Arabidopsis thaliana computationally-generated next-state gene interaction models*

David J. John Bree
Ann LaPointe
djj@wfu.edu / Wake Forest University. NC, USA

James L. Norris
norris@wfu.edu / Wake Forest University. NC, USA

Alexandria F. Harkey
Joëlle K. Muhlemann
Gloria K. Muday
muday@wfu.edu / Wake Forest University. NC, USA

Receipt: 6-8-2018 / Acceptance: 3-9-2018

Abstract. The construction of gene interaction models must be a fully collaborative and intentional effort. All aspects of the research, such as growing the plants, extracting the measurements, refining the measured data, developing the statistical framework, and forming and applying the algorithmic techniques, must lend themselves to repeatable and sound practices. This paper holistically focuses on the process of producing gene interaction models based on transcript abundance data from Arabidopsis thaliana after stimulation by a plant hormone.

Keywords: OpenFlow, data center, artificial neural network, Knowledge-Defined Networking

Modelo generado computacionalmente de interacción genética del próximo estado basado en *Arabidopsis thaliana*

Resumen. La elaboración de modelos de interacción genética debe ser un esfuerzo totalmente intencional y colaborativo. Todos los aspectos de la investigación, tales como el cultivo de las plantas, la obtención de las mediciones, el refinamiento de los datos recopilados, el desarrollo del marco estadístico, y la formulación y aplicación de técnicas algorítmicas, deben colaborar entre sí para establecer prácticas reproducibles y eficaces. Este artículo se centra, de manera holística, en el proceso de creación de modelos de interacción genética basados en los datos de la abundancia de transcritos obtenidos de la estimulación de la planta *Arabidopsis thaliana* mediante hormonas vegetales.

Palabras clave: OpenFlow, Flow, centro de procesamiento de datos, red neuronal artificial, Defined Networking

* Bree LaPointe thanks the Wake Forest University Center for Molecular Signaling for her support Aarch Assistant. The authors thank the National Science Foundation for their support with grant NSF#1716279.

1. Introduction

Our process of creating gene interaction models from Arabidopsis thaliana gene transcript abundance data involves multiple specialized steps supervised by biologists, biochemists, computer scientists, mathematicians, and statisticians. This has been a long-term interdisciplinary collaborative commitment, which ultimately has yielded and continues to yield models that provide testable hypotheses of gene pathways. The research into the construction of gene interaction models is an active area. Various groups of researchers have taken a number of different modeling approaches for time-course measurements. The modeling approach and setting presented herein is based on the stimulation of Arabidopsis thaliana with either the plant hormone auxin or ethylene at time 0, the collection of three replicates of gene transcript abundance measurements taken at 8 time points, and the creation of interaction models by rigorously-developed computational techniques guided by relative posterior probabilities of directed acyclic graphs.

The creation of gene interaction models is an active area of research. Algebraic techniques (Allen, Fetrow, Daniel, Thomas, & John, 2006; Laubenbacher & Stigler, 2004; Liang & Han, 2012; Stigler, 2007), differential equations (Cao, Qi, & Zhao, 2012), and partial correlations (de la Fuente, Bing, Hoeschele, & Mendes, 2004; Krämer, Schäfer, & Boulesteix, 2009; Li & Gai, 2008; Wille et al., 2004) are some of the approaches applied to this important problem. The techniques discussed herein are all based on our mathematically rigorous Bayesian probabilistic techniques (Patton, John, & Norris, 2012; Patton, John, Norris, Lewis, & Muday, 2013, 2014; Norris, Patton, Huang, John, & Muday, 2015; John, Fetrow, & Norris, 2011).

The biological thrust of this research is to understand lateral root development. Arabidopsis thaliana, the lab rat for plants, is specifically studied in this research. Among the reasons for using this plant are the ease in which it can be grown and propagated for similarity, as well as the extensive literature and databases on the plant, its genes and proteins.

The one of this research is to create hierarchical gene interaction models. Essentially, this means that there is an overall truth about the interaction of the genes, and any model should capture some elements of that truth. Also, various time paradigms can be applied to time-course data. The next-state paradigm is used exclusively in this paper. It proposes that, if there is a directed edge from gene A to gene B in the true biochemical network, the measurement of A's expression at time t has an influence on B's value at time t +1.

