BAYESIAN MODELING OF PLANT DROUGHT RESISTANCE PATHWAY

A Thesis

by

ADITYA LAHIRI

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

MASTER OF SCIENCE

Chair of Committee,	Aniruddha Datta
Co-Chair of Committee,	Krishna Narayanan
Committee Members,	Yang Shen
	Charles D. Johnson
Head of Department,	Miroslav M. Begovic

August 2018

Major Subject: Electrical Engineering

Copyright 2018 Aditya Lahiri

ABSTRACT

Plants are sessile organisms and are unable to relocate to favorable locations under extreme environmental conditions, and hence they have no choice but to acclimate and eventually adapt to the severe conditions to ensure their survival. With climate change affecting the environment adversely, it is of utmost importance to make plants and crops robust enough to withstand harsh conditions and safeguard global food production. As traditional methods of bolstering plant defense against stressful conditions come to their biological limit, we require newer methods that can allow us to strengthen the plant's internal defense mechanism. This motivated us to look into the genetic networks of plants. In this thesis, we lay out a method to analyze genetic networks in plants that are activated under abiotic stress, specifically drought conditions. This method is based on the analysis of Bayesian networks and should ultimately help in finding genes in the genetic networks of the plant that play a key role in its defense response against drought.

The WRKY transcription factor is well known for its role in plant defense against biotic stresses, but recent studies have shed light on its activity against abiotic stresses such as drought. Therefore, it is logical to study the various components of the WRKY gene network in order to maximize a plant's defense mechanism. The data used to learn the parameters of the Bayesian network consisted of both real world and synthetic data. The synthetic data was generated using the dependencies in the Bayesian network model. The network parameters were learned using a Bayesian approach and the frequentist approach. The estimated parameters are then used to build a Bayesian

decision network, where nodes are selected one at a time for intervention and the utility (score) for the upregulation of a downstream abiotic stress response gene is computed. The node that maximizes this utility is recommended for biological intervention.

DEDICATION

This thesis is dedicated to my parents, who believed in me and sent me halfway across the world to pursue my dreams and to my brother who has always been there to support and encourage me in my endeavors.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Aniruddha Datta who not only gave me the opportunity to work at his lab but also the independence to discover my research interests. I would also like to thank my former colleague and Texas A&M alumnus, Dr. Priyadharshini S. Venkatasubramani who brought me up to speed with the research work at the lab.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a thesis committee consisting of Professor Aniruddha Datta (advisor), Professor Krishna Narayanan (co-advisor), Professor Yang Shen from the Department of Electrical and Computer Engineering and by Professor Charles D. Johnson from the Department of Genomics and Bioinformatics Services.

All other work for the thesis was completed by the student independently.

Funding Sources

This work was made possible in part by the TEES-AgriLife Center for Bioinformatics and Genomic Systems Engineering (CBGSE) and in part by the National Science Foundation under Grant ECCS-1404314.

NOMENCLATURE

ABA	Abscisic Acid
BN	Bayesian Network
MLE	Maximum Likelihood Estimation

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1. INTRODUCTION

The global population is set to rise by 35% by the year 2050 and increasing crop yields to ensure food security has become a grand challenge [1]. The rise in temperature worldwide due to global warming has increased the risk of drought affecting crop yields and has further complicated this challenge. It is estimated that by the year 2100, the global drought affected area will rise from 15.4% to 44%, and the global crop yield will reduce by more than 50% by 2050 and by 90% in 2100 [2]. The unprecedented rise in worldwide population accompanied by a rise in demand for crop supply comes at a time when traditional approaches of maximizing crop production are coming to their biological limits. Hence, developing drought resistant crops has become a global priority to ensure food security. Fortunately, plants have multiple innate stress sensing mechanisms that are able to detect unfavorable changes in the environment and deploy appropriate defense responses. Therefore, it is of great interest to understand the genetic networks behind a plant's defense mechanism in order to augment its genetic yield potential while reducing its susceptibility to harsh conditions.

Abscisic acid (ABA) is a well-known plant hormone that is induced under drought stress conditions and regulates a plant's gene expression through the action of transcription factors [3,4]. The family of WRKY transcription factors is traditionally associated with plant defense mechanisms against pathogens; however, many recent studies have highlighted WRKY's role in abiotic stress responses [5,6,7]. Since WRKY is one of the largest families of transcription factors in plants with such diverse roles in

plant defense mechanisms, it will be practical to model the interaction among various components of the WRKY's signaling pathway to gain valuable insights into these interactions [8]. In this paper, we use Bayesian networks (BN) to model the ABA-induced WRKY transcription factor network. We then apply a utility based inference technique to determine the significant regulators of drought stress response genes in the BN. This approach allows us to integrate existing biological knowledge into our model.

2. BIOLOGICAL BACKGROUND & WRKY SIGNALING PATHWAY

Similar to the way adrenaline functions as a stress hormone in animals, plants respond to harsh environmental changes, pathogen attacks or wounding by secreting plant hormones such as Abscisic Acid, Cytokinins, Salicylic Acid and Ethylene to trigger its own defense mechanisms. In the context of plants, drought is characterized by the unavailability of water which can prevent plants from performing basic survival processes such as photosynthesis. When a plant faces water deficit conditions, it can defend itself either by the process of avoidance or tolerance. In the case of avoidance, a plant may complete its life cycle in the wet season. Whereas in the case of tolerance, the plant may initially acclimate to the change in conditions by introducing reversible changes into its physiology through altering its gene expression; however, if the drought conditions still persist, then the plant passes its altered genes to its next generation so that these new generations of plants are already adapted to the drought conditions [9].

To adopt either of these defense mechanisms the plant must undergo a process of signal transduction when it gets the initial cue of droughts such as a drop in the water potential in the apoplast and a rise in the ion concentration [9]. All these signals along with many others cause a rapid rise in the level of the plant hormone ABA which acts as a stress sensor and it subsequently activates secondary messengers such as Ca^{2+}, Reactive Oxygen Species (ROS) and Cyclic Adenosine Monophosphate (cAMP). These secondary messengers turn on their own respective signaling pathways (e.g. MAPK, CDPK, etc.) where protein phosphorylation (addition of phosphate $(PO_4)^{3-}$) and

dephosphorylation may take place via the actions of kinases (enzymes) and phosphatases (enzymes) respectively [10]. Following the signaling actions of kinases and phosphatases, transcription factors are either activated or deactivated to regulate downstream gene expression [10]. Transcription factors are proteins that bind to a specific DNA sequence of the gene(s) in order to activate or deactivate them. Finally, transcription factors are directly responsible for turning on the stress response genes and turning off any other nonessential genes.

