APPLICATIONS OF PROBABILISTIC GRAPHICAL MODELS IN GENOMIC NETWORKS FOR AGRICULTURE

A Dissertation

by

ADITYA LAHIRI

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Aniruddha Datta
Committee Members,	Shankar P. Bhattacharyya
	Yang Shen
	Ping He
Head of Department,	Miroslav M. Begovic

May 2022

Major Subject: Electrical Engineering

ABSTRACT

Agricultural productivity is severely limited by environmental stresses that affect plants. Environmental stresses can be classified as abiotic or biotic. This study focuses on drought and saline stress, the two significant abiotic stresses causing crop loss worldwide. Crop loss due to drought and saline stress are major factors that threaten global food security. This problem is exacerbated by the growing world population, which is expected to rise by 2 billion in the next thirty years. Fortunately, plants have internal mechanisms to defend against environmental stresses. These mechanisms are deployed through complex networks of molecules known as signaling pathways. Environmental stress stimuli can trigger signaling pathways that activate or inhibit downstream genes to implement defensive measures and restore homeostasis. Signaling pathways are not only limited in their capability to defend against stresses but are also responsible for mediating other activities, including protein synthesis, cell death, and differentiation. Thus understanding the signaling pathways in plants is key to developing plants that can defend against environmental stresses and are nutritionally valuable.

We studied the drought signaling pathways in Arabidopsis to identify the genetic regulators of drought-responsive genes. Additionally, we examined the lysine biosynthesis pathway in rice under normal and saline stress conditions. Lysine is an essential amino acid present in the lowest quantity compared to all the other amino acids in rice. Amino acids are the building block of proteins and play a crucial role in maintaining the human body's healthy functioning. Thus, increasing the lysine content in rice will help improve global health. We modeled both the drought signaling and lysine synthesis pathways using Bayesian networks. We chose Bayesian networks as they allow us to integrate pathway information from literature with experimental data. Using Bayesian networks, we identified that ATAF1 is a negative regulator of drought and DAPF is the most potent regulator of lysine. These regulators can be targeted using genetic intervention methods such as CRISPR-CAS9 to make plants robust against drought and increase lysine content in rice. Our work with drought signaling pathways was validated through wet-lab experiments.

DEDICATION

To my mother (Nabanita Lahiri), father (Debkumar Lahiri), brother (Anirban Lahiri), sister (Lucia Mattiangeli) and my nephews (Matteo and Leonardo Lahiri)

ACKNOWLEDGMENTS

I consider myself fortunate to have been mentored by Professor Aniruddha Datta. Armed with no discernible career goals or research experience, I met Professor Datta in his office in August 2016 to ask for a research position at his lab. He was generous and took a chance on me and gave me the opportunity to attend his weekly lab meetings, and the rest is history. I am deeply thankful for his patience, motivational support, and trust in me to work independently toward my research objectives. I have grown incredibly both as a researcher and a person under his mentorship, and it has been an honor and privilege of my life to have him as my PhD Advisor.

I would also like to take this opportunity to express my sincerest gratitude and thanks to Professor Yang Shen, Professor Shankar Bhattacharyya, and Professor Ping He for their mentorship, feedback, discussions, and serving on my dissertation committee.

Most importantly, I would like to thank my parents, who trusted me at age 18 and decided to send me a continent and an ocean away from home to pursue my dreams. Finally, I would like to recognize my brother, who has been an absolute source of inspiration and motivation all my life; I wouldn't be here in this foreign land had he not come here.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor Aniruddha Datta (advisor), Professor Shankar Bhattacharyya and Professor Yang Shen of the Department of Electrical & Computer Engineering and Professor Ping He of the Department of Biochemistry & Biophysics.

The validation work done in section 2 was carried out by Ms. Lin Zhou of the Department of Biochemistry & Biophysics. The lysine biosynthesis pathway in section 3 was constructed in part by Ms. Khushboo Rastogi of the Department of Soil and Crop Sciences.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

This work was supported in part by the TEES-AgriLife Center for Bioinformatics and Genomic Systems Engineering (CBGSE) startup funds, the Texas A&M X-Grant Program, and in part by the National Science Foundation under Grant ECCS-1609236 (to Aniruddha Datta). This work was partially supported by the USDA NIFA Grant 2020-67013-31615 (to Ping He). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

NOMENCLATURE

IPCC	Intergovernmental Panel on Climate Change
bZIP	basic-domain leucine zipper
ABA	Abscisic acid
AREB/ABF	ABA-responsive element-binding protein/factor
JA	Jasmonic Acid
MAPK	Mitogen-Activated Protein Kinase
MLE	Maximum Likelihood Estimate
LKR	Lysine ketoglutarate reductase
SDH	Saccharopine dehydrogenase
DHPS	Dihydrodipicolinate synthase
AK	Aspartate kinase
GRN	Gene regulatory network
GMO	Genetically modified organisms
MSU	Michigan State University
TF	Transcription factor
BN	Bayesian network
PGM	Probabilistic graphical model
LPD	Local probability distribution
i.i.d	Independent and identically distributed
LW	Likelihood weighting

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION AND LITERATURE REVIEW *

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, developing drought-resistant crops to ensure food security is essential. In section 2 of this dissertation, 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 that 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.

In section 3, we focus on modeling the lysine biosynthesis pathway in rice. Lysine is the first limiting essential amino acid in rice because it is present in the lowest quantity compared to all the other amino acids. Amino acids are the building block of proteins and play an essential role in maintaining the human body's healthy functioning. Rice is a staple food for more than half of the global population; thus, increasing the lysine content in rice will help improve global health. To this end, we study the lysine biosynthesis pathway in rice (Oryza sativa) to identify the regulators of the lysine reporter gene LYSA (LOC_Os02g24354). Genetically intervening

*Parts of this section are reprinted with permission from Lahiri A, Zhou L, He P, Datta A (2021) Detecting drought regulators using stochastic inference in Bayesian networks. PLoS ONE 16(8): e0255486. https://doi.org/10.1371/journal.pone.0255486 and Lahiri A, Rastogi K, Datta A, Septiningsih EM. Bayesian Network Analysis of Lysine Biosynthesis Pathway in Rice. Inventions. 2021; 6(2):37. https://doi.org/10.3390/inventions6020037

at the regulators has the potential to increase the overall lysine content in rice. We model the lysine biosynthesis pathway in rice seedlings under normal, and saline (NaCl) stress conditions using Bayesian networks. We estimate the model parameters using experimental data and have identified the gene DAPF(LOC_Os12g37960) as a positive regulator of the lysine reporter gene LYSA under normal and saline stress conditions. Based on this analysis, we conclude that the gene DAPF is a potent candidate for genetic intervention. Upregulating DAPF using methods such as CRISPR-Cas9 gene editing strategy has the potential to upregulate the lysine reporter gene LYSA and increase the overall lysine content in rice.

2. DETECTING DROUGHT REGULATORS USING STOCHASTIC INFERENCE IN BAYESIAN NETWORKS *

2.1 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 the droughts which occurred from 2012 to 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, 7, 8, 9, 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,

*Parts of this section are reprinted with permission from Lahiri A, Zhou L, He P, Datta A (2021) Detecting drought regulators using stochastic inference in Bayesian networks. PLoS ONE 16(8): e0255486. https://doi.org/10.1371/journal.pone.0255486

frequent and more intense droughts in the future only serve to exacerbate this problem [14]. Thus, it is of paramount importance to develop crops that are robust against drought.

While the risk of imminent droughts has motivated the scientific communities' efforts in developing drought resilient plants, it has also led plants to develop and evolve their internal defense mechanisms to protect against droughts. 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 plant breeding methods 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 section 2.4, 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].

2.2 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, 23, 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, 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.

2.3 Drought Signaling Networks

In this section of the dissertation, 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]. ABA acts a sensor of drought in plants. Under drought conditions, the ABA levels increase rapidly in plants which allows them to subsequently respond by closing their 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 and another plant phytohormone Jasmonic Acid (JA) 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* [34, 33]. *DREB1A*, *DREB1B*, and *DREB1C* are negatively regulated by a transcription factor *MYB15* and positively regulated by another transcription factor, *ICE1* [35, 36, 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 independently 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, 41, 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 analysis 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, 45, 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 *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 fac-

tors (*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 Figure 2.1. In section 2.4, 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 Figure 2.1).

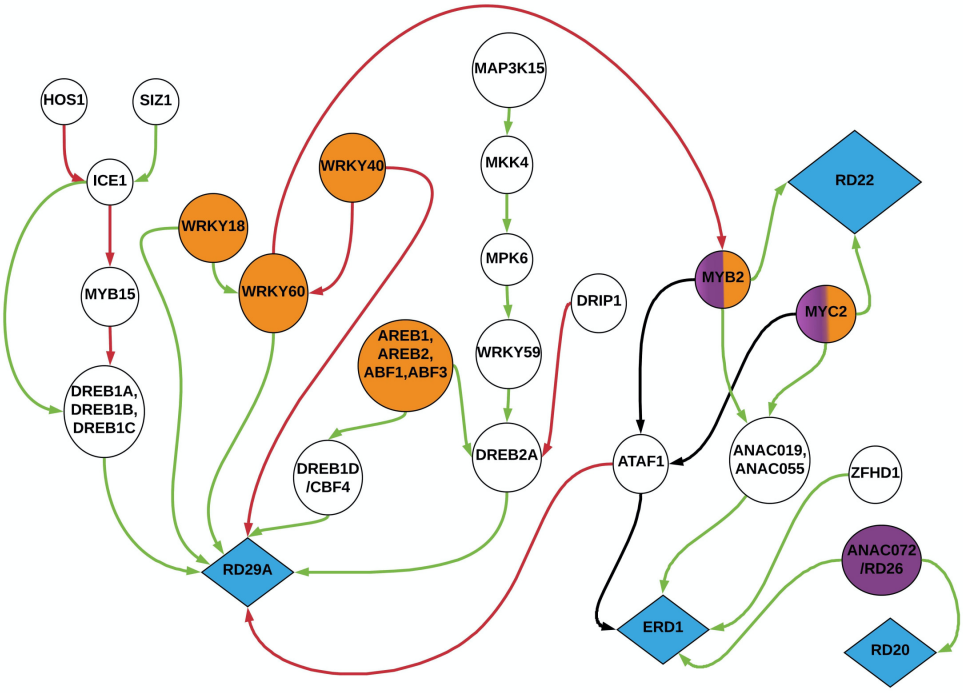

Figure 2.1: **Drought signaling pathways in Arabidopsis.** The orange circular nodes represent elements directly regulated by ABA whereas the purple nodes represent elements regulated by JA. The two nodes colored with a mix of orange and purple represent elements regulated by both JA and ABA pathways (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. (Reprinted from [128])

2.4 Materials and Methods

2.4.1 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, 53, 54, 55, 56, 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 section, our objective is to create a BN model of the drought signaling pathways outlined in Figure 2.1 and use this model to determine which transcription factor, protein or gene is the best regulator of drought responsive reporter genes (blue diamonds in Figure 2.1). The predictions made by the model can help us identify potential targets for genetic intervention techniques like CRISPR-Cas9 to create drought resistant crops.

Figure 2.2 represents the BN model of the signaling pathways shown in Figure 2.1. 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 that can assume 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 conditional probability distribution is given by $\theta_{ICE1|HOS1,SIZ1}$. Henceforth, we will refer to local conditional or marginal probability distribution as just local probability distributions (LPD). We learn these LPDs from experimental biological 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.

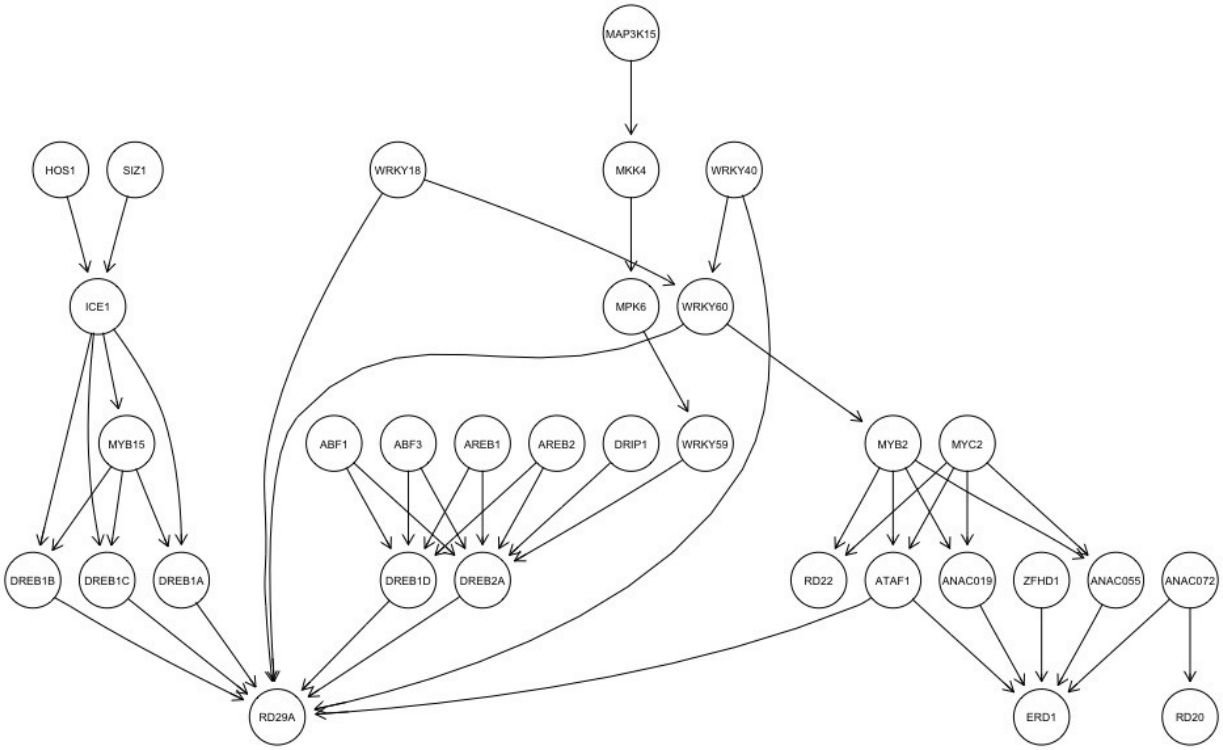

Figure 2.2: **Bayesian Network Model of Drought Signaling Pathway.** Every circular node represents a biological element in the drought signaling pathway. Every edge or black arrow represents the causal biological relationship between the nodes. Associated with every node is a θ parameter that represents the local probability distribution of the node.(Reprinted from [128])

2.4.2 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 in the prior sections 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 Figure 2.2 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 the prior distribution is subjective and there exists 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 section, we use a Bayesian approach to estimate the local probability distributions for the BN model outlined in Figure 2.2. We assumed that the nodes are binary random variables, which implies that for any node **X** in the BN, **X**=1

(success) when the node is activated and **X**=0 (failure) when the node is inhibited. Then for a single observation for any node **X** in the BN can be modeled as a Bernoulli random variable.

Let us suppose that we have a BN model with **N** nodes. Then the probability with which any node **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 **X**, we observe *k* instances when the node attains a state of 1, then the likelihood for node **X** is given by:

$$P(X|P_a(X), \theta_X) \sim Binomial(n, \theta_X) \tag{2.1}$$

$$Binomial(n, \theta_X) = \frac{n!}{k!(n-k)!}\theta_X^k(1-\theta_X)^{n-k} \tag{2.2}$$

P_a(X) in Equation(2.1) refers to the parents, if any, of node **X**. Since we are using a Bayesian approach to estimate the LPD of Node **X**, we need to select a prior distribution on the node **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:

$$\theta_X \sim Beta(\alpha_X, \beta_X) \tag{2.3}$$

$$P(\theta_X|X) \sim Beta(\alpha_X^{'}, \beta_X^{'}) \tag{2.4}$$

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

In equation (2.3), α_X and β_X represent the shape parameters of the Beta distribution, and in equation (2.4) 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 distri-

bution 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 (2.4) 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 (2.5).

$$\theta_X \simeq E[\theta_X|X] = \frac{\alpha_X^{'}}{\alpha_X^{'} + \beta_X^{'}} \tag{2.5}$$

Once these parameters are learned the BN is complete as we have both the DAG and the set of conditional probabilities. In section 2.4.3, we study the effect on drought responsive genes for intervening (activating/ inhibiting) at various nodes, then summarize our findings in the results section.

