ABSTRACT

NAIK, PUNITH PAVOOR. A Systems Biology approach towards understanding the Regulation of Monolignol Biosynthesis in *Populus trichocharpa.* (Under the direction of Dr. Joel Ducoste).

Lignin is the most abundant polymer after cellulose and hemicelluloses found naturally in the secondary cell walls of all vascular plant. Lignin is entangled with cellulose and hemicelluloses forming an impermeable matrix. The primary purpose of lignin is to transport nutrients, provide protection against pathogens and provide upright support. With the recent push towards utilization of plant biomass as a source of biofuel, the rigidity of lignin unfortunately acts as a barrier in the utilization of high energy sugars like hemicellulose and cellulose. With the advancement in high throughput technologies, the biosynthetic pathway of monolignol continues to be experimentally characterized. However, since most of the biological networks are characterized by highly non-linear interactions with multiple substrates competing with multifunction enzymes and proteins interacting with each other forming complexes, it may be challenging to predict the effect of genetic perturbations on the monolignol biosynthetic pathway. This gap in knowledge could be filled with the development of mathematical modeling that characterizes these non-linear interactions. The overall objective of this research was to develop mathematical models to enhance the understanding of monolignol biosynthesis in *Populus trichocharpa.* These novel mathematical models can then be used to predict the effect of genetic perturbations on the monolignol biosynthetic pathway, especially the lignin content and structure. The models can also be used to gain insights about regulatory control, generate testable hypotheses, and

genetically engineer plants with desired lignin content and structure after experimental validation.

As a result of this extensive modeling effort, the following outcomes of this thesis have been drawn:

1. The model was able to assess the role of protein-protein interactions on the lignin biosynthetic pathway as well as the lignin composition and structure. The analysis of the model suggested that the presence of the protein-protein interactions improves the robustness of the pathway.
2. Analysis of the RNA Seq data revealed the existence of a modular structure of the genes that regulate proteins involved in the monolignol biosynthesis pathway and its role on the lignin composition and structure. The analysis of this modular structure suggested that the plants maintain this compartmentalized structure to improve their resiliency against perturbations. The modular information was incorporated into the model which was then used to quantify the effect of perturbation of the proteins involved in each module on the lignin structure and composition. The analysis of the model suggest that the compartmentalization of genes that regulate proteins involved in lignin biosynthesis improves the stability of the pathway against perturbations. The same genes within these modules can also be used as viable targets, which may enable researchers to more effectively tailor the lignin structure.
3. The predictive model was validated using the experimental data and showed that the model was able to account for 74% of the variations in the S:G ratio. For cases where detailed kinetics information is not available, a purely data driven

model could be developed. In this study, an Artificial Neural Network (ANN) was used to predict the variation of S:G ratio and the total lignin content (S+G) as a function of protein concentration. The ANN model was able to account for 80% of the variation in S:G ratio and 74% of the variation in (S+G) values.

4. Finally, a network inference algorithm was used to develop a gene regulatory network. The gene regulatory network was then combined with the metabolic network model to predict the role of perturbation of Transcription Factors (TF's) and genes on the monolignol biosynthetic pathway as well as the lignin composition and structure. The complete model was able to predict that the upregulation of MYB TF`s result's in an increase in the S:G ratio and a reduction in total lignin content. The knowledge of the interactions between the various transcription factors, genes and proteins would enhance our knowledge about the regulation of the monolignol biosynthetic pathway and identify the key transcription factors that affect the lignin composition and structure.

A Systems Biology approach towards understanding the Regulation of Monolignol Biosynthesis in *Populus trichocharpa*

by
Punith Pavoor Naik

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Civil Engineering

Raleigh, North Carolina

2016

APPROVED BY:

_____	_____
Dr. Joel Ducoste	Dr. Cranos Williams
_____	_____
Dr. Ranji Ranjithan	Dr. Francis L. de los Reyes III
_____	
Dr. Vincent Chiang	

DEDICATION

To my wife Sree, for all your love and support

BIOGRAPHY

Punith was born and raised in Mangalore, a coastal town in the state of Karnataka, India. He completed his undergraduate degree in Chemical Engineering from National Institute of Technology Karnataka, Surathkal. He obtained his M.S in Chemical Engineering from Mississippi State University, where he worked on the remediation of nuclear waste at the Savannah River Site, under the direction of Dr. Rebecca Toghiani. He also developed a keen interest in mathematical modeling during the course of this project. This led him to pursue doctoral research on modeling of complex systems, under the direction of Dr. Joel Ducoste at North Carolina State University. He is currently working as a Quantitative Analytics Senior for Federal Home Loans and Mortgage Corporation (Freddie Mac), where he reviews and validates various capital and counterparty credit risk models.

ACKNOWLEDGEMENTS

My dad always encouraged and supported me, no matter how trivial the task was or how impossible the task may be. Even though he is not here today, I am sure he is happy for all the things I have accomplished. Growing up, my mom made sure that I stayed focused and out of trouble. I want to thank my mom for all the sacrifices she made so that I could be successful in life. Hence, I wouldn't have been here if it weren't for my parents. I would also like to thank my uncle, Dr.Bellipady Chinmaya Rai and my aunt, Pavithra Rai for their unwavering support. I am also indebted to the love and guidance of my late grandmother, Indira Rai. It is because of you guys, I am the person I am today.

Dr. Joel Ducoste has always believed in me and has pushed me to do better. I am thankful for your patience with me. Dr. Williams provided valuable guidance on alternative modeling strategies which helped me move ahead with my dissertation. I learnt a lot from both of you. Dr. Chiang, Dr. Sederoff, Dr. Ranjithan, and Dr. Reyes were instrumental in shaping the course of my doctoral work. This dissertation wouldn't be where it is today, without all your guidance and your direction. Thank you for your time and effort.

Dr. Jack Wang helped me better understand the biology behind my modeling efforts. I am thankful to Jack, "Cotton", and other members of the Forest biotechnology group at NCSU for feeding me excellent data, which helped me develop and validate

mathematical models for the lignin biosynthesis pathway. Special thanks to my friends, Aishwarya, Abhijit, Kirsten, Siddharth, Senganal, Sashikanth and Venu, who made my stay in Raleigh enjoyable.

I am grateful to Ozzie, my four legged friend for his companionship and for entertaining me with his antics. Last but not the least, I wouldn't be here without the love and support of my best friend and love of my life, Sree. You were there for me whenever I needed you .Words cannot express my gratitude for everything you have done over the years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Background

With increasing global energy demands, the search for alternative fuel sources have dramatically increased over the past decade. Plant biomass has been identified as a viable source of biofuels as it contains energy rich cellulose and hemicellulose sugars, which can be converted into ethanol under fermentation (Chapple et al, 2007). The secondary cell wall of plants is primarily made up of lignin, cellulose and hemicellulose. Lignin binds into cellulose and hemicellulose via covalent bonds that then forms lignocellulosic biomass (Yang et al, 2007). The conversion of lignocellulosic biomass into biofuels, pulp, and paper involves chemical pretreatment followed by enzymatic hydrolysis. However, this pretreatment process can be expensive and environmentally hazardous (Mosier et al, 2005).

An alternative approach to pretreatment that would result in low pre-processing cost is to genetically modify the plant cell wall resulting in less lignin formation (Lee, 2010). Although the lignin biosynthetic pathway is well studied, details about the regulation and underlying topology is still unclear (Lee et al, 2012, Bhatia and Bosch, 2014). In order to enhance the understanding of the regulatory control involved in monolignol biosynthesis, the objective of this dissertation is to use mathematical modeling and simulation for systematic analysis of monolignol biosynthesis in wildtype and mutant plants. Mathematical models for biosynthetic pathways will be developed, validated, and analyzed, yielding insights about lignin regulation that are difficult to

obtain with traditional molecular and biochemical experimental approaches alone. The insights obtained from the model can then be used to generate testable hypotheses and suggest targets to engineer the lignin composition and structure.

1.2 Lignin Biosynthesis:

Lignin is found in the secondary cell wall of all vascular plants and is a polymerized product of three monolignols namely p-coumaryl alcohol, coniferyl alcohol and sinapyl alcohol. The monomers are synthesized in the monolignol biosynthetic pathway inside the cytoplasm and then transported to the cytoplasm, where it undergoes polymerization to form syringyl (S), guaiacyl (G) and p-hydroxyphenyl (H) units. The lignin composition and content varies among plants depending on the environmental conditions (Boerjan et al, 2003). The lignin biosynthetic pathway involves 24 metabolites along with 21 enzymes belonging to 10 protein families that convert substrates to products involving 32 intermediate metabolic reactions.

Although the topology of the monolignol biosynthetic pathway is well studied, the information about the mechanisms of regulations is not complete. The lignin biosynthetic pathway is regulated in several different ways: (1) Substrate inhibitions that exists between metabolites, (2) Proteins are multifunctional, hence they display different substrate specificity, and (3) Proteins interact with each other forming protein complexes. Due to these complexities, experimental perturbations made to the pathway may sometime result in counter-intuitive results, which cannot be explained by knowledge of the topology alone. Although there are several reported literatures

pertaining to the study of monolignol biosynthesis (Dixon et al, 2001; Humpreys and Chapple, 2002; Boerjan et al, 2003), only one study reported literature has focused on developing a mathematical model of the monolignol biosynthesis pathway (Lee et al, 2010).However, the major drawback of Lee et al.'s model is that it uses a static and constraint based approach to gain an understanding of the pathway. The drawbacks of such methods will be explored in the literature review. The overall goal of this dissertation is to develop a computational framework that would enable researchers to gain insights that are otherwise difficult to obtain with traditional molecular and biochemical experimental approaches.

1.3 Significance of this research:

Given the complexity of biological systems, mathematical modeling has been extensively applied to biological pathway to gain a better understanding of the system. Computational modeling might be advantageous to explain the sometimes counterintuitive results obtained from genetic perturbation of the pathway. The computational model can be used to explain the changes in the steady state flux distribution when the pathway is subjected to enzymatic perturbations and the role of protein interactions on the monolignol biosynthesis pathway. The model can also be used to gain insights about the regulation of monolignols and formulate testable hypotheses. Users of the model can perform in silico transgenics saving thousands of dollars in experimental test costs.

1.4 Literature Review

In this section, we will review the methods used in developing a computational framework for modeling monolignol biosynthesis. A biosynthetic pathway can be broadly divided into a regulatory network and a metabolic network. The regulatory network primarily involves interactions between genes, mRNA and proteins, whereas the metabolic network is made up of interactions between proteins and metabolites. The next few sections will involve a brief review of the methods used in developing these networks.

1.4.1 Mathematical Modeling of Metabolic Network:

With the emergence of high throughput technologies, the current challenge in studying biological systems insilco is to develop a computational framework to gain a systematic understanding of the underlying mechanisms regulating the pathway. The current modeling approaches for metabolic networks are classified into static (Schuster et al, 199; Schilling et al, 2000; Palsoon, 2006) and dynamic based models (Reich and Selkov, 1981; Palsoon and Lightfoot, 1984). Static based model uses only the information about the stoichiometry of the network. Using mass conservation principles, the rate of change of the metabolite concentration as a function of stoichiometry and reaction rate can be expressed as follows

$$\frac{dS}{dt} = Nv \tag{1.1}$$

where S is the vector of substrate concentrations of all the metabolites involved in the pathway, N is the stoichiometric matrix and v is the reaction flux. Each row of the

stoichiometric matrix indicates all the reactions in which a given substrate is involved. The reaction flux, v, depends on the metabolite concentrations and other non-metabolic factors like pH and temperature. The constraint based methods and the kinetic based methods are distinguished based on the functional form of the metabolic flux, v.

Constraint based methods are based on the stoichiometry structure of metabolic networks. The constraint based model assumes a quasi-steady state, which means that the systems reaches equilibrium relatively fast, thus reducing the left hand side of the above equation to zero. The system of equations can then be solved to obtain the steady state flux. However, the main challenge faced in the above approach is that the system is overdetermined, i.e the number of reactions are more than the number of metabolites. Methods such as extreme pathways (Schilling et al, 2000) and elementary flux modes (Shuster et al, 1999) have been developed to identify a finite set of fluxes, which is a basis vector of all possible steady state flux in the pathway.

One of the most widely used constraint base approaches is the Flux Balance Analysis (FBA) (Palsson et al, 2001). FBA has been used in modeling the metabolic network of a variety of biological systems including the metabolic network of monolignol biosynthesis in alpha-alpha (Raman and Chandra, 2000, Lee et al, 2010). The major drawback of this approach is that there is no best solution, rather there is an optimal solution based on imposed constraints. The other drawback is that the definition of the objective function is not always clear. Depending on the external environment, the cells may maximize biomass, maintain homeostasis, maximize metabolic product, or maximize growth rate (Almaas et al., 2004). FBA does not include the enzyme

efficiency, which may be different for different substrates based on the concentration of the substrates as well as the substrate inhibition that exists in the metabolic network. The dynamics of the network, which is crucial in understanding the effect of perturbations, cannot be known. Nonetheless, in the absence of available experimental data, FBA has proven to be a powerful approach to model the metabolic networks of the biological systems (Papin et al., 2002; Almaas et al., 2004; Holzhutter 2004).

1.4.2 Kinetic Based Models:

The kinetic based modeling approach are by far the most widely used modeling tool to simulate the metabolic networks. In the kinetics based approach, the metabolic flux is expressed as a function of substrate concentrations and kinetic parameters. The next step is to utilize the mass conservation principle to quantify the rate of change of metabolite concentration as a function of the net metabolic flux. The reaction rate can then be expressed using either Michaelis Menten kinetics, Generalized Mass Action (GMA) (Horn and Jackson, 1972) or S-systems (Savageau, 1976). Although the Michaelis Menten kinetics provides an accurate description of the changes in reaction flux involved in the metabolic networks, specification of Michaelis Menten kinetics require detailed knowledge of the activation and inhibition parameters, which may be difficult to obtain. Michaelis Menten kinetics have been used to model a variety of biological systems ranging from glycolytic pathway (Johnson, 2013) to monolignol biosynthesis (Wang et al, 2014).

1.4.3 Gene Regulatory Networks

Gene regulatory networks involves the interactions between genes, mRNA and proteins that interact with each other in the cell to regulate the expression levels of mRNA and proteins. The gene regulatory network provides crucial insights about the various regulatory interactions that are essential for day to day cellular function, growth and stress response (MacNeil and Walhout, 2011). The knowledge about how the genes interact with each other is extremely important to quantify the effect of perturbations on the phenotype (Davidson and Levine, 2005; Walhout, 2006 and Long et al, 2008). With advances in high throughput technologies, large number of databases have been developed containing information about genes, protein-protein interactions and transcription factor data. However, one of the biggest challenges is to use the vast amount of data to infer interactions between the transcription factors, mRNA and proteins (Friedman, 2004).

1.4.3.1 Modeling Gene Regulatory Networks:

Several methods have been reported to build a gene regulatory network using high throughput experimental data. The most commonly used methods are Bayesian networks and other network inference methods (i.e., Mutual Information, PC algorithm). The Bayesian network is a graphical model, which uses the conditional probability distribution between random variables to infer relationship (Pearl, 2000). The random variables can be genes, mRNA or proteins (Friedman et al, 1998). The advantage of the Bayesian network is that they can capture linear or non-linear, relationships among participating variables (Chen et al., 2008). In the Bayesian network approach for

reconstructing a GRN from expression data, the best network is determined by a scoring function, which selects the best network based on the criterion of maximizing likelihood given the experimental data. However, finding an optimal network is computationally intensive since each additional node increases the number of parameters by 2^N, where N is the number of nodes. The objective function can only be evaluated using heuristic approach, which may not result in a global maximum. An additional concern with the Bayesian approach is that well-connected networks tend to result in over fitting due to the increased number of parameters. Using Bayesian networks, Burell et al. (2008) were able to identify the genes responsible for transducing cold stress genes to other genes. Other applications of Bayesian networks in gene regulatory networks are discussed in detail elsewhere (Beal et al., 2005). The drawback of Bayesian networks that was identified in the literature was that the Bayesian networks couldn't model feedback interactions, which are a major part of these regulatory networks.

Mutual information based network inference relies on the knowledge of joint probability distribution between random variables to infer their interaction (Meyer et al, 2010). The mutual information measures the uncertainty reduction between a pair of random variables, given information about one of the variables (Altay and Emmert, 2010). The advantage of mutual information is that it can account for both non-linear and linear relationship between the random variables. They have distinct advantage over correlation based networks since they are free of any assumptions of normality

distributions (Wang et al, 2013). A detailed review of mutual information methods used for network inference can be found in Marbach et al (2012).

1.4.3.2 Peter Clark (PC) Algorithm

The PC algorithm uses partial correlation or partial mutual information to infer causal relationships between random variables (Spirtes and Glymour, 1991). It assumes a Bayesian causal network model and it makes use of a set of constraints to produce a Directed Acyclic Graph (DAG) as an output. The PC algorithm comprises of three steps. In the first step, it applies the conditional independence test to discover relationships between variables. In subsequent steps, it tries to orientate these relationships without creating cyclic structures. More information on the PC algorithm can be found in Harris and Drton (2013). The gene network inference was performed using the package *pacalg* in R©. The advantage of the PC algorithm over the Bayesian network is that, PC algorithm used a constraint based approach rather than a scoring approach to infer the final DAG, which makes it less computationally intensive (Abellan et al, 2006). The drawback is that in some cases it may not arrive at a fully developed DAG; rather the final network may have some undirected DAG`s (Raskutti, 2013).

1.4.3.3 A joint regulatory and metabolic network model:

Although the methods to model the regulatory and metabolic network has been developed, little effort has been devoted towards bridging the two networks. By bridging the two networks, it would enable researchers to quantify the effect of regulatory based perturbations on the monolignol biosynthetic pathway as well as lignin composition and

structure. In one of the recently proposed methodology, Covert et al (2007) developed a framework to combine the regulatory information into the metabolic network with a small ODE model for E. coli model. One of the drawbacks of the above method is that it simplifies the relationship between the transcriptional regulation and metabolic process to a binary process, where the regulation can either be present or absent. The model fails to capture the intermediate levels that are known to exist in biological networks (Chandrashekaran, 2010), as well as quantify the effect of the perturbation of genes in the regulatory network on the phenotypes. To overcome such drawbacks, Wittman et al (2010) developed a model that converts a discrete Boolean interaction into a continuous process using the Hill equation.

Based on the literature review, it can be concluded that although several computational approaches have been developed towards modeling biological networks, most of the focus has been in studying the metabolic and regulatory networks separately. The cellular phenotype is an emergent property of a system, which is a result of interactions between the various components involved in a biological network (genes, mRNA, proteins and metabolite) (Hellerstien, 2003). The overall process of biosynthesis cannot be determined from information generated at a single level; rather it requires information at multiple levels, including gene expression levels, absolute protein concentrations, and metabolite concentrations.

In this dissertation, a Michaelis Menten kinetics approach was used to model the metabolic network and the PC algorithm was used to infer the regulatory interactions

between the genes, mRNA and proteins. The two networks will then be linked using the continuous Hill cube transformation.

1.4.4 Dissertation Overview and Objectives

The main objective of this dissertation is to use computational and statistical approaches to develop a predictive model for the systematic analysis of monolignol biosynthesis in *Populus trichocharpa* (*P. trichocharpa*) wild-type and genetically engineered plants. The computational framework will help to improve our understanding of the regulation of monolignol biosynthesis. The results generated by this model would demonstrate that insilco investigation is an effective and predictive complement to traditional biotechnological and transgenic experimental methods in plants. Mathematical models will be constructed and analyzed, yielding insights about regulatory control that are difficult to obtain with traditional experimental approaches alone and allowing the formulation of more effective hypotheses with regards to the regulation of the pathway. The validated model can then be used to draw insights that will form a solid foundation for the design of genetic modification strategies towards the generation of lignin-modified crops.

The dissertation research questions that the model will answer are: (1) what are the key regulators that affect the lignin composition and structure. (2) What is the role of these key regulators on the lignin composition and structure? (3) How can we devise strategies to change the lignin structure? (4) What is the role of multi-enzyme interactions on the network? (5) How does the steady state concentration of

monolignols change when the pathway is perturbed? The hypotheses that will be tested in this research are as follows (1) the biochemical network is resilient and hence robust to internal and external perturbations. (2) The genes that control the proteins involved in monolignol biosynthesis form modules that provide resilience against perturbations. The specific objectives of this dissertation are:

(1) To identify and quantify the role of regulatory controls on the steady state flux distribution of the monolignol biosynthetic pathway.(2) Identify the effect of perturbations on the lignin composition and structure.(3) Identify the role of co-regulatory protein structure and its overall effect on the lignin composition and structure under enzymatic perturbations.(4) Validate the predictive power of the model by comparing the results with the experimentally determined phenotypic data.(5) Develop an artificial neural network (ANN) model to predict the S/G ratio by using transgenic data.(6) Develop a joint regulatory metabolic network model to quantify the role of perturbing the Transcription factors on the lignin composition and structure.

The overall contributions of this dissertation are as follows: (1) a computational framework is proposed to develop a joint regulatory and metabolic model for lignin biosynthesis. The model would enable researchers to gain a better understanding of the monolignol biosynthesis process as well as reveal new regulatory mechanisms that affect the lignin composition and structure. (2) The knowledge about the role of complex and protein modules would enhance the overall understanding of the regulation of the monolignol biosynthetic pathway. (3) The knowledge about the steady state behavior of the system also enables us to understand the effect of perturbations on the network and

how the system switches from one steady state to another in the presence of these perturbations. Phenotypes resulting from the perturbations can be predicted by knowing the steady states. (4) Due to the inherent complexities of biochemical pathways, the predictive model would allow researchers to analyze lignin biosynthesis in wild-type and genetically engineered plants. (5) The model would yield insights that are difficult to obtain with traditional experimental molecular and biochemical approaches and would allow researchers to formulate new, testable hypotheses about the regulatory mechanisms involved in the pathway. Although the model was developed for monolignol biosynthesis, the framework can be applied to any biochemical network of interest

References

AJ, Kohane IS: Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements.

Akasu T, Munakata Y, Tsurusaki M & Hasuo H (1999). Role of GABAA and GABAC receptors in the biphasic GABA responses in neurons of the rat major pelvic ganglia. J Neurophysiol 82, 1489– 1496.

Alvira P, Tomás-Pejó E, Ballesteros M, Negro MJ. Pretreatment technologies for an efficient bioethanol production process based on enzymatic hydrolysis: A review. Bioresource Technol. 2010 Jul;101(13):4851-61.

Amos Tanay and Ron Shamir. “Computational Expansion of Genetic Networks”. Bioinformatics 17 Supp 1 S270-S278 (2001) (Proceedings of ISMB 2001).

C.H. Yeang, M. Vingron. A joint model of regulatory and metabolic networks. BMC Bioinformatics 7:332 2006.

Chen, F. and He, Y. (2009) Caspase-2 mediated apoptotic and necrotic murine macrophage cell death induced by rough Brucella abortus, PLoS One, 4, e6830.

Chen, T, He, H.L., and Church, G.M. 1999. Modeling gene expression with differential equations.

Chiang, VL, Tsai CJ & Hu, WJ. (2002). Methods of modifying lignin in plants by transformation with a 4-coumarate coenzyme a ligase nucleic acid. U.S. Patent No. 6455762. Washington, DC: U.S. Patent and Trademark Office.

Covert MW, Schilling CH, Palsson BØ. 2001. Regulation of gene expression in flux balance models of metabolism. Journal of theoretical biology. 213(1):73-88.

D. Heckerman and C. Meek. Models and Selection Criteria for Regression and Classification. In Proceedings of Thirteenth Conference on Uncertainty in Artificial Intelligence, Providence, RI, pages 223-228. Morgan Kaufmann, August 1997. Also appears as Technical Report MSR-TR-97-08, Microsoft Research, May, 1997.

Dojer, N., et al. (2006) Applying dynamic Bayesian networks to perturbed gene expression data, BMC Bioinformatics, 7, 249.

EO Voit, MA Savageau. Accuracy of alternative representations for integrated biochemical systems. Biochemistry 26 (21), 6869-6880.

Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS: Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles.

Florbela Carvalheiro, Luís C. Duarte and Francisco M Gírio. Hemicellulose biorefineries: a review on biomass pretreatments. Journal of Scientific & Industrial Research, Vol. 67, November 2008, pp.849-864.

Gerosa L, Sauer U (2011) Regulation and control of metabolic fluxes in microbes. Curr Opin Biotechnol 22: 566–575.

Hendricks, A.T.W.M., Zeeman, G., 2008. Pretreatments to enhance the digestibility of lignocellulosic biomass. Bioresource Technology. 100, 10-18.

Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P: Inferring regulatory networks from expression data using tree-based methods.

Idekar, T., Galitski, T., Hood, L., 2001. A new approach to decoding life: systems biology. Ann. Rev. Genom. Hum. Genet. 2,343–372.

Josef T. Kittler and Peter L. Oliver. Chapter 15. Genomic and Post-Genomic Tools for Studying Synapse Biology.

Kirimasthong K, Manorat A: Inference of gene regulatory network by Bayesian network using Metropolis-Hastings Algorithm. In Proceedings of 3rd International Conference on Advanced Data Mining and Alications: 06-08 Aug, 2007; Harbin Edited by Alhajj R, Gao H. 2007, 276-286.

Langfelder P, Horvath S: WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008, 9:559

Maki, Y., Tominaga, D., Okamoto, M., Watanabe, S., Eguchi, Y.,2001. Development of a system for the inference of large scale genetic networks. Pac. Symp. Biocomput. 446–458.

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A: ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.

Meyer PE, Kontos K, Lafitte F, Bontempi G: Information-theoretic inference of large transcriptional regulatory networks.

Meyer PE, Lafitte F, Bontempi G: minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information.

Mordelet F, Vert JP: SIRENE: supervised inference of regulatory networks. Bioinformatics 2008, 24:i76-82.

Ong, I.M., Glasner, J.D. and Page, D. (2002) Modelling regulatory pathways in E. coli from time series expression profiles, Bioinformatics, 18 Suppl 1, S241-248.

Pearl J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo, CA, USA: Morgan Kaufmann Publishers.

Pe'er, D., et al. (2001) Inferring subnetworks from perturbed expression profiles, Bioinformatics, 17 Suppl 1, S215-224.

Proc. Pac. Symp. Biocomuting (PSB'99), vol. 4, 29–40, Singapore, World Scientific Publishing.

Ragauskas AJ, et al., "The path forward for biofuels and biomaterials," Science 311(5760): 484-9, 27 January 2006.

Reverter A, Chan EK: Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks.

Rubén Armañanzas, Iñaki Inza, Pedro Larrañaga. Detecting reliable gene interactions by a hierarchy of Bayesian network classifiers. Computer Methods and Programs in Biomedicine - August 2008 (Vol. 91, Issue 2, Pages 110-121, DOI: 10.1016/j.cmpb.2008.02.010)

Taherzadeh, M. J. and K. Karimi (2008). "Pretreatment of lignocellulosic wastes to improve ethanol and biogas production: A review." International Journal of Molecular Sciences 9(9): 1621-1651.

Wittmann DM, Krumsiek J, Saez-Rodriguez J, Lauffenburger DA, Klamt S, Theis FJ. Transforming Boolean Models to Continuous Models: Methodology and Application to T-Cell Receptor Signaling,2009 Sep 28;3:98.

CHAPTER II

METHODS

This section discusses the numerical methods that were used to perform this research study. Modeling approaches include: 1) a methodology to build a gene regulatory network to uncover relationships between the transcription factors and the transcript abundance and 2) a model to predict the reaction flux through the metabolic network to develop a predictive model for monolignol synthesis. Combining these two networks would enable researchers to quantify the changes in transcription factors on the lignin composition and structure. The methodology used to infer gene regulatory network will be discussed in section 2.1, the regulatory interaction will be quantified in section 2.2 and the methodology to quantify the reaction flux through the metabolic network will be discussed in 2.3.

2.1 Inferring gene regulatory network:

The causal interactions between TF`s and genes involved in the monolignol biosynthetic pathway can be inferred using graphical Gaussian method (GGM). The associative interactions between the proteins involved in the monolignol biosynthesis were inferred using information theory. The causal relationship between the genes and proteins were inferred using Graphical Gaussian Network (GGN). A GGN involves the development of a directed acyclic graph (DAG) by utilizing a heuristic search method. The objective function used for the heuristic search is to develop a DAG that maximizes

the likelihood function. Peter Clarke (PC) algorithm is a type of GGM that uses conditional independence for selection of DAG structure (Kalish, 2008; Harris and Drton, 2013). The conditional independence between the random variables are assessed using mutual information theory. Once the relationship between the TF`s and genes are inferred, the next step is to quantify the relationship between the TFS and genes and the genes and proteins. The regulatory relationship can be quantified by converting each interaction within the gene regulatory network into a discrete Boolean function. The discrete Boolean function can then be converted into a continuous function using a hillcube approximation (Wittman et al., 2008).

2.1.1 Mutual information

In the presence of non-monotonic dependencies between genes, the correlation based methods fail to capture the relationship between genes. These drawbacks can be overcome by using mutual information to measure the degree of association between the random variables. The advantage of mutual information based methods is that it can be used for cases when the relationship between genes is non-monotonic (Butte and Kohane, 2000). The mutual information between the discrete variables *A* and *B* is defined in equation 2.1

$$I(A,B) = \sum_{a_i \in A_i} \sum_{b_j \in Bj} p(\mathrm{a}_i, \mathrm{b}_j) \log \left(\frac{p(\mathrm{a}_i, \mathrm{b}_j)}{p(\mathrm{a}_i)\, p(\mathrm{b}_j)} \right) \qquad (2.1)$$

where p(a_i, b_j) is the joint probability distribution of A_i and B_j and $p(a_i)$ and $p(b_j)$ are the marginal probabilities. A_i and B_j are required to be discrete variables, which in the case of data, require a discretization step prior to the computation of mutual information. Mutual information was used to quantify non-linear relationships between variables but does not provide information about causality.

2.1.2 CLR (context likelihood of relatedness)

CLR (Faith et al., 2007) extends the relevance network method (RN) by taking the background distribution of the mutual information values $I(A_i, B_j)$ into account. The interactions that deviate most from the background distribution are defined as the most probable interaction. A maximum z-score, z_i, is calculated for each gene *i* as shown in Equation 2.2.

$$z_i = \max_j \left(0, \frac{I(A_i, B_j) - \mu_i}{\sigma_i} \right) \tag{2.2}$$

where μ_i and σ_i and are the mean value and standard deviation of the mutual information values $I(A_i, B_k)$, $k=1,...,n$, respectively. The interaction score z_{ij} between two genes *i* and *j* is then defined as shown in Equation 2.3.

$$z_{ij} = \sqrt{z_i^2 + z_j^2} \tag{2.3}$$

The main goal of the background correction step described is to reduce the prediction of false positives based on false correlations and indirect interactions that might be observed in the case of relevance networks.

2.1.3 Discrete to Continuous Boolean Transformation

As described previously, the Boolean approach has been used significantly in previous years to build a gene regulatory network, based on the interactions between the various components within the network. Assuming that we already have a fully connected network (gene or metabolic) that has been justified by previous inference methods, the first step is to define a discrete Boolean function for each interaction and then convert it into a continuous function using a hill equation approximation for each component in the network. The methodology is shown in a simple network below showing interaction of 3 genes A, B and C (Figure 2.2). As shown in Figure 2.1, gene C is activated by A and inhibited by B. The Boolean function for C in terms of A and B can be written as shown in Equation 2.4.

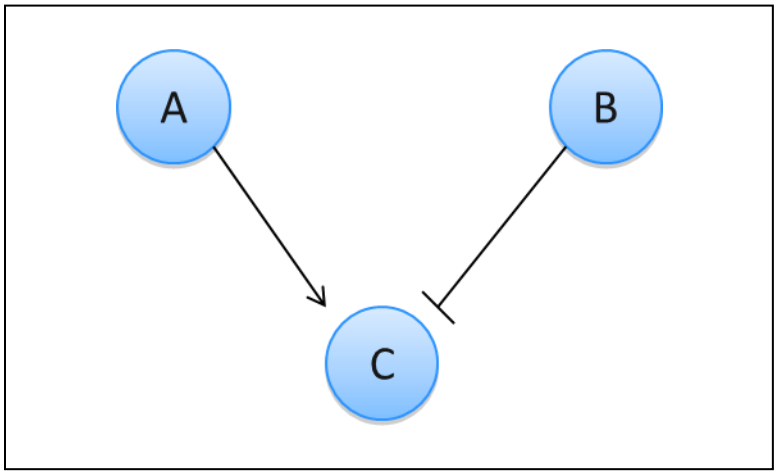

Figure 2.1: Depiction of a toy network that includes three genes A, B and C, with A influencing C and B down regulating C.

Discrete Boolean Logic $C = f(A \cap !B)$ (2.4)

Although gene C can be formed in the presence of A alone or B alone, the resulting equation would require a large number of parameters that needs to be estimated using experimental data. Since the amount of available experimental data is usually limited, we assumed that the simplest representation of the interaction graph can be represented using an AND function.

The next step is to convert the discrete Boolean function into a continuous function using a transformation proposed by Wilttman et al (2010). The transformation uses Hill equation to quantify the changes in levels of each of the component (TF, gene, protein) involved in the interactions. The continuous function showing the concentration of C as a function of genes A and B is shown in the equation 2.5

Continuous Function

$$\text{Activation Function: } \left[\frac{A^{n_A}}{A^{n_A}+k_A^{n_A}}\right]$$

$$\text{Inhibition Function: } \left[\frac{k_B^{n_B}}{B^{n_B}+k_B^{n_B}}\right] \qquad (2.5)$$

$$C = C_{\max}\left[\frac{A^{n_A}}{A^{n_A}+k_A^{n_A}}\right]\left[\frac{k_B^{n_B}}{B^{n_B}+k_B^{n_B}}\right]$$

2.3 Michaelis-Menten kinetics:

When the chemical reaction is catalyzed by an enzyme, the enzyme is not consumed or produced by this reaction, but it may form a temporary complex with the substrate in the reaction. In such cases, we can use Michaelis-Menten kinetics to describe its reaction rate under the key assumption of quasi steady state, which is valid when the enzyme concentration is much lower than the substrate concentration and when the enzyme is not allosteric. Michaelis-Menten kinetics with no inhibition has the following formulation:

$$V = \frac{V_{\max}[S]}{k_m + [S]} \tag{2.6}$$

where K_m is the Michaelis constant and is equal to the substance concentration that causes the half-maximal reaction rate V_{max}.

The Ordinary Differential Equation (ODE) modeling approach represents the concentration change of a substrate/product over time by an ordinary differential equation. For a certain substrate concentration $[S_i]$ varying as a function of time, the rate of change in substrate concentration is calculated using conservation of mass principle. The conservation of mass suggests that the rate of change in concentration $[S_i]$ is the difference between the rate of formation of S and the rate of consumption of S. The resulting rate of change in substrate concentration is shown in equation 2.7.

$$\frac{dS_i}{dt} = \sum V_{production} - \sum V_{consumption} \tag{2.7}$$

2.4 Model simulation:

Once the mass balance equations for all the metabolites involved in the pathway are obtained, the next step is to solve the series of ordinary differential equations (ODE) from an initial concentration level to a steady state concentration. The ODE's were solved using the Matlab® function ode15s. The rate of change of each metabolite is expressed in terms of Michaelis Menten kinetics. The initial concentrations for each of the metabolites are assumed to be a very low concentrations and the parameters of the Michaelis Menten kinetics were obtained experimentally. The steady state concentrations can then plugged back into the original metabolic flux equation to obtain the steady state flux of all the reactions involved in the pathway.

2.5 Sensitivity Analysis:

Global sensitivity analysis is performed to estimate the changes in the output as a result of changes made to the model input. Sensitivity analysis is usually performed by evaluating the output of the model by varying the parameters of the model. The parameters are assumed to follow a particular probability distribution function. The most commonly used distributions are normal, uniform, and triangular (Staltelli et al, 2000). A Monte Carlo sampling method is employed to ensure the parameter values are randomly sampled (Mckay et al, 1979).

To calculate the sensitivity between the parameters and the output of interest, the most commonly used approach is to calculate the partial rank correlation coefficient

(PRCC) (Blower and Dowlatabadi, 1994; Saltelli, 2004). PRCC provides a measure of the linear relationship between the parameters and model output and varies from -1 to 1. For the PRCC method, the output corresponding to the different values of the parameters are tabulated. The partial correlation between x and y is defined as the correlation coefficient between x and y, as shown in Equations 2.7 and 2.8.

$$\hat{x}_j = c_o + \sum_{p \neq j} c_p x_p \tag{2.7}$$

$$\hat{y} = b_o + \sum_{p \neq j} b_p x_p \tag{2.8}$$

Once $\hat{x}_j$ and $\hat{y}$ are obtained, the residuals $x - \hat{x}_j$ and $y - \hat{y}$ were obtained and the partial rank correlation coefficient is calculated using Equation 2.9, where $\bar{x}$ and $\bar{y}$ are the respective sample means.

$$prcc(x_j, y) = \frac{\mathrm{cov}(x_j, y)}{\sqrt{\mathrm{var}(x_j)\,\mathrm{var}(y)}} \qquad j = 1, 2, \ldots k \tag{2.9}$$

2.6 Parameter Estimation

Parameter estimation is an inverse problem where the goal is to estimate the parameters of the model using experimental data given the model structure. The parameter estimation is an optimization procedure with a clearly defined objective function and constraints on the parameters. The objective function that was used in the

parameter estimation was the minimization of mean squared error between the experimental data and the results predicted by the model. The optimization problem is described in Equation 2.10, where *p* is the parameters to be estimated and *n* is the number of data points. A particle swarm optimization (PSO) algorithm (Kennedy and Eberhart, 1997) was used in this research. The advantage of PSO is that it is a heuristic based algorithm, hence the direction of the search does not involve the computation of the gradients, which is computationally intensive for non-linear objective functions.

$$\min_{p \in R} \left(\frac{(y_{\text{exp}} - y_{\text{model}}(p))^2}{n} \right) \tag{2.10}$$

2.7 System Dynamics

As mentioned earlier, biological networks are composed of nonlinear interactions between the various TF's, genes, proteins and metabolites. The interactions between these components determines the resilience/robustness of the network to perturbations. The system dynamics is the study of the effect of perturbations on the stability of the system providing information about the long-term behavior of the system. Among the most well-known instances of complex dynamics are temporal variations in the concentrations of metabolic intermediates of the yeast glycolytic pathway (Selkov, 1968; Dano et al., 1999) and the photosynthetic Calvin cycle (Laisk and Walker, 1986; Giersch, 1986).

The rate of enzymatic reaction can be represented as a function the substrate concentration and kinetic parameters, shown in Equation 2.11.

$$\frac{dS}{dt} = Nv(S) \tag{2.11}$$

N is the stoichiometric matrix and *v* is the reaction flux. For small perturbations around the steady state, the behavior of a nonlinear system can be approximated by a linear system. Using the Taylor series expansion for a differential equation and ignoring higher order terms, we obtain Equation 2.12.

$$\frac{dS}{dt} = Nv(S^0) + N\left.\frac{\partial v}{\partial S}\right|_{S^0}(S - S^0) \tag{2.12}$$

At steady state $\mathrm{Nv(S^0)}$=0

For a small perturbation, ΔS = S – S_0 in the vicinity of S_0 can be described by a linear differential equation shown in Equation 2.13.

$$\frac{d}{dt}\Delta S = M\Delta S \tag{2.13}$$

where M is the Jacobian matrix

The solution to the above equation can be specified by the Eigenvectors and Eigenvalues of the Jacobian matrix. The stability of the system can be determined by the trace and determinant of the Jacobian matrix as shown in Figure 2.2.

