG databases at GenomeNet," *Nucleic Acids Res.*, 30, 42–46.
- Kiani, N. a. and L. Kaderali (2014): "Dynamic probabilistic threshold networks to infer signaling pathways from time-course perturbation data," *BMC Bioinformatics*, 15, 250.
- Langfelder, P. and S. Horvath (2008): "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, 9, 559.
- Langfelder, P. and S. Horvath (2012): "Fast R functions for robust correlations and hierarchical clustering," *J. Stat. Softw.*, 46, 11.
- Mardia, K., J. Kent, and J. Bibby (1979): *Multivariate analysis*, New York: Academic Press.
- Rahmatallah, Y., F. Emmert-Streib, and G. Glazko (2014): "Gene Sets Net Correlations Analysis (GSNCA): a multivariate differential coexpression test for gene sets," *Bioinformatics*, 30, 360–8.
- Schäfer, J. and K. Strimmer (2005): "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Stat. Appl. Genet. Mol. Biol.*, 4, 32.
- Segal, E., N. Friedman, D. Koller, and A. Regev (2004): "A module map showing conditional activity of expression modules in cancer," *Nat. Genet.*, 36, 1090–1098.
- Segal, E., M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman (2003): "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data," *Nat. Genet.*, 34, 166–176.
- Shendure, J. and H. Ji (2008): "Next-generation DNA sequencing," *Nat. Biotechnol.*, 26, 1135–1145.
- Song, L., P. Langfelder, and S. Horvath (2012): "Comparison of co-expression measures: mutual information, correlation, and model based indices," *BMC Bioinformatics*, 13, 328.
- Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov (2005): "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Natl. Acad. Sci. USA*, 102, 15545–15550.
- Ulitsky, I. and R. Shamir (2007): "Identification of functional modules using network topology and high-throughput data," *BMC Syst. Biol.*, 1, 8.
- Yang, I. V., E. Chen, J. P. Hasseman, W. Liang, B. C. Frank, S. Wang, V. Sharov, A. I. Saeed, J. White, J. Li, N. H. Lee, T. J. Yeatman, and J. Quackenbush (2002): "Within the fold: assessing differential expression measures and reproducibility in microarray assays," *Genome Biol.*, 3, 62.
- Zhu, J. and M. Q. Zhang (2000): "Cluster, function and promoter: analysis of yeast expression array," *Pac. Symp. Biocomput.*, 5, 476–487.

Supplemental Material: The online version of this article offers supplementary material (<https://doi.org/10.1515/sagmb-2016-0059>).