Internship report: **Using Bayesian networks to predict some regulatory systems components of Escherichia coli from microarray gene expression data**

Diego Mauricio Garcia Jimenez, PhD(**c**), Universidad del Valle, Colombia

Centro de ciencias genomicas – UNAM, México

July 2017

1. **Introduction**

Identification and analysis of signature genes and its interactions are keys on pharmacist design process, improve physiological of plants and research of disease as oncogenesis, among others. it is possible to analyze the collective behavior of genes using models that represent the interaction between them. Some of these models are gene co-expression networks and gene regulatory networks; where we can observe the expression patterns that follow gene groups under given biological conditions and the impact of the behavior of one group of genes over other. The networks in question have been represented with models like Bayesian network and we can infer its structure and parameters using statistical methods like Markov Chain Monte Carlo simulation (MCMC) and Bayesian inference.

Purpose of this work is to valid the useful of Bayesian networks to predict several bacterial regulatory system components.

First, we obtain gene expression data of 4297 genes of Escherichia coli and 466 observations in several biological conditions. We use a subset of modular genes and then discretize the data into three categories.

Second, we learned Bayesian networks of gene interactions adopting MCMC simulation and Bayesian inference methods. Specifically, we use Metropolis Hasting algorithm and Posterior distribution to approximate the parameters and structure of network. To avoid overfitting in the learning we include two constraints during simulations: maximum in-grade in nodes and minimum dependence in edges of network.

Third, we compare interactions found with interactions published of modular genes in bacterial databases with strong evidence. For some subsets of genes, we use Jaccard index in the comparison.

Finally, we found Bayesian networks that correspond to some components of modular genes like Lactose in Escherichia coli, among others. Predictions was successfully valided with respect to gene interactions published.

In this work, we see that Bayesian networks can help us to understand some bacterial regulatory systems components, mainly components of modular genes, however, only its learning method from microarray gene expression data can’t say us about other components such as global regulators, inter-modular regulators and basal machinery.

1. **Methods**
   1. **Selecting cluster of genes from gene co-expression network**

Before applying Bayesian learning, we must get a gene co-expression network then select one or more network’s cluster. Selection can be queried by biologist depending of study or also we can do it randomly or deterministically.

Firsts, we need obtain microarray gene expression data from a public source like GEO (Barrett 2013), ArrayExpress (Kolesnikov 2015), Colombos (Moretto et al. 2016) or (Faith et al. 2008), among others.

After, we can calculate correlation coefficient (like Pearson or Kendall) between each pair of expression vector of genes selected then we joint it to get a correlation matrix that will represent a gene co-expression network. See (Wikipedia 2017).

Last, we must choose a cutoff for correlation coefficient values and each cluster will be conformed with those pair genes that exceeds cutoff threshold chosen.

* 1. **Applying Bayesian framework to each cluster**

Now, for each cluster of genes selected, we proceed loading its samples from data of expression, later we apply our Bayesian learning method to the samples loaded and next we take the results and graph its networks and traces, respectively.

In this work, each gene expression vector must be submitted to a discrete process before Bayesian learning stage. We will use quantiles discretization method with three levels (quantiles). See (Faming 2010).

Our Bayesian learning method consist on an adaptation of Metropolis Hasting (MH) algorithm. Such adaptation consists on doing a random walking over search space of possible structures and evaluate it from posterior distribution to data with accept criteria of MH algorithm (Koller and Friedman 2009).

Once finished MCMC simulation, we analyze results of each iteration looking those graphs that have greatest posterior density function value (score), among other properties (like minimum in-grade in its nodes and maximum conditional dependence in its edges).

Finally, we convert the sub-set of graphs selected to a weighted network and its interactions proposed is compared with interactions known (Ibarra-Arellano 2016). i.e. we do it averaging occurrence of each edge over cardinality of sub-set and again we choose a cutoff for weight values and weighted network will be conformed with those edges that exceeds cutoff threshold chosen.