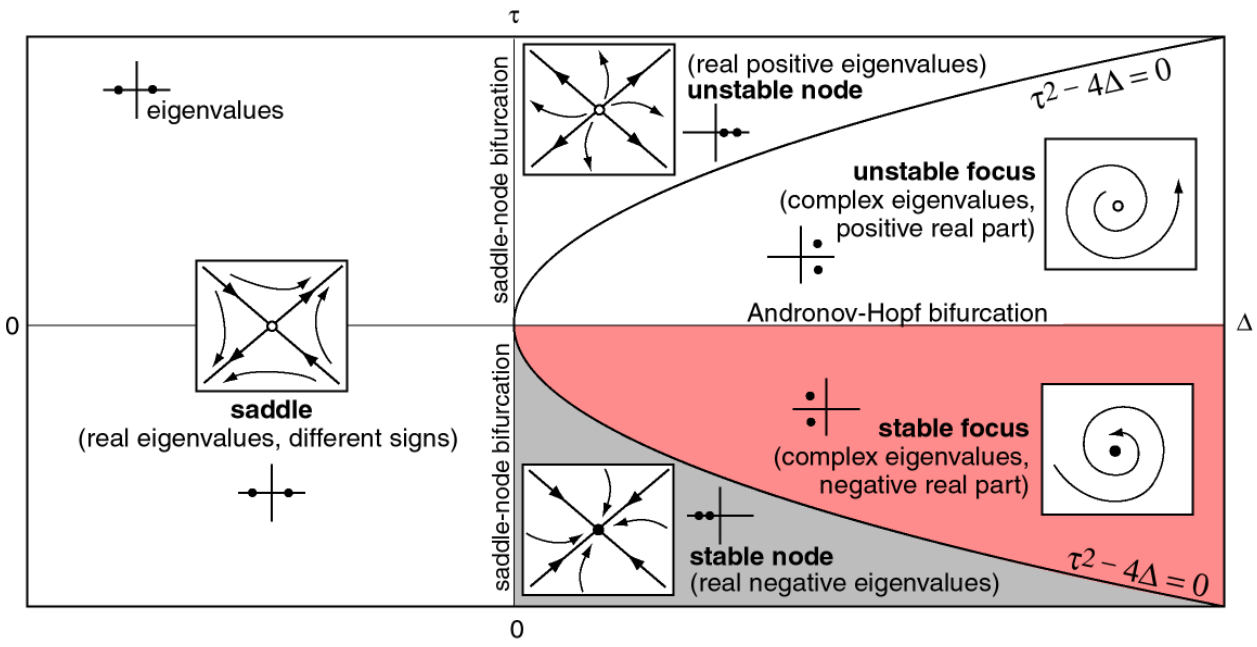

Figure 2.2: Figure showing the regions of qualitative behavior of dynamic systems based on the range of trace and determinant (Keith, 2012). The x axis represents the determinant and y axis represents the trace.

When Δ ≥ 0 and Eigenvalues both positive, this results in an unstable node, which means that all the all trajectories in the neighborhood of the fixed point will be in the outward direction. When Eigenvalues both negative, its results in a stable node, where all the fields in the neighborhood of the steady state will be directed towards the fixed point. When Eigenvalues have opposite signs, it results in a saddle node, where the fields in the direction of the negative eigenvector will converge toward the fixed point but will be outwards as they approach a region dominated by the positive eigenvalue. A stable node means that when a system is at equilibrium, then small changes made to the system would not result in any changes in the steady state behavior of the system. On the other hand, an unstable node means that a small change might result in a system shifting to another stead state, which may be stable or unstable.

2.8 Artificial Neural Networks (ANN):

ANN models have been predominantly used in quantifying nonlinear relationships between the dependent and independent variables. One of the advantages of ANN is that the knowledge about the functional form of the relationship between the dependent and independent variables is not required (Morgan and Scofield, 2012). The simplest ANN network is typically composed of 3 layers namely the input, hidden and the output layer. The most commonly used architecture for ANN is the multilayered neural network with backpropagation. The backpropagation training is composed of three steps: (i) Input is passed into the input layer and passed through the hidden layer and realized through the output layer (ii) the difference in values between the output and the target is back-propagated through the network and finally (iii) the weights are adjusted using optimization such that the error is minimized (Rumelhart et al, 1986). The number of neurons in the hidden layers are identified such that the training and testing errors are minimized. A neural network with more than optimum nodes results in overfitting and hence lacks generalization of patterns observed in the data (Hussain et al., 1992).

The data to the input layer of the ANN is normalized and then passed from input layer through the hidden layer and finally to the output layer of the network (Hussain et al., 2002). The data from the input layer is assigned a random weight and then combined with the other inputs in the hidden layer. The output from the hidden layer undergoes a similar linear weighted transformation process. Each neuron can be viewed as a transfer function that converts an input into a sigmoidal output. The output

from a neuron is again transformed using an appropriate transfer function (Razavi et al., 2003).

References:

Agunda BD and Clarke BL. Bistability in chemical reaction networks: Theory and application to the peroxidase reaction. J. Chem. Phys., 87(6):3461–3469, 1987.

Bailey JE. Lessons from metabolic engineering for functional genomics and drug discovery. Nature Biotechnology, 17(616–618), 1999.

Camacho D, A. de la Fuente, and P.Mendes. The origin of correlations in metabolomics data. Metabolomics, 1:53–63, 2005.

Carrier TA and Keasling JD. 1999. "Investigating autocatalytic gene expression systems through mechanistic modeling." J. Theor. Biol. 201:25-36.

Chassagnole C, NoisommitRizzi N, Schmid JW, Mauch K, and Reuss M. Dynamic modeling of the central carbon metabolism of Escherichia coli. Biotechnology and Bioengineering, 79(1):53 – 73, 2002.

De Hoon et al. (2003). Inferring gene regulatory networks from time ordered gene expression data using differential equations Pac Symp Bioc , 267274.

Edwards JS and Palsson BO. The Escherichia coli mg1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. Proc. Nat. Acad. Sci., 97(10):5528–5533, 2000.

Endy D, and Brent R. 2001. Modelling cellular behavior. Nature 409, 391–395.

Erdi P and Toth J. Mathematical models of chemical reactions: theory and applications of deterministic and stochastic models. Manchester University Press ND,1989.

Fiehn O, Kopka J, D'ormann P, Altmann T, Trethewey RT, and Willmitzer L. Metabolite profiling for plant functional genomics. Nat. Biotechnol., 18:1157–1161, 2000.

Goldbeter, A. 1995. A model for circadian oscillations in the Drosophila Period protein (PER). Proc. R. Soc. Lond. B261, 319–324.

Heinrich R and Rapoport TA. A linear steady state treatment of enzymatic chains. General properties, control and effector strength. Eur. J. Biochem., 42:89–95, 1974

Hellerstein MK, Hoh RA, Hanley MB, Cesar D, Lee D, Neese RA, McCune JM. Subpopulations of long-lived and short-lived T cells in advanced HIV-1 infection. J Clin Invest 112(6):956-66, 2003.

Junker JH, Lonien J, Heady LE, Rogers A, and Schwender J. Parallel determination of enzyme activities and in vivo fluxes in brassica napus embryos grown on organic or inorganic nitrogen source. Phytochemistry, 68:2232–2242, 2007.

Koh, BT, Tan, RBH and Yap, MGS. 1998. Genetically structured mathematical modeling of trp attenuator mech- anism. Biotechnol. Bioeng. 58, 502–509.

Leloup, JC and Goldbeter, A. 1998. A model for the circadian rhythms in Drosophila incorporating the formation of a complex between the PER and TIM proteins. J. Biol. Rhythms 13(1), 70–87.

Mahaffy JM and Pao CV (1984), Models of genetic control by repression with time delays and spatial effects, J. Math. Biol. 20, 39-58.

Mayo AE, Setty Y, Shavit S, Zaslaver A, Alon U (2006) Plasticity of the cis-Regulatory Input Function of a Gene . PLoS Biol 4(4): e45. doi:10.1371/journal.pbio.0040045

Prokudina EI, Valeev RY and Tchuraev RN. 1991. A new method for the analysis of the dynamics of the molecular genetic control systems. II. Application of the method of generalized threshold models in the investigation of concrete genetic systems. J. Theor. Biol. 151, 89–110.

R. Alves, F. Antunes, and A. Salvador. Tools for kinetic modeling of biochemical networks. Nature Biotechnology, 24(6):667–672, 2006.

Ruoff, P., Vinsjevik, M., Monnerjahn, C., and Rensing, L. 2001. The Goodwin model: Simulating the effect of light pulses on the circadian sporulation rhythm of Neurospora crassa. J. Theor. Biol. 209, 29–42.

Sanglier M., and Nicolis, G., 1976. Sustained oscillations and threshold phemomena in an operon control circuit. Biophys. Chem. 4, 113–121.

Santillán M and Mackey MC. 2001. Dynamic regulation of the tryptophan operon: A modeling study and comparison with experimental data. Proc. Natl. Acad. Sci. USA 98(4), 1364–1369.

Savageau MA. 1989. Are there rules governing patterns of gene regulation? In B.C. Goodwin and P.T. Saunders, eds. Theoretical Biology: Epigenetic and Evolutionary Order, 42–66. Edinburgh University Press, Edinburgh.

Ueda HR., Hagiwara M and Kitano, H. 2001. Robust oscillations within the interlocked feedback model of Drosophila circadian rhythm. J. Theor. Biol. 210, 401–406.

Wong P., Gladney S., and Keasling JD. 1997. Mathematical model of the lac operon: Inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose. Biotechnol. Prog. 13, 132–143.

Xiu, ZL, Zeng AP and Deckwer WD. 1997. Model analysis concerning the effects of growth rate and intracellular tryptophan level on the stability and dynamics of tryptophan biosynthesis in bacteria. J. Biotechnol. 58, 125–140.

You, L., and Yin, J. 2001. Simulating the growth of viruses. In R.B. Altman, A.K. Dunker, L. Hunter, K. Lauderdale, and T.E. Klein, eds. Proc. Pac. Symp. Biocomput. (PSB'01), vol. 6, 532–543, Singapore, World Scientic Publishing.

CHAPTER III

ROLE OF *4CL* COMPLEX ON THE ROBUSTNESS OF MONOLIGNOL BIOSYNTHETIC PATHWAY: A SYSTEMS BIOLOGY APPROACH

Punith Naik[1], Jack Wang[2], Jie Liu, Hsi-Chuan Chen[2], Rui Shi[2], Christopher M. Shuford, Quanzi Li[2], C.Y. Lin[2], David C. Muddiman, Ronald Sederoff[2], Vincent Chiang[2], Cranos Williams[3], and Joel Ducoste[1*]

[1] Civil, Construction, and Environmental Engineering, North Carolina State University, Raleigh, North Carolina 27695

[2] Forest Biotechnology Group, Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, North Carolina 27695

[3]Electrical and Computer Engineering, North Carolina State University, Raleigh, North Carolina 27695

*Corresponding author; e-mail jducoste@ncsu.edu.

Abstract:

Lignin is a polymer present in the secondary cell walls of all vascular plants. It is a known barrier to pulping and the extraction of high energy sugars from cellulosic biomass. While enzymatic perturbations have resulted in plants reduced lignin, some of these transgenic plants resulted in lignin monomers that were different from expectations. To address these issues, we performed an in depth analysis of the lignin biosynthesis pathway by incorporating the regulatory information about the enzymes involved in the pathway. In particular, we investigated the role of a specific protein complex formation between Ptr4CL3 and Ptr4CL5 enzymes on the monolignol biosynthesis pathway. The role of this Ptr4CL3-Ptr4CL5 enzyme complex on the steady state flux distribution was quantified by performing Monte Carlo simulations to assess the variability in lignin composition and structure resulting from variations in the enzyme concentration. Simulation results suggest that the presence of the Ptr4CL3-Ptr4CL5 complex leads to variability in the overall steady state flux distribution. The effect of this complex on the robustness and the homeostatic properties of the pathway were identified by performing sensitivity and stability analyses, respectively. Results from these robustness and stability analyses suggest that the monolignol biosynthetic

pathway is resilient to mild perturbations in the presence of the Ptr4CL3-Ptr4CL5 complex. The presence of Ptr4CL3-Ptr4CL5 complex increased the stability of the pathway to 92% compared to 70% in the absence of the Ptr4CL3-Ptr4CL5 complex. The robustness in the pathway is maintained due to the presence of multiple enzyme isoforms as well as the presence of alternative pathways resulting from enzymatic perturbations.

3.1 Introduction:

Lignin is the second most abundant complex polymer that is synthesized in the secondary cell walls of all vascular plants (Sarkanen, 1971). It enables the transport of water and nutrients through the stem, upright growth, and protection against pathogens. Recalcitrance of lignin hinders the usage of plant biomass as the viable source of biofuels or in the use of plant biomass in pulp and paper. The lignin polymer is primarily composed of derivatives of the H, G, and S monolignols, which are synthesized via the phenylpropanoid pathway (Higuchi, 1997). The structure of the lignin polymer is directly linked to variations in lignin content and composition. The removal of lignin for the above-mentioned uses requires a chemical pretreatment and consequently, strong knowledge of the regulatory and metabolic framework that influences its formation (Chiang, 2002; Ragauskas et al, 2006). The lignin biosynthetic pathway consists of twenty-four metabolites, twenty-one enzymes belonging to ten gene families and thirty-five reactions that convert various reactants to products (Figure 3.1). The reactions and the enzymes catalyzing each of the reactions are outlined in supplemental Table 1. Substantial effort has been directed towards understanding the lignin biosynthesis pathway in plants to lower the amount of total lignin and/or change its relative ratio of monolignol subunits. Although all the metabolites, enzymes, and reactions involved in

the monolignol biosynthesis has been identified, critical details about the pathway regulation, however, remain un-clear (Lee et al, 2011).

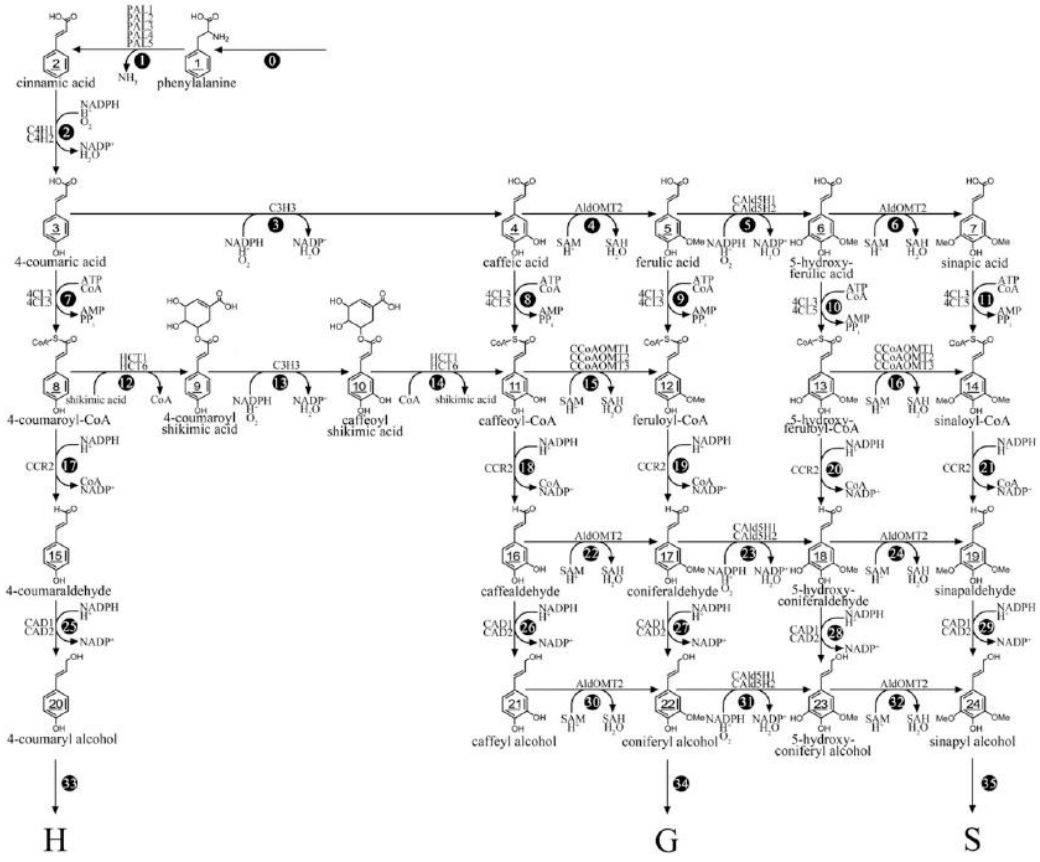

Figure 3.1: The Monolignol Biosynthetic Pathway in *P. trichocarpa*. Thirty-five metabolic fluxes (V0 to V35, represented by the circled numbers) mediate the conversion of 24 metabolites (underlined numbers) for monolignol synthesis by the 21 pathway enzymes (Wang et al, 2014).

Previous attempts towards quantitatively analyzing the monolignol biosynthesis pathway primarily relied on the usage of a static based kinetic model (Flux Balance Analysis) (Lee and Voit, 2010; Lee et al, 2011). The results from these analyses are largely hypothetical as the output is primarily based on constraints that are not

experimentally proven (Schallau and Junker, 2010). Results from prior simulations of the kinetic model for the metabolic network of monolignol biosynthesis (Wang et al., 2014) provided useful information about the role of various enzymes on the flux distribution within the pathway, as well as the role of individual enzymes on the lignin content and structure. However, the model did not incorporate the enzyme complex formation resulting from the interaction between Ptr4CL family enzymes. In another recently completed study on the lignin biosynthesis pathway, researchers revealed that Ptr4CL3 and Ptr4CL5 enzymes react with each other forming an enzyme-enzyme complex (Chen et al., 2013; Song et al., 2014).

3.2 Complex Formation:

It has been argued that for proper function, most proteins form complexes that may be composed of monomers or heteromers (Marsh et al, 2013). The formation of protein complexes provides an evolutionary advantage. For the case of the monolignol biosynthetic pathway, it was experimentally determined that the 4CL enzymes that convert the various hydroxycinnamic acids into CoA esters, exhibited protein - protein interactions. Chemical crosslinking coupled with immune-detection and mass spectrometry suggested that the Ptr4CL3 and Ptr4CL5 are present at a ratio of 3:1 in a complex (Chen et al, 2014). Using this information and assuming mass action kinetics, a mechanistic model was developed to quantify the rate of reaction resulting from the complex (Song et al., 2014). In single substrate reactions, Ptr4CL5 showed broader substrate affinity as compared to Ptr4CL3. Ptr4CL3 displayed competitive inhibition

while Ptr4CL5 showed both allosteric regulation and substrate self-inhibition (Chen et al., 2011). The rate equations describing the role of the complex were developed only for *4-coumaric acid* and *caffeic acid*, because Ptr4CL3 and Ptr4CL5 showed competitive and uncompetitive inhibition by substrates *4-coumaric acid* and *caffeic acid*. The rate equations were not developed for *Ferulic acid, 5-hydroxyferulic acid* and *sinapic acid* since they were found to be weak inhibitors (Chen et al., 2011, 2014).

The results from the 4CL transgenic experiments suggest that the down regulation of 4CL leads to a reduction in lignin content in tobacco, Arabidopsis, and aspen (Hu et al, 1999; Kajita et al, 1997; Lee at al, 1999) and to a higher amounts of cell wall–bound hydroxycinnamic acids (*p-coumaric, ferulic, and sinapic acid*) in tobacco and poplar (Hu et al, 1999, Kajita et al, 1997). The effects on S/G lignin composition are contradictory. In tobacco, a reduction in S units is reported (Kajita et al, 1997, Kajita et al, 1996) while in Arabidopsis, only G units are reduced (Lee et al, 1999). In transgenic aspen, the S/G ratio was shown to be similar to that of the control (2:1) (Hu et al, 1999). These results suggest that the exact role of 4CL on the lignin structure and composition is still not clear and a more detailed quantification of the effects of 4CL on the lignin biosynthetic pathway would greatly enhance our understanding of the role of the complex on the pathway.

In this work, we analyzed the mathematical model that was previously developed to simulate the lignin biosynthetic pathway (Wang et al., 2014) by incorporating the model developed to quantify the role of complex on the reaction rate (Chen et al, 2014

and Song et al, 2014). Using this model, we posed several questions to assess the role of the complex: (1) How does the steady state flux and steady state metabolite concentration change with uncertainty in initial metabolite concentrations? (2) What is the role of protein complex on the steady state flux and metabolite concentration under enzymatic perturbations? Finally, (3) Does the lignin biosynthetic pathway exhibit increased robustness/ homeostatic behavior in the presence of protein complex?

3.3 Results and Discussion

3.3.1 Steady State Metabolite Concentration Variation under WT conditions:

The steady state concentrations for the models with and without the complex were calculated for randomly sampled initial metabolite concentrations from a pre-specified range as mentioned in the methods section. The concentrations of all 21 enzymes belonging to 10 enzyme families were fixed at their WT concentrations, which can be found in Wang et al (2014). In addition, the phenylalanine concentration was fixed such that the resulting S/G ratio was 2:1. The mean steady state concentrations of the 24 metabolites with and without complex are shown in Figure 3.2. From the figure, the model predicts that all the metabolite concentrations in the pathway are similar in the absence and presence of Ptr4CL3-Ptr4CL5 complex except for p-*coumaric acid*, *caffeic acid* and *ferulic acid*. The steady state concentration of p-*coumaric acid*, *caffeic acid*, and *ferulic acid* in the presence of the complex are 20 fold, 24 fold, and 100 fold higher, respectively, than for the model without the complex. The primary reason for the increased steady state concentration in the presence of complex is because the

substrate specificity of Ptr4CL3 is higher for the acids as compared to Ptr4CL5, as well as the self-inhibition of caffeic acid in the presence of Ptr4CL5. In the presence of complex, the inhibitory effect of ferulic acid on *p-coumaric acid* and *caffeic acid* is higher, due to the accumulation of ferulic acid, hence resulting in an accumulation of *p-coumaric acid* and *caffeic acid* (Chen et al, 2014). As with the steady state concentrations, the steady state flux results are identical to the model without the complex except for the steady state fluxes V_3, V_7 and V_8. The total lignin content and composition was also found to remain unchanged. These results support the recent findings that have demonstrated that lignin biosynthesis is more resistant to perturbations (Weng et al, 2010). This resistance to perturbation is also due to redundancies in the monolignol biosynthesis pathway where six of the ten enzyme families have functional redundancies (Shi et al, 2010).

Under WT conditions, the presence of the Ptr4CL3-Ptr4CL5 complex provides an additional path for CoA ligation, which is in addition to *p-coumaric acid*, *caffeic acid* is also a preferred substrate of Ptr4CL in the presence of this complex. In the absence of the Ptr4CL3-Ptr4CL5 complex, the *P-coumaric acid* substrate specificity of Ptr4CL3 is 4 fold higher than Ptr4CL5. In the presence of the Ptr4CL3-Ptr4CL5 complex, the majority of the flux in the pathway is routed through p-*coumaric acid* with a small portion of the flux flowing through *caffeic acid* (300% increase). The small portion of the flux that flows from *cinnamic acid* to *caffeic acid* and *caffeic acid* to *caffeoyl-CoA* as observed for the case of the Ptr4CL3-Ptr4CL5 complex is due to the controlling role that Ptr4CL5 plays in

lignin formation. Ptr4CL5 in the complex controls the amount of activation and inhibition (Chen et al, 2013, 2014).

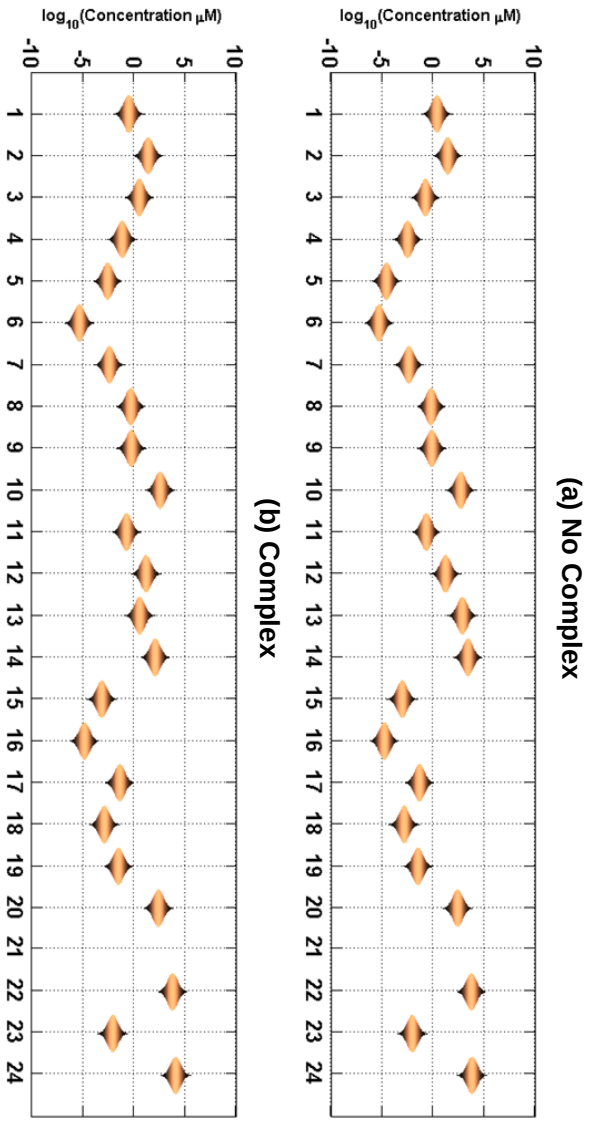

Figure 3.2: Steady state metabolite concentrations observed for the model without the complex (a) and model with the complex (b) under WT enzyme concentrations. The concentrations are in the log scale because of the variability in the steady state concentrations for different metabolites. The distribution of steady state values is a result of 10,000 runs performed under varying initial concentrations of the metabolites.

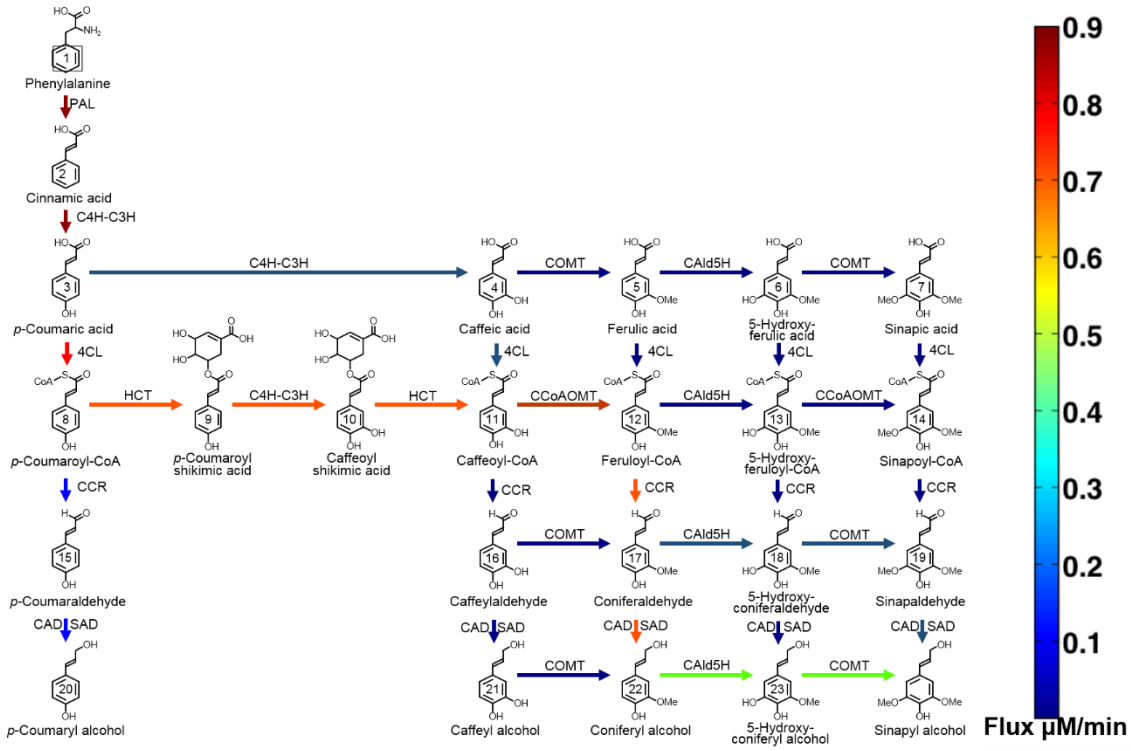

Figure 3.3: Steady state flux pattern observed for the model without the complex WT enzyme concentrations. Colored arrows represent the magnitude of flux and the colors can be mapped to their flux values with the colorbar.

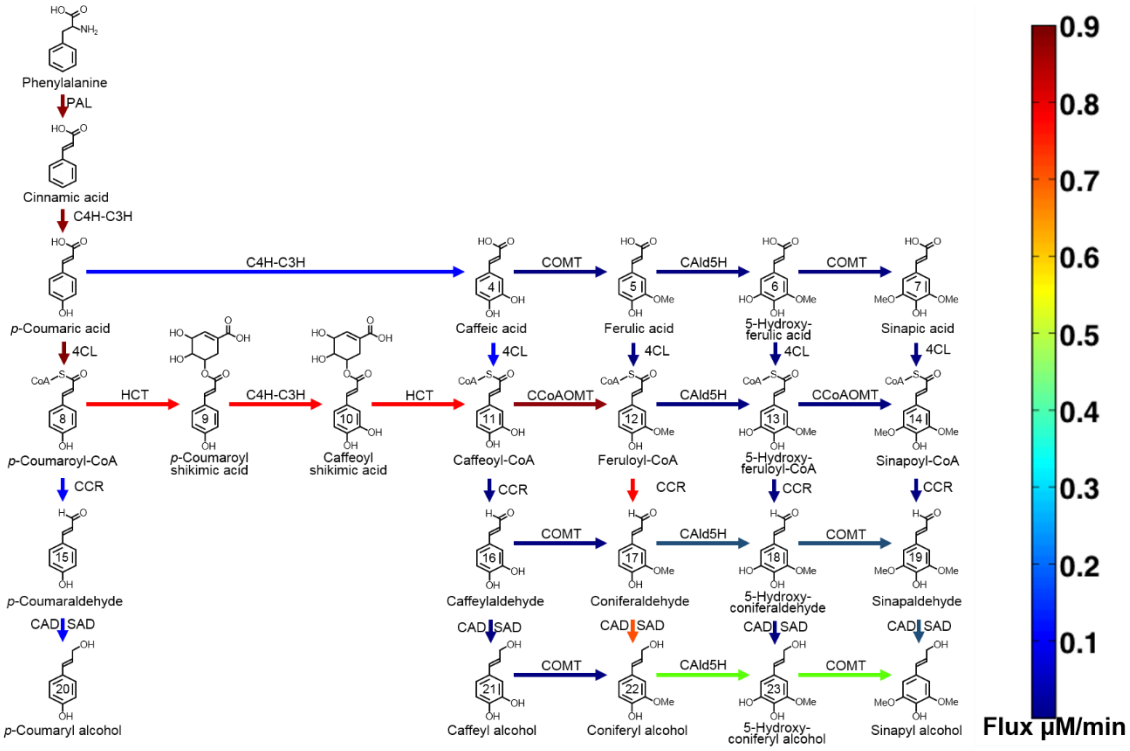

Figure 3.4: Steady state flux pattern observed for the model with the complex under WT enzyme concentrations colored arrows represent the magnitude of the flux as shown in the color bar.

3.3.2 Role of Ptr4CL3-Ptr4CL5 complex on the flux distribution when Ptr4CL enzymes are perturbed:

In this section, we assessed the role of Ptr4CL3-Ptr4CL5 complex on the lignin biosynthesis pathway in the presence of enzymatic perturbations. The enzyme concentration of Ptr4CL3 and Ptr4CL5 were varied from 0 to WT and 10,000 pairs of different combinations of Ptr4CL3 and Ptr4CL5 concentrations were randomly selected. For each pair, the resulting steady state concentrations of the twenty-four metabolites were calculated.

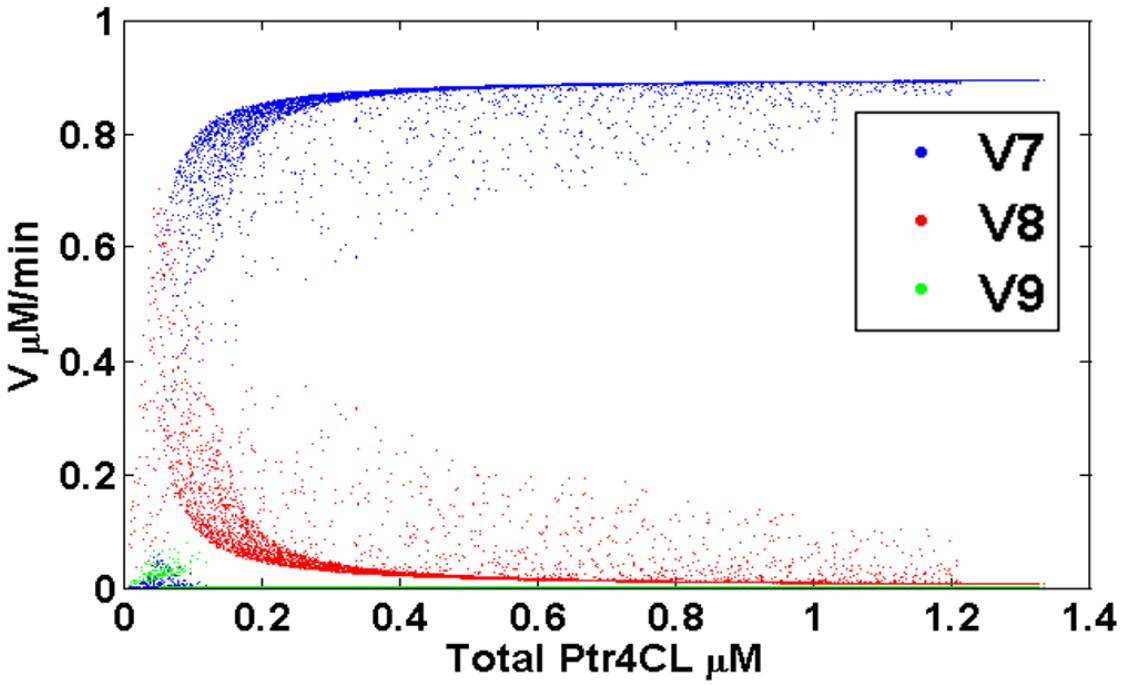

Figure 3.5: Steady state flux (V_7, V_8 and V_9) variation as a function of total Ptr4CL concentration for models without the Ptr4CL3-Ptr4CL5 complex.

Because the Ptr4CLs primarily mediate the conversion of hydroxycinnamic acids to CoA derivatives, which are represented by the reactions V_7, to V_{11}, we tried to assess the localized effect of varying Ptr4CL concentrations on the reaction flux when subjected to enzymatic perturbation. Prior research showed that the fluxes V_{10} and V_{11} did not contribute towards monolignol biosynthesis (Chen et al, 2013). The variation of reaction flux V_7 as a function of total Ptr4CL in the absence of the Ptr4CL3-Ptr4CL5 complex is shown in Figure 3.5. In absence of the complex under WT concentration, the flux through the pathway is primarily routed through V_7, hence the steady state flux V_7 corresponds to the input flux 0.9 μM/min and the flux V_8 and V_9 are 0 μM/min. As the concentration of Ptr4CL3 and Ptr4CL5 are reduced from WT concentration, we observe

a decrease in the steady state flux V_7 with a corresponding increase in flux V_8. At concentrations corresponding to 10% of WT concentration, the steady state flux V_7 and V_8 are equal and further decrease in Ptr4CL concentration results in caffeic acid being the primary substrate. These results are consistent with findings by Lin et al. (2015) that 3-hydroxylation for monolignol biosynthesis can occur at both V_3 and V_{13}.

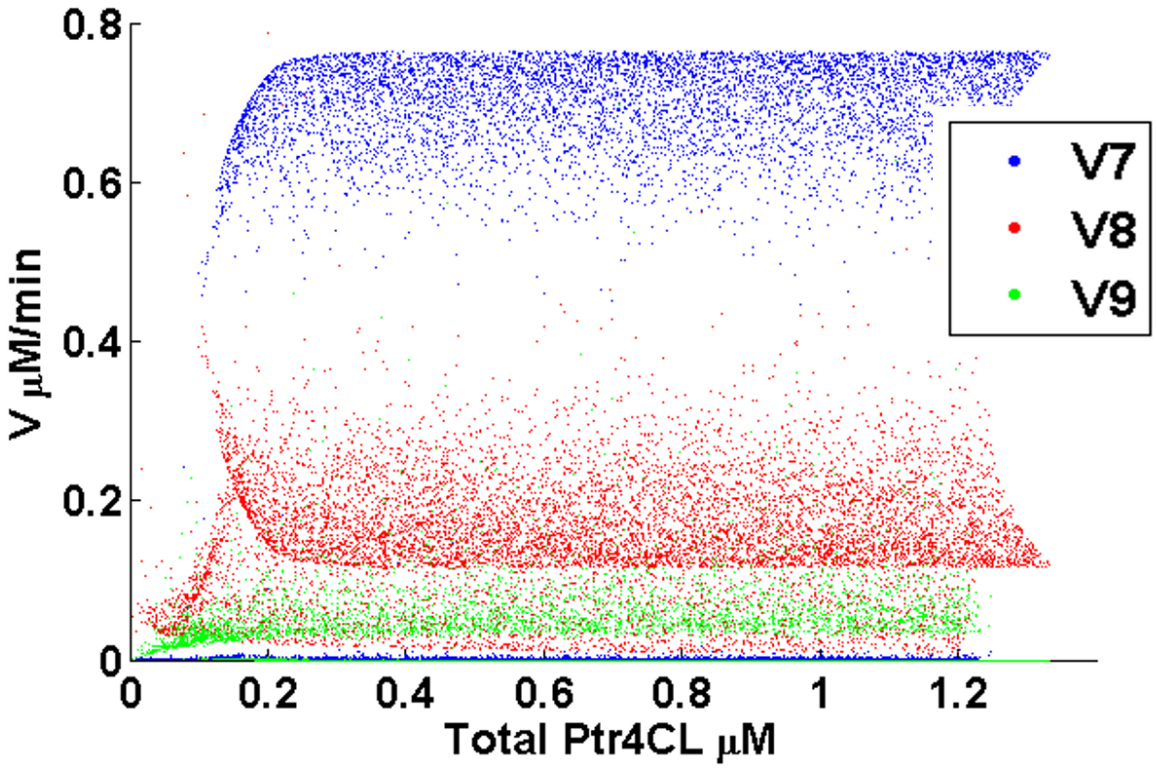

Figure 3.6: Steady state flux (V7, V8 and V9) variation as a function of total Ptr4CL concentration for models without the Ptr4CL3-Ptr4CL5 complex.