2.4.3 Sampling Based Inference in Bayesian Networks

In this section, we are interested in using the BN model in determining 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 ef-

ficient 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 use an exact inference technique to calculate the LPDs in our BN model. However, exact inference techniques will be computationally expensive to implement as our network is multiply connected, i.e., there are at least two nodes in our BN model connected by more than one path. For instance, we can see that the nodes *DREB1A* and *ICE1* are directly connected and are also connected through *MYB15*, hence making our 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 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:

$$Pr(X_Q = x_q | X_E = x_e) \simeq \lim_{M \to \infty} \frac{\sum_i^M \mathbb{1}[x_q^{(i)} = x_q] Pr(X_E = x_e | (P_a(X_E))^{(i)})}{\sum_i^M Pr(X_E = x_e | (P_a(X_E))^{(i)})} \tag{2.6}$$

The proof for Equation (2.6) is not trivial and is presented in a paper by Menon [71]. A pseudo code for estimating the conditional probabilities using LW is presented in algorithm 1.We will now demonstrate LW on an example BN.

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 $\in$ 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] * $P(X_i=x \mid P_a(X_i))$

    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

Figure 2.3 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 Figure 2.3. 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.

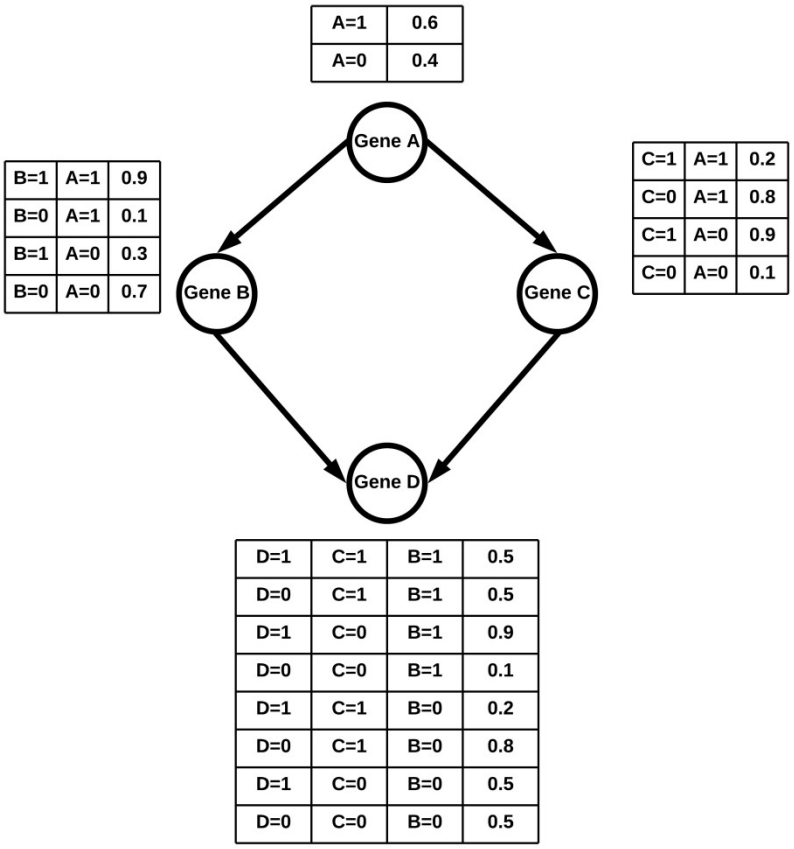

Figure 2.3: **Example BN with LPDs.** Gene A positively regulates Gene B and negatively regulates Gene C. Gene B positively regulates Gene D and Gene C negatively regulates Gene D. (Reprinted from [128])

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 2.1: **Sample Data from Example Bayesian Network.** (Reprinted from [128])

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

2.5 Dataset and Simulation

To estimate the LPDs for the nodes in the BN model, we needed gene expression data (e.g., microarray, RNA-Seq, eQTL, etc.) for Arabidopsis under drought conditions. We searched the NCBI GEO database and selected the dataset GSE42408 [72, 73]. We chose this dataset as it had gene expression data for the genes of interest in our BN model from 104 recombinant inbred lines of Arabidopsis under drought conditions. Furthermore, this dataset had the most number of data points per gene compared to other datasets found during the search of the NCBI GEO database, which also led to its selection for our analysis. This dataset contains 104 eQTL (expression quantitative trait loci) 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. 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]. All the code and data files are also made available publicly at the following GitHub repository: https://github.com/adilahiri/Drought_Regulators.

2.6 Results

Figure 2.4 displays the dataset GSE42408 after it was normalized and binarized. Each bar in Figure 2.4 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 (2.7) below.

$$\begin{aligned} &Score(Evidence = \{0,1\}) = \\ &Pr(\textbf{\textit{RD29A}} = 0 | Evidence = \{0,1\}) \\ &Pr(\textbf{\textit{RD22}} = 0 | Evidence = \{0,1\}) \\ &Pr(\textbf{\textit{RD20}} = 0 | Evidence = \{0,1\}) \\ &Pr(\textbf{\textit{ERD1}} = 1 | Evidence = \{0,1\}). \end{aligned} \tag{2.7}$$

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 suitable candidate for intervention. In figures 2.5 and 2.6, 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 Figures 2.5, whereas in Figure 2.6, they are inhibited. From Figure 2.5, it is clear that when *MYC2* is activated, it results

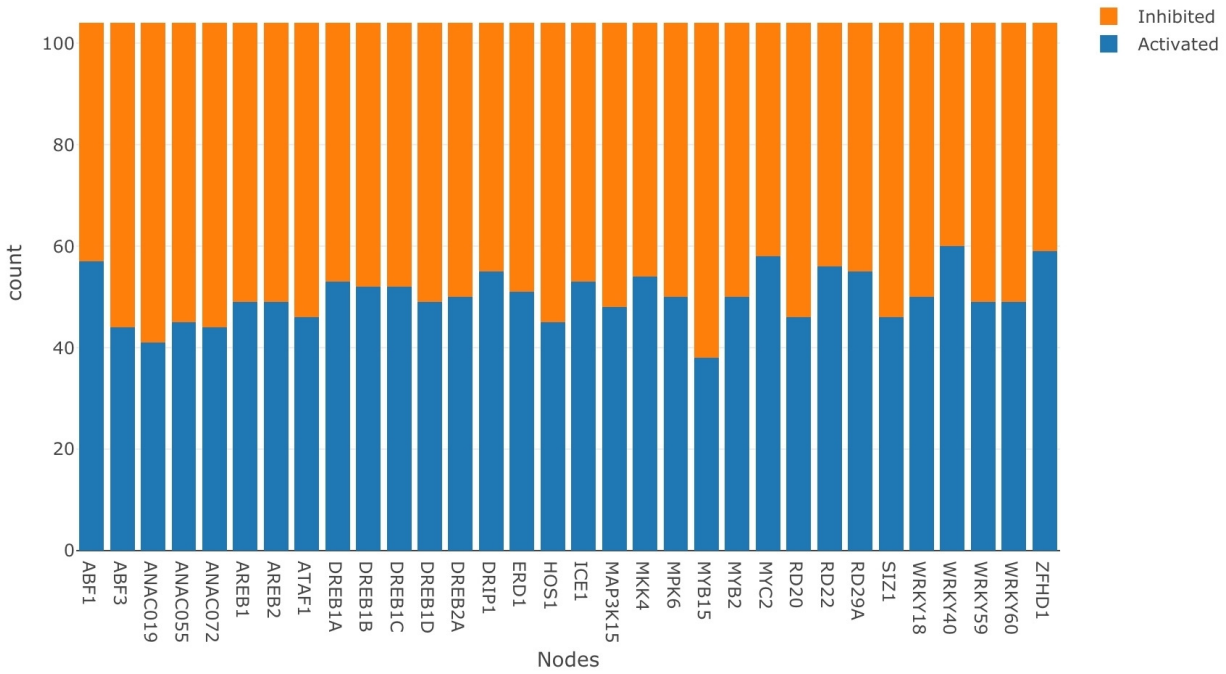

Figure 2.4: **Activation vs Inhibition plot.** This figure represents the data after it has been normalized and then binarized. There are a total of 104 data points per node. The blue part of each bar represents activation counts whereas the orange part represents the inhibition counts.(Reprinted from [128])

in the highest score, whereas *ANAC072* and *ZFHD1* have the second and third highest scores, respectively. On the other hand, in Figure 2.6, *ATAF1* has the highest score for inhibition, followed by *ANAC019*. Our analysis shows that activating *MYC2* or inhibiting *ATAF1* maximizes the scores under single node intervention. Thus these are the best strategies to activate *ERD1* and inhibit *RD29A*, *RD20*, and *RD22*. We observe that the score for *MYC2* is the lowest when it is inhibited (Figure 2.6) and the score for *ATAF1* is lowest when it is activated (Figure 2.5), this makes logical sense for the analysis.

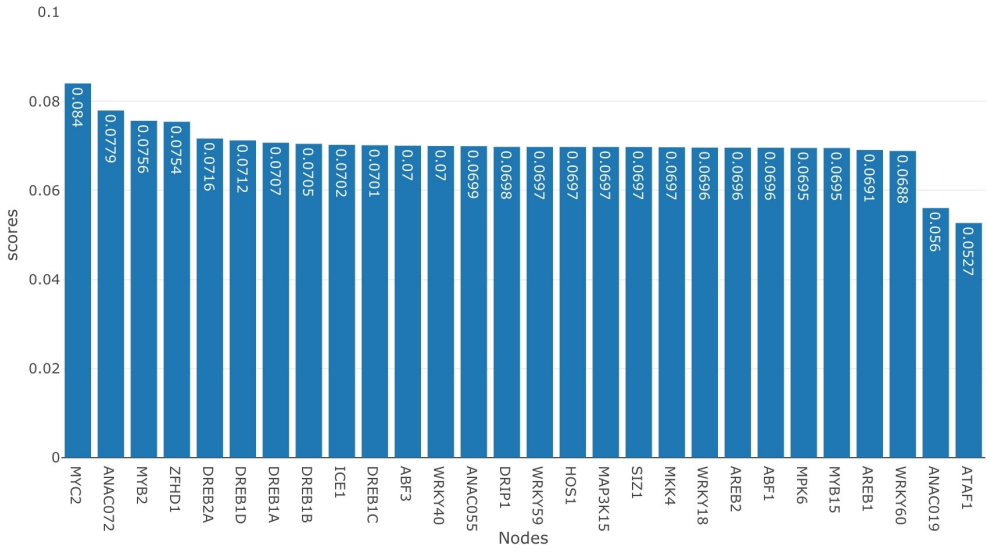

Figure 2.5: **Activation Scores for non-reporter gene nodes.** Associated with each node is a blue bar which represents the score for activating that node.(Reprinted from [128])

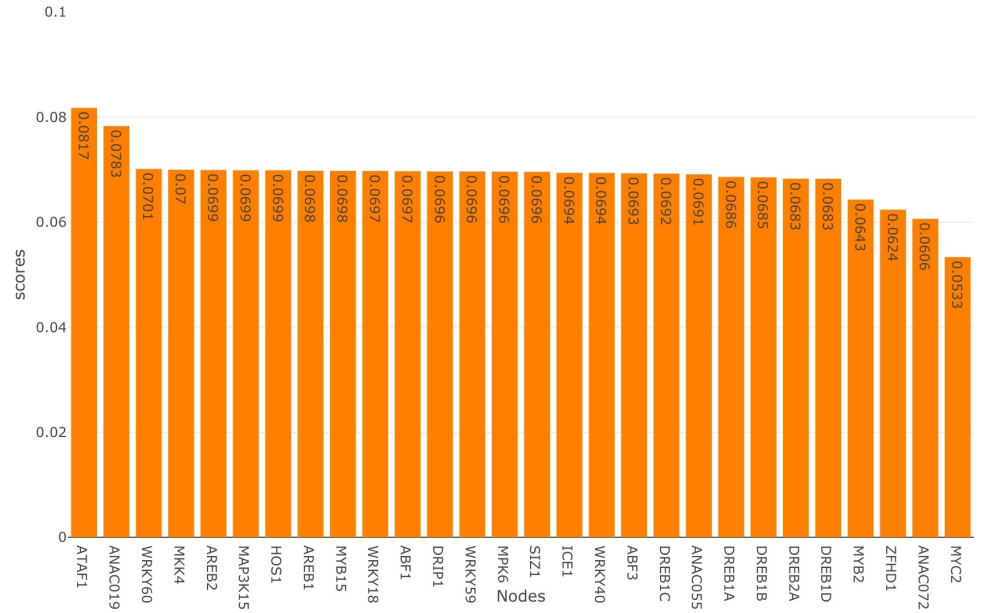

Figure 2.6: **Inhibition Scores for non-reporter gene nodes.** Associated with each node is an orange bar which represents the score for activating that node.(Reprinted from [128])

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 Figure 2.7, 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, 79, 80]. A study found *MYC2* to have no significant regulatory effect on *RD29A* in Arabidopsis [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.

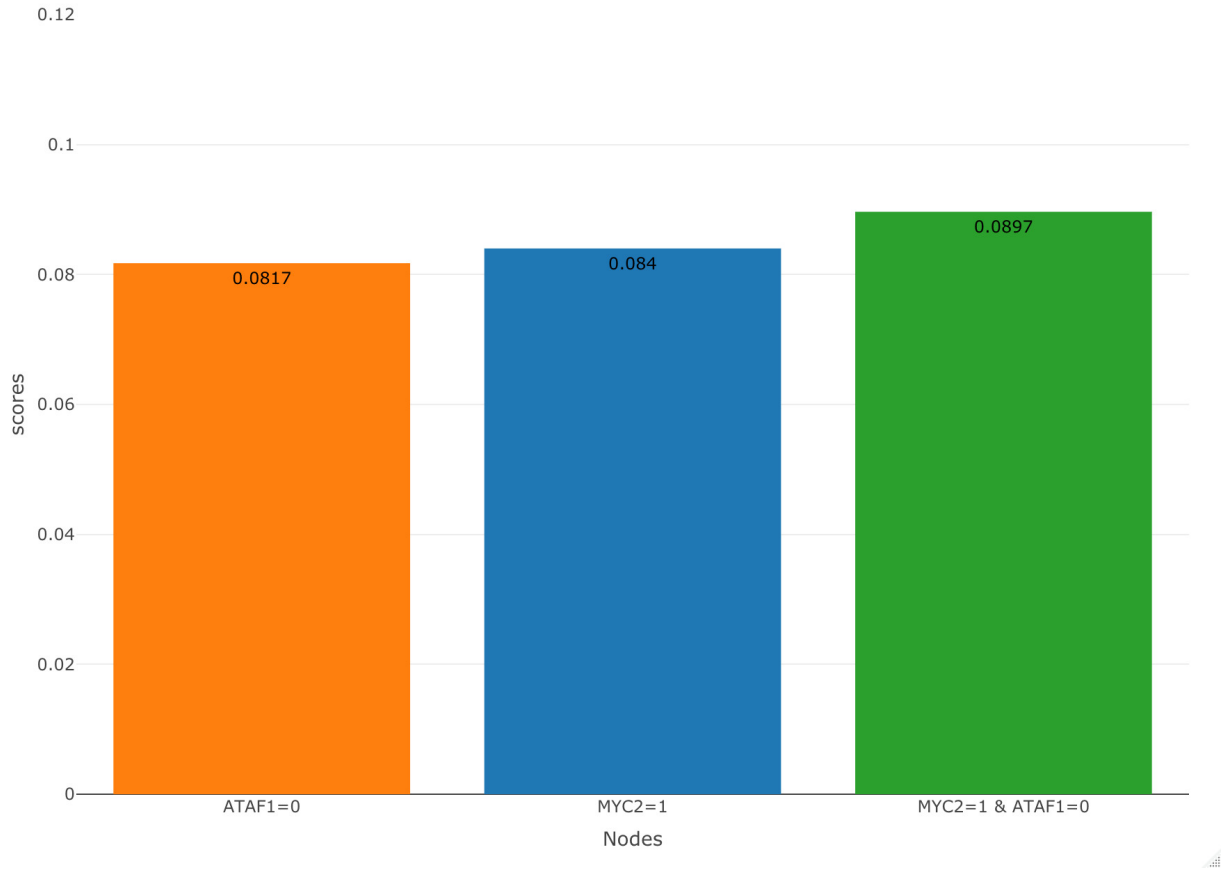

Figure 2.7: **Comparing the scores of multi-node and single node intervention under optimal response case.** Simultaneous (multi-node) intervention on *MYC2* and *ATAF1* has a slightly higher score than single node intervention.(Reprinted from [128])

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 (Figure 2.8 A). 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 (Figure 2.8 B).

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 (Figure 2.8 B) and higher green cotyledon rate (Figure 2.8 C) 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 (Figure 2.8 D). We also germinated WT and *myc2* mutants on the MS medium with or without 300 mM mannitol treatment (Figure 2.8 E). However, there is no significant difference in the green cotyledon rate between WT and *myc2* mutants with or without mannitol treatment (Figure 2.8 F). Similarly, the green cotyledon inhibition rate between WT and *myc2* mutants also did not show a significant difference (Figure 2.8 G). 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.

