Simple Bayesian Gene Network Learning in Populus Draught Transcriptome Data

Amir Almasi Zadeh Yaghuti
Nanjing Forestry University
Ali Movahedi (ali_movahedi@njfu.edu.cn)
Nanjing Forestry University https://orcid.org/0000-0001-5062-504X
Hui Wei
Nanjing Forestry University
Weibo Sun
Nanjing Forestry University
Mohaddeseh Mousavi
Nanjing Forestry University
Qiang Zhuge
Nanjing Forestry University

Research Article

Keywords: Gene network, Populus, microarray data, NCBI

Posted Date: March 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-287640/v1

Simple Bayesian Gene Network Learning in *Populus* Draught Transcriptome Data

Amir Almasi Zadeh Yaghuti [1,†], Ali Movahedi [1,†], Hui Wei [1], Weibo Sun [1], Mohaddeseh Mousavi [1], and Qiang Zhuge [1,*]

[1] Co-Innovation Center for Sustainable Forestry in Southern China, Key Laboratory of Forest Genetics & Biotechnology, Ministry of Education, College of Biology and the Environment, Nanjing Forestry University, Nanjing 210037, China; amir_20364@yahoo.com (AAZY); ali_movahedi@njfu.edu.cn (A.M.); 15850682752@163.com (HW); czswb@njfu.edu.cn (WS); m.m2132@yahoo.com (MM)

* Correspondence: qzhuge@njfu.edu.cn; Fax: +86-25-8542-8701

† These authors are contributed equally to the first author.

Abstract

Populus is not only important for wood-based products, such as paper and timber, but also for metabolic-based production, for instance, bioethanol and biofuels. Constructing a sensibly functional gene interaction network is highly appealing to better understand system-level biological processes governing various *Populus* traits. Bayesian network learning provides an elegant and compact statistical approach for modeling causal gene-gene relationships in microarray data. Therefore, it could come with the illumination of functional molecular playing in Biology Systems. In this study, different forms of gene Bayesian networks were learned on *Populus* cellular transcriptome data. We addressed that Markov blankets, separating genes external to a regulatory Bayesian network from its internal genes, would likely be emerging at every possible gene regulatory Bayesian network level. The results have also shown that *PtpAffx.1257.4.S1_a_at,1.0* hypothetical protein is the most important in its possible regulatory program. This paper illustrates that the gene network regulatory inference is possible to encapsulate within a single BN model. Therefore, such a BN model can serve as a promising training tool for *Populus* gene expression data to better prepare future experimental scenarios.

Keywords: Gene network; *Populus*; microarray data; NCBI

Introduction

The development of the theory of causal modeling roots back to the 1950s, which has left some scientific community controversies. The core of these debates and controversies is the Markov condition/ assumption, an assumption made in the Bayesian probability theory. Almost three decades later, causal modeling caught up with many researchers' attention, devoted to developing the theory of Bayesian networks(Gillies and Sudbury 2013). BN has been used to model chlordecone bioaccumulation in plants (Liber, Cornet, et al. 2020) to discover the best regulators of drought response (Lahiri, Venkatasubramani, et al. 2019) and to infer gene regulatory networks (Vignes, Vandel, et al. 2011). A Bayesian gene network consists of a digraph, which connecting regulatory genes to their targets, and elegantly encodes conditional independence between genes. Technically, BN is a combination of Bayesian theorem with the directed acyclic graph (DAG). The DAG decomposes the joint probability distribution. Given the DAG, the joint probability distribution of the expression of 6 nodes/genes factorizes as follows: $P(A,S,E,O,R,T) = P(A)P(S)P(E \mid A,S)P(O \mid E)P(R \mid E)P(T \mid O,R)$

In general terms, this can be written as: $p(x) = \prod_{i=1}^{n} p(x_i \mid parents(x_i))$. In a BN, every gene must have a conditional probability table (CPT). For each gene, CPT indicates all the possible combinations of values of the parent genes. Each possible combination is called an Instantiation of the parent set. A BN structure is usually obtained from the data using score-based or constraint-based approaches (Koller and Friedman, 2009). Score-based algorithms maximize the BN likelihood, using Markov chain Monte Carlo (MCMC); to search the space of network structures, it operates on edge additions, deletions, or inversions. Score-based algorithms have been reported to fit well on simulated genetics and genomics data (Zhu, Wiener, et al. 2007) (Tasaki, Sauerwine, et al. 2015). However, we initially learned the network's undirected skeleton using repeated conditional independence tests in constraint-based approaches. Then by resolving directional constraints, each edge direction is assigned (v-structures and acyclicity) to the skeleton. Genetic mapping of multiple complex traits has been done(Scutari, Howell, et al. 2014). Due to its complexity, Bayesian gene network inference is feasible for systems of at most a few hundred genes or variables through conventional algorithms(Beckmann, Lin, et al. 2018, Wang, Audenaert, et al. 2019). The topology of the network of BN encodes conditional independence assertions. Therefore, we could logically find a set of possible gene regulators out of entire gene expression data. In this study, we try to find those genes in the *Populus* genome.

Drought is pivotal abiotic stress that affects plant development and poplar productivity (Hamanishi, Barchet, et al. 2015). Therefore, poplar breeding requires an understanding of the underpinning molecular regulatory machines controlling poplar resistance to drought stress. Molecular studies have revealed drought stress might take place in different plant tissues. For example, many transcriptomic studies, in poplar and also other plant genomes, have addressed transcriptional changes in roots (Molina, Rotter, et al. 2008, Cohen, Bogeat-Triboulot, et al. 2010, Lorenz, Alba, et al. 2011, Stolf-Moreira, Lemos, et al. 2011, Dash, Yordanov, et al. 2018) due to drought stress over different genotypes (Cohen, Bogeat-Triboulot, et al. 2010, Stolf-Moreira, Lemos, et al. 2011, Cao, Jia, et al. 2014, Hamanishi, Barchet, et al. 2015, Jia, Li et al. 2016) or at different levels of drought stress (Cohen, Bogeat-Triboulot, et al. 2010, Stolf-Moreira, Lemos et al. 2011). Gene network analysis on poplar root transcriptome identified a hierarchical-like gene network structure in which the highest hierarchical level 2,934 genes were affected by9 super hubs (supergenes) (Hamanishi, Barchet, et al. 2015).

Here, we report microarray-based transcriptomic Bayesian gene network analysis of drought-resistant in black poplar genotypes measured in well-watered, moderate drought, severe drought, and post-drought re-watering conditions. Understanding the characteristics of poplar gene regulation of drought resistance, including various molecular interaction processes, shall elaborate functional knowledge of genes underlying this stress-induced phenomenon.

Materials and Methods

Data

Using the GEOquery package (Davis and Meltzer 2007), information directly from the GEO database with accession number GSE76322 was downloaded in this research. Initially, data dimensions were 61413×18. To run the Bayesian regulatory network, the probes with the highest variances were selected. Finally, the data dimension was reduced to 2210 × 18.

Program

We used several programs. The bnlearn package was used to infer the regulatory Bayesian network on our data. To determine the best network structure, the hill-climbing algorithm was used. The result of bnlearn was considered as a regulatory Bayesian network. The number of nodes, edges, Markov blanket size, neighborhood size, and the learned network's branching factor was calculated. In general, we assumed that the wider the Markov blanket size, the giant module network could be detected. With the bnlearn package help, the adjacency matrix with Cytoscape-based aMatReader software was imported to Cytoscape. Using the Network

Analyzer plugin, a comprehensive set of topological parameters such as number of nodes, edges, network diameter, radius and clustering coefficient, neighborhood connection, shortest path length, number of familiar neighbors, the degree distribution, the centripetal proximity parameter, etc. were calculated for our network.