The presence of Ptr4CL3-Ptr4CL5 complex results in two steady state flux regimes due to the random sampling of different concentrations of Ptr4CL3 and Ptr4CL5 as seen in Figure 3.6. The flux regime close to zero corresponds to very low

concentrations of Ptr4CL5 and such low concentrations would not result in a viable plant. The spread in the predicted steady state flux is attributed to the random sampling of the Ptr4CL3 and Ptr4CL5 concentrations and depending on the concentration of Ptr4CL5, the resulting flux value may fall into one of the regimes. As the concentration of Ptr4CL3 and Ptr4CL5 is reduced, we see a decrease in the steady state flux V_7 and a corresponding increase in the flux V_8. As the concentration of the enzymes is reduced further up to 10% of its WT concentration, the flux through V_7 reduces to zero. Any further reduction in the enzyme concentration results in the decrease in the flux V_8 that eventually reduces to 0 µM/min. The major difference in flux variations for the models with and without the Ptr4CL3-Ptr4CL5 complex is that the complex results in the increased affinity of caffeic acid as the substrate; hence providing an alternate pathway towards the synthesis of S and G monolignols. This result is consistent with the observation that caffeic acid is the most abundant hydroxycinnamic acid in Stem Differentiating Xylem (SDX) of P. trichocarpa (Chen et al., 2013).

Two main feedback inhibitions primarily affect the steady state flux distribution when the Ptr4CL3 and Ptr4CL5 concentrations are perturbed: (1) the feedback inhibition between p-*coumaric acid* and *caffeic acid*, and (2) the feedback inhibition between p-*coumaric acid* and *ferulic acid*. The steady state distribution of metabolic flux V_7, V_8 and V_9 is shown in Figure 3.7. The violin plot shows the distribution of the steady state flux. The violin plot assumes that the steady state flux follows a normal distribution. The mean and median of the steady state flux distribution is shown in black and red lines respectively. The presence of two distinct flux regimes can be clearly seen for the case

of flux V_7. The two flux regimes were generated from the presence of the bimodal steady state distribution of the *caffeic* and *ferulic acid* concentrations. The high concentration of Ptr4CL5 results in low steady state concentrations of *caffeic acid* and *ferulic acid*. Reduction in concentration of Ptr4CL5 results in an accumulation of *caffeic acid* and *ferulic acid* as seen in Figure 3.8.

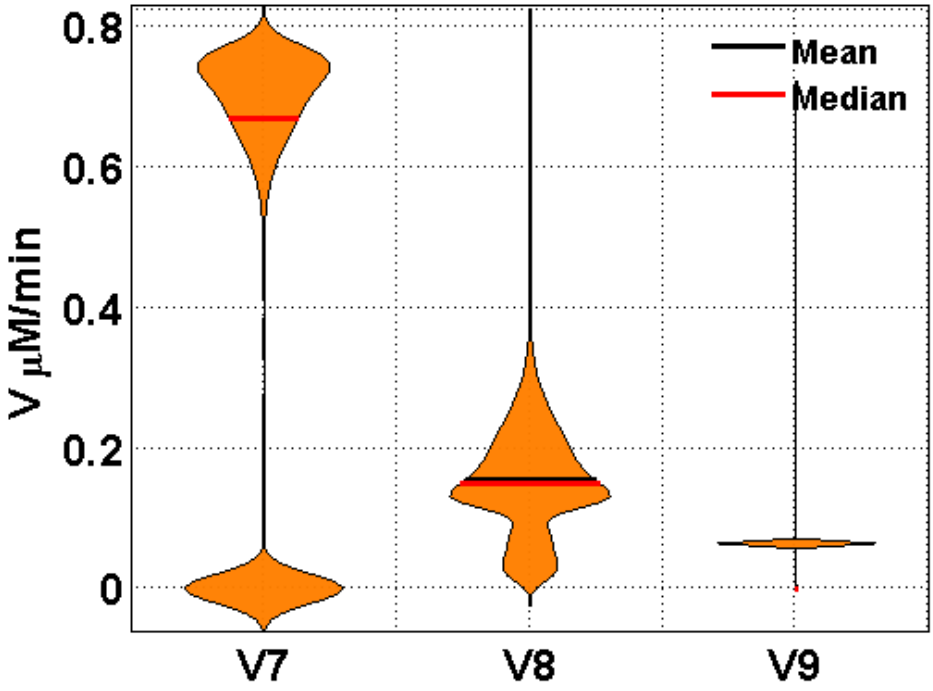

Figure 3.7: Violin plot showing the steady state flux (V_7, V_8 and V_9) variation as a function of total Ptr4CL concentration for models with the Ptr4CL3-Ptr4CL5 complex.

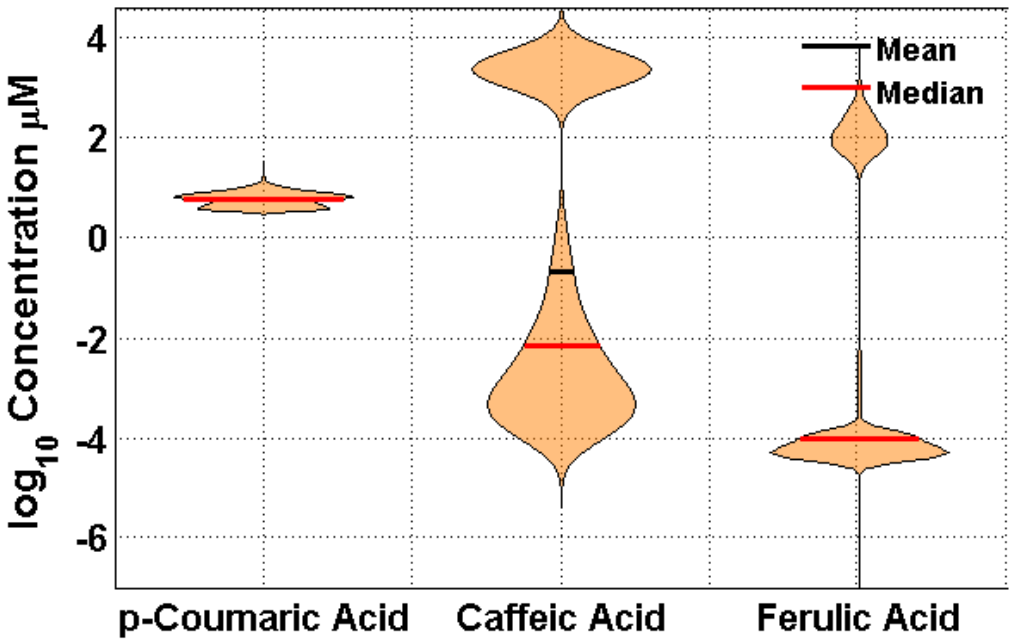

Figure 3.8: Variation of P-coumaric acid (µM) concentration as a function of total cinnamic acid concentration (µM**)** in the presence of complex

The role of individual enzymes on the steady state flux distribution of V_7 can be visualized with the help of contour plots presented in Figures 3.9 and 3.10. In the absence of the Ptr4CL3-Ptr4CL5 complex, the changes in steady state flux (V_7) is primarily brought about by changes in levels of Ptr4CL3 as seen in Figure 3.9. Under WT levels of Ptr4CL3 and Ptr4CL5, the steady state flux corresponds to a maximum value of 0.8 µM/min shown in red, which represents almost 75% of the steady state solutions. In order to observe a change in the steady state flux by 0.1 µM/min unit, a reduction in Ptr4CL3 concentration of up to 90% of the WT levels would be required and achieved by holding the Ptr4CL5 concentration at WT. On the other hand, if the concentration of Ptr4CL5 were reduced by 90% of its WT concentration, then the

concentration of Ptr4CL3 would have to be reduced by 80% of the WT concentration to observe a 10% reduction in the steady state flux. These results suggest that in the absence of Ptr4CL3 – Ptr4CL5 complex, the model is robust to perturbations, requiring a high degree of enzymatic perturbations to observe significant change in the steady state flux. The robustness is primarily due to the abundance of Ptr4CL3 levels when compared to Ptr4CL5 levels. These results provide a mechanistic explanation to why severe downregulation of *4CL* genes is needed to affect lignin biosynthesis in transgenic *P. trichocarpa* (Wang et al., 2014).

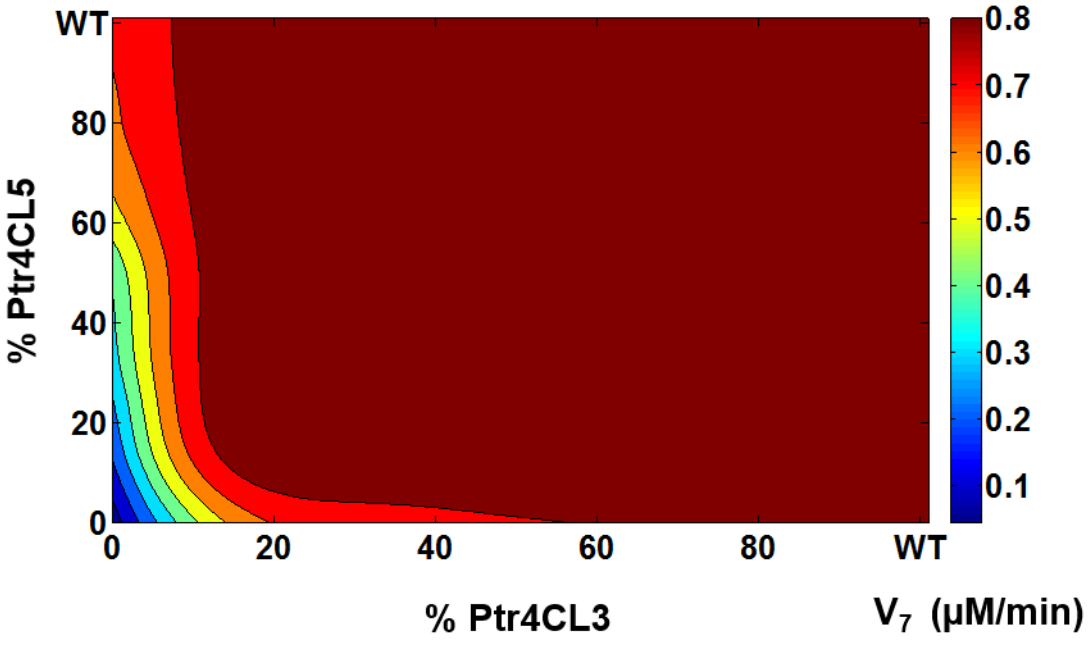

Figure 3.9: Contour plot showing the variation of steady state flux (V7) as a function of Ptr4CL3 and Ptr4CL5 concentration in the absence of a complex. The axis values represents the percentage of the protein concentration as a function of the wild type concentration.

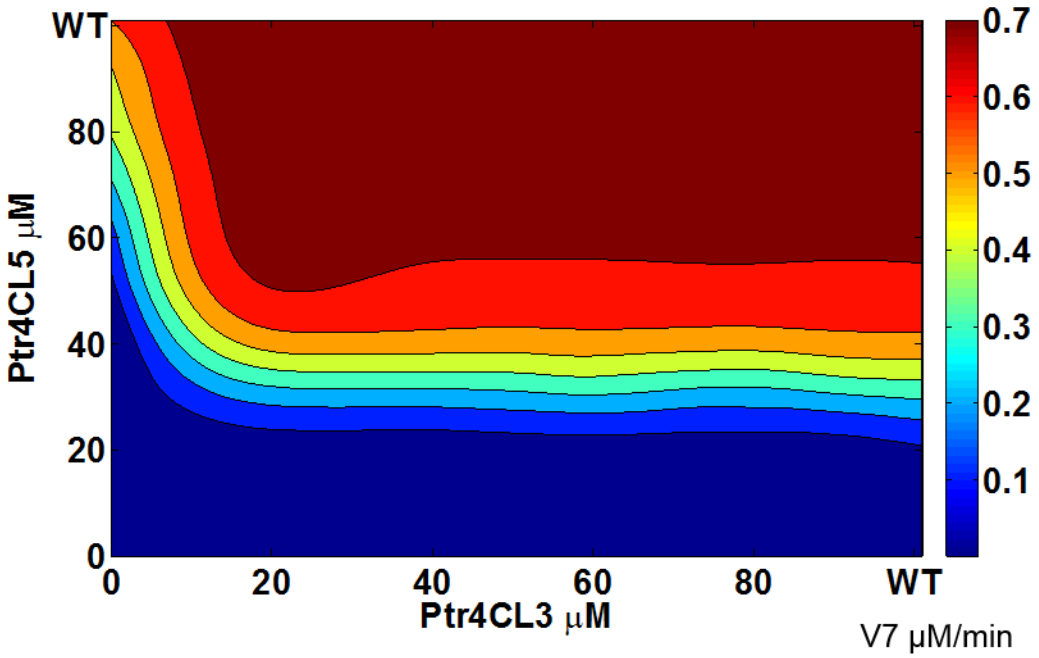

Figure: 3.10 Contour plot showing the variation of steady state flux (V7) as a function of Ptr4CL3 and Ptr4CL5 concentration in the presence of a complex, the colorbar represents the flux values (μM/min).

In the presence of the Ptr4CL3-Ptr4CL5 complex, the variation in the steady state flux V_7 can be visualized in Figure 3.10. Under WT Ptr4CL3 and Ptr4CL5 concentrations, the steady state flux is at its maximum value of 0.7 μM/min. As seen in the Figure 3.10, the change in flux V_7 is primarily due to the changes in Ptr4CL5 concentration. When the concentration of Ptr4CL5 is reduced by 50% of its WT concentration while holding the concentration of Ptr4CL3 at its WT level, we observe a 10% reduction in the steady state flux. On the other hand, if the concentration of Ptr4CL3 was reduced by 80% of its WT, then the Ptr4CL5 concentration needs to be reduced by 20% to observe a 10% change in the steady state flux.

The variations of all the metabolic fluxes involved in the biosynthesis are summarized in supplementary Figure A1. In the presence of a complex, the variation of Ptr4CL concentration results in a large variability of steady state flux. The variation is primarily due to the changes in concentrations of Ptr4CL3-Ptr4CL5 complex bimodal flux regimes. The model with the complex is sensitive to the concentrations of Ptr4CL5 enzyme hence variation in Ptr4CL5 concentration results in large variability in steady state flux.

3.3.3 Role of Ptr4CL complex on S and G Monolignols, Lignin Content and Composition:

So far, we discussed the model's prediction on the impact of varying Ptr4CL concentrations on the steady state metabolite and flux distributions for the model with and without the complex. In this section, we extend the discussion to the total lignin content and ratio. Down-regulation of 4CL activity results in decreased lignin content in alfalfa, arabidopsis, tobacco, aspen (*Populus tremuloides*) and hybrid white poplar (*Populus tremula X Populus alba*) (Dixon et el, 2013). This decrease in lignin content is accompanied by a slight increase in the S/G ratio in alfalfa, a greater increase in S/G ratio in Arabidopsis, an unchanged S/G ratio in aspen, and a decreased S/G ratio in hybrid white poplar. The reduction in lignin content in Arabidopsis following down-regulation of 4CL is achieved by a decrease in G units but not S units, leading to a higher S/G ratio and the plants appearing phenotypically normal (Lee et al, 1997). Researchers have shown that a down regulation of Ptr4CL enzyme in *poplar* results in a decrease in total lignin content and an increase in S/G ratio (Voelker et al, 2010; Wang et al, 2014). At low concentrations of Ptr4CL, lignin is primarily composed of S units,

which results in an elevated S/G ratio. In the absence of a complex and for a wide range of Ptr4CL concentrations, the S and G subunits do not change from the WT condition. At very low concentrations of Ptr4CL, the S and G subunits approach 0. In the presence of a complex and at higher concentrations of Ptr4CL, lignin is primarily composed of S and G monolignols; but as the concentration of Ptr4CL is reduced, the model predicted an increase in the S/G ratio. This result is in agreement with the results obtained by Li et al (1997) and Hu et al (2001). From the above results, we can conclude that the inclusion of the Ptr4CL complex in the model provides us with a more comprehensive model that better predicts the experimental lignin composition results for *P. trichocharpa,* when the pathway is subjected to *4CL* perturbations.

The changes in Ptr4CL3 and Ptr4CL5 did not affect the total lignin content and lignin composition for the model in the absence of a complex. This lack of impact on lignin content and lignin composition was expected since the changes in the concentrations of Ptr4CL3 and Ptr4CL5 did not significantly affect the steady state distribution of the metabolic flux. However, the model with the complex does display some deviation from the wildtype, leading to a 20 percent reduction in lignin content that does not have wildtype properties. The variation of S and G monolignol units when the Ptr4CL3 and Ptr4CL5 concentrations are perturbed in the presence of the complex is shown in Figures 3.11.a and 3.11.b respectively. About 80% of the S and G units correspond to the WT levels, while at low concentrations of Ptr4CL the majority of lignin is composed of S units and very low levels of G units

The effect of variation of Ptr4CL3 and Ptr4CL5 on the S/G ratio in the absence of a complex can be seen in the contour plot shown in Figure 3.12. As seen in Figure 3.12, changes in the Ptr4CL concentration do not result in a significant change in the S/G ratio over a wide range of values. When the Ptr4CL3 concentration was reduced to less than 10% of its WT concentration, the S/G ratio increased from its WT ratio. However, these results are not biologically achievable because reduction of Ptr4CL concentrations would result in plants with very low lignin content, a result that would produce plants lacking mechanical strength.

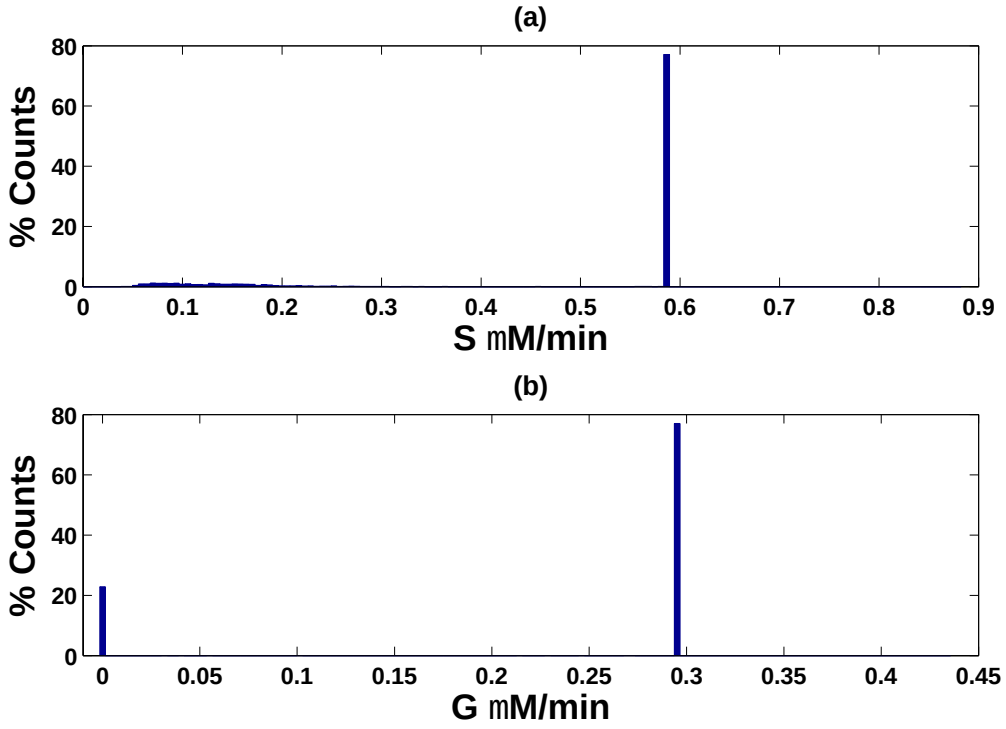

Figure 3.11: Variation of S and G monolignol units resulting due to changes in levels of Ptr4CL3 and Ptr4CL5 concentrations in the presence of a complex.

The variation in S/G ratio for the model with Ptr4CL3-Ptr4CL5 complex as a function of the Ptr4CL3 and Ptr4CL5 concentrations is shown in Figure 3.13. The

contour plot suggests that for the model with the complex, changes in S/G ratio is primarily due to variation in Ptr4CL5. The S/G ratio is robust to perturbations until the concentration of Ptr4CL5 is reduced to 60% of the WT concentration; any further change in the Ptr4CL5 results in a linear change in S/G ratio. These results confirm the regulatory role of Ptr4CL5 in the presence of the Ptr4CL3-Ptr4CL5 complex (Chen et al., 2013 and 2014).

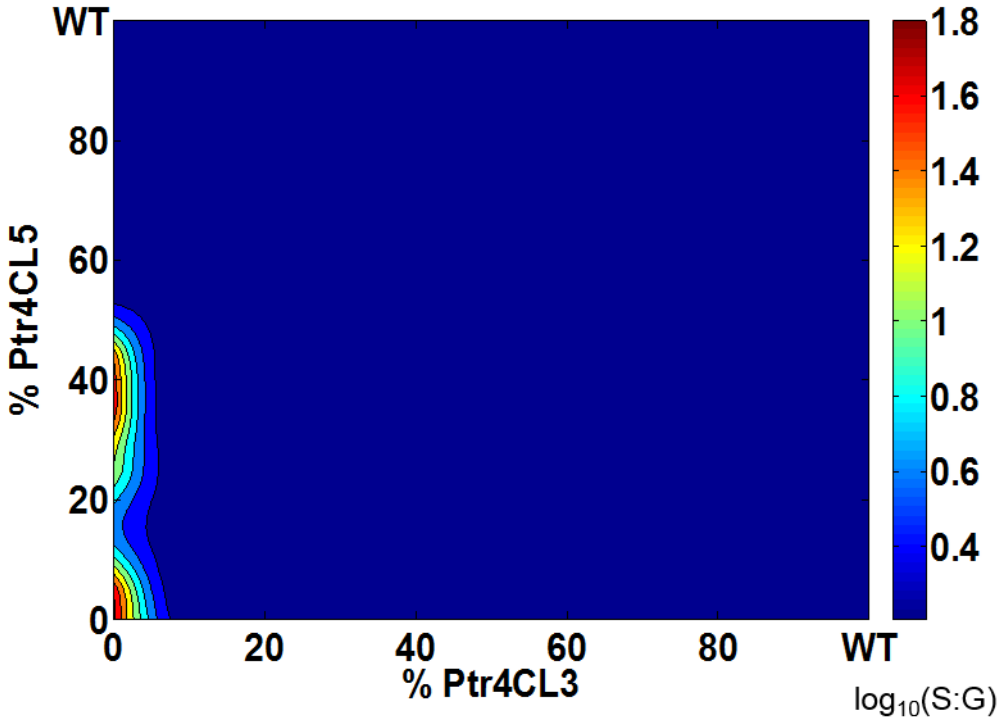

Figure 3.12: Variation of S/G ratio as a function of Ptr4CL3 and Ptr4CL5 Concentration in the absence of a complex. The color bar shows the variation of S/G ratio in log scale, where S/G ratio of 2 corresponds to a value of 0.3 in log scale.

From the above results, the variation in lignin content and composition as a function of changes in concentrations of Ptr4CL is significantly influenced by the presence of the Ptr4CL3-Ptr4CL5 complex. Hence, the inclusion of the complex into the

model enhances our understanding about the regulation of metabolic flux through the pathway, which is otherwise not apparent in the absence of the Ptr4CL3-Ptr4CL5 complex. The role of multi-enzyme complex on the genetic manipulation of lignin was previously studied by Campbell and Sederoff (1996). However, the exact role of the complex was not evident at that time. The simulation results of the PKMF model suggest that there is an opportunity to genetically modify the lignin content and structure by targeting the Ptr4CL5 enzyme. Although the role of the Ptr4CL3-Ptr4CL5 complex on the lignin content and structure were identified, the effect of changing levels of Ptr4CL3-Ptr4CL5 enzyme concentrations on the individual reaction rates can be quantified by performing a sensitivity analysis.

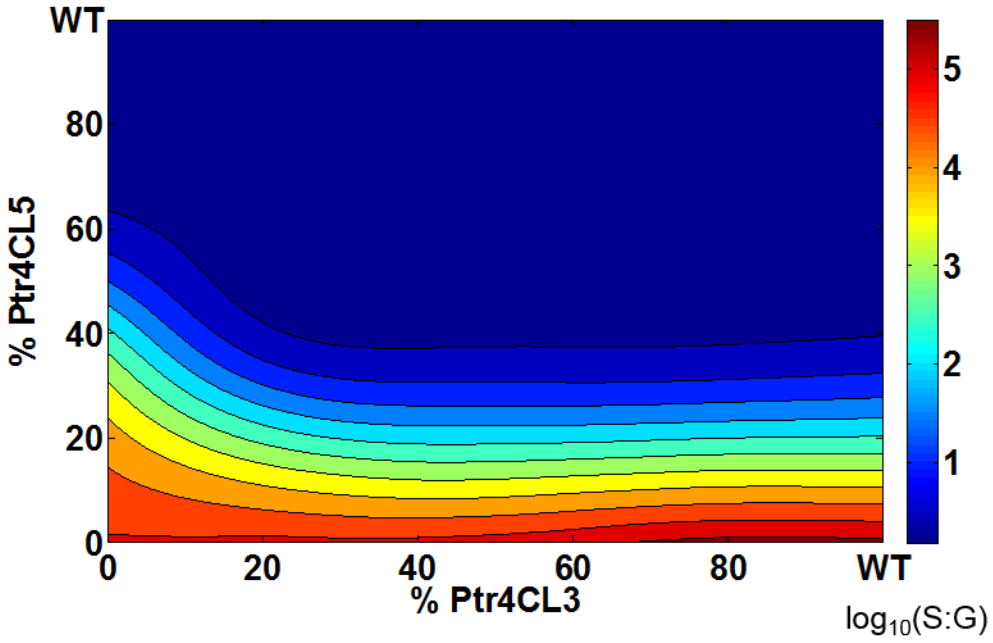

Figure 3.13: Variation of S/G ratio as a function of Ptr4CL3 and Ptr4CL5 concentration in the presence of a complex. The color bar shows the variation of S/G ratio in log scale.

3.3.4 Sensitivity Analysis of Ptr4CLs on the Monolignol Biosynthetic Pathway:

As discussed previously, the sensitivity analysis considers changes to one parameter at a time, whereas in biological systems multiple parameters may act together to produce an effect. Thus, in a biosynthetic pathway, it is important to understand the role of multiple parameters in the pathway. Since the variation of the reaction flux with Ptr4CL enzymes were non monotonic, we used the variance decomposition method (Sobol, 1990) to assess the sensitivity of steady state reaction flux to Ptr4CL concentration. The concentrations of other enzymes were fixed at their WT levels.

The sensitivity of the steady state flux to changes in the concentrations of Ptr4CL3 and Ptr4CL5 enzymes in the absence and presence of the Ptr4CL3-Ptr4CL5 complex is shown in Figures 3.14 and 3.15, respectively. The first order sensitivity index measures the percentage of variance explained by varying the enzymes independently. To quantify the role of individual enzymes on the resulting flux or substrate concentration, the first order sensitivity indices for Ptr4CL3 and Ptr4CL5 were compared against a dummy parameter, which is an arbitrary parameter that has no influence on the model (Stalteli et al, 2000). The dummy parameter has no effect on the model as it does not appear in any of the equations and hence should have a very low sensitivity. The sensitivity index for Ptr4CL3 and Ptr4CL5 that is significantly different from the dummy parameter indicates that the particular enzyme has an influence on the resulting flux or substrate concentration. In Figure 3.14, the reaction flux is primarily affected by changes in the levels of Ptr4CL3 in the absence of a complex. The fluxes that are

affected mainly involve the reactions mediated by Ptr4CL enzymes and the reactions in the vicinity of the Ptr4CL pathway.

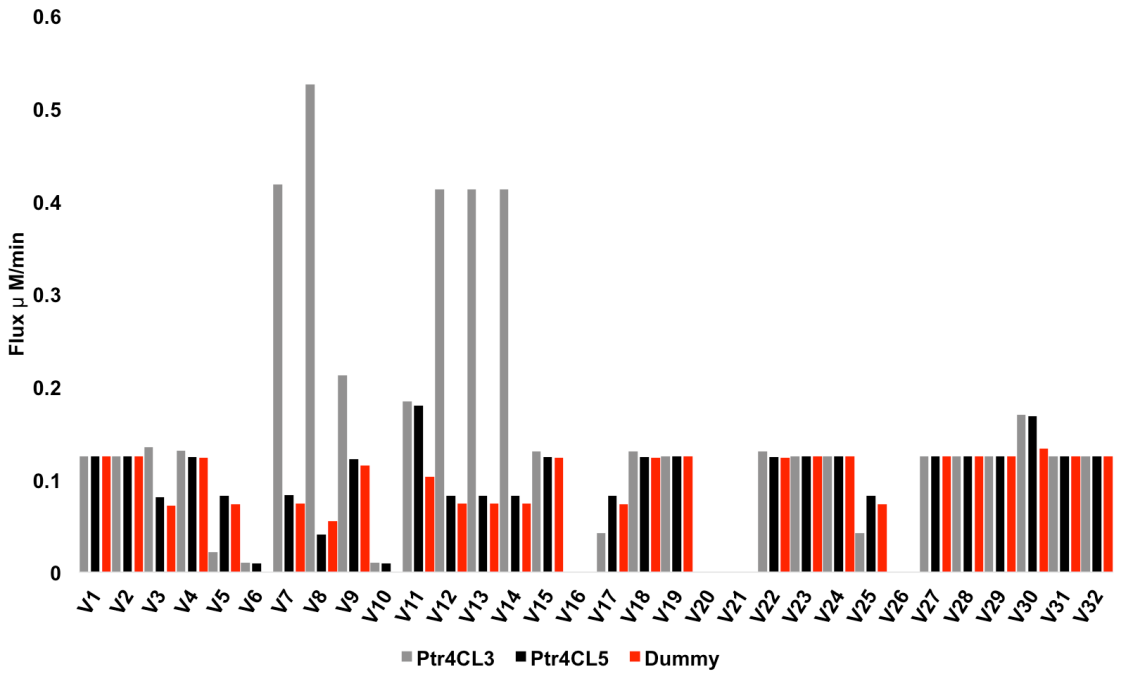

Figure 3.14: The first order sensitivity index for monolignol flux with respect to Ptr4CL3 and Ptr4CL5 concentrations for the model without the complex.

In the presence of a Ptr4CL3-Ptr4CL5 complex, the reactions mediated by Ptr4CLs (V_7, V_8 and V_9) are especially sensitive to the changes in Ptr4CL5 (Figure 3.15). The terminal reactions towards the end of the biosynthetic pathway, which are directly responsible for the lignin content and structure (V_{27}, V_{29}, V_{31}, and V_{32}) all show a slightly higher sensitivity to variation in Ptr4CL3 and Ptr4CL5 concentrations for the model with the complex. The changes in the terminal metabolic flux might indicate the reason as to why we observe a distribution of S/G ratios around the WT Ptr4CL concentrations (see Figure 3.13).

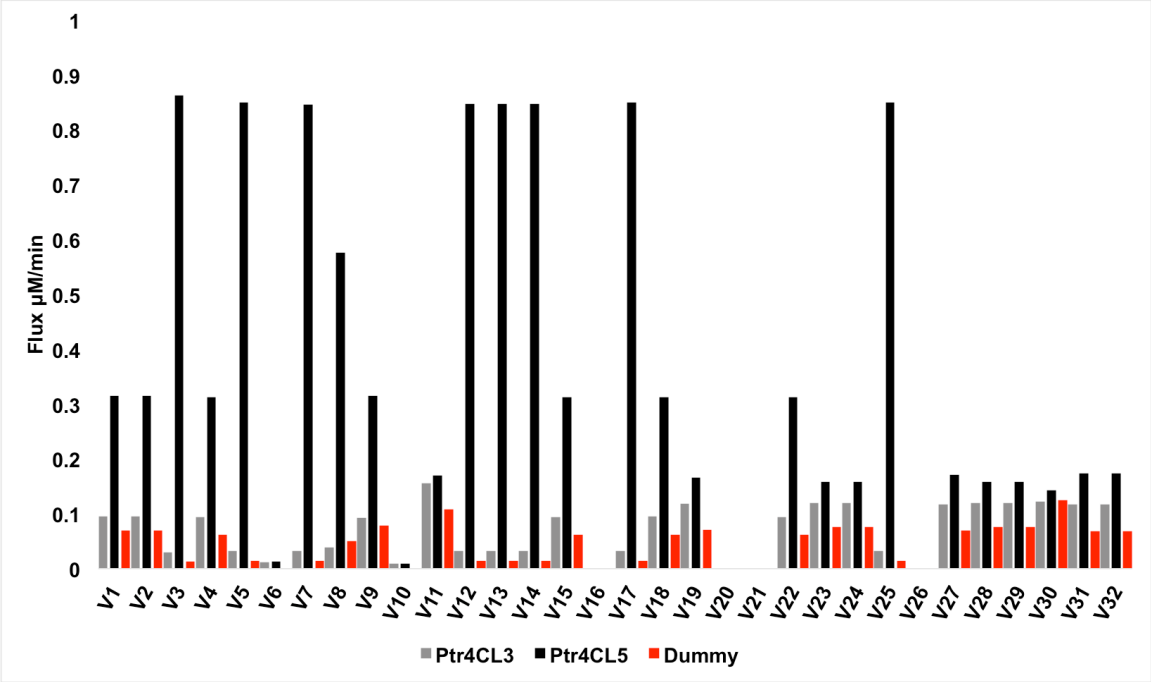

Figure 3.15: The first order sensitivity index for monolignol flux with respect to Ptr4CL3 and Ptr4CL5 concentrations for the model with complex.

3.3.5 Robustness of Monolignol Biosynthetic Pathway:

Robustness, a property that allows a system to maintain its functions in the presence of large environmental perturbations (Kitano, 2004, 2007), is a widely discussed topic in systems biology (Zhou et al, 1998; Steur, 2004; Kitano, 2004, 2007). The primary function in our case represents maintaining the levels of S/G ratio and the total lignin content (S+G). The biosynthesis of monolignols is a metabolic grid, which means that there is more than one route in the synthesis of monolignol subunits (Figure 3.1). This potential flexibility in synthesis routes gives plants the ability to maintain near wild type lignin structure and composition when a step is inhibited. In the presence of a complex and depending on the level of Ptr4CL perturbation, the flux through the monolignol pathway can be routed through *P-coumaric acid* or *caffeic acid* (Figure 3.6).

In case of extreme perturbation, plants are known to have incorporated other phenolic components into the monolignol subunits (Ralph et al, 2004). As seen for the case of the model with the complex, the changes in Ptr4CL levels resulted in a proportional change in the S and G monolignol levels, with a majority of S and G monolignol subunits distributed around their WT concentrations, suggesting that the presence of a Ptr4CL complex results in increased plasticity of the pathway.

Robustness applies to a few variables or a particular part of the model and not necessarily for the entire system, a key difference from stability where it is a property of the entire system. It has been suggested that there is some form of equilibrium in robustness; where some part of the pathway may be sensitive to perturbations while the remaining pathway remains unaffected (Zhou et al, 1998). A similar scenario was observed in our case, where the reaction flux corresponding to the early steps of the monolignol biosynthesis pathways were more sensitive to the changes in the levels of Ptr4CL3 and Ptr4CL5 concentrations. As we move downstream towards the terminal reactions, the sensitivity of those reactions was considerably reduced. From the results of sensitivity analysis, the pathway is fairly robust to changes in levels of Ptr4CL3 and Ptr4CL5 concentrations under mild perturbations. These results explain the role of the complex as well as provide insights into why enzymes form a complex.

3.3.6 Stability Analysis

The role of Ptr4CL3-Ptr4CL5 complex on the overall monolignol biosynthetic pathway can also be assessed by performing stability analysis. Since the presence of

the complex affects the steady state flux distribution through feed forward and feedback inhibition, the complex affects the dynamic properties of the biosynthetic pathway. The Jacobian matrix was evaluated for models with and without the complex based on the i^{th} $\lambda_{Re}^{max} > 0$ implying instability of the metabolic state (Steur et al, 2007). A total of 10,000 model evaluations were performed for different steady state metabolite concentrations resulting from different enzyme concentrations of Ptr4CL3 and Ptr4CL5 for the model with and without the Ptr4CL3-Ptr4CL5 complex. Based on the sign of the Eigenvalue for each set of metabolite and enzyme concentrations, the model was classified into stable or unstable conditions. The resulting Eigenvalues were then plotted as a histogram of Eigenvalues for 10000 iterations. The cumulative distribution function (CDF) of Eigen value distribution is shown in Figure 3.16 for the case of model with and without the complex. From the Figure 3.16, it can be seen that the lignin biosynthetic model is inherently stable for both cases, which suggests that under perturbations, the system is robust to small changes in enzyme levels. These results support the experimental findings that the S/G ratio of lignin in most transgenic plants is around the WT S/G ratio. The presence of complex increased the percentage of model stable conditions from 70% to 92%. The percentage of stable models were calculated by counting the number of models with negative eigenvalues (λ_{max}) divided by the total number of models (10000). Although a large percentage of the models are stable, a small percentage of unstable models cannot be neglected, as these steady states indicate that the network could be driven out of the observed steady state when Ptr4CL3 and/or Ptr4CL5 concentrations are perturbed. The increased stability in the model containing the enzyme complex enables plants to maintain the lignin composition and structure under

normal external or internal perturbation. The results from stability analysis suggests that although the efficiency of the Ptr4CL enzymes increase when they form a complex, the primary reason for the complex formation may be because they are able to help the plant maintain homeostasis more efficiently. This claim is supported by the results from the model that, in the presence of the Ptr4CL3-Ptr4CL5 complex, the pathway is able to use *P-coumaric acid* and *caffeic acid* as substrates to produce lignin. The results of varying the enzyme concentrations suggest that the monolignol biosynthetic pathway is homeostatic either in the absence and presence of the complex. The flux distribution for the model without complex appears to be unaffected with the variation in Ptr4CL concentrations. However in the presence of a complex, the model exhibits multi-stability; i.e., the steady state flux corresponding to high enzyme concentrations are similar to the steady state flux observed for the model without complex. At very low concentrations of Ptr4CL, however, the model results in an additional steady state.

