DOI: https://doi.org/10.24297/jbt.v8i0.8298

Graphic Model-Based Gene Regulatory Network Reconstruction Using RNA Sequencing Count Data

Cristian Andrés González Prieto, Liliana López-Kleine

Universidad Nacional de Colombia, Bogotá D.C., Colombia

cragonzalezpr@unal.edu.co, llopezk@unal.edu.co

Abstract

Interactions between genes, such as regulations are best represented by gene regulatory networks (GRN). These are often constructed based on gene expression data. Few methods for the construction of GRN exist for RNA sequencing count data. One of the most used methods for microarray data is based on graphical Gaussian networks. Considering that count data have different distributions, the negative binomial distribution is much more likely for RNA-sequencing data. For this distribution, no model-based method for the construction of GRN has been proposed until now. Here, we present a graphical, model-based method for the construction of GRN assuming a negative binomial distribution of the RNA sequencing count data. The R code is available under request. We used the method proposed both on simulated RNA sequencing count data and on real data. The graph is showen, and its descriptive measurements were assessed. Some interesting biological conclusions were found. We confirm that using negative binomial distribution for fitting the model is suitable because RNA sequencing data presents overdispersion.

Keywords: Gene Regulatory Network, RNA-Seq, Graphical Model, Negative Binomial Distribution, Overdispersion, Networks.

Introduction

Gene regulatory networks are very informative in order to understand the underlying mechanisms of biological functions at the molecular level. Their reconstruction is based on several assumptions about the data (often gene expression data) and the model describing relationships between genes. When the statistical distribution of the variables, from which the gene expression samples for each gene are obtained, is known, graphical models allow establishing relationships between them [1–4]. Graphical models have been traditionally used for the construction of GRN. Networks constructed using a Graphical Gaussian Model (GGM) are probabilistic graphical models that describe the conditional relationship between genes under the assumption that the gene expression data comes from a multivariate normal distribution. There are directed and undirected GGM. The undirected GGMs, are represented by a non-directed graph$\mathrm{G} = (\mathrm{V}; \mathrm{U})$, where the nodes $\mathrm{v}_1; \ldots; \mathrm{v}_\mathrm{p}$ represent the ensemble of variables (in gene networks, they are the expression profiles of a gene across all samples) and U is the ensemble of vertices. When a random vector $\boldsymbol{X}$ has a distribution $\mathrm{N}_\mathrm{P}(\mu, \Sigma)$, the graph G represents a model where $\mathrm{K} = \Sigma^{-1}$ is a semidefinite positive matrix with $\mathrm{k}_{\mathrm{il}} = 0$ when no edge between the nodes i and l exists. This kind of graph is called a dependence graph because when i and l are adjacent, then $\mathrm{i} \perp\!\!\!\perp \mathrm{l}|\mathrm{U}\text{-}\{\mathrm{i}, \mathrm{l}\}$. These graphs comply the global Markov property [5]. This models have been widely used for the construction of GRN on continuous microarray data as reviewed by Liu [6].

Very few methods have been proposed for discrete RNA-sequencing count data. Jia et al. (2017) [7] proposed a methodology for RNA sequencing data after a transformation that conducts to a continuous distribution. The transformation is based on models of random effects and after transformation, a GGM can be applied to these data. For RNA sequencing count data without transformation, Allen and Liu (2013), have proposed a method assuming a Poisson distribution of the RNA sequencing counts. Their proposal is based on the local Markov property in which each variable has a Poisson distribution conditional to all others. For the construction of the network, they propose an algorithm, which selects a node (a gene) and starts searching for relationships with all other gens via penalized regressions. Gene neighbors, for which the coefficients of the penalized regression are

not significative, no edge is established [1]. Nevertheless, this model is not completely adapted for RNA-sequencing data, because overdispersion is an important issue, when Poisson distribution is assumed [8].

Karbalayghareh and Hu (2015) [9] proposed a method for RNA-sequencing time series data based on a log-linear model for the temporal evolution of the gene expressions and modeled the gene expression levels by a negative binomial distribution. Their method also controls the network sparseness. Thorne (2018) [10], developed a model-based method, assuming negative binomial distribution and uses sparse regression with a horseshoe prior to learn a dynamic Bayesian network of interactions between genes. This method was also developed for time series data.