Each family of transcription factors such as WRKY, bZIP, and NAC regulates a large number of genes. Hence, learning their activities is critical for understanding plant stress response mechanisms. WRKY is a large family of transcription factors and has roles in plant defense mechanisms against both abiotic and biotic stress. Until recently, the role of WRKY in dealing with abiotic stresses was not as extensively explored as in the case of biotic stresses, because of which there is a lack of available experimental data [3]. In this paper, we are interested in studying the interaction among various members of the WRKY transcription factor signaling pathway (figure 1) which are rapidly induced by ABA under drought stress. Learning these interactions will give us deeper insights into the functioning of this pathway which will further aid us in developing intervention strategies for breeding drought-resistant plants.

It has been shown that the transcription factors WRKY18, WRKY40, and WRKY60 are induced by ABA under water deficit and salt stress conditions (Chen et al.) [11]. Furthermore, it has also been reported that WRKY18 and WRKY60 have positive sensitivity for ABA in inhibition of seed germination, root growth and

enhancing plant sensitivity to water deficit stress; in contrast, WRK40 antagonizes WRKY18 and WRKY60 to affect a plant's ABA sensitivity and abiotic stress responses (Chen et al.). Experiments were carried out with WRKY18 and WRKY40 deficient mutants which showed that the expression of WRKY60 was negligible and indicated that WRKY18 and WRKY40 directly induced WRKY 60 by recognizing a cluster of W-BOX sequences in the promoter of WRKY60 (Chen et al.). In addition to the various regulatory behaviors of these three WRKY transcription factors, it has been noted that these three transcription factors not only interact with themselves to form three homocomplexes; but also, interact amongst each other to form heterocomplexes [12].

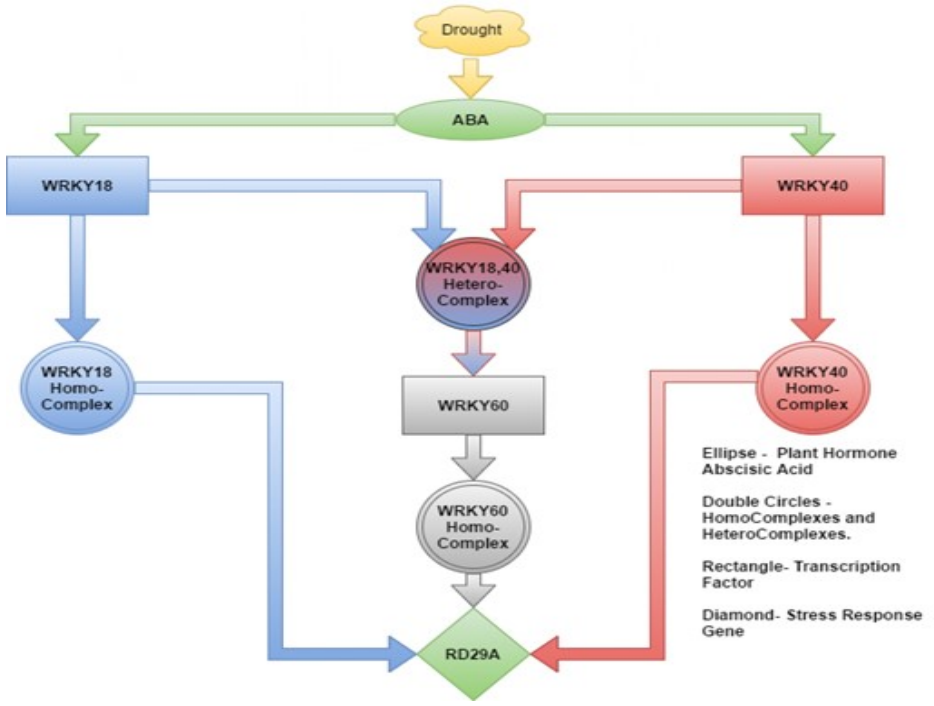

Figure 1: The induction of WRKY transcription factor signaling pathway by ABA.

3.WRKY BAYESIAN NETWORK MODELING

Biological networks are inherently tortuous and stochastic in nature. It is often difficult to interpret the multivariate interaction among different components of the network. A BN is a directed acyclic graph that determines the conditional decomposition of the joint probability distributions of a set of random variables in the network and thus simplifies the computation of their joint probability distribution (Sinoquet and Mourad) [13]. Therefore, we are interested in using BNs to model the interactions in a biological network as they provide a clean and compact framework for representing the joint probability distributions and for drawing inferences from these networks [14]. Inspection of BNs can help enhance our beliefs about relationships among different elements in the network and provide insights into the causality of the network.

In this thesis, we will be modeling the WRKY signaling pathway (figure 1) involved in the drought stress responses of the model plant, Arabidopsis. Based on the signal transduction pathway outlined in figure 1, we have constructed a BN as shown figure 2. Each circular node (A,B,C,..., H) represents a gene, transcription factor or protein complex and every directed edge between the nodes represents a causal relationship that exists in the WRKY signaling pathway. Attached to every node is a rectangle which represents the parameter or local probability model associated with that node. For instance, for node C the $\theta_{C|A,B}$ represents the conditional probability density of node C given its parent nodes A and B. These parameters can be learned from data and are important in the understanding the overall graph structure.