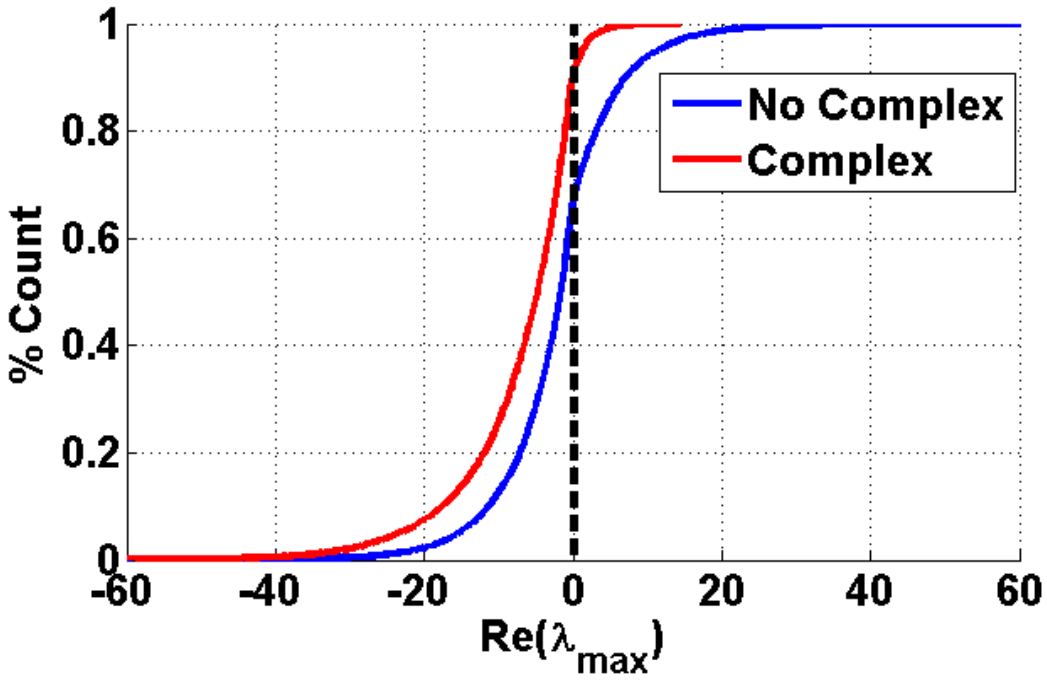

Figure 3.16: Cumulative distribution function plot of Eigenvalues for the model with and without the complex under WT enzyme concentrations.

3.4 Conclusion:

The simulations performed using the Monte Carlo simulation of PKMF model enables us to identify all the feasible steady state concentrations of the metabolites involved in the pathway. This approach provides a comprehensive analysis of the entire monolignol biosynthesis pathway involving all the metabolites, enzymes, and the multi-enzyme complex. The results from the analysis enable us to quantify the effect of perturbation of Ptr4CL enzymes on the steady state flux distribution within the pathway as well as its effect on the lignin content and structure. The overall steady state flux

distribution within the pathway was mostly similar for the models in the absence and presence of the Ptr4CL3-Ptr4CL5 complex under WT conditions. In the presence of perturbations, the simulation results suggest that the presence of Ptr4CL3-Ptr4CL5 complex in the model results in a redundant pathway towards the biosynthesis of S and G monolignols. The perturbation of the model in the presence of the Ptr4CL3-Ptr4CL5 complex results in a bimodal distribution of steady state flux. This bimodal distribution is primarily due to the feedback inhibition that exists between the various hydroxy cinnamic acids involved in the Prt4CL pathway. The analysis of the PKMF model also suggests that the Ptr4LC5 enzyme plays a key regulatory role in the modulation of the metabolic flux. These results further confirm the experimental findings that Ptr4CL5 is essential for regulating the levels of caffeic acid that is essential to moderate the CoA flux ligation, which is an essential step in the monolignol biosynthesis (Chen et al, 2013). Sensitivity analysis of the PKMF model suggests that in the absence of a complex, the perturbation of Ptr4CL3 and Ptr4CL5 enzymes results in only a localized change in the steady state metabolic flux. Whereas, in the presence of complex, the steady state flux was sensitive to changes in Ptr4CL5 concentrations, further confirming the regulatory role of Ptr4CL5 enzyme. An interesting result that was observed for the model with the Ptr4CL3-Ptr4CL5 complex is the presence of redundant pathways in the CoA ligation step. Previous results on redundancy in pathways, suggests that biochemical pathways display a high level of redundancy against random failures (Stelling et al., 2002). It also explains why, despite the perturbations, the metabolic networks are unaffected (Kitano, 2002, 2004a, 2004b). Thus, the presence of the complex enhances the robustness of the pathway in the presence of perturbations. The

robustness of a metabolic network can be attributed to the stability of the metabolic states of the different metabolites involved in the pathway. The local stability analysis suggests that in the presence of Ptr4CL3-Ptr4CL5 complex, the stability of the metabolic network increases by 15%.

3.5 Methodology:

3.5.1 Computing Steady State Metabolite Concentration:

Biological networks are characterized by highly nonlinear interactions between various components within the network. One of the results of these nonlinear interactions in biological networks is the presence of multiple steady states, which correspond to metabolic flux conditions that the cell achieves during growth or when perturbed by external stressors. It's because of the presence of multiple steady states, that the biological systems are able to perform the required functions, even in the presence of external perturbations (Steur et al, 2007). For the monolignol biosynthetic pathway, the existence of multiple steady states corresponds to different phenotypes (lignin structure). The identification of multiple steady states would aid plant biologists in determining the biological conditions that would lead to a particular lignin structure.

Identification of multiple steady states is challenging and computationally intensive, mainly due the high dimensionality of biological networks and the inherent nonlinearity associated with the networks (Grimbs, 2010). In this study a Latin Hypercube Sampling (LHS) procedure is employed. In LHS, random initial metabolite

concentrations are sampled and used to simulate the system of Ordinary Differential Equations (ODE) to achieve a steady state concentration.

The above procedure was repeated with 10000 different initial metabolite concentrations, thus enabling us to identify all possible steady states of the system. The ranges of initial sampled concentrations were specified based on the maximum K_m values for each substrate enzyme reaction (i.e., values of metabolites ranged from 0 to 10 times the K_m values) (Wang et al, 2014). A uniform distribution and 10000 different concentrations of the metabolites were sampled to perform the simulations.

The reaction fluxes for all the reactions in the pathway were expressed in the form of Michaelis Menten kinetics (Wang et al, 2014). The rate equations developed by Song et al. (2014) that capture the interactions between Ptr4CL3, Ptr4CL5, and the Ptr4CL3/Ptr4CL5 complex were incorporated in the lignin biosynthesis pathway model from Chen et al (2014).

3.5.2 Sensitivity Analysis of Ptr4CLs on the Monolignol Biosynthetic Pathway:

Sensitivity analysis provides a measure of the influence of model parameters on the model output (Helton and Iman, 1985). We used global sensitivity analysis to quantify the overall impact of model inputs on the model output by perturbing model input parameters that have a large range of values (Sumner, 2010). In this study, variance decomposition method (Marino et al, 2008) was used. The algorithm divides

the output variance into explained and unexplained variance. The explained variance is a result of the variations in the output as a result of variations in the input parameters.

In this study, we used Ptr4CL3 and Ptr4CL5 as the factors and the steady state flux (V1 to V32) as the output. Assuming uniform distribution, 10,000 values of Ptr4CL3 and Ptrr4CL5 were randomly sampled using the LHS procedure. The range of the enzyme concentration was assumed to be ±50% of the WT concentrations. All remaining enzymes were fixed at their respective WT concentrations. The first order and total sensitivity index were calculated using *eFAST* technique, which is based on Fourier transforms (Marino et al, 2008).

3.5.3 Stability Analysis of the Monolignol Biosynthetic Pathway:

One of the important aspects of the biochemical pathway modeling is stability of the steady states resulting from the interconnected kinetic models. It has been hypothesized that metabolic stability is the key mechanism through which the biological systems maintain homeostasis. Stability is the qualitative behavior of the systems resulting from the perturbation of the system. Since the overall goal of this study is to identify the role of the 4CL enzyme complex on the monolignol biosynthetic pathway, performing stability analysis on the kinetic model would provide insights about the role of complex on the change in the stability of the system under perturbation.

The local stability of steady states can be assessed using the linearization principle (Strogatz, 1997). In the presence of multiple steady states, the Jacobian matrix

is evaluated over a range of steady state metabolite concentrations and for each steady state concentration, the Eigenvalues are calculated. In order to assess the role of the complex, the concentrations of Ptr4CL3 and Ptr4CL45 were varied from 0 to WT levels. The steady state metabolite concentrations for all 24 metabolites were calculated for each pair of Ptr4CL3 and Ptr4CL5 values and the Jacobian matrix was evaluated over the range of steady state concentrations. The Jacobian matrix was calculated numerically using the forward difference method and the Eigenvalues were calculated using the MATLAB *eig* function. More information about steady state analysis and its application on biochemical pathways can be found in Steur et al (2005), Grimbs et al (2007) and Girbig et al (2012).

Table 3.1: List of all the reactions involved in the monolignol biosynthesis pathway.

Flux	Substrate	Product	Protein
V_1	Phenylalanine	Cinnamic Acid	PAL
V_2	Cinnamic Acid	p-Coumaric acid	C4H
V_3	p-Coumaric acid	Caffeic Acid	C3H
V_4	Caffeic Acid	Ferulic Acid	COMT
V_5	Ferulic Acid	5-Hydroxy ferulic Acid	CAld5H
V_6	5-Hydroxy ferulic Acid	Sinapic Acid	COMT
V_7	p-coumaric acid	p-Coumary-CoA	4CL
V_8	Caffeic Acid	Caffeoyl-CoA	4CL
V_9	Ferulic Acid	Feruloyl-CoA	4CL
V_{10}	5-Hydroxy Ferulic Acid	5-Hydroxy feruloyl-CoA	4CL
V_{11}	Sinapic Acid	Sinapoyl-CoA	4CL
V_{12}	p-Coumaryl-CoA	p-Coumaroyl Shikimic Acid	HCT
V_{13}	p-Coumaroyl Shikimic Acid	Caffeoyl Shikimic Acid	C3H
V_{14}	Caffeoyl Shikimic Acid	Caffeoyl-CoA	HCT
V_{15}	Caffeoyl-CoA	Feruloyl-CoA	CCoAOMT
V_{16}	5-Hydroxy feruloyl-CoA	Sinapoyl-CoA	CCoAOMT
V_{17}	p-Coumaryl-CoA	p-Coumaraldehyde	CCR
V_{18}	Caffeoyl-CoA	Caffealdehyde	CCR
V_{19}	Feruloyl-CoA	Coniferaldehyde	CCR
V_{20}	5-Hydroxy ceruloyl-CoA	5-Hydroxy-coniferaldehyde	CCR
V_{21}	Sinapoyl-CoA	Sinapaldehyde	CCR
V_{22}	Caffealdehyde	Coniferaldehyde	COMT
V_{23}	Coniferaldehyde	5-Hydroxy-coniferaldehyde	CAld5H
V_{24}	5-Hydroxy-coniferaldehyde	Sinapaldehyde	COMT
V_{25}	p-Coumaraldehyde	p-Coumaryl Alcohol	CAD
V_{26}	Caffealdehyde	Caffeyl Alcohol	CAD
V_{27}	Coniferaldehyde	Coniferyl Alcohol	CAD
V_{28}	5-Hydroxy-coniferaldehyde	5-Hydroxy coniferyl Alcohol	CAD
V_{29}	Sinapaldehyde	Sinapyl Alcohol	CAD
V_{30}	Caffeyl Alcohol	Coniferyl Alcohol	COMT
V_{31}	Coniferyl Alcohol	5-Hydroxy coniferyl Alcohol	CAld5H
V_{32}	5-Hydroxy coniferyl Alcohol	Sinapyl Alcohol	COMT

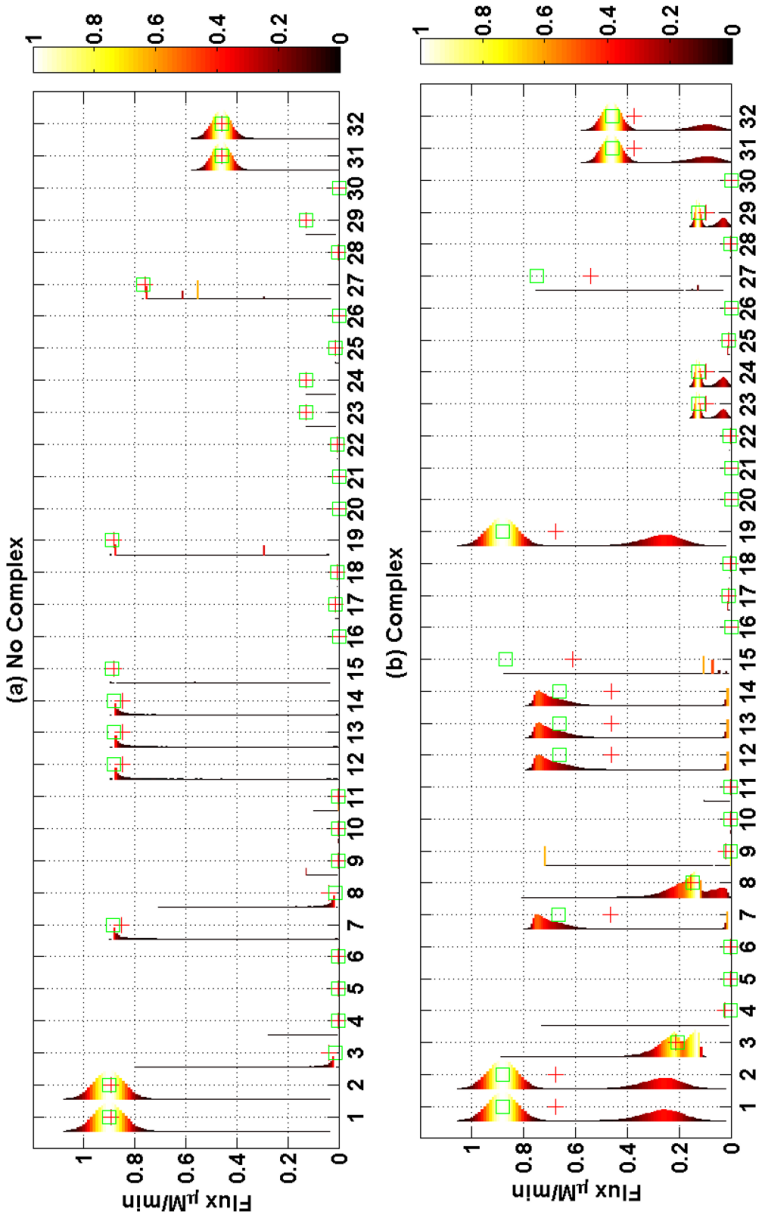

Figure A1: The steady state distribution of all the metabolic flux involved in the monolignol biosynthetic pathway. (a) The steady state flux distribution in the absence of Ptr4CL3-Ptr4CL5 complex; (b) The steady state flux distribution in the presence of Ptr4CL3-Ptr4CL5 complex. As seen in the figure, the presence of complex induces a bimodal steady state flux distributions. The green box represents the median steady state flux and the red + sign represents the mean steady state flux values

References:

Almaas E, Kovacs B, Vicsek T, Oltvai ZN and Barabási AL (2004). Global organization of metabolic fluxes in the bacterium Escherichia coli. Nature 427, 839-843.

Antonio OR (2009). Filling kinetic gaps: dynamic modeling of metabolism where detailed kinetic information is lacking. PloS one 4(3), e4967

Barton HA, Chiu WA, Setzer RW, Andersen ME, Bailer AJ, Bois FY, DeWoskin RS, Hays S, Johanson G, Jones N, Loizou G, MacPhail RC, Portier CJ, Spendiff M, Tan YM: Characterizing Uncertainty and Variability in Physiologically Based Pharmacokinetic Models: State of the Science and Needs for Research and Implementation. Toxicol Sci 2007, 99(2):395-402.

Berthet S, Demont-Caulet N, Pollet B, Bidzinski P, Cézard L, Le Bris P, Borrega N, Hervé J, Blondet E, Balzergue S, Lapierre C and Jouanin L (2011). Disruption of LACCASE4 and 17 results in tissue-specific alterations to lignification of Arabidopsis thaliana stems. Plant Cell, 23, 1124–1137.

Blower SM, Dowlatabadi H: Sensitivity and Uncertainty Analysis of Complex Models of Disease Transmission: an HIV Model, as an Example. Int Stat Rev 1994, 62(2):229-243.

Bruggeman, FJ, Westerhoff, HV (2007). The nature of systems biology. Trends in Microbiology 15 (1), 45-50.

Campbell MM and Sederoff RR: Variation in Lignin Content and Composition (Mechanisms of control and implications for the genetic improvement of plants). Plant Physiology 1996 110:3-13.

Chiang VL (2002). From rags to riches. Nat Biotechnol 20: 557-558.

Covert MW, Schilling CH, Palsson BØ (2001).Regulation of gene expression in flux balance models of metabolism. Journal of theoretical biology. 213(1):73-88.

Dixon R, Chen F, Guo D and Parvathi K (2001). The biosynthesis of monolignols: a “metabolic grid”, or independent pathways to guaiacyl and syringyl units? Phytochemistry, 57, 1069–1084.

Helton JC, Iman RL and Brown JB. 1985, Sensitivity Analysis of the Asymptotic Behavior of a Model for the Environmental Movement of Radionuclides, Ecol. Modelling. 28, 243-278.

Higuchi T (1997). Biochemistry and Molecular Biology of Wood. Springer-Verlag, Berlin-Heidelberg-New York.

Higuchi T (2003). Pathways for monolignol biosynthesis via metabolic grids: coniferyl aldehyde 5-hydroxylase, a possible key enzyme in angiosperm syringyl lignin biosynthesis. Proc. Jpn. Acad. Ser. B-Phys. Bio. Sci. 79, 227–236.

Holzhütter HG (2004). The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. Eur. J. Biochem., 271(14), 2905-22.

Humphreys JM, Hemm MR and Chapple C (1999). New routes for lignin biosynthesis defined by biochemical characterization of recombinant ferulate 5-hydroxylase, a multifunctional cytochrome P450-dependent monooxygenase. Proc. Natl Acad. Sci. U.S.A. 96, 10045–10050.

Iman RL, Davenport JM and Zeigler, DK. Latin Hypercube Sampling (A Program User's Guide): Technical Report SAND79-1473, Sandia Laboratories, Albuquerque (1980). Ingalls BP, Sauro HM: Sensitivity analysis of stoichiometric networks: an extension of metabolic control analysis to non-steady state trajectories. J Theor Biol 2003, 222:23-36.

Kitano H. 2004. Biological robustness. Nat Rev Genet, 5 826–837

Kitano H. 2007. Towards a theory of biological robustness. Mol Syst Biol, 3: 137.

Krewski D, Wang Y, Bartlett S, Krishnan K: Uncertainty, variability, and sensitivity analysis in physiological pharmacokinetic models. J Biopharm Stat 1995, 5(3):245-271. Lander, ES and Schork, NJ (1994). Genetic Dissection of Complex Traits. Science 265, 2037-2048.

Lee Y, Chen F, Gallego-Giraldo L, Dixon RA, Voit EO. 2011. Integrative analysis of transgenic alfalfa (Medicago sativa L.) suggests new metabolic control mechanisms for monolignol biosynthesis. PLOS Comput. Biol. 7: e1002047.

Lee Y, Voit EO. 2010. Mathematical modeling of monolignol biosynthesis in Populus xylem. Math. Biosci. 228: 78–89.

Lin CY, Wang JP, Li Q, Chen HC, Liu J, Loziuk P, Song J, Williams C, Muddiman DC, Sederoff RR and Chiang VL. 4-Coumaroyl and Caffeoyl Shikimic Acids Inhibit 4-Coumaric Acid: Coenzyme A Ligases and Modulate Metabolic Flux for 3-Hydroxylation in Monolignol Biosynthesis of Populus trichocarpa 2015. Molecular Plant, 8, 176-187.

Liu CJ (2012). Deciphering the enigma of lignification: precursor transport, oxidation, and the topochemistry of lignin assembly. Mol. Plant, 5, 304–317.

Lu S, Li Q, Wei H, Chang MJ, Tunlaya-Anukit, S, Kim H, Liu J, Song J, Sun YH, Yuan L, Yeh TF, Peszlen I, Ralph J, Sederoff RR and Chiang VL (2013). Ptr-miR397a is a negative regulator of laccase genes affecting lignin content in Populus trichocarpa. Proc. Natl Acad. Sci. U.S.A. 110, 10848–10853.

Miao YC and Liu CJ (2010). ATP-binding cassette-like transporters are involved in the transport of lignin precursors across plasma and vacuolar membranes. Proc. Natl Acad. Sci. U.S.A. 107, 22728–22733.

Nicholson JK, Holmes E, Lindon JC, Wilson ID (2004). The challenges of modeling mammalian biocomplexity. Nature biotechnology 22 (10), 1268-1274

Papin J, Price N, and Palsson B (2002). Extreme pathway lengths and reaction participation in genome-scale metabolic network. Genome Re, 12(12):1889 - 1900.Almaas et al., 2004;

Rabitz H, Kramer M, Dacol D: Sensitivity Analysis in Chemical Kinetics. Annu Rev Phys Chem 1983, 34:419-461.

Ragauskas AJ, Williams CK, Davison BH et al. (2006). The Path Forward for Biofuels and Biomaterials. Science 311:484-489.

Ralph J, Lundquist K, Brunow G, Lu F, Kim H, Schatz PF, Marita JM, Hatfield RD, Ralph SA, Christensen JH., Boerjan, W. (2004) Lignins: natural polymers from oxidative coupling of 4-hydroxyphenyl-propanoids. Phytochem. Rev. 3, 29–60.

Rodríguez A, Infante D (2009) Network models in the study of metabolism. Electronic Journal of Biotechnology 12: 1–19.

Salerno L, Cosentino C, Merola A, Bates DG and Amato F (2014). Robustness Model Validation of Bistability in Biomolecular Systems. A Systems Theoretic Approach to Systems and Synthetic Biology II: Analysis and Design of Cellular Systems. ISBN: 978-94-017-9046-8

Saltelli A, Marivoet J. Nonparametric Statistics in Sensitivity Analysis for Model Output - a Comparison of Selected Techniques. Reliability Engineering & System Safety. 1990;28(2):229–253.
Sarkanen KV and Ludwig CH (1971). Lignin: Occurrence, Formation, Structure and Reactions. C.H. Wiley-Interscience: New York.

Schallau K, Junker BH. 2010. Simulating plant metabolic pathways with enzyme-kinetic models. Plant Physiol. 152: 1763–1771.
Schäuble S, Stavrum AK, Puntervoll P, Schuster S (2013), Ines Heiland Effect of substrate competition in kinetic models of metabolic networks,FEBS Letters, 587(17):2818-2824

Shi R, Sun YH, Li Q, Heber S, Sederoff R. and Chiang VL (2010). Towards a systems approach for lignin biosynthesis in Populus trichocarpa: transcript abundance and specificity of the monolignol biosynthetic genes. Plant Cell Physiol. 51, 144–163.

Smith RA, Schuetz M, Roach M, Mansfield SD, Ellis, B. and Samuels, L. (2013) Neighboring parenchyma cells contribute to Arabidopsis xylem lignification, while lignification of interfascicular fibers is cell autonomous. Plant Cell, 25, 3988–3999.

Sobol I. 1993. Sensitivity analysis for non-linear mathematical models. Mathematical Modeling & Computational Experiment, 1, 407–414.
Vanholme R, Morreel K, Ralph J and Boerjan W (2008). Lignin engineering. Curr. Opin. Plant Biol. 11, 278–285.

Volcke EIP, Sbarciog M, Loccufier M, Vanrolleghem 4- COUMARIC ACID, Noldus EJL. 2007. Influence of microbial growth kinetics on steady state multiplicity and stability of a

two-step nitrification (SHARON) model. Biotechnology and Bioengineering 98(4): 882-89.

Wang JP, Naik PP, Chen HS, Shi R, Lin CY, Liu J, Shuford CM, Li Q, Sun YH, Anukit ST, Williams CM, Muddiman DC, Ducoste JJ, Sederoff RR, Chiang VL (2014). Complete Proteomic-Based Enzyme Reaction and Inhibition Kinetics Reveal How Monolignol Biosynthetic Enzyme Families Affect Metabolic Flux and Lignin in *Populus trichocarpa*. Plant Cell, 26 (3), 894-914.

Wooley, J. and Lin, H (2005) Catalyzing Inquiry at the Interface of Computing and Biology, National Academy of Science Press, Washington, D.C.

Zhao J. and Tiede C. 2011: Using a variance-based sensitivity analysis for analyzing the relation between measurements and unknown parameters of a physical model. Nonlin. Processes Geophys. 18, 269276.

Zhao Q, Nakashima J, Chen F, Yin Y, Fu C, Yun J, Shao H, Wang X, Wang, ZY and Dixon RA (2013b). LACCASE is necessary and nonredundant with PEROXIDASE for lignin polymerization during vascular development in Arabidopsis. Plant Cell, 25, 3976–3987.

CHAPTER IV

A MODULARITY BASED COMMUNITY DETECTION APPROACH TO IDENTIFY REGULATORY CO-EXPRESSION IN THE MONOLIGNOL BIOSYNTHETIC PATHWAY

Punith Naik[1], Jack Wang[2], Liu Jie[2], Hsi-Chuan Chen[2], Rui Shi[2], Christopher M. Shuford, Quanzi Li[2], Kevin Lin[2], David C. Muddiman, Ronald Sederoff[2], Vincent Chiang[2], Cranos Williams[3], and Joel Ducoste[1*]

[1] Civil, Construction, and Environmental Engineering, North Carolina State University, Raleigh, North Carolina 27695
[2] Forest Biotechnology Group, Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, North Carolina 27695
[3]Electrical and Computer Engineering, North Carolina State University, Raleigh, North Carolina 27695

Abstract:

The resistance of lignocellulosic biomass to sugar release is a key challenge in the utilization of biomass as an economically viable source of energy. The variation in lignin content and composition has been directly linked to the recalcitrant property of plant biomass. While recent advances in genomics, proteomics and metabolomics have broadened our understanding of lignin biosynthesis significantly, the regulation of monolignol biosynthesis pathway is still not fully understood. In the past decade, computational and mathematical methods have been used extensively in understanding the regulation of complex metabolic networks. In this study, we used mutual information theory and community detection methods to identify the genes that control the groups of proteins that appear to form functional modules and regulate the monolignol biosynthetic pathway. The role of these modules on lignin content and composition were quantified using the previously developed Proteomic Kinetic Mass Flux (PKMF) model and performing a Monte Carlo simulation. We hypothesize that the cells maintain a modular co-expression structure to protect from detrimental environmental stressors. The robustness of the pathway in the presence of such modular topology was quantified using stability analysis. The results from stability analysis suggests that the modular topology provides a stable architecture when the cell is subjected to external perturbations. The results further suggest that each module corresponds to a particular abiotic/biotic stress that is experienced by plant cells and that the modular structure acts as a protective mechanism to limit the effect of these specific perturbations. The analysis of the PKMF model also enables us to identify the most suitable proteins that can be targeted to engineer the lignin content and structure.

4.1 Introduction:

Lignin is a second most abundant polymer found in cells walls of all vascular plants and plays a vital role on the growth and development of plants (Dixon et al, 2001; Rogers and Campbell, 2004). Lignin is composed primarily of three main alcohol precursors or monolignols, namely *p-coumaryl*, *coniferyl*, and *sinapyl alcohols*, which later undergo polymerizations by peroxidase (PER) and laccase (LAC) to form *p-hydroxyphenyl* (H), *guaiacyl* (G) and *syringyl* (S) lignin, respectively (Weng et al, 2008). The relative percentage of each lignin unit varies with species, plant type, and developmental stages. The monolignol biosynthetic pathway has undergone several revisions over the past decade, primarily resulting from the various transgenic studies performed (Boerjan et al, 2003).

In addition to lignin, the secondary cell walls of vascular plants contain abundant polysaccharides, particularly as cellulose and hemicelluloses. Extraction of these polysaccharides from lignocellulosic biomass for biofuel production has been of great interest. However, the natural resistance of lignocellulosic biomass to enzymatic hydrolysis has been an impediment in the utilization of high energy sugars. While advances have been made toward the reduction of biomass recalcitrance, (Lee et al, 2012; Raguskas et al, 2006) rational design of less recalcitrant plant cells walls would require a deeper understanding of monolignol biosynthesis in plants, which is still lacking.

Phenotypic changes resulting from stress (abiotic or biotic) is due to the changes in levels of multiple genes and proteins that are interconnected by complex networks. Thus, the phenotypic properties cannot be determined from information generated at a single level. Rather, information at multiple levels, including gene expression, enzymatic activity, mRNA, protein, and metabolite concentrations, contribute to the phenotypic properties (Ovacik and Androulakis, 2008). Studying the interactome, which is the interactions between biological components in cells and organisms, is essential in understanding how gene function and regulation are integrated at the level of an organism. Therefore, the main goal is to integrate information from different levels and provide a holistic view of the structural and functional organization of biological systems (Zhang et al, 2008). The metabolic flux, which quantifies the rate of conversion of biochemical molecules in a metabolic network, has been identified as one of the most critical parameters in understanding the regulation of biological pathways (Stephanopoulos, 1999). The mechanisms involved in the changes of metabolic fluxes is extremely important in gaining a deeper understanding of the responses and emerging phenotypic properties of the cell.

The advances in high throughput technology has enabled us to generate large amount of genomic, proteomic and metabolomic data. These technologies have exposed new ways to use this information for analyzing how groups of genes/proteins are connected in pathways that might explain how organisms accomplish the integration on an organismal level (Phizicky and Fields, 1995; Ma and Bohnert, 2007; Raman, 2010). The protein functions can be better understood by analyzing the large-scale

interactomic datasets based on using graphical methods, in which the proteins are represented by nodes and the edges represent the interactions between the proteins. (Rivas and Fontanillo, 2010). One way of understanding the cellular function and organization is to identify a group of proteins (modules) in these networks that share more of the common functionalities (Royer et al, 2008; Navlakha et al, 2009 and Pinkert et al, 2010; Nepusz et al, 2012). Based on RNA Seq transgenic data, if two proteins are associative, then they are more likely to share the same cellular functions than proteins that are not associative. Thus, highly interconnected set of proteins within a network can be viewed as potential functional modules. Several modularity-based algorithms (Newman, 2006; Newman and Girvan, 2004) have been successfully developed and utilized to identify functional modules in Protein-Protein Interaction (PPI) networks.

One of the fundamental goals of systems biology is to link structure of biological network to its function. The approaches used to integrate regulatory interactions into metabolic network modeling are however limited. Understanding how to best formulate a joint modeling framework for metabolism and its regulatory control constitutes an important ongoing challenge (Yeang and Vingron, 2006). Although most of the proteins involved in monolignol biosynthesis have been identified, there is little information about how the proteins are co-regulated following perturbations. Knowledge about how the various proteins are co-regulated could provide valuable information about how the monolignol biosynthesis pathway is regulated under stress or in transgenics. Another important observation is that since both biological and non-biological networks are insensitive to random node removal, however, they are extremely sensitive to the

targeted removal of hubs (Jeong et al, 2001). Such hubs would also yield insightful information about the robustness of the pathway to perturbations. Analysis of the functional modules allows the understanding of both the functionality of proteins and the biological processes that involve such proteins.

In this paper, we identified protein communities resulting from the proteins involved in monolignol biosynthesis and integrated the network into a previously developed PKMF (Wang et al, 2014) model for lignin biosynthesis metabolic pathway modified to include the Ptr4CL3-Ptr4CL5 enzyme complex (Song et al, 2014) . The effect of the modules on lignin content and structure were quantified using a Monte Carlo Simulation of the PKMF model. We hypothesize that the protein modules provide robustness to the monolignol biosynthetic pathway in the presence of various perturbations. The effect of the modules on the robustness of the pathway to perturbations were quantified using stability analysis (Grimbs et al, 2010). The absolute concentrations of the proteins involved in the monolignol biosynthesis pathway were obtained from RNA Seq data. RNA Seq data has been widely used in the protein co-expression network inference. Using RNA Seq data of DNA damage response in S*accharomyces cerevisiae*, Ulitysky and Shamir (2009), were able to identify novel modules and predict novel protein functions. RNA Seq data was used to identify protein modules associated with the drought response in Rice (Zhang et al, 2012). Similarly, Ding et al (2014) used gene expression data to predict the protein- protein interaction network for sweet orange. Using the RNA Seq data, we hope to uncover the regulatory mechanisms that affect the proteins involved in the monolignol biosynthetic pathway.

4.2 Materials and Methods:

4.2.1 Mutual Information Relevance Networks:

Before identifying the communities in the network, we need to obtain a network that represents all the interactions that exist between the different proteins in the pathway. Several methods ranging from simple correlation to more complex approaches like Bayesian networks have been used to identify putative interactions, each with its own advantages and disadvantages. Correlation based approaches do not accommodate non-linear interactions, which is common in biological networks. The mutual information/relevance network approach is similar to the correlation-driven network inference methods, but measures mutual information (MI) between proteins instead of correlation to predict interactions. The mutual information between random variables X_i and X_j is defined as

$$MI(X_i, X_j) = \sum_{x_i \in X_i} \sum_{x_j \in X_j} p(x_i, x_j) \log\left(\frac{p(x_i, x_j)}{p(x_i)\, p(x_j)} \right) \qquad (4.1)$$

where p(xi, xj) is the joint probability distribution of Xi and Xj, and p(xi) and p(xj) are the marginal probabilities. Mutual information has the advantage over the correlation-based methods to identify non-linear relationships between variables but does not inform about the type of interaction such as activating or inhibitory. Mutual information between a gene pair can be due to random background effects or a regulatory relationship. Context Likelihood Relatedness (CLR) (Faith et al, 2007) extends the mutual

information approach by taking the background distribution of the mutual information values $I(X_i, X_j)$ into account. The most probable interactions are those that deviate most from the background distribution.

4.2.2 Community Detection and Modularity:

Once the network of interactions between various proteins in the pathway was obtained, the next step was to partition the nodes of the network into disjoint sets that maximizes the modularity of the network. Modularity detection has received a lot of attention over the past few years especially in biological networks (Badger and Hogue, 2003; Rives and Galitski, 2003; Newman and Girvan, 2004; Dunn et al, 2005 and Sharan et al, 2005). In general, the detection of functional modules would enhance our understanding about the organization and orientation of biological networks and also serves as the basis for other network analysis. Several approaches have been developed for community identification and summarized by Arenas et al. (2004). These approaches include (1) link removal methods; (2) agglomerative methods; (3) modularity maximization and (4) spectral methods. In this paper, we used modularity maximization to identify the different protein co-regulatory modules, primarily due to their computational speed as well as their applicability to larger networks (Newman, 2003).

Modularity (Q) is the measure of strength of division of a network into smaller sub networks or modules. Networks that are highly modular have a high degree of connectivity between the nodes belonging to the same modules, but are sparsely

connected between the nodes that belong to different modules. Modularity varies from 0 to 1, with values approaching 1 representing increased modularity found by partitioning in the network. The mathematical representation of modularity is calculated using equation (4.2).

$$Q = \frac{1}{4m} \sum_{ij} \left(A_{ij} - \frac{k_i^{in} k_j^{out}}{2m} \right) \bullet \delta_{ij} \qquad (4.2)$$

where, A_{ij} is the weighted adjacency matrix, which is a matrix indicating the weight of interaction between two nodes *i* and *j*. k_i and k_j are the degrees of nodes *i* and *j*, *m* is the total number of edges in the network, δ_{ij} is 1 if nodes *i* and *j* are in the same community and -1 if nodes belong to different communities. However, it is computationally expensive to search for partitions that would yield a maximum modularity since modularity optimization is known to be (Non-deterministic Polynomial-time) (N-P) hard (Brandes et al, 2008). Many heuristic methods were introduced to find high-modularity partitions, such approaches include greedy algorithms (Newman et al, 2004; Blondel et al, 2008) spectral methods (Newman, 2006), simulated annealing (Guimera and Amaral, 2005; Massen and Doye, 2005), sampling technique (Sales et al, 2007), and mathematical programming (Agarwal and Kempe, 2008). In this paper, we used the spectral partitioning algorithm proposed by Newman (2006) to determine the optimal number of communities and the simulations were performed in R using the package *igraph.* To ensure that the resulting community was not sensitive to a particular detection algorithm, we compared the results of 3 different community detection

algorithms (1) Leading Eigen value community; (2) Optimal modularity and (3) Fast greedy community.

To identify the protein modules, we used three different approaches that try to maximize the modularity in a given network to ensure that the resulting community is consistent despite the approaches used to calculate modularity. All three methods aim to maximize modularity by dividing a network into its disjoint sets. Because the community detection methods are stochastic (Trevino et al, 2012), there is a level of uncertainty associated with the community detection algorithms (Good et al, 2010), which might result in a different community structure for the same modularity score (Trevino et al, 2012). To verify that the resulting modular structure is robust to noise, we added noise to the experimental data set with Gaussian noise varying from 0% to 20% of the mean for each protein in the pathway obtained experimentally. The CLR algorithm was rerun for each noisy dataset and the resulting communities were obtained.

4.2.3 Monte Carlo Simulation of PKMF model:

Once the different protein co-expression modules were identified, the next step was to quantify the role of the individual modules on the lignin content and composition. This analysis was achieved by randomly sampling the concentration values for each of the enzymes from a uniform distribution belonging to a particular module. The enzyme concentrations were varied from ± 50% of their WT concentrations and the protein concentrations of the other enzymes belonging to other modules were

fixed at their respective WT concentrations. We used the Latin Hypercube Sampling procedure (LHS) (Iman et al, 1980) to sample the 10000 concentration values of different proteins in the module. The PKMF model was simulated to steady state and the resulting metabolite concentrations and pathway fluxes were calculated for the randomly selected protein concentrations.

4.3 Results and Discussion:

4.3.1 Inferring association networks from expression data

The protein interaction network was obtained using the mutual information approach as outlined in the methods section. The CLR method was used to infer association strength between the various proteins in the pathway. The CLR matrix resulting from the analysis provides pairwise related strength between the proteins and is shown in Figure 4.1. In Figure 4.1, there is a strong association between proteins belonging to the PAL family. Similarly, the edges between proteins belonging to 4CL and HCT families are higher as compared to the rest of the network. The interactions between the proteins belonging to the cytochrome P450 families namely CAld5H, C3H and C4H are dense as compared to the other proteins in the pathway. The association network showing the interactions between the proteins involved in the monolignol biosynthesis is shown in Figure 4.2. The nodes indicate the proteins and the edges indicate the interaction (association) between the proteins. Once we have an association network between all the proteins involved in the pathway, the next step is to apply the community detection algorithm to obtain disparate modules such that the

number of edges observed within the module is greater than the edges between the modules.