Results and Discussion

In general, for high-dimensional genetics and genomics data, BN learning is challenging since the number of expected networks scales up exponentially with the number of genes. The computational cost of conventional BN inference could be a burden prohibitive (Wang, Audenaert, et al. 2019). Table 1 gives the general parameters of the observed probe-based Bayesian regulatory network. The number of probes (nodes), as can be seen in table 1, is 2210, which is higher than the number of directed edges (1000) that is logical in this context. This would indicate that some orphan probes in the network or some nodes are associated averagely high with other nodes.

One of the main parameters of detected BN is the Markov blanket (0.45).

Table 1: Estimation of structural Bayesian network parameters with Hill Climbing algorithm (MB: Markov Blanket, NS: Neighborhood Size, BF: Branching Factor, HC: Hill Climbing)

Parameters	values
No. of Nodes	2210
No. of Arcs (Edges)	1000
Undirected Edges	0
Directed Arcs	1000
MB	0.92
NS	0.90
BF	0.45
Penalization coefficient	1.445186
No. of Tests	4647736

To give a general biologically motived expression of Markov blanket (MB), we could say that MB implicitly and explicitly is a minimum set of genes that would block all back-door paths that could be discovered from learned BN. The MB of a gene includes its parents, children, and co-parents. Parents in the MB separate the impact of specific disturbance on the outcome from the rest of the gene network. Therefore, this concept can be used to select a smaller set of relevant genes in high-dimensional problems. The MB is proven to be highly effective for feature reduction in high-dimensional problems, sometimes reducing the number of variables a thousandfold without any loss of accuracy (Aliferis, Tsamardinos, et al. 2003, Shen, Li et al.

2008, Fu and Desmarais 2010, Tan and Liu 2013). We put forward this idea that the MB establishment is also instrumental in creating gene module networks. In other words, we could say that number of MBs and their sizes in BN may reflect the sense of modularity in the network. Parents of genes in the BN are the connections that would reflect causalities in the gene network. By identifying the MBs in different sizes over all genes in the gene network, we can quickly draw different gene sizes that cohesively function together, e.g., module network.

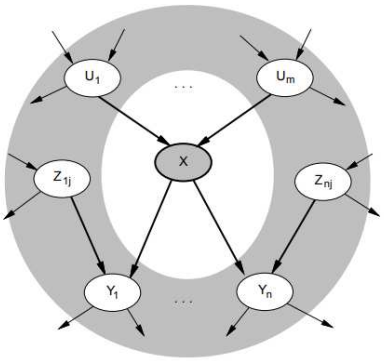

Figure 1: MB of a given postulated gene BN shows that MB of a node/probe/gene is the set containing the node/gene's parents, children, and co-parents.

Figure 1 illustrates seven genes in a postulated gene regulatory BN. These seven genes constitute a structure which is called MB. The U1, Um genes are parents of the X gene, and Y1 and Yn genes are Childs of the X gene in which the other two parents, e.g., Z1j and Znj genes they share paternity with the X gene. This structure demonstrates some cohesive acting of genes, which we call them up as gene modules. Finding these structures in BN would be biologically appealing. In our study, it was turned out that many MB are cohesively overlapped. Therefore, a general picture of gene modules was not seen. We are expected to see some grape-like structures, i.e., gene modules due to MB; such structures were not visually turned up on the learned BN. The main finding of this study, e.g., *PtpAffx.1257.4.S1_a_at,1.0,* hardly bordered as a specific set of probes as gene module. In other words, we could say that this probe (and the gene it belongs to) might be involved in many other MB as well. This may indicate the level of complication of doughtiness on the poplar genome. In other words, drought stress has caused many different biological pathways actions. In every gene regulatory BN network, we could find the total connectivity for each gene/probe, and this measure in BN can be split up into out-degree and in-degree. Table 2 and figure 2 address some topological aspects of learned BN. Putting this in a biological context, we could see that nodulizing BN has been

hardly achieved. One reason could be the fact that drought stress affected many different functional parts of the poplar genome.

However, the distribution of gene out-degree has been much higher than gene in-degree. This is biologically supported as out-degree probes/genes could be considered as a regulator, which patently indicates that a small number of genes had many out-degrees. Some of these high-out-degree genes are shown in table3. The results of this table can be used for further analysis, for example, utilizing suitable gene set enrichment analysis. A general feature of many gene networks is their nature of scale-free topologies addressing a gene degree distribution that can be shown with a function of power-law, e.g., $P(k) = Ck^{-\alpha}$, in which P(k) is the randomly selected gene having degree k (or k connections), α is the power-law exponent, and the constant C is a Riemann's zeta function e.g. $\sum_{n=1}^{\infty} \frac{1}{n^s}$ (s = complex variable and n = integer) normalizing the power law probability distribution, e.g. $\sum_{k=1}^{\infty} P(k) = 1$. In scale-free gene network topologies, most genes generally have relatively few interactions reflected as a lower degree. In contrast, a small number of genes, so-called 'hubs' genes, have a higher degree.

Table 2"simple parameters in network Analyzer.

Clustering coefficient	00.00
Number of nodes	2211
Connected components	1
Network diameter	1
Network radius	1
Shortest paths	2210(0%)
Characteristics path length	1
The average number of neighbors	1.999
Network density	0.0
Isolated nodes	426
number of self-loops	0
multi-edge node pairs	0
Analysis times (Sec)	0.874

Table 3: The CytoHubba plugin chose the first 30 genes in the MCC method.

PtpAffx.1257.4.S1_a_at,1.0	**Hypothetical protein /// hypothetical protein**
PtpAffx.1258.1.S1_s_at,1.0	**Hypothetical protein**
PtpAffx.1258.4.S1_s_at,1.0	**Aktin9**
PtpAffx.1259.1.S1_s_at,1.0	**Hypothetical protein**
PtpAffx.12595.1.A1_s_at,1.0	**Hypothetical protein**
PtpAffx.126167.1.S1_at,1.0	**Branched-chain amino acid aminotransferase-like protein**
PtpAffx.126235.1.S1_at,1.0	**Auxin-responsive family protein**
PtpAffx.12628.1.S1_at,1.0	**Metal handling. Binding, chelation, and storage**
PtpAffx.126599.1.A1_s_at,1.0	**Arginine decarboxylase**
PtpAffx.1263.1.A1_a_at,1.0	**Hypothetical protein**
PtpAffx.12675.3.A1_s_at,1.0	**Unknown protein; predicted by gene finder**
PtpAffx.12675.3.A1_a_at,1.0	**Hypothetical protein**
PtpAffx.12684.1.S1_at,1.0	**pyrophosphate-dependent 6-phosphofructose-1-kinase, putative**
PtpAffx.12676.1.S1_at,1.0	**MAC/Perforin domain-containing protein**
PtpAffx.1269.1.A1_x_at,1.0	**Putative thioredoxin**
PtpAffx.1269.1.A1_a_at,1.0	**Putative thioredoxin**
PtpAffx.12712.1.S1_at,1.0	**Hypothetical protein**
PtpAffx.127056.1.A1_at,1.0	**Signal transducer activity**
PtpAffx.12745.1.A1_at,1.0	**Unknown protein**
PtpAffx.1273.1.S1_at,1.0	**Copper ion binding, laccase activity,**
PtpAffx.127847.1.A1_at,1.0	**DNA binding protein**
PtpAffx.127644.1.A1_s_at,1.0	**Hypothetical protein**
PtpAffx.12800.1.S1_a_at,1.0	**Hypothetical protein**
PtpAffx.127878.1.A1_at,1.0	**Unknown protein**
PtpAffx.1286.2.S1_s_at,1.0	**S-adenosylmethionine decarboxylase**
PtpAffx.1286.1.S1_s_at,1.0	**S-adenosylmethionine decarboxylase, putative**
PtpAffx.12874.1.S1_at,1.0	**Hypothetical protein**
PtpAffx.1286.5.S1_s_at,1.0	**Hypothetical protein**
PtpAffx.128895.1.S1_at,1.0	**Lysine-ketoglutarate reductase/saccharopine dehydrogenase**

This may prohibit seeing clean-cut gene modules in studies similar to our study.