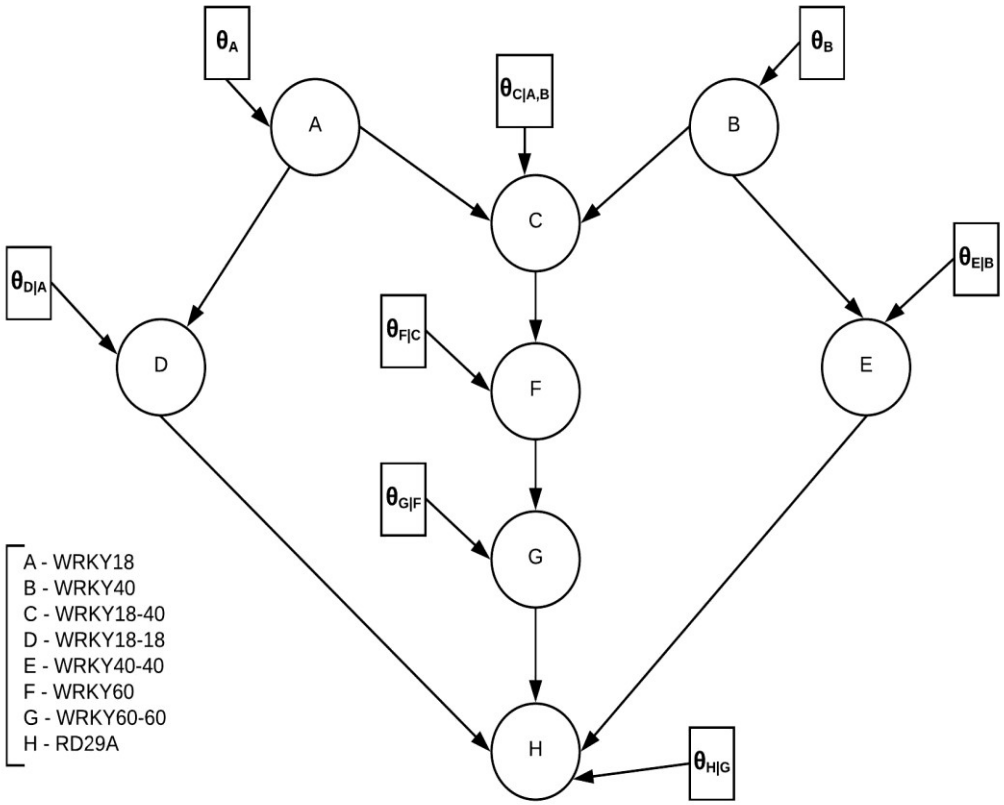

Figure 2: BN model of WRKY signaling pathway with conditional probabilities depicted in rectangles.

4. PARAMETER ESTIMATION IN BAYESIAN NETWORKS

Depending on the availability of prior knowledge in a given application, one may use a frequentist approach or the Bayesian approach for learning the parameters of a BN. Frequentist approaches such as Maximum Likelihood Estimation (MLE) assume that the parameter being learned is fixed and produce a point estimate without taking into account prior information. On the other hand, Bayesian Estimation treats the parameter as a random variable and uses the data and prior distribution of the parameter to obtain the parameter's posterior distribution. Furthermore, the Bayesian approach takes into account the problem of zero probability estimates which may affect the learning algorithm. Bayesian methods provide a non-zero probability estimate even when the prior information follows a uniform distribution (non-informative prior). This is because the posterior belief is being governed both by data and prior knowledge and hence zero estimates of probability will only be associated with nonoccurrence of an event. However, the Bayesian estimation process is computationally challenging as it requires performing integration in order to obtain the probability of the evidence (data). Due to this reason, we will exploit the concept of conjugate priors for a given likelihood function in the process of Bayesian estimation [15].

In the BN in figure 2, we assume that each of the nodes X in the BN can attain only binary values, X= 0 or X= 1. When X=1 for a node, it indicates that the gene, transcription factor or protein complex represented by that node is activated whereas if X=0, it indicates just the opposite (gene, transcription factor or protein complex is

inhibited). This formulation allows us to model the state of each node in the network, given the state of its parent nodes, using a Bernoulli distribution.

Suppose that there are N nodes in the BN and let θ_X be the probability that X=1 (success) and 1-θ_X be the probability that X=0 (failure). Assume that we make n (> 0) observations regarding the state of each node and let k be the number of times the state of a node is 1. We further assume that the sequence of random variables $X_1, X_2, \ldots, X_n$ obtained after n observations for each node to be independent and identically distributed. So, the probability distribution of a node given its parent nodes (Pa(X)) follows a Binomial distribution and is given by:

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

$$\text{Binomial}(n, \theta_X) = \frac{n!}{(n-k)!\,k!} * \theta_X^k * (1-\theta_X)^{n-k} \tag{1.b}$$

To estimate the posterior distribution, we need to define the prior over the parameter θ_X for our model. Since the likelihood function associated with our model is binomial, we choose the prior distribution to follow a Beta distribution with some shape parameters (α_X, β_X) and this results in the representation:

$$\theta_X \sim \text{Beta}(\alpha_X, \beta_X) \tag{2}$$

Due to the modeling of the priors as a beta distribution under Binomial likelihood, it follows from the properties of conjugate families that the posteriors will

also follow a beta distribution with shape parameter (α'_X, β'_X) [15]. In our model, the posterior distribution of the parameter θ_X is given by:

$$P(\theta_X|X) \sim \text{Beta}(\alpha'_X, \beta'_X) \tag{3}$$

where $\alpha'_X = (\alpha_X + k)$ and $\beta'_X = (\beta_X + n - k)$. The expected value of this distribution is given by:

$$E(\theta_X|X) = \frac{\alpha'_X}{(\alpha'_X + \beta'_X)} \tag{4}$$

We can use experimental data to iteratively update α_X and β_X to obtain the posterior distribution. With more data, the posterior distribution will converge towards the actual posterior distribution. We modeled the prior as a Beta under Binomial likelihood and, this allowed us to obtain a closed form solution for the posterior. Other non-conjugate priors may be used; but, a closed form solution may not be guaranteed. Note that this approach gives us the marginal and conditional posterior distribution associated with every node and not their probabilities (θ_X or $\theta_{Y|X}$). In this thesis for the purpose of learning these probabilities, we approximate the probabilities by the expected value (equation (4)) of the posterior distribution for their respective nodes. Furthermore, we also learn the probabilities using the frequentist approach of MLE, in order to compare the final results, we get by using both the approaches. Ideally, when data is abundant the Bayesian approach and MLE estimate converge to the same point[16]. The marginal probabilities and the conditional probabilities for binary random variables can be estimated using MLE by equation (5) and (6) respectively.

$$\theta_{X_1} = \frac{M[X^1]}{M[X^1]+M[X^0]} \quad (5)$$

$$\theta_{Y_1|X_0} = \frac{M[Y^1,X^0]}{M[Y^1,X^0]+M[Y^0,X^0]} \quad (6)$$

Where $M[X^1]$ is number of times the random variable X is 1, $M[X^0]$ is the number of time X is 0, $M[Y^1, X^0]$ is the number of times X is 0 and Y is 1 and $M[Y^0, X^0]$ is the number of times X is 0 and Y is 0.