2. From plant to refined data

The first stage for producing a gene interaction model involves growing the Arabidopsis thaliana plants in a controlled environment. The experiments that generate the gene transcript

abundance data require thousands of plants. These plants should be as genetically similar as possible. To accomplish this, an initial stand of plants is grown in the laboratory. After the plants mature, all visually dissimilar plants are removed from the population. Then, the plants cross pollinates to create the next generation. This process is repeated until obtaining the third plant generation.

Using this third generation of plants, the *Arabidopsis thaliana* is stimulated by a plant hormone at time 0, and then gene transcript abundance measurements are collected at times 0, 0.5, 1, 2, 4, 8, 12 and 24 hours. At each of the 8 time points, some of the plants are harvested and analyzed. The Affymetrix technology is used to assess the gene transcript abundance measurements for the studied genes.

Each experiment is repeated three times, leading to three sets of data for each experiment. The three replicates should be similar, but certainly not identical.

The transcript abundance data consists of either 1, 246 or 449 transcripts depending on the specific experiment. The first step in data refinement will remove genes that have incomplete Affymetrix measurements. Next, data with too large p-value measurements is removed: these correspond to measurements that the Affymetrix technology reports as unreliable.

From the remaining genes, the biologists and biochemists select subsets of genes, many of which have the same functional relationship. For the present paper, this ultimately results in three sets of gene transcript abundance data known as the IAA12, ACC26 and IAA37 data sets.

Next, these transcript abundance data sets are further culled based on their numerical properties (Lewis et al., 2013). Then, these reduced genes are clustered into classes that reflect classes of similar gene stimulation or repression across the 8 time points. Finally, representatives of the equivalence classes are chosen for the final data sets IAA12, ACC26 and IAA37. There are twelve (12) and thirty seven (37) genes represented in the IAA12 and IAA37 sets, respectively. All of these genes have been stimulated with the plant hormone auxin (IAA). In ACC26, there are twenty six (26) genes, all of which have been stimulated by the plant hormone ethylene (ACC) (Harkey et al., 2018). Each of these data sets has three replicates that were incorporated in a hierarchical manner as detailed in Patton et al. (2014).

3. From data sets to gene interaction models

Each data set consists of three replicates, r1, r2, r3, for n gene's transcript abundance measurements. The goal is to produce a directed graph, or network, with vertices representing the genes, where each edge, gi ✳ gj, is labeled with the probability of a next-state relationship between gene gi and gene gj. There are a number of steps required to achieve this goal. A mathematical model is needed to represent a set of possible next-state relationships between the n genes.

This mathematical model represents one possible set of next-state relationships between pairs of genes. Subsequently, a statistical development is required to provide an optimum way to compare two of the mathematical models, i.e., which of the two models is more likely given the three sets of observations. Lastly, an algorithmic mechanism for searching through the mathematical models is desired: one that is guided by the relative posterior probabilities.

A directed acyclic graph (DAG) provides the structure to model a possible next-state relationship between the n genes. Reflexive and circular relations are not supported by DAGs. The directed edges of the DAGs are not labeled. The number of DAGs (Moon, 1970) is given by

$$f_{n,k} = \binom{n}{k} \sum_{i=0}^{k} \left(-\frac{1}{2}\right)^{i} (k+i)i! \binom{k}{i} \binom{n-k}{i} n^{n-k-i-1}$$

where n and k are the numbers of vertices and components. Clearly, any search over DAGs will require substantial sophistication.

The Norris-Patton likelihood (NPL) of a DAG (Patton, 2012; Norris et al., 2015; Patton et al., 2012, 2013, 2014), shown in Equation 1, was specifically developed to compute the likelihood that replicates r1, r2, r3 are described by a DAG D, NPL (r1, r2, r3 | D). The DAG has j genes with at least one parent, and w genes in the data. ixn is the time-course data for the parents of child n from replicate i. n is the concatenated average time-course data over all replicates from each parent of child n. iyn is the time-course data for child n in replicate i.