1. **Results**
   1. **Gene co-expression network**

Deterministically, we selected a subset of 16 genes that correspond to 5 different modules and its global regulators, see (Strong Evidences) Escherichia coli str. K-12 substr. MG1655 (Ibarra-Arellano 2016):

* Module 34 – Lactose transport (GO:0015767)

lacA – galactoside O-acetyltransferase monomer

lacI – LacI DNA-binding transcriptional repressor

lacY – lactose / melibiose:H+ symporter LacY

lacZ – β-galactosidase monomer

Global regulators:

rpoD – RNA polymerase, sigma 70 (sigma D) factor

hns – H-NS DNA-binding transcriptional dual regulator

crp – CRP transcriptional dual regulator

* Module 37 – Zinc ion transport (GO:0006829)

znuA – Zn2+ ABC transporter - periplasmic binding protein

* Module 40 – Sodium ion transport (GO:0006814)

ttdA – L-tartrate dehydratase, α subunit

ttdB – L-tartrate dehydratase, β subunit

ttdR – Dan

ttdT – tartrate:succinate antiporter

* Module 52 – Phosphorelay signal transduction system (GO:0000160)

zraP – zinc homeostasis protein

zraR – ZraR transcriptional activator

zraS – ZraS sensory histidine kinase

* Module 55 – Response to arsenic-containing substance (GO:0046685)

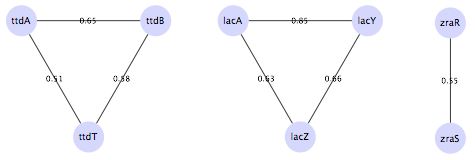
arsB – ArsB

Constructing of gene co-expression network was done jointing the correlation coefficient values calculated using the function “cor” of R with “pearson” argument. Gene expression data was taken from (Faith et al. 2008) and the resulting matrix is the following:



**Figure 1** Adjacency matrix of gene co-expression network.

The values highlighted in yellow (***Figure 1***) correspond to a cutoff of 0.5, i.e. values greater than 0.5 are into threshold and its genes will conform the clusters, so clusters selected are the following:



**Figure 2** Clusters selected from gene co-expression network.

* 1. **Bayesian learning**

Clusters showed in ***Figure 2*** are the input for Bayesian learning but before we must discretize its expression values with the “discretize” function in the “bnlearn” package of R. See (Nagarajan, Scutari and Lbre 2013).

Later, we execute three simulations of 1000 iterations, one per cluster, and we obtain the following results:



**Figure 3** Simulation of 1000 iterations for lacA, lacY and lacZ genes.



**Figure 4** Simulation of 1000 iterations for ttdA, ttdB and ttdT genes.



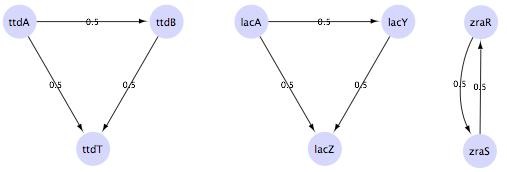
**Figure 5** Simulation of 1000 iterations for zraR and zraS genes.

We can see in the figures of above that optimal level of the density function posterior to the data is achieved during the firsts iterations in all simulations. Also, we can observe that the number of edges and the maximum in-grade are constant during simulations.

In summary, we obtained 6 different Bayesian networks in the first and second simulation and 2 Bayesian networks in third simulation.

* 1. **Weighted network**

In previous sub-section, we obtained several Bayesian networks as result of learning process. Now we take the sub-set of networks and averaging each edge over total of networks obtained, we get a weighted network per simulation. The result is shown to continuation:



**Figure 6** Weighted networks learned corresponding to modules 40, 34 and 52 of E. coli.

1. **Discussion**
   1. **Relation between co-expression network and Bayesian network**

We can observe in ***Figure 2*** that co-expressed genes inside each cluster also it present dependence relations between them, as it is shown in ***Figure 6***. If we see this from point of view mathematics seems logic that two variables whose values present high correlation, also it shows conditional dependence, talking probabilistically, i.e. these results are an effect numeric and no biological. May be this the reason because from Bayesian paradigm of statistics is not enough for to predict other bacterial regulatory system component like regulators or basal machinery if not there are high numeric correlation between them from expression data. In conclusion, we recommend articulate other machine learning approaches to complement this paradigm and achieve better predictions.