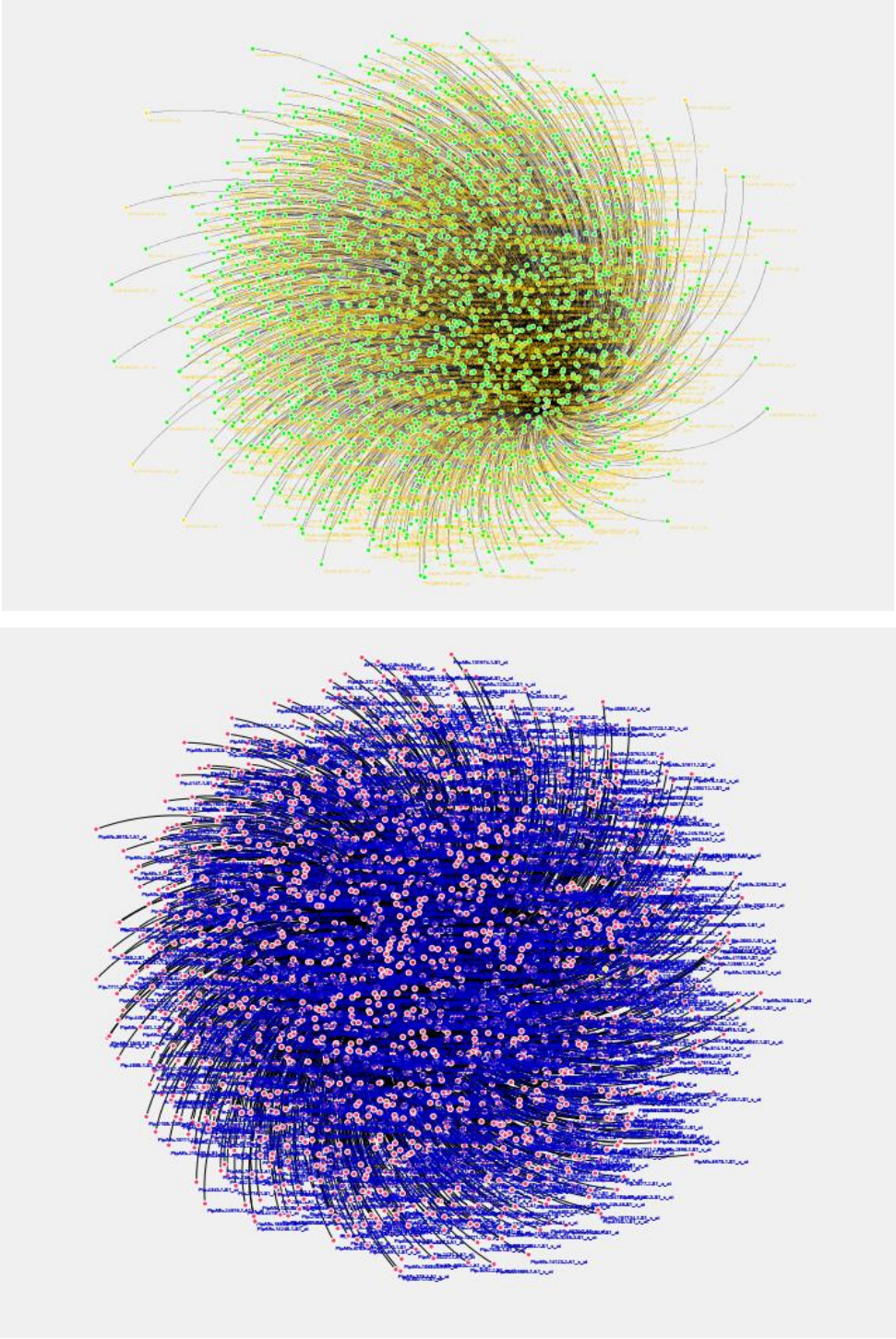

Figure 2: The snapshot of gene connectivity in learned BN in two graph forms.

To predict gene function in poplar gene expression data using DNA microarray datasets, private or publicly available databases are needed. Our method allows users to extract genes involved in biological processes, which could be valuable for understanding similar tree species mechanisms (Ogata, Suzuki, et al. 2009). The mode of genes in terms of being regulator or regulated is not reflected in this type of database. This could make the practical biological

application of such databases restrictive. We used public domain data in this study and to check up a new hypothesis in popups genome. This sort of data is being used to create a gene network (Cai, Li et al. 2014). In (Cai, Li et al. 2014), candidate *Populus* cell wall genes using a combination of a genome-wide *Populus* gene co-expression network (PGCN) and module detection and gene ontology (GO) enrichment analysis was performed to assign the functional category to these modules. However, our approach in this study was a probabilistic one instead of a Pearson-based gene correlation matrix. Our data did not find any topological properties of the network indicating scale-free and modular behavior. As it turned out, we were intensely relied on BN topological properties, e.g., Markov blanket, to address possible modules in our data. Figures 1 and 2 showed no sign of modules.

In the domain of gene expression high throughput data, probabilistic dependencies arise because genes are related in different ways (e.g., through common phenotype, logical connections, pathway connections, through (non-causal) physical laws, constrained of thermodynamic laws or structural boundary conditions). This may prohibit seeing clean-cut gene modules in studies like the current study. Figure 1 demonstrates that no sign of modules is trackable to pick one out of the gene regulatory network. Our method could likely help to identify and characterize cell wall-related genes in *Populus*. Using graph-based theory, (Zhang and Yin 2016) identified 14 probe-sets/genes related to the GO cellular component term of B plasmodesmata and many other genes for other tissues. However, the graph-based networks of the different tissues have shown different topological properties. They illustrate that genes in the root network were the most highly co-expressed. Whereas the leaf genes were the weakest co-expressed, and those in wood were in the intermediate. These topological network differences provide some unseen epigenetic mechanisms in different tissues. The genetic mechanism underlying the different topological properties among these tissues remains unknown and worthy of further investigation in future studies. In general, finding gene modules in *Populus* data has an almost long history (Grönlund, Bhalerao, et al. 2009, Liu, Ding et al. 2016, Han, An et al. 2020). In most of them, the co-expression analyses have been the cornerstone in finding modules of co-expressed genes; gene modules display distinct expression profiles. However, addressing hub genes in this kind of study is a difficult task. For example, figure 2 shows how genes are cohesively related, and therefore, it is quite challenging to pick up and address some genes which actively responded to drought.

255 Table 4. The identified DEC probe sets in leaf tissue involved in photosynthesis-related processes and
256 light reactions

