Detecting Drought Regulators using Stochastic Inference in Bayesian Networks

Aditya Lahiri (adi441994@gmail.com)
Texas A&M University College Station https://orcid.org/0000-0001-9352-1312
Lin Zhou
Texas A&M University College Station
Ping He
Texas A&M University College Station
Aniruddha Datta
Texas A&M University College Station https://orcid.org/0000-0003-1213-3807

Research article

Keywords: Drought, Bayesian Networks, Inference, ATAF1, MYC2

Posted Date: April 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-73056/v2

Detecting Drought Regulators using Stochastic Inference in Bayesian Networks

Aditya Lahiri[1*], Lin Zhou[2], Ping He[2,3], Aniruddha Datta[1,4]

1 Department of Electrical and Computer Engineering, Texas A&M University, College Station, Texas, USA
2 Department of Biochemistry and Biophysics, Texas A&M University, College Station, Texas, USA
3 Institute for Plant Genomics and Biotechnology, Norman E. Borlaug Center, College Station, Texas, USA
4 TEES-AgriLife Center for Bioinformatics and Genomic Systems Engineering (CBGSE),College Station, Texas, USA

* alahiri2@tamu.edu

Abstract

Drought is a natural hazard that affects crops by inducing water stress. Water stress, induced by drought, accounts for more loss in crop yield than all the other causes combined. With the increasing frequency and intensity of droughts worldwide, it is essential to develop drought-resistant crops to ensure food security. In this paper, we model multiple drought signaling pathways in Arabidopsis using Bayesian networks to identify potential regulators of drought-responsive reporter genes. Genetically intervening at these regulators can help develop drought-resistant crops. We create the Bayesian network model from the biological literature and determine its parameters from publicly available data. We conduct inference on this model using a stochastic simulation technique known as likelihood weighting to determine the best regulators of drought-responsive reporter genes. Our analysis reveals that activating *MYC2* or inhibiting *ATAF1* are the best single node intervention strategies to regulate the drought-responsive reporter genes. Additionally, we observe simultaneously activating *MYC2* and inhibiting *ATAF1* is a better strategy. The Bayesian network model indicated that *MYC2* and *ATAF1* are possible regulators of the drought response. Validation experiments showed that *ATAF1* negatively regulated the drought response. Thus intervening at *ATAF1* has the potential to create drought-resistant crops.

Author summary

Drought is a natural climatic phenomenon characterized by prolonged periods of dry conditions. It leads to a reduction in precipitation, streamflow, groundwater and creates a general water shortage in the areas it afflicts, thereby making crops difficult to irrigate and sustain. Droughts are the leading cause of crop loss globally, and with climate change, droughts are increasing in both frequency and intensity. Hence, droughts pose a significant food security risk, and there is a need for drought-resistant plants. Fortunately, plants possess drought signaling pathways that function as their internal defense mechanism. These signaling pathways enable plants to detect drought conditions and develop defensive traits through complex and stochastic interactions of

genes and transcription factors present in them. Thus modeling these pathways can provide us with valuable insights needed for creating drought resilient plants. Therefore, we modeled various drought signaling pathways using a graph and probability theory-based technique known as Bayesian networks. The model identified two genes that were significant in responding against drought. We performed a validation experiment that established one of these genes to be a drought regulator. Intervening at this gene (ATAF1) using methods such as gene editing has the potential for creating drought-resistant plants.

Introduction

Drought is a natural hazard characterized by prolonged periods of dry conditions which can lead to economic, humanitarian, and ecological crises. In the context of agriculture, drought occurs when the amount of water available is not enough to sustain crops; such deficiency of water may arise from the lack of precipitation, soil water deficit and reduced levels of ground or reservoir water [1, 2]. It is important to study the effect of droughts on agriculture as it is usually one of the first sectors to be impacted [3]. The United Nations Food and Agriculture organization reported that between 2005-2015 the agricultural sector of the developing countries suffered a loss of $ 29 Billion due to droughts [4]. In the United States, the state of California alone incurred a loss of $ 3.8 billion dollars from 2014-2016 due to recent droughts in California(2012-2016) [5]. Although the long term global drought trends have been a subject of debate, recent regional studies have shown an increasing trend of intensity and frequency of droughts across the Mediterranean, Western Africa, Central China and Southwest and Central Plains of Western North America [6–10]. According to the special report published by the Intergovernmental Panel on Climate Change (IPCC) in 2018, human activities have contributed to global warming, and, at the current rate of warming, temperatures will rise by 1.5 °C between 2030 and 2052 [11]. This warming of the climate is projected to increase the frequency and intensity of droughts, especially in the southern African and Mediterranean regions [12]. Droughts are not caused by global warming alone; recent studies have shown that in the southwestern regions of the United States, droughts are expected to be more frequent and hotter due to structural changes in forested ecosystems and mass mortality of trees [13]. Along with being expensive events, droughts also threaten food security by affecting the global crop yield. With food security being a grand challenge due to a rising global population, frequent and more intense droughts in the future only serve to exacerbate this problem [14]. Thus, it is of paramount importance to develop crops which are robust against drought.

While the risk of imminent droughts has motivated the agricultural community to search for novel methods to develop drought resilient crops, it is important to recognize that plants have adapted and evolved their internal defense mechanisms over the years to counter against harsh climatic phenomena such as drought. Under drought conditions, plants can implement various strategies to conserve water to ensure their survival. For instance, plants can develop longer roots to search for water, shed their leaves early, slow their growth, or develop spines to conserve water in response to drought [15]. In addition to a plant's internal defense mechanism against drought, farmers have relied on traditional methods of plant breeding such as selection and hybridization to combat drought. These methods have been successful in developing drought resistant plants in the past; however, progress has been slow due to the limited understanding of genetic and molecular interactions in the signaling pathways involved in the defense response of plants against drought [16]. Thus it is essential to develop a strong understanding of these signaling pathways. In this paper, we use Bayesian networks (BNs) to model the various drought signaling pathways of the model plant

Arabidopsis. We use BNs as they allow us to combine biological pathway information along with experimental data, which is essential for developing a complete understanding of the interactions that take place inside a plant under drought conditions. We then perform inference using likelihood weighting in the BN model to identify targets in the pathways that regulate drought responsive genes. Genetically intervening (activating/inhibiting) at these target sites using methods such as CRISPR-Cas9 can help develop drought resistant plants [17].

Plant Defense Mechanisms

Most living organisms can escape harsh environments by seeking refuge in favorable locations however, plants are immobile organisms and have to adapt to these conditions. If plants do not adapt to stressful conditions then their growth, development, yield, and seed quality may be hampered [18]. Plant stress can be categorized into two groups, biotic and abiotic. Biotic stress includes attacks on the plant by herbivores, bacteria, fungi, and other pathogens, whereas under abiotic stress the plant faces detrimental environmental conditions such as extreme temperatures, droughts,and mineral toxicity. Plants defend against such stress by activating complex networks of signaling pathways. These pathways are often activated with the help of small molecules such as Ca^{2+}, reactive oxygen species, nitrogen or phytohormones such as ethylene, jasmonic acid, abscisic acid, and salicylic acid which serve as biological stress sensors [19]. These pathway activators often initiate a protein phosphorylation cascade to directly target defensive proteins or transcription factors to regulate the stress responsive genes [20]. Under stressed conditions, the natural metabolic homeostasis of plants is disrupted and, by activating the stress signaling pathways, plants achieve a new state of homeostasis; this process is commonly referred to as acclimation [21].

When a plant comes under drought conditions, it typically responds by implementing drought escape, avoidance, and tolerance strategies [22]. Drought escape strategies involve the plant developing high plasticity and completing its life cycle before the onset of drought whereas under drought avoidance the plant learns to maintain high water content in its tissues by increasing water uptake and reducing water loss [22–24]. Drought tolerant strategies are characterized by the plant developing traits such as epicuticular wax formation, osmotic adjustment, cellular elasticity, and protoplasmic resistance. These strategies allow the plant to survive in drought conditions with low tissue water content [24]. Plants do not deploy these defensive responses one at a time; instead they implement a combination of these strategies to cope against drought [23]. Such a diverse range of defensive responses is achieved through the actions of Gene Regulatory Networks (GRNs) [24, 25]. GRNs are complex networks of genetic regulators called Transcription factors and their target genes; GRNs are directly responsible for altering the gene expression of plants when they receive environmental cues such as drought [26]. Due to these reasons, in this paper, we are interested in modeling the various GRNs, that are activated in plants in response to drought. Modeling these genetic interactions will help us establish a deep understanding of how plants deploy phenotypical defensive behavior through the actions of genes and transcription factors. Such a model will also help us identify the key regulators of drought response. The various GRNs involved in drought response in Arabidopsis are described in the following section.

Drought Signaling Networks

In this paper, we build a BN model from several signaling pathways, involved in the drought response of Arabidopsis. Since the plant's response to drought happens in a complex manner, it is necessary to build a comprehensive network model that can

capture the multivariate and stochastic interactions taking place under drought conditions. Drought responses in plants are largely regulated by Abscisic acid (ABA) dependent and independent pathways [27]. Under drought conditions, the levels of ABA which serves as a stress sensor rapidly increase in a plant, and the plant subsequently responds by closing its stomata and inducing drought responsive genes [28]. ABA regulates the expression of these genes through transcription factors in its drought signaling pathway. The basic-domain leucine zipper (*bZIP*) transcription factor and its subfamily of ABA-responsive element-binding protein/factor (*AREB*/*ABF*) constitute the primary transcription factors through which ABA regulates drought responsive genes [29, 30]. Under drought conditions, ABA induces *AREB1*(*ABF2*), *AREB2*(*ABF4*), *ABF1*, and *ABF3* from this transcription factor family in the vegetative tissues of Arabidopsis [31]. ABA along with another plant phytohormone Jasmonic Acid (JA) is known to regulate the expression of the drought responsive gene *RD22* in Arabidopsis via the transcription factors *MYB2* and *MYC2* [32, 33]. *MYB2* and *MYC2* act as a point of crosstalk between the ABA and JA signaling pathways. On the other hand, Dehydration-responsive element binding protein 1 (*DREB1*)/*CBF* (C-repeat binding factor) and *DREB2* transcription factor families operate independently of the ABA dependent pathway to regulate the drought responsive gene *RD29A*. This is achieved by the actions of transcription factors *DREB1A*(*CBF3*), *DREB1B*(*CBF1*), *DREB1C*(*CBF2*), and *DREB2A* [33, 34]. *DREB1A*, *DREB1B*, and *DREB1C* are negatively regulated by a transcription factor *MYB15* and positively regulated by another transcription factor *ICE1* [35–37]. While *ICE1* negatively regulates *MYB15*, it is suppressed by transcription factors *HOS1* and upregulated by transcription factor *SIZ1* [38]. Among the various members of the DREB1 and *DREB2* family, *DREB2A* and *DREB1D*(*CBF4*) play an interesting role in regulating drought response. Unlike the other *DREB* transcription factors discussed here which function independent of the ABA pathway, *DREB2A* and *DREB1D* can be induced by the ABA pathway through the *ABRE* transcription factor family under drought conditions [33, 39, 40]. Therefore *DREB2A* and *DREB1D* serve as another point of crosstalk for both ABA dependent and independent pathways in regulating drought responsive genes. *DREB2A* was found to be further regulated by *DRIP1*. Singh et al. (2015) found that transgenic Arabidopsis overexpressing *DRIP1* delayed the expression of drought responsive genes regulated by *DREB2A* [33]. Downstream of the *DREB* and *ABRE* transcription factors is the drought responsive gene *RD29A* which is heavily regulated by these transcription factors [29, 40–42].