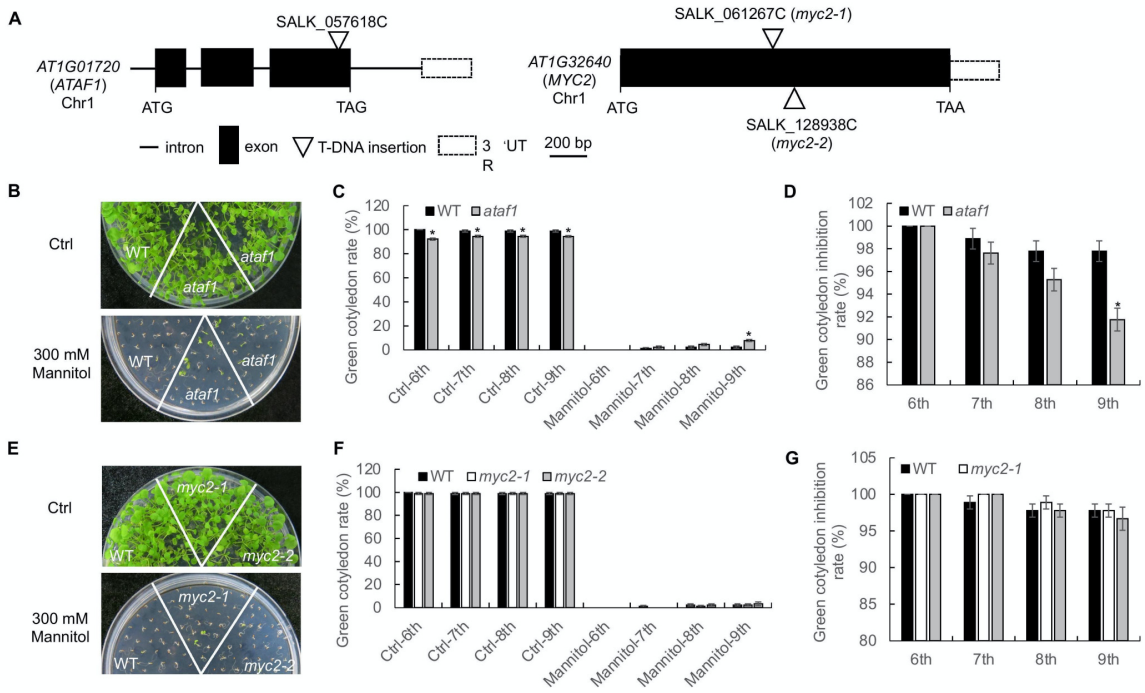

Figure 2.8: **Results from validation experiments.** **A.**The scheme of the *ATAF1* and *MYC2* genomic DNA and T-DNA insertion. The panel is a schematic illustration of the *ATAF1* and *MYC2* genomic DNA with exons (solid box), intron (lines) and 3' untranslated region (open box). The position of T-DNA 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).(Reprinted from [128])

2.7 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 $\pm$ SD from three independent repeats (n=3, *, $p<0.05$, Student's t-test). The photos were taken four-weeks post-germination.

3. BAYESIAN NETWORK ANALYSIS OF LYSINE BIOSYNTHESIS PATHWAY IN RICE*

3.1 Introduction

Proteins are one of the primary building blocks of all life on Earth and are present in every cell in the human body. Proteins are a crucial macro-nutrient in the human diet; they help build and repair cells and are essential for the human body's growth and development[85]. Proteins are comprised of long chains of amino acids; once the human body digests the proteins, they are broken down into their constituent amino acids [86]. There are twenty naturally existing amino acids that encode the 20,000 (approximate) unique proteins in the human body [87]. Among these amino acids, nine are classified as essential, and eleven are classified as non-essential [86, 87]. Amino acids produced by the human body are considered non-essential, whereas the amino acids that cannot be synthesized by the body are considered essential [87]. Essential amino acids include phenylalanine, valine, tryptophan, threonine, isoleucine, methionine, histidine, leucine, and lysine [88]. Since essential amino acids cannot be synthesized, they need to be introduced to the human body through diets rich in *complete proteins*. A protein food source is considered a complete protein if it contains all the essential amino acids [89]. Typically animal-based proteins are considered sources of complete protein. On the other hand, plant-based proteins are considered incomplete as they do not contain all the essential amino acids [89, 90].

According to the National Academy of Medicine, the recommended dietary allowance (RDA) of protein intake is 0.8 g/kg/day [91, 92]. A diet deficient in protein can cause edema, thinning of hair, and muscle mass loss in adults [93]. Though protein deficiency is rare in the developed world, it is still prevalent in impoverished and underdeveloped countries, especially among children [93, 94]. Plant-based proteins accounted for 57 % of the global protein supply and were followed by protein sourced meat and dairy, which accounted for 18% and 10%, respectively [95]. Even though plant-based proteins constitute a majority of the global protein supply, according to the

*Parts of this section are reprinted with permission from Lahiri A, Rastogi K, Datta A, Septiningsih EM. Bayesian Network Analysis of Lysine Biosynthesis Pathway in Rice. Inventions. 2021; 6(2):37. https://doi.org/10.3390/inventions6020037

World Health Organization (WHO) the demand for animal-based protein has been on the rise due to urbanization, population growth, and rising economies. The WHO predicts the annual meat production to reach 376 million tonnes by 2030, a 72% increase since 1997-1999 when the yearly meat production was 218 million tonnes [96]. This global increase has placed a burden on the livestock sector, especially in Europe and the Americas, where animal-based protein intake is higher than plant-based proteins [97]. In the USA and European countries, proteins from animal-based sources ranged from 55% to 71 % (depending on countries) of the total protein intake, a significant proportion of which were from red meat [98].

Animal-based protein sources such as meat, milk, and eggs are richer in essential amino acids and have a higher food protein quality in terms of digestibility, net protein utilization, and biological value compared to plant-based protein sources like legumes and cereals [97]. However, animal-based proteins, specifically processed and red meats, have been linked with cancer, type 2 diabetes, and cardiovascular diseases [99, 100, 101]. Apart from health concerns, proteins sourced from animals have a significant impact on climate change. According to the Food and Agriculture Organization of the United Nations, the livestock supply chain accounts for 14.5% of global anthropogenic greenhouse gas emissions[102]. With the global population set to reach 9.8 billion by 2050 and the increasing demand for animal-based proteins, the challenges associated with food security and climate change will only be exacerbated[96, 103]. Hence, a shift towards plant-based protein sources may help reduce the carbon footprint, risks of chronic illness, and food insecurity. While plant-based proteins may not contain all the necessary essential amino acids, a diet containing a diverse range of plant proteins can help overcome this limitation [104]. Cereal plants such as wheat, rice, and maize constitute the primary protein sources in developing countries [105, 106]. With the majority of the world's population living in developing countries, it will therefore be beneficial to increase the protein content in cereal plants to ensure food security and prevent malnutrition.

3.1.1 Lysine Content in Rice

Lysine is produced in the aspartate pathway along with three other essential amino acids threonine, methionine, and isoleucine [107]. Lysine is also the first limiting essential amino acid in cereal and legume crops because it is present in the lowest quantity [107, 108, 109]. This is why lysine deficiency is a common problem in developing nations that rely heavily on cereal crops [107, 110]. A lysine deficient diet can reduce immunity, decrease protein levels in the blood, and cause retardation of mental and physical development in children [108]. Rice is a cereal plant that is an important food source for more than 50% of the global population [111]. About 95% of global rice is produced in developing countries, among which 92% are countries in Asia [112]. Rice accounts for 50% of the dietary caloric supply for 520 million living in poverty in Asia [113]. Like most cereal crops, rice is deficient in lysine, so in this study, we are interested in identifying the genetic regulators of lysine in rice, since intervening at these regulators has the potential to increase the free lysine content in rice [114]. Enriching lysine content in rice will be a step towards ensuring food security and preventing malnutrition especially in the vulnerable sectors of the global population.

Over the last 50 years, lysine metabolism has been extensively studied. It has been shown that lysine is a self-regulating amino acid as the lysine biosynthesis pathway has two inhibition feedback loops. These feedback loops are activated by the free lysine content, which negatively regulates the enzymes dihydrodipicolinate synthase (DHPS) and aspartate kinase (AK) [108, 115]. AK is the first enzyme of the lysine biosynthesis pathway and is also inhibited by threonine, another essential amino acid synthesized by the aspartate pathway [108, 115]. Lysine is also degraded through the enzymes lysine ketoglutarate reductase (LKR) and saccharopine dehydrogenase (SDH) bifunctional enzymes [115]. The LKR and SDH enzymes are present in the saccharopine pathway and they initiate the lysine catabolism process through the TCA cycle (tricarboxylic acid cycle) [108]. The metabolic pathway of lysine biosynthesis and catabolism in presented in Figure 3.1 [116, 117, 118]. Thus lysine can be enriched in cereal plants by enhancing its production in the biosynthesis pathway, preventing its catabolism, or combining these two approaches. A study by

Long et al. (2013) focused on enhancing lysine through metabolic engineering of rice.

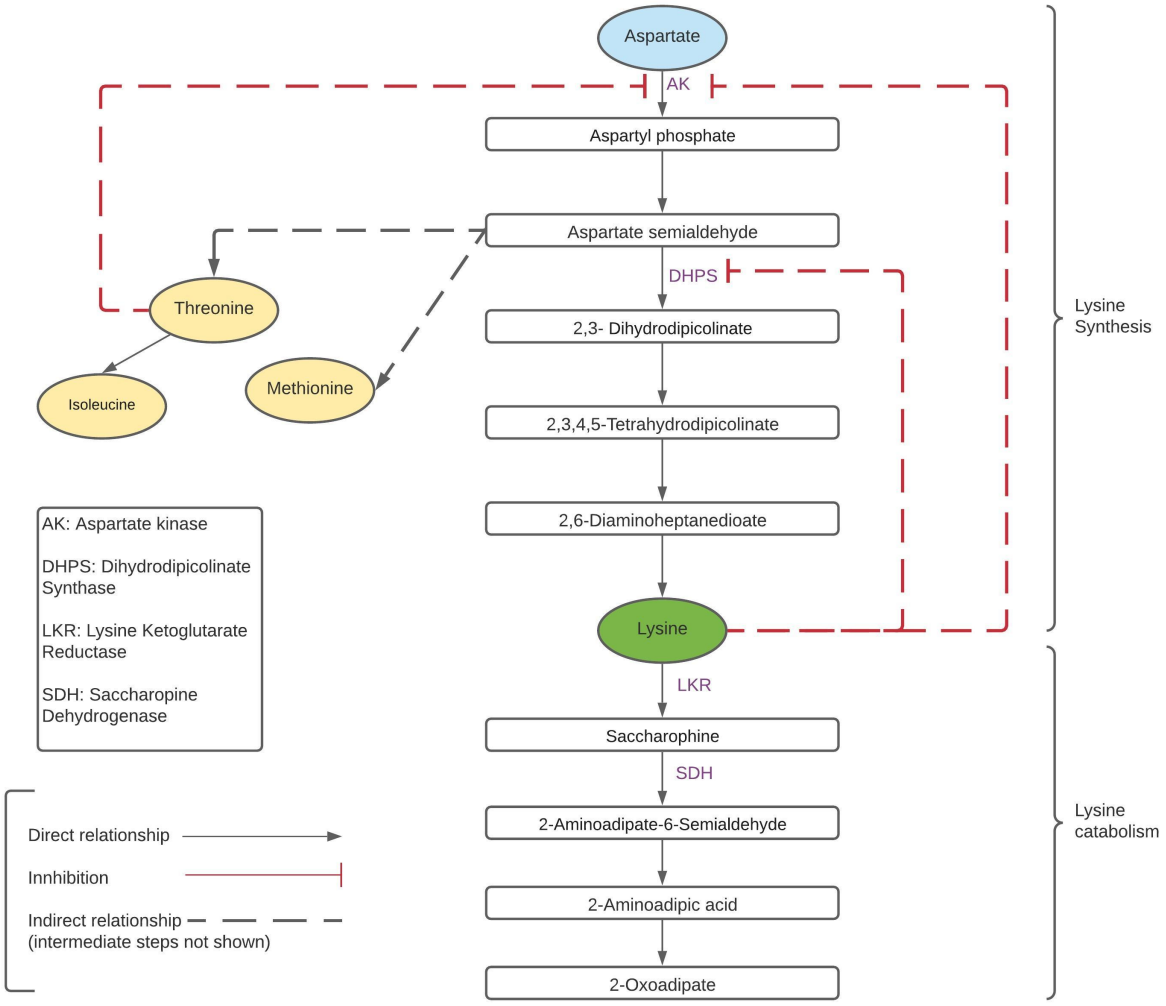

Figure 3.1: **Lysine metabolic pathway for synthesis and catabolism**.(Reprinted from [56])

These transgenic lines of rice overexpressed AK and DHPS. They observed that LKR and SDH levels were significantly higher in seeds of these rice lines, implying that the catabolic enzymes LKR and SDH were counteracting the effects of transgene AK and DHPS [118]. This method increased the free lysine content by 1.1 times in transgenic lines compared to the wild type. This study also implemented a LKR-RNAi line, which showed a 10 fold increase in lysine content, and

a combination LKR-RNAi with AK/DHPS overexpressing lines led to a 60 fold increase in free lysine content. In a different study, *Yang et al*. (2016) developed two pyramid transgenic lines in rice. The lysine content in these transgenic lines showed increased lysine content up to 25 fold. This was achieved by enhancing the biosynthesis pathway and suppressing the catabolism pathway at the same time [119]. Unlike many lysine enhancement studies, which lead to reduced yield, oil content, and phenotype change, no significant trait changes were observed in this case, and the developed transgenic rice was deemed favorable for commercialization [120, 121, 122].

While these studies have demonstrated that lysine content can be enhanced through careful metabolic engineering of high-lysine transgenic lines, these are not yet commercialized. Furthermore, transgenic crops rely on introducing foreign genes (transgenes) into the host crop, making them vulnerable to public rejection. That is why in this section, we are interested in understanding the underlying genetic regulatory networks (GRNs) that govern these complex interactions. The GRNs can help us identify the genetic regulators of lysine which can be targeted using gene-editing methods such as CRISPR-Cas9. Unlike transgenic crops, the final product of gene editing can be cleared of any foreign DNA segments. Instead of relying on transgenic insertions, gene editing instead knocks out or replaces targeted native genes in the genome of the crop to give rise to desirable traits. The United States Department of Agriculture (USDA) has allowed gene edited crops to be labeled as non-GMO, which will make gene edited crops significantly less controversial than transgenic crops [123]. A recent study by *Shew et al.* showed that gene edited crops were preferred over GMO crops in multiple countries [124]. Thus by studying the underlying GRN involved in lysine regulation in rice we can identify potential targets for gene editing.

LKR and SDH are known regulators of lysine in the catabolic pathway, and genetically intervening at them can prevent lysine degradation [125]. Therefore in this section, we focus on identifying lysine regulators in the biosynthesis pathway. Overexpressing the regulators in the biosynthesis pathway through gene editing techniques such as CRISPR-Cas9 has the potential to increase the free lysine content in rice. In Figure 3.2, we derive the GRN of the lysine biosynthesis pathway in rice (*Oryza Sativa*) from the KEGG pathways database [126]. Each rectangular box

in Figure 3.2 represents a gene in the lysine biosynthesis pathway. The gene names are annotated according to their respective MSU IDs (LOC_Os##g#####) [127].