Biological process	Probe sets	Gene model	Description	TAIR gene model	Alias
Photosynthesis	PtpAffx.71066.5.A1_s_at	Potri.010G089400	Photosystem I 20kD family protein	AT1G03130	PSAD-2
Photosynthesis	Ptp.5240.1.S1_s_at	Potri.008G151600	Photosystem I 20kD family protein	AT1G03130	PSAD-2
Photosynthesis	Ptp.2148.1.S1_at	Potri.001G081500	Photosystem I chain III family protein	AT1G31330	PSAF
Photosynthesis	PtpAffx.249.164.A1_s_at	Potri.002G055700	Photosystem II oxygen-evolving complex protein two precursor	AT1G06680	OE23, OEE2, PSBP-1, PSII-P
Photosynthesis	Ptp.1584.4.S1_x_at	Potri.005G239300	Chlorophyll a-b-binding protein 2	AT2G34430	LHB1B1, LHCB1
Photosynthesis	PtpAffx.689.1.S1_at	Potri.010G255500	Hypothetical protein	AT5G08410	FTRA2
Photosynthesis	Ptp.5386.1.S1_s_at	Potri.006G144000	Membrane protein	AT2G06520	PSBX
Photorespiration	PtpAffx.638.1.S1_s_at	Potri.010G045100	Alanine-2-oxoglutarate aminotransferase 1	AT1G70580	AOAT2, GGT2
Photorespiration	PtpAffx.638.1.S1_s_at	Potri.008G187400 Alanine-2-oxoglutarate aminotransferase 1	Photorespiration	PtpAffx.638.1.S1_s_at	Potri.008G187400 Alanine-2-oxoglutarate aminotransferase 1
Photorespiration	Ptp.1238.1.S1_at	Potri.009G081600	Malate dehydrogenase family protein	AT2G22780	PMDH1
Photosystem I	PtpAffx.7681.2.A1_at	Potri.009G065900	Hypothetical protein		
photosystem II	Ptp.6095.1.S1_at	Potri.006G088200	Hypothetical protein	AT5G02120	OHP
Light reaction	Ptp.8143.1.S1_s_at	Potri.011G099400	Cytochrome P450 family protein	AT3G14690	CYP72A15
Light reaction	Ptp.6095.1.S1_at	Potri.006G088200	Hypothetical protein	AT5G02120	OHP
Light reaction	Ptp.1584.4.S1_x_at	Potri.005G239300	Chlorophyll a-b-binding protein 2	AT2G34430	LHB1B1, LHCB1.4
Response to UV	Ptp.4373.1.S1_x_at	Potri.005G246100	Plastocyanin family protein	AT1G76100	PETE1

257

258

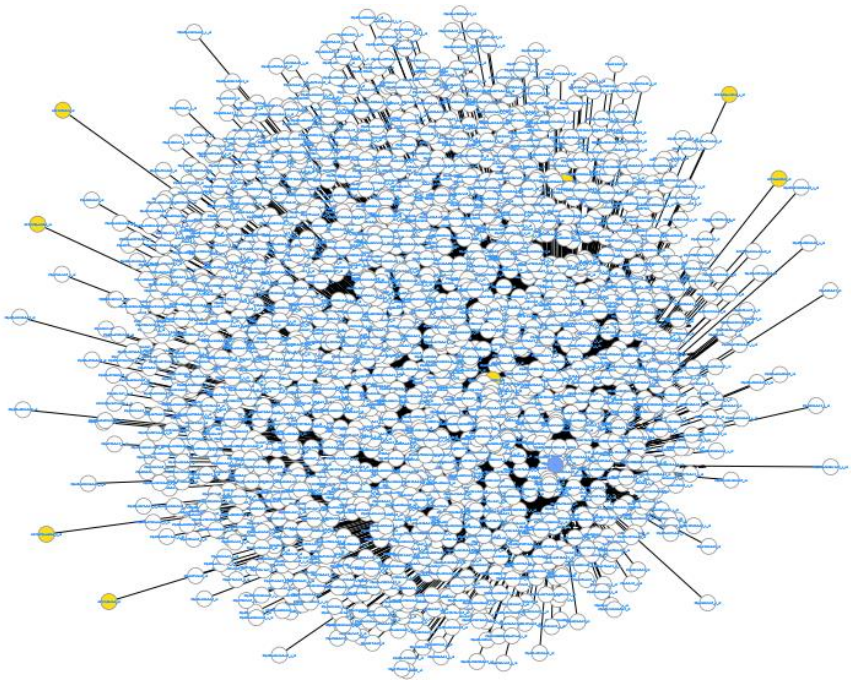

Figure 3: Highlighting top 10 genes with their neighbors using MCC algorithm

In the MCC algorithm, it is possible to highlight motivated biological molecules in the network. Some yellow circle nodes in figure 3 display the top ten most connected hub genes. As it can be seen, there is a close relationship between other genes and these top 10 genes, which in turn makes it challenging to report some general gene module structures in our data. However, in this study, the applied BN model should be regarded as a conjecture, which should be tested using more and better posed high dimensional data.

Poplar has emerged as an ideal model system for studying woody plants. To better understand the biological processes underlying various poplar traits, e.g., wood development, a comprehensive functional gene interaction network is highly needed. To more effectively screening down, we have used some highly connected gens in this study. Poplar Gene offers comprehensive functional interactions and extensive poplar gene functional annotations. Seminal research on the poplar genome shows that relatively small genome, quick growth trend, and simple genome clonal manipulation are the interesting biological features that have made poplar a long-lived forest tree model system (Liu, Ding, et al. 2016). Drought stress is the leading cause of plant loss worldwide, and drought is one of the essential environmental fluctuations that affect almost all plant species (Dash, Yordanov, et al. 2018). As a model, *Populus* provides an opportunity to study the stress response in a perennial tree growing as a commercial biomass product to produce carbon-neutral energy (Street, Skogström, et al., 2006).

In this study, as can be seen in table 3, many genes turned out to be highly connected ones: hub genes. Moreover, their annotations reveal that they do not have comprehensive biological support (as indicated by hypothetic term). Therefore, it was not easy to see their possible protein interaction in the STRING database. Protein-protein interaction networks play an essential role in understanding the system level of cellular processes. These networks can filter and evaluate functional genomic data and create an intuitive platform for annotating proteins' structural, functional, and evolutionary properties. In this study, we took the aktin 3 genes (Table 3) and fed them to the STRING database, and the grid of protein interaction was extracted. STRING is a protein-protein interactions database (the known and predicted one) and a retrieval search tool. The results of drawing interactive networks in the STRING database are shown in figure 4. The result of the protein network was architecturally different from those obtained from the Bayesian gene network. In other words, some inconsistency and discrepancy in the results of the network learned by STRING were observed compared to the learned BN network, which shows that post-transcriptional modification may play crucial roles in regulating gene expression in the *Populus* genome.

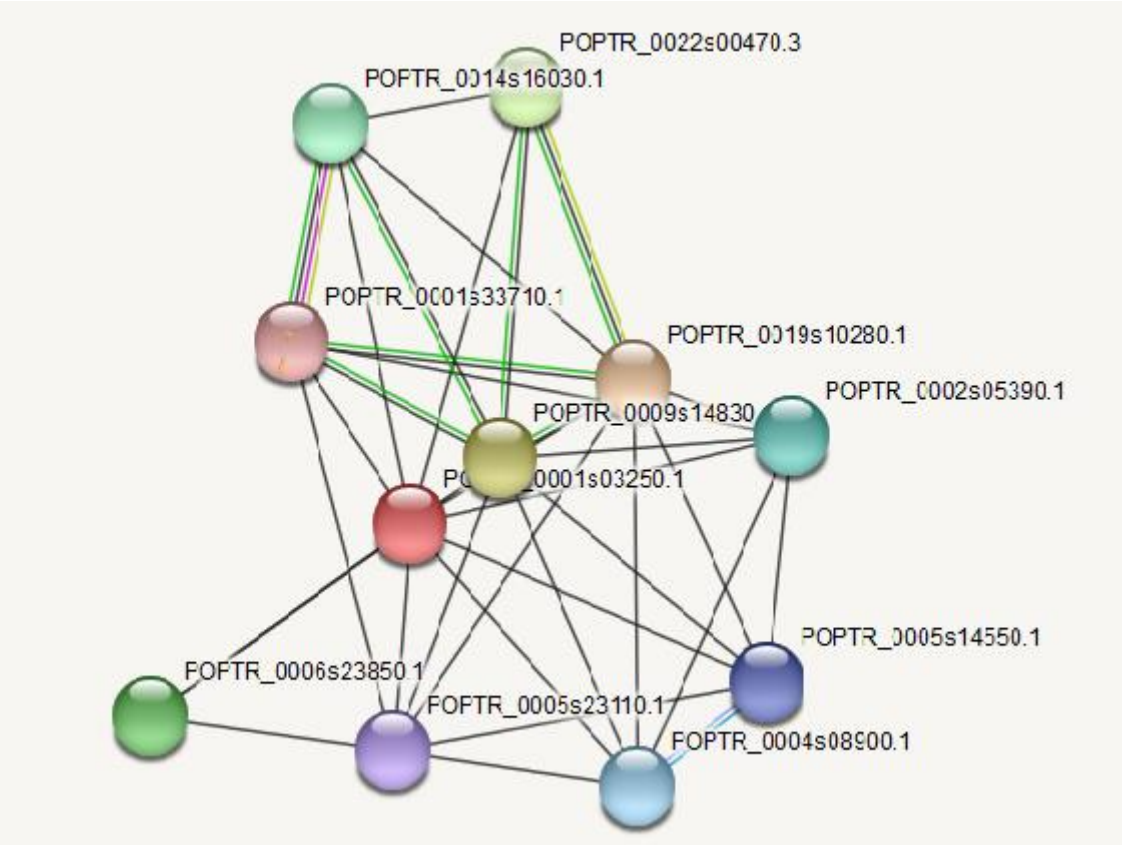

Figure 4: Network of protein-protein interactions by STRING database information.