A recent study by Li et al. (2017), identified a drought stress-activated mitogen-activated protein (MAP) kinase cascade in cotton that regulates the expression of a drought responsive transcription factor *GhWRKY59*. *GhWRKY59* directly binds to the W-boxes of the transcription factor *GhDREB2* to regulate drought response in cotton [43]. We include this ABA independent pathway in our study of the drought regulatory network in Arabidopsis where the MAP Kinase cascade is known to converge at the transcription factor *DREB2A*. In building our network model, we also study the *WRKY* transcription factor family which is traditionally associated with defense response against pathogens. However, many studies have now shown that *WRKY* transcription factor is involved in the defense response against drought [44–46]. The *WRKY* transcription factors *WRKY40*, *WRKY60*, *WRKY18* are induced by ABA to regulate the expression of *RD29A* [47]. *WRKY18*, *WRKY60* are known to positively regulate the expression of *RD29A* whereas *WRKY40* inhibits *RD29A* and *WRKY60* [48]. Our previous paper on modeling the *WRKY* transcription factor in Arabidopsis under drought further confirmed these regulatory behaviors of the *WRKY* transcription factor family [49]. It should be noted that there is often crosstalk between ABA dependent and other independent pathways, we noted two instances of this earlier.

Another instance of the crosstalk between the JA and ABA pathways was highlighted by Mintgen et al. (2014), where *WRKY60* from the ABA pathway suppresses the expression of *MYB2* in the JA pathway to regulate the drought responsive gene *RD22* [50]. Other than *RD22*, *MYB2* and along with *MYC2* also regulate the expression of another drought responsive gene *ERD1* [33]. According to a study by Ollas et al. (2016), *MYB2* and *MYC2* regulated the expression of *ERD1* through a cluster of transcription factors (*ANAC019*, *ANAC055*, and *ATAF1*) belonging to the *NAC* transcription factor family. *ERD1* was found to be further regulated by the transcription factor zinc finger homeodomain 1 (*ZFHD1*) and the gene *RD26* (*ANAC072*) in the ABA pathway [51]. In addition to the drought responsive genes *RD29A*, *ERD1*, and *RD22*, we also consider the gene *RD20* in our network model. *RD20* was found to be directly upregulated by the gene *RD26*(*ANAC072*) [51]. The biological interactions discussed above are summarized in Fig.1. In the next section, we create a Bayesian network model based on these signaling pathways to predict the best regulator(s) for the drought responsive genes (marked in blue in Fig.1).

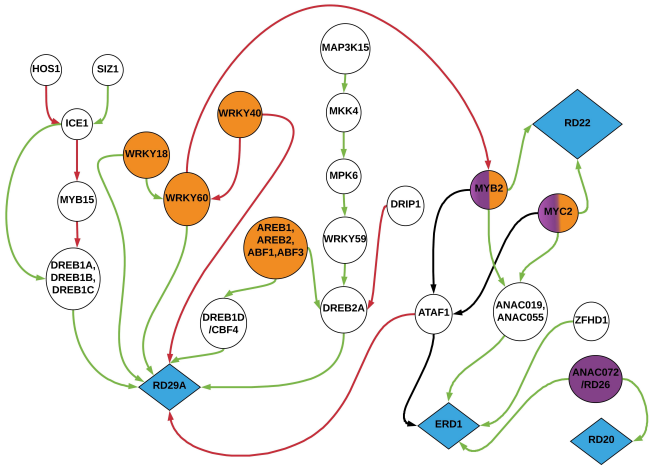

Fig 1. Drought signaling pathways in Arabidopsis. The orange circular nodes represent elements directly regulated by ABA whereas the purple nodes represent elements regulated by JA. The two nodes colored with a mix of orange and purple represent elements regulated by both JA and ABA pathways (Crosstalk). The blue diamonds represent drought responsive reporter genes. The plain circular nodes with no colors represent the transcription factors, genes and proteins involved in the regulation of drought responsive reporter genes in an ABA independent manner. The green and red arrows represent positive and negative regulation. The arrows going into and out of *ATAF1* are marked black to indicate that the nature of regulation is not known at this time.

Materials and methods

Bayesian Network Model

We observed in the previous section that plants deploy a diverse range of defense mechanisms to survive under drought conditions. These phenotypical defense responses are mediated through complex networks of signaling pathways at the genomic level. Biological signaling pathways have been successfully modeled using methods such as linear models, Boolean networks, probabilistic Boolean networks, Bayesian networks,

and small molecule level models [52–57].In order to develop a thorough understanding of these multivariate and stochastic interactions, we create a BN model of the drought signaling pathways. Unlike some modeling techniques which are solely driven by data, a BN model allows us to integrate pathway information in the form of prior knowledge along with experimental data [58]. BNs are directed acyclic graphs that represent the causal probabilistic relationships among a set of random variables and provide the conditional decomposition of the joint probability distribution of these random variables [59, 60]. Thus BNs serve as an ideal modeling paradigm to study the drought signaling pathways [58]. In this paper, our objective is to create a BN model of the drought signaling pathways outlined in Fig.1 and use this model to determine which transcription factor, protein or gene is the best regulator of drought responsive reporter genes (blue diamonds in Fig.1). The predictions made by the model can help us identify potential targets for genetic intervention techniques like CRISPR-Cas9 to create drought resistant crops.

Fig.2 represents the BN model of the signaling pathways shown in Fig.1. Every node (circle) in the network represents a gene, protein or transcription factor in the drought signaling pathway. The black arrows or edges connecting the nodes represent the causal biological relationships we discussed in the previous section. We assume each of the nodes are binary random variables which can assume the state of 1 for activation and 0 for inhibition. Since the nodes are random variables, associated with each of them is a parameter θ which describes the local marginal or conditional probability distribution for that node. For instance, the conditional probability parameter associated with the node representing *MKK4* is given by $\theta_{MKK4|MAP3K15}$. This parameter represents the activation or inhibition probability of the node representing *MKK4* conditioned on the state of the node representing *MAP3K15*. Similarly, for the node representing the transcription factor *ICE1*, the local marginal probability distribution is given by θ_{ICE1}. Henceforth, we will refer to local conditional or marginal probability distribution as just local probability distributions (LPD). We learn these LPDs from biological experimental data; once these LPDs are learned the BN model is complete and can be used for carrying out inference simulations to determine the best modulator for the drought responsive genes.

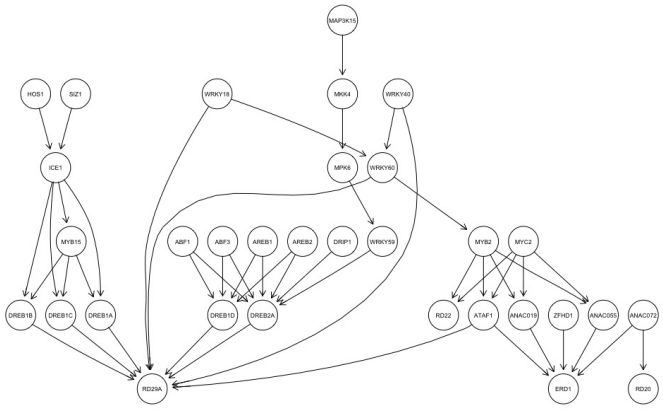

Fig 2. Bayesian Network Model of Drought Signaling Pathway. Every circular node represents a biological element in the drought signaling pathway. Every edge or black arrow represents the causal biological relationship between the nodes. Associated with every node is a θ parameter that represents the local probability distribution of the node.

Parameter estimation in Bayesian networks

BNs consist of two major components: a directed acyclic graph (DAG) and a set of local probability distributions. The DAG can be learned from data or constructed from domain knowledge. Learning BNs from data, also known as structure learning in the literature, is an NP-Hard problem and requires us to choose a DAG from several candidate DAGs [61]. This is not very practical as we observed in section 2 that pathway interactions are well defined and there can only be a single DAG representing them. Furthermore, in the context of Arabidopsis under drought, we are limited by the sizes of publicly available datasets. These datasets are not large enough to construct a reliable DAG, so we elected to create the BN model in Fig.2 using pathway information from the existing biological literature. While a DAG can be learned either using data or from domain knowledge,the local probability distributions associated with the DAG have to be estimated from experimental data. There are several ways to estimate the local probability distribution in a BN model. Typically, either a frequentist approach such as a Maximum Likelihood Estimate (MLE) or a Bayesian approach is employed. Though methods such as MLE are simple and provide a point estimate, they are only driven by data and do not take any relevant prior information into account [62]. On the other hand, a Bayesian approach provides us with the posterior distribution, which is driven by both data in the form of likelihood and relevant information in the form of a prior distribution. However, the Bayesian approach has two significant drawbacks. The first one is computing the normalizing constant or the probability of data (evidence) [63]. The normalizing constant very rarely has a closed form solution and hence can be computationally expensive to determine. The second drawback pertains to the choice of a prior distribution. Since the choice of distribution is subjective and there is no established method to select one, different choices of prior distribution will lead to different results [64]. Nonetheless, the Bayesian approach is logically rigorous and unlike frequentist approaches, once the prior distribution is established the Bayesian approach follows deductive logic. In this paper, we use a Bayesian approach to estimate the local probability distributions for the BN model outlined in Fig.2. We assumed that the nodes are binary random variables, which implies that for any node $\mathbf{X}$ in the BN, $\mathbf{X}$=1 (success) when the node is activated and $\mathbf{X}$=0 (failure) when the node is inhibited. Then for a single observation for any node $\mathbf{X}$ in the BN be modeled as a Bernoulli random variable.

Let us suppose that we have a BN model with $\mathbf{N}$ nodes. Then the probability with which any node $\mathbf{X}$ attains a state of 1 is given by θ_X. Thus if we make n (>0) independent and identically distributed observations (i.i.d) observations for each node in the BN, and if for a given node $\mathbf{X}$, we observe k instances when the node attains a state of 1, then the likelihood for node $\mathbf{X}$ is given by:

$$P(X|P_a(X),\theta_X) \sim Binomial(n,\theta_X) \tag{1}$$

$$Binomial(n,\theta_X) = \frac{n!}{k!(n-k)!}\theta_X^k(1-\theta_X)^{n-k} \tag{2}$$

P_a(X) in Equation(1) refers to the parents, if any, of node $\mathbf{X}$. Since we are using a Bayesian approach to estimate the LPD of Node $\mathbf{X}$, we need to select a prior distribution on the node $\mathbf{X}$. Considering the computational complexity required in calculating the normalizing constant, and since the likelihood function associated with our model follows a binomial distribution by design, we assume the prior distribution on θ_X to follow a Beta distribution. Since the Beta and Binomial distributions belong to conjugate families, we know that the posterior distribution of θ_X will also follow a Beta distribution [65]. This is formulated as follows:

$$\theta_X \sim Beta(\alpha_X, \beta_X) \tag{3}$$

$$P(\theta_X|X) \sim Beta(\alpha_X^{'}, \beta_X^{'}) \tag{4}$$

where $\alpha_X^{'}$ =α_X +k and $\beta_X^{'} = \beta_X$ + (n-k).