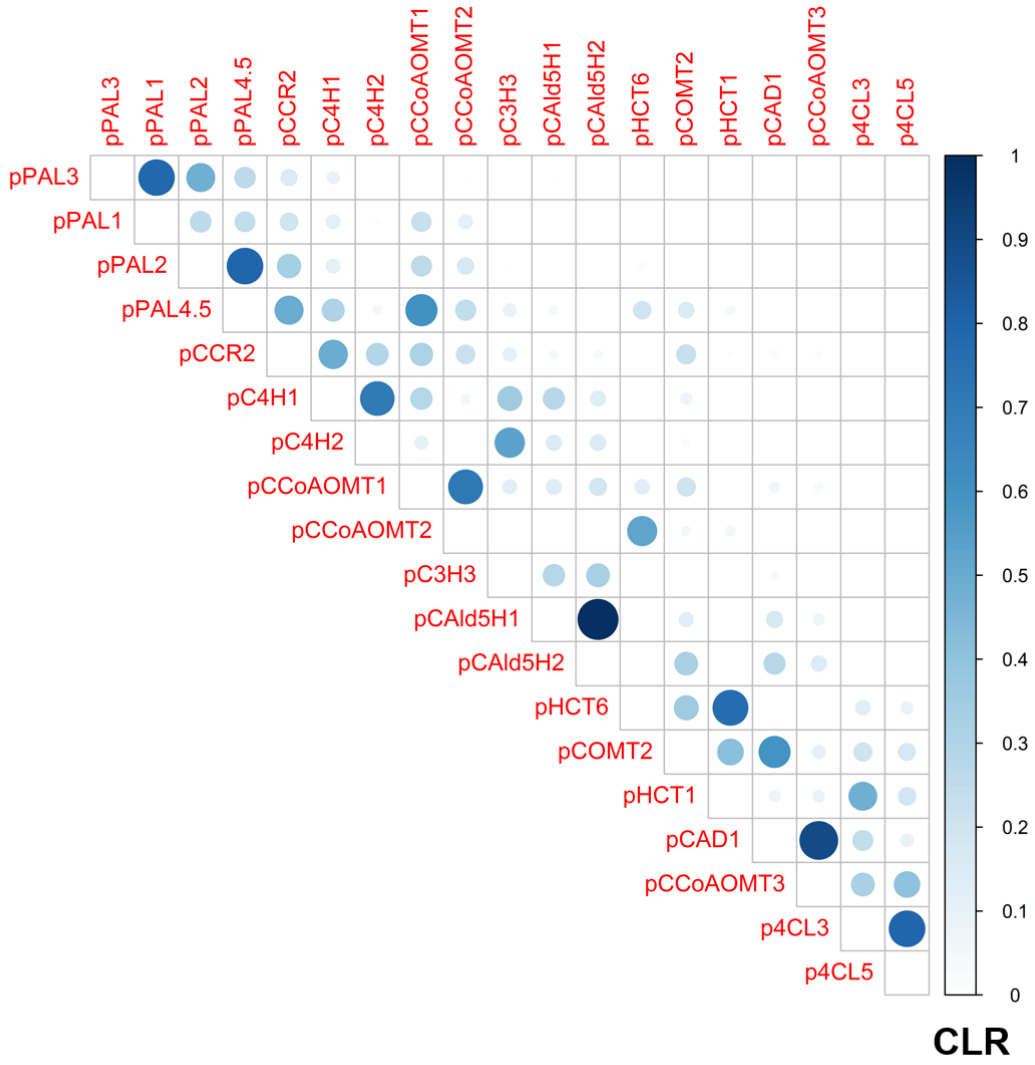

Figure 4.1: The Context Likelihood Relatedness (CLR) matrix shows the strength of association between different proteins in the pathway, and the strength ranges from 0 to 1. The size and intensity of each circle corresponds to the strength of association between the pair of proteins. Smaller circles and lighter colors indicate weak associations, and as the association between proteins approaches 1 (indicating strong association) the size of the circle increases correspondingly and the intensity of the circle also increases.

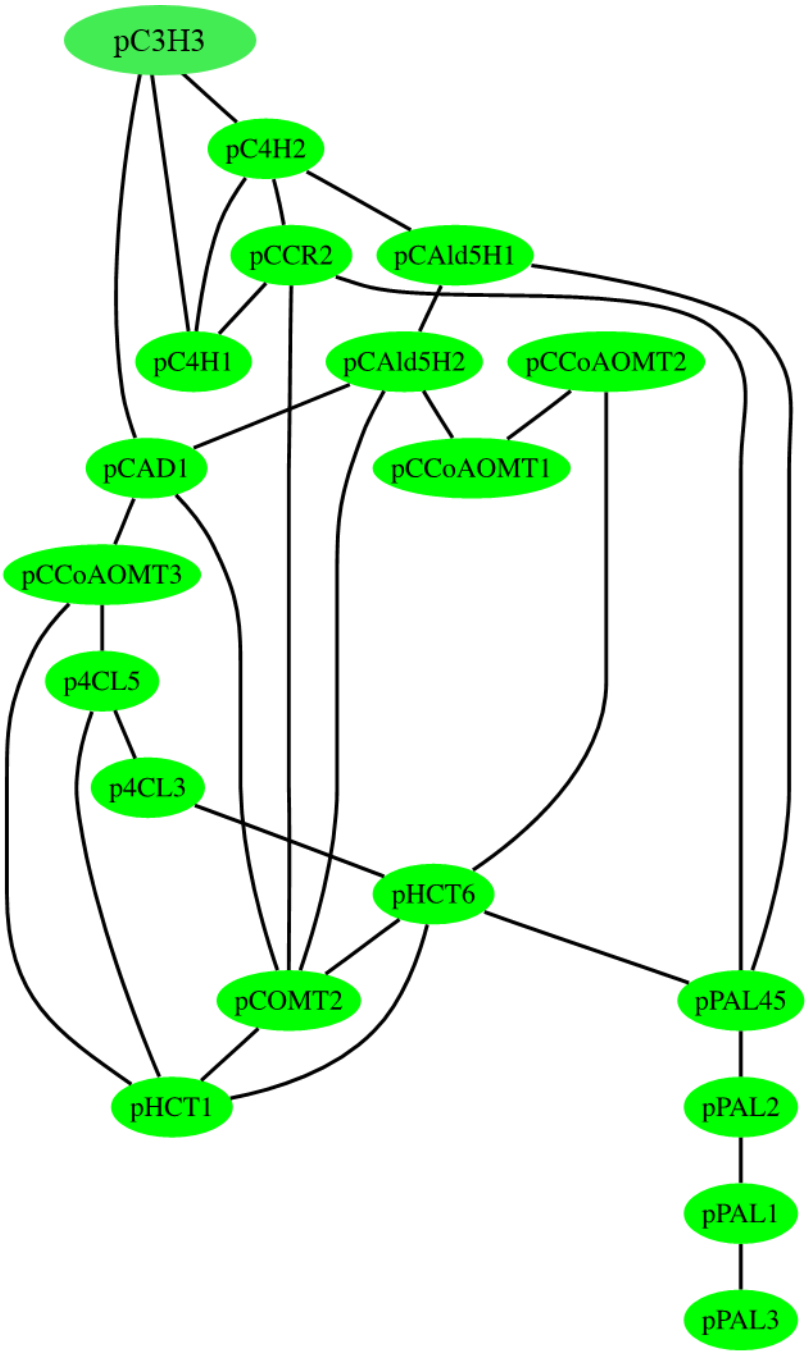

Figure 4.2: The protein interaction network of all the proteins involved in the monolignol biosynthetic pathway.

4.3.2 Robustness of community structure to experimental noise:

We added a Gaussian noise equivalent to the standard error that was reported for each measurement to assess the robustness of the community detection to noise. The community detection algorithm was then re-run to detect the community. The resulting community structure due to noise clearly had some impact on the resulting community structure. In the absence of any noise, there were 4 distinct modules present in the network (Figure 4.3). As the level of noise increased to 10%, the change in community structure lead to the merging of one community reducing the number to 3 communities (Figure 4.4). Any further increase in the noise did not result in any shift in communities or changes in number of modules. These results suggest that the community structure is robust and the three core communities were preserved despite the addition of noise.

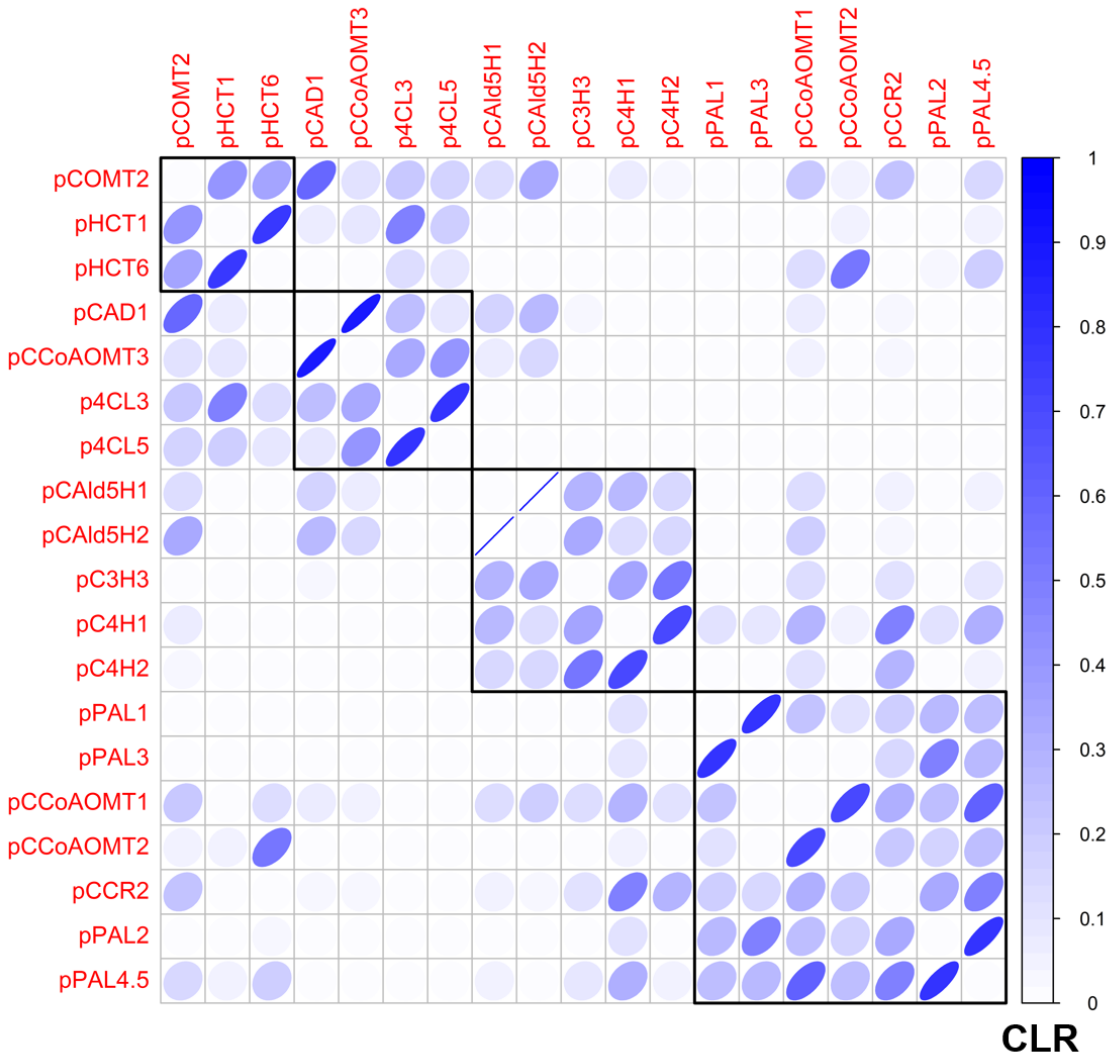

Figure 4.3: The modular structure obtained for the data without any Gaussian noise. The colorbar quantifies the strength of association in terms of Context Likelihood Relatedness (CLR).

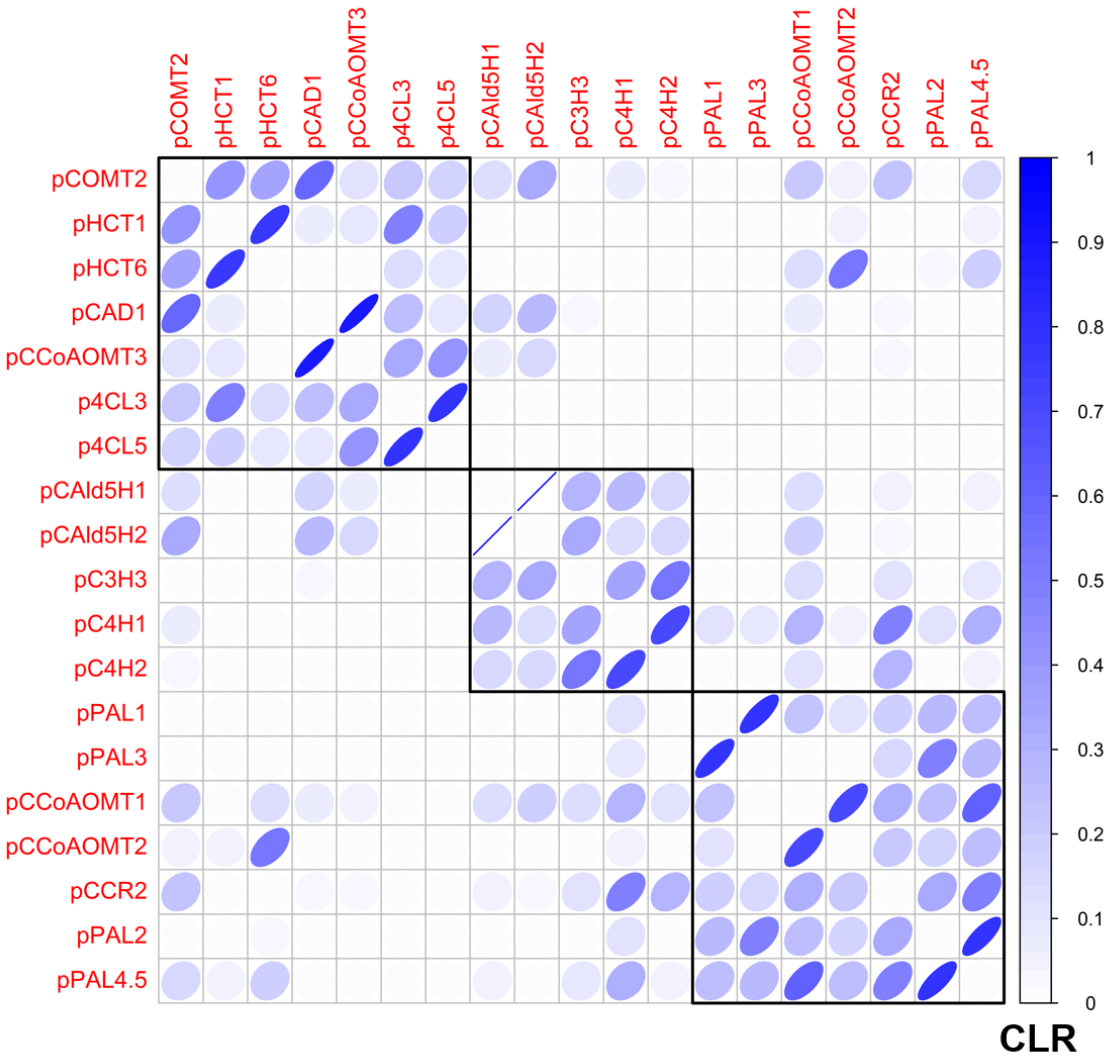

Figure 4.4: The modular structure obtained for the data in the presence of Gaussian noise. The colorbar quantifies the strength of association in terms of Context Likelihood Relatedness (CLR).

4.3.3 Role of perturbing Pt4CL3, Ptr4CL5, PtrCCoAOMT3 and PtrHCT-PtrCOMT2 on the lignin content and composition:

When the dataset was subjected to a noise level of 10% of the original dataset, the resulting community structure suggested that Ptr4CL3, Ptr4CL5 and PtrCCoAOMT3 were part of a distinct community. In order to assess the role of the perturbation of

Ptr4CL and PtrCCoAOMT3 concentrations on the lignin content and ratio, the concentrations Ptr4CL and PtrCCoAOMT3 were varied and the resulting steady state flux was then calculated. The lignin content and composition are displayed in the histogram plot below (Figure 4.5). About 95% of the steady state total lignin content and S/G ratio is distributed around the WT levels, suggesting that the changes in levels of Ptr4CL3, Ptr4CL5 and PtrCCoAOMT3 does not significantly affect the lignin content (mean = 0.9 $\mu Mmin^{-1}$) and results in a slight increase in lignin composition (mean = 2.6). Although the perturbation of Ptr4CL and PtrCCoAOMT3 individually results in a reduction in lignin content and increased S/G ratio, when both enzymes are down regulated simultaneously, they do not significantly affect the lignin content and lignin composition steady state values.

Similarly the role of PtrHCT and PtrCOMT2 on the lignin content and composition was quantified. Since there was an extra module identified when the dataset was subjected to noise, we decided to perturb the enzymes belonging to the additional module separately. The variation of the lignin content and composition as a function of the variation in protein concentrations is shown in the histogram plot in Figure 4.6. These results suggest that changes in concentrations of PtrHCT and PtrCOMT2 did not significantly alter the lignin content and composition when compared to the WT levels. Arounf 70% of the steady state lignin content values and about 70% of the S/G ratio steady state values are distributed around the WT levels. The resulting mean and median values of the lignin content as a result of perturbing PtrHCT and PtrCOMT2 are 0.8 µM/min and 0.9 µM/min, respectively. Similarly, the mean and the median values for

the lignin composition values are 3.6 and 2.0, respectively. The remaining 30% of the steady state lignin content and ratio values are a result of very low values of PtrHCT concentrations. The variation in S/G ratio as a function of the PtrHCT and PtrCOMT2 enzymes are shown in Figure 4.7 to assess the contribution of each enzyme. In Figure 4.7, a majority of the S/G ratio corresponds to the WT levels with any change in S/G ratio coming primarily from changes in the PtrHCT concentration. From these results, we can conclude that the changes in S/G ratio can be achieved by varying the concentrations of PtrHCT enzymes alone.

Although the role of perturbing multiple enzyme families on the lignin content is unknown, the effects of perturbing single enzyme families have been well studied (Wang et al, 2014). When levels of Ptr4CL enzymes are perturbed, the transgenic plants resulted in lower lignin content and an increased S/G ratio compared to the WT levels (Lee et al, 1997). On the other hand, the down-regulation of PtrCOMT2 levels in transgenic tobacco and hybrid poplar resulted in a decrease in total lignin content as well as a decrease in the S/G ratio (Dwivedi et al, 1994). Perturbation of PtrHCT concentration from wild type levels in tobacco results in a reduction in total lignin content and an increase in S/G ratio (Shadle et al, 2007).

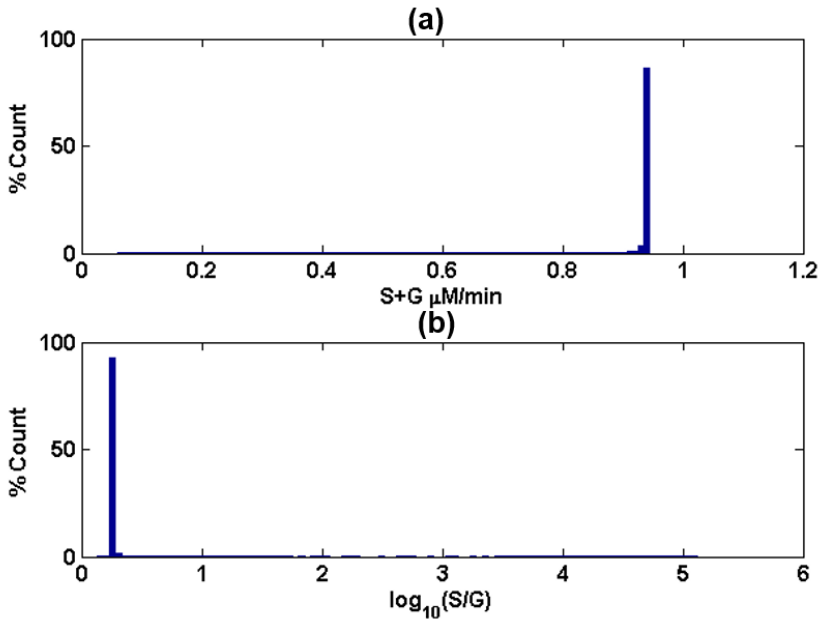

Figure 4.5: Steady state distribution of (a) total lignin content and (b) the lignin ratio (b) from 10000 runs as a result of perturbing Ptr4CL and PtrCCoAOMT3.

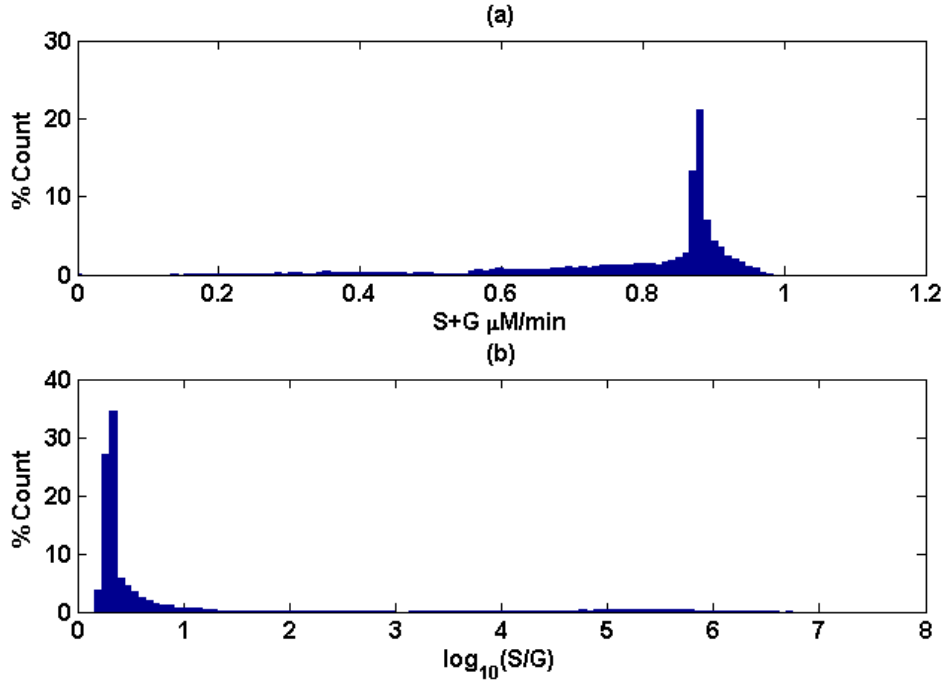

Figure 4.6: Steady state distribution of total lignin content (a) and the lignin ratio (b) from 10000 runs as a result of perturbing PtrHCT and PtrCOMT2.

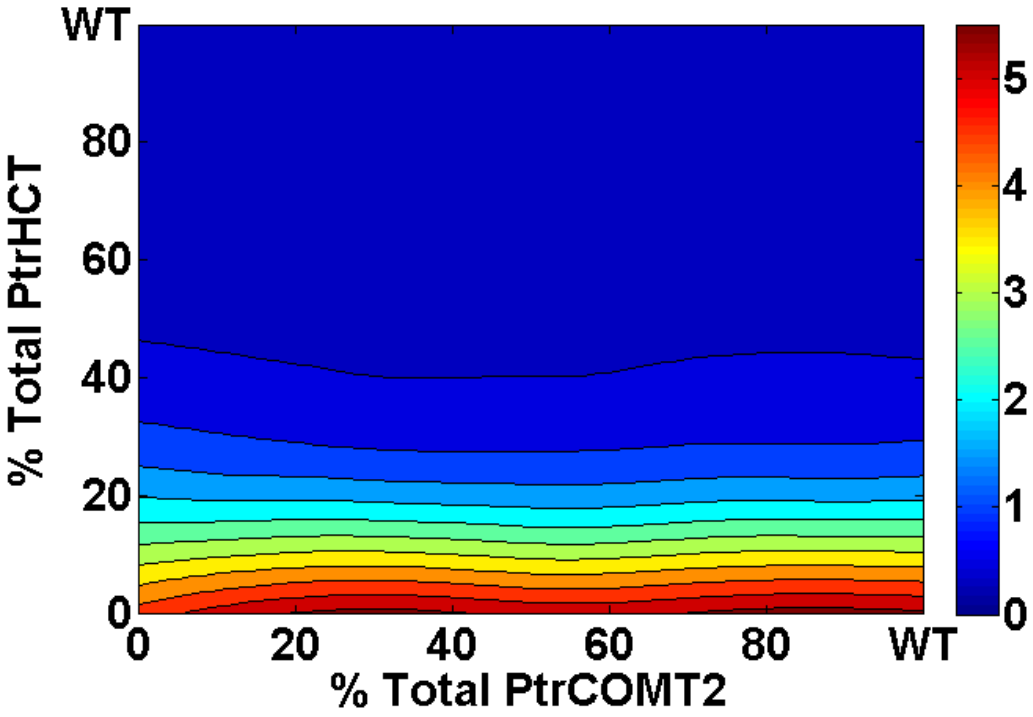

Figure 4.7: Contour plot showing the variation of S/G ratio as a function of the changes in concentrations of PtrHCT and PtrCOMT2.

4.3.4 Role of the co-expressed protein communities on lignin content:

The histogram plot of the variation of total lignin content and lignin ratio as a function of changes in the enzyme concentrations in module 1 (Ptr4CL, PtrHCT, PtrCOMT2, PtrCAD and PtrCCoAOMT3) is shown in Figure 4.8. At WT protein concentrations, the lignin content and composition distribution is spread around the peak corresponding to a high total lignin content and low lignin ratio. About 80% of the steady state total lignin content values and 85% of the steady state S/G ratios are distributed around the WT levels. The mean S/G ratio values as predicted by the PKMF model is 2.75 with a median of 2, similarly the mean lignin flux predicted by the model is 0.84 µM/min and the median is 0.9 µM/min. The spread in lignin content and distribution

suggests that a very low fraction of the steady state lignin content and S/G ratios deviate significantly from the WT levels. This deviation is a result of very low levels of protein concentrations of enzymes involved in the module.

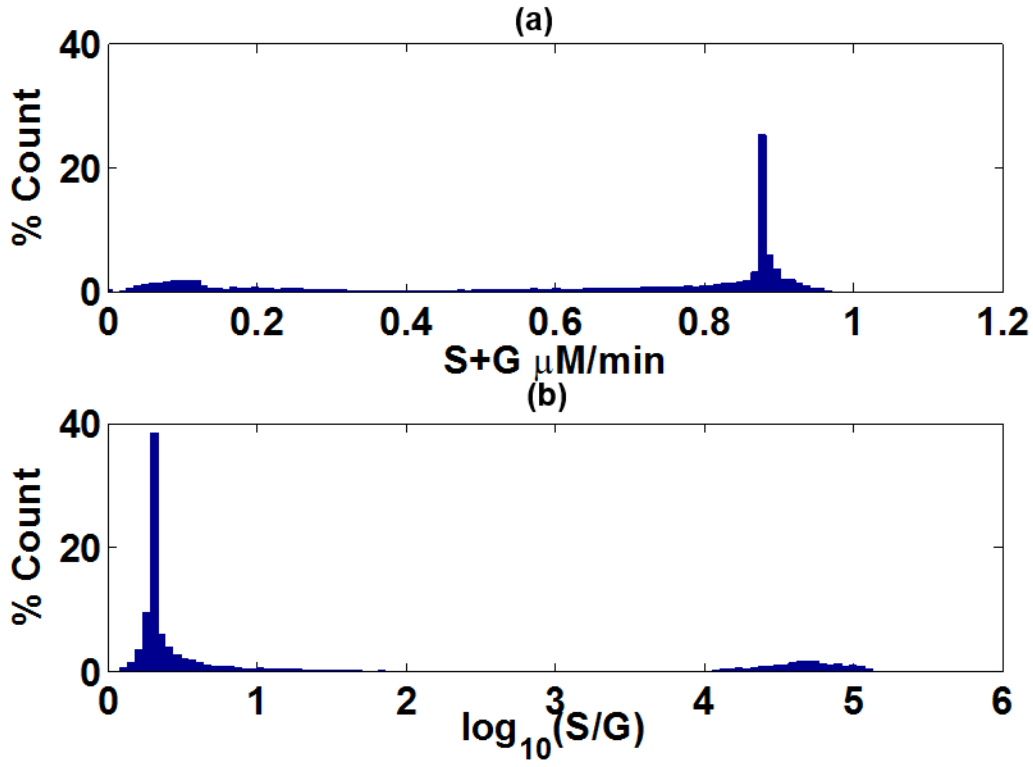

Figure 4.8: Steady state distribution of total lignin content (a) and the lignin ratio (b) from 10000 runs as a result of perturbing proteins in module 1.

The role of individual enzyme families in module 1 on the steady state S/G ratio can be visualized using a contour plot (Figure 4.9). From Figure 4.9, when concentrations of Ptr4CL and PtrHCT are plotted as a function of % WT concentration, the majority of the S/G ratio is around the WT value of 2:1. The variation in the S/G ratio in module 1 is primarily from variations in PtrHCT and Ptr4CL concentrations. As the total concentration of PtrHCT is reduced from WT, the S/G ratio varies from WT to very

high values but depends on the concentration of Ptr4CL. When the concentration of Ptr4CL is fixed at WT, the S/G ratio remains unchanged until the total concentration of PtrHCT is reduced by 60% from its WT concentration. Additional reduction in total PtrHCT concentration results in an increase in S/G ratio. Similarly, when the concentration of Ptr4CL is reduced to less than 10% of its WT concentration, an increase in S/G ratio is predicted when the total PtrHCT concentration is reduced to 60% of its WT concentration.

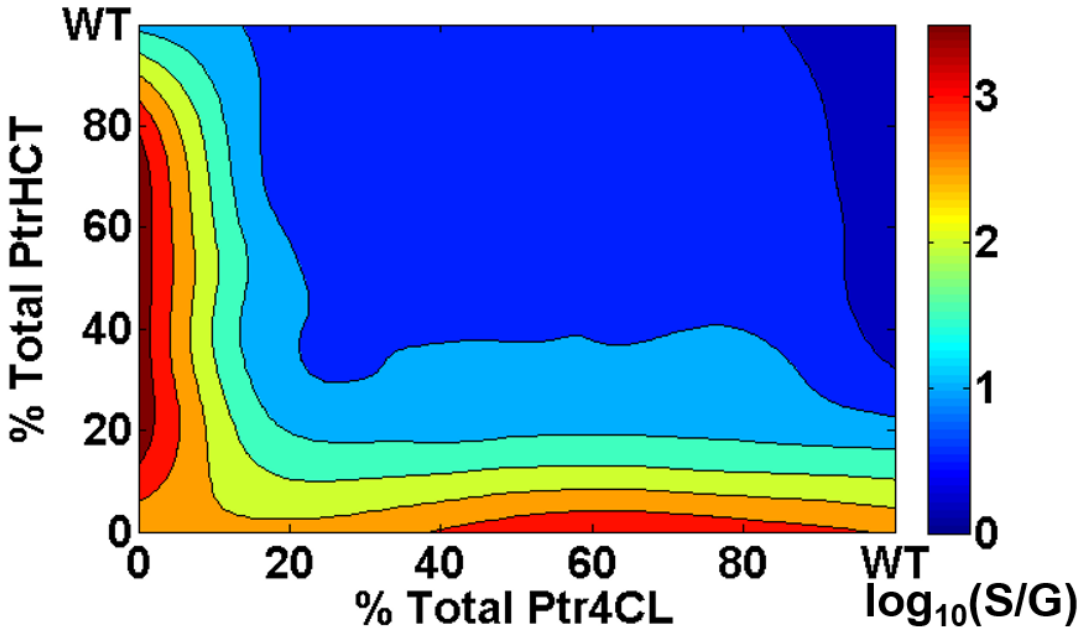

Figure 4.9: Variation of S/G ratio as a function of changes in Ptr4CL and PtrHCT concentrations in module 1.

Similarly, the variation of S/G ratio when the total concentrations of PtrHCT and PtrCOMT2 are perturbed are shown in Figure 4.10. In Figure 4.10, the variation is S/G ratio is primarily a function of the changes in the total PtrHCT concentrations. From these results, we can conclude that for module 1, the major changes in S/G ratio is due

to changes in levels of PtrHCT concentrations. A combination of PtrHCT and Ptr4CL down-regulation would be useful in fine-tuning the desired S/G ratio required in a transgenic plant.

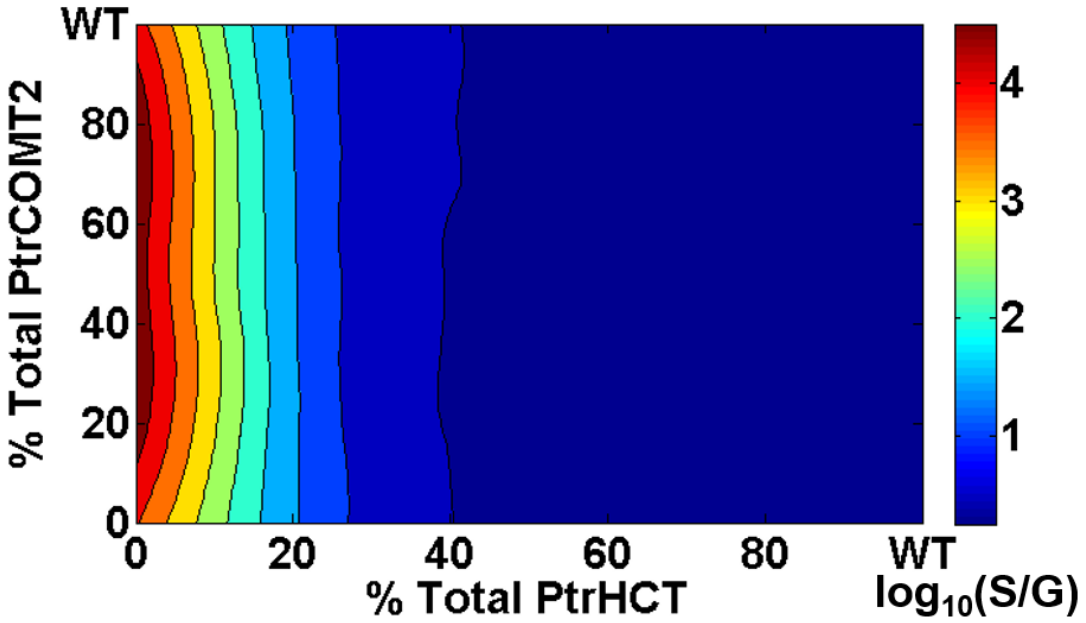

Figure 4.10: Variation of S/G ratio as a function of changes in PtrHCT and PtrCOMT2 concentrations in module 1.

The enzymes belonging to module 2 were perturbed in a similar manner described earlier. Module 2 is composed of PtrCCoAOMT1, PtrCCoAOMT2, PtrCCR2 and the entire PtrPAL family. The variation of the lignin content and composition as a result of varying all the proteins in the module is shown in Figure 4.11. More than 90 % of the distribution in lignin content and composition is distributed around the WT levels from the perturbation of module 2. The mean and the median lignin content as predicted by the model is 0.88 µM/min and 0.9 µM/min, respectively. Similarly, the mean and median S/G ratios resulting from the perturbations of all the enzymes in module 2 are

2.52 and 2, respectively. Both the PKFM model predictions and the literature suggest that when PtrPAL family is perturbed, the pathway is robust to the changes in levels of PtrPAL enzymes because of the presence of redundant PAL isoforms (Wang et al, 2014). We also observe a slight variation in lignin content and S/G ratio around the WT lignin content levels. This variation in total lignin content is due to the changes in the levels of PtrCCR2 (Figures 4.12 and 4.13). The contour plot shows that the changes in S/G ratios in module 2 is primarily a result of the changes in concentrations of PtrCCR2.

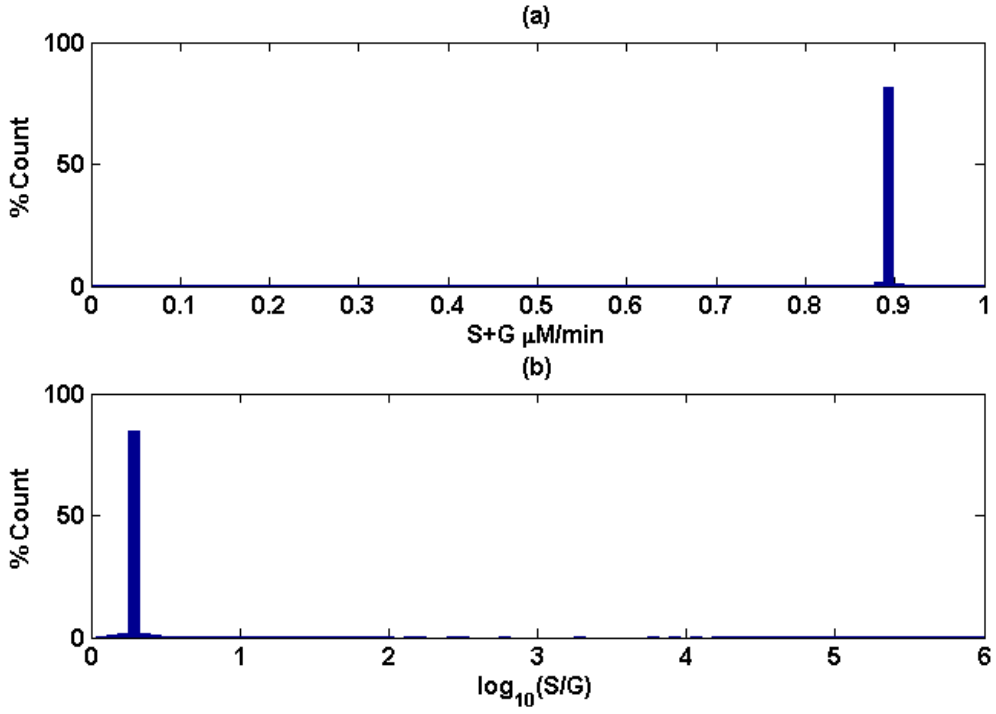

Figure 4.11: Steady state distribution of (a) total lignin content and (b) lignin ratio as a result of perturbing concentrations all the proteins involved in module 2.

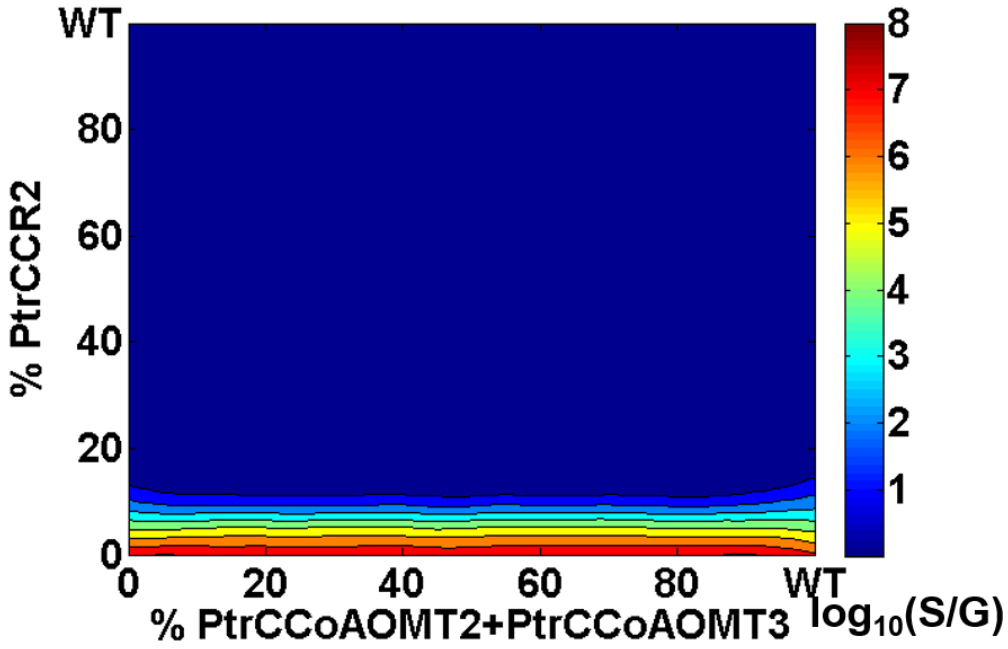

Figure 4.12: Variation of S/G ratio as a function of changes in PtrCCR2, PtrCCoAOMT2 and PtrCCoAOMT3 concentrations in module 2.