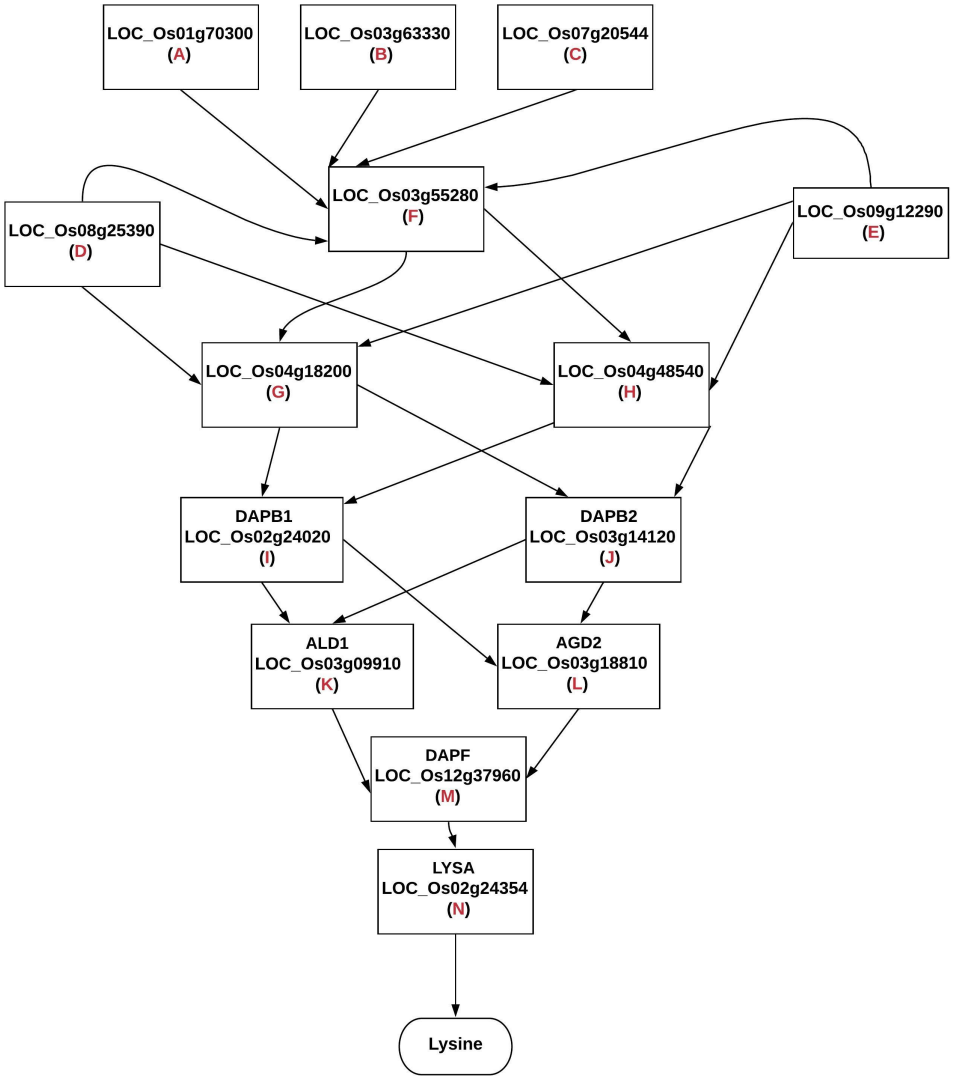

Figure 3.2: **Gene regulatory network for lysine biosynthesis pathway in rice.** The gene names are presented according to their MSU IDs. The letters in red font are aliases for the respective genes. For e.g. LOC_Os01g70300 will be referred to as gene A. Genes I-N have been given names in the literature, these have been mentioned in the figure alongside their respective MSU IDs.(Reprinted from [56])

In addition to the MSU IDs, the boxes contain letters in red font within parenthesis. These

letters are used as aliases for genes in the later sections of this dissertation. Genes I-N have been given names in the literature and these names have been mentioned in the boxes alongside their MSU IDs, for e.g. gene K (LOC_Os03g09910) is also known as *ALD1*. The genetic interactions converge at *LYSA* (LOC_Os02g24354 or gene N) which positively regulates the amino acid lysine (L-Lysine, where the α carbon is in the S configuration). This makes *LYSA* (gene N) a reporter gene of lysine. Thus our objective is to identify genes that will upregulate *LYSA*.

To identify the *LYSA* regulators, we will model the GRN of the lysine biosynthesis pathway using Bayesian Networks (BN). We will then use publicly available data to infer the BN model's parameters. The model can then be used to identify the genes that upregulate *LYSA*. This modeling pipeline is similar to our previous work where we identified regulators of drought response in *Arabidopsis* [49, 128]. We identify the *LYSA* regulators under normal and saline stress (NaCl) conditions. Soil salinity is one of the significant environmental constraints on the crop life cycle. Nearly 5% (77 million hectares) of the global arable land has excess salinity [129]. Due to various factors such as climate change and irrigation malpractices, the soil salinity is predicted to increase by 16.2 million hectares by 2050 [130, 131]. Among abiotic stresses, soil salinity is the second largest cause of crop loss in rice after drought [132, 133].Saline stress primarily affects rice during its seedling, early vegetative, and reproductive stages [134, 132]. We have extensively studied and identified regulators of drought response in our previous works [49, 128]. This is why in our current study we shift our attention to saline stress in rice. We are specifically interested in observing if the *LYSA* regulators change under saline stress. Stewart et al. showed that saline stress leads to the accumulation of aspartic acid (aspartate), which is the first element in the lysine biosynthesis pathway [135]. Furthermore, it has been reported that under stressed conditions, aspartic acid catabolizes into asparagines, threonine, lysine, isoleucine, and methionine [136]. Studies involving maize and wheat showed increased Lysine content under saline stress; however, the precise effect of saline stress on the Lysine content in rice remains to be explored [137, 138].

3.2 Materials and Methods

GRNs describe the complex interactions taking place between regulators and their target genes. Typically regulators consist of transcription factors (TFs),genes, RNA binding proteins, and regulator RNAs that can control the gene expression of the target genes [139, 140, 141]. GRNs govern the decision-making process in response to endogenous and external stimuli; thus, understanding their behavior at the genomic level can give us critical insights into achieving desirable phenotypical traits like increased lysine content [142, 143]. GRNs have been modeled extensively in the past for a wide range of applications such as discovering novel biological relationships, studying complex diseases, drug design, and developing pathogen-resistant crops [144, 145, 146, 147, 148]. Common modeling techniques include differential equations, linear models, Boolean networks, probabilistic Boolean networks, Bayesian networks, and small molecule level models [52, 67, 53, 54, 55]. Each technique has its set of advantages and limitations. Therefore, we must consider the nature of the interactions in the GRN and the overall domain of the study while selecting a modeling method. In this section, we are interested in modeling the lysine biosynthesis pathway in rice under normal (unstressed) and saline stress conditions. The interactions taking place in the pathway are sparse, multivariate, and stochastic in nature. Furthermore with the advent of high throughput technologies, publicly available genomic data have become easily accessible [149]. Due to these factors, we will model the lysine biosynthesis pathway using Bayesian networks (BNs). BNs provide a stochastic framework and allow integration of pathway knowledge and data.

3.2.1 Bayesian Network Modeling

BNs are a class of Probabilistic Graphical Models (PGM) that integrate probability and graph theory to represent stochastic and causal relationships among variables in a system [150, 151]. BNs consist of two main components (i) a directed acyclic graph (DAG) and (ii) local probability distributions (LPD) or the network parameters [152]. The DAG is a map that describes the causal relationships among the system variables, also known as nodes. DAGs specify the dependencies among the nodes and explain the flow of cause and effect in the overall network. The DAG can

be derived from the literature or estimated from data using structure learning algorithms [153]. Associated with each node in the DAG is a local probability distribution (LPD) which describes the stochastic nature of interaction among the connected nodes [151]. The LPDs and the DAGs together describe the factorization of the joint probability distribution of all the nodes in terms of their LPDs. In order to formalize this notion consider a BN with N nodes such that it has a DAG structure $\mathcal{G}$(X,E), where X_i represents the i^{th} node in the set of nodes X and E represents the set of casual edges between the nodes. Now suppose the LPD for each node X_i is given by P(X_i | $P_a(X_i)$), where $P_a(X_i)$ is the set of parent nodes of X_i. Then by the local Markov independence assumption, each node given its parent nodes, is independent of its nondescendant nodes. We can then factorize the joint probability of all the nodes in X as:

$$P(X = \{X_1, X_2, .., X_i, ...X_N\}) = \prod_{i=1}^{N} P(X_i|P_a(X_i)) \tag{3.1}$$

To model the lysine biosynthesis pathway using BN, we construct a DAG from the Kegg pathway we discussed in Figure 3.2. Learning the DAG from data is an NP-Hard problem and often requires selecting a graph structure from a candidate of possible DAGs [154, 155]. This is a computationally expensive task, and the size of publicly available genomic datasets is not sufficiently large to produce a reliable DAG. Therefore we use pathway information (see Figure 3.2) to construct the DAG for the lysine biosynthesis pathway in Figure 3.3. Every node (represented by circles) in the DAG represents a gene present in the lysine biosynthesis pathway. These genes are referenced by their aliases; for instance, gene N represents *LYSA*. The nodes are connected by arrows that represent actual biological relationships as described in the pathway. We assume that genes in the network can be active, dormant, or inhibited. Thus we model each node as a categorical random variable with three states 1(active), 0 (dormant), and -1 (inhibited). Associated with each node is a rectangular box that describes the LPD (network parameter). For Node A, θ_A is vector representing the marginal probability of gene A being active, dormant, or inhibited. Similarly, $\theta_{M|L,K}$ is a vector representing the conditional probability of gene M being active, dormant, or in-

hibited given the states of its parents, gene L and gene K. This completes our discussion of the DAG for the lysine biosynthesis pathway. In the following section, we will discuss how to estimate the LPDs. Once all the LPDs have been calculated the Bayesian network model is complete and can be used to perform gene intervention simulation under normal and saline stress conditions. These simulations will help us gain insight into the effect of intervening at the various genes. Genes that upregulate *LYSA* (gene N) will be considered ideal targets for genetic intervention. Interventions in the GRN can be carried out using gene editing methods such as CRISPR-Cas9 [143]. A simple example BN with its LPDs has been shown in section 3.2.3 for the purpose of demonstrating inference in BN.This example might be useful in developing a better understanding of the DAG structure and the LPDs.

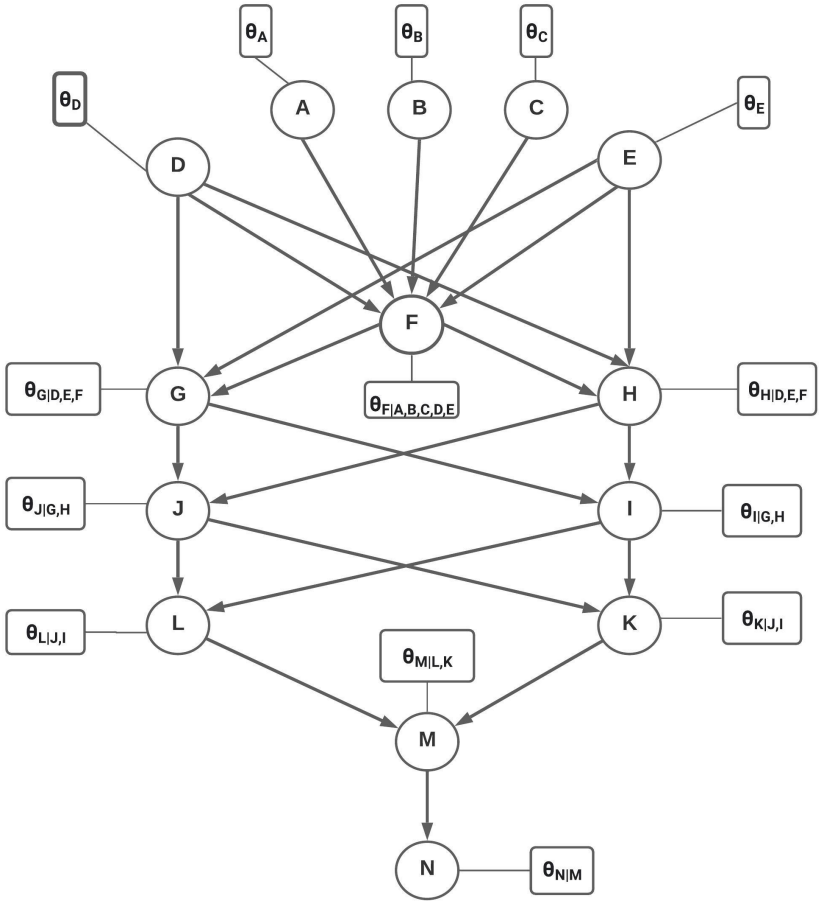

Figure 3.3: **Directed Acyclic Graph (DAG) of the lysine biosynthesis pathway.** Each node (circle) represents a gene in the pathway. The rectangular boxes represent the local probability distributions of the respective nodes. Each node is modeled as a categorical random variable with the following states: active(1), dormant(0), and inhibited (-1).(Reprinted from [56])

3.2.2 Parameter Estimation

Several methods can be employed to estimate the LPDs (network parameters) in a BN. Frequentist approaches such as Maximum Likelihood Estimation (MLE) are common when estimating the LPDs in a BN [156]. However, we will use a Bayesian approach to estimate the LPDs for the DAG constructed in the previous section. This is because the sizes of publicly available datasets are not sufficiently large to be reliably used by data-driven frequentist approaches. Unlike frequentist approaches, Bayesian estimation produces a posterior probability distribution for the

LPDs based on data and prior knowledge [157]. The point estimate for the LPDs can be obtained by approximating the posterior distributions by their expected value or mode [158]. The Bayesian estimation process is based on Bayes rule where the posterior distribution of a random variable X, for a dataset $\mathcal{D}$, is given by:

$$P(X|D) = \frac{P(\mathcal{D}|X)P(X)}{P(\mathcal{D})} \tag{3.2}$$

where P(X) is the prior distribution of X

We will now use this approach to derive the general expression for estimating the LPDs for a BN where the nodes are modeled as categorical random variables. We can then extend our derivation to the DAG in Figure 3.3.

Consider a BN with a DAG denoted by $\mathcal{G}$ containing N(N is a Natural number) nodes. Each node X_i in $\mathcal{G}$ is modeled as a categorical random variable with the following states: active (1), dormant (0), and inhibited (-1). Thus for any node X_i in $\mathcal{G}$, $X_i \in$ **S**={1,0,-1}, so if X_i =0 , it implies that the node X_i is dormant. Let the probability with which X_i assumes any of the states in set **S** be given by the probability vector θ_{X_i}. Then θ_{X_i} is of the form $[\theta_{X_i=1},\theta_{X_i=0},\theta_{X_i=-1}]^T$, where $\theta_{X_i=s}$ represents the probability of X_i=s for s $\in$ **S** and $\Sigma_s\ \theta_{X_i=s}$ =1. Now, suppose we have a dataset $\mathcal{D}$ which contains *n* (*n* is natural number) independent and identically distributed(i.i.d) observations for each of the N nodes in $\mathcal{G}$. For a node X_i in $\mathcal{G}$, let M_{X_i}[**S**=s] represent the frequency of X_i=s in $\mathcal{D}$ (Σ_s M_{X_i}[s]=n). Then the likelihood under the dataset $\mathcal{D}$ can be modeled as:

$$P(X_i|P_a(X_i), \theta_{X_i}) \sim Multinomial(\theta_{X_i}, n) \tag{3.3}$$

$$Multinomial(\theta_{X_i}, n) = n! \prod_{s \in \mathbf{S}} \frac{\theta_{X_i}^{M_{X_i}[s]}}{M_{X_i}[s]!} \tag{3.4}$$

The Bayesian estimation process requires selecting a prior distribution. Prior distributions can be selected based on domain knowledge; however in its absence, there are no fixed methods to choose a prior. The subjective selection of the prior distribution is often cited as a drawback of

the Bayesian estimation process, as different priors lead to different results for the posterior distribution [159]. We set the prior distribution on θ_{X_i} for each node $X_i \in \mathcal{G}$ to follow a Dirichlet distribution. A Dirichlet prior under a multinomial likelihood causes the posterior distribution also to follow a Dirichlet distribution. This is because the multinomial and Dirichlet distributions belong to conjugate families of distributions[160, 161]. Therefore we have the following formulation for the posterior distribution on θ_{X_i}:

$$\theta_{X_i} \sim Dirichlet(\boldsymbol{\alpha}) \tag{3.5}$$

$$\boldsymbol{\alpha} = [\alpha_{s=1}, \alpha_{s=0}, \alpha_{s=-1}]$$

$$Dirichlet(\theta_{X_i}; \boldsymbol{\alpha}) = \frac{1}{\beta(\boldsymbol{\alpha})} \prod_{s \in \mathbf{S}} [\theta_{X_i=s}]^{\alpha_s - 1} \tag{3.6}$$

where $\beta(\boldsymbol{\alpha})$ is the Multivariate Beta function

$$P(\theta_{X_i}|X_i) = Dirichlet(\boldsymbol{\alpha}') \tag{3.7}$$

and

$$\boldsymbol{\alpha}' = [\alpha_{s=1} + M_{X_i}[s=1], \alpha_{s=0} + M_{X_i}[s=0], \alpha_{s=-1} + M_{X_i}[s=-1]]$$

$$\boldsymbol{\alpha}' = [\alpha'_{s=1}, \alpha'_{s=0}, \alpha'_{s=-1}]$$

In our study we specifically set the prior distribution on each node X_i to be Dirichlet($\alpha_{s=1}$ =1,$\alpha_{s=0}$=1,$\alpha_{s=-1}$=1), which corresponds to uniform distribution over the open standard 2-simplex and is a non informative prior [162, 163]. This is an appropriate choice for the prior distribution in our study as we do not have prior knowledge regarding the distribution of each node in the BN. Furthermore, this assumption on the prior distribution of the nodes allows us to obtain a closed form solution for the posterior distribution. Selecting a different prior will often lead to non-closed form solution for the posterior distribution and calculating the probability of data (P($\mathcal{D}$)) can be computationally expensive [164]. The formulation in Equation (3.7) represents the posterior distribution of the node parameter θ_{X_i}. We approximate θ_{X_i} by its expected value in order to obtain a point estimate for the LPDs in the BN. The expectation of a Dirichlet distribution is given by