In equation (3), α_X and β_X represent the shape parameters of the Beta distribution, and in equation (4) these parameters get updated for the posterior distribution on θ_X. We assume α_X =1 and β_X =1 for our calculations as the Beta(1,1) distribution corresponds to the standard uniform distribution over the interval [0,1] [66]. Setting the prior distribution to the standard uniform distribution guarantees that we have no information regarding the prior distribution of θ_X. We chose the Beta(1,1) distribution as our prior because we do not have any domain knowledge information regarding the prior distribution of every node in the BN model. If we had such information regarding the prior distribution, they could be incorporated into this model. However, it is to be noted that choosing a different prior distribution may not allow us to reach a closed form solution for the posterior distribution on θ_X. Since the result we get in Equation (4) is a distribution and not a point estimate like what we would have obtained had we used a frequentist approach, we approximate the values for θ_X with the expected value of the posterior distribution. We do this approximation for the posterior distributions estimated at every node in the BN. This approximation for the node **X** has been presented in Equation (5).

$$\theta_X \simeq E[\theta_X|X] = \frac{\alpha_X^{'}}{\alpha_X^{'} + \beta_X^{'}} \tag{5}$$

Once these parameters are learned the BN is complete as we have both the DAG and the set of conditional probabilities. In the next section, we study the effect on drought responsive genes for intervening (activating/ inhibiting) at various nodes, then summarize our findings in the results section.

Sampling based inference in Bayesian networks

In this section we are interested in using the BN model, to determine which nodes are the best regulators of the drought responsive reporter genes *RD29A*,*RD20*,*RD22*,and *ERD1*. Specifically, we want to study the effect on the reporter genes of intervening at the non-reporter genes. In other words, we will fix the state of every non-reporter gene node one at a time, to either 0 or 1 and observe how this action (intervention) affects the LPDs for the nodes representing the drought responsive reporter genes. This kind of simulation in BNs is known as inference. Inference techniques are categorized as either exact or approximate. Exact inference techniques such as Enumeration, Variable Elimination, and Pearl's Message Passing Algorithm are particularly efficient in polytrees or singly connected networks. One such application of exact inference was demonstrated by Vundavilli et al. to find significant nodes in the breast cancer signaling pathway [67]. Ideally, we would like to implement an exact inference technique for the BN model we have developed in this paper; however, such a technique will be computationally expensive as our network is multiply connected, i.e. there are at least two nodes in our BN model that are connected by more than one path. For instance, we can see that the node *DREB1A* and *ICE1* are directly connected and also connected through *MYB15*, hence making the BN model multiply connected. While exact inference algorithms work in polynomial time in polytrees, it has been shown to be NP-Hard in more generalized BNs, hence implementing them in multiply connected

networks may not be practical [68]. Therefore, the size and structure of the BN would govern our choice of inference techniques. This is the reason why, for determining the regulators of drought responsive reporter genes, we employ an approximate inference technique known as likelihood weighting.

Likelihood Weighting (LW) is an approximate inference technique based on stochastic simulations. Inference techniques based on stochastic simulations usually involve drawing samples from a sampling distribution, calculating an approximate posterior probability based on the samples, and then showing that the posterior probability converges to the actual probability [69]. In the context of our model, the sampling distribution will be specified by the BN in the form of LPDs. Unlike exact inference techniques, LW is generally insensitive to the network topology, however, convergence in estimating the posterior probabilities can be slow if they are close to 0 or 1 [70]. We will now describe the mathematical formulation for LW.

Consider a BN consisting of N nodes such that the DAG follows a topological ordering of $\{X_1,X_2,..,X_N\}$. Suppose we make an observation on the node X_E in the BN, we will refer to X_E as the evidence node. Now suppose our objective is to find the effects of this observation on another node X_Q, known as the query node in the BN. Specifically, we want to estimate the posterior probability $Pr(X_Q{=}x_q|X_E{=}x_e)$, where 'x_q' and 'x_e' are some instantiation of nodes X_Q and X_E. At this step we begin performing LW by drawing M samples from the BN for every node except for the evidence node X_E, in topological order. The generated dataset (ξ) will be a matrix with M rows and N columns, where each row represents an N-dimensional sample (datapoint) and columns represent nodes in the BN. Thus after the first iteration of the sample generation process, the datapoint will be of the form $\xi^{(1)} = \{ {x_1}^{(i=1)}$,${x_2}^{(i=1)}$,....,${x_e}^{(i=1)}$,..,${x_N}^{(i=1)}\}$. We will repeat this process M-1 more times to obtain M such samples , thus that dataset will be of the form $\xi = \xi^{\{i=1,2,..,M\}} = \{$ ${x^{(i)}}_1$,${x^{(i)}}_2$,....,x_e,..,${x^{(i)}}_N\}$. It should be noted that x_e, does not change across the M samples. This is because X_E is the evidence variable that has been observed and fixed. The samples for the rest of the non-evidence nodes are generated according to the LPDs associated with those nodes. For example we draw a sample x_1 for root node X_1 according to $Pr(X_1)$. Similarly we draw a sample x_2 for the node X_2 according to $Pr(X_2 \mid X_1 = x_1)$ and so on. It should be noted that all the children of node X_E have a fixed instantiation for X_E, that is x_e. We then approximate $Pr(X_Q{=}x_q|X_E{=}x_e)$ as follows:

$$Pr(X_Q = x_q|X_E = x_e) \simeq \lim_{M\to\infty} \frac{\sum_i^M \mathbb{1}[x_q^{(i)} = x_q] Pr(X_E = x_e|(P_a(X_E))^{(i)})}{\sum_i^M Pr(X_E = x_e|(P_a(X_E))^{(i)})} \quad (6)$$

The proof for Equation (6) is not trivial and is presented in a paper by Menon [71]. A psudo code for estimating the conditional probabilities using LW is presented in algorithm 1.We will now demonstrate LW on an example BN.

Fig.3 describes an example BN consisting of four genes A,B,C, and D. We consider the nodes representing the genes as binary random variables which can take on the values of 1 for activation and 0 for inhibition. The LPDs for this example BN are already estimated and are presented in Fig.3. For the purpose of this example, we assume that Gene A positively regulates gene B, while it negatively regulates gene C. Gene D is upregulated by gene B, while gene C downregulates it. These effects are reflected in the LPDs for each node. Now suppose, we are interested in gene D being positively regulated and we decide to intervene at Gene B and set it to 1. Therefore, node B=1 serves as the evidence variable, and let us consider node D as the query variable. Then we are interested in finding the probability P(D|B=1) using LW.

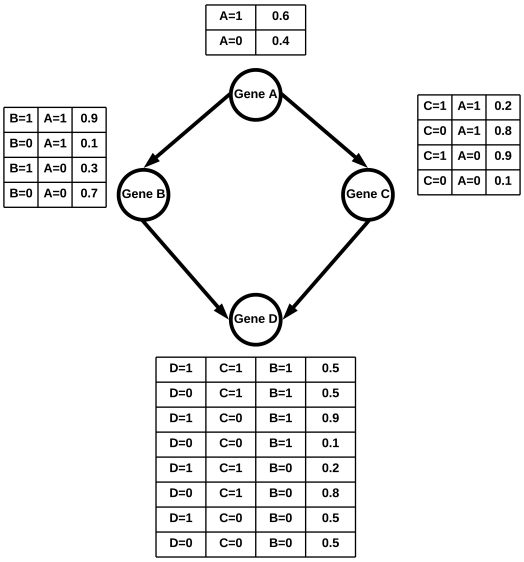

Fig 3. Example BN with LPDs. Gene A positively regulates Gene B and negatively regulates Gene C. Gene B positively regulates Gene D and Gene C negatively regulates Gene D.

Algorithm 1: Psudo Code for likelihood weighting in Bayesian Networks

Input:

BN : The Bayesian Network
Q: The Query Variable, Let Q=q,
that is node Q is instantiated to some value of interest q.
E: The Evidence variable. Let E=e,
that is node E is instantiated to some observed value e.
M: Number of Samples.

Output: Probability: Estimate of P(Q=q|E=e)

Initialization: X_1,X_2,..,X_N Topological Ordering of BN
Sampled_Data= {} {} ,M by N matrix to store sampled data
Weight= {1,...,1}, an array of size M, consisting of weights
with values initialized to 1.
Counts[k]=0, where k ∈ domain of Q
while iter= 1 to M **do**
  for each node X in BN in topological order **do**
    if X=X_i is in E **then**
      Sampled_Data[iter][X_i]= x , where x is the value of X_i
      Weight[iter]= Weight[iter] * P(X_i=x | P_a (X_i))
    else
      Sampled_Data[iter][X_i]= Generate random sample from P(X_i=x|P_a(X_i))
    end if
  end for
  iter=iter+1
end while
k = List of row indices in Sampled_Data where Q=q
Probability = Sum (Weights [k])/ Sum(Weights)
return Probability

In order to estimate this probability, we will need to query the BN and generate samples first. We use the topological ordering of{A,B,C,D}, another valid ordering is {A,C,B,D}. The sample generation process is described in the following steps:

1. Set the weight variable 'W_i' to 1. $W_{(i)}$=1
2. The matrix Sampled_Data[iter][X_i]= is empty. This matrix will store the value of nodes A,B,C,D.
3. We start topologically at node A. Since A is not an evidence node, we sample it according to its LPD, specifically P(A). Assume this sample results in A=1.
4. We now move on to node B. Since B is an evidence node, we do not sample it. We update, $W_{(i)}$=1. P(B=1|A=1)= 1. (0.9)= 0.9.
5. We now go to node C. Since C is not an evidence node, we sample it according to its LPD, specifically P(C|A=1). Let us assume the result of this process is C=0.
6. We now sample node D with its LPD of P(D|B=1,C=0). Assume that this results in D=1.
7. The sample generated is (A=1,B=1,C=0,D=1) with $W_{(i=1)}$ =0.9. Thus Sampled_Data[1][All Columns] = [1,1,0,1]
8. We repeat steps 1-7, M-1 more times to obtain a total of M samples.
9. We can then calculate P(D|B=1) as follows:

$$P(D=1|B=1) = \frac{\sum_{i=1}^{M} W_i \mathbb{1}[D_{(i)}=1]}{\sum_{i=1}^{M} W_i}$$
$$P(D=0|B=1) = \frac{\sum_{i=1}^{M} W_i \mathbb{1}[D_{(i)}=0]}{\sum_{i=1}^{M} W_i}$$

Therefore for M=5 , if we generated sample, it would result in a 5 by 4 matrix (Sampled_Data[iter],[X_i]). Table 1 shows this matrix with an extra column for weights belonging to each sample. From the samples and weights in Table 1, we can now estimate P(D=1|B=1) and P(D=0|B=1) as follows:

$$\begin{aligned} P(D=1|B=1) &= \frac{\sum_{i=1}^{5} W_i \mathbb{1}[D_{(i)}=1]}{\sum_{i=1}^{5} W_i} \\ &= \frac{W_1*1+W_2*0+W_3*1+W_4*0+W_5*1}{W_1+W_2+W_3+W_4+W_5} \\ &= \frac{0.9*1+0.3*0+0.9*1+0.9*0+0.3*1}{0.9+0.3+0.9+0.9+0.3} \\ &= \frac{2.1}{3.3} \\ &= 0.636364 \end{aligned}$$

$$\begin{aligned} P(D=0|B=1) &= \frac{\sum_{i=1}^{5} W_i \mathbb{1}[D_{(i)}=0]}{\sum_{i=1}^{5} W_i} \\ &= \frac{W_1*0+W_2*1+W_3*0+W_4*1+W_5*0}{W_1+W_2+W_3+W_4+W_5} \\ &= \frac{0.9*0+0.3*1+0.9*0+0.9*1+0.3*0}{0.9+0.3+0.9+0.9+0.3} \\ &= \frac{1.2}{3.3} \\ &= 0.363636 \end{aligned}$$

Table 1. Sample Data from Example Bayesian Network.

index	A	B	C	D	Weight(W_i)
1	1	1	0	1	0.9
2	0	1	1	0	0.3
3	1	1	1	1	0.9
4	1	1	0	0	0.9
5	0	1	0	1	0.3

Dataset and Simulation

To estimate the LPDs for the BN model, we required data for Arabidopsis under drought conditions. We use the dataset GSE42408, which is publicly available at the NCBI GEO database [72, 73]. This dataset contains 104 eQTL data points for Arabidopsis under drought conditions. The data for each node is normalized using min-max feature scaling. We further compute the normalized means for each node and use it as a threshold for binarizing the data. Additional details on the normalization and binarization process can be found in the R scripts provided in the supporting information section. All the code and data files are also made available publicly at the following GitHub repository: https://github.com/adilahiri/Drought_Regulators. The processed data was then used to learn the LPDs for each node and perform inference using LW. We chose a sample size (M) of 600,000 in the LW algorithm to ensure convergence in estimating the conditional probabilities. The model building and all the associated data processing tasks were completed using the R programming language [74]. The Bnlearn package was used to perform inference using LW [75].