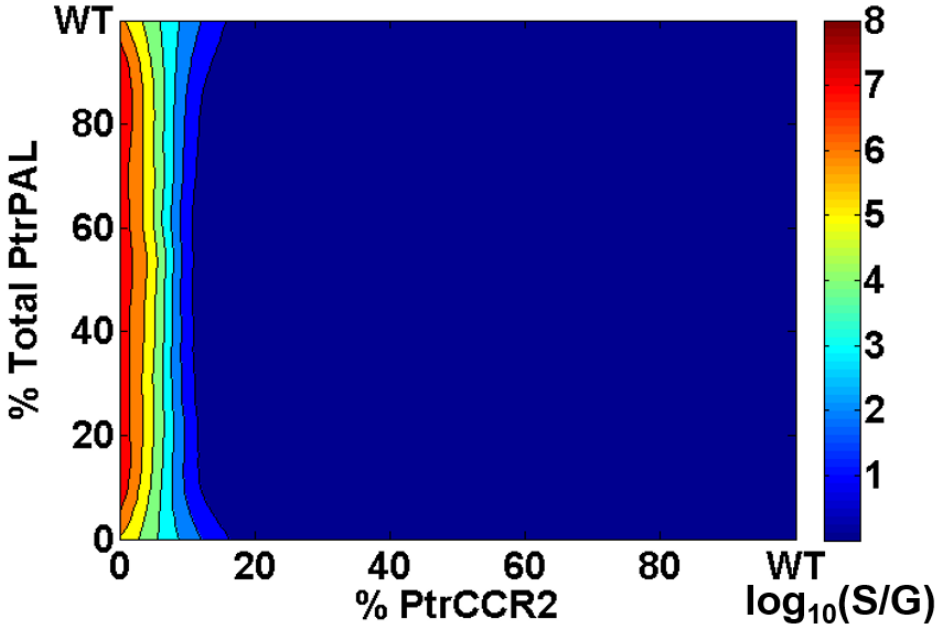

Figure 4.13: Variation of S/G ratio as a function of changes in PtrPAL and PtrCCR2 concentrations in module 2.

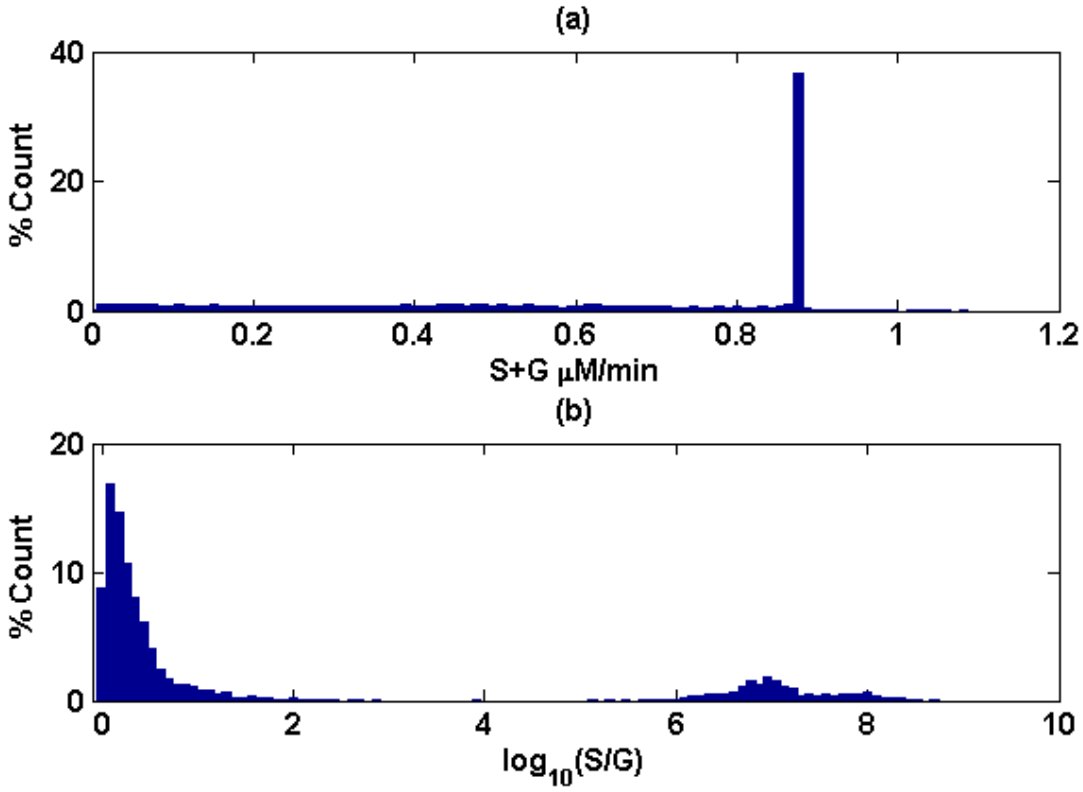

Figure 4.14: Steady state distribution of (a) total lignin content and (b) lignin ratio as a result of perturbing concentrations all the proteins involved in module 3 (C3H, C4H and CAld5H).

The third module primarily consists of PtrC3H, PtrC4H, and PtrCAld5H1 and PtrCAld5H2 enzymes. An interesting observation is that all 3 proteins belong to the cytochrome P450 family (Reddy et al, 2005). Here PtrC3H and PtrC4H primarily catalyze the initial steps of the monolignol biosynthetic pathway, while PtrCAld5H catalyzes the last step in the pathway. It has been proposed by several authors that the enzymes C3H, C4H and CAld5H are rate limiting and any possible changes in these enzyme concentrations would lead to the regulation of not only the genes involved in monolignol biosynthesis as well as the lignin composition and structure, but also of their rates of polymerization. The results from perturbing the cytochrome P450 enzymes

suggest that the reduction in levels of those enzymes lead to a significant variation in the S/G ratio without a significant change in total lignin content (Figure 4.14). The changes in S/G ratio is mainly a result of PtrCAld5H enzymes. The down-regulation of PtrCAld5H leaves the total lignin content unaffected but the concentration of S subunit varies over a wide range as seen in Figure 4.15. The wide spread in the total lignin content is primarily from PtrCAld5H involvement in the conversion of coniferaldehyde and coniferyl alcohol to 5-hydroxy coniferaldehyde and 5-hydroxy coniferyl alcohol, which are the last steps in the monolignol biosynthesis pathway.

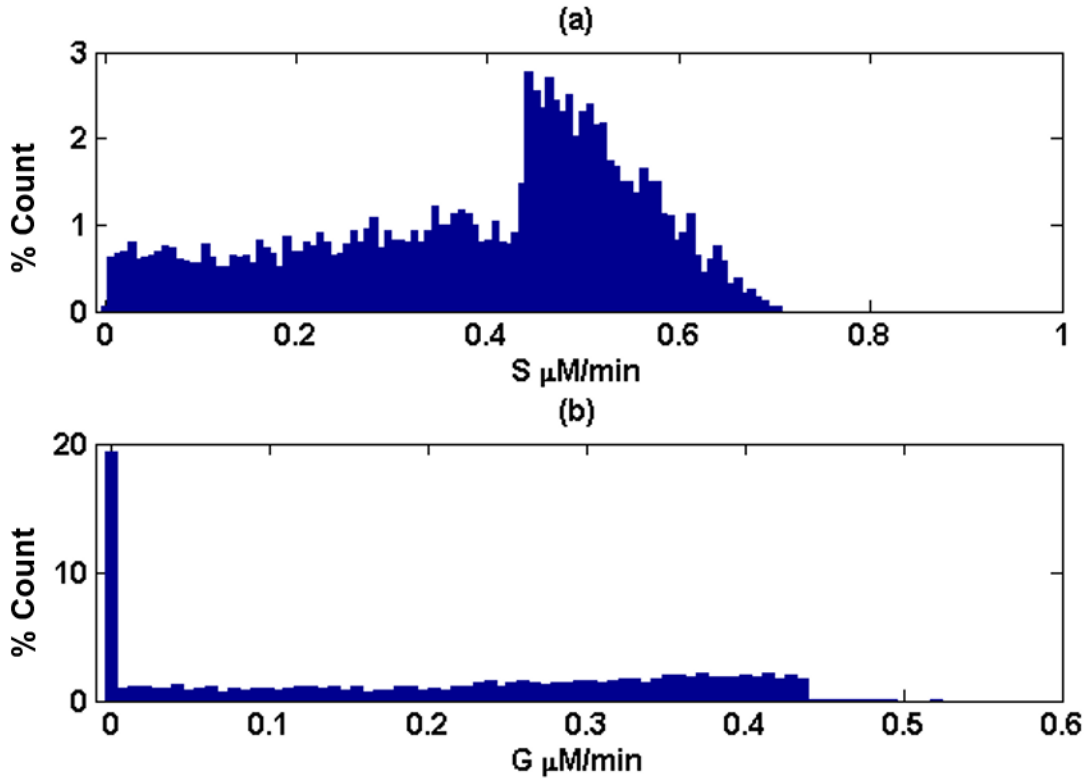

Figure 4.15: Steady state distribution of (a) S subunit (b) G subunit resulting from variation in concentrations of all the proteins (C3H, C4H and CAld5H) involved in module 3.

The variation of S/G ratio as a function of enzymes in module 3 is shown as a contour plot in Figure 4.16. As shown in Figure 4.16 under WT concentrations of PtrCAld5H, PtrC3H and PtrC4H, the steady state S/G corresponds to the ratio 2:1. When the concentration of PtrCAld5H is reduced from the WT and the concentrations of other enzymes are fixed at WT, an increase in S/G ratio is seen for very low levels of PtrCAld5H, which is consistent with the experimental results. As the concentration of PtrC3H and PtrC4H is decreased, we observe an increase in S/G ratio with little to no change in PtrCAld5H concentration. When all the concentrations of the enzymes in the module are decreased, an increase in the S/G ratio is observed. The increase in S/G is primarily due to the decrease in levels of PtrC3H and PtrC4H.

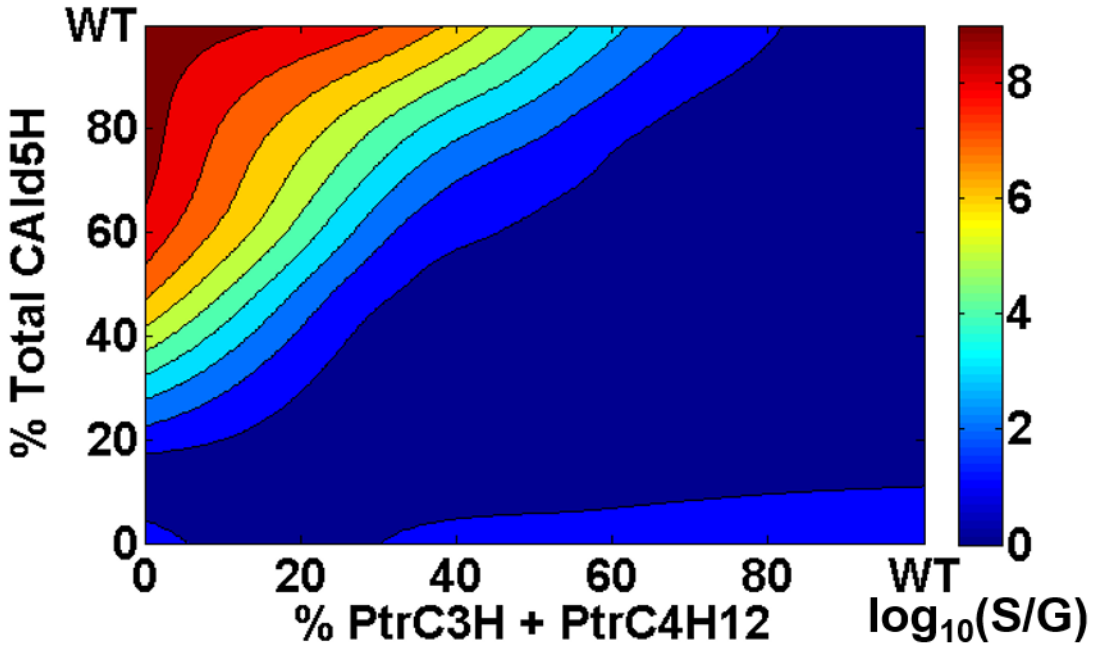

Figure 4.16: Steady state distribution of total lignin content (a) and the lignin ratio (b) from 10000 runs as a result of perturbing proteins in module 3.

From the simulations reported above, the results suggest that module 2 (PtrPAL, PtrCCoAOMT2, PtrCCoAOMT3, CCR2) is most resilient to transgenic perturbations with respect to the lignin content and structure. Transgenic perturbations of module 3, on the other hand, results in large variations in lignin content and structure. When the enzyme concentrations in module 1 are varied, the majority of changes in S/G ratio is a result of changes in PtrHCT levels. In order to modify the lignin composition without altering the lignin content, the most probable targets should be PtrHCT and Ptr4CL enzymes which are part of Module 1. The perturbation results of the PKMF model suggest that the lignin content and structure can be tailored for specific needs by targeting a set of enzymes in the pathway as opposed to targeting an individual enzyme.

4.3.5 Importance of Protein Community Structure on Biosynthetic Pathways

In the previous section, we showed how the information about the protein community structure provides resiliency to changes in the lignin content and composition from their perturbation. However, the most important question that needs to be answered is why do cells prefer such community structure/topology and what are the evolutionary advantages of this particular topology? A previous study conducted by Moura et al (2010) suggest that the modular structure of monolignol biosynthesis enzymes acts as a self-defense mechanism in which only a few enzymes are affected. Thus, a modular enzyme structure based on co-expression enables cells to perform the cellular functions and also acts as a mechanism that provides robustness against perturbations.

To begin our understanding of the role of modules on regulatory robustness, let's assume that all the proteins in the pathway are a part of a single highly connected network (i.e., no modularity as suggested in the previous section). In order to quantify the effect of changing the protein concentrations on the S/G ratio, we performed a simulation by simultaneously changing the concentrations of all the enzymes in the pathway. All 21 enzyme concentrations in the pathway were sampled between 0 to WT levels and the resulting lignin content and composition is shown in Figure 4.17. The results show a high degree of variability in lignin content and composition and suggest that in the absence of a modular framework, the pathway would be unable to withstand large perturbations eventually leading to an un-survivable condition or a plant that is not resilient to environmental stressors.

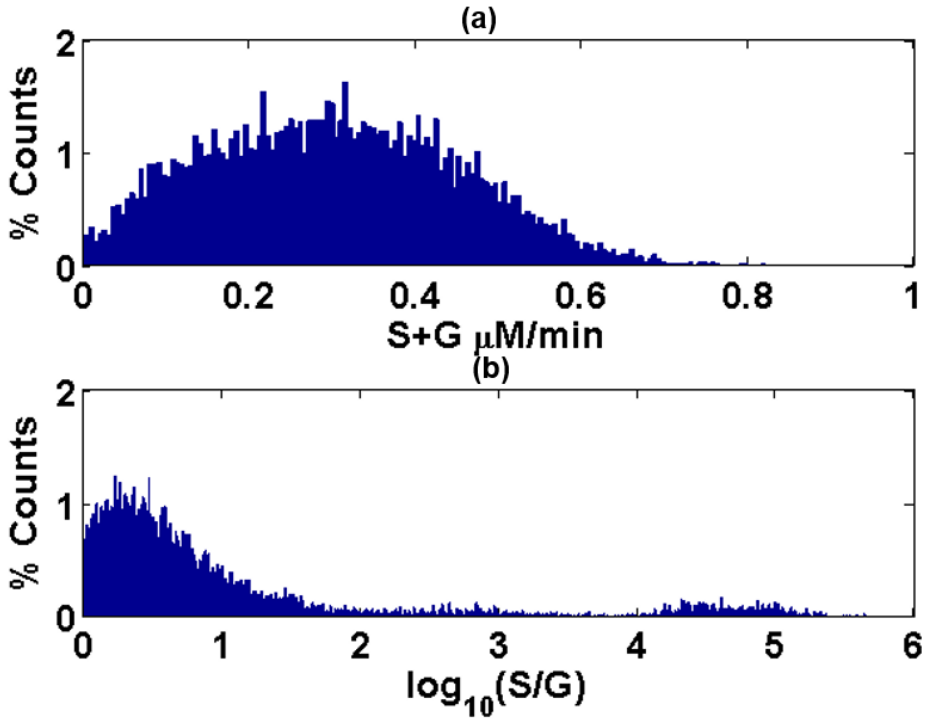

Figure 4.17: Steady state distribution of (a) total lignin content and (b) lignin ratio as a result of perturbing concentrations all the proteins involved in the monolignol biosynthetic pathway.

4.3.6 Role of perturbing individual enzyme at a module level on lignin composition and structure:

In the previous section all the concentrations of all the enzymes belonging to a particular module were perturbed and their effect on the lignin composition and structure were quantified. However, under that assumption the associative strength between the proteins were not taken into consideration, rather the proteins concentration can take any random value irrespective of the concentration of the other proteins in the module. In this section we perturbed concentration of individual protein and the concentrations of other proteins are varied based on the associative strength between two pair of proteins. The concentrations of the proteins are perturbed by ± 50% of their WT concentrations and 10000 concentration values were randomly sampled from a uniform distribution. The results from the simulations are summarized in the Table 4.1 and 4.2, which shows the targeted protein in each community and the corresponding mean, median and standard deviations of the simulates lignin composition and lignin content distributions respectively. From the results it can be seen that when the proteins in modules 1 and 2 are perturbed, a small variation in the lignin content and composition are observed as indicated by the standard deviation values. However, when the proteins in module 3 are perturbed, a large variation in the lignin composition is observed. Suggesting that for the plants to maintain resiliency, the proteins in module 3 should remain unaffected.

Table 4.1: Distribution summary of the variation in lignin composition (S/G) when individual enzymes in the community are perturbed.

Community	Target	Mean	Median	Std. dev
Community 1	Ptr4CL3	1.94E+00	1.90E+00	3.36E-01
	Ptr4CL5	2.01E+00	2.00E+00	3.56E-01
	PtrHCT1	2.00E+00	1.98E+00	3.51E-01
	PtrHCT6	1.96E+00	1.92E+00	3.40E-01
	PtrCAD1	1.87E+00	1.81E+00	3.17E-01
	PtrCCoAOMT3	1.93E+00	1.90E+00	3.35E-01
	PtrCOMT2	1.91E+00	1.90E+00	3.21E-01
Community 2	PtrCCoAOMT1	1.88E+00	1.85E+00	4.26E-01
	PtrCCoAOMT2	1.88E+00	1.85E+00	4.66E-01
	PtrCCR2	1.88E+00	1.85E+00	4.27E-01
	PtrPAL1	1.90E+00	1.85E+00	5.02E-01
	PtrPAL2	1.89E+00	1.86E+00	3.96E-01
	PtrPAL3	1.88E+00	1.86E+00	2.75E-01
	PtrPAL45	1.87E+00	1.85E+00	2.41E-01
Community 3	PtrC3H3	1.40E+00	1.06E+00	1.18E+00
	PtrC4H1	1.47E+00	1.09E+00	1.29E+00
	PtrC4H2	1.31E+00	1.01E+00	1.10E+00
	PtrCAld5H1	2.45E+00	1.80E+00	1.96E+00
	PtrCAld5H2	2.50E+00	1.82E+00	2.05E+00
	PtrCAD2	6.69E-01	5.89E-01	4.41E-01

Table 4.2: Distribution summary of the variation in lignin content (S+G) when individual enzymes in the community are perturbed.

Community	Target	Mean	Median	Std. Dev
Community 1	Ptr4CL3	9.04E-01	9.06E-01	8.32E-02
	Ptr4CL5	9.04E-01	9.06E-01	8.75E-02
	PtrHCT1	9.04E-01	9.06E-01	8.64E-02
	PtrHCT6	9.05E-01	9.06E-01	8.25E-02
	PtrCAD1	9.04E-01	9.06E-01	7.78E-02
	PtrCCoAOMT3	9.04E-01	9.06E-01	8.09E-02
	PtrCOMT2	9.04E-01	9.06E-01	8.00E-02
Community 2	PtrCCoAOMT1	9.04E-01	9.06E-01	1.84E-02
	PtrCCoAOMT2	9.04E-01	9.06E-01	1.79E-02
	PtrCCR2	9.03E-01	9.05E-01	1.73E-02
	PtrPAL1	9.03E-01	9.06E-01	2.18E-02
	PtrPAL2	9.03E-01	9.05E-01	3.32E-02
	PtrPAL3	9.03E-01	9.05E-01	2.12E-02
	PtrPAL45	9.04E-01	9.06E-01	1.76E-02
Community 3	PtrC3H3	9.04E-01	9.06E-01	2.94E-01
	PtrC4H1	9.04E-01	9.06E-01	3.05E-01
	PtrC4H2	9.04E-01	9.06E-01	2.88E-01
	PtrCAld5H1	9.04E-01	9.06E-01	3.24E-01
	PtrCAld5H2	9.04E-01	9.06E-01	3.29E-01
	PtrCAD2	9.01E-01	9.06E-01	2.22E-01

4.3.7 Stability Analysis of protein co-expression modules:

An alternate view of robustness displayed by these modular configured lignin biosynthesis proteins can be performed using stability analysis (Girbig et al, 2014). The enzyme protein concentrations belonging to each module was perturbed by +/- 50% of their respective WT concentrations to assess the stability of the modules. The steady state concentrations of all the 24 metabolites were calculated corresponding to the enzyme concentrations. The stability of the system over the range of enzyme concentrations were assessed based on the Eigenvalues of all the metabolites in the system.

The variation of Eigenvalues for different proteins are displayed in the histogram plot (Figure 4.18). The protein concentrations of all the enzymes in the three modules were perturbed simultaneously from their WT concentrations to assess the effect of the plant cell developing a modular co-expression framework for the lignin biosynthesis pathway. In Figures 4.18.a – 4.18.c, the majority of the Eigenvalues are negative (~90%), suggesting that the modular structure confers stability to the pathway as compared to the scenario when all the proteins in the pathway are perturbed (Figure 18.d) (~50%). These results further support the hypothesis that, the presence of modules in the pathway provides a mechanism through which the cells maintain robustness with respect to lignin structure.

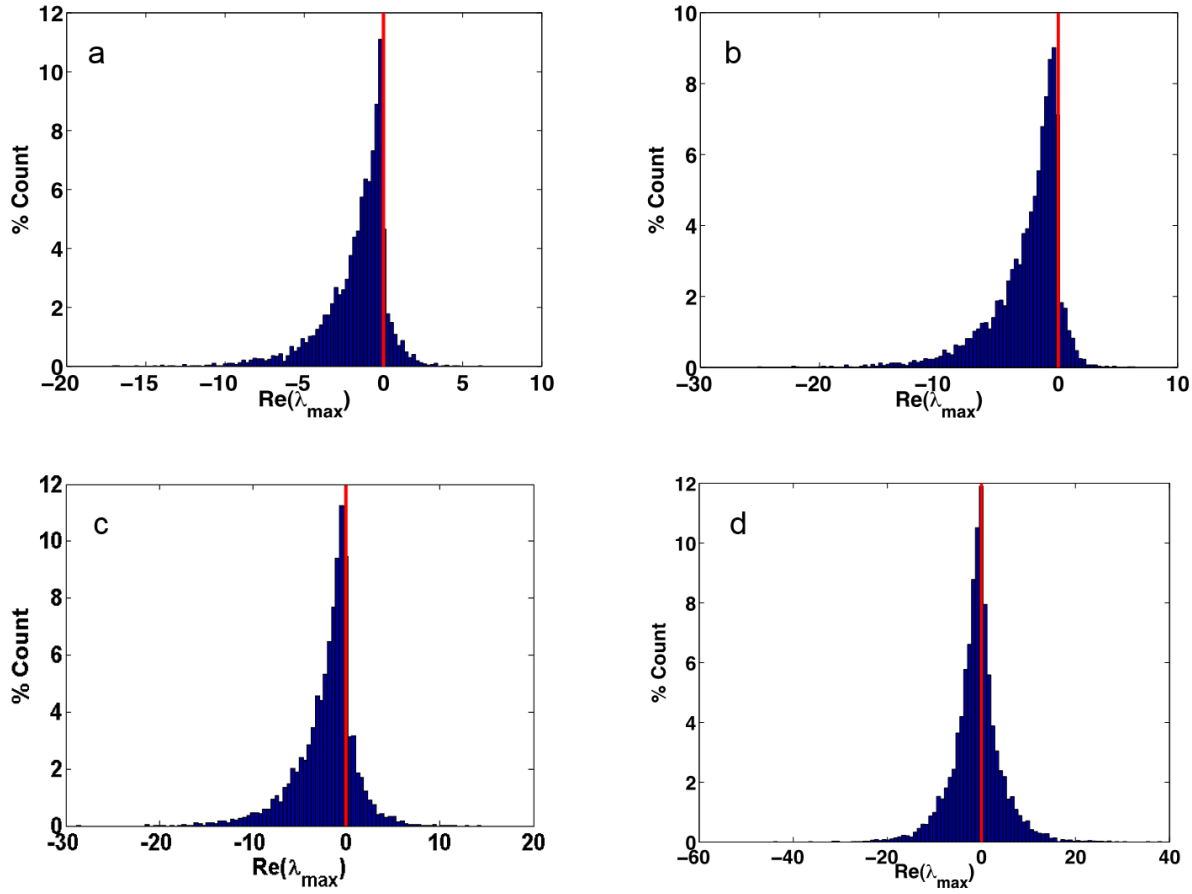

Figure 4.18: Eigen value distribution for different topology of proteins in the monolignol biosynthetic pathway. (a) Module 1 (b) Module 2 (c) Module 3 (d) all proteins concentrations were varied.

4.3.7 The role of modules on the robustness of the pathway in the presence of stress:

Complex cellular processes are a result of functional modules working together in unison (Ravasz et al, 2002; Hartwell et al, 1999). One important property of networks is their ability to maintain the overall network topology despite the loss of individual components (Barbarasi, 2004). The modular organization makes a system relatively robust to perturbations that target individual components. By performing stability analysis, we were able to establish that one of the primary reasons why the regulatory

control that result in protein communities is to maintain robustness to perturbations. The reason why proteins may form co-expression communities is that, the community architecture serves as a protective mechanism through which the pathway is able to function despite some of the communities being adversely affected.

During their entire life cycle, plants are faced by a myriad of environmental stresses. Two major categories of environmental stress are namely the abiotic stress, which corresponds to environmental conditions, such as drought, submergence, salinity, and biotic stress which is caused by infectious living organisms, such as bacteria and viruses. Both biotic and abiotic stress negatively affect the productivity and survival of plants (Shaik and Ramakrishna, 2014).

Suppressing single and multiple genes involved in lignin biosynthesis leads to an accumulation of secondary metabolites (Wang et al, 2014). Many of these compounds play a vital role in plant-environment interactions and altering their biosynthesis can influence the response of plants to such stresses (Baxter and Stewart, 2013). Additionally, transcriptomic and proteomic studies have indicated that lignin modifications can trigger an increase in expression of stress response genes (Gallego et al, 2011).

Since the lignin biosynthesis pathway is robust to moderate perturbation, we hypothesize that the modules act as a protective mechanism by which each module reacts to a particular type of stress, thus leaving the other modules unaffected. This

modularity enables the plant to absorb stress and still continue to function. Our hypothesis is supported by recent findings that report the effect of abiotic and biotic stresses on the various genes involved in the monolignol biosynthetic pathway.

In a recent study, it was observed that under iron depravation, an increase in CCoAOMT1, 4CL and HCT gene expression occurred in Arabidopsis (Teeples et al, 2015). In rice, proteomic studies suggested that drought stress resulted in an increased amount of PAL levels (Pandey et al, 2010) while in Grey-maranta (*Ctenanthe setosa*), there was a16-fold increase in PAL activity as a result of drought stress (Terzi et al, 2013). Drought stress also resulted in an increase in level of genes encoding *CCoAOMT* (Yoshimura, 2007).

Cold stress results in changes gene expression and plant metabolism which ultimately effects the biological functions (Stitt and Hurry, 2002). Under lower temperatures, the activity of PAL enzyme was higher in *Brassica napus (*Domon et al, 2013). In leaves, the accumulation rate of phenolic compounds in leaves depended on the temperatures the plant was subjected to (Solecka and Acperska, 2003). The expression levels of the genes involved in monolignol biosynthesis such as PAL, CCR, and CCoAOMT were higher in pea in response to cold (Zhu et al, 2013; Badowiec et al, 2013).

Salt stress has been shown to impact secondary cell wall formation and structure, as revealed by an altered lignin biosynthesis. Prior studies have shown that

the expression levels of PAL and CCoAOMT in response to salt (Salekdeh et al, 2002; Zhao et al, 2013). The inclusion of heavy metals in soil and the contamination of water with heavy metals have detrimental effects on plants (D'Emilio et al, 2012; Kelepertziz et al, 2014). The activities of PAL and CAD enzymes were altered when *Panax ginseng* was exposed to a higher level of copper (Cu). The change in enzyme levels was correlated with an accumulation of lignin (Ali et al, 2006). Similarly, the genes encoding *4CL* and *COMT* in *Medicago truncatula* where were up-regulated in response to Al stress in root tips (Chandran et al, 2008).

Although light, which contains Ultra Violet (UV) is essential for plants to ensure normal growth, excess light can cause cellular damage. Plants have therefore developed mechanisms to counter the effect of light through changes in cell wall. The effect of light was countered through pigmentation, regulating levels of antioxidant, metabolites, and enzymes. Studies have shown that the lignin content in several plants were altered depending on the intensities of light. As an example, the lignin content in the seedlings of *Ebenus cretica* L. grown in presence of light doubled when compared to the lignin content of the seedlings grown in the shade (Syros et al, 2005). A Similar result was observed in leaves of coffee exposed to sunlight, where the lignin content in leaves was higher when compared to the leaves that were grown in the shade (Cabane et al, 2013). The activity of PAL enzymes were induced in the leaves of orchid plants when it was exposed to different light intensities (Ali et al, 2005). In Arabidopsis, expression of 7000 genes were altered when plants were subjected to high light intensity. Several genes involved in the lignin biosynthetic pathway had shown an

increase in expression levels (Kimura et al, 2003).

Ozone (O_3) is one of consequences of air pollution and the plant cell walls have been shown to be an early target of ozone (Gunthardt-Goerg et al, 1997). O_3 induces leaf and root injuries, resulting in biomass reduction, as observed in tobacco, poplar and birch (Racherla et al, 2008; Paoletti et al, 2010). In Manna ash (*Fraxinus ornus* L.), one of the main structural impacts of O_3 is a puncture within the cell wall, as observed on leaf mesophyll cells (Paoletti et al, 2010). In poplar, ozone damage results in the reduction in stem diameter and the structural properties of wood (Richet N, 2011). O_3 affects the lignin biosynthesis process due to the reduced cambial growth and xylem differentiation, which in turn results in a reduction in cellulose and lignin (Richet et al, 2011). Changes in lignin structure were seen in poplar trees exposed to O_3, the lignin structure contained higher proportion of H-units that were higher than the WT levels (Cabane M, 2004). The enzyme activity of CAD was strongly increased during the developmental stage, and the activity of PAL was increased in old and mid-aged leaves (Cabane M, 2004). In non-woody plants, the transcript level of *PAL* mRNA increased 3-fold compared to control Arabidopsis or parsley plants (Sharma et al, 1994; Eckey-Kaltenbach et al, 1994). O_3 was also found to modify cell wall components by de-polymerization of lignin, which released small phenolic compounds (Wiese et al, 2000).

From this literature, a summary of various stresses affecting the protein co-expression modules can be developed (Figure 4.19). An interesting observation from this analyses is that the proteins belonging to module 2 are the ones that are most affected by abiotic and biotic stress. At the same time, module 2 is most resilient to enzymatic perturbations as shown by earlier simulations. Similarly, module 3 is least

affected by the abiotic/biotic stress suggesting that the enzymes belonging to module 3 were shielded because these enzymes significantly impact lignin content and structure. As a result, the enzymes belonging to module 3 may have evolved after the divergence of angiosperms and gymnosperms compared to other enzymes in the monolignol biosynthetic pathway (Xu et al, 2009).

The results from the Monte Carlo Simulation of the PKMF model revealed that except for module 3, the majority of the steady state S/G ratios are distributed around the WT S/G levels when the protein concentrations of enzymes were perturbed. In the presence of perturbations, the pathway was able to maintain the overall function despite the changes in levels of some of the enzymes involved in the monolignol biosynthesis. It can be concluded that the primary reason why plants produced a regulatory mechanism that results in enzymes forming modules is simply homeostatic. In other words, regulatory modularity was needed as a defense mechanism to provide resiliency against perturbations. The data from literature supports this hypothesis, where different modules correspond can be induced by different kind of stressors. Modules 1 and 2 can be influenced by a majority of stressors, while module 3 can be influenced by few stressors, thus enabling the cells to maintain function. The results from stability analysis also suggest that when a few proteins belonging to a particular module are perturbed, the system is robust to perturbations and when all the proteins are perturbed in the pathway, the system loses its stability.

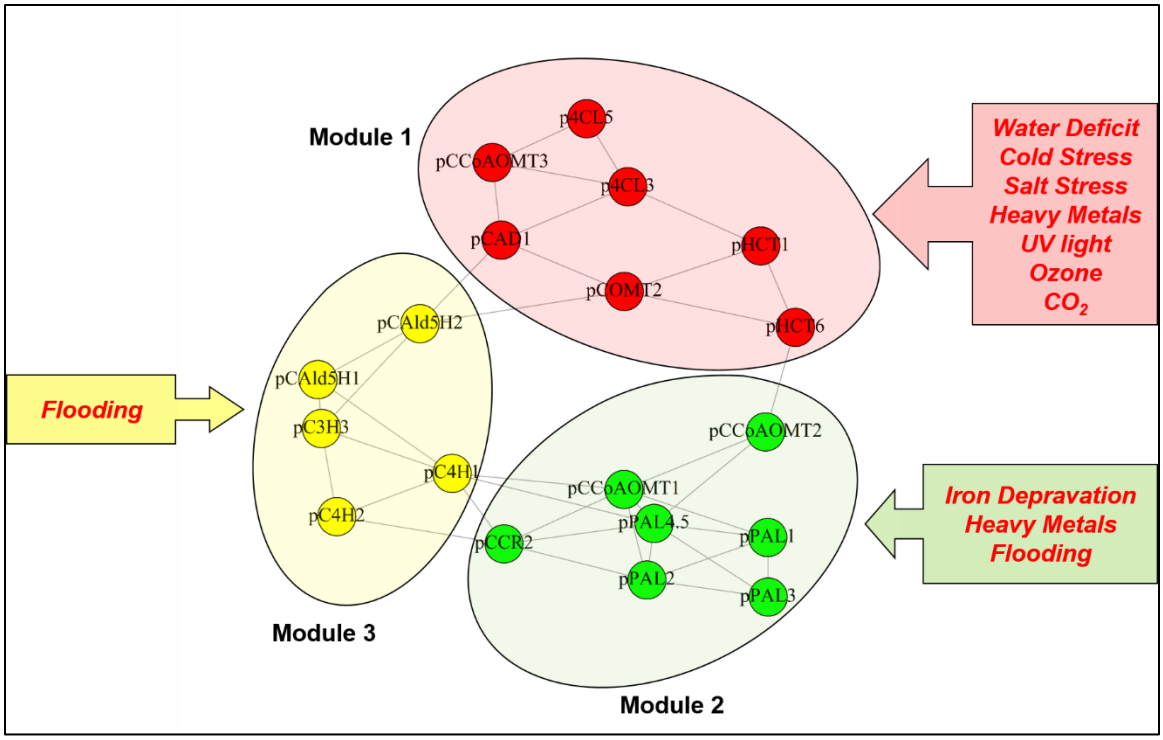

Figure 4.19: The protein modules and the stresses that affect the expression levels of proteins on the modules.

Conclusions:

In this study we were able to identify the modular structure of proteins that are regulated by the genes involved in the monolignol biosynthetic pathway. The hypothesis is that the modular structure results in an improved resiliency of the pathway under perturbations. From the community wise perturbation analysis of the PKMF model, we were able to confirm that the pathway is resilient/robust to the changes in the total lignin content and composition when modules 1 and 2 perturbed. However, perturbation of module 3 results in large variations in the lignin structure. The changes in lignin structure is primarily because the module 3 is composed of CAld5H1 and CAld5H2 proteins that catalyze the last step in the monolignol biosynthesis pathway. Hence,

changes made to module 3 would result in the changes in S:G ratio. It has also been suggested that the enzymes in module 3 belongs to the cytochrome p-450 enzyme family, which have evolved recently compared to the other enzymes in the pathway. Recent studies on the effect of various abiotic and biotic stresses suggest that under variety of stresses the levels of enzymes belonging to module 3 remain unaffected suggesting that the plant is able to maintain levels of enzymes

References:

Agarwal G, Kempe D. (2008). Modularity-Maximizing Graph Communities via Mathematical Programming. In European Physics Journal B, 66:3.

Ali MB, Singh N, Shohael AM, Hahn EJ, Paek KY. (2006). Phenolics metabolism and lignin synthesis in root suspension cultures of Panax ginseng in response to copper stress. Plant Sci.171, 147–154.

Appenroth KJ. (2010).What are “heavy metals” in plant sciences? Acta Physiol. Plant, 32, 615–619.

Bader GD, Hogue CW. (2003). An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics. 4:2.

Badowiec A and Weidner S. (2014). Proteomic changes in the roots of germinating Phaseolus vulgaris seeds in response to chilling stress and post-stress recovery. J. Plant Physiol. 171, 389–398.

Badowiec A, Swigonska S and Weidner S. (2013). Changes in the protein patterns in pea (Pisum sativum L.) roots under the influence of long- and short-term chilling stress and post-stress recovery. Plant Physiol. Biochem. 71, 315–324.

Barabasi AL and Oltvai ZN. (2004). Network biology: Understanding the cell’s functional organization. Nat. Rev. Genet. 5: 101–113.

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. Theory E 10.

Brandes U., Delling D., Gaertler M., Goerke R., Hoefer M., Nikoloski Z., Wagner D. (2006).Maximizing modularity is hard. arXiv:physics, 0608255.