A key assumption we make in the BN modeling is that the joint distribution for the set of nodes factorizes according to the BN in figure 2. This assumption basically implies that dependencies in the biological structure are reflected in the data from which we are learning the network parameters. One can employ constraint-based or score-based learning techniques to derive the graph structure from data and then subsequently learn the network parameters[17]. However, in the context of this thesis, we avoid learning the graph structure as publicly available experimental data is highly limited for the WRKY transcription factor under abiotic stress and also our network contains protein complexes (nodes C, D, E, and G) for which expression data doesn't exist alongside with gene expression data (nodes A, B, F, and H). Generally, datasets that contain gene expression data do not contain expression data for protein complexes and vice versa. Hence synthetic data were generated for the protein complexes using the dependencies in the BN and the experimental data for other non-protein complex nodes in the network for which data were available.

5. UTILITY BASED INFERENCE IN BAYESIAN NETWORKS

After the network parameters or the local probabilities associated with every node are inferred from the data the BN has sufficient information for carrying out inference. Our objective is to find a single node in the WRKY BN that maximizes the upregulation of the downstream expression of the drought resistant gene. In other words, we are interested in finding a single node (nodes A-G) in BN which when up or downregulated will maximize the chances of our stress response gene (node H) being upregulated.

There are multiple ways to perform inference in a BN. Pearl's message passing algorithm is favored whenever we have a singly connected graph as it allows us to perform exact inference [18]. However, the BN in consideration here is not singly connected and also has loops which cannot be handled using Pearl's algorithm. There are other non-exact sampling-based techniques that require a large amount of data to provide reliable inference. Hence in this thesis, we have considered another type of approximate inference technique that computes a score, commonly known as an expected utility, based on an action taken at a specific node. Utility measures the efficacy of that action. To implement utilities into our BN, we first need to understand the concept of Bayesian decision networks and how we can create one from a BN.

In order to illustrate these concepts, consider the following example involving a simple BN as shown in figure 3.

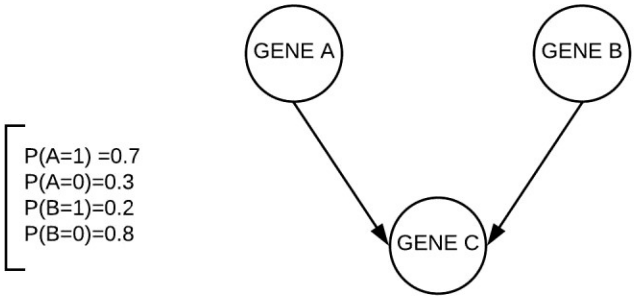

Figure 3: Example BN, with marginal probabilities of parent nodes.

The BN in figure 3 has three nodes gene A, gene B and gene C and we assume each can take on a binary value of 0 (inhibited) or 1 (activated). Gene A and gene B are parents nodes of gene C and have marginal probabilities associated with them as shown in figure 3. Also, let us assume that when gene A is active it activates gene C and when gene B is active it inhibits gene C. Based on this BN we construct a Bayesian decision network as shown in figure 4. The rectangular node acts as decision (action) node, the diamond-shaped node serves as a utility node and the circular nodes represent chance (nature or probabilistic) nodes. In this example we are interested in having gene C to take on the value of 1, this what the utility will measure. In this case we have the option to take action at either of the chance nodes gene A or gene B. Once we decide to either activate or inhibit the chance node, it no longer remains a chance node but becomes a deterministic node.

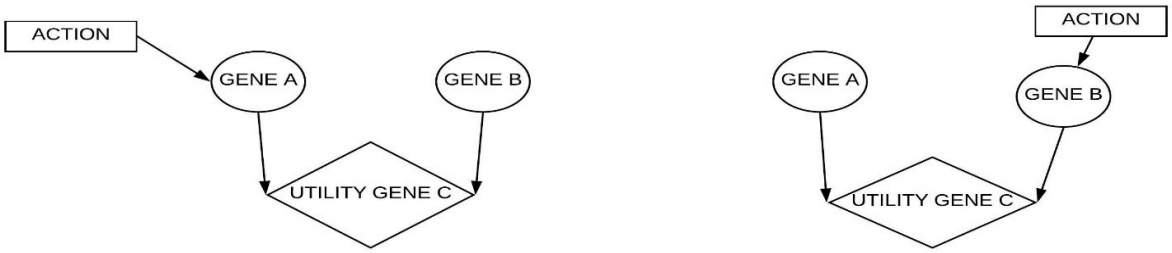

Figure 4: Bayesian Decision networks for intervention at gene A and at gene B.

Depending on the action taken the expected utility can be calculated by equation (7).

$$EU(A) = \sum_i P(O_i \mid A) * U(O_i) \quad (7)$$

where $P(O_i \mid A)$ represents the probabilities of the outcomes (O_i) that are consistent with action A, and $U(O_i)$ represents the utility value for that outcome under action A. The utility table is defined in table 1, where the first row represents the best-case scenario when gene A is active and gene B is inhibited and the last row represents the worst-case scenario when gene A is inhibited and gene B is active. The rest of the rows represent the other possible scenarios. The utility scores assigned are relative to best (highest utility) and worst (lowest utility) case scenarios, these values are not unique and can be redefined differently, however, the scores must reflect the scenario depicted in the decision network.

Table 1: Utilities for example Bayesian decision network

GENE A	GENE B	UTILITY GENE C
1	0	100
1	1	50
0	0	50
0	1	0

Using equation (7) we first calculate the expected utility for taking action at gene A as follows:

Case 1: Action taken: gene A is activated (A= 1).

$$EU(\text{Gene } A=1) = P(A=1,B=1 \mid A=1)*U(A=1, B=1) + P(A=1, B=0 \mid A=1)*U(A=1, B=0)$$

$$= P(B=1)*U(A=1, B=1) + P(B=0)*U(A=1, B=0)$$

$$= 0.2 * 50 + 0.8 * 100 = 90$$

Case 2: Action taken: gene A is inhibited (A= 0).

$$EU(\text{Gene } A=0) = P(A=0,B=1 \mid A=0)*U(A=0, B=1) + P(A=0,B=0 \mid A=0) * U(A=0,B=0)$$

$$= P(B=1) * U(A=0, B=1) + P(B=0) * U(A=0,B=0)$$

$$= 0.2 * 0 + 0.8 * 50 = 40$$

So, when gene A=1 or activated the expected utility is greater. Similarly, let us calculate the expected utilities for taking action at gene B.