[165]:

$$\theta_{X_i} = \begin{bmatrix} \theta_{X_i=1} \\ \theta_{X_i=0} \\ \theta_{X_i=-1} \end{bmatrix} \approx E[\theta_{X_i}|X_i] = \begin{bmatrix} \frac{\alpha'_{s=1}}{\sum_{\mathbf{S}} \alpha'_s} \\ \frac{\alpha'_{s=0}}{\sum_{\mathbf{S}} \alpha'_s} \\ \frac{\alpha'_{s=-1}}{\sum_{\mathbf{S}} \alpha'_s} \end{bmatrix} \tag{3.8}$$

Similarly if we have a node X_i with a parent node Y_i=s (s $\in$ **S**) under the same Dirichlet and Multinomial framework, then the LPD associated with $\theta_{X_i|Y_i}$ can formulated as follows:

$$\theta_{X_i|Y_i=s} = \begin{bmatrix} \theta_{X_i=1|Y_i=s} \\ \theta_{X_i=0|Y_i=s} \\ \theta_{X_i=-1|Y_i=s} \end{bmatrix} \approx E[\theta_{X_i|Y_i=s}|(X_i \mid Y_i = s)] = \begin{bmatrix} \frac{\alpha_{s=1}+M_{X_i|Y_i}[X_i=1,Y_i=s]}{\sum_{\mathbf{S}} \alpha_s+M_{X_i|Y_i}[X_i=1,Y_i=s]} \\ \frac{\alpha_{s=0}+M_{X_i|Y_i}[X_i=0,Y_i=s]}{\sum_{\mathbf{S}} \alpha_s+M_{X_i|Y_i}[X_i=0,Y_i=s]} \\ \frac{\alpha_{s=-1}+M_{X_i|Y_i}[X_i=-1,Y_i=s]}{\sum_{\mathbf{S}} \alpha_s+M_{X_i|Y_i}[X_i=-1,Y_i=s]} \end{bmatrix} \tag{3.9}$$

In equation (3.9), $M_{X_i|Y_i}$[X_i=1,Y_i=s] represents the frequencies when X_i=1 and Y_i=s simultaneously in the dataset $\mathcal{D}$. Similarly, $M_{X_i|Y_i}$[X_i=0,Y_i=s] is the frequency of datapoints in $\mathcal{D}$ when X_i=0 and Y_i=s simultaneously, and so on for X_i=-1.Once the node parameters are estimated, gene intervention simulations can be carried out using inference in the BN. Inference computes the effect of intervening at each node on the reporter gene *LYSA* (gene N).

3.2.3 Gene Intervention Simulations

BNs represent the cause and effect relationship among the nodes of the system being modeled. Inference quantifies the cause and effect relationship by allowing us to compute conditional probability queries. Then for a node of interest X, also known as the query node and an intervention(or evidence) node E in the BN, we can compute the conditional probability P(X|E) using inference algorithms. This implies, if we instantiate (fix) node E, we can calculate its effect on node X. Inference algorithms use the network parameters and structural dependencies to compute the required conditional probabilities. To further elucidate this notion, consider the BN shown in

Figure 3.4. Let each node of the BN be a binary random variable with states 0 and 1. Suppose we have estimated the LPDs P(A), P(B|A), P(C|A), P(D|B,C), then we can use inference in this BN to answer conditional probability queries such as P(D=1|A=1).

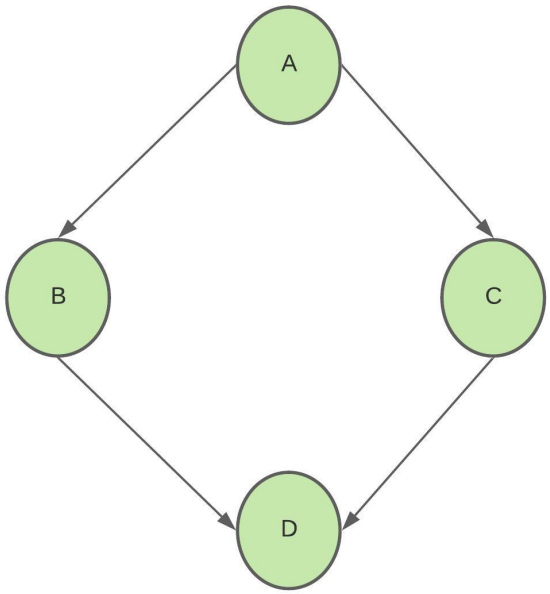

Figure 3.4: **Example BN with binary nodes.**(Reprinted from [56])

We compute P(D=1|A=1) as follows:

$$P(D=1|A=1) = \frac{P(D=1, A=1)}{P(A=1)}$$
$$= \frac{\sum_B \sum_C P(A=1, B, C, D=1)}{P(A=1)}$$

Using the properties of the BN, all nodes are independent of any non descendant nodes

$$P(D=1|A=1) = \frac{\sum_B \sum_C P(A=1)P(B|A=1)P(C|A=1)P(D=1|B,C)}{P(A=1)}$$
$$= \sum_B \sum_C P(B|A=1)P(C|A=1)P(D=1|B,C)$$

We can use the LPDs to calculate the exact probability P(D=1|A=1).

Inference techniques such as the one applied in the BN in Figure 3.4 are classified as "exact" because they compute the true values for the conditional probability query. However, exact inference in BNs has been shown to be NP-hard [166, 167]. While there exist efficient algorithms for exact inference, they are often limited to simpler DAG structures[166]. For example, Pearl's message-passing algorithm works efficiently for singly connected DAG structures [168].Therefore for larger DAGs, exact inference is not ideal as the computational cost of calculating the conditional probabilities can be expensive. In such cases, we employ approximate inference algorithms, which produce estimates of the exact conditional probabilities [169]. Approximate inference can include wide-ranging techniques such as model simplification methods, loopy belief propagation methods, search based methods, utility based methods, and stochastic simulation methods [170]. In this section, we implement a stochastic simulation-based inference technique called Likelihood Weighting (LW) to estimate the conditional probability queries in the BN model for the lysine Biosynthesis pathway. Stochastic simulation techniques estimate the conditional probabilities by drawing samples from the LPDs. These estimates typically converge to the true conditional probabilities as the number of samples drawn increases. LW can efficiently handle inference of large multiply connected BNs and is based on forward sampling [170, 171]. Since our BN model is multiply connected and we are only interested in estimating P(N=1 | E $\in$ {A,B,C,...,M}), i.e., the probability of upregulating *LYSA* (gene N), while conditioning on other genes (evidence or intervention nodes), LW turns out be a suitable method for performing inference.

The LW algorithm estimates the conditional probability, P(X=x | E=e) for a query node X and an evidence node E by generating samples from a BN model. We fix the sample size (m) and a topological ordering at the start of the algorithm. The algorithm iterates through a sample generation process m times, and then computes the conditional probability from the generated samples. During the sample generation process, the algorithm generates values for the nonevidence nodes only; it sets the value of the evidence node to its observed (e, in this case) value. The node values for each sample are generated in the established topological ordering. Each sample

is assigned a weight of 1 at the start of the sample generation process. The weight is updated only when an evidence node is encountered while traversing the topological ordering. When this happens, the sample's weight is updated by multiplying the current weight with the likelihood of the evidence node conditioned on the state of its parent nodes. The likelihood is given by the probability P(E=e | P_a(E)). The process is repeated until m samples are generated. Following this step, conditional probability is estimated by dividing the sum of the weights of the samples where X=x by the sum of all the sample's weights. The pseudo code for the LW algorithm by Stuart Russell and Peter Norvig is presented in Algorithm 1 [172].

Algorithm 2: Likelihood-Weighting Algorithm

Function LIKELIHOOD-WEIGHTING(X, e, bn,N)**:**
- **outputs** an estimate of P(X|e)
- **inputs:** X, the query variable
  - **e**, observed values for variables **E**
  - bn, a Bayesian network specifying joint distribution P(X_1,..,X_n)
  - N, the total number of samples to be generated
- **local variables:W,** a vector of weighted counts for each value of X, initially zero
- **for** *j=1 to N* **do**
  - **x**,w ← WEIGHTED-SAMPLE(bn,**e**)
  - **W**[x] ← **W**[x] + w where x is the value of X in **x**
- **end**
- **return** NORMALIZE(**W**)

Function WEIGHTED-SAMPLE(*bn*, ***e***)**:**
- **outputs** an event and a weight
- w ← 1; **x** ← an event with n elements initialized from **e**
- **for** ***each*** *variable* X_i ***in*** X_1,..,X_n **do**
  - **if** X_i *is an evidence variable with value* x_i *in* ***e*** **then**
    - w ← w × P(X_i =x_i | parents($_i$))
  - **else**
    - **x**[i]← a random sample from P(X_i | parents(X_i))
  - **end**
- **end**
- **return x**,w

3.2.4 Dataset

To estimate the LPDs in the BN model, we use the dataset GSE98455, which is publicly available from the NCBI GEO database [173, 174, 175]. This dataset was selected as it contains RNA-Seq counts for rice seedlings under saline stress and normal (unstressed or control) conditions and had the highest number of samples (data points) per gene available among publicly available datasets. The entire dataset contained 57846 rows (genes) and 368 columns (control and saline stress). Since our BN model contains nodes modeled as categorical variables, the RNA-Seq data had to be preprocessed. The data preprocessing steps are outlined as follows:

1. The entire dataset was normalized using the ratio of medians methods.

2. We selected the data for the genes A-N, as these were the genes in the BN model. We identified the data for each of the genes by mapping their dataset IDs to their respective MSU IDs. This reduced our dataset to a size of 14 rows (Gene A-N) and 368 columns.

3. We further segregated the normalized dataset based on saline stress and normal conditions. Since the number of columns for saline stress and normal conditions were the same, each of the resulting datasets had 14 rows and 184 columns.

4. We ran K-means clustering separately on both the saline stress and normal conditions dataset to convert them from normalized to categorical values. The clustering process categorized the data in both the datasets into the following values 1 (active), 0 (dormant), and -1 (inhibited). The low expression values were categorized to the value of -1, the high expression values were categorized to the value of 1, and the remaining expression values in the middle were categorized to a value of 0.

Once the categorical values were obtained for both the treatment and control datasets, the LPDs were estimated under each case using the Bayesian approach described in the *Parameter Estimation* section. We then ran LW to simulate gene intervention.The ratio of medians methods used for normalization is described in the DESeq2 data processing protocols by *Love et al*.[176]. DESeq2 is one of the most commonly used RNA-Seq data processing protocols and is easily accessible on the R programming language as a package (DESeq2)[177, 178, 179, 180]. The file for mapping dataset IDs to MSU IDs was provided to us by the authors of the dataset GSE98455. We have highlighted their contribution in the acknowledgment section. A visual representation of the data processing pipeline has been presented in Figure 3.5. Figures 3.6 and 3.7 show the discretized categorical data for each node in the BN under normal and saline stress conditions, respectively.

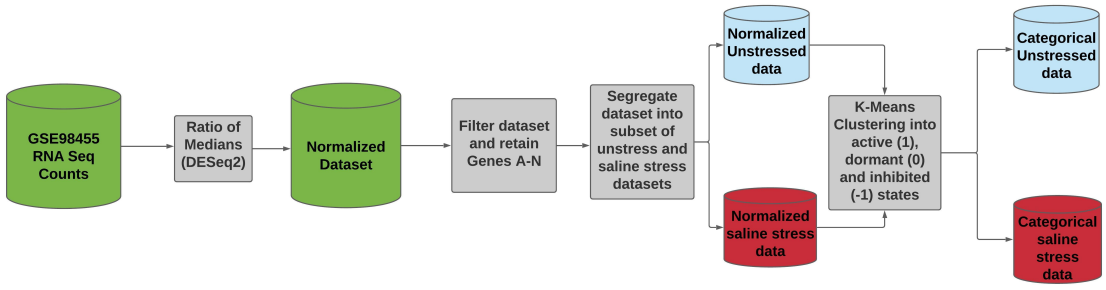

Figure 3.5: **Data processing pipeline for RNA-Seq dataset GSE98455.**(Reprinted from [56])

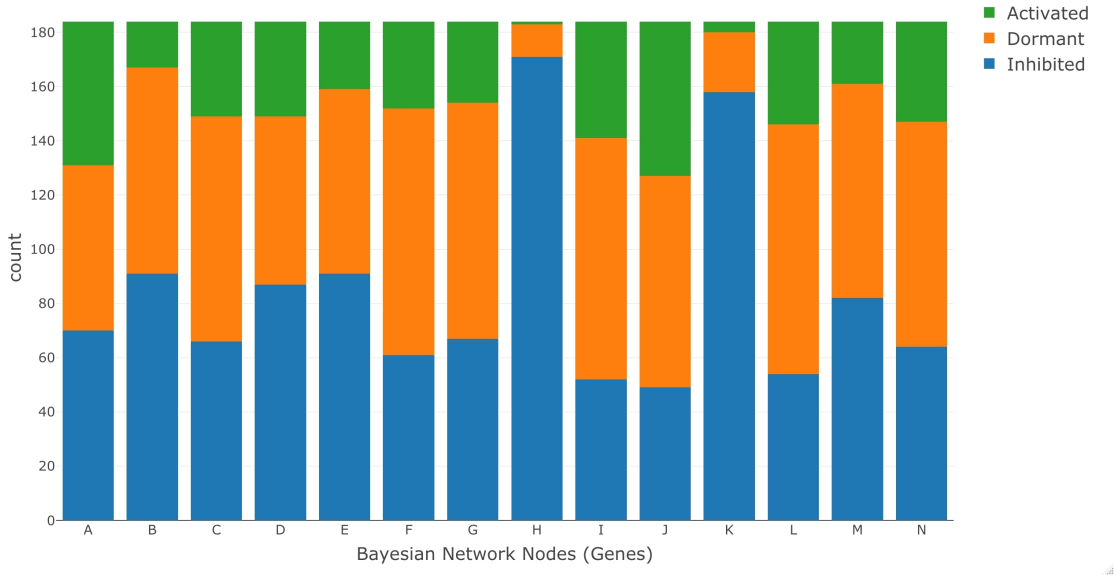

Figure 3.6: **Discretized RNA-Seq data under normal conditions.**(Reprinted from [56])

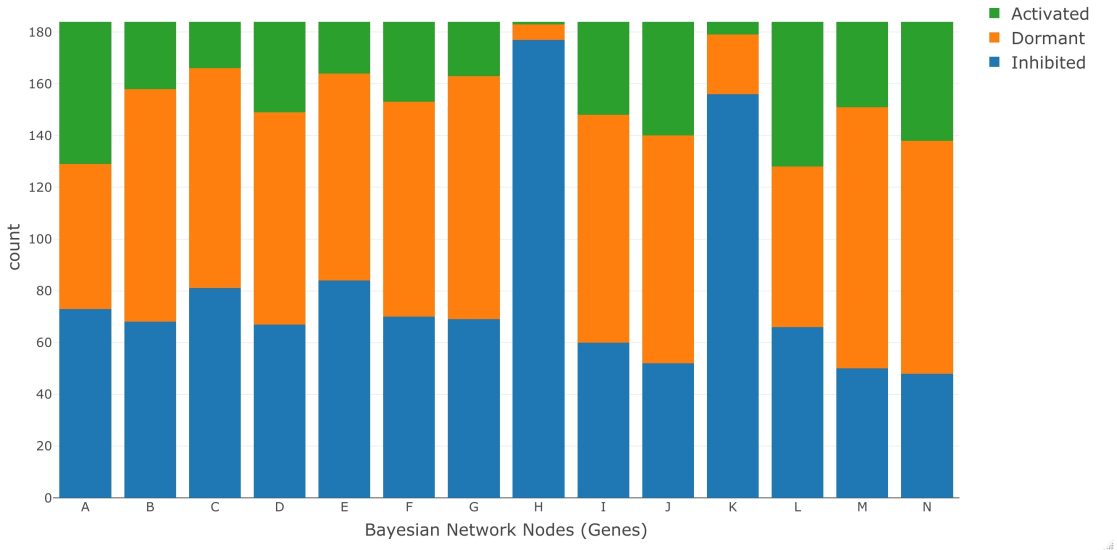

Figure 3.7: **Discretized RNA-Seq data under saline stress conditions.**(Reprinted from [56])

3.3 Results

The LPDs estimated from the RNA-Seq dataset were used to simulate gene intervention in the BN. When intervening at a gene, the node representing that gene in the BN was instantiated to a status of active (1), dormant (0), or inhibited (-1). We applied the LW algorithm with a large sample size of 600,000 to compute the probability P(N=1 | Gene Intervention) and ensure convergence of the probabilities being estimated. Gene N (*LYSA*) is set as the query node because it is the reporter gene for lysine production, thus upregulating gene N (*LYSA*) may lead to increased lysine production. We perform intervention at genes A-M one at a time and then in combinations of two (pairs) at a time. These gene intervention strategies were applied under both normal and saline stress conditions. In order to measure the causal effect of intervention, we subtract the marginal probability P(N=1) from P(N=1 | Gene Intervention), for all the possible gene intervention strategies. This difference is defined as the score metric and is used to compare the effectiveness of each gene intervention strategy. The data processing and probability computation pipeline was written in the R programming language, and the Bnlearn package was used to perform LW [181, 182, 183]. So,

$$score = P(N = 1 \mid \text{Gene Intervention}) - P(N = 1). \tag{3.10}$$

Since there are many possible combinations under single and pairwise gene interventions, we only include the top five intervention strategies with the highest scores in Figure 3.8 and Tables 3.1 and 3.2. In Figures 3.8 (a) and 3.8(b), we present the scores for single node intervention under normal and saline stress conditions.