$$NLP(data|DAG) = (2\pi)^{-rtj/2} 2^{j(rt+1)/2} g^{-rk/2} 2^{-j/2} \Gamma(1/2)^{-j} \Gamma[(rt+1)/2]^{j} \quad (1)$$

$$\times \prod_{n=1}^{j} \frac{|\bar{x}_n^T \bar{x}_n|^{r/2}}{\prod_{i=1}^{r} |{}_i x_n^T {}_i x_n + \bar{x}_n^T \bar{x}_n|^{1/2}}$$

$$\times \left[1 + \sum_{i=1}^{r} \left({}_i y_n^T {}_i y_n - ({}_i x_n^T {}_i y_n)^T [({}_i x_n^T {}_i x_n + \bar{x}_n^T \bar{x}_n)^{-1}]^T {}_i x_n^T {}_i y_n\right)\right]^{-(rt+1)/2}$$

$$\times (2\pi)^{-rt(w-j)/2} exp^{-r(t-1)(w-j)/2}$$

Throughout this paper, we assume uniform priors on the DAGs so that a DAG's likelihood is its relative posterior probability. Cotemporal and next-state versions of the NPL have been developed, in both hierarchical and independent situations. Given two DAGs, D1 and D2, the NPL provides the ability to say the degree to which D1 is better than D2. Replicates r1, r2 and r3 are hierarchically incorporated to obtain their NPL.

Two algorithmic DAG search methodologies have been developed to produce the final next-state gene interaction models. The first one is based on a Metropolis-Hastings (MH) algorithm and the second one is a specialized genetic algorithm (BCHC). For both of these approaches, the chosen algorithm samples the DAG space guided by the NPL. The execution time complexity of the MH algorithm severely restricts the problem size, whereas the BCHC scales work reasonably well with the problem size. For either MH or BCHC, unique DAGs are collected across the entire executions of the algorithm. From all these unique DAGs, the final next-state gene interaction model is created using the classical Bayesian model averaging under equal DAG priors (Hoeting, Madigan, Raftery, & Volinsky, 1999). Specifically, the posterior probability of a directed edge e, M(e) in the model is computed.

$$M(e) = \frac{\sum_{d \in AL} \chi_d(e) L(d_1, d_2, d_3|d)}{\sum_{d \in AL} L(d_1, d_2, d_3|d)} \qquad (2)$$

where χd (e) = 1 if and only if e is a directed edge in the DAG d; otherwise χd (e) = 0. Even though DAGs do not allow cycles, it is certainly possible for cycles, but not loops, to appear in the final next-state gene interaction model.

The Metropolis-Hastings (MH) approach is a search governed by the decision process shown in Algorithm 1. After a suitable initialization, the MH approach guided the exploration of the DAG space for 500,000,000 steps in each of 10 independent and parallel executions (John et al., 2011; Norris et al., 2015). The 200 DAGs with highest likelihood were collected across these steps, and a final gene interaction model was produced using Equation 2.

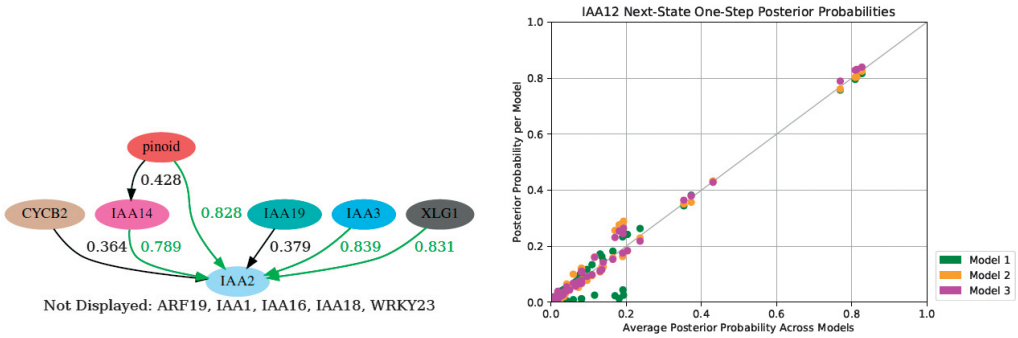
The BCHC modeling algorithm is a specialized genetic algorithm specifically designed to handle a population of DAGs (LaPointe, 2017; LaPointe et al., submitted). Each BCHC population is a set of DAGs. At the ith step, the BCHC uses the current population of 200 DAGs to produce the next population of 200 DAGs. It performs this using the specially adapted genetical algorithm operators of selection, crossover, mutation, and repair. The selection operator pairs the 200 DAGs to be parents, the crossover allows dissimilar parents to exchange (genetic) information, and the mutation operator is applied to the entire population when the population has essentially become stagnant. Unfortunately, the crossover of two DAGs and the mutation of a DAG can result in a directed graph containing a cycle. A repair operator is required to convert a directed graph containing a cycle into a DAG. Every genetic algorithm has many parameters, for example, the population size (200) and the total number of generations (250). For all the executions in the BCHC in this paper, the BCHC parameters were fixed.