Case 1: Action taken: gene B is activated (B= 1).

$$EU(\text{Gene } B=1) = P(A=1,B=1 \mid B=1) * U(A=1, B=1) + P(A=0,B=1 \mid B=1) * U(A=0, B=1)$$

$$= P(A=1) * U(A=1, B=1) + P(A=0) * U(A=0, B=1)$$

$$= 0.7 * 50 + 0.3 * 0 = 35$$

Case 2: Action taken: gene B is inhibited (B= 0).

$$EU(\text{Gene } B=0) = P(A=1,B=0 \mid B=0) * U(A=1, B=0) + P(A=0,B=0 \mid B=0) * U(A=0,B=0)$$

$$= P(A=1) * U(A=1, B=0) + P(A=0) * U(A=0,B=0)$$

$$= 0.7 * 100 + 0.3 * 50 = 85$$

Hence when gene B is inhibited the expected utility is larger. However, the utility of gene A being activated is larger than gene B being inhibited. So, we must select activating gene A over inhibiting gene B to maximize the chances of gene C being activated.

6. DATASET AND SIMULATION

To estimate the parameters and carry out utility calculation in the BN, we need to obtain data for WRKY transcription factor under drought stress condition. Since the WRKY transcription factor has only recently been implicated for its role in abiotic stress response, it is difficult to obtain large scale data that is publicly available. However, we were able to obtain real world microarray gene expression data for all the genes and transcription factors (Nodes A, B, F, and H) in the BN from the datasets GSE46365, GSE65046, and GSE76827 which are publicly available from the NCBI GEO database [19,20,21]. These datasets were individually normalized and binarized and aggregated into one composite dataset which contained 116 data points for each of the non protein complex nodes (genes and transcription factors). Once the real world data were binarized, they were used along with the dependencies in the BN to generate data for the protein complexes denoted by nodes C, D, E, and G. For example, in order to generate the dataset for node D the expression values for node A and node H were observed, i.e. all the parents and children of node D. If both node A and H were observed to be upregulated (state =1) node D was assigned deterministically to be upregulated (state =1), if both the nodes A and H were both observed to be downregulated (state =0) then node D was assigned deterministically to be downregulated (state =0). This is because we know from the biological literature (Chen et. al) that node A upregulates node D, and node D in turn upregulates Node H. If the expression status of node A was upregulated (state=1) and that of node H was downregulated (state=0) then node D was assigned a value of 1 (upregulated) with a probability larger than 0.5. This is because if node A is

upregulated it is highly likely that node D is also upregulated but not enough to counter the downregulatory effect of node E on node H which might have caused node H to be downregulated. The probability with which node D was upregulated was randomly selected from a set of discrete probability values of [0.6,0.7,0.8 0.9 and 1] where each value had an equally likely chance of being selected. Similarly, when node A=0 and node H=1 node D was probabilistically assigned a value of 0 (downregulated). In this fashion the data for nodes C, G, and E were also generated, so that these synthetic data reflected the network dependencies and the real data for the nonprotein complex nodes. Tables 2-5, pseudocode 1, given below along with the R code attached in the supplemental section further explain the data generation for nodes D, E, G and C. Generally, gene expression datasets do not contain protein-protein interaction data which is needed for avoiding generating synthetic data for protein complexes in our network. Though we can find protein-protein interaction datasets, however those datasets will not contain gene expression data, so in order to circumvent this issue, we considered generating synthetic data for the protein complexes in our BN.

Once synthetic data for all the protein complex nodes were generated, they were aggregated along with real world data in a single dataset. This dataset was used for the purpose of estimating the network parameters using the Bayesian approach and the maximum likelihood approach as outlined in section 4. For the Bayesian approach, the prior for every node was first initialized to a beta (1,1) distribution which is a uniform distribution over the interval [0,1]. Using the data and equations (3) the posterior distribution for every node was updated and the expected values, computed using

equation (4). The expected values were approximated to be the conditional probability for the nodes. A separate set of parameters were learned using the MLE approach as well using equations (5) and (6).

Then, the utilities for single point intervention were computed. The utility node was set at node H, and the utility analysis was carried out to find single intervention points that upregulated node H. A utility table (table 6) was defined based on the Bayesian decision network. The utility at node H depended directly on nodes D, G and E. The best case scenario was when nodes D and G were upregulated and node E was downregulated, whereas the worst case scenario was when nodes D and G were downregulated and node E was upregulated. These scenarios were representative of the actual biological processes in the network and the utility scores were defined relative to these scenarios, with a high utility score being favorable. Simulations were carried out using R software [22] and the utility calculations were done using Netica [23].

Table 2: Synthetic Data generation for Node D

NODE A	**NODE H**	**NODE D**
1	1	1
1	0	Assign (Value=1)
0	1	Assign (Value=0)
0	0	0

Table 3: Synthetic Data generation for Node E

NODE B	NODE H	NODE E
1	1	Assign (Value =1)
1	0	1
0	1	0
0	0	Assign (Value =0)

Table 4: Synthetic Data generation for Node G

NODE F	NODE H	NODE G
1	1	1
1	0	Assign (Value=1)
0	1	Assign(Value =0)
0	0	0

Table 5: Synthetic Data generation for Node C

NODE A	NODE B	NODE F	NODE C
0	0	0	0
0	0	1	Assign(Value =0)
0	1	0	Assign(Value =0)
0	1	1	Assign(Value =1)
1	0	0	Assign(Value =0)
1	1	0	Assign(Value =1)
1	0	1	Assign(Value =1)
1	1	1	1

Pseudocode1: To assign a node a value of 0 or 1.

1.Function = Assign(Value)

2. Define a set of probabilities = {0.6,0.7,0.8,0.9,1}

3. Probability (P) = sample one value randomly from Probability Set, with every element of the set having an equally likely chance of being picked

4. If Value is 0

5. Assign the Node a Value of 0 with the Probability "P" selected in step 4, a value of 1 is assigned with probability 1-P.