In this study, we logically relied on *Populus* mRNA data; therefore, talking about the PPI network does not make that much sense. However, we believe that the whole body or tissue-specific protein-protein interactome (PPI) map could highlight drought tolerance in the *Populus*. Large-scale studies of PPIs accompanying the species-specific database support the interplay of cascading translational protein levels for drought-related stress adaptive mechanisms in *Populus*. Interestingly, many genes that became hub genes in this study were biologically involved in photosynthesis-related processes and light reactions. These are the key findings of this study. Most likely abiotic processes like drought act on the biological system level of *Populus*.

Boxplot is a visualization tool of microarray data. Each box stretches from the lower hinge - defined as the 25th percentile to the upper hinge - the 75th percentile- and the median is delineated as a line across the box. This figure has appeared in microarray experiment data at the top of other figures. However, we took some so-called downside-up approach, in a way that is shown at the bottom of the table to make it visually much more transparent than the treatments used in applied GSE76322, which did not impart much more diversity on the gene expression (figure 5). However, this sort of view should not mislead researchers, as in this study, many probes turned out to be differently expressed. We believe that BN has lots of potential in deciphering information on *Populus* microarray data, a matter that should be figured out in subsequent studies.

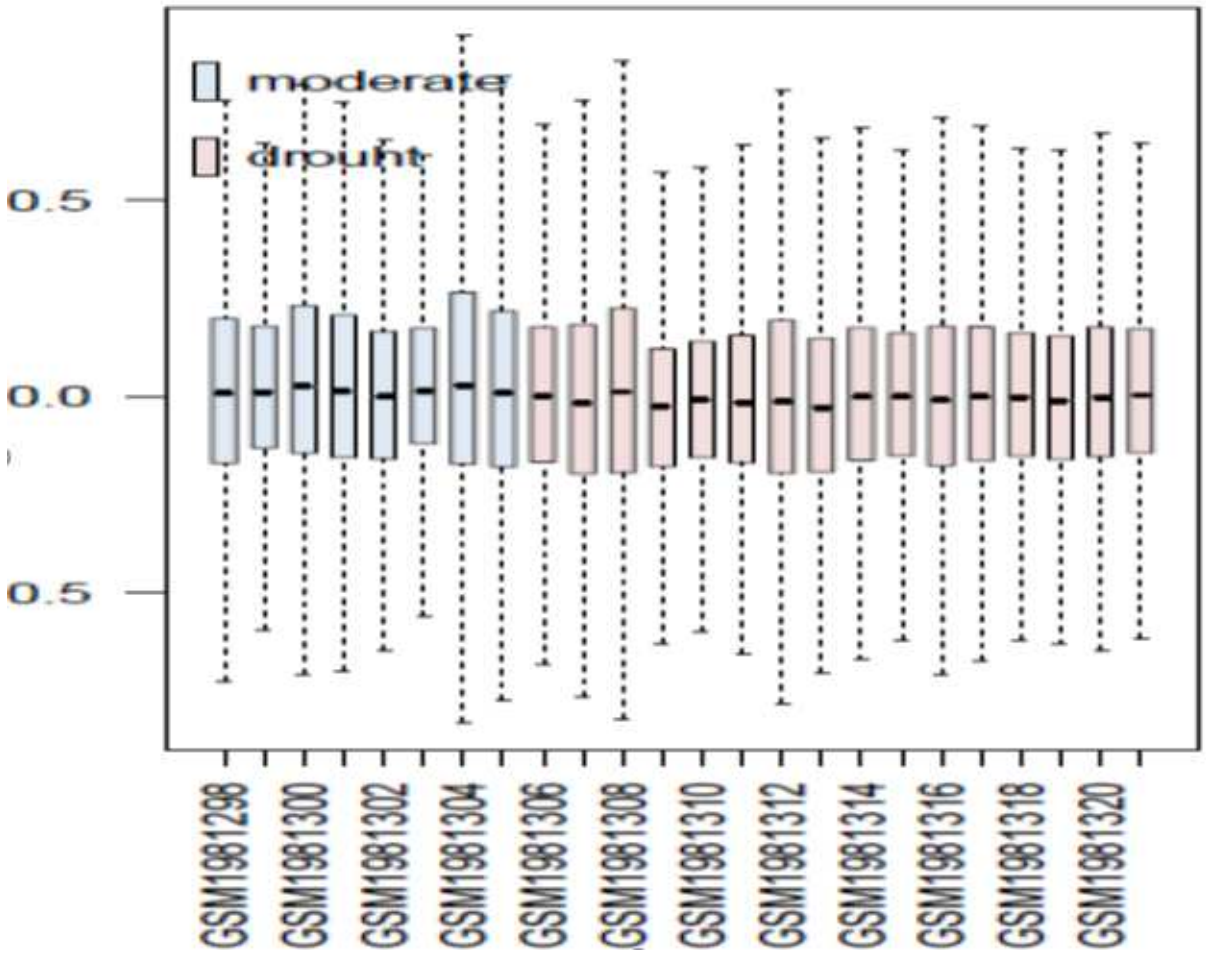

Figure 5: Boxplot of GSE764322 using in this study.

Conclusion

For the first time, a BN was learned out on *Populus* DNA microarray data. In this study, we identified a gene network in the *Populus* to highlight candidate genes used as regulator genes. We identified the top 30 hub genes, in which some of them had low if any, valid biological information. We noted that the MB in BN of gene regulation could be assumed to narrow down the whole gene network complexity. Statistically speaking, the existence of an MB means external genes are conditionally independent of internal genes and vice versa. This is biologically appealing. Despite extensive physiological and morphological descriptions of the *Populus* response to drought, little work has been done to explain the differences in gene levels and examine the similarity of the stress response between this perennial and the annual crop. Because poplar genome sequences and poplar microarrays are now available, a bridge can be made between quantitative trait locus mapping approaches, the candidate gene approach, and transcription.

Acknowledgment

This study was supported by the National Science Foundation of China (No. 31570650) and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Aliferis, C. F., et al. (2003). HITON: a novel Markov Blanket algorithm for optimal variable selection. AMIA annual symposium proceedings, American Medical Informatics Association.

Beckmann, N. D., et al. (2018). "Multiscale causal network models of Alzheimer's disease identify VGF as a key regulator of disease." bioRxiv: 458430.

Cai, B., et al. (2014). "Systematic identification of cell-wall related genes in *Populus* based on analysis of functional modules in co-expression network." PloS one 9(4): e95176.

Cao, X., et al. (2014). "Anatomical, physiological and transcriptional responses of two contrasting poplar genotypes to drought and re-watering." Physiol Plant 151(4): 480-494.

Cohen, D., et al. (2010). "Comparative transcriptomics of drought responses in *Populus*: a meta-analysis of genome-wide expression profiling in mature leaves and root apices across two genotypes." BMC genomics 11(1): 630.

