# Detecting Key Nodes in the MAPK Plant Defense Pathway using a Bayesian Network Based Approach

Priyadharshini S. Venkatasubramani, *Student Member, IEEE,* Krishna R. Narayanan, *Fellow, IEEE,* and Aniruddha Datta, *Fellow, IEEE*

## I. INTRODUCTION

The world's growing population has made food security a global concern. One of the key factors that impact food security is the loss of crops due to diseases caused by plant pathogens. Many plant-associated microbes are parasitic organisms that impair plant growth and reproduction. Plants possess inherent immune receptors that detect the presence of microbial pathogens and trigger defense responses against a multitude of harmful pathogens. But, due to adaptive evolution and the fight for survival, pathogens have developed various strategies to invade host plant tissue undetected and cause infections that render food crops unfit for consumption. A number of plant biological studies have implicated the Mitogen Activated Protein Kinase (MAPK) cascade in plant cell signaling as the point of convergence of various stress stimuli. Hence, there has been a lot of interest in targeting specific components of the MAPK pathway in an effort to improve disease resistance in crop plants. A mathematical model of the interactions among the components of the MAPK pathway is critical to obtain a better understanding of the nature of these interactions and to advocate intervention strategies for breeding disease-resistant crops.

A number of methods have been proposed for the selection of 'important' genes from a large set of genes. The most widely mentioned among these are based on classification of huge volumes of genotypic and phenotypic data using numerous data mining techniques such as support vector machines (SVM), regression techniques, random forest, clustering, gene ranking, etc. However, the genes selected from these methods may be irrelevant to the biological phenomena being studied or it may even be difficult to ascribe biological connotations to the genes selected from many of these methods, since gene expression patterns alone may not convey the essence of interactions among genes.

Another popular approach for analyzing and making sense of microarray gene expression data is the construction of genetic regulatory networks. There has been a lot of interest in looking at the interaction among genes in a holistic manner because the activity of genes are not isolated or independent of each other. Hence, network perspectives are integral to our understanding of biological interactions and to channel this insight in order to develop successful intervention methodologies.

## II. GENE SELECTION

Given a genetic regulatory network, it is important to differentiate between those genes that have a major influence on the regulation of the child gene and those that only have a minor influence. Biologically, a gene with stronger influence on another has the potential to regulate the dynamics of the network, overshadowing the effect of other genes that have minimal influences. Many such biological relationships are known to exist. For instance, the activation of the p53 oncogene, which is a well-known tumor suppressor, actively leads to the expression of various genes that promote apoptosis, whereas p73, another tumor suppressor belonging to the same group of signaling pathway elements is less effective in activating apoptotic genes.

We consider the problem of deducing the effect of individual genes on the behavior of the network in a statistical framework. We make use of biological information from the literature to develop a Bayesian network (BN) and provide a novel method to select significant nodes in the network using a decision theoretic approach. The proposed method is applied to the analysis of the Mitogen Activated Protein Kinase (MAPK) pathway in plant defense response to pathogens. Results from applying the method to experimental data show that the proposed approach is effective in selecting genes that play crucial roles in the biological phenomenon being studied.

This work was supported 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.

The authors are with the TEES-AgriLife Center for Bioinformatics and Genomic Systems Engineering, 101-A Gateway Blvd., 3004 TAMU, College Station, TX 77845, USA, and the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA. (e-mail: priya26@email.tamu.edu).

Our objective is to maintain some reporter nodes at certain states in the Bayesian network. In such a scenario, we have different options for the choice of point(s) of intervention to prod the network towards a specific behavior which we are interested in. For each gene, we have an associated probability distribution over its possible states and its influence on a desired node in the network. Therefore, the gene section problem is essentially a problem of decision making under uncertainty. In order to find optimal decisions, we assign numerical utilities to all possible outcomes and then choose the decisions that result in maximal utility value.

Utility is a subjective notion and the design of the utility function depends on the objective of the action (gene intervention in our case) and the nature and preference for tools available to cause the action. For instance, in the case of gene intervention, gene knockouts may be easier to implement than gene activation, and may have stronger downstream effects and hence have more utility to the biologist, depending on the type of genetic network under consideration. An important consideration in our approach is that biologists are always looking for one crucial lever gene that can be manipulated in real experiments rather than a bunch of genes that have some influence on the desired response, since it is both time-consuming and expensive to test the effect of manipulating multiple genes.

## III. BAYESIAN NETWORK APPROACH

BNs are a natural fit to the problem of gene selection since they can be used to represent causal relationships, similar to the nature of relations in biological signaling pathways. The state of each node in a Bayesian network is described by a probability distribution. The nodes that have no parents have marginal probability distributions whereas the other nodes have conditional probability distributions describing their state, conditioned on the states of their parents. Estimating network parameters using a Bayesian approach requires us to define priors over the parameters. The observed data points form the likelihood in the Bayesian setting. Given the distribution of the parameters, the observed values for a node are independent of each other. For simplicity, we choose a beta-binomial distribution to model our BN. The priors are beta distributed and the likelihood follows a binomial distribution. Consequently, the conditional posterior probability distributions of the nodes are beta distributed. As more data are observed, we can update the values of the network parameters so that the posterior probabilities approach the true underlying distribution.

## IV. RESULTS

We specifically look at the interaction of the plant *Arabidopsis Thaliana* with bacterial pathogens. Arabidopsis is widely used as a model plant since it is easily manipulated and genetically tractable. We use the plant-pathogen signaling pathway from the biological literature to build the graph of our BN. Three real gene expression data sets from the NCBI GEO database were used for simulations. Each of these datasets contain gene expression data obtained from experiments where the Arabidopsis plant was exposed to bacterial molecules and the gene expression changes induced by this stimulus were measured using microarrays.

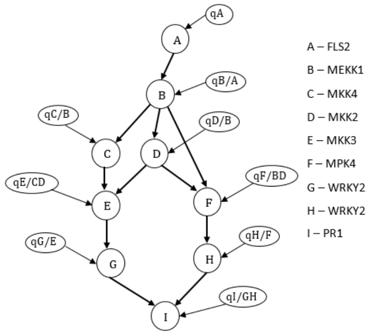

Fig. 1. BN model illustrating the conditional dependencies in our network

The parameters that encode the probability distributions of nodes in the network are shown in Fig. 1. The Bayesian network and the conditional probabilities associated with the Bayesian network are used to model the utility function in order to select significant nodes in the network. We consider two important factors in the design of the utility function: the first factor is the effect of the gene on the utility variable; and second, the fact that the effort involved in flipping a gene depends on its predisposition to be activated or inactivated, given that other genes that influence it are in a certain state. Let us consider that our goal is to achieve transcription of the defense response gene Pathogenesis-related protein 1 (PR1). This is the leaf node of the network shown in Fig. 1. We evaluate the utility obtained by manipulating every gene in the network (other than PR1) to achieve this goal, and select the genes with maximal expected utility to be the points of intervention. The inference from the application of the gene selection algorithm is that WRKY25 is the most preferred node for intervention. WRKY25 belongs to the WRKY I group of proteins, which play significant roles in plant development and response to biotic and abiotic stresses in various plant species.