6. Else if the Value is 1

7. Assign the Node a Value of 1 with the Probability "P" selected in step 4, a value of 0 is assigned with probability 1-P.

Table 6: Utility values used to calculate the maximum expected utilities in figure 6 and figure 7.

Node G (Status)	**Node D (Status)**	**Node E(Status)**	**Utility at Node H**
1	1	1	50
1	1	0	100
1	0	1	10
1	0	0	50
0	1	1	10
0	1	0	50
0	0	1	0
0	0	0	10

7. RESULTS

The divided bar plot in figure 5 represents the activation and inhibition status for every gene in the BN, after the data have been preprocessed. Table 7 displays the conditional probabilities estimated using the Bayesian and MLE approaches. The maximum expected utilities using the parameters obtained from the Bayesian and MLE approaches are displayed in figures 6 and in 7 respectively.

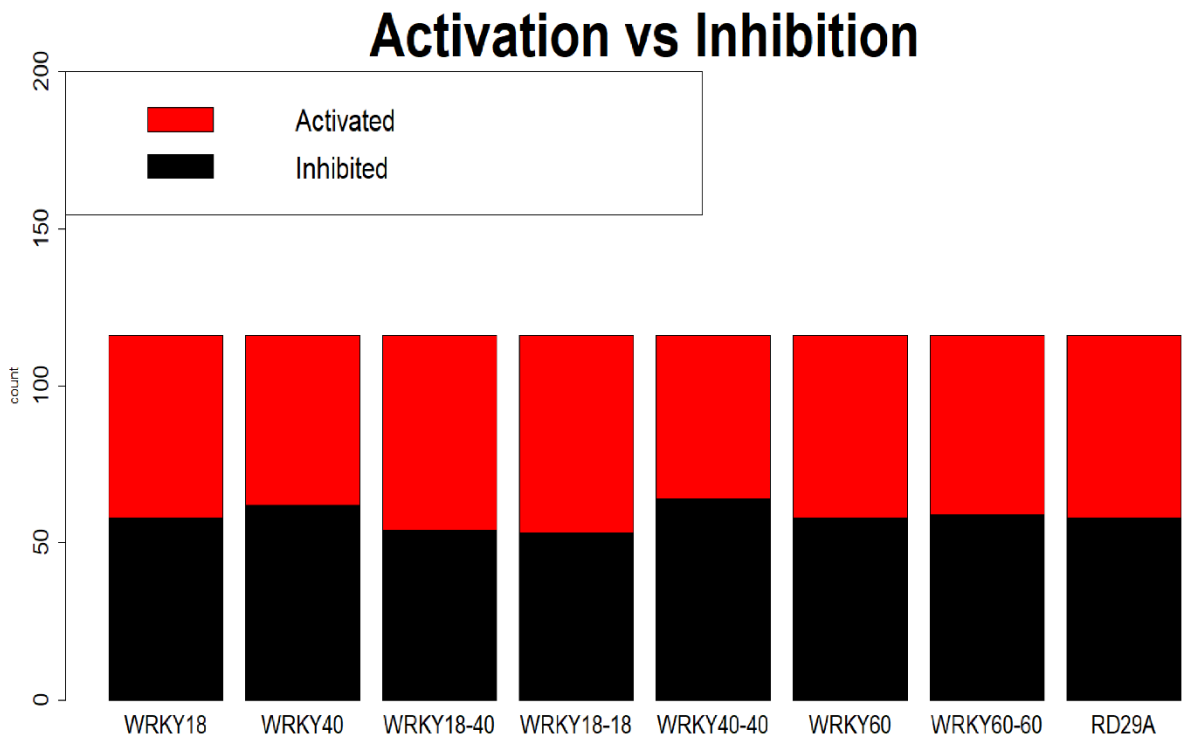

Figure 5: Node activation vs inhibition plot. Red region in the plot shows the counts for which a particular node was activated and the black region in the plot shows the counts for which the node was inhibited. The dataset contained 116 data points per node.

Table 7: Marginal and Conditional Probabilities using the Bayesian and MLE approaches.

Local Probabilities	**Bayesian Approach**	**MLE Approach**
$P(A_1)$	0.50	0.50
$P(B_1)$	0.466	0.465
$P(C_1\|A_1, B_1)$	0.905	0.925
$P(C_1\|A_0, B_1)$	0.625	0.645
$P(C_1\|A_1, B_0)$	0.60	0.611
$P(C_1\|A_0, B_0)$	0.13	0.113
$P(D_1\|A_1)$	0.983	1
$P(D_1\|A_0)$	0.10	0.086
$P(E_1\|B_1)$	0.857	0.870
$P(E_1\|B_0)$	0.093	0.080
$P(F_1\|C_1)$	0.766	0.774
$P(F_1\|C_0)$	0.196	0.185
$P(G_1\|F_1)$	0.867	0.879
$P(G_1\|F_0)$	0.117	0.103

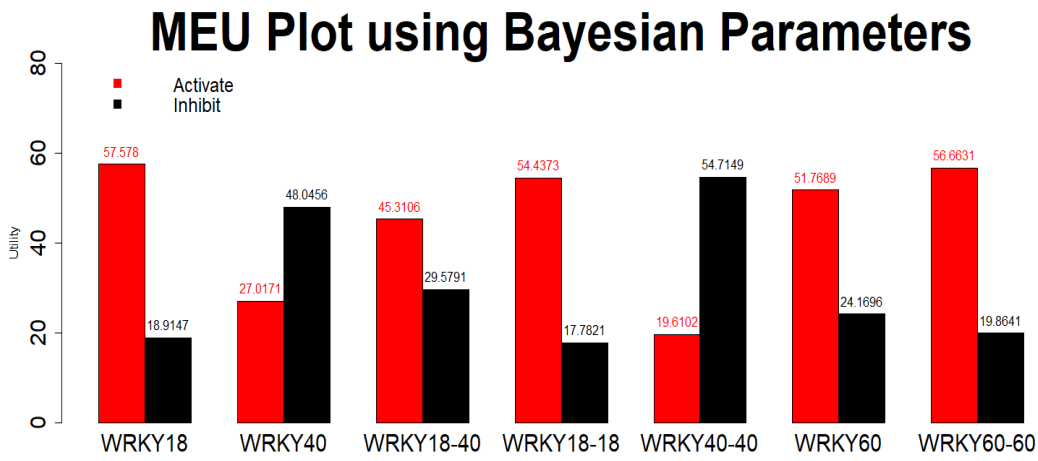

Figure 6: Maximum expected utility values when using parameters from the Bayesian approach. Red bar plot shows utility for activating that node and black bar plot shows the utility for inhibiting that node.