Bussotti F, Agati G, Desotgiu R, Matteini P, Tani C. (2005). Ozone foliar symptoms in woody plant species assessed with ultrastructural and fluorescence analysis. New Phytol. 166, 941–955.

Cabane M, Afif D, Hawkins S. (2013). Lignins and abiotic stresses. Adv. Bot. Res. 61, 220.

Cabané M, Pireaux JC, Leger E, Weber E, Dizengremel P, Pollet B, Lapierre C. (2004). Condensed lignins are synthesized in poplar leaves exposed to ozone. Plant Physiol. 134, 586–594.

Carpita NC and Gibeaut DM. (1993). Structural models of primary cell walls in flowering plants: Consistency of molecular structure with the physical properties of the walls during growth. Plant J., 3, 1–30.

Chandran D, Sharopova N, Ivashuta S, Gantt JS, VandenBosch KA, Samac DA. (2008). Transcriptome profiling identified novel genes associated with aluminium toxicity, resistance and tolerance in Medicago truncatula. Planta 228, 151–166.

Chen F, Dixon RA. (2007). Lignin modification improves fermentable sugar yields for biofuel production. Nat Biotechnol 25 759–761.

Chen F, Duran AL, Blount JW, Sumner LW, Dixon RA (2003) Profiling phenolic Chinnusamy V, Zhu J, Zhu JK. (2004). Salt stress signaling and mechanisms of plant salt tolerance. Genet. Eng. (NY), 27, 141–177.

Chinnusamy V, Zhu J, Zhu JK. (2007). Cold stress regulation of gene expression in plants. Trends Plant Sci. 12, 444–451.

Christianson JA, Llewellyn DJ, Dennis ES, Wilson IW. (2010). Globalgene expression responses to waterlogging in roots and leaves of cotton (Gossypium hirsutum L.). Plant Cell Physiol. 51, 21–37.

Claus H. (2004). Laccases: Structure, reaction, distribution. Micron, 35, 93–96.
D'Emilio M, Caggiano R, Macchiato M, Ragosta M, Sabia S. (2012). Soil heavy metal contamination in an industrial area: Analysis of the data collected during a decade. Environ. Monit. Asses.185, 5951–5964.

Ding YD, Chang JW, Guo J et al. (2014). Prediction and functional analysis of the sweet orange protein-protein interaction network. BMC Plant Biology 14:213.

Dixon RA, Chen F, Guo D, and Parvathi K. (2001). The biosynthesis of monolignols: A “metabolic grid”, or independent pathways to guaiacyl and syringyl units? Phytochemistry 57, 1069–1084.

Domon J-M, Baldwin L, Acket S, Caudeville E, Arnoult S, Zub H, Gillet F, Lejeune-Hénaut I, Brancourt-Hulmel M, Pelloux J. et al. (2013). Cell wall compositional modifications ofMiscanthus ecotypes in response to cold acclimation. Phytochemistry, 85, 51–61.

Doncheva S, Georgieva K, Vassileva V, Stoyanova Z, Popov N, Ignatov G. (2005). Effects of succinate on manganese toxicity in pea plants. J. Plant Nutr. 28, 47–62.
Doye JPK and Massen CP. (2005). Characterizing the network topology of the energy landscapes of atomic clusters. J. Chem. Phys. 122, 084105.

Dunn R, Dudbridge F, Sanderson CM. (2005). The use of edge-betweenness clustering to investigate biological function in protein interaction networks. BMC Bioinformatics.

Eckey-Kaltenbach H, Ernst D, Heller W, Sandermann H. (1994). Biochemical plant responses to ozone: IV. Cross-induction of defensive pathways in parsley (Petroselinum crispum L.) plants. Plant Physiol.104, 67–74.

Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, and Gardner TS. (2007). Largescale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 5, 8.

Fan L. (2006). Progressive inhibition by water deficit of cell wall extensibility and growth along the elongation zone of maize roots is related to increased lignin metabolism and progressive stelar accumulation of wall phenolics. Plant Physiol.140, 603–612.

Feigl G, Kumar D, Lehotai N, Tugyi N, Molnor A, Ordog A, Szepesi A, Gemes K, Laskay G, Erdei L. et al. (2013). Physiological and morphological responses of the root system of Indian mustard (Brassica jumcea L. Czern.) and rapeseed (Brassica napus L.) to copper stress. Ecotox. Environ. Safe, 91, 179–189.

Good BH, Montjoye YAD, and Clauset A. (2010). Performance of modularity maximization in practical contexts. Physical Review E, 81(4):046106.

Goulas, E.; Schubert, M.; Kieselbach, T.; Kleczkowski, L.A.; Gardeström, P.; Schröder, W.; Hurry, V. The chloroplast lumen and stromal proteomes of Arabidopsis thaliana show differential sensitivity to short- and long-term exposure to low temperature. Plant J. 2006, 47, 720–734.

Guimerà R., Amaral L. A. N. (2005b). Functional cartography of complex metabolic networks.Nature 433, 895–90010.1038/nature03288.

Günthardt-Goerg MS, Mc Quattie CJ, Scheidegger C, Rhiner C, Matyssek R. (1997). Ozone-induced cytochemical and ultrastructural changes in leaf mesophyll cell walls. Can. J. For. Res. 27, 453–463.

Günthardt-Goerg MS. (1996). Different responses to ozone of tobacco, poplar, birch, and alder. J. Plant Physiol. 48, 207–214.

Guo J, Wang MH. (2009). Characterization of the phenylalanine ammonia-lyase gene (SlPAL5) from tomato (Solanum lycopersicum L.). Mol. Biol. Rep. 36, 1579–1585.
Hartuv E and Shamir R. (1999). Clustering algorithm based graph connectivity. Information Processing Letters, 76(4-6): 175 –181.
Hartwell LH, Hopfield JH, Leibler S and Murray AW. (1999). From molecular to modular cell biology. Nature 402:C47–C52.

Hellerstein MK. (2003). In vivo measurement of fluxes through metabolic pathways: The missing link in functional genomics and pharmaceutical research. Annual Review of Nutrition 23: 379–402.

Huang TL, Nguyen QTT, Fu SF, Lin CY, Chen YC, Huang HJ. (2012). Transcriptomic changes and signaling pathway induced by arsenic stress in rice roots. Plant Mol. Biol. 80, 587–608.

Iman RL, Davenport JM, and Zeigler DK.(1980). Latin Hypercube Sampling (A Program User's Guide): Technical Report SAND79-1473, Sandia Laboratories, Albuquerque.

Kelepertzis E. (2014). Accumulation of heavy metals in agricultural soils of Mediterranean: Insights from Argolida basin, Peloponnese, Greece. Geoderma, 221, 82–90.

Kimura M, Yamamoto Y, Seki M, Sakurai T, Sato M, Abe T, Yoshida S, Manabe K, Shinozaki K, Matsui M. (2003). Identification of Arabidopsis genes regulated by high light-stress using cDNA microarray. Photochem. Photobiol. 77, 226–233.

Komatsu S, Kobayashi Y, Nishizawa K, Nanjo Y, Furukawa K. (2010). Comparative proteomics analysis of differentially expressed proteins in soybean cell wall during flooding stress. Amino Acids, 39, 1435–1449.

Lee Y, Treviño, E, Dixon RA, Voit EO. (2012). Functional Analysis of Metabolic Channeling and Regulation in Lignin Biosynthesis: A Computational Approach. PLoS Computational Biology. Vol. 8 Issue 11.Ragauskas AJ, Williams CK, Davison BH, Britovsek G, Cairney J, et al. (2006). The path forward for biofuels and biomaterials. Science 311: 484–489.

Li L, Zhou Y, Cheng X, Sun J, Marita JM, Ralph J, Chiang VL. (2003). Combinatorial modification of multiple lignin traits in trees through multigene cotransformation. Proc Natl Acad Sci USA 100 4939–4944.

Li X, Weng JK, Chapple C. (2008). Improvement of biomass through lignin modification. Plant J 54: 569–581.

Li XJ, Yang MF, Chen H, Qu LQ, Chen F, Shen SH. (2010). Abscisic acid pretreatment enhances salt tolerance of rice seedlings: Proteomic evidence. Biochim. Biophys. Acta, 1804, 929–940.

Ma S and Bohnert HJ. (2007). Gene Networks for the Integration and better Understanding of Gene Expression Characteristics. Weed Science Journal. 56:314-321. metabolites in transgenic alfalfa modified in lignin biosynthesis. Phytochemistry 64: 1013-1021.

Moura JC, Bonine CA, de Oliveira Fernandes Viana J, Dornelas MC, Mazzafera P. (2010). Abiotic and biotic stresses and changes in the lignin content and composition in plants. J Integr Plant Biol 52:360–376.

Moura JCMS, Bonine CAV, de Oliveira Fernandes Viana J, Dornelas MC, Mazzafera P. (2010). Abiotic and biotic stresses and changes in the lignin content and composition in plants. J. Integr. Plant Biol. 52, 360–376.

Moura-Sobczak J, Souza U, Mazzafera P. (2011). Drought stress and changes in the lignin content and composition in Eucalyptus. BMC Proc. 2011, 5, 103.

Munns R. (2005). Genes and salt tolerance: Bringing them together. New Phytol.167, 645–663.

Neves GYS, Marchiosi R, Ferrarese MLL, Siqueira-Soares RC, Ferrarese-Filho O. (2010). Root growth inhibition and lignification induced by salt stress in soybean. J. Agron. Crop. Sci. 196, 467–473.

Newman M, Girvan M. (2004). Finding and evaluating community structure in networks. Phys. Rev. E 69, 026113.

Newman M. E. J. (2006). Modularity and community structure in networks. Proc. Natl. Acad. Sci. U.S.A. 103, 8577–858210.1073/pnas.0601602103.

Ovacik M and Androulakis I. (2008). On the Potential for Integrating Gene Expression and Metabolic Flux Data.Current Bioinformatics 3(3): 142-148.

Pandey A, Rajamani U, Verma J, Subba P, Chakraborty N, Datta A, Chakraborty S, Chakraborty N. (2010). Identification of extracellular matrix proteins of rice (Oryza sativa L.) involved in dehydration-responsive network: A proteomic approach. J. Proteome Res. 9, 3443–3464.

Paoletti E, de Marco A, Beddows DCS, Harrison RM, Manning WJ. (2014). Ozone levels in European and USA cities are increasing more than at rural sites, while peak values are decreasing.Environ. Pollut. 192, 295–299.

R Sharan et al. (2005). Conserved patterns of protein interaction in multiple species. PNAS, 102(6):1974–1979.

Racherla PN, Adams PJ. (2008). The response of surface ozone to climate change over the eastern United States. Atmos. Chem. Phys. 871–885.

Ralph J, Akiyama T, Kim H, Lu F, Schatz PF, Marita JM, Ralph SA, Reddy MSS, Chen F, Dixon RA. (2006). Effects of coumarate-3-hydroxylase downregulation on lignin structure. J Biol Chem281 8843–8853.

Ravasz E, Somera A, Mongru D, Oltvai Z, Barabasi A-L. Kitano H. (2002). Systems biology: a brief overview. Science 295. Hierarchical organization of modularity in metabolic networks. 1662–1664. Science 297, 1551–1555.

Reddy MS, Chen F, Shadle G, Jackson L, Aljoe H, Dixon RA (2005) Targeted downregulation of cytochrome P450 enzymes for forage quality improvement in alfalfa (Medicago sativa L.). Proc Natl Acad Sci USA 102: 16573-16578.

Richet N, Afif D, Huber F, Pollet B, Banvoy J, Zein R, Lapierre C, Dizengremel P, Perré P, Cabané M. (2011). Cellulose and lignin biosynthesis is altered by ozone in wood of hybrid poplar (Populus tremula × alba). J. Exp. Bot. 62, 3575–3586.

Rives AW and Galitski T. (2003). Modular organization of cellular networks. Proc. Natl. Acad. Sci. 100: 1128–1133.

Rogers LA, Campbell MM. (2004). The genetic control of lignin deposition during plant growth and development. New Phytol 164:17-30.

Salekdeh GH, Siopongco J, Wade LJ, Ghareyazie B, Bennett J. (2002). Proteomic analysis of rice leaves during drought stress and recovery. Proteomics 2, 1131–1145.

Seki, M, Narusaka M, Ishida J, Nanjo T, Fujita M, Oono Y, Kamiya A, Nakajima M, Enju A, Sakurai T. (2002). Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. Plant J. 2002, 31, 279–292.

Shaik R and Ramakrishna W. (2014). Machine learning approaches distinguish multiple stress conditions using stress-responsive genes and identify candidate genes for broad resistance in rice, Plant Physiol, 164, pp. 481–495.

Shamir R and Ulitsky I. (2009). Identifying functional modules using expression profiles and confidence-scored protein interactions. Bioinformatics 25 (9): 1158-1164.

Sharma YK, Davis KR. (1994). Ozone-induced expression of stress-related genes in Arabidopsis thaliana. Plant Physiol.105, 1089–1096.

Solecka D, Kacperska A. (2003). Phenylpropanoid deficiency affects the course of plant acclimation to cold. Physiol. Plant, 119, 253–262.

Stephanopoulos G. (1999). Metabolic fluxes and metabolic engineering, Metabolic Engineering 1:1-11.
Stitt M, Hurry V. (2002). A plant for all seasons: Alterations in photosynthetic carbon metabolism during cold acclimation in Arabidopsis. Curr. Opin. Plant Biol. 5, 199–206.

Syros TD, Yupsanis TA, Economou AS. (2005). Expression of peroxidases during seedling growth inEbenus cretica L. as affected by light and temperature treatments. Plant Growth Regul. 46, 143–151.

Terzi R, Güler NS, Çaliskan N, Kadioglu A. (2013). Lignification response for rolled leaves of Ctenanthe setosa under long-term drought stress. Turk. J. Biol. 37, 614–619.
Treviño S III, et al. (2012). Robust detection of hierarchical communities from Escherichia coli gene expression data. PLoS Comput. Biol.

Verslues PE, Agarwal M, Katiyar-Agarwal S, Zhu J, Zhu JK. (2006). Methods and concepts in quantifying resistance to drought, salt and freezing, abiotic stresses that affect plant water status. Plant J. 45, 523–539.

Wang JP, Naik PP, Chen H-C, Shi R. et al. (2014). Complete proteomic-based enzyme reaction and inhibition kinetics reveal how monolignol biosynthetic enzyme families affect metabolic flux and lignin in Populus trichocarpa. Plant Cell, 26, 894–914.

Weng J, Aphasizheva I, Etheridge RD, Huang L, Wang X, Falick AM and Aphasizhev R. (2008). Guide RNA-binding complex from mitochondria of trypanosomatids. Mol. Cell 32: 1–12.

Wiese CB, Pell EJ. (2003). Oxidative modification of the cell wall in tomato plants exposed to ozone.Plant Physiol. Biochem. 41, 375–382.

Yeang CH and Vingron M. (2006). A joint model of regulatory and metabolic networks. BMC Bioinformatics, 7:332.

Yoshimura K, Masuda, A, Kuwano M, Yokota A, Akashi K. (2007). Programmed proteome response for drought avoidance/tolerance in the root of a C3 xerophyte (wild watermelon) under water deficits. Plant Cell Physiol. 49, 226–241.

Zabotin I, Barisheva TS, Trofimova OI, Toroschina TE, Larskaya IA, Zabotina OA. (2009). Oligosaccharin and ABA synergistically affect the acquisition of freezing tolerance in winter wheat. Plant Physiol. Biochem.47, 854–858.

Zhang L, Yu S, Zuo K et al. (2015). Identification of Gene Modules Associated with Drought Response in Rice by Network-Based Analysis. PLOS one DOI: 10.1371/journal.pone.0033748.

Zhang LV, King OD, Wong SL, Goldberg DS, Tong AH, Lesage G, Andrews B, Bussey H, Boone C and Roth FP. (2005). Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. J. Biol. 4: 6

Zhao Q, Zhang H, Wang T, Chen S, Dai S. (2013). Proteomics-based investigation of salt-responsive mechanisms in plant roots. J. Proteomics, 82, 230–253.
Zhou J, Lee C, Zhong R, Ye ZH. (2009). MYB58 and MYB63 are transcriptional activators of the lignin biosynthetic pathway during secondary cell wall formation in Arabidopsis. Plant Cell 21: 248–266.

Zhu YN, Shi DQ, Ruan MB, Zhang LL, Meng ZH, Liu J, Yang WC. (2013). Transcriptome analysis reveals crosstalk of responsive genes to multiple abiotic stresses in cotton (Gossypium hirsutum L.). PLoS One. 8, e80218.

CHAPTER V

MATHEMATICAL MODELING AND VALIDATION OF LIGNIN BIOSYNTHESIS IN *POPULUS TRICHOCHARPA* USING PHENOTYPIC DATA

Punith Naik[1], Jack Wang[2], Liu Jie[2], Hsi-Chuan Chen[2], Rui Shi[2], Christopher M. Shuford, Quanzi Li[2], Kevin Lin[2], David C. Muddiman, Ronald Sederoff[2], Vincent Chiang[2], Cranos Williams[3], and Joel Ducoste[1*]

[1] Civil, Construction, and Environmental Engineering, North Carolina State University, Raleigh, North Carolina 27695
[2] Forest Biotechnology Group, Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, North Carolina 27695
[3]Electrical and Computer Engineering, North Carolina State University, Raleigh, North Carolina 27695

Abstract:

Lignin is the second most abundant polymer found primarily in the cell walls of all vascular plants. With the recent shift in focus on renewable energy, plant biomass has been identified as a future energy source. However, the presence of lignin hinders the release of sugar, which can be fermented to ethanol. The variation in lignin content and composition among different species is primarily due to the difference in proportions of H, G and S monolignols. The variability in these monolignols has a direct correlation on the structural rigidity of lignin. While recent advances in metabolomics and proteomics have significantly enhanced our understanding of lignin biosynthesis, the regulation of monolignol biosynthetic pathway is still not clear. Several computational and mathematical methods have been used as effective tools for understanding the structure and regulation of complex biochemical networks. As increasing amounts of data are being published, the application of these methods to study lignin biosynthesis may be crucial in developing genetically engineered crops with reduced recalcitrance. Here, we experimentally validate a mass action kinetic model using transgenic data. In addition to the mass action model, we also developed an Artificial Neural Network (ANN) model to predict the changes in lignin composition as a function of enzyme concentrations to evaluate the effectiveness of purely data driven modeling approaches. The mass action model was able to predict the changes in lignin composition as a function of enzyme concentrations and were quantitatively consistent with several transgenic experiments. The Predictive Kinetic Metabolic Flux (PKMF) model was able to account for 73% of the variations in S/G ratio resulting from changes in protein concentrations. The ANN model, however, was able to predict 80% of the variations in the lignin composition. The results generated by our model of *invivo* lignin biosynthesis demonstrate that mathematical modeling can be an effective complement to experimental biotechnological and transgenic approaches in plants. The ANN model may also serve as a complementary tool to predict the changes in lignin composition resulting from enzymatic perturbations. Although it predicted a higher percent accuracy,

ANN, cannot be used to explain the changes in the metabolic flux and metabolite concentrations resulting from the perturbations.

5.1 Introduction:

Lignin is the second most abundant polymer found naturally in the biosphere. The primary role of lignin is to provide strength and rigidity to the plant cell wall for mechanical support, water transport, and resistance against pathogen attack. With the recent focus on plant biomass as a source of biofuels, the recalcitrance of lignin can hinder the use of plant biomass as biofuels. Overcoming this ‘‘recalcitrance” is essential fermentation of sugars from hemicellulose and cellulose into ethanol, butanol or other biofuels (Raguskas et al, 2006). Since lignin plays a central process in this process, much attention has been focused on understanding lignin biosynthesis and on exploring the potential of developing transgenic plants with reduced lignin content or modified lignin which would greatly reduce or even eliminate the acid pretreatment step (Chen et al, 2007).

Extensive research efforts have enabled researchers to identify the specific roles of most genes involved in the monolignol biosynthetic pathway (Lee et al, 2012). Complete genome sequences along with functional annotation of relevant genes for two model plants, *Arabidopsis thaliana* and the black cottonwood *Populus trichocarpa*, are available (Somerville, 1999). Although the information from the genome sequences are valuable, on its own, it is insufficient for predicting how the monolignol biosynthetic pathway would respond to changes in enzyme activities or gene expression.

Systems modelling of intracellular biochemical processes can provide quantitative insight into a cell's response to stimuli and perturbations (Lee et al, 2010; Kholodenko et al, 2012). Understanding the dynamic behavior of biological systems has been one of the challenges in the post-genomic era. Quantitative models describing the changes in the dynamic of metabolic network has been a valuable tool to understand complex biological systems properties and to guide experimentation (Kitano, 2007). In this context, ordinary differential equations have been used predominantly to simulate the dynamics of metabolic network (Bakker et al., 1999; Klipp et al., 2007) that require prior knowledge of the network structure and a large amount of experimental information, such as initial concentrations of metabolites and kinetic parameters. Mechanistic kinetic rate expressions have been the typical approach in metabolic networks modeling. The advantages of using kinetic based ODE models in comparison to constraint based pathway models have been discussed (Rohwer et al, 2012). The metabolic networks of plants have been genetically engineered in the past to achieve a desired outcome (Sewalt et al., 1994; Ye et al., 2000; Kebeish et al., 2007; Aluru et al., 2008). However, there has been instances where genetic modification of plants has resulted in unanticipated or little or no net change to the system (Lee et al, 2010). As an example, down-regulation of the enzymes involved in the terminal steps of the monolignol biosynthetic pathway did not result in a reduction of lignin content (Dwivedi et al., 1994; Atanassova et al., 1995; Van Doorsselaere et al., 1995). Some genetic modifications of plants have resulted in unanticipated changes in cellular metabolite pools or have affected pathways that are not directly involved in the targeted network. These unexpected alterations sometimes lead to negative effects on plant growth and

development (Wu et al., 2006; Dauwe et al., 2007; Napier, 2007). In many cases, these problems are due to an incomplete understanding of network dynamics and control even though the general network structure may be known. Thus, the current challenge is to find ways to detect such changes in order to prevent or overcome these difficulties.

One of the ways to make genetic engineering a powerful tool is through the use of predictive models that provide accurate description of metabolic networks kinetic equations (Daae et al., 1999; Libourel and Shachar-Hill, 2008). Mass action based kinetic models have been used to understand the dynamics and steady state behavior as well as regulatory control mechanisms (Pearcy et al., 1997; Wang et al, 2015).

In this paper, we used the previously developed Predictive Kinetic Metabolic Flux (PKMF) model (Wang et al, 2014) that was modified to incorporate the protein interactions between Ptr4CL enzymes (Song et al, 2014 and Chen et al, 2014), to validate the experimentally determined S/G ratio from transgenic experiments (Sermsawat et al, 2015). In order to obtain a confidence interval for the predictions, we performed simulations on the PKMF model by randomly sampling the enzyme concentrations using the standard errors of the experimentally determined protein concentrations for all 21 enzymes involved in the monolignol biosynthesis pathway.

Although the mechanistic model is an effective approach towards understanding the role of perturbations on the lignin composition and structure as well as quantifying the changes in the steady state flux distributions in the pathway, developing a detailed model for every biological process is tedious and time consuming. Alternatively, if the

only goal of the model is to predict the changes in lignin structure as a result of perturbations, researchers can employ statistical learning techniques to carry out that goal. In this paper, we also developed an artificial neural network model to predict the changes in S/G ratio as a function of protein concentrations of all the enzymes involved in the monolignol biosynthesis. Artificial Neural Networks (ANN) are one of the most popular statistical learning tools due to their ability to learn/approximate complex, nonlinear behavior of relatively poorly understood processes without any prior knowledge of the relationship between the inputs and outputs (Richardson et al., 2002). The reliability of using neural networks in practice has been affirmed in many different applications ranging from pattern recognition (Basu et al, 2010), in chromatographic spectra (Jalali-Heravi et al, 2009; Cartwright, 2009), and expression profiles (Urda et al, 2010; Lancashire et al, 2010), to functional analyses of genomic and proteomic sequences (Choe et al, 2010) to QSAR models (Devilliers, 1996; Agrafiotis et al, 2002).

5.2 Methodology:

In this paper, we used a previously developed kinetic model whose parameters were determined experimentally (Wang et al, 2014) and the metabolic fluxes for the pathways involved in the enzyme complex were incorporated into the PKMF model (Song et al, 2014). In order to compare the experimentally obtained S/G ratio to the simulated results, we used the protein concentrations obtained from transgenic experiments as input to the PKMF model and the corresponding S/G ratio from the transgenic experiments were compared to the output from the PKMF model. We

assumed that the measured protein concentrations was subjected to Gaussian noise with the mean corresponding to the average of the replicates and the standard errors (S.E) as the standard deviation for the measured data. The input flux to the pathway model was calculated based on the WT S/G ratio for each batch of transgenic data. The initial concentrations of the metabolites along with the input flux for each batch is shown in the Table 5.S.1. The protein concentration for each enzyme was then sampled from the Gaussian distribution, which was then input to the model. The PKMF model was simulated to steady state and repeated for a 1000 different sampled protein concentration data. A box and whiskers plot was used to plot the S/G distributions predicted by the PKMF model for each construct.

The steady state concentrations of all the metabolites in the pathway were computed by solving the set of ODE equations using the MATLAB ode15s function. The computation of steady state metabolite concentrations using ODEs requires the specification of initial metabolite concentrations, which is usually unknown in plant biosynthetic pathways. To overcome this, we employed a Monte Carlo sampling procedure to sample the initial concentrations from a pre-specified range and then simulated the ODE to steady state under WT enzyme concentrations. The procedure was repeated for 10000 different initial concentrations such that we have a large enough sample to draw conclusions about the nature of steady states. The steady state metabolite concentrations were then used as initial concentrations to compute the steady state metabolite concentrations for the pathway when subjected to enzymatic perturbation. The outline of the modeling approach is shown in Figure 5.1. For step 1

shown in Figure 5.1a, the initial concentration of the metabolites were sampled from a range of 0-10k_m using Latin Hypercube Sampling (LHS) (Mckay et al, 1979 ; Iman et al, 1980). The parameter k_m is the Michaelis-Menten constant for each enzyme substrate combination. The k_m values are listed in Wang et al (2014). The ODE model was then simulated to steady state using the initial metabolite concentrations, kinetic parameters and WT enzyme concentrations. The above procedure was repeated for 10000 different sets of initial concentrations. The mode of the resulting steady state metabolite concentrations were then used as an initial concentration for the second step of the simulation, which is shown in Figure 5.1b. In the second step, the model is simulated for different enzyme concentrations resulting from targeting a particular gene involved in monolignol biosynthesis. The enzyme concentrations were sampled from a Gaussian distribution, with an experimentally determined mean and standard deviation. The input flux is fixed such that the resulting S/G ratio for WT enzyme concentrations match the experimentally determined WT S/G ratio for each transgenic batch. The model is simulated to steady state for different enzyme concentrations and the resulting S/G distribution is then plotted as a box and whiskers plot.

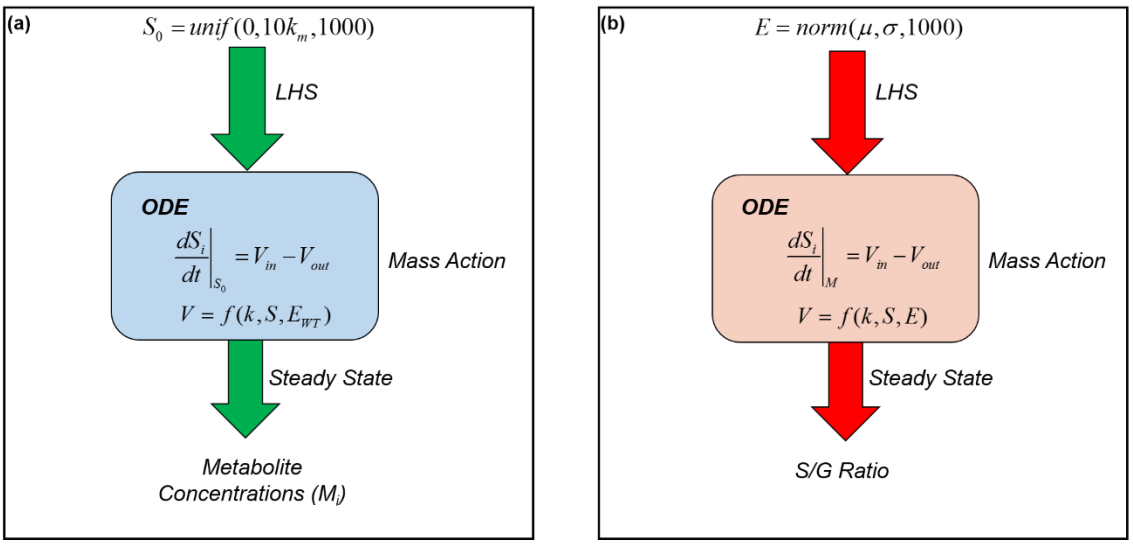

Figure 5.1: (a) An outline of the methodology used to assess the effect of uncertainty in initial metabolic concentrations on the steady state metabolic concentrations. (b) The simulation procedure used to obtain the distributions of S/G values for different transgenic experiments.

5.2.1 Data:

The protein concentrations and the corresponding lignin composition data was obtained from transgenic experiments. Each construct in the transgenic experiment corresponds to a perturbation of a single or group of enzymes involved in the monolignol biosynthesis. The transgenic experiments were split into 4 different batches. For a more detailed description of the transgenic data and constructs, the reader is referred to Wang et al (2015). The constructs for each batch of the transgenic experiments along with the experimentally measured S/G ratio are shown in the Tables 5.1, 5.2 and 5.3. Although, the experimental data was measured for five batches, we compared results from batch1, batch 2, batch 4, and batch 5 transgenics with the PKMF model. We did not perform simulations on batch 3 data primarily because batch 3 data consisted of only CAD1 and CAD2 transgenics. Because WT concentration of CAD2

was very low, the S/G values predicted by the PKMF model resulting from the perturbation of CAD2 enzymes did not show any variation with respect to the WT S/G levels.

Table 5.1: The description of constructs in batch 1 and batch 2 along with the corresponding S: G ratio determined experimentally.

Tree ID	Targets	S:G
a1-1-1	pal1	2.04
a1-8-1	pal1	1.86
a2-1-1	pal3	2.08
a2-3-1	pal3	2.18
i7-10-1	pal2,pal4,pal5	2.29
i7-2-1	pal2,pal4,pal5	2.44
i7-6-1	pal2,pal4,pal5	2.09
i7-8-1	pal2,pal4,pal5	2.17
o67-3-1	oCAD2	2.14
o67-9-1	oCAD2	2.17
NSF1_WT-1	WT	1.97
a3-3-1	pal4	2.38
a3-4-1	pal4	2.38
a4-3-1	pal5	3.08
a5-6-1	pal2	2.32
i6-5-1	pal1,pal3	2.22
i6-9-1	pal1,pal3	2.46
i8-1-1	pal1,pal2,pal3,pal4,pal5	2.59
i8-10-1	pal1,pal2,pal3,pal4,pal5	2.40
NSF2_WT-1	WT	2.33

Table 5.2: The description of constructs in batch 4 and batch 5 along with the corresponding S: G ratio determined experimentally

Tree ID	Targets	S:G
a10-8-1	c4h1	1.98
a9-2-1	c4h2	2.21
a9-6-1	c4h2	1.96
i26-4-1	ccr2	1.43
i26-9-1	ccr2	1.77
NSF4_WT-1	WT	1.81
a17-1-1	hct1	2.18
a17-10-1	hct1	2.25
a17-4-1	hct1	2.09
a17-9-1	hct1	2.18
a18-11-1	hct6	2.20
a18-4-1	hct6	1.99
a18-5-1	hct6	2.08
a18-9-1	hct6	1.99
a22-10-1	ccoaomt1	2.65
a22-17-1	ccoaomt1	2.21
a27-M-1	cald5h1	0.65
a28-H-1	cald5h2	2.14
a28-L-1	cald5h2	0.71
i19-15-1	hct1,hct6	2.26
i19-4-1	hct1,hct6	2.49
i19-7-1	hct1,hct6	1.99
i29-H-1	cald5h1,cald5h2	1.18
i29-L-1	cald5h1,cald5h2	0.24
i29-M-1	cald5h1,cald5h2	0.12
NSF5_WT-1	WT	2.44

Table 5.3: The description of constructs in batch 3 along with the corresponding S: G ratio determined experimentally.

Tree ID	Target	S:G
i20-10-1	c3h3	3.73
i20-2-1	c3h3	2.27
i20-5-1	c3h3	9.93
i33-10-1	cad1	2.26
i33-2-1	cad1	2.59
i33-5-1	cad1	2.45
i34-2-1	cad2	2.09
i34-6-1	cad2	2.38
i34-8-1	cad2	1.86
i35-1-1	cad1,cad2	2.00
i35-7-1	cad1,cad2	3.02
i69-10-1	c3H3,c4h1,c4h2	2.55
i69-13-1	c3H3,c4h1,c4h2	3.06
i69-4-1	c3H3,c4h1,c4h2	2.77
o65-11-1	cad1,oCAD2	2.30
o65-16-1	cad1,oCAD2	1.90
o65-7-1	cad1,oCAD2	2.49
o66-21-1	cad1,oCAD2	2.15
o66-7-1	cad1,oCAD2	1.98
NSF3_WT	WT	2.56

5.2.2 Artificial Neural Network (ANN):

ANN models have been widely used in quantifying nonlinear relationship between the dependent and independent variables. One of the advantages of ANN is that it does not require prior information about the functional form of the relationship between the dependent and independent variables. In its simplest form, ANN is composed of 3 layers namely the input, hidden and the output layer. The most commonly used architecture for ANN is the multilayered neural network with backpropagation. The backpropagation training is composed of three steps: (i) Input is

passed into the input layer and passed through the hidden layer and realized through the output layer (ii) the difference between the output and the target is back-propagated through the network and finally (iii) the weights are adjusted using optimization such that the error is minimized. The number of neurons in the hidden layers are identified such that the training and testing errors are minimized. A neural network with more than optimum nodes results in overfitting and hence lacks generalization of patterns observed in the data (Hussain et al., 1992).

The data to the input layer of the ANN is normalized and then passed from input layer through the hidden layer and then to the network output layer (Hussain et al., 2002). The data from the input layer is assigned a random weight and then combined with the other inputs in the hidden layer. The output from the hidden layer undergoes a similar linear weighted transformation process. Each neuron can be viewed as a transfer function that converts an input into a sigmoidal output. The output from a neuron is again transformed using an appropriate transfer function (Razavi et al., 2003).

In this study, we used ANN to estimate the changes in S/G ratio and total lignin content (S+G) as a function of protein concentrations. The network was trained using experimentally determined S/G ratios and (S+G) values from transgenic experiments. The network consists of an input layer with 21 neurons, hidden layers and an output layer. Inputs for the network are the 21 enzyme concentrations, while the output is the S/G ratio and (S+G) values. The structure of the proposed ANN is shown in Figure 5.2.

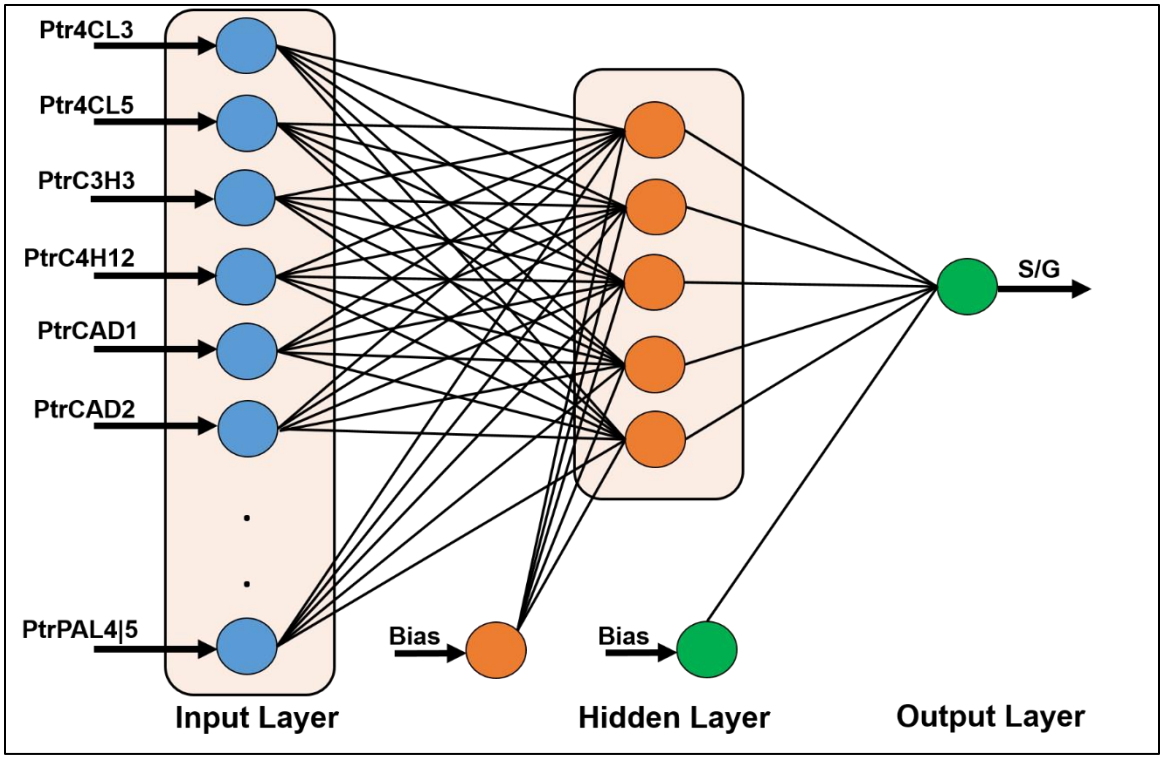

Figure 5.2: The architecture of the ANN used for the prediction of lignin composition consisting of twenty one inputs, one hidden layer with five neurons and one output.

The artificial neural network was built using the Neural Network Toolbox, MATLAB® Release 14 (The Mathworks, Natick, MA). Transfer function between the input and the hidden layer was '*tansig*' (Vogl et al, 1988) and the one between the hidden layer and the output layer was '*logsig*'. Training of the network was performed with the function '*trainlm*' (Hagan and Menhaj, 1994), which updates weight and bias values according to Levenberg–Marquardt optimization (Marquardt, 1963). The parameters of the neural network were obtained using training data such that the mean square error (MSE) between targets and outputs was minimum (Basri et al, 2007; Izadifar et al, 2007; Wang et al, 2008).

The number of neurons in hidden layer, were determined using an iterative procedure, in which the number of hidden layers was varied from 1 to 25. The mean square error (MSE) was used as the error function, the number of neurons resulting in the minimum MSE was chosen as an optimum. The ANN model was trained using the experimental data, which was split into training (70%) and testing (30%) sets. Due to the limited amount of data, we used a 10 fold cross-validation of training data. The data set as divided into 10 equal subsets and for each run, one of the 10 subsets is used as the test set and the other 9 subsets were used as training sets The advantage of cross validation is that it reduces the generalization effect of the model, hence resulting in an unbiased model, which performs well both on training and testing data and avoids over-fitting (Cawley and Talbot, 2010). The training data was used to compute the network parameters. The testing data was used to assess the predictive ability of the generated model.