* 1. **Trend to overfitting when we try select more than one cluster**

Other observation is trend to overfitting in the Bayesian learning process. For instance, we try adding other component (lacI regulator gene) in the simulation of the lacA, lacY and lacZ genes and we saw the addition of edges from lacI to each modular gene. The following is the weighted network of the simulation:



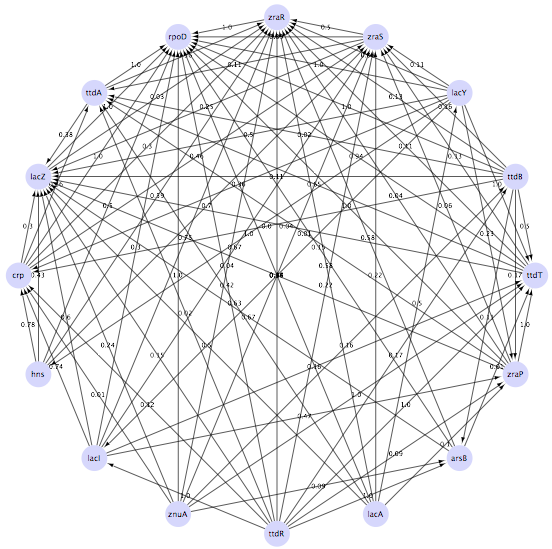
**Figure 7** Weighted network of module 34 after of adding lacI regulator gene.

The same effect is observed when adding rpoD and znuA genes:



**Figure 8** Weighted network of module 34 after adding rpoD and znuA genes.

Overfitting can become chronic if we include more than one cluster in one simulation. The ***Figure 9*** shows this situation:



**Figure 9** Trend to overfitting when several clusters are included inside one MCMC simulation.

In conclusion, we recommend highly to apply Bayesian learning process only to clusters of gene co-expression networks to avoid overfitting of Bayesian networks.

1. **Conclusions and future work**

* We found Bayesian networks that correspond to some components of modular genes like Lactose in Escherichia coli, among others. Predictions was successfully confirmed with respect to gene interactions published.
* We see that Bayesian networks can help us to understand some bacterial regulatory systems components, mainly components of modular genes, however, its learning method only from microarray gene expression data cannot say us something about other components such as global regulators, inter-modular regulators or basal machinery.
* A future work may be analyzing possibility of articulate Bayesian learning with other learning approaches like neural networks, fractal analysis, contact maps or dependences networks (based on constraints) to do prediction of other bacterial regulatory system components.

**References**

Barrett, T. a. W. S. E. a. L. P. a. E. C. a. K. I. F. a. T. M. a. M. K. A. a. P. K. H. a. S. (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research,* 41**,** D991-D995.

Faith, J. J., M. E. Driscoll, V. A. Fusaro, E. J. Cosgrove, B. Hayete, F. S. Juhn, S. J. Schneider & T. S. Gardner (2008) Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res,* 36**,** D866-70.

Faming, L. 2010. Learning Bayesian Networks for Gene Expression Data. In *Bayesian Modeling in Bioinformatics*, 271-291. Chapman and Hall/CRC.

Ibarra-Arellano, M. A. a. C.-G. A. I. a. T.-Q. L. G. a. T. A. a. F.-G. J. A. (2016) Abasy Atlas: a comprehensive inventory of systems, global network properties and systems-level elements across bacteria. *Database,* 2016**,** baw089.

Kolesnikov, N. a. H. E. a. K. M. a. M. O. a. T. Y. A. a. W. E. a. D. M. a. K. N. a. B. M. a. B. T. a. (2015) ArrayExpress update—simplifying data submissions. *Nucleic Acids Research,* 43**,** D1113-D1116.

Koller, D. & N. Friedman. 2009. *Probabilistic graphical models : principles and techniques*.Cambridge, MA: MIT Press.

Moretto, M., P. Sonego, N. Dierckxsens, M. Brilli, L. Bianco, D. Ledezma-Tejeida, S. Gama-Castro, M. Galardini, C. Romualdi, K. Laukens, J. Collado-Vides, P. Meysman & K. Engelen (2016) COLOMBOS v3.0: leveraging gene expression compendia for cross-species analyses. *Nucleic Acids Res,* 44**,** D620-3.

Nagarajan, R., M. Scutari & S. Lbre. 2013. *Bayesian Networks in R: with Applications in Systems Biology*. Springer Publishing Company, Incorporated.

Wikipedia. 2017. Gene co-expression network --- Wikipedia, The Free Encyclopedia.