The graphical models based on generalized linear models (GM-GLM) appear when node conditional distributions of the exponential family are considered [11]. Let $\mathrm{X} = (\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{p})$, be a random vector where each X_i takes values from an ensemble X and $G = (U; V)$ be a non-directed graph on p nodes. The graphical model is a set of distributions that satisfy the independence Markov assumptions of the graph. Based on the Hammersley-Clifford theorem, it can be assumed that the node-conditional distributions make part of the exponential family. Let us assume that the distribution of X_s given all the other nodes X_{V-s} is given by the exponential family but the canonical parameter of the exponential family is a linear combination of order k of the univariate functions. This allows defining the joint distribution as shown in [11]. Therefore, a regression can be applied to adjust a model explaining the RNA measurements for each gene (its gene expression profile), based on all other genes of the data set.

As this procedure is computationally very costly, a strategy to avoid considering all remaining genes is to use a regularized regression. Regularization methods try to find the most parsimonious model selecting only some of the predictor variables (in this case: potential neighbors of each gene) [12–14]. Some of the most known regularization methods are Ridge, LASSO (Least Absolute Shrinkage and Selection Operator) or Elastic Net [12,14,15].

Here, we propose to apply a regularized regression for each gene, considering all possible neighbors, assuming that the gene expression count data are independent between samples and follow a negative binomial distribution.

Materials and Methods

2.1 Assumptions for Negative Binomial model

A random vector $\mathrm{X} = (\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{p})$ exists and is associated to a graph $G = (V; U)$. Each X_i corresponds to a genetic profile e_i. Independence is assumed between columns. Assuming that the gene/node conditional distribution is negative binomial and therefore belongs to the exponential family. These distributions can be combined to obtain a negative binomial random Markov field:

$$P(\boldsymbol{X}; \boldsymbol{\Theta}) = \exp\left\{\sum_{i \in V} \theta_i X_i + \frac{\Gamma(\phi + x_i)}{\Gamma(x_i + 1)\Gamma(\phi)} + \sum_{(i,j) \in U} \theta_{ij} X_i X_j - A(\boldsymbol{\Theta})\right\}$$

With $\mathrm{A}(\Theta) = \phi \log(1 - \exp(\theta))$, but by construction $\theta < 0$. This means that only inverse relationships between nodes can be found. The solution we propose follows the one proposed by Allen y Liu (2013) [1] and consists in proposing a local graphic binomial model, as the properties of a local Markovian field are maintained, but no joint distribution can be specified. The local Markov property, that defines the independence between two variables conditioned to its neighborhood is still valid. The conditionally independent relationships allow estimating the complex dependency structures [1].

2.2 Quality assessment of the model

The quality of the model is assessed using a pseudo regression coefficient as follows [16]:

$$Pseudo - R^2 = 1 - \frac{D(x_{ik}, \hat{\mu}_{ik})}{D\left(x_{ik}, \hat{\mu}_{ik}^{(0)}\right)}$$

where $D(x_{ik}, \hat{\mu}_{ik})$ is the deviance of the adjusted model and $\mathrm{D}\left(\mathrm{x_{ik}}, \hat{\mu}_{\mathrm{ik}}^{(0)}\right)$ is the deviance of the saturated model.

Departure from the assumed distribution was also tested using confidence bands obtained via simulations [17] and visualization of them on an envelope plot. These bands can give information about big distances from the expected normality line. Moreover, this plot can show if influential points or heteroscedastic variance exists [18].

2.3 GRN construction

Based on the above exposed theoretical assumptions, the construction of the GRN based on the expression profiles obtained from an RNA sequencing count data set is as follows:

1. For each gene, a negative binomial general linear model $X_{ik}|X_{V-i} \sim NB(\mu_{ik}, \theta)$ is adjusted using a penalized regression. Here, we used the Elastic Net penalization method. The adjustment is evaluated using pseudo R^2 regression coefficient and the plots of residuals vs. adjusted values.

2. Coefficients of genes that are non-zero after regularized regression are retained to establish relationship between genes. The sign indicates the direction of the relationship.