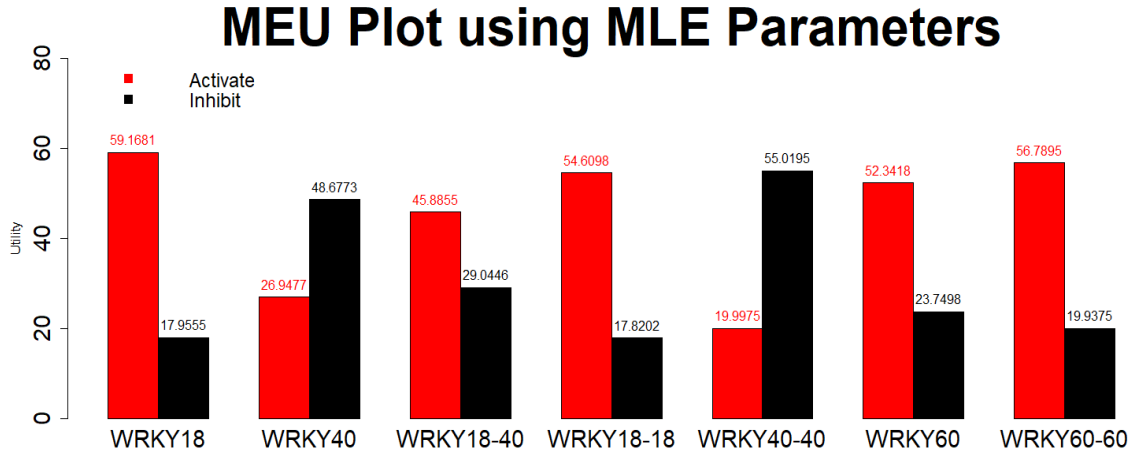

Figure 7: Maximum expected utility values when parameters are estimated using MLE. Red bar plot shows utility for activating that node and black bar plot shows the utility for inhibiting that node.

From the utility analysis using both the Bayesian and MLE approaches, we find that WRKY18 has the highest utility score for its activation. This means that the

upregulation of WRKY18 is the most effective single-gene intervention in bringing about the upregulation of the downstream drought stress response gene. This result is consistent with the biological literature which suggests that WRKY18 has a positive sensitivity to ABA under drought stress conditions and plays a critical role in the upregulation of downstream gene expression. We can also see from the bar plots in figures 6 and 7 that the second-best point for intervention is at the protein complex WRKY60-60 which also upregulates the expression of the downstream drought stress response genes. Also, consistent with the literature we see that both WRKY40 and WRKY40-40 have high utilities for inhibition, as they are responsible for downregulating the downstream drought stress response gene. We also see that our utility scores in both the Bayesian and MLE approaches are comparable which is due to the fact that the estimated probabilities (table 7) using both the approaches are very similar.

8. SUMMARY AND CONCLUSION

In this thesis, we presented the WRKY signaling pathway, which is traditionally associated with plant defense response against biotic stresses, but recently it has been shown to play a significant role in plant defense response against abiotic stresses such as drought. Due to its diverse role in plant defense, it was an interesting pathway choice to investigate. We modeled the WRKY pathway using a BN, where every node in the network represented a gene, transcription factor or protein complex from the pathway and every edge between the nodes represents a causal relationship that exists in the pathway. Associated with each node is a conditional or marginal probability which represents the probability with which the node is activated or inhibited. For our analysis, we assumed that nodes in the network can only assume binary values of 1 (activation) or 0 (inhibition). Since a BN can capture both the causal biological relationships and the probabilistic nature of biological pathways, it was an ideal choice for modeling purposes.

In order to learn the parameters in the network, we used real world gene expression data and generated synthetic data which reflected the network dependencies. To estimate the local conditional and marginal probabilities both a Bayesian approach and a frequentist approach were used. In the Bayesian approach, we assumed every node to have a prior distribution of Beta (1,1) (uniform distribution in the range [0,1]) which signified we had no prior knowledge about our model. Since our likelihood followed a binomial distribution, we were able to use a closed form formula through the properties

of conjugate families to arrive at a posterior distribution for each node. The expected value of the posterior distribution was used as an estimate for the local probabilities. We selected conjugate families in order to simplify our calculations and arrive at a closed form solution, however, it may not always be the best choice to select a conjugate prior. If sufficient information is available the prior can be modeled using non-conjugate family distributions and the posterior can be estimated using (Markov-Chain-Monte-Carlo) MCMC techniques, although, this may be computationally expensive. In the frequentist approach, we simply employed the maximum likelihood estimate to obtain the local probabilities. The probabilities obtained using both the methods were found to be very similar to each other.

Once the parameters from each method were learned, the task of inferring the best node for intervention was carried out using the concept of utilities. We used a non-exact inference technique in our model as we could not employ exact techniques such as Pearl's message passing algorithm in our Bayesian network as the former works only for singly connected and loopless networks. Also, the number of data points was quite limited which made the choice of utility for the purpose of inference quite sensible as opposed to data intensive sampling-based inference techniques. The utility analysis carried out using parameters from the Bayesian and MLE approaches revealed that WRKY18 served as the most potent node for intervention, and upregulating WRKY18 would further upregulate the downstream stress response genes in the WRKY signaling pathway. This result was consistent with the biological literature which says that WRKY18 actively upregulates the gene expression of drought response genes under

drought conditions. Our next step in this research will be to explore and implement more informative priors in the Bayesian parameter estimation approach rather than the Beta (1,1) prior that we have used here. We would also like to investigate other signaling pathways that are implicated in plant defense response against drought and find the key the regulators in those networks and compare their efficacy to that of WRKY18. We are also interested in expanding our research to networks consisting of multiple stress response regulators and intervention nodes.