Dash, M., et al. (2018). "Gene network analysis of poplar root transcriptome in response to drought stress identifies a PtaJAZ3PtaRAP2.6-centered hierarchical network." PLOS ONE 13(12): e0208560.

Davis, S. and P. S. Meltzer (2007). "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor." Bioinformatics 23(14): 1846-1847.

Fu, S. and M. C. Desmarais (2010). Markov blanket-based feature selection: a review of past decade. Proceedings of the world congress on engineering, Newswood Ltd.

Gillies, D. and A. Sudbury (2013). "Should causal models always be Markovian? The case of multi-causal forks in medicine." European Journal for Philosophy of Science 3(3): 275-308.

Grönlund, A., et al. (2009). "Modular gene expression in Poplar: a multilayer network approach." New Phytologist 181(2): 315-322.

Hamanishi, E. T., et al. (2015). "Poplar trees reconfigure the transcriptome and metabolome in response to drought in a genotype- and time-of-day-dependent manner." BMC genomics 16(1): 329.

Han, X., et al. (2020). "Comparative transcriptome analyses define genes and gene modules differing between two *Populus* genotypes with contrasting stem growth rates." Biotechnology for biofuels 13(1): 1-21.

Jia, J., et al. (2016). "Physiological and transcriptional regulation in poplar roots and leaves during acclimation to high temperature and drought." Physiol Plant 157(1): 38-53.

Lahiri, A., et al. (2019). "Bayesian modeling of plant drought resistance pathway." BMC Plant Biology 19(1): 96.

Liber, Y., et al. (2020). "A Bayesian network approach for the identification of relationships between drivers of chlordecone bioaccumulation in plants." Environmental Science and Pollution Research: 1-6.

Liu, Q., et al. (2016). "PoplarGene: poplar gene network and resource for mining functional information for genes from woody plants." Scientific Reports 6(1): 31356.

Lorenz, W. W., et al. (2011). "Microarray analysis and scale-free gene networks identify candidate regulators in drought-stressed roots of loblolly pine (P. taeda L.)." BMC genomics 12: 264.

Molina, C., et al. (2008). "SuperSAGE: the drought stress-responsive transcriptome of chickpea roots." BMC genomics 9(1): 553.

Ogata, Y., et al. (2009). "A database for poplar gene co-expression analysis for systematic understanding of biological processes, including stress responses." Journal of Wood Science 55(6): 395-400.

Scutari, M., et al. (2014). "Multiple quantitative trait analysis using Bayesian networks." Genetics 198(1): 129-137.

Shen, J., et al. (2008). Markov Blanket Feature Selection for Support Vector Machines. AAAI.

Stolf-Moreira, R., et al. (2011). "Transcriptional Profiles of Roots of Different Soybean Genotypes Subjected to Drought Stress." Plant Molecular Biology Reporter 29(1): 19-34.

Street, N. R., et al. (2006). "The genetics and genomics of the drought response in *Populus*." The Plant Journal 48(3): 321-341.

Tan, Y. and Z. Liu (2013). Feature selection and prediction with a Markov blanket structure learning algorithm. BMC bioinformatics, Springer.

Tasaki, S., et al. (2015). "Bayesian network reconstruction using systems genetics data: comparison of MCMC methods." Genetics 199(4): 973-989.

Vignes, M., et al. (2011). "Gene regulatory network reconstruction using Bayesian networks, the Dantzig Selector, the Lasso and their meta-analysis." PloS one 6(12): e29165.

Wang, L., et al. (2019). "High-dimensional Bayesian network inference from systems genetics data using genetic node ordering." Frontiers in genetics 10: 1196.

Zhang, H. and T. Yin (2016). "Identifying candidate genes for wood formation in poplar based on microarray network analysis and graph theory." Tree Genetics & Genomes 12(3): 61.

Zhu, J., et al. (2007). "Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations." PLoS Comput Biol 3(4): e69.

Figures

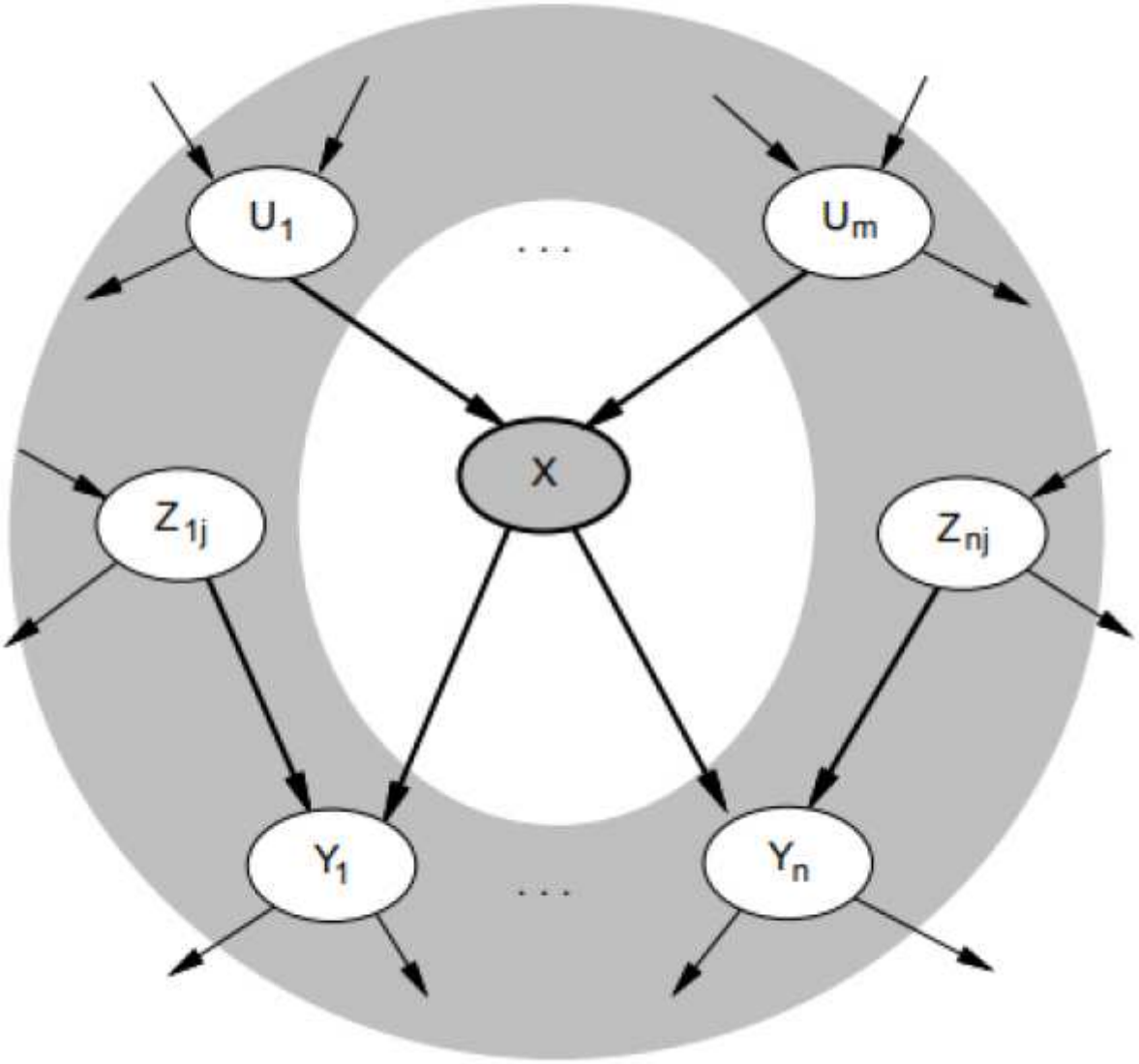

Figure 1

MB of a given postulated gene BN shows that MB of a node/probe/gene is the set containing the node/gene's parents, children, and co-parents.