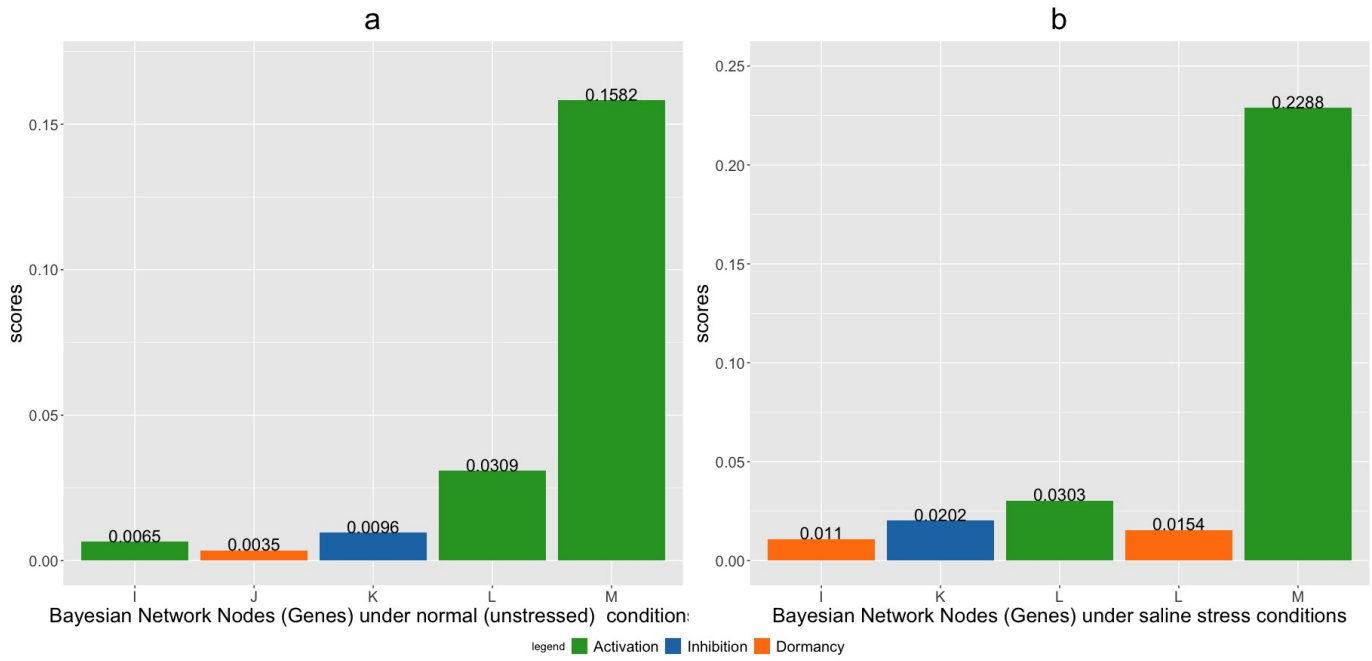

Figure 3.8: **Single node intervention under (a) normal and (b) saline stress conditions.**(Reprinted from [56])

It is clear from Figures 3.8(a) and 3.8(b) that activating gene M (*DAPF*) has the maximum score. This implies that under both normal and saline stress conditions, genetically activating gene M (*DAPF*) has the best chance for upregulating the reporter gene N (*LYSA*). We also notice that gene L (*AGD2*) is also fairly active in its role in upregulating gene N (*LYSA*). Activating gene L (*AGD2*) achieves the second-largest score under normal conditions. Under saline stress conditions, activating gene L (*AGD2*) or keeping it dormant also ranks among the top five gene intervention strategies. Inhibiting gene K (*ALD1*) achieves the third-highest scores under normal and saline stress conditions. Additionally, we also observe that midstream genes such as gene I (*DAPB1*) and gene J (*DAPB2*) also play an active role in upregulating gene N (*LYSA*). However, activating gene M (*DAPF*) has a significantly higher score under both conditions, thus gene M (*DAPF*) servers as an ideal candidate for gene intervention.

Table 3.1: **Top five pair wise intervention strategies under normal conditions.**(Reprinted from [56])

Index	Gene Name/Alias	Intervention	Gene Name/Alias	Intervention	Score
1	*ALD1* (Gene K)	Active	*DAPF* (Gene M)	Active	0.1657
2	*ALD1* (Gene K)	Dormant	*DAPF*(Gene M)	Active	0.1653
3	*AGD2* (Gene L)	Inhibited	*DAPF* (Gene M)	Active	0.1639
4	Gene A	Active	*DAPF* (Gene M)	Active	0.1637
5	Gene C	Active	*DAPF* (Gene M)	Active	0.1634

Table 3.2: **Top five pair wise intervention strategies under saline stress conditions.** (Reprinted from [56])

Index	Gene Name/Alias	Intervention	Gene Name/Alias	Intervention	Score
1	Gene F	Dormant	*DAPF* (Gene M)	Active	0.2322
2	*ALD1* (Gene K)	Dormant	*DAPF*(Gene M)	Active	0.2321
3	Gene B	Inhibited	*DAPF* (Gene M)	Active	0.2312
4	Gene E	Inhibited	*DAPF* (Gene M)	Active	0.2306
5	Gene A	Dormant	*DAPF* (Gene M)	Active	0.2305

Tables 3.1 and 3.2 represent the five highest-scoring pairwise intervention strategies for normal and saline stress conditions. Each table contains gene names or aliases along with their intervention strategies. The tables are arranged in descending order of the score. Under each condition, the score amongst different five highest-scoring strategies are almost similar with marginal differences. Under normal conditions in table 3.1, we observe that activating both gene K (*ALD1*) and gene M (*DAPF*) maximized the scores. While under saline stress keeping gene F (LOC_Os03g55280)

dormant and activating gene M (*DAPF*) achieved the highest score. This implies that under each of the conditions, the respective pairwise intervention strategy with highest scores maximize the likelihood of upregulating gene N (*LYSA*).From tables 3.1 and 3.2, it can also be seen that upstream genes such as genes A,B,C and E are also involved in the upregulation of gene N (*LYSA*) and produce comparable scores to those produced by the regulation of downstream genes like gene K (*ALD1*) and gene L (*AGD2*). Across both the conditions, we also observe that gene M (*DAPF*) is always upregulated, which serves to be a further indicator of the high regulatory effect of gene M (*DAPF*) on gene N (*LYSA*). We should note that these rankings in Tables 3.1 and 3.2 might vary slightly upon rerunning the simulations, as LW is based on a stochastic simulation process, which may cause minor variation in estimating the probabilities required for computing the score metric. However, this does not affect our overarching conclusion that *DAPF* is the most potent regulator of *LYSA*, as it is present and upregulated in all the top five strategies under pairwise intervention. Furthermore, under single intervention, *DAPF* scores significantly higher than the rest of the genes. The proteins encoded by each of the genes in tables 3.1 and 3.2 are summarized in Table 3.3.

Table 3.3: **Protein encoded by intervention genes in Figure 3.8 and Tables 3.1 and 3.2.** (Reprinted from [56])

Gene Alias/Name	MSU IDs	Protein
Gene A	LOC_Os01g70300	Aspartokinase 3, chloroplast precursor, putative, expressed
Gene B	LOC_Os03g63330	Aspartokinase, chloroplast precursor, putative, expressed
Gene C	LOC_Os07g20544	Aspartokinase, chloroplast precursor, putative, expressed
Gene E	LOC_Os09g12290	Bifunctional aspartokinase/homoserine dehydrogenase, chloroplast precursor, putative, expressed
Gene F	LOC_Os03g55280	Semialdehyde dehydrogenase, NAD binding domain containing protein, putative, expressed
Gene I/*DAPB1*	LOC_Os02g24020	Dihydrodipicolinate reductase, putative, expressed
Gene J/ *DAPB2*	LOC_Os03g14120	Dihydrodipicolinate reductase, putative, expressed
Gene K/*ALD1*	LOC_Os03g09910	Aminotransferase, classes I and II, domain containing protein, expressed
Gene L/*AGD2*	LOC_Os03g18810	Aminotransferase, classes I and II, domain containing protein, expressed
Gene M/*DAPF*	LOC_Os12g37960	Diaminopimelate epimerase, chloroplast precursor, putative, expressed

4. 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 section 2, 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 selected the model plant Arabidopsis for our study because the genes and proteins in drought response pathways are well defined and identified for Arabidopsis compared to major crops. 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, we observed that the scores were maximized upon activating *MYC2* or inhibiting *ATAF1*. The maximization

*Parts of this section are reprinted with permission from Lahiri A, Zhou L, He P, Datta A (2021) Detecting drought regulators using stochastic inference in Bayesian networks. PLoS ONE 16(8): e0255486. https://doi.org/10.1371/journal.pone.0255486 and Lahiri A, Rastogi K, Datta A, Septiningsih EM. Bayesian Network Analysis of Lysine Biosynthesis Pathway in Rice. Inventions. 2021; 6(2):37. https://doi.org/10.3390/inventions6020037

of scores implied that *MYC2* and *ATAF1* were potential drought regulators, and 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 these interventions at the same time provides a slightly improved score value, indicating the synergistic effect of the strategic interventions. These simulation results indicated that *ATAF1* and *MYC2* were the most potent regulators of drought response compared to the other drought regulatory genes modeled in the BN.

From biological literature we note 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 algorithms, 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 results in section 2, 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.

In section 3, we studied the lysine biosynthesis pathway in rice to identify the genetic regulators of lysine content. Rice is a staple food source for 50% of the global population; with lysine being the first limiting essential amino acid in rice, it is vital to identify gene regulators that can boost lysine content. We modeled the lysine biosynthesis pathway in rice using BNs under normal and saline stress conditions to identify these regulators. We used BNs because they allow us to integrate domain knowledge in the form of pathway information with experimental data. We used real-world RNA-Seq data to estimate the LPDs in the BN and run the gene intervention simulations. We intervened at the genes one at a time and then in pairwise combinations using the LW inference

algorithm. We calculated a score metric to measure the efficacy of the gene intervention strategies.

Our analysis revealed that upregulating *DAPF* (gene M) maximized the probability of the lysine reporter gene *LYSA* (gene N) being upregulated under both normal and saline stress conditions. When *DAPF* (gene M) was upregulated, it not only achieved the highest score under single gene intervention but was also present in all the five highest-scoring gene intervention strategies under pairwise intervention. This implies that *DAPF* (gene M) is a positive regulator of *LYSA* (gene N) and serves as an ideal candidate for genetic intervention. Gene editing can be used to target and upregulate *DAPF* (gene M) in rice. Field experiments involving *DAPF* overexpressing rice can confirm if this intervention strategy upregulates *LYSA* and increases the overall lysine content. We further observed under single gene intervention that midstream genes such as *DAPB1* (gene I) and *DAPB2* (gene J) also played significant roles in upregulating *LYSA* (gene N). On the other hand, under pairwise intervention, we found upstream genes such as genes A, B, C, and E were also involved in upregulating *LYSA* (gene N).

The future steps in our study of lysine will include confirming our finding in this dissertation by performing validation experiments in the field. We would also like to improve our choice of the prior distribution on each node. In our current analysis, we used a noninformative prior as we did not have any knowledge regarding the prior distribution of the nodes in the BN. Using informative priors may increase the computational costs but has the potential to improve our predictions of lysine regulators. Furthermore, we are also interested in studying how other essential amino acids such as Threonine, Methionine, and Isoleucine in the larger aspartate pathway regulate lysine content. Threonine is known to downregulate the enzyme AK in the lysine biosynthesis pathway; thus, studying the multilevel regulation among the different amino acids in the aspartate pathway will help understand lysine production.

REFERENCES

[1] M. Denchak, "Drought: Everything you need to know," Nov 2019. https://www.nrdc.org/stories/drought-everything-you-need-know.

[2] NOAA, "Drought public fact sheet," Aug 2006. https://gml.noaa.gov/obop/mlo/educationcenter/students/brochures%20and%20diagrams/noaa%20publications/Drought%20Fact%20Sheet.pdf.

[3] Q. Wang, J. Wu, T. Lei, B. He, Z. Wu, M. Liu, X. Mo, G. Geng, X. Li, H. Zhou, and et al., "Temporal-spatial characteristics of severe drought events and their impact on agriculture on a global scale," *Quaternary International*, Jul 2014.

[4] FAO, "Disasters causing billions in agricultural losses, with drought leading the way," 2018.

[5] J. Lund, J. Medellin-Azuara, J. Durand, and K. Stone, "Lessons from california's 2012–2016 drought," *Journal of Water Resources Planning and Management*, vol. 144, no. 10, p. 04018067, 2018.

[6] J. Sheffield, E. F. Wood, and M. L. Roderick, "Little change in global drought over the past 60 years," *Nature*, vol. 491, no. 7424, pp. 435–438, 2012.

[7] A. Dai, "Increasing drought under global warming in observations and models," *Nature Climate Change*, vol. 3, p. 52–58, May 2012.

[8] J. Spinoni, G. Naumann, and J. V. Vogt, "Pan-european seasonal trends and recent changes of drought frequency and severity," *Global and Planetary Change*, vol. 148, p. 113–130, 2017.

[9] Z. Wang, J. Li, C. Lai, Z. Zeng, R. Zhong, X. Chen, X. Zhou, and M. Wang, "Does drought in china show a significant decreasing trend from 1961 to 2009?," *Science of The Total Environment*, vol. 579, p. 314–324, 2017.

[10] B. I. Cook, T. R. Ault, and J. E. Smerdon, "Unprecedented 21st century drought risk in the american southwest and central plains," *Science Advances*, vol. 1, no. 1, 2015.

[11] M. R.Allen, M. Kainuma, H. Otto Pörtner, M. Babiker, K. de Kleijne, A. Revi, . , and T. Gabriel Johansen, "Special report: Global warming of 1.5 ° c," 2018.

[12] A. Arneth, H. Barbosa, T. Benton, K. Calvin, E. Calvo, S. Connors, . , and Z. Zommers, "Climate change and land," Aug 2019.

[13] P. Szejner, S. Belmecheri, J. R. Ehleringer, and R. K. Monson, "Recent increases in drought frequency cause observed multi-year drought legacies in the tree rings of semi-arid forests," *Oecologia*, vol. 192, p. 241–259, Apr 2019.

[14] A. D. Tripathi, R. Mishra, K. K. Maurya, R. B. Singh, and D. W. Wilson, "Estimates for world population and global food availability for global health," *The Role of Functional Food Security in Global Health*, p. 3–24, 2019.

[15] N. G. Society, "Drought," Sep 2019.

[16] M. A. Hossain, S. H. Wani, S. Bhattacharjee, D. J. Burritt, and L.-S. P. Tran, *Drought stress tolerance in plants*. Springer, 2016.

[17] A. F. Gilles, J. B. Schinko, and M. Averof, "Efficient crispr-mediated gene targeting and transgene replacement in the beetle tribolium castaneum," *Development*, vol. 142, p. 2832–2839, Sep 2015.

[18] A. Gull, A. A. Lone, and N. U. I. Wani, "Biotic and abiotic stresses in plants," *Abiotic and Biotic Stress in Plants*, 2019.