Algorithm 1 The decision process of the Metropolis-Hastings algorithm, searching for best models and high probability edges using Norris-Patton likelihood. The function *random*() returns a random value between 0 and 1, uniformly.

1: Generate *New* DAG from the immediate neighbors of *Current* DAG
{If *New* is an improvement over *Current* then unconditionally accept *New*, else probabilistically accept *New*}
2: **if** $NPL(data \mid New) > NPL(data \mid Current)$ **then**
3: $Current \leftarrow New$
4: **else if** random() $< \frac{NPL(data|New)}{NPL(data|Current)}$ **then**
5: $Current \leftarrow New$
6: **end if**

The MH algorithm was implemented in MATLAB, and the BCHC algorithm in Python. For both programs, in order to minimize numerical errors, most computations involved the likelihood logarithm. The implementation of both the MH and BCHC algorithms involved distributed computing.

The 12 IAA12 genes were stimulated by auxin (IAA) and were chosen from the 1, 246 Arabidopsis thaliana genes that respond to IAA treatment (Lewis et al., 2013). For IAA12, both MH and BCHC next-state gene interaction models are produced. The BCHC IAA12 next-state gene interaction model is shown in Figure 1(a). Figure 1(b) shows that the BCHC algorithm is very consistent across multiple runs of the IAA12 data. The MH next-state gene interaction model is presented in Norris et al. (2015), Table 1, column H1. In this instance, there was not much agreement between the MH- and BCHC-based models. However, in numerous simulation studies, both the MH and BCHC next-state models do closely agree with the respective simulated networks (LaPointe, 2017; Norris et al., 2015).



(a) IAA12 gene interaction model

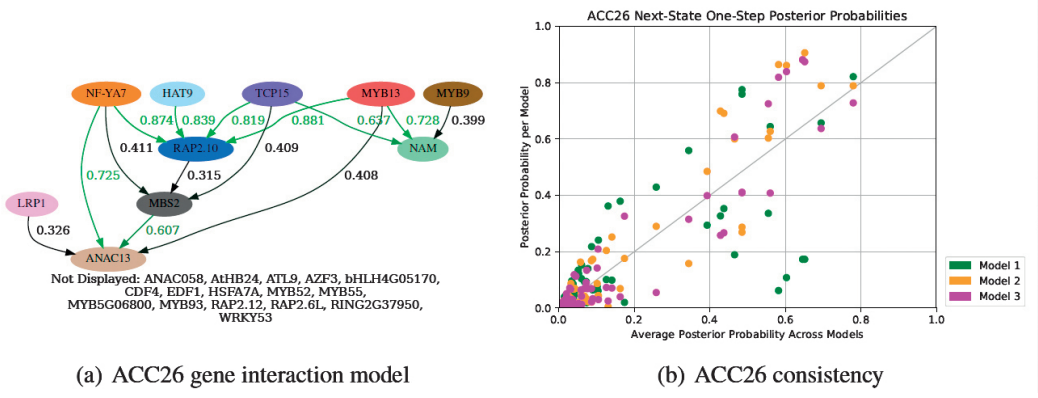
(b) IAA12 consistency

Note. A next-state gene interaction model and consistency plot across three interaction models for the IAA12 data set. Only directed edges with posterior probabilities of at least 0.35 are shown. Each directed edge, gi → gj, is labeled with the posterior probability of gi influencing gj. The three models in the consistency plot are from three executions of the modeling algorithm on the IAA12 data.

Figure 1. A next-state gene interaction model and consistency plot for the IAA12 data set
Elaborated by the authors

The MH algorithm has not been applied to any data set with more than 12 genes. The execution time required for MH to complete is prohibitive for moderately more than 12 genes. In fact, this restriction is one of the main motivations for the development of the BCHC algorithm.