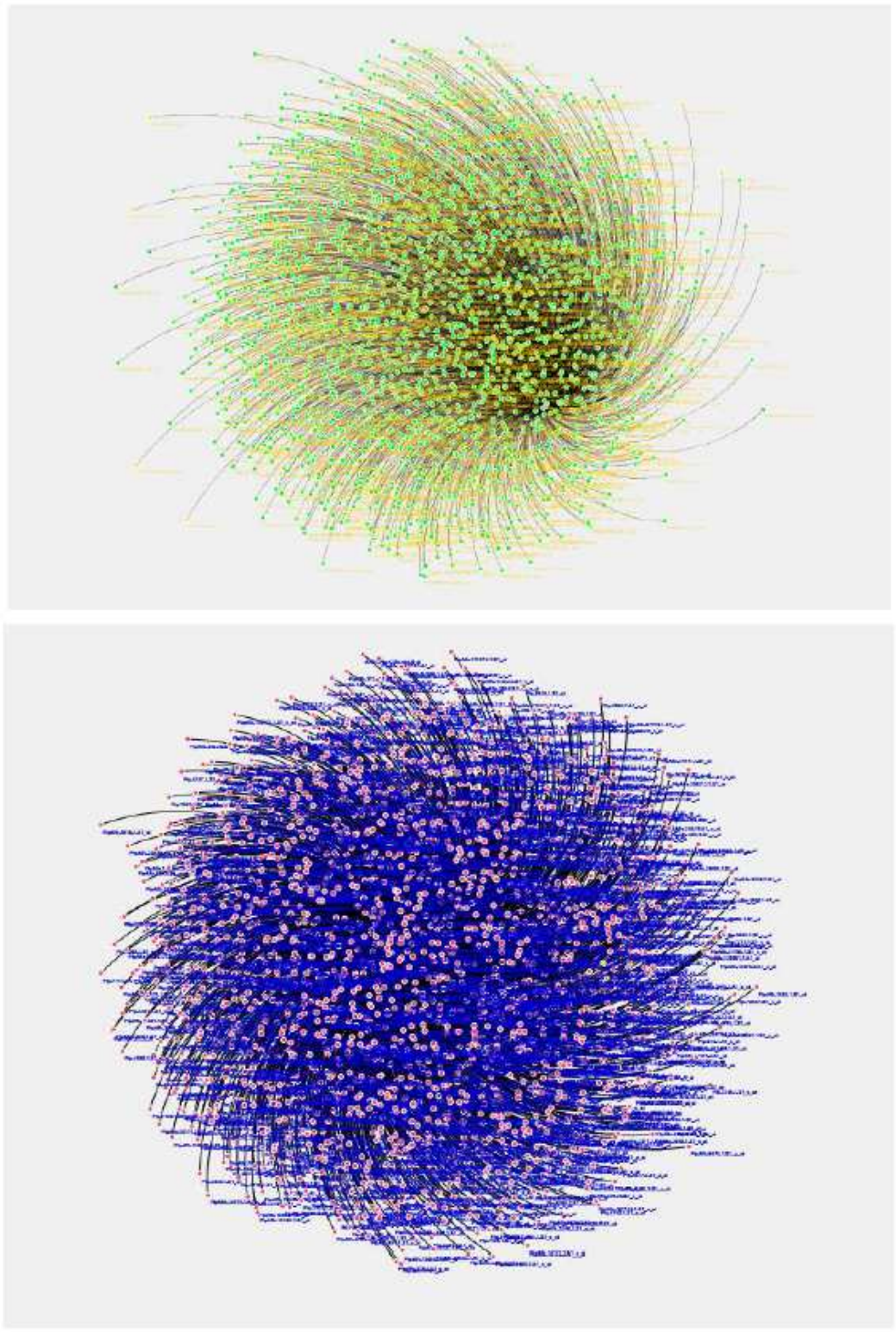

Figure 2

The snapshot of gene connectivity in learned BN in two graph forms.