Results

Fig 4. displays the dataset GSE42408, after it was normalized and binarized. Each bar in Fig.4 represents the inhibition and activation counts for each node in the BN. We use the Bayesian approach as discussed in section 3.1, with Beta (1,1) as the prior distribution for each node to estimate the LPDs. For the inference analysis, the query nodes were the drought responsive reporter genes *RD29A*, *RD20*, *RD22*, and *ERD1*. We were interested in the activation of *ERD1*, and the inhibition of *RD29A*, *RD20*, and *RD22*. Though all these reporter genes have been shown to confer drought resistant characteristics, they also impart undesirable traits such as sterility, reduced seed yield, and dwarfing [51]. Thus activating all of them is not optimal, hence for our analysis we are interested in finding a single node which upon intervention would increase the chances of the reporter gene *ERD1* being activated and the reporter genes *RD29A*,*RD20*, and *RD22* being inhibited. Since the LW yields a probability for the status of every drought reporter node based on performing an intervention at an evidence node, we establish a composite scoring metric defined in Equation (7) below.

$$\begin{aligned} &Score(Evidence = \{0,1\}) = \\ &Pr(RD29A = 0|Evidence = \{0,1\}) \\ &Pr(RD22 = 0|Evidence = \{0,1\}) \\ &Pr(RD20 = 0|Evidence = \{0,1\}) \\ &Pr(ERD1 = 1|Evidence = \{0,1\}). \end{aligned} \tag{7}$$

This metric multiplies the conditional probability for all the drought responsive reporter genes into a single number which is easy to interpret. A high score represents a

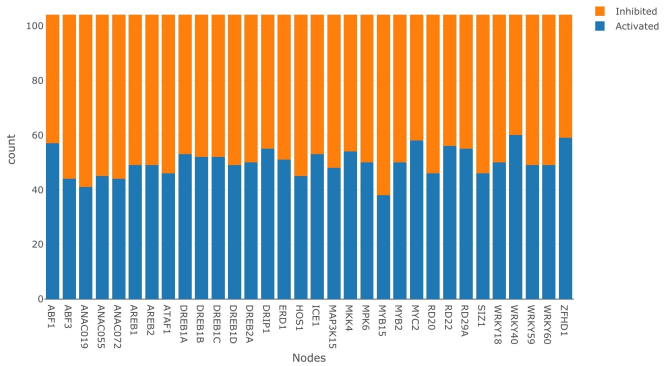

Fig 4. Activation vs Inhibition plot. This figure represents the data after it has been normalized and then binarized. There are a total of 104 data points per node. The blue part of each bar represents activation counts whereas the orange part represents the inhibition counts.

suitable candidate for intervention. In Fig 5. and Fig 6., we present the score for intervening at each of the non-reporter nodes one at a time in the BN. The non-reporter nodes are activated in Fig. 5, whereas in Fig. 6 they are inhibited. From Fig.5 it is clear that when *MYC2* is activated, it results in the highest score, whereas *ANAC072* and *ZFHD1* have the second and third highest scores respectively. On the other hand, in Fig.6, *ATAF1* has the highest score for inhibition followed by *ANAC019*. Based on our analysis, either activating *MYC2* or inhibiting *ATAF1* is the best strategy to activate *ERD1* and inhibit *RD29A*, *RD20*, and *RD22*. We observe that the score for *MYC2* is the lowest when it is inhibited (Fig.6) and the score for *ATAF1* is lowest when it is activated (Fig.5), this makes logical sense for the analysis.

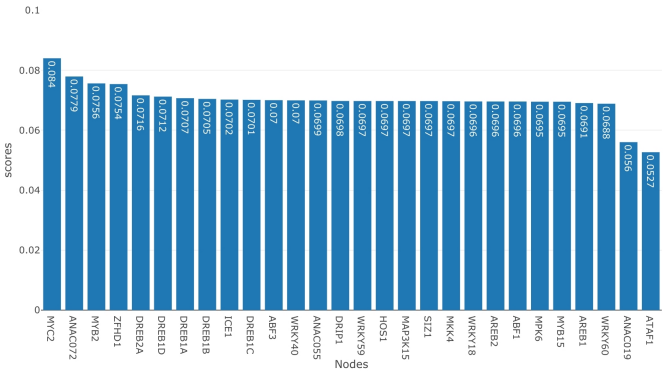

Fig 5. Activation Scores for non-reporter gene nodes. Associated with each node is a blue bar which represents the score for activating that node.

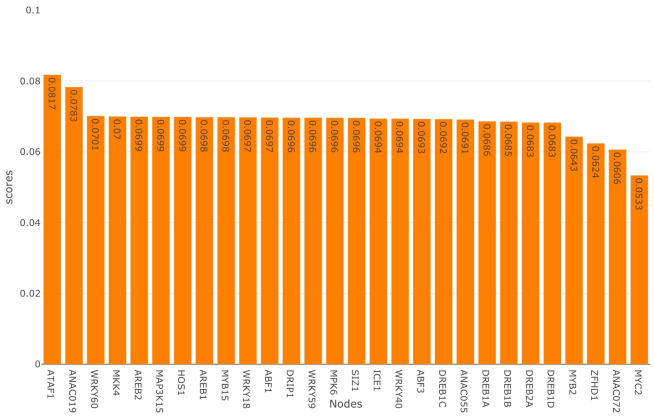

Fig 6. Inhibition Scores for non-reporter gene nodes. Associated with each node is an orange bar which represents the score for activating that node.

The above results from the single node intervention analysis, motivated us to study effects on the drought reporter genes when we simultaneously intervened at *MYC2* and *ATAF1*. In Fig.7, we present the score of simultaneously activating *MYC2* and inhibiting *ATAF1*. Upon comparing this score to the individual scores of activating *MYC2* and inhibiting *ATAF1*, we notice that the score for the combined intervention is slightly higher, indicating the synergistic effect of intervening strategically at the two nodes. Furthermore, both *MYC2* and *ATAF1* are established regulators of the drought response [76, 77]. *MYC2* is known to be a positive regulator of the drought responsive reporter genes *RD20*,*RD22*,and *ERD1* [78–80]. However, it was found to have no significant regulatory effect on *RD29A* [81]. In contrast to the positive drought regulatory nature of *MYC2*, *ATAF1* is known to negatively regulate the expression of *RD29A* and *RD22* [82]. The regulatory effects of *ATAF1* on *RD20* and *ERD1* are not yet known. Due to *MYC2* being a positive regulator for most of the drought responsive reporter genes, and *ATAF1* being a negative regulator for two of the drought responsive reporter genes, it is biologically consistent for them to be the best regulators under activation and inhibition respectively.

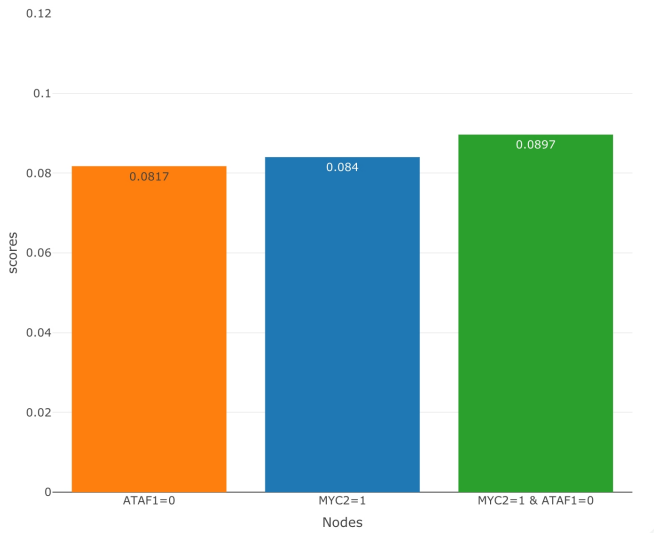

Fig 7. Comparing the scores of multi-node and single node intervention under optimal response case. Simultaneous (multi-node) intervention on *MYC2* and *ATAF1* has a slightly higher score than single node intervention.

Experimental Validation

To validate the conclusions from the Bayesian network model, we isolated Arabidopsis *ataf1* (SALK_057618C) and *myc2* (*myc2*-1, SALK_061267C; *myc2*-2, SALK_128938C) mutants from the Arabidopsis Biological Resource Center (ABRC) [83]. The *ataf1* mutant has a T-DNA insertion in the third exon of the *ATAF1* (AT1G01720) genomic DNA, both *myc2* mutants have a T-DNA insertion in the exon of the *MYC2* (AT1G32640) genomic DNA (Fig. 8A). We germinated wild-type (WT) Col-0 and *ataf1* mutant on the half-strength Murashige and Skoog (MS) medium with or without 300 mM mannitol treatment (Fig. 8B). The addition of mannitol reduces water potential of growth media, which is often used to mimic drought stress (Mu et al., 2019) [84]. Although the germination rate of the *ataf1* mutant was lower than WT in the medium without mannitol, the *ataf1* mutant had more green cotyledon seedlings (Fig. 8B) and higher green cotyledon rate (Fig. 8C) than WT seedlings under 300 mM mannitol treatment. The difference became significant at nine days after germination. We also compared the green cotyledon inhibition rate of WT and *ataf1* mutant on MS medium with or without mannitol. Consistently, the *ataf1* mutant showed lower green cotyledon inhibition rate than WT, and the tendency became more pronounced with the increase of growth time (Fig. 8D). We also germinated WT and *myc2* mutants on the MS medium with or without 300 mM mannitol treatment (Fig. 8E). However, there is no significant difference in the green cotyledon rate between WT and *myc2* mutants with or without mannitol treatment (Fig. 8F). Similarly, the green cotyledon inhibition rate between WT and *myc2* mutants also did not show a significant difference (Fig. 8G). Thus, our data show that the *ataf1* mutant was more tolerant to the mannitol treatment, and suggests that *ATAF1* plays a role in plant drought stress response. Our test conditions, such as plant growth stage, treatment, or the combination, may not be suitable to reveal the difference between WT and *myc2* mutants.