REFERENCES

[1] "US Census Bureau, Demographic Internet Staff ," *International Programs, International Data Base*, June 2011.[Online] Available: https://www.census.gov/ population /international/data/idb/worldpopgraph.php. [Accessed: 07-Apr-2017].

[2] Y. Li, W. Ye, M. Wang, and X. Yan, "Climate change and drought: a risk assessment of crop-yield impacts," *Climate Research*, vol. 39, pp.31–46, 2009.

[3] R. Finkelstein, "Abscisic Acid Synthesis and Response," *The Arabidopsis Book*, vol.11, 2013.

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

[5] S.P. Pandey and I.E. Somssich, "The Role of WRKY Transcription Factors in Plant Immunity," *Plant Physiology*, vol. 150, no. 4, pp.1648–1655, Jun. 2009.

[6] 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, pp. 199–206, Jan. 2000.

[7] L. Chen, Y. Song, S. Li, L. Zhang, C. Zou, and D. Yu, "The role of WRKY transcription factors in plant abiotic stresses," *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, vol.1819, no.2, pp.120–128, 2012.

[8] G. He, J. Xu, Y. Wang, J. Liu, P. Li, M. Chen, Y. Ma and Z. Xu, "Drought-responsive WRKY transcription factor genes TaWRKY1 and TaWRKY33 from wheat confer drought and/or heat resistance in Arabidopsis," *BMC Plant Biology*, vol.16, no.1, 2016.

[9] L. Taiz, E. Zeiger, I.M. Møller, and A. Murphy, *Plant physiology and development*, 6th ed. Sinauer Associates, Inc, Sunderland, MA, 2015.

[10] H. Wang, H. Wang, H. Shao, and X. Tang, "Recent Advances in Utilizing Transcription Factors to Improve Plant Abiotic Stress Tolerance by Transgenic Technology," *Frontiers in Plant Science*, vol.7, Sep. 2016.

[11] 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.

[12] X. Xu, C. Chen, B. Fan, and Z. Chen, "Physical and Functional Interactions between Pathogen-Induced Arabidopsis WRKY18, WRKY40, and WRKY60 Transcription Factors," *The Plant Cell Online*, vol.18, no.5, pp.1310–1326, Jan. 2006.

[13] C. Sinoquet and M. Raphaël. *Probabilistic graphical models for genetics, genomics and postgenomics*, Oxford: Oxford University Press, 2014.

[14] CJ. Needham, JR. Bradford, AJ. Bulpitt, and DR. Westhead, "A Primer on Learning in Bayesian Networks for Computational Biology," *PLoS Computational Biology*, vol. 3, no.8, 2007.

[15] A. Kak, "ML, MAP, and Bayesian — The Holy Trinity of Parameter Estimation and Data Prediction," 04-Jan-2017. [Online]. Available: https://engineering.purdue.edu/kak/Tutorials/Trinity.pdf. [Accessed: 07-Apr-2017].

[16] R. J. Williams, "Maximum Likelihood vs. Bayesian Parameter Estimation." [Online]. Available: http://www.ccs.neu.edu/home/rjw/csg220/lectures/MLE-vs-Bayes.pdf [Accessed: 07-Apr-2017].

[17] D. Koller and N. Friedman, *Probabilistic graphical models principles, and techniques*. Cambridge, Mass.: MIT Press, 2012.

[18] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Francisco, CA: Morgan Kaufmann, 2009.

[19] J.-M. Kim, T. K. To, A. Matsui, K. Tanoi, N. I. Kobayashi, F. Matsuda, Y. Habu, D. Ogawa, T. Sakamoto, S. Matsunaga, K. Bashir, S. Rasheed, M. Ando, H. Takeda, K. Kawaura, M. Kusano, A. Fukushima, T. A. Endo, T. Kuromori, J. Ishida, T. Morosawa, M. Tanaka, C. Torii, Y. Takebayashi, H. Sakakibara, Y. Ogihara, K. Saito, K. Shinozaki, A. Devoto, and M. Seki, "Acetate-mediated novel survival strategy against drought in plants," *Nature Plants*, vol. 3, no. 7, p. 17097, 2017.

[20] U. Bechtold, C. A. Penfold, D. J. Jenkins, R. Legaie, J. D. Moore, T. Lawson, J. S. Matthews, S. R. Vialet-Chabrand, L. Baxter, S. Subramaniam, R. Hickman, H. Florance, C. Sambles, D. L. Salmon, R. Feil, L. Bowden, C. Hill, N. R. Baker, J. E. Lunn, B. Finkenstädt, A. Mead, V. Buchanan-Wollaston, J. Beynon, D. A. Rand, D. L. Wild, K. J. Denby, S. Ott, N. Smirnoff, and P. M. Mullineaux, "Time-Series Transcriptomics Reveals That AGAMOUS-LIKE22 Affects Primary Metabolism and Developmental Processes in Drought-Stressed Arabidopsis," *The Plant Cell*, vol. 28, no. 2, pp. 345–366, 2016.

[21] S. Rasheed, K. Bashir, A. Matsui, M. Tanaka, and M. Seki, “Transcriptomic Analysis of Soil-Grown Arabidopsis thaliana Roots and Shoots in Response to a Drought Stress,” *Frontiers in Plant Science*, vol. 7, 2016.

[22] R Development Core Team , “R: A language and environment for statistical computing. R Foundation for Statistical Computing,” Vienna, Austria. ISBN 3-900051-07-0, 2018.

[23] Norsys Software Corp, Netica 6.04 Bayesian Network Software from Norsys,2018.

APPENDIX

Supplemental File Name	**Content**	**Access**
PlotofUtilities.R	R code for plotting utilities.	This file can be accessed using the programming software RStudio.
Parameter_Estimation.R	R code for estimating parameters. Also serves as the main execution file.	This file can be accessed using the programming software RStudio.
MLE_Estimates.R	R code for finding MLE.	This file can be accessed using the programming software RStudio.
minmax_normalize.R	R code to normalize the data.	This file can be accessed using the programming software RStudio.
calc_shape_param.R	R code to calculate shape parameters of Beta distribution.	This file can be accessed using the programming software RStudio.
binarize_mean_median.R	R code to binarize the data using mean or median	This file can be accessed using the programming software RStudio.
assign_value.R	R code to generate synthetic data	This file can be accessed using the programming software RStudio.