3. These coefficients are retained in a gene-by-gene coefficient matrix, which is not symmetric due to the fact that relationships to all genes are adjusted given all remaining genes. The rules for the construction of the coefficient matrix we propose are as follows:

   a. If the coefficient between $Gene_i \sim Gene_j$ is above or below 0 and the coefficient between $Gene_j \sim Gene_i$ is 0, this indicates no regulatory relationship between genes and the corresponding fields in the coefficient matrix will be 0.

   b. When the coefficients $Gene_i \sim Gene_j$ and $Gene_j \sim Gene_i$ have the same sign in positions (i, j) and (j, i) he sum of both will appear in the coefficient matrix in these two fields.

   c. When the coefficients $Gene_i \sim Gene_j$ and $Gene_j \sim Gene_i$ have different signs in positions (i, j) and (j, i) the sum of both will appear in the coefficient matrix in these two fields.

4. Once the coefficient gene by gene matrix is established, the adjacency matrix describing the graph is constructed using the following adjacency function:

$$\boldsymbol{A}(\beta_{ij}) = \begin{cases} a_{ij} = 1, & if\ \beta_{ij} > 0 \\ a_{ji} = 0, & if\ \beta_{ji} > 0 \\ a_{ij} = 0, & if\ \beta_{ji} < 0 \\ a_{ji} = 1, & if\ \beta_{ij} < 0 \end{cases}$$

The regularization models are implemented in the R package mpath [19] and all other functions were coded in R (see supplementary data).

2.4 Simulated data

The simulated RNAseq count data was generated using the function makeExampleCountDataSet from the DeSeq package [8] of Bioconductor [20] for 100 genes and 30 samples (15 for a control condition and 15 for a treatment condition). 30% of these genes have a differential expression between conditions. Mean expression values where obtained from an exponential distribution with parameter 1/250.

2.5 Real data

The real data used to illustrate the methodology is freely available at the Gene Expression Omnibus of the NCBI data bank under the accession number GSE72548. RNA was obtained from *Arabidopsis thaliana* wild-type roots and mutant roots non-infected and infected with *Heterodera schachtii* nematodes [21]. This dataset consists of 24 samples. We constructed a GRN for 100 genes selected for their highest mean expression values (Shanks et al 2016) [21] and identified the 10 most connected genes (hubs).

Results and Discussion

The pseudo R^2 was calculated for all models (all genes) in order to assess adjustment quality. The mean of the pseudo R^2 for the models on simulated data sets ranged from 0.5715 to 0.9998; a very low proportion of low-quality models was observed. A similar behavior was observed for the real data set: pseudo R^2 ranged from 0.9791 to 1. The mean of residuals was also close to zero in all simulations and only a very few genes showed high residuals (Figure **1**), which confirms the findings based on pseudo R^2. Similar results were obtained for real data (Figure **2**). A summary of the simulated network's characteristics can be found in supplementary Table **1**. The obtained networks are very consistence having similar number of nodes and average degree, which indicates that the model is robust.

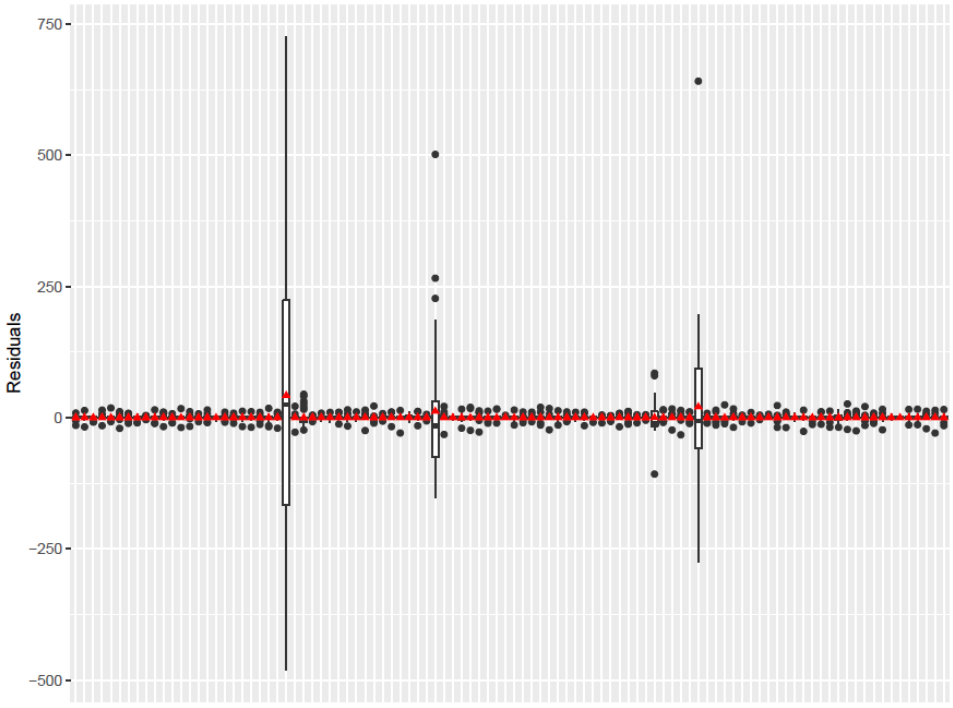

Figure 1: Boxplots of residuals for each fitted model in simulated data.