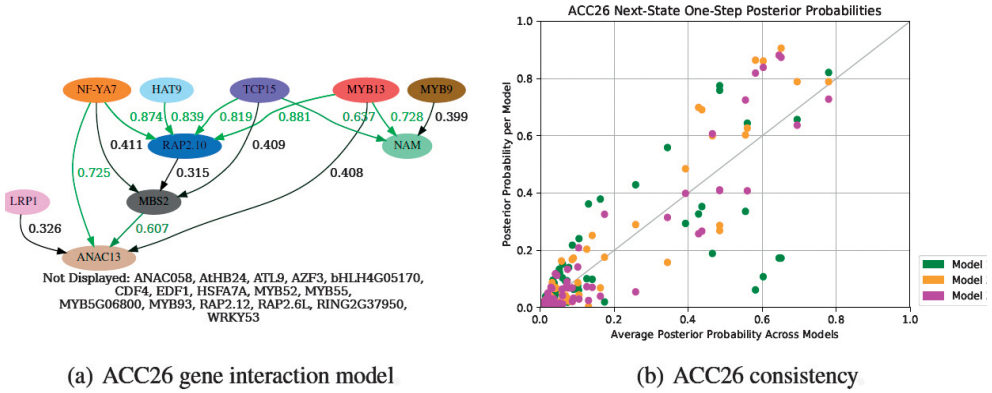
The ACC26 data set contains information about 26 genes from the 449 that responded to treatment with the ethylene precursor ACC. For these 26 genes, specific forbidden gene interactions well known from the biological literature were incorporated into the BCHC model algorithm (O'Malley et al., 2016). The BCHC gene interaction model is shown in Figure 2(a), and the indication of the consistency of similar models is shown in Figure 2(b). Comparing Figures 1(b) and 2(b), as the number of genes increased from 12 to 26, the overall consistency, though still good, diminished.



(a) ACC26 gene interaction model (b) ACC26 consistency

Figure 2. A next-state gene interaction model and consistency plot for the ACC26 data set
Elaborated by the authors

The 37 genes in the IAA37 data set were identified as IAA-dependent transcriptional changes dependent on auxin response factor19, ARF19. The chosen transcripts are in one of two functional groups: transcription factors (TF) or cell wall (CW) remodelers. It is known that an ARF19 gene can never be a child, a TF gene can only be the child of another TF gene, and a CW gene cannot be a parent and can only be a child of a TF gene. These give rise to another set of forbidden relationships that have been incorporated into the BCHC algorithm. Figure 3 shows both the IAA37 gene interaction model and the consistency information across three BCHC executions on the IAA37 data.



(a) ACC26 gene interaction model (b) ACC26 consistency

Figure 3. A next-state gene interaction model and consistency plot for IAA37 data set
Elaborated by the authors

4. Conclusions

The BCHC algorithm, guided by the next-state likelihood, produces consistent gene interaction models. The number of genes increases the execution time of the BCHC algorithm scales reasonably, in fact, linearly. Since the rigorous relative posterior probabilities of each visited DAG is known, the BCHC algorithm is a specialized genetic algorithm which aggressively searches the DAG space for DAGs with high likelihood. The gene interaction model should be a reasonably good estimate of the underlying biochemical relationships. Laboratory testing of the proposed directed edges suggested by these models is the next important step in this interdisciplinary collaborative journey.

Clearly, as the number of genes increases, the BCHC parameters should be adjusted. In particular, two parameters that should be adjusted as a function of the number of genes are the size of each population and the total number of generations. The increased variance of the consistency plots, as a function of the number of genes, is certainly partially caused by the fixed BCHC parameters, so future works should include adapting parameters for different situations.

References

Allen, E. E., Fetrow, J. S., Daniel, L. W., Thomas, S. J., & John, D. J. (2006, January). Algebraic dependency models of protein signal transduction networks from time-series data. *Journal of Theoretical Biology, 238*(2), 317-330. DOI: 10.1016/j.jtbi.2005.05.010

Cao, J., Qi, X., & Zhao, H. (2012). Modeling gene regulation networks using ordinary differential equations. In *Next generation microarray bioinformatics, 802*, 185-197. Springer. DOI:10.1007/978-1-61779-400-1_12

De la Fuente, A., Bing, N., Hoeschele, I., & Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics, 20*(18), 3565-3574.