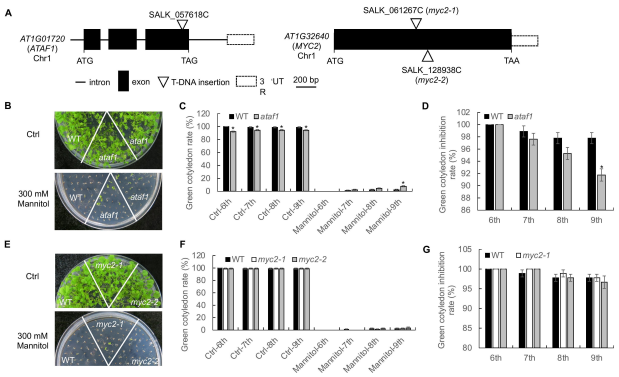

Fig 8. Results from validation experiments. A. The scheme of the *ATAF1* and *MYC2* genomic DNA and T-DNA insertion. The panel is a schematic illustration of the *ATAF1* and *MYC2* genomic DNA with exons (solid box), intron (lines) and 3' untranslated region (open box). The position of T-DNA insertion of *ataf1* (SALK_057618C), *myc2* (SALK_061267C, SALK_128938C)was labeled. **B.**The *ataf1* mutant is more resistant to mannitol treatment. Wild-type (WT) Col-0 and *ataf1* mutant seeds were germinated on 1/2 MS medium with or without 300 mM mannitol. 30 seeds per genotype were used for each replicate. The photos were taken four-week post-germination. **C.**Quantification of cotyledon greening on plates corresponding to B. Seedlings with green cotyledon expansion were counted at 6-9 days post-germination. Data are shown as means ± SD (standard deviation) from three independent replicates (n=3, *, $p<0.05$, Student's t-test). **D.**Quantification of cotyledon greening inhibition rate on plates corresponding to B. Seedlings with green cotyledon expansion were counted at 6-9 days post-germination. Data are shown as means ± SD from three independent replicates (n=3, *, $p<0.05$, Student's t-test). **E.**Growth of WT and *myc2* mutants on MS plates. WT and *myc2* mutant seeds were germinated on 1/2 MS medium with or without 300 mM mannitol. 30 seeds per genotype were used for each replicate. The photos were taken four-week post-germination. **F.**Quantification of cotyledon greening on plates corresponding to E. Seedlings with green cotyledon expansion were counted at 6-9 days post-germination. Data are shown as means ± SD from three independent replicates (n=3, no statistical significance with Student's t-test). **G.** Quantification of cotyledon greening on plates corresponding to E. Seedlings with green cotyledon expansion were counted at 6-9 days post-germination. Data are shown as means ± SD from three independent replicates (n=3, no statistical significance with Student's t-test).

Experimental setup

A. thaliana mutants *ataf1* (SALK_057618C) and *myc2* (SALK_061267C, SALK_128938C) were obtained from the Arabidopsis Biological Resource Center (ABRC). The wild-type (Col-0) and mutant plants were grown in a growth room at 23 °C, 45 % humidity, and 75 μE m^{-2} s^{-1} light with a 12-hr light /12-hr dark photoperiod. To detect cotyledon greening rate, 30 seeds per genotype were sterilized and germinated on half-strength Murashige and Skoog (MS) medium with or without 300 mM Mannitol treatment in each replicate. Seedlings with green cotyledon expansion were counted at 6-9 d post-germination, data are shown as means ± SD from three independent repeats (n=3, *, $p<0.05$, Student's t-test). The photos were taken four-weeks post-germination.

Discussion

As the severity and duration of droughts around the world are predicted to rise in the coming years, developing drought resistant crops is increasingly becoming a priority for ensuring global food security. Thus to develop drought resistant crops, it is necessary for scientists to identify the potent regulators of the drought response in plants. In this paper, we have presented the drought signaling pathway in Arabidopsis and observed that drought response is mediated by the ABA dependent or several ABA-independent pathways. We modeled these pathways using BNs, as it provides a framework to integrate both biological prior knowledge in the form of pathway information along with experimental data. This feature of BNs was a key factor in our selection of this modeling technique. In the BN model, we assumed each node to be a binary random variable with the states of activation or inhibition. We then used the Bayesian approach along with publicly available experimental data to estimate the LPDs associated with the nodes of the BN model. The prior distribution for each node was assumed to follow a Beta(1,1) distribution as this corresponds to the non-informative Uniform distribution on the interval [0,1]. This choice of prior was logical as we did not know the prior distribution for each of the nodes. Furthermore, choosing a Beta prior with Binomial likelihood provides us with a closed form solution for the posterior distribution and reduces our computational requirements. Once the LPDs were learned, we applied an approximate inference technique called likelihood weighting to perform simulations for intervening at the non-reporter gene nodes.

After intervening at the nodes representing the non- reporter genes, one at a time, the results from our simulations indicated that activating *MYC2* or inhibiting *ATAF1* was the best strategy to regulate the drought-responsive reporter genes. We also observed that the score for implementing both the interventions at the same time provides a slightly improved value, indicating the synergistic effect of the strategic interventions. Furthermore, we noted from the biological literature that both *MYC2* and *ATAF1* are known regulators of drought response. However, from the validation experiments, we found that *MYC2* did not have any obvious drought regulatory response as neither the green cotyledon rate nor the green cotyledon inhibition rate between WT and *myc2* mutants with or without mannitol treatment had significant differences. On the other hand, *ataf1* mutants had more green cotyledon seedlings and higher green cotyledon rates than the WT seedlings under mannitol treatment, suggesting that *ATAF1* negatively regulated drought response. We were unable to show that *MYC2* was a drought regulator; this could be due to test conditions or limitations of the Bayesian network model. Testing factors such as plant growth stage, treatment may have been unfavorable for finding the difference between WT and *myc2* mutants. Besides testing factors, we must also consider some of the limitations of the BN model. While we have considered numerous drought-responsive pathways in our BN model, there may be other pathways outside our model's scope, which may interact with the pathways considered in our BN model. These undiscovered interactions may have potentially influenced the drought regulators during the validation experiments. In order to avoid neglecting such interactions, BNs are learned from data using structure learning algorithms. However, this process typically requires large volumes of data, which is currently unavailable. Furthermore, if any previously unaccounted interactions are discovered using structure learning algorithm, we cannot validate them using existing biological literature, and we will need to conduct additional experiments to validate them. Another reason that might have prevented us from proving *MYC2* as a drought regulator is the difference between the experimental setup of our validation experiments and the publicly available dataset(GSE42408) used to learn the parameters of the BN model. The methods used to induce drought in the dataset GSE42408 are different from the methods used in our validation experiments; this might have been

unfavorable in establishing *MYC2* as drought regulator.

This paper's results build upon our previous paper, where we modeled only the WRKY transcription factor signaling pathway in Arabidopsis under drought and found the transcription factor *WRKY18* to be the best regulator of the drought-responsive gene *RD29A* [49]. In our current model, we take into account multiple other pathways, including the WRKY signaling pathway, and observe that the scores across the WRKY transcription factor family are approximately the same and are not as high as the scores for *MYC2* and *ATAF1*. The score for *WRKY18* may be low due to crosstalk happening across multiple pathways, which may negatively impact the regulatory effects of *WRKY18*. Additionally, we tracked multiple drought-responsive reporter genes in our current study, so the score of *WRKY18* in this study reflects its ability to regulate all the drought-responsive reporter genes, unlike in the previous paper, where the score is for the regulation of *RD29A* only. In the future, we would like to extend our research to include more informative priors instead of the non-informative Beta (1,1) distribution. We want to explore new methods to incorporate continuous data into the BN model, rather than to binarize it and lose valuable information. We noticed that multi-node intervention gave a slightly improved score than single node interventions; thus, exploring other node combinations for intervention will be an interesting path for future research.

Conclusion

We modeled several drought-responsive pathways in Arabidopsis using Bayesian Networks and real-world experimental data. Our computational analysis indicated that the transcription factors *MYC2* and *ATAF1* are the most potent candidates for regulating drought-responsive reporter genes. However, we were only able to validate the drought regulatory response of *ATAF1* experimentally. Since *ATAF1* had the highest score for inhibition and validation experiments showed all *ataf1* mutants had a higher green cotyledon rate than WT, it implies that *ATAF1* negatively regulates drought response. Thus genetically inhibiting *ATAF1* with techniques such as CRISPR-Cas9 has the potential to develop drought-resistant crops.

Supporting information

S1 File. MainScript.R Main R code file for executing the Bayesian network simulation.

S2 File. minmax_normalize.R Supporting R code file for normalizing the data.

S3 File. binarize_mean_median.R Supporting R code file for binarizing the data.

S4 File. Shap_Param_Calc.R Supporting R code file for calculating shape parameters.

S5 File. rename_matrix.R Supporting R code file for renaming the dataset with appropriate gene names.

S6 File. GSE_Subset.csv This file contains the subset of the dataset GSE42408, which supports the conclusion of this article. This subset includes the data under drought conditions for pertinent genes involved in the Bayesian network analysis. The

complete dataset can be publicly accessed online from the NCBI GEO database with the accession number of GSE42408.

Acknowledgments

Not applicable.

References

1. Denchak M. Drought: Everything You Need to Know; 2019. Available from: urlhttps://www.nrdc.org/stories/drought-everything-you-need-know.

2. NOAA. Drought Public Fact Sheet; 2006. Available from: urlhttps://www.esrl.noaa.gov/gmd/obop/mlo/educationcenter/students/brochures%20and%20diagrams/noaa%20publications/Drought%20Fact%20Sheet.pdf.

3. Wang Q, Wu J, Lei T, He B, Wu Z, Liu M, et al. Temporal-spatial characteristics of severe drought events and their impact on agriculture on a global scale. Quaternary International. 2014;.

4. FAO. Disasters causing billions in agricultural losses, with drought leading the way; 2018. Available from: urlhttp://www.fao.org/news/story/en/item/1106977/icode/.

5. Lund J, Medellin-Azuara J, Durand J, Stone K. Lessons from California's 2012–2016 Drought. Journal of Water Resources Planning and Management. 2018;144(10):04018067. doi:10.1061/(asce)wr.1943-5452.0000984.

6. Sheffield J, Wood EF, Roderick ML. Little change in global drought over the past 60 years. Nature. 2012;491(7424):435–438. doi:10.1038/nature11575.

7. Dai A. Increasing drought under global warming in observations and models. Nature Climate Change. 2012;3(1):52–58. doi:10.1038/nclimate1633.

8. Spinoni J, Naumann G, Vogt JV. Pan-European seasonal trends and recent changes of drought frequency and severity. Global and Planetary Change. 2017;148:113–130. doi:10.1016/j.gloplacha.2016.11.013.

9. Wang Z, Li J, Lai C, Zeng Z, Zhong R, Chen X, et al. Does drought in China show a significant decreasing trend from 1961 to 2009? Science of The Total Environment. 2017;579:314–324. doi:10.1016/j.scitotenv.2016.11.098.

10. Cook BI, Ault TR, Smerdon JE. Unprecedented 21st century drought risk in the American Southwest and Central Plains. Science Advances. 2015;1(1). doi:10.1126/sciadv.1400082.