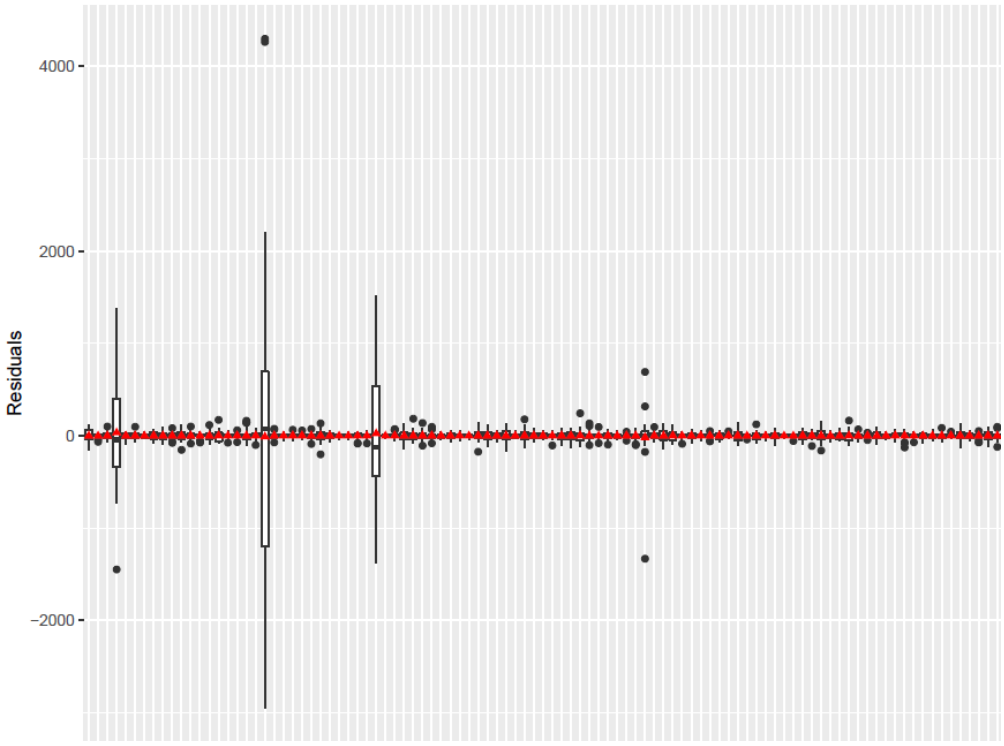

Figure 2: Boxplots of residuals for each fitted model in real data.

In order to confirm if an overdispersion exists and that the right model to adjust is NB and not Poisson (in which case they should not exceed a value of one), the scale parameter for each model was calculated, these scaling parameters that ranged from 2.8 to 56790.8 for simulated data and from 368.2 to 31229539 for the real data set, confirming that a NB is more suitable due to overdispersion. Moreover, envelope plots show that departure from the assumed distribution is very low (Supplementary figures **1** for simulated data and **2** for real data).

The resulting networks showed few highly connected genes as expected for a real biological network. For the 10 simulated datasets between 69 and 88 genes were connected to other genes and the mean degree of genes ranged from 6.41 to 13.19. For more details on these networks see supplementary table **1**.

On the real data set between 1 and 46 genes were connected to other genes; the mean indegree of genes ranged from 0 to 26 and the mean outdegree of genes ranged from 0 to 23. Also, in this network, some highly connected genes, or hubs were present (See table **1**). These hubs have an important biological meaning. They are related to biotic stress (AT4G33720, AT3G01420, AT4G21960), signaling (AT4G33720, AT2G05440, AT2G05510, AT3G09260, AT3G01420, AT4G21960, AT2G43150) and metal binding (AT4G30190, AT3G01420, AT4G21960, AT1G20620). The first two biological functions were identified through detection of differentially expressed genes in this data set when originally analyzed [21]. They are related to the different behavior of mutants in response to the presence of nematodes. This indicates that the negative binomial model-based gene regulatory network reconstruction using RNA sequencing count data is adequate.