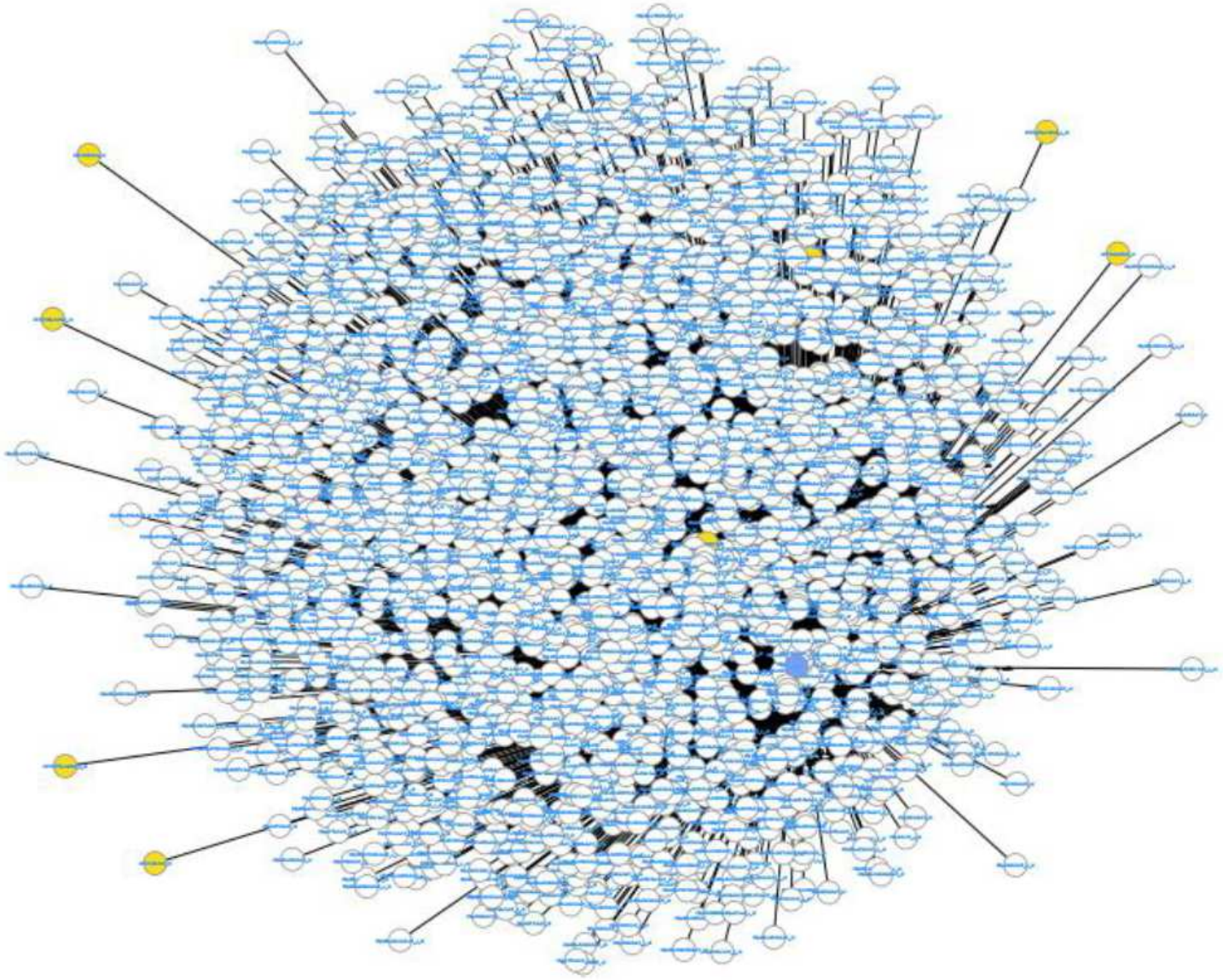

Figure 3

Highlighting top 10 genes with their neighbors using MCC algorithm

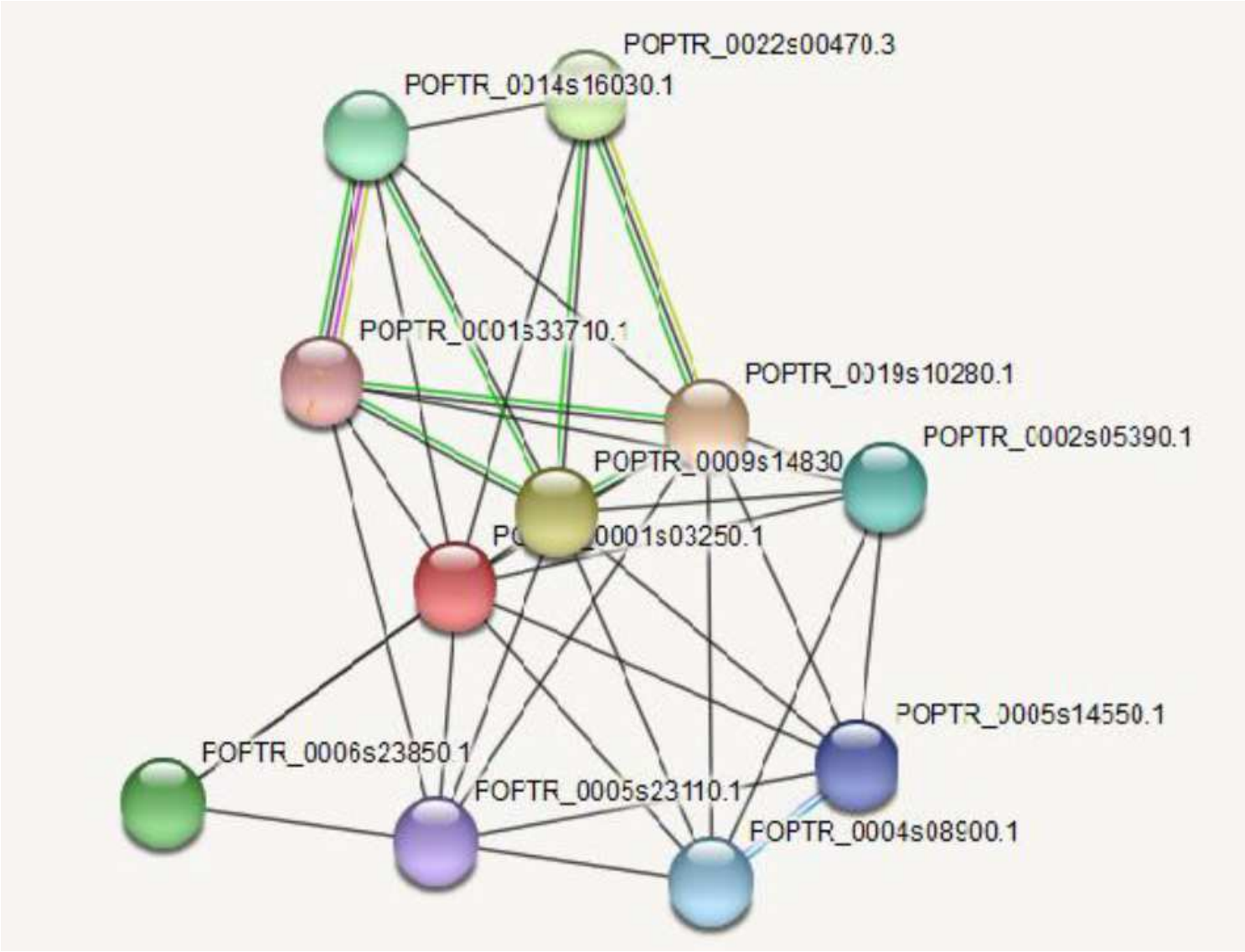

Figure 4

Network of protein-protein interactions by STRING database information.

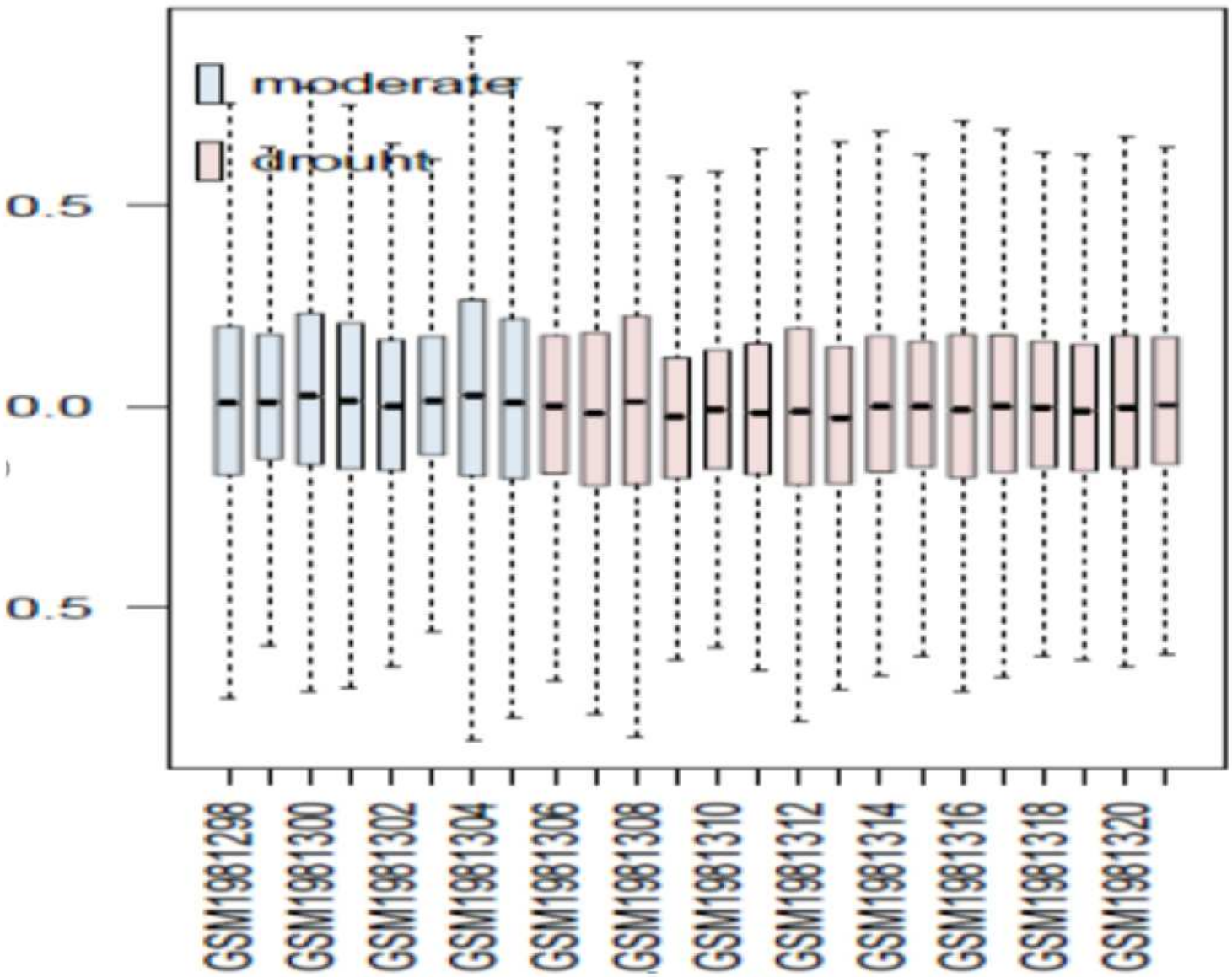

Figure 5

Boxplot of GSE764322 using in this study.