11. R Allen M, Kainuma M, Otto Pörtner H, Babiker M, de Kleijne K, Revi A, et al.. SPECIAL REPORT: GLOBAL WARMING OF 1.5 ° C; 2018. Available from: urlhttps://www.ipcc.ch/sr15/chapter/spm/.

12. Arneth A, Barbosa H, Benton T, Calvin K, Calvo E, Connors S, et al.. Climate Change and Land; 2019. Available from: urlhttps://www.ipcc.ch/site/assets/uploads/2019/08/4.-SPM_Approved_Microsite_FINAL.pdf.

13. Szejner P, Belmecheri S, Ehleringer JR, Monson RK. Recent increases in drought frequency cause observed multi-year drought legacies in the tree rings of semi-arid forests. Oecologia. 2019;192(1):241–259. doi:10.1007/s00442-019-04550-6.

14. Tripathi AD, Mishra R, Maurya KK, Singh RB, Wilson DW. Estimates for World Population and Global Food Availability for Global Health. The Role of Functional Food Security in Global Health. 2019; p. 3–24. doi:10.1016/b978-0-12-813148-0.00001-3.

15. Society NG. Drought; 2019. Available from: urlhttps://www.nationalgeographic.org/encyclopedia/drought/.

16. Hossain MA, Wani SH, Bhattacharjee S, Burritt DJ, Tran LSP. Drought stress tolerance in plants. Springer; 2016.

17. Gilles AF, Schinko JB, Averof M. Efficient CRISPR-mediated gene targeting and transgene replacement in the beetle Tribolium castaneum. Development. 2015;142(16):2832–2839. doi:10.1242/dev.125054.

18. Gull A, Lone AA, Wani NUI. Biotic and Abiotic Stresses in Plants. Abiotic and Biotic Stress in Plants. 2019;doi:10.5772/intechopen.85832.

19. Nguyen D, Rieu I, Mariani C, Dam NMV. How plants handle multiple stresses: hormonal interactions underlying responses to abiotic stress and insect herbivory. Plant Molecular Biology. 2016;91(6):727–740. doi:10.1007/s11103-016-0481-8.

20. Xiong L, Schumaker KS, Zhu JK. Cell Signaling during Cold, Drought, and Salt Stress. The Plant Cell. 2002;14(suppl 1). doi:10.1105/tpc.000596.

21. Shulaev V, Cortes D, Miller G, Mittler R. Metabolomics for plant stress response. Physiologia Plantarum. 2008;132(2):199–208. doi:10.1111/j.1399-3054.2007.01025.x.

22. Tiwari S, Lata C, Chauhan PS, Prasad V, Prasad M. A Functional Genomic Perspective on Drought Signalling and its Crosstalk with Phytohormone-mediated Signalling Pathways in Plants. Current Genomics. 2017;18(6). doi:10.2174/1389202918666170605083319.

23. Chaves MM, Maroco JP, Pereira JS. Understanding plant responses to drought — from genes to the whole plant. Functional Plant Biology. 2003;30(3):239. doi:10.1071/fp02076.

24. Yildirim K, Kaya Z. Gene regulation network behind drought escape, avoidance and tolerance strategies in black poplar (Populus nigra L.). Plant Physiology and Biochemistry. 2017;115:183–199. doi:10.1016/j.plaphy.2017.03.020.

25. Takahashi F, Kuromori T, Sato H, Shinozaki K. Regulatory Gene Networks in Drought Stress Responses and Resistance in Plants. Advances in Experimental Medicine and Biology Survival Strategies in Extreme Cold and Desiccation. 2018; p. 189–214. doi:10.1007/978-981-13-1244-1_11.

26. Sun Y, Dinneny JR. Q&A: How do gene regulatory networks control environmental responses in plants? BMC Biology. 2018;16(1). doi:10.1186/s12915-018-0506-7.

27. Liu S, Lv Z, Liu Y, Li L, Zhang L. Network analysis of ABA-dependent and ABA-independent drought responsive genes in Arabidopsis thaliana. Genetics and Molecular Biology. 2018;41(3):624–637. doi:10.1590/1678-4685-gmb-2017-0229.

28. Shinozaki K, Yamaguchi-Shinozaki K. Gene networks involved in drought stress response and tolerance. Journal of Experimental Botany. 2006;58(2):221–227. doi:10.1093/jxb/erl164.

29. Uno Y, Furihata T, Abe H, Yoshida R, Shinozaki K, Yamaguchi-Shinozaki K. Arabidopsis basic leucine zipper transcription factors involved in an abscisic acid-dependent signal transduction pathway under drought and high-salinity conditions. Proceedings of the National Academy of Sciences. 2000;97(21):11632–11637. doi:10.1073/pnas.190309197.

30. Fujita Y, Nakashima K, Yoshida T, Katagiri T, Kidokoro S, Kanamori N, et al. Three SnRK2 Protein Kinases are the Main Positive Regulators of Abscisic Acid Signaling in Response to Water Stress in Arabidopsis. Plant and Cell Physiology. 2009;50(12):2123–2132. doi:10.1093/pcp/pcp147.

31. Fujita Y, Fujita M, Shinozaki K, Yamaguchi-Shinozaki K. ABA-mediated transcriptional regulation in response to osmotic stress in plants. Journal of Plant Research. 2011;124(4):509–525. doi:10.1007/s10265-011-0412-3.

32. Abe H, Urao T, Ito T, Seki M, Shinozaki K, Yamaguchi-Shinozaki K. Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) Function as Transcriptional Activators in Abscisic Acid Signaling. The Plant Cell. 2002;15(1):63–78. doi:10.1105/tpc.006130.

33. Singh D, Laxmi A. Transcriptional regulation of drought response: a tortuous network of transcriptional factors. Frontiers in Plant Science. 2015;6. doi:10.3389/fpls.2015.00895.

34. Nakashima K, Ito Y, Yamaguchi-Shinozaki K. Transcriptional Regulatory Networks in Response to Abiotic Stresses in Arabidopsis and Grasses. Plant Physiology. 2009;149(1):88–95. doi:10.1104/pp.108.129791.

35. Agarwal M, Hao Y, Kapoor A, Dong CH, Fujii H, Zheng X, et al. A R2R3 Type MYB Transcription Factor Is Involved in the Cold Regulation of CBF Genes and in Acquired Freezing Tolerance. Journal of Biological Chemistry. 2006;281(49):37636–37645. doi:10.1074/jbc.m605895200.

36. Dong CH, Agarwal M, Zhang Y, Xie Q, Zhu JK. The negative regulator of plant cold responses, HOS1, is a RING E3 ligase that mediates the ubiquitination and degradation of ICE1. Proceedings of the National Academy of Sciences. 2006;103(21):8281–8286. doi:10.1073/pnas.0602874103.

37. Miura K, Jin JB, Lee J, Yoo CY, Stirm V, Miura T, et al. SIZ1-Mediated Sumoylation of ICE1 Controls CBF3/DREB1A Expression and Freezing Tolerance in Arabidopsis. The Plant Cell. 2007;19(4):1403–1414. doi:10.1105/tpc.106.048397.

38. Jiao L, Zhang Y, Wu J, Zhang H, Lu J. A Novel U-Box Protein Gene from "Zuoshanyi" Grapevine (Vitis amurensis Rupr. cv.) Involved in Cold Responsive Gene Expression in Arabidopsis thanliana. Plant Molecular Biology Reporter. 2014;33(3):557–568. doi:10.1007/s11105-014-0783-4.

39. Kim JS, Mizoi J, Yoshida T, Fujita Y, Nakajima J, Ohori T, et al. An ABRE Promoter Sequence is Involved in Osmotic Stress-Responsive Expression of the DREB2A Gene, Which Encodes a Transcription Factor Regulating Drought-Inducible Genes in Arabidopsis. Plant and Cell Physiology. 2011;52(12):2136–2146. doi:10.1093/pcp/pcr143.

40. Nakashima K, Yamaguchi-Shinozaki K. Molecular Studies on Stress-Responsive Gene Expression in Arabidopsis and Improvement of Stress Tolerance in Crop Plants by Regulon Biotechnology. Japan Agricultural Research Quarterly: JARQ. 2005;39(4):221–229. doi:10.6090/jarq.39.221.

41. Pandey GK. Mechanism of plant hormone signaling under stress. Wiley Blackwell; 2017.

42. Ensminger I, Yao-Yun Chang C, Bräutigam K. Advances in Botanical Research. Advances in Botanical Research Land Plants - Trees. 2015;74:243. doi:10.1016/s0065-2296(15)00041-5.

43. Li F, Li M, Wang P, Cox KL, Duan L, Dever JK, et al. Regulation of cotton (Gossypium hirsutum) drought responses by mitogen-activated protein (MAP) kinase cascade-mediated phosphorylation of GhWRKY59. New Phytologist. 2017;215(4):1462–1475. doi:10.1111/nph.14680.

44. Pandey SP, Somssich IE. The Role of WRKY Transcription Factors in Plant Immunity. Plant Physiology. 2009;150(4):1648–1655. doi:10.1104/pp.109.138990.

45. Eulgem T, Rushton PJ, Robatzek S, Somssich IE. The WRKY superfamily of plant transcription factors. Trends in Plant Science. 2000;5(5):199–206. doi:10.1016/s1360-1385(00)01600-9.

46. Rahaie M, Xue GP, M P. The Role of Transcription Factors in Wheat Under Different Abiotic Stresses. Abiotic Stress - Plant Responses and Applications in Agriculture. 2013;doi:10.5772/54795.

47. Bakshi M, Oelmüller R. WRKY transcription factors. Plant Signaling & Behavior. 2014;9(2). doi:10.4161/psb.27700.

48. Chen H, Lai Z, Shi J, Xiao Y, Chen Z, Xu X. Roles of arabidopsis WRKY18, WRKY40 and WRKY60 transcription factors in plant responses to abscisic acid and abiotic stress. BMC Plant Biology. 2010;10(1):281. doi:10.1186/1471-2229-10-281.

49. Lahiri A, Venkatasubramani PS, Datta A. Bayesian modeling of plant drought resistance pathway. BMC Plant Biology. 2019;19(1). doi:10.1186/s12870-019-1684-3.

50. Mintgen MAC. Genetic Analysis of Plant Responses to Combinatorial Stress in Arabidopsis thaliana Natural Variation [Master's Thesis]. Wageningen University. the Netherlands; 2014.

51. Ollas CD, Dodd IC. Physiological impacts of ABA–JA interactions under water-limitation. Plant Molecular Biology. 2016;91(6):641–650. doi:10.1007/s11103-016-0503-6.

52. Vijesh N, Chakrabarti SK, Sreekumar J. Modeling of gene regulatory networks: A review. Journal of Biomedical Science and Engineering. 2013;06(02). doi:10.4236/jbise.2013.62A027.

53. Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. Nature Reviews Molecular Cell Biology. 2008;9(10). doi:10.1038/nrm2503.

54. Vundavilli H, Datta A, Sima C, Hua J, Lopes R, Bittner M. Cryptotanshinone Induces Cell Death in Lung Cancer by Targeting Aberrant Feedback Loops. IEEE Journal of Biomedical and Health Informatics. 2020;24(8). doi:10.1109/JBHI.2019.2958042.

55. Kapoor R, Datta A, Sima C, Hua J, Lopes R, Bittner ML. A Gaussian Mixture-Model Exploiting Pathway Knowledge for Dissecting Cancer Heterogeneity. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2019; p. 1–1. doi:10.1109/TCBB.2018.2869813.