Table 1: Hubs genes in the experiment and some descriptive measurements of the network.

Gene names	Edge Count	Indegree	Outdegree
AT3G09260	46	26	20
AT4G21960	42	23	19

AT2G05440	41	21	20
AT1G20620	35	12	23
AT2G05510	32	19	13
AT4G33720	31	16	15
AT2G43150	29	16	13
AT4G30190	29	16	13
AT2G21660	29	21	8
AT3G01420	28	16	12

Discussion

The here proposed method using on a model-based gene regulatory network reconstruction for RNA sequencing count data is a novel approach that is adapted for the most likely negative binomial (NB) distribution of counts and expands therefore methods based on other distributions like the Poisson distribution [1]. As was already pointed out earlier [7] and we could confirm based on real and simulated data, dispersion of RNAseq count data is higher than could be expected for Poisson distribution. Therefore, it is correct to assume the negative binomial distribution and apply methods developed for it as the one we propose here.

Other methods have been proposed for RNAseq time course data, assuming NB distribution [8,9], but are restricted to this type of data, which is very rare. Most experiments are replicates of two conditions being compared because they were designed originally to detect differentially expressed genes. The flexibility or our approach allows obtaining GRN for most of the available RNAseq count data.

The results indicate that adjustment to the model is good and therefore describes correctly regulatory relationships between genes. Moreover, the network obtained on real data confirms that hubs among most related genes reflect the central role of those genes in the phenotype analyzed in the A. thaliana dataset.

Conclusions

We showed that using negative binomial distribution for fitting the model is suitable because RNA sequencing data present overdispersion. The network obtained on real data confirms that hubs among most related genes reflect the central role of those genes in the experiment.

Data Availability (excluding Review articles)

The real datasets are publicly available and accession numbers have been indicated. The code for the simulated data is also available.

Conflicts of Interest

The authors have no competing interests to declare.

Funding Statement

The here presented work is the result of a Master final work at the Statistics Department of Universidad Nacional de Colombia – sede Bogotá

Acknowledgments

This article is the result of CG Master final work in Statistics under the direction of LLK.

References

1. Allen GI, Liu Z. A local poisson graphical model for inferring networks from sequencing data. IEEE Trans Nanobioscience 2013;12:189–98.
2. Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. Nat Rev Mol Cell Biol 2008;9:770–80.
3. Luo F, Yang Y, Zhong J, Gao H, Khan L, Thompson DK, et al. Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. BMC Bioinformatics 2007;8:299.
4. Liang F, Song Q, Qiu P. An Equivalent Measure of Partial Correlation Coefficients for High-Dimensional Gaussian Graphical Models. J Am Stat Assoc 2015;110:1248–65.
5. Højsgaard S, Edwards D, Lauritzen S. Graphical models with R. Springer Science & Business Media; 2012.
6. Liu Z-P. Reverse Engineering of Genome-wide Gene Regulatory Networks from Gene Expression Data. Curr Genomics 2015;16:3–22. doi:10.2174/1389202915666141110210634.
7. Jia B, Xu S, Xiao G, Lamba V, Liang F. Learning gene regulatory networks from next generation sequencing data. Biometrics 2017;73:1221–30.
8. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010;11:R106.
9. Karbalayghareh A, Hu T. Inference of Sparse Gene Regulatory Network from RNA-Seq Time Series Data. IEEE Glob Conf Signal Inf Process 2015:967–71. doi:10.1186/s12859-018-2125-2.
10. Thorne T. Approximate inference of gene regulatory network models from RNA-Seq time series data. BMC Bioinformatics 2018;19:1–12. doi:10.1186/s12859-018-2125-2.
11. Yang E, Allen G, Liu Z, Ravikumar PK. Graphical models via generalized linear models. Adv. Neural Inf. Process. Syst., 2012, p. 1358–66.
12. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Statistical Methodol 2005;67:301–20. :57–68.

Supplementary Materials

Table S1: Global networks measures obtained on simulated data

Figure S1: Envelope for models adjusted to genes of simulated data

Figure S2: Envelope for models adjusted to genes of real *A. thaliana* data