Harkey, A. F., Watkins, J. M., Olex, A. L., DiNapoli, K. T., Lewis, D. R., Fetrow, J. S., .Muday, G. K. (2018). Identification of transcriptional and receptor networks that control root responses to ethylene. *Plant Physiology, 176*(3), 2095-2118. Retrieved from http://www.plantphysiol.org/content/176/3/2095. DOI: 10.1104/pp.17.00907

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and E.I. George, and a rejoinder by the authors). *Statistical Science, 14*(4), 382-417.

John, D. J., Fetrow, J. S., & Norris, J. L. (2011, September/October). Continuous cotemporal probabilistic modeling of systems biology networks from sparse data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8*(5), 1208-1222. DOI:10.1109/TCBB.2010.95

Krämer, N., Schäfer, J., & Boulesteix, A.-L. (2009). Regularized estimation of large-scale gene association networks using graphical Gaussian models. *BMC Bioinformatics, 10*(384). DOI:10.1186/1471-2105-10-384

LaPointe, B. A. (2017). *Arabidopsis thaliana gene interaction exploration with CHC genetic algorithm (Unpublished master's thesis)*. Wake Forest University, Department of Computer Science.

LaPointe, B. A., John, D. J., Norris, J. L., Harkey, A. F., Muhlemann, J. K., & Muday, G. K. (submitted). *A specialized genetic algorithm to model cotemporal hierarchical Arabidopsis thaliana gene interactions.*

Laubenbacher, R., & Stigler, B. (2004, August). A computational algebra approach to the reverse engineering of gene regulatory networks. *Journal of Theoretical Biology, 229*(4), 523-537. DOI: 10.1016/j. jbi2004.04.037

Lewis, D. R., Olex, A. L., Lundy, S. R., Turkett, W. H., Fetrow, J. S., & Muday, G. K. (2013, September). A kinetic analysis of the Auxin transcriptome reveals cell wall remodeling proteins that modulate lateral root development in Arabidopsis. *The Plant Cell, 25*, 3329-3346. DOI:10.1105/tpc.113.114868

Li, H., & Gai, J. (2008). Gradient directed regularization for sparse Gaussian, concentration graphs with applications to inference of genetic networks. *Biostatistics, 7*(2), 302-317.

Liang, J., & Han, J. (2012). Stochastic Boolean networks: an efficient approach to modeling gene regulatory networks. *BMC Systems Biology, 6*(113), 1-20. Retrieved from http://www.biomedcentral.com/1752-0509/6/113

Moon, J. W. (1970). Counting labelled trees, Canadian mathematical monographs, n.º 1. In *Canadian Mathematical Congress, Montreal, Quebec*.

Norris, J. L., Patton, K. L., Huang, S., John, D. J., & Muday, G. K. (2015, April). First and second order Markov posterior probabilities on multiple time-course data sets. In *SoutheastCon 2015* (pp. 1-8). Norfolk, Virginia: IEEE. DOI: 10.1109/SECON.2015.7132880

O'Malley, R. C., Huang, S. S. C., Song, L., Lewsey, M. G., Bartlett, A., Nery, J. R., ... & Ecker, J. R. (2016). Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell*, *165*(5), 1280-1292.

Patton, K. L. (2012). *Bayesian interaction and association networks from multiple replicates of sparse time-course data* (Doctoral dissertation, Wake Forest University).

Patton, K. L., John, D. J., & Norris, J. L. (2012, June). Bayesian probabilistic network modeling from multiple independent replicates. *BMC Bioinformatics, 13*(Supplement 9), 1-13.

Patton, K. L., John, D. J., Norris, J. L., Lewis, D., & Muday, G. (2013). Hierarchical Bayesian system network modeling of multiple related replicates. *BMC Bioinformatics, 7*, 803-812.

Patton, K. L., John, D. J., Norris, J. L., Lewis, D. R., & Muday, G. K. (2014, March/April). Hierarchical probabilistic interaction modeling for multiple gene expression replicates. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 11*(2), 336-346. DOI: 10.1109/TCBB.2014.2299804

Stigler, B. (2007). Polynomial dynamical systems in system biology. *2006 AMS Proceedings of Symposia in Applied Mathematics, 64*, 59-84.

Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., ... & Zitzler, E. (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome biology*, *5*(11), R92. DOI: 10.1186/gb-2004-5-11-r92