56. Lahiri A, Rastogi K, Datta A, Septiningsih EM. Bayesian Network Analysis of Lysine Biosynthesis Pathway in Rice. Manuscript submitted for publication.; 2021. Available from: urlhttps://www.preprints.org/manuscript/202104.0344/v1.

57. Vundavilli H, Tripathi LP, Datta A, Mizuguchi K. Network Modeling and Inference of Peroxisome Proliferator-Activated Receptor Pathway in High fat diet-linked Obesity. Journal of Theoretical Biology. 2021;doi:10.1101/2020.09.15.298356.

58. Needham CJ, Bradford JR, Bulpitt AJ, Westhead DR. A Primer on Learning in Bayesian Networks for Computational Biology. PLoS Computational Biology. 2007;3(8). doi:10.1371/journal.pcbi.0030129.

59. Sinoquet C. Probabilistic Graphical Models for Next-generation Genomics and Genetics. Probabilistic Graphical Models for Genetics, Genomics, and Postgenomics. 2014; p. 3–29. doi:10.1093/acprof:oso/9780198709022.003.0001.

60. Murphy K. A Brief Introduction to Graphical Models and Bayesian Networks; 1998. Available from: urlhttps://www.cs.ubc.ca/ murphyk/Bayes/bnintro.html.

61. Scanagatta M, Salmerón A, Stella F. A survey on Bayesian network structure learning from data. Progress in Artificial Intelligence. 2019;8(4):425–439. doi:10.1007/s13748-019-00194-y.

62. Zhang NL. Introduction to Bayesian Networks; 2018. Available from: urlhttps://www.cse.ust.hk/bnbook/pdf/l06.h.pdf.

63. Robert CP. Bayesian Computational Tools; 2013. Available from: urlhttps://arxiv.org/abs/1304.2048.

64. Orlof J, Bloom JO, Jonathan. Comparison of frequentist and Bayesian inference.; 2014. Available from: urlhttps://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading20.pdf.

65. Kak A. ML, MAP, and Bayesian — The Holy Trinity of Parameter Estimation and Data Prediction; 2017. Available from: urlhttps://engineering.purdue.edu/kak/Trinity.pdf.

66. Jensen PA. Project Management; 2004. Available from: urlhttps://www.me.utexas.edu/ jensen/ORMM/omie/computation/unit/project/beta.html.

67. Vundavilli H, Datta A, Sima C, Hua J, Lopes R, Bittner M. Bayesian Inference Identifies Combination Therapeutic Targets in Breast Cancer. IEEE Transactions on Biomedical Engineering. 2019;66(9):2684–2692. doi:10.1109/tbme.2019.2894980.

68. Cooper GF. The computational complexity of probabilistic inference using bayesian belief networks. Artificial Intelligence. 1990;42(2-3):393–405. doi:10.1016/0004-3702(90)90060-d.

69. Lozano-Perez T. Inference in Bayesian networks; 2006. Available from: urlhttps://courses.csail.mit.edu/6.034s/handouts/ spring12/chapter14_mod_b.pdf.

70. Chiarandini M, Russell S, Norvig P. Inference in Bayesian Networks; 2012. Available from: urlhttps://imada.sdu.dk/ marco/Teaching/AY2011-2012/DM828/Slides/dm828-lec7.pdf.

71. Menon A. Rejection Sampling and Likelihood Weighting; 2012.

72. Lowry DB, Logan TL, Santuari L, Hardtke CS, Richards JH, Derose-Wilson LJ, et al. Expression Quantitative Trait Locus Mapping across Water Availability Environments Reveals Contrasting Associations with Genomic Features in Arabidopsis. The Plant Cell. 2013;25(9):3266–3279. doi:10.1105/tpc.113.115352.

73. National Library of Medicine. National Center for Biotechnology Information; 1988. Available from: urlhttps://www.ncbi.nlm.nih.gov/.

74. R Core Team. R: A Language and Environment for Statistical Computing; 2013. Available from: urlhttp://www.R-project.org/.

75. Scutari M. Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software. 2010;35(3):1–22. doi:10.18637/jss.v035.i03.

76. Dombrecht B, Xue GP, Sprague SJ, Kirkegaard JA, Ross JJ, Reid JB, et al. MYC2 Differentially Modulates Diverse Jasmonate-Dependent Functions in Arabidopsis. The Plant Cell. 2007;19(7):2225–2245. doi:10.1105/tpc.106.048017.

77. Wu Y, Deng Z, Lai J, Zhang Y, Yang C, Yin B, et al. Dual function of Arabidopsis ATAF1 in abiotic and biotic stress responses. Cell Research. 2009;19(11):1279–1290. doi:10.1038/cr.2009.108.

78. Seo KI, Lee JH, Nezames CD, Zhong S, Song E, Byun MO, et al. ABD1 Is an Arabidopsis DCAF Substrate Receptor for CUL4-DDB1–Based E3 Ligases That Acts as a Negative Regulator of Abscisic Acid Signaling. The Plant Cell. 2014;26(2):695–711. doi:10.1105/tpc.113.119974.

79. Kazan K, Manners JM. MYC2: The Master in Action. Molecular Plant. 2013;6(3):686–703. doi:10.1093/mp/sss128.

80. Li Y, Yang X, Li X. Role of jasmonate signaling pathway in resistance to dehydration stress in Arabidopsis. Acta Physiologiae Plantarum. 2019;41(6). doi:10.1007/s11738-019-2897-7.

81. Liu N, Avramova Z. Molecular mechanism of the priming by jasmonic acid of specific dehydration stress response genes in Arabidopsis. Epigenetics & Chromatin. 2016;9(1). doi:10.1186/s13072-016-0057-5.

82. Lu PL, Chen NZ, An R, Su Z, Qi BS, Ren F, et al. A novel drought-inducible gene, ATAF1, encodes a NAC family protein that negatively regulates the expression of stress-responsive genes in Arabidopsis. Plant Molecular Biology. 2006;63(2):289–305. doi:10.1007/s11103-006-9089-8.

83. Scholl R, Anderson M. Arabidopsis Biological Resource Center. Plant Molecular Biology Reporter. 1994;12(3):242–244. doi:10.1007/bf02668747.

84. Mu C, Zhou L, Shan L, Li F, Li Z. Phosphatase GhDs PTP 3a interacts with annexin protein Gh ANN 8b to reversely regulate salt tolerance in cotton (Gossypium spp.). New Phytologist. 2019;223(4):1856–1872. doi:10.1111/nph.15850.

Figures

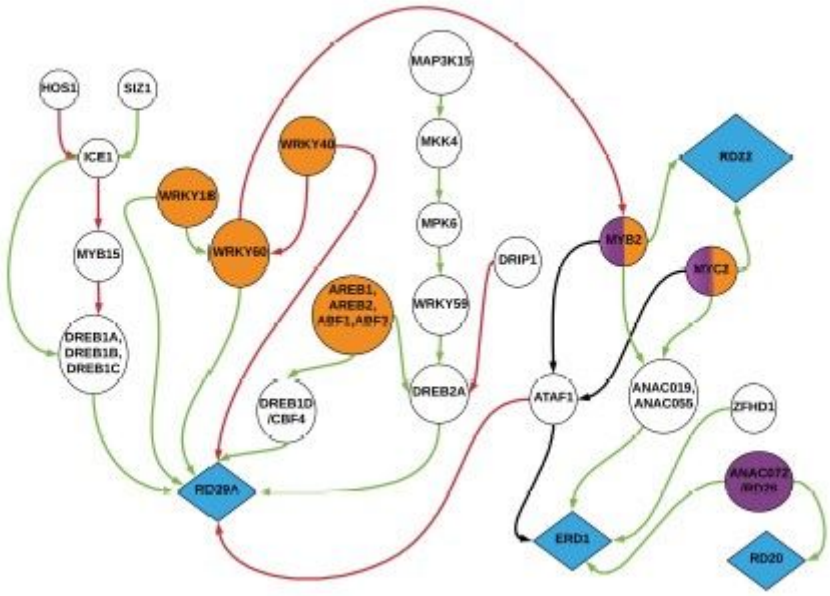

Figure 1

Drought signaling pathways in Arabidopsis. The orange circular nodes represent elements directly regulated by ABA whereas the purple nodes represent elements regulated by JA. The two nodes colored with a mix of orange and purple represent elements regulated by both JA and ABA pathways (Crosstalk). The blue diamonds represent drought responsive reporter genes. The plain circular nodes with no colors represent the transcription factors, genes and proteins involved in the regulation of drought responsive reporter genes in an ABA independent manner. The green and red arrows represent positive and negative regulation. The arrows going into and out of ATAF1 are marked black to indicate that the nature of regulation is not known at this time.

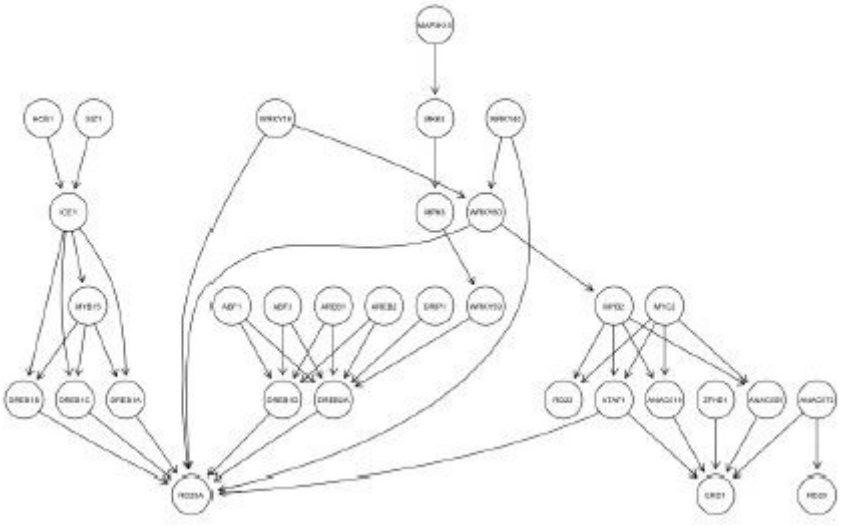

Figure 2

Bayesian Network Model of Drought Signaling Pathway. Every circular node represents a biological element in the drought signaling pathway. Every edge or black arrow represents the causal biological relationship between the nodes. Associated with every node is a θ parameter that represents the local probability distribution of the node.

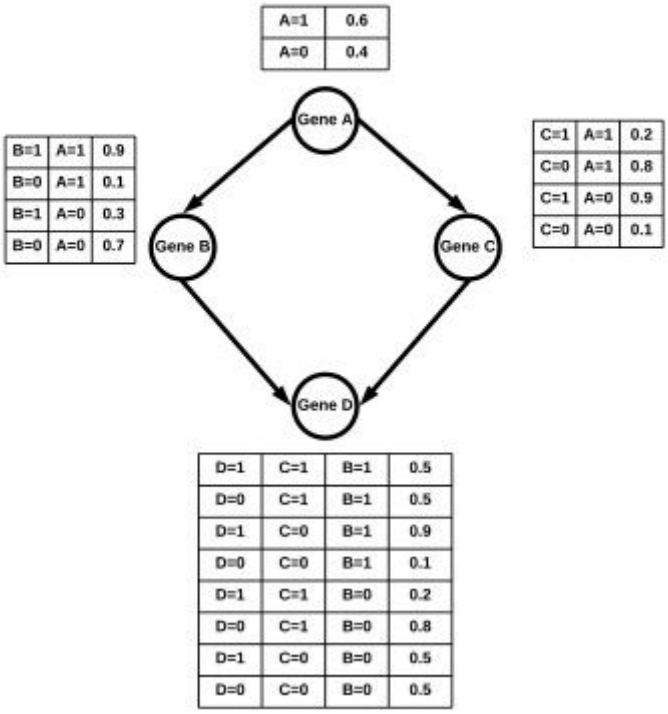

Figure 3

Example BN with LPDs. Gene A positively regulates Gene B and negatively regulates Gene C. Gene B positively regulates Gene D and Gene C negatively regulates Gene D.