[19] D. Nguyen, I. Rieu, C. Mariani, and N. M. V. Dam, "How plants handle multiple stresses: hormonal interactions underlying responses to abiotic stress and insect herbivory," *Plant Molecular Biology*, vol. 91, no. 6, p. 727–740, 2016.

[20] L. Xiong, K. S. Schumaker, and J.-K. Zhu, "Cell signaling during cold, drought, and salt stress," *The Plant Cell*, vol. 14, no. suppl 1, 2002.

[21] V. Shulaev, D. Cortes, G. Miller, and R. Mittler, "Metabolomics for plant stress response," *Physiologia Plantarum*, vol. 132, p. 199–208, Sep 2008.

[22] S. Tiwari, C. Lata, P. S. Chauhan, V. Prasad, and M. Prasad, "A functional genomic perspective on drought signalling and its crosstalk with phytohormone-mediated signalling pathways in plants," *Current Genomics*, vol. 18, no. 6, 2017.

[23] M. M. Chaves, J. P. Maroco, and J. S. Pereira, "Understanding plant responses to drought — from genes to the whole plant," *Functional Plant Biology*, vol. 30, no. 3, p. 239, 2003.

[24] K. Yildirim and Z. Kaya, "Gene regulation network behind drought escape, avoidance and tolerance strategies in black poplar (populus nigra l.)," *Plant Physiology and Biochemistry*, vol. 115, p. 183–199, 2017.

[25] F. Takahashi, T. Kuromori, H. Sato, and K. Shinozaki, "Regulatory gene networks in drought stress responses and resistance in plants," *Advances in Experimental Medicine and Biology Survival Strategies in Extreme Cold and Desiccation*, p. 189–214, 2018.

[26] Y. Sun and J. R. Dinneny, "Q&a: How do gene regulatory networks control environmental responses in plants?," *BMC Biology*, vol. 16, Nov 2018.

[27] S. Liu, Z. Lv, Y. Liu, L. Li, and L. Zhang, "Network analysis of aba-dependent and aba-independent drought responsive genes in arabidopsis thaliana," *Genetics and Molecular Biology*, vol. 41, no. 3, p. 624–637, 2018.

[28] K. Shinozaki and K. Yamaguchi-Shinozaki, "Gene networks involved in drought stress response and tolerance," *Journal of Experimental Botany*, vol. 58, p. 221–227, Jun 2006.

[29] Y. Uno, T. Furihata, H. Abe, R. Yoshida, K. Shinozaki, and K. Yamaguchi-Shinozaki, "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*, vol. 97, no. 21, p. 11632–11637, 2000.

[30] Y. Fujita, K. Nakashima, T. Yoshida, T. Katagiri, S. Kidokoro, N. Kanamori, T. Umezawa, M. Fujita, K. Maruyama, K. Ishiyama, and 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*, vol. 50, no. 12, p. 2123–2132, 2009.

[31] Y. Fujita, M. Fujita, K. Shinozaki, and K. Yamaguchi-Shinozaki, "Aba-mediated transcriptional regulation in response to osmotic stress in plants," *Journal of Plant Research*, vol. 124, no. 4, p. 509–525, 2011.

[32] H. Abe, T. Urao, T. Ito, M. Seki, K. Shinozaki, and K. Yamaguchi-Shinozaki, "Arabidopsis atmyc2 (bhlh) and atmyb2 (myb) function as transcriptional activators in abscisic acid signaling," *The Plant Cell*, vol. 15, no. 1, p. 63–78, 2002.

[33] D. Singh and A. Laxmi, "Transcriptional regulation of drought response: a tortuous network of transcriptional factors," *Frontiers in Plant Science*, vol. 6, 2015.

[34] K. Nakashima, Y. Ito, and K. Yamaguchi-Shinozaki, "Transcriptional regulatory networks in response to abiotic stresses in arabidopsis and grasses.," *Plant Physiology*, vol. 149, no. 1, p. 88–95, 2009.

[35] M. Agarwal, Y. Hao, A. Kapoor, C.-H. Dong, H. Fujii, X. Zheng, and J.-K. Zhu, "A r2r3 type myb transcription factor is involved in the cold regulation of cbf genes and in acquired freezing tolerance," *Journal of Biological Chemistry*, vol. 281, p. 37636–37645, Feb 2006.

[36] C.-H. Dong, M. Agarwal, Y. Zhang, Q. Xie, and J.-K. Zhu, "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*, vol. 103, no. 21, p. 8281–8286, 2006.

[37] K. Miura, J. B. Jin, J. Lee, C. Y. Yoo, V. Stirm, T. Miura, E. N. Ashworth, R. A. Bressan, D.-J. Yun, P. M. Hasegawa, and et al., "Siz1-mediated sumoylation of ice1 controls cbf3/dreb1a expression and freezing tolerance in arabidopsis," *The Plant Cell*, vol. 19, no. 4, p. 1403–1414, 2007.

[38] L. Jiao, Y. Zhang, J. Wu, H. Zhang, and J. Lu, "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*, vol. 33, p. 557–568, Oct 2014.

[39] J.-S. Kim, J. Mizoi, T. Yoshida, Y. Fujita, J. Nakajima, T. Ohori, D. Todaka, K. Nakashima, T. Hirayama, K. Shinozaki, and 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*, vol. 52, no. 12, p. 2136–2146, 2011.

[40] K. Nakashima and K. Yamaguchi-Shinozaki, "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*, vol. 39, no. 4, p. 221–229, 2005.

[41] G. K. Pandey, *Mechanism of plant hormone signaling under stress*. Wiley Blackwell, 2017.

[42] I. Ensminger, C. Yao-Yun Chang, and K. Bräutigam, "Advances in botanical research," *Advances in Botanical Research Land Plants - Trees*, vol. 74, p. 243, 2015.

[43] F. Li, M. Li, P. Wang, K. L. Cox, L. Duan, J. K. Dever, L. Shan, Z. Li, and P. He, "Regulation of cotton (gossypium hirsutum) drought responses by mitogen-activated protein (map) kinase cascade-mediated phosphorylation of ghwrky59," *New Phytologist*, vol. 215, p. 1462–1475, Dec 2017.

[44] S. P. Pandey and I. E. Somssich, "The role of wrky transcription factors in plant immunity," *Plant Physiology*, vol. 150, p. 1648–1655, Jun 2009.

[45] T. Eulgem, P. J. Rushton, S. Robatzek, and I. E. Somssich, "The wrky superfamily of plant transcription factors," *Trends in Plant Science*, vol. 5, no. 5, p. 199–206, 2000.

[46] M. Rahaie, G.-P. Xue, and P. M., "The role of transcription factors in wheat under different abiotic stresses," *Abiotic Stress - Plant Responses and Applications in Agriculture*, 2013.

[47] M. Bakshi and R. Oelmüller, "Wrky transcription factors," *Plant Signaling & Behavior*, vol. 9, no. 2, 2014.

[48] H. Chen, Z. Lai, J. Shi, Y. Xiao, Z. Chen, and X. Xu, "Roles of arabidopsis wrky18, wrky40 and wrky60 transcription factors in plant responses to abscisic acid and abiotic stress," *BMC Plant Biology*, vol. 10, no. 1, p. 281, 2010.

[49] A. Lahiri, P. S. Venkatasubramani, and A. Datta, "Bayesian modeling of plant drought resistance pathway," *BMC Plant Biology*, vol. 19, 12 2019.

[50] M. A. Mintgen, "Genetic analysis of plant responses to combinatorial stress in arabidopsis thaliana natural variation," master's thesis, Wageningen University, the Netherlands, Mar 2014.

[51] C. D. Ollas and I. C. Dodd, "Physiological impacts of aba–ja interactions under water-limitation," *Plant Molecular Biology*, vol. 91, no. 6, p. 641–650, 2016.

[52] N. Vijesh, S. K. Chakrabarti, and J. Sreekumar, "Modeling of gene regulatory networks: A review," *Journal of Biomedical Science and Engineering*, vol. 06, no. 02, 2013.

[53] G. Karlebach and R. Shamir, "Modelling and analysis of gene regulatory networks," *Nature Reviews Molecular Cell Biology*, vol. 9, 10 2008.

[54] H. Vundavilli, A. Datta, C. Sima, J. Hua, R. Lopes, and M. Bittner, "Cryptotanshinone Induces Cell Death in Lung Cancer by Targeting Aberrant Feedback Loops," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, 8 2020.

[55] R. Kapoor, A. Datta, C. Sima, J. Hua, R. Lopes, and M. L. Bittner, "A gaussian mixture-model exploiting pathway knowledge for dissecting cancer heterogeneity," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, p. 1–1, 2019.

[56] A. Lahiri, K. Rastogi, A. Datta, and E. M. Septiningsih, "Bayesian network analysis of lysine biosynthesis pathway in rice," *Inventions*, vol. 6, no. 2, 2021.

[57] H. Vundavilli, L. P. Tripathi, A. Datta, and K. Mizuguchi, "Network modeling and inference of peroxisome proliferator-activated receptor pathway in high fat diet-linked obesity," *Journal of Theoretical Biology*, Jun 2021.

[58] C. J. Needham, J. R. Bradford, A. J. Bulpitt, and D. R. Westhead, "A primer on learning in bayesian networks for computational biology," *PLoS Computational Biology*, vol. 3, no. 8, 2007.

[59] C. Sinoquet, "Probabilistic graphical models for next-generation genomics and genetics," *Probabilistic Graphical Models for Genetics, Genomics, and Postgenomics*, p. 3–29, 2014.

[60] K. Murphy, "A brief introduction to graphical models and bayesian networks," 1998.

[61] M. Scanagatta, A. Salmerón, and F. Stella, "A survey on bayesian network structure learning from data," *Progress in Artificial Intelligence*, vol. 8, no. 4, p. 425–439, 2019.

[62] N. L. Zhang, "Introduction to bayesian networks," 2018.

[63] C. P. Robert, "Bayesian computational tools," Jun 2013.

[64] J. Orlof, J. O. Bloom, and Jonathan, "Comparison of frequentist and bayesian inference.," 2014.

[65] A. Kak, "Ml, map, and bayesian — the holy trinity of parameter estimation and data prediction," Jan 2017.

[66] P. A. Jensen, "Project management," 2004.

[67] H. Vundavilli, A. Datta, C. Sima, J. Hua, R. Lopes, and M. Bittner, "Bayesian inference identifies combination therapeutic targets in breast cancer," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 9, p. 2684–2692, 2019.

[68] G. F. Cooper, "The computational complexity of probabilistic inference using bayesian belief networks," *Artificial Intelligence*, vol. 42, no. 2-3, p. 393–405, 1990.

[69] T. Lozano-Perez, "Inference in bayesian networks," 2006.

[70] M. Chiarandini, S. Russell, and P. Norvig, "Inference in bayesian networks," 2012.

[71] A. Menon, "Rejection sampling and likelihood weighting," Oct 2012.

[72] D. B. Lowry, T. L. Logan, L. Santuari, C. S. Hardtke, J. H. Richards, L. J. Derose-Wilson, J. K. Mckay, S. Sen, and T. E. Juenger, "Expression quantitative trait locus mapping across water availability environments reveals contrasting associations with genomic features in arabidopsis," *The Plant Cell*, vol. 25, no. 9, p. 3266–3279, 2013.

[73] National Library of Medicine, "National Center for Biotechnology Information," 1988.

[74] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[75] M. Scutari, "Learning bayesian networks with the bnlearn R package," *Journal of Statistical Software*, vol. 35, no. 3, pp. 1–22, 2010.

[76] B. Dombrecht, G. P. Xue, S. J. Sprague, J. A. Kirkegaard, J. J. Ross, J. B. Reid, G. P. Fitt, N. Sewelam, P. M. Schenk, J. M. Manners, and et al., "Myc2 differentially modulates diverse jasmonate-dependent functions in arabidopsis," *The Plant Cell*, vol. 19, no. 7, p. 2225–2245, 2007.

[77] Y. Wu, Z. Deng, J. Lai, Y. Zhang, C. Yang, B. Yin, Q. Zhao, L. Zhang, Y. Li, C. Yang, and et al., "Dual function of arabidopsis ataf1 in abiotic and biotic stress responses," *Cell Research*, vol. 19, no. 11, p. 1279–1290, 2009.

[78] K.-I. Seo, J.-H. Lee, C. D. Nezames, S. Zhong, E. Song, M.-O. Byun, and X. W. Deng, "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*, vol. 26, no. 2, p. 695–711, 2014.

[79] K. Kazan and J. M. Manners, "Myc2: The master in action," *Molecular Plant*, vol. 6, no. 3, p. 686–703, 2013.

[80] Y. Li, X. Yang, and X. Li, "Role of jasmonate signaling pathway in resistance to dehydration stress in arabidopsis," *Acta Physiologiae Plantarum*, vol. 41, no. 6, 2019.

[81] N. Liu and Z. Avramova, "Molecular mechanism of the priming by jasmonic acid of specific dehydration stress response genes in arabidopsis," *Epigenetics & Chromatin*, vol. 9, no. 1, 2016.

[82] P.-L. Lu, N.-Z. Chen, R. An, Z. Su, B.-S. Qi, F. Ren, J. Chen, and X.-C. Wang, "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*, vol. 63, p. 289–305, Oct 2006.

[83] R. Scholl and M. Anderson, "Arabidopsis biological resource center," *Plant Molecular Biology Reporter*, vol. 12, p. 242–244, Sep 1994.

[84] C. Mu, L. Zhou, L. Shan, F. Li, and Z. Li, "Phosphatase ghds ptp 3a interacts with annexin protein gh ann 8b to reversely regulate salt tolerance in cotton (gossypium spp.)," *New Phytologist*, vol. 223, p. 1856–1872, Jun 2019.

[85] B. Alberts, D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential Cell Biology*. Garland Science, 3 ed., 2010.

[86] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, "The shape and structure of protein," in *Molecular Biology of the Cell*, Garland Science, 4 ed., 3 2002.

[87] M. Lopez and S. Mohiuddin, "Biochemistry, Essential Amino Acids.," *StatPearls Publishing*, 4 2020.

[88] J. P. F. D'Mello, "Amino acids as multifunctional molecules.," in *Amino Acids in Animal Nutrition*, pp. 2–2, CABI Publishing, 2 ed., 4 2003.

[89] J. Hoffman and M. Falvo, "Protein - Which is Best?," *Journal of Sports Science and Medicine*, 9 2004.

[90] D. Tien Lea, H. Duc Chua, and N. Quynh Lea, "Improving Nutritional Quality of Plant Proteins Through Genetic Engineering," *Current Genomics*, vol. 17, 3 2016.

[91] C. Stencel and C. Dobbins, "Report Offers New Eating and Physical Activity Targets To Reduce Chronic Disease Risk," 9 2002.

[92] Y. Zha and Q. Qian, "Protein Nutrition and Malnutrition in CKD and ESRD," *Nutrients*, vol. 9, 2 2017.

[93] National Research Council, *Recommended Dietary Allowances*. Washington, D.C.: National Academies Press, 10 ed., 1 1989.

[94] A. Titchenal, S. Hara, N. Arceo Caacbay, W. Meinke-Lau, Y.-Y. Yang, M. Ksinoa Fialkowski Revilla, J. Draper, G. Langfelder, C. Gibby, C. Nicole Chun, and A. Calabrese, *Human Nutrition*. University of Hawaii at Manoa Food Science and Human Nutrition Program, 2020 ed., 2020.

[95] M. Henchion, M. Hayes, A. Mullen, M. Fenelon, and B. Tiwari, "Future Protein Supply and Demand: Strategies and Factors Influencing a Sustainable Equilibrium," *Foods*, vol. 6, 7 2017.

[96] A. Vasileska and G. Rechkoska, "Global and Regional Food Consumption Patterns and Trends," *Procedia - Social and Behavioral Sciences*, vol. 44, 2012.

[97] I. Berrazaga, V. Micard, M. Gueugneau, and S. Walrand, "The Role of the Anabolic Properties of Plant- versus Animal-Based Protein Sources in Supporting Muscle Mass Maintenance: A Critical Review," *Nutrients*, vol. 11, 8 2019.

[98] E. de Gavelle, J.-F. Huneau, C. Bianchi, E. Verger, and F. Mariotti, "Protein Adequacy Is Primarily a Matter of Protein Quantity, Not Quality: Modeling an Increase in Plant:Animal Protein Ratio in French Adults," *Nutrients*, vol. 9, 12 2017.

[99] I. Abete, D. Romaguera, A. R. Vieira, A. Lopez de Munain, and T. Norat, "Association between total, processed, red and white meat consumption and all-cause, CVD and IHD mortality: a meta-analysis of cohort studies," *British Journal of Nutrition*, vol. 112, 9 2014.