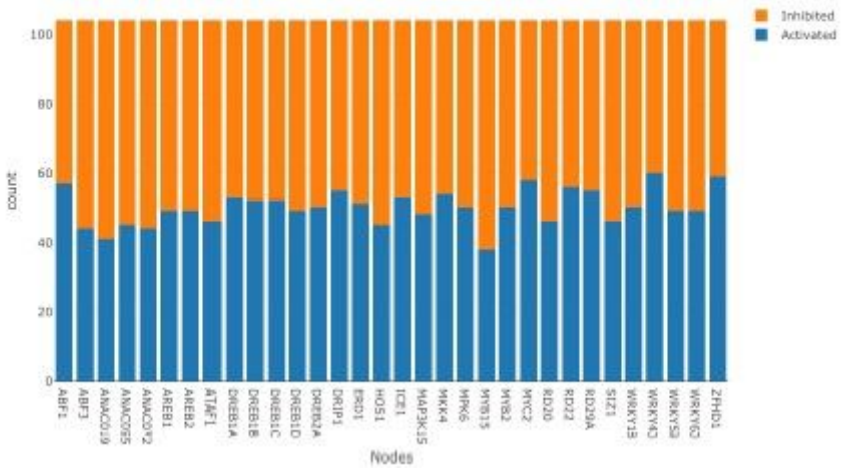

Figure 4

Activation vs Inhibition plot. This figure represents the data after it has been normalized and then binarized. There are a total of 104 data points per node. The blue part of each bar represents activation counts whereas the orange part represents the inhibition counts.

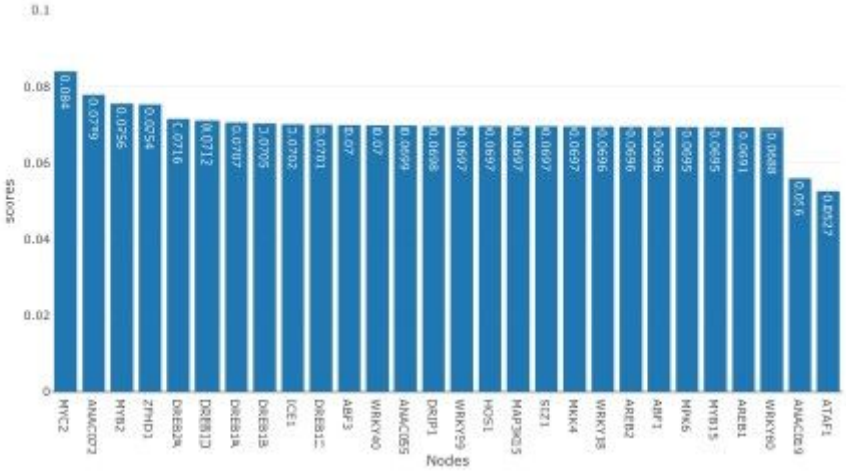

Figure 5

Activation Scores for non-reporter gene nodes. Associated with each node is a blue bar which represents the score for activating that node.

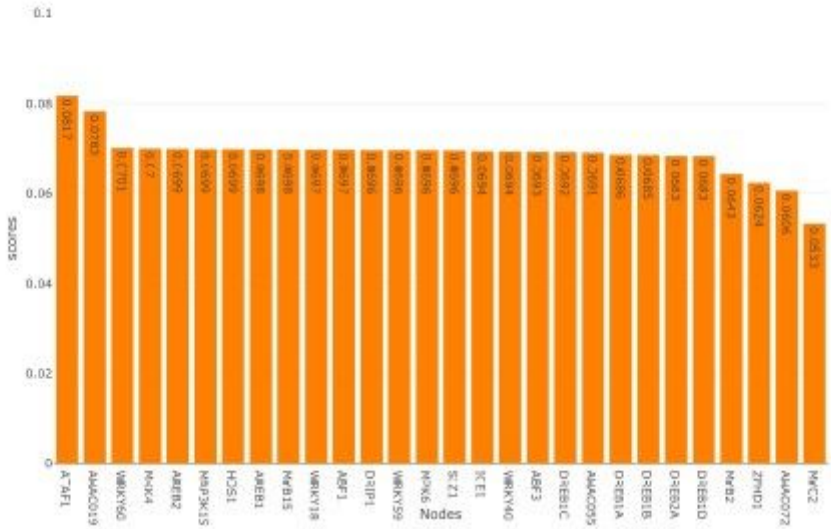

Figure 6

Inhibition Scores for non-reporter gene nodes. Associated with each node is an orange bar which represents the score for activating that node.

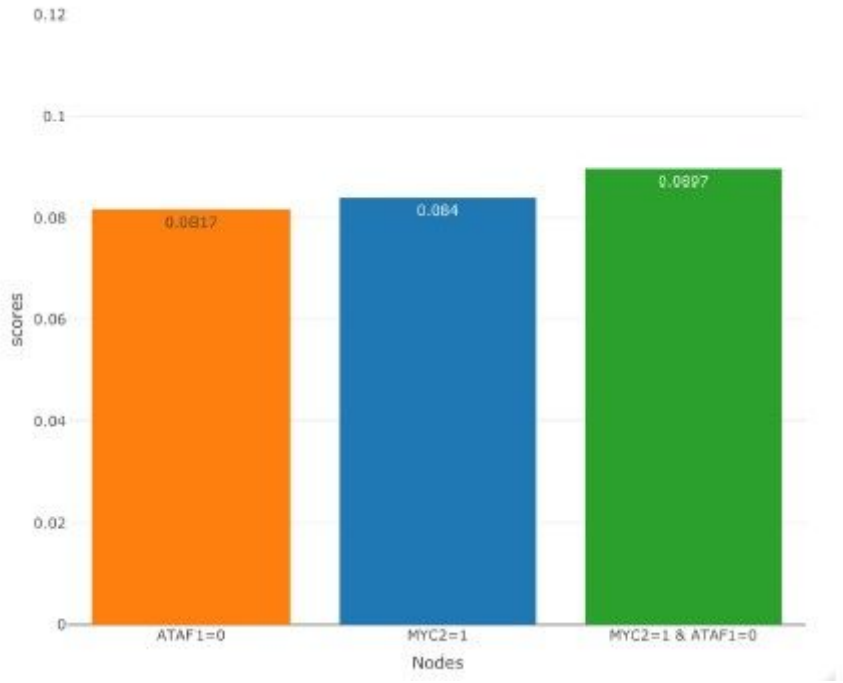

Figure 7

Comparing the scores of multi-node and single node intervention under optimal response case. Simultaneous (multi-node) intervention on MYC2 and ATAF1 has a slightly higher score than single node intervention.

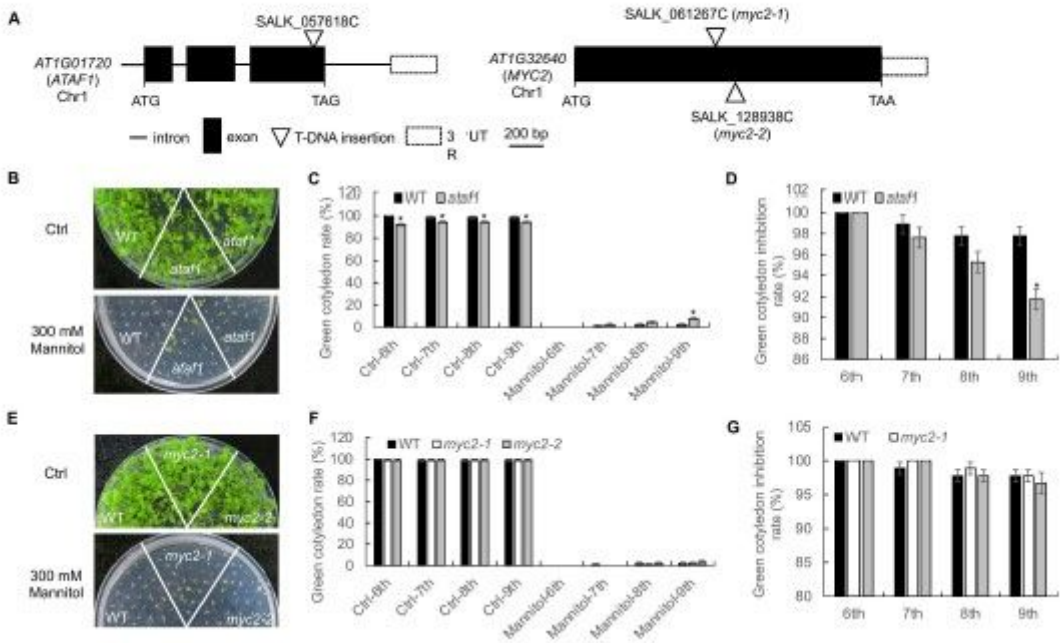

Figure 8

Results from validation experiments. A. The scheme of the ATAF1 and MYC2 genomic DNA and T-DNA insertion. The panel is a schematic illustration of the ATAF1 and MYC2 genomic DNA with exons (solid box), intron (lines) and 3' untranslated region (open box). The position of T-DNA insertion of ataf1 (SALK 057618C), myc2 (SALK 061267C, SALK 128938C)was labeled. B. The ataf1 mutant is more resistant to mannitol treatment. Wild-type (WT) Col-0 and ataf1 mutant seeds were germinated on 1/2 MS medium with or without 300 mM mannitol. 30 seeds per genotype were used for each replicate. The photos were taken four-week post-germination. C. Quantification of cotyledon greening on plates corresponding to B. Seedlings with green cotyledon expansion were counted at 6-9 days post-germination. Data are shown as means ± SD (standard deviation) from three independent replicates (n=3, *, $p<0.05$, Student's t-test). D. Quantification of cotyledon greening inhibition rate on plates corresponding to B. Seedlings with green cotyledon expansion were counted at 6-9 days post-germination. Data are shown as means ± SD from three independent replicates (n=3, *, $p<0.05$, Student's t-test). E. Growth of WT and myc2 mutants on MS plates. WT and myc2 mutant seeds were germinated on 1/2 MS medium with or without 300 mM mannitol. 30 seeds per genotype were used for each replicate. The photos were taken four-week post-germination. F. Quantification of cotyledon greening on plates corresponding to E. Seedlings with green cotyledon expansion were counted at 6-9 days post-germination. Data are shown as means ± SD from three independent replicates (n=3, no statistical significance with Student's t-test). G. Quantification of cotyledon greening on plates corresponding to E. Seedlings with green cotyledon expansion were counted at 6-9 days post-germination. Data are shown as means ± SD from three independent replicates (n=3, no statistical significance with Student's t-test).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MainScript.r
- minmaxnormalize.r
- binarizemeanmedian.r
- ShapParamCalc.r
- renamematrix.r
- GSESubset.csv