[100] D. Demeyer, B. Mertens, S. De Smet, and M. Ulens, "Mechanisms Linking Colorectal Cancer to the Consumption of (Processed) Red Meat: A Review," *Critical Reviews in Food Science and Nutrition*, vol. 56, 12 2016.

[101] V. S. Malik, Y. Li, D. K. Tobias, A. Pan, and F. B. Hu, "Dietary Protein Intake and Risk of Type 2 Diabetes in US Men and Women," *American Journal of Epidemiology*, vol. 183, 4 2016.

[102] The Food and Agriculture Organization of the United Nations, "Livestock solutions for climate change," tech. rep., United Nations, 2017.

[103] United Nations, "World population projected to reach 9.8 billion in 2050, and 11.2 billion in 2100," 6 2017.

[104] L. Day, "Proteins from land plants – Potential resources for human nutrition and food security," *Trends in Food Science & Technology*, vol. 32, 7 2013.

[105] M. W. Rosegrant, N. Leach, and R. V. Gerpacio, "Alternative futures for world cereal and meat consumption," *Proceedings of the Nutrition Society*, vol. 58, 5 1999.

[106] D. J. Millward and A. A. Jackson, "Protein/energy ratios of current diets in developed and developing countries compared with a safe protein/energy ratio: implications for recommended protein and amino acid intakes," *Public Health Nutrition*, vol. 7, 5 2004.

[107] M. Kusano, Z. Yang, Y. Okazaki, R. Nakabayashi, A. Fukushima, and K. Saito, "Using Metabolomic Approaches to Explore Chemical Diversity in Rice," *Molecular Plant*, vol. 8, 1 2015.

[108] G. Galili and R. Amir, "Fortifying plants with the essential amino acids lysine and methionine to improve nutritional quality," *Plant Biotechnology Journal*, vol. 11, 2 2013.

[109] W. Wang and G. Galili, "Transgenic high-lysine rice – a realistic solution to malnutrition?," *Journal of Experimental Botany*, vol. 67, 7 2016.

[110] G. Galili, H. Karchi, O. Shaul, A. Perl, A. Cahana, I. B.-T. Tzchori, X. Z. Zhu, and S. Galili, "Production of transgenic plants containing elevated levels of lysine and threonine," *Biochemical Society Transactions*, vol. 22, 11 1994.

[111] D. Grigg, "The pattern of world protein consumption," *Geoforum*, vol. 26, 2 1995.

[112] B. O. Juliano and The Food and Agriculture Organization of the United Nations, "World rice production compared to other cereals," in *Rice in human nutrition*, Rome: International Rice Research Institute of the United Nations, 1993.

[113] S. Muthayya, J. D. Sugimoto, S. Montgomery, and G. F. Maberly, "An overview of global rice production, supply, trade, and consumption," *Annals of the New York Academy of Sciences*, vol. 1324, 9 2014.

[114] T. Kawakatsu and F. Takaiwa, "Differences in Transcriptional Regulatory Mechanisms Functioning for Free Lysine Content and Seed Storage Protein Accumulation in Rice Grain," *Plant and Cell Physiology*, vol. 51, 12 2010.

[115] P. Arruda, E. L. Kemper, F. Papes, and A. Leite, "Regulation of lysine catabolism in higher plants," *Trends in Plant Science*, vol. 5, 8 2000.

[116] A. Frizzi, S. Huang, L. A. Gilbertson, T. A. Armstrong, M. H. Luethy, and T. M. Malvar, "Modifying lysine biosynthesis and catabolism in corn with a single bifunctional expression/silencing transgene cassette," *Plant Biotechnology Journal*, 2007.

[117] P. Arruda and P. Barreto, "Lysine catabolism through the saccharopine pathway: Enzymes and intermediates involved in plant responses to abiotic and biotic stress," *Frontiers in Plant Science*, vol. 11, 2020.

[118] X. Long, Q. Liu, M. Chan, Q. Wang, and S. S. M. Sun, "Metabolic engineering and profiling of rice with increased lysine," *Plant Biotechnology Journal*, vol. 11, 5 2013.

[119] Q.-q. Yang, C.-q. Zhang, M.-l. Chan, D.-s. Zhao, J.-z. Chen, Q. Wang, Q.-f. Li, H.-x. Yu, M.-h. Gu, S. S.-m. Sun, and Q.-q. Liu, "Biofortification of rice with the essential amino acid

lysine: molecular characterization, nutritional evaluation, and field performance," *Journal of Experimental Botany*, vol. 67, 7 2016.

[120] X. Zhu and G. Galili, "Increased Lysine Synthesis Coupled with a Knockout of Its Catabolism Synergistically Boosts Lysine Content and Also Transregulates the Metabolism of Other Amino Acids in Arabidopsis Seeds," *The Plant Cell*, vol. 15, 4 2003.

[121] R. Angelovici, A. Fait, A. R. Fernie, and G. Galili, "A seed high-lysine trait is negatively associated with the TCA cycle and slows down Arabidopsis seed germination," *New Phytologist*, vol. 189, 1 2011.

[122] I. B.-T. Tzchori, A. Perl, and G. Galili, "Lysine and threonine metabolism are subject to complex patterns of regulation in Arabidopsis," *Plant Molecular Biology*, vol. 32, 11 1996.

[123] M. Rappe, "CRISPR Plants: New Non-GMO Method to Edit Plants," 5 2020.

[124] A. M. Shew, L. L. Nalley, H. A. Snell, R. M. Nayga, and B. L. Dixon, "Crispr versus gmos: Public acceptance and valuation," *Global Food Security*, vol. 19, p. 71–80, 2018.

[125] K. Rastogi, O. Ibarra, M. Molina, M. Faion-Molina, M. Thomson, and E. M. Septiningsih, "Using crispr/cas9 genome editing to increase lysine levels in rice," in *ASA-CSSA-SSSA International Annual Meeting, San Antonio, TX*, (San Antonio, TX), 2019.

[126] M. Kanehisa, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, 1 2000.

[127] Y. Kawahara, M. de la Bastide, J. P. Hamilton, H. Kanamori, W. R. McCombie, S. Ouyang, D. C. Schwartz, T. Tanaka, J. Wu, S. Zhou, K. L. Childs, R. M. Davidson, H. Lin, L. Quesada-Ocampo, B. Vaillancourt, H. Sakai, S. S. Lee, J. Kim, H. Numa, T. Itoh, C. R. Buell, and T. Matsumoto, "Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data," *Rice*, vol. 6, no. 1, 2013.

[128] A. Lahiri, L. Zhou, P. He, and A. Datta, "Detecting drought regulators using stochastic inference in bayesian networks," *PLOS ONE*, vol. 16, no. 8, 2021.

[129] M. Sheng, M. Tang, H. Chen, B. Yang, F. Zhang, and Y. Huang, "Influence of arbuscular mycorrhizae on photosynthesis and water status of maize plants under salt stress," *Mycorrhiza*, vol. 18, 9 2008.

[130] R. Tisarum, C. Theerawitaya, T. Samphumphuang, K. Polispitak, P. Thongpoem, H. P. Singh, and S. Cha-um, "Alleviation of Salt Stress in Upland Rice (Oryza sativa L. ssp. indica cv. Leum Pua) Using Arbuscular Mycorrhizal Fungi Inoculation," *Frontiers in Plant Science*, vol. 11, 3 2020.

[131] I. N. B. L. Reddy, B.-K. Kim, I.-S. Yoon, K.-H. Kim, and T.-R. Kwon, "Salt Tolerance in Rice: Focus on Mechanisms and Approaches," *Rice Science*, vol. 24, 5 2017.

[132] N. Kakar, S. H. Jumaa, E. D. Redoña, M. L. Warburton, and K. R. Reddy, "Evaluating rice for salinity using pot-culture provides a systematic tolerance assessment at the seedling stage," *Rice*, vol. 12, 12 2019.

[133] V. Deshmukh, S. P. Mankar, C. Muthukumar, P. Divahar, A. Bharathi, H. B. Thomas, A. Rajurkar, R. Sellamuthu, R. Poornima, S. Senthivel, and et al., "Genome-wide consistent molecular markers associated with phenology, plant production and root traits in diverse rice (oryza sativa l.) accessions under drought in rainfed target populations of the environment," *Current Science*, vol. 114, no. 02, p. 329–340, 2018.

[134] S. Razzaque, S. M. Elias, T. Haque, S. Biswas, G. M. N. A. Jewel, S. Rahman, X. Weng, A. M. Ismail, H. Walia, T. E. Juenger, and Z. I. Seraj, "Gene Expression analysis associated with salt stress in a reciprocally crossed rice population," *Scientific Reports*, vol. 9, 12 2019.

[135] G. Stewart and F. Larher, "Accumulation of amino acids and related compounds in relation to environmental stress," *Amino Acids and Derivatives*, p. 609–635, 1980.

[136] Q. Ali, H.-U.-R. Athar, M. Z. Haider, S. Shahid, N. Aslam, F. Shehzad, J. Naseem, R. Ashraf, A. Ali, S. M. Hussain, and et al., "Role of amino acids in improving abiotic stress tolerance to plants," *Plant Tolerance to Environmental Stress*, p. 175–204, 2019.

[137] M. Wang, C. Liu, S. Li, D. Zhu, Q. Zhao, and J. Yu, "Improved nutritive quality and salt resistance in transgenic maize by simultaneously overexpression of a natural lysine-rich protein gene, sbglr, and an erf transcription factor gene, tsrf1," *International Journal of Molecular Sciences*, vol. 14, no. 5, p. 9459–9474, 2013.

[138] S. Saeedipour, "Stress-induced changes in the free amino acid composition of two wheat cultivars with difference in drought resistance," *African Journal Of Biotechnology*, vol. 11, no. 40, 2012.

[139] C. A. Jackson, D. M. Castro, G.-A. Saldi, R. Bonneau, and D. Gresham, "Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments," *eLife*, vol. 9, 1 2020.

[140] E. H. Davidson and D. H. Erwin, "Gene regulatory networks and the evolution of animal body plans," *Science*, vol. 311, 2 2006.

[141] M. Kærn, W. J. Blake, and J. Collins, "The engineering of gene regulatory networks," *Annual Review of Biomedical Engineering*, vol. 5, no. 1, 2003.

[142] A. Bonnaffoux, U. Herbach, A. Richard, A. Guillemin, S. Gonin-Giraud, P.-A. Gros, and O. Gandrillon, "WASABI: a dynamic iterative framework for gene regulatory network inference," *BMC Bioinformatics*, vol. 20, 12 2019.

[143] Y. Sun and J. R. Dinneny, "Q&A: How do gene regulatory networks control environmental responses in plants?," *BMC Biology*, vol. 16, 12 2018.

[144] F. Emmert-Streib, M. Dehmer, and B. Haibe-Kains, "Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks," *Frontiers in Cell and Developmental Biology*, vol. 2, 8 2014.

[145] O. A. Arshad and A. Datta, "Towards targeted combinatorial therapy design for the treatment of castration-resistant prostate cancer," *BMC Bioinformatics*, vol. 18, 3 2017.

[146] H. Vundavilli, A. Datta, C. Sima, J. Hua, R. Lopes, and M. Bittner, "Targeting oncogenic mutations in colorectal cancer using cryptotanshinone," *PLOS ONE*, vol. 16, 2 2021.

[147] T. Timmermann, B. González, and G. A. Ruz, "Reconstruction of a gene regulatory network of the induced systemic resistance defense response in Arabidopsis using boolean networks," *BMC Bioinformatics*, vol. 21, 12 2020.

[148] P. S. Venkat, K. R. Narayanan, and A. Datta, "A Bayesian Network-Based Approach to Selection of Intervention Points in the Mitogen-Activated Protein Kinase Plant Defense Response Pathway," *Journal of Computational Biology*, vol. 24, 4 2017.

[149] C. Sinoquet and R. Mourad, "Probabilistic Graphical Models for Next-generation Genomics and Genetics," in *Probabilistic Graphical Models for Genetics, Genomics, and Postgenomics*, pp. 1–16, Oxford University Press, 12 2014.

[150] D. Heckerman and J. Breese, "Causal independence for probability assessment and inference using Bayesian networks," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 26, no. 6, 1996.

[151] M. E. Borsuk, C. A. Stow, and K. H. Reckhow, "A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis," *Ecological Modelling*, vol. 173, 4 2004.

[152] V. Sevinc, O. Kucuk, and M. Goltas, "A Bayesian network model for prediction and analysis of possible forest fire causes," *Forest Ecology and Management*, vol. 457, 2 2020.

[153] R. E. Neapolitan, *Learning Bayesian networks*. Prentice Hall, 2004.

[154] R. Kabli, F. Herrmann, and J. McCall, "A chain-model genetic algorithm for Bayesian network structure learning," in *Proceedings of the 9th annual conference on Genetic and evolutionary computation - GECCO '07*, (New York, New York, USA), ACM Press, 2007.

[155] M. Scanagatta, A. Salmerón, and F. Stella, "A survey on Bayesian network structure learning from data," *Progress in Artificial Intelligence*, vol. 8, 12 2019.

[156] N. L. Zhang, "COMP538: Introduction to Bayesian Networks Lecture 6: Parameter Learning in Bayesian Networks," 2008.

[157] D. Spiegelhalter, "Lecture 6: Bayesian estimation," 1 2016.

[158] Z. Fan and A. Chin, "Lecture 20 — Bayesian analysis," 2016.

[159] J. Orlof and J. Bloom, "Comparison of frequentist and Bayesian inference," 2014.

[160] A. J. Storkey, "Machine Learning and Pattern Recognition: Note on Dirichlet Multinomial," 2020.

[161] H. Liu and L. Wasserman, "Bayesian Inference," in *Statistical Machine Learning*, pp. 299–305, Carnegie Mellon University, 2014.

[162] D. Alvares, C. Armero, and A. Forte, "What Does Objective Mean in a Dirichlet-multinomial Process?," *International Statistical Review*, vol. 86, 4 2018.

[163] D. Kelly and C. Atwood, "Finding a minimally informative Dirichlet prior distribution using least squares," *Reliability Engineering & System Safety*, vol. 96, 3 2011.

[164] C. P. Robert, "Bayesian computational tools," *Annual Review of Statistics and Its Application*, vol. 1, pp. 153–177, 4 2014.

[165] D. Koller and F. Friedman, "Bayesian Parameter Estimation," in *Probabilistic Graphical Models*, pp. 738–739, The MIT Press, 7 2009.

[166] C. Bielza and P. Larrañaga, "Bayesian networks in neuroscience: a survey," *Frontiers in Computational Neuroscience*, vol. 8, 10 2014.

[167] S. E. Shimony, "Finding MAPs for belief networks is NP-hard," *Artificial Intelligence*, vol. 68, pp. 399–410, 8 1994.

[168] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann Publishers, INC, 1 ed., 1988.

[169] T. Lozano-Pérez and K. Kaelbling, "6.825 Techniques in Artificial Intelligence (SMA 5504)," 2002.

[170] H. Guo and W. Hsu, "A Survey of Algorithms for Real-Time Bayesian Network Inference," tech. rep., Association for the Advancement of Artificial Intelligence, 10 2002.

[171] M. Shwe and G. Cooper, "An empirical analysis of likelihood-weighting simulation on a large, multiply connected medical belief network," *Computers and Biomedical Research*, vol. 24, pp. 453–475, 10 1991.

[172] S. Russell and Norvig Peter, *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3 ed., 2010.

[173] National Library of Medicine, "National Center for Biotechnology Information," 1988.

[174] R. Edgar, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, 1 2002.

[175] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, "NCBI GEO: archive for functional genomics data sets—update," *Nucleic Acids Research*, vol. 41, 11 2012.

[176] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, dec 2014.

[177] H. Varet, L. Brillet-Guéguen, J.-Y. Coppée, and M.-A. Dillies, "SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data," *PLOS ONE*, vol. 11, jun 2016.

[178] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi, "A survey of best practices for RNA-seq data analysis," *Genome Biology*, vol. 17, dec 2016.

[179] G. Wen, "A Simple Process of RNA-Sequence Analyses by Hisat2, Htseq and DESeq2," in *Proceedings of the 2017 International Conference on Biomedical Engineering and Bioinformatics - ICBEB 2017*, (New York, New York, USA), ACM Press, 2017.

[180] H.-H. Jeong and Z. Liu, "Are HHV-6A and HHV-7 Really More Abundant in Alzheimer's Disease?," *Neuron*, vol. 104, dec 2019.

[181] R. Nagarajan, M. Scutari, and S. Lèbre, *Bayesian Networks in R*. Springer New York, 2013.

[182] M. Scutari, "Learning bayesian networks with the bnlearn r package," *Journal of Statistical Software*, vol. 35, no. 3, 2010.

[183] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.