5.3 Results:

5.3.1 A Kinetic Model to Predict Lignin Composition

The predictive capabilities of the PKMF model was evaluated by comparing the simulated and experimental data. The comparisons were made batch wise for four batches. Because the protein concentration measurements were subjected to experimental errors, we incorporated those errors into our simulations as outlined in the methods section.

The batch 1 and 2 transgenics primarily involve downregulation of PAL genes either individually or a combination of more than one PAL gene. The description of the constructs in batch 1 and 2 is shown in the Table supplemental table 5.S.1. The S/G variation for the PKMF model, along with the corresponding experimental values for different constructs are shown in Figure 5.3. In Figure 5.3, the PKMF model was able to capture the increase and decrease in S/G values compared to the WT for different constructs. The variations of S/G ratio in batch shows that majority of the S/G values are around WT levels. The PKMF model predicted mean S/G values distributed around the WT levels, with some extreme values/outliers, which may be due to the low concentrations of other enzymes involved in the pathway as a result of PAL downregulation.

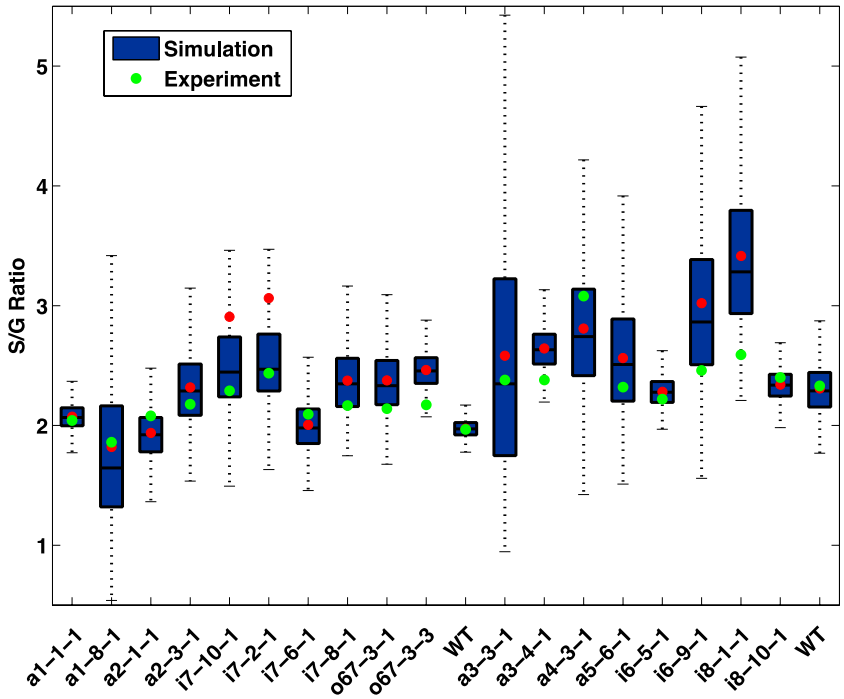

Figure 5.3: Comparison of the predicted S/G ratios shown in box plots to the experimentally measured S/G ratio shown as red dot for batch 1 and batch 2 transgenics.

The distributions of S/G values predicted by PKMF model are in close agreement with the experimentally determined S/G values. The model was able to capture the changes in S/G ratio resulting from the enzymatic perturbations for different constructs. The transgenic S/G values determined experimentally are distributed around the WT levels as expected because lignin biosynthetic pathway is robust to changes in levels of PAL enzymes as seen in Wang et al (2014).

The changes in the S/G levels resulting from the transgenics in batch 4 is shown in Figure 5.4. Batch 4 transgenics primarily consist of C4H1 (a10), C4H2 (a9) and

CCR2 (i29) as targets. The S/G values predicted by the PKMF model shows a large variation for some of the transgenic constructs involving perturbations of C4H1, C4H2 and CCR2. Downregulation of C4H1, C4H2 and CCR2 are known to increase the S/G values with respect to the WT levels (Wang et al, 2014). The increase in S/G ratio is due to re-routing a majority of the metabolic flux through sinapyl alcohol pathway. At low levels of C4H1 and C4H2, the S/G ratio increases with respect to the WT levels, similarly low levels of CCR2 results in an increase in S/G ratio. The range of predicted S/G distribution contains the experimentally measured S/G ratio and in most cases, they lie within the 1st and 3rd quantile of the box plots.

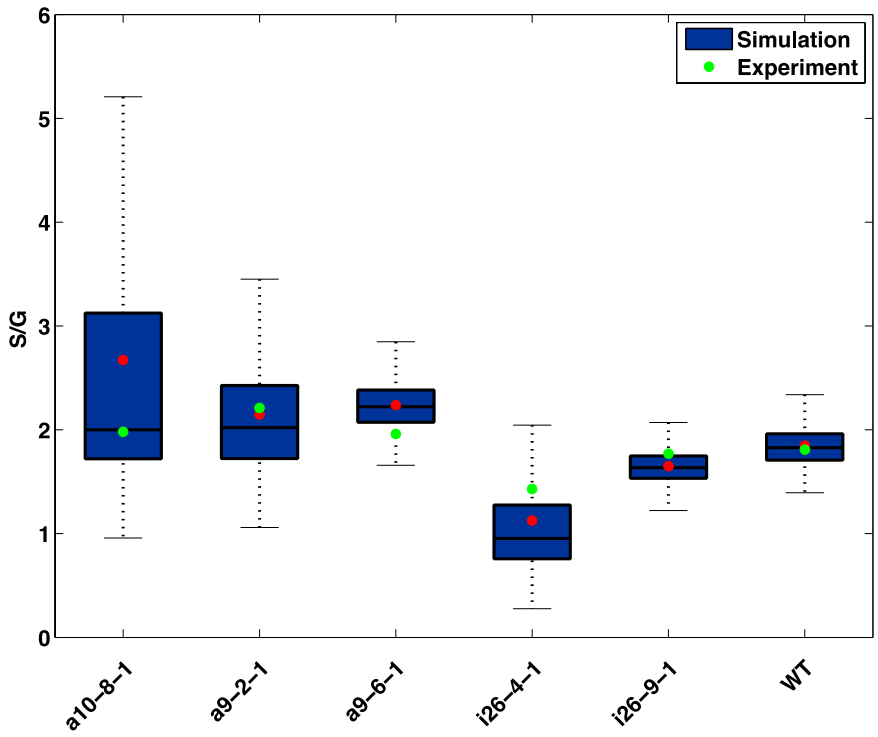

Figure 5.4: Comparison of the predicted S/G ratios shown in box plots to the experimentally measured S/G ratio shown as red dot for batch 4 transgenics.

The final transgenic batch consists of HCT1 (a17), HCT6 (a18), CCoAOMT1 (a22), Cald5H1 (a27), CAld5H2 (a28), HCT1 and HCT6 (i19), CAld5H1 and CAld5H2 (i29) as targets. The transgenic constructs for batch 5 is shown in Table 5.S.2. The S/G predictions from the PKMF model along with the experimental results in Figure 5.5 show that the S/G values vary from 0.3 to 2.5 depending on the targets. The PKMF model is able to predict the S/G values, which is in agreement with the experimental values. The S/G ratio increases when PtrHCT1 and PtrHCT6 enzyme concentrations are down regulated while a decrease in S/G ratio is observed when concentrations of PtrCAld5H1 and PtrCAld5H2 concentrations are down regulated.

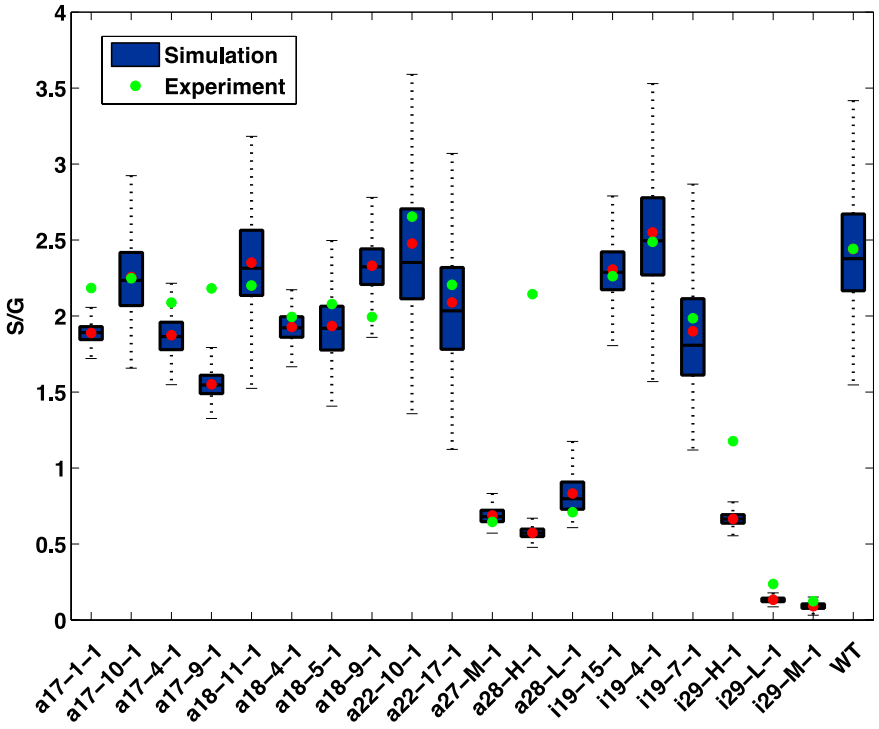

Figure 5.5: Comparison of the predicted S/G ratios shown in box plots to the experimentally measured S/G ratio shown as red dot for batch 5 transgenics.

A comparison of the experimentally measured S/G values and the mean S/G values predicted by the PKMF model is shown in Figure 5.6. In Figure 5.6, the PKMF model is able to predict 73% (R^2) of the variation in the experimentally measured S/G values. The slope of the resulting regression line is 1 suggesting that the model predictions are in close agreement with the experimental data.

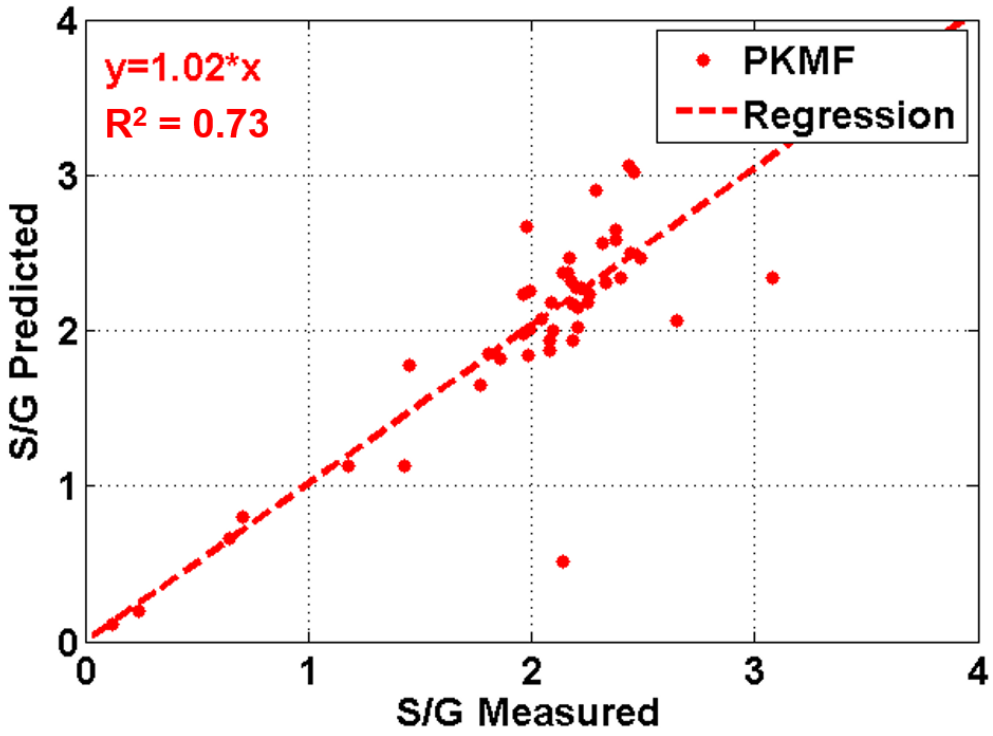

Figure 5.6: The scatter plots of PKMF predicted S/G ratio versus actual S/G ratio for training dataset, blue line is the regression line. The distance of each data from the blue line corresponds to its deviation from the related experimental value.

5.3.2 An Artificial Neural Network for prediction of lignin composition:

In the previous section, the results of the PKMF model were compared to the experimentally determined S/G ratios. The PKMF model was able to predict the changes in S/G ratios due to the changes in the underlying protein concentrations of the enzymes involved in the monolignol biosynthetic pathway. Although the mass action model provides an accurate description of the dynamics and control mechanisms of the pathway (including metabolite concentration, reaction flux, and protein concentrations), developing a kinetic model, requires the experimental determination of all inhibition and activation parameters.

If the only goal of this network model is to solely predict the changes in S/G ratio or the total lignin content (S+G) induced by changes in the protein concentrations, then one can use ANN or some other statistical learning technique to make these predictions. The input to the ANN model is the 21 protein concentrations of enzymes belonging to 10 different families while the output of the model is the S/G ratio and the (S+G) values. The same procedure was repeated to predict the variation in (S+G) as a function of protein concentrations.

Figure 5.7 illustrates the performance of the ANN network for testing data versus the number of neurons in the hidden layer using LM algorithm. The number of neurons in the hidden layer were changed from 1 to 25 and for each Neural Network architecture, the MSE between the experiment and model were compared. A network with 9 hidden neurons for the S/G model and 14 hidden neuron for the (S+G) model

resulted in the minimum MSE for the training and testing data, when LM algorithm was employed.

The performance of the ANN model on the training data is shown in Figure 5.8. In Figure 5.8 (a), the resulting R^2 value for the ANN model on the training data was 0.9, suggesting that the ANN model is able to account for a majority of the variations in the training data. The majority of the S/G ratio are spread around the WT level which is about 2:1. The predictive power of the ANN model can be further explored when using the model on the testing data, the plot of the variation of measured S/G data and the S/G values predicted by ANN using the testing data is shown in Figure 5.8 (a). Results in Figure 5.8 (a) show that the ANN model was able to accurately predict 80% of the variation in the S/G ratio on the testing data. Similarly the performance of the ANN model developed to predict the (S+G) values on the training and testing data is shown in Figure 5.8 (b). The resulting R^2 value for the model on the training data was 0.88, suggesting that the model was able to explain majority of the variation in the total lignin content. When the model was input with the testing data, the model was able to predict 74% of the variation of (S+G) values.

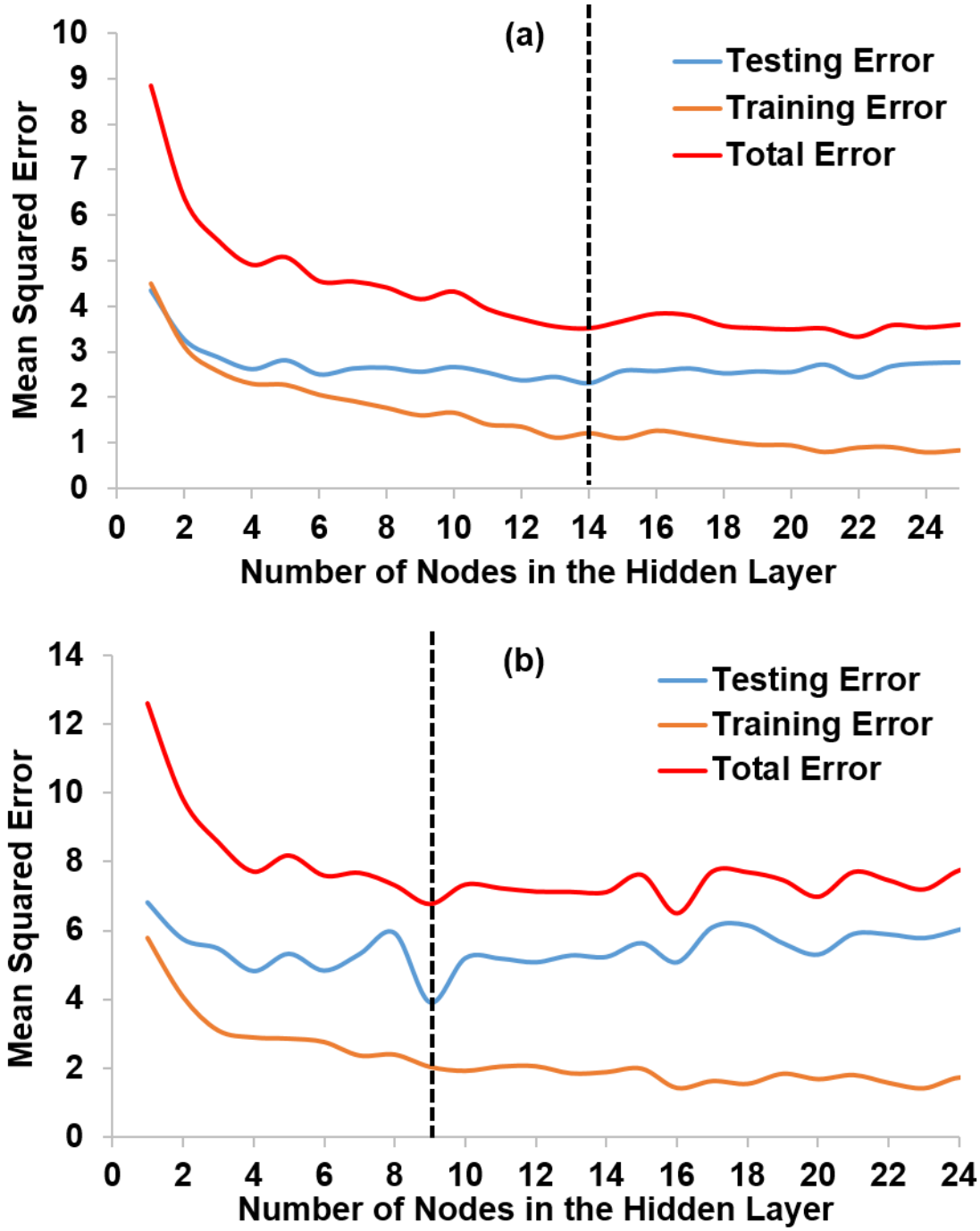

Figure 5.7: Network performance under different number of nodes in the hidden layer. (a) The variation of Mean Squared Error (MSE) as a function number of nodes in the hidden layer for the ANN to predict the total lignin content (S+G). (b) The variation of Mean Squared Error (MSE) as a function number of nodes in the hidden layer for the ANN to predict the lignin composition (S/G)

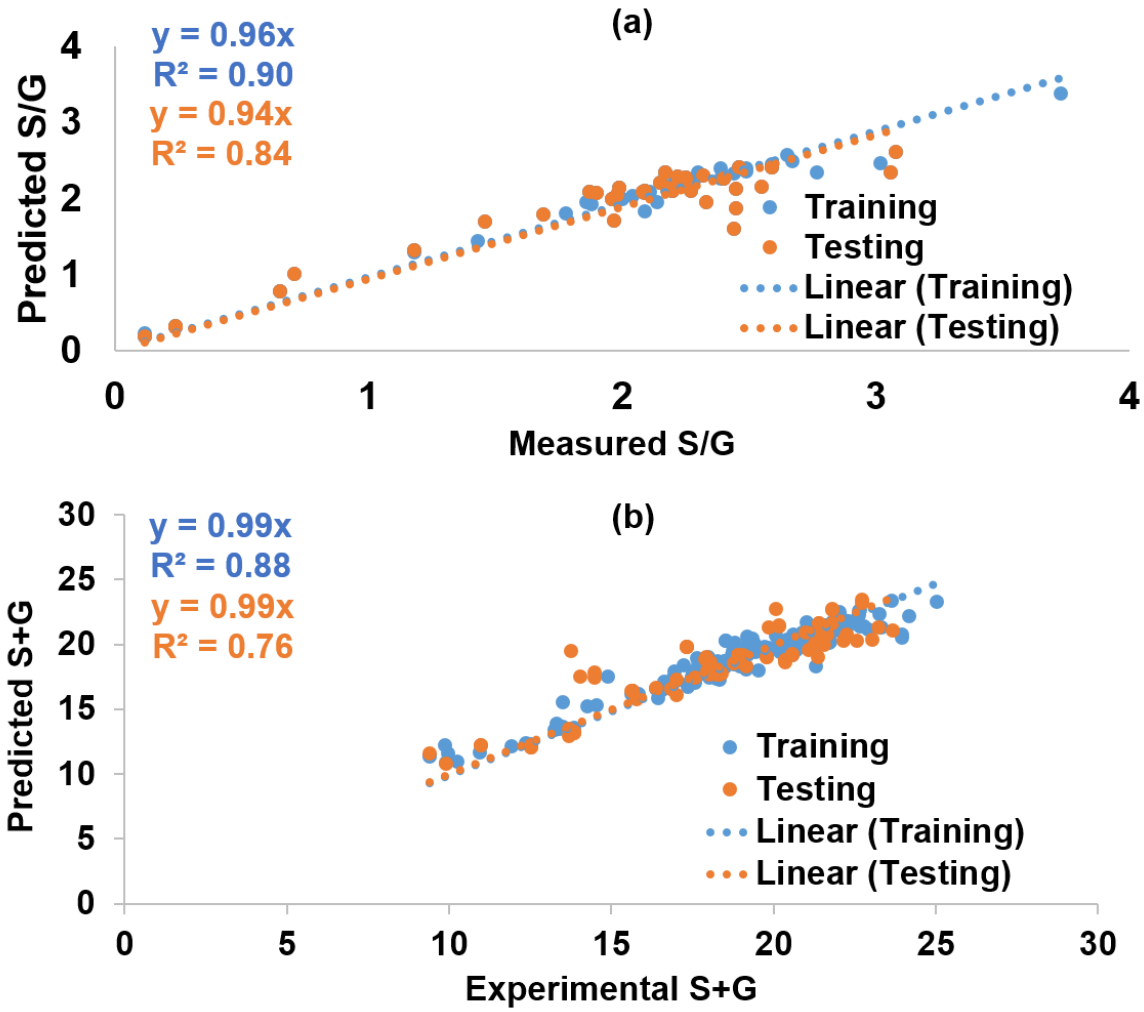

Figure 5.8: (a) The scatter plots of ANN predicted S/G ratio versus actual S/G ratio or testing dataset plotted along with the training dataset, blue line is the regression line for the training dataset and orange line is the regression line for the testing dataset. The distance of each data from the orange line corresponds to its deviation from the related experimental value. (b) The scatter plots of ANN predicted S+G values versus actual S+G values.

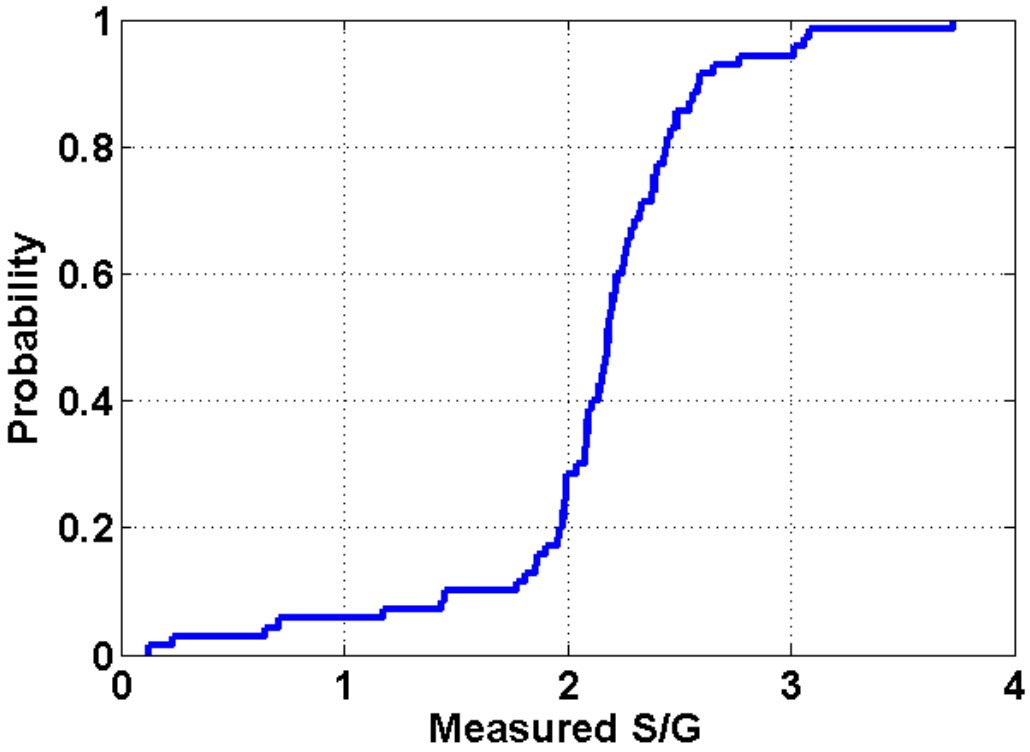

Figure 5.9: A cumulative distribution plot showing the probability of measured S/G ratios.

The discrepancy between the predicted and measured S/G values for lower S/G values is primarily caused by the majority of S/G values measured experimentally are around 2 or greater and hence the ANN model was presented with such values in the training data. The cumulative distribution plot for the experimentally measured S/G values are shown in the Figure 5.9. Results in Figure 5.9 show that 70% of the S/G values lie between 2 to 2.5, hence the training data provided to the ANN would contain more of those values compared to other transgenic results. Similarly for the ANN model developed to predict the S+G values, about 80% of the values lie between 17 and 25, whereas from the Figure 5.8b majority of the discrepancies between the predicted and measured values are for S+G values below 15. Although, ANN is a black box model in

terms of the variable transformation, the role of the inputs on the output of the ANN can be assessed by the weights associated with each input. An input with a larger input suggests that that particular input is important in explaining the variation in the output function as compared to an input with a smaller weight. The weights associated with each of the inputs in the ANN model used to predict the S/G ratio and (S+G) values are shown in the Table 5.4. As seen in the table, based on the weights associated with each of the protein concentrations for the ANN model used to predict the S/G ratio, the S/G ratio is most negatively affected by the variation in concentration of C4H2 and positively affected by the variation of CAld5H1. Similarly, the variation in the total lignin content is negatively affected by the variation in the CCoAOMT3 concentration and positively affected by the changes in the C3H3 concentration. These results are mostly consistent with the findings from chapter 4, where most of the variation in S/G ratio are brought about by the perturbations of C3H3, C4H1. C4H2, CAld5H1 and CAld5H2 proteins. However perturbation of PKMF model suggests that the decrease in composition of PAL, 4CL and HCT enzymes should result in an increase in the S/G ratio. The discrepancies between model is primarily because the ANN model is a purely data driven model with a limited amount of transgenic data and does not take into consideration the various inhibition and activation kinetics. A similar argument can be made for the ANN model to predict the total lignin content, where the total lignin content in most positively affected by C3H3 concentration followed by C4H2, HCT6 and 4CL3 which is in agreement with the PKMF model.

Table 5.4: The weights associated with the inputs to the ANN model to predict the variation in the lignin composition (S/G) and the structure (S+G). The protein concentrations are arranged in ascending order based on the weights.

S:G		(S+G)	
Protein	**Weight**	**Protein**	**Weight**
pC4H2	-0.82	pCCoAOMT3	-0.46
pCCoAOMT2	-0.65	pPAL1	-0.26
pCAD1	-0.45	pCOMT2	-0.26
pC3H3	-0.42	pHCT1	-0.21
pPAL3	-0.39	pCCR2	-0.19
pHCT1	-0.28	pCAld5H1	-0.11
pCAD2	-0.21	pCCoAOMT1	0.00
pCCoAOMT3	-0.20	pCAld5H2	0.08
p4CL5	-0.19	p4CL5	0.14
pPAL4|5	-0.15	pCCoAOMT2	0.15
pPAL1	-0.05	pPAL2	0.15
pC4H1	-0.05	pCAD2	0.20
pPAL2	-0.02	pC4H1	0.21
pCAld5H2	0.13	pCAD1	0.24
pCCoAOMT1	0.16	pPAL3	0.28
pHCT6	0.21	pPAL4|5	0.32
pCOMT2	0.22	p4CL3	0.37
p4CL3	0.32	pHCT6	0.40
pCCR2	0.44	pC4H2	0.42
pCAld5H1	0.46	pC3H3	1.00

5.4 Discussion:

The PKMF model was able to capture the majority of the variation in the S/G ratios for different batches as shown in the Figure 5.6. In some cases, however, the mean S/G value predicted by the PKMF model for certain constructs deviated from the experimentally determined S/G values. The cause for the discrepancies can be analyzed with the help of scatter plots of the enzyme concentration variations from the WT for the specific construct. Figure 5.10 displays the variation of enzyme concentrations relative to their WT levels for the overexpression of CAD2 construct (O67-9-1) in batch 1.

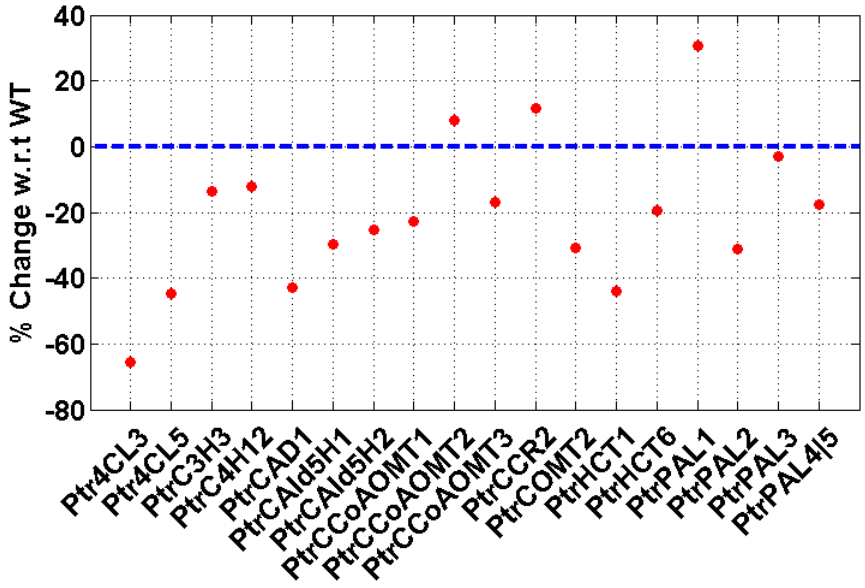

Figure 5.10: Variation of protein concentrations of all the enzymes involved in monolignol biosynthesis as a result of overexpressing PtrCAD2 (O67-9-1). The blue dash line represents the WT level.

In Figure 5.10, the majority of the enzyme concentrations involved in the monolignol biosynthetic pathway were reduced relative to their WT levels during the overexpression of PtrCAD2. Previous results from the analysis of PKMF model (Wang et al, 2014 and Naik et al, 2016(Chapter 3)) have shown that the reduction in the majority of the enzyme concentrations involved in the monolignol biosynthesis resulted in an increase in S/G ratio with the exception of PtrCAld5H and PtrCOMT2.

Similarly for the case of batch 2, the S/G mean predictions from the PKMF model for the construct i8-1-1 (Figure 5.3), which corresponds to the downregulation of PAL1, PAL2, PAL3, PAL4 and PAL5 are higher when compared to the experimentally determined S/G values. The variation in enzyme concentrations relative to WT levels are shown in Figure 5.11. In Figure 5.11, the majority of the proteins concentrations are below

the WT levels with the exception of PtrCAld5H1, PtrCCoAOMT3 and PtrHCT1. Because the downregulation of PtrPALs results in a significant increase in S/G ratio (Wang et al, 2014), the resulting steady state S/G ratio distribution from the PKMF model are higher as compared to the WT levels.

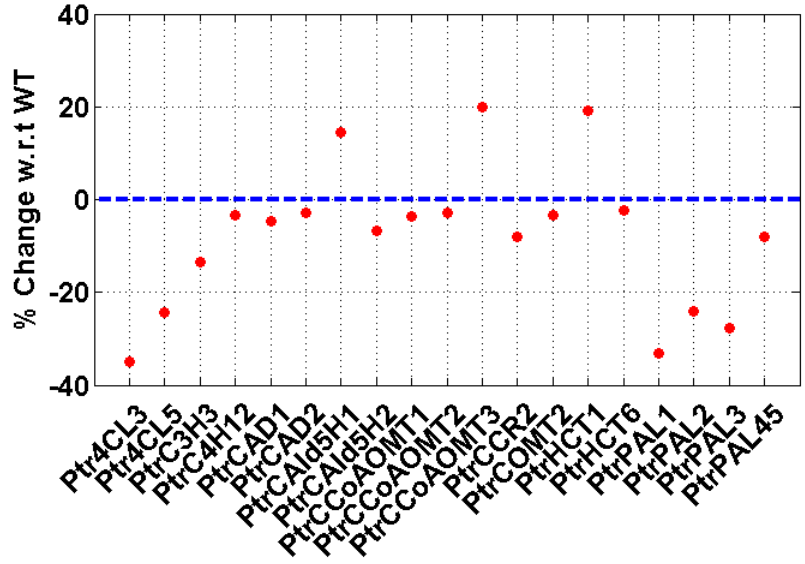

Figure 5.11: Variation of protein concentrations of all the enzymes involved in monolignol biosynthesis as a result of downregulation of PtrPAL enzymes. The blue dash line represents the WT level.

For the batch 5 transgenics, the PKMF predicted S/G values followed the same trend as the experimentally determined S/G values. The PKMF model results for two constructs (a28-H-1 and i29-L-1), however, deviated from the experimental results. For the case of construct a28-H-1, which corresponds to the downregulation of PtrCAld5Hs, the experimental S/G value was 2.14, which was over three times the PKMF model predicted mean S/G of 0.57. The cause for the deviation can be assessed with the help of Figure 5.12. As seen in Figure 5.12, the downregulation of PtrCAld5H1 results in an overall decrease in the levels of PtrCAld5H1 by 100% and PtrCAldH2 by 50%. Because

downregulation of PtrCAld5H results in a decrease in the lignin content as well as S/G ratio, the PKMF model predicts a low S/G value compared to the WT. Similarly for the construct i29-L-1, which corresponds to downregulation of both PtrCAld5H1 and PtrCAld5H2, the predicted S/G value is different from the experimental S/G value.

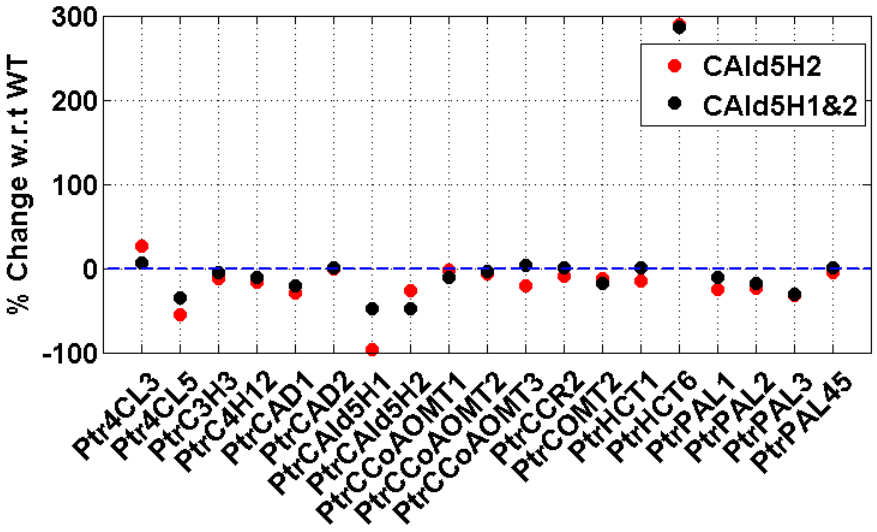

Figure 5.12: Variation of protein concentrations of all the enzymes involved in monolignol biosynthesis as a result of downregulating PtrCAld5H2 and PtrCAld5H 1 and CAld5H2. The blue line represents the WT level.

The discrepancies in some of the predicted S/G values may be attributed to the simplifying assumptions made while developing the model. One of the most important assumptions we made was the measured S/G ratios correspond to the steady state concentrations of sinapyl alcohol and coniferyl alcohol. Another assumption was that because the mechanism of polymerization of monolignols to lignin is unclear, we assumed that there is a 1:1 relationship between the steady state flux of formation of S and G monolignols to their polymerized products. The PKMF model does not

incorporate any information about transcription and assumes that only Ptr4CL forms a multi-protein complex. Studies on lignin biosynthesis formation in *P. trichocharpa* has revealed that other multi enzyme complex exists such as between PtrHCT1 and PtrHCT6 (Lin et al, 2015) and between PtrC3H and PtrC4H (Chen et al, 2012). In addition, the PKMF model does not incorporate feedback mechanism between the metabolites and the transcription factors, which are known to regulate the levels of metabolites involved in the monolignol biosynthesis pathway (Tamagnone et al, 1998). Future models that include these additional characteristics will need to be evaluated to determine if they explain the discrepancies noted in the current study.

5.5 Conclusions:

In this paper, we utilized experimentally measured S/G ratio data to validate a previously developed PKMF model. The PKMF model was able to capture the majority of the experimental S/G ratios as a result of changes in enzyme concentrations. A linear fit of the PKMF model S/G values against the observed S/G values revealed a slope of 1 and the R^2 value of 0.7 indicating that the model predictions are in close agreement with the observed data. For a few transgenic constructs, the S/G values predicted by the PKMF value were distributed over a wide range of ratios. The wide variation in S/G ratio was primarily caused by low concentrations of Ptr4CL and PtrHCT enzymes. For constructs involving downregulation of enzymes belonging to PtrCAld5H family, the PKMF model underestimated the S/G ratio compared to the measured S/G ratio. This particular model discrepancy was primarily due to the inhibition of PtrCAld5H enzymes,

which results in the accumulation of aldehydes and alcohols that in turn are incorporated into the lignin structure. The PKMF model assumes a 1:1 relationship between the rate of formation of lignin polymer and monolignol concentrations. Overall, the PKMF model was able to predict a majority of the variations in the S/G ratios. The model would serve as a valuable tool to predict changes in lignin composition due to changes in the protein concentrations. The model can then be used to engineer lignin content and composition in plants for its use in pulp and paper and biofuels. We also developed an ANN model to predict the S/G ratio and the total lignin content as a function of protein concentrations. The model was able predict 84% of the variation in the S/G ratios and 77% of the variation in the total lignin content. Although such models are advantageous in predicting the changes in S/G ratios and total lignin content as a function of changes in protein concentrations, the model does not explain the changes in steady state flux distribution or the changes in the metabolite concentrations. The ANN model is still a potential tool for researchers in predicting the different S/G ratios and total lignin content resulting from transgenic perturbations and does not rely on the information regarding the reaction rate or enzyme kinetics.

References:

Affourtit C, Krab K and Moore AL. (2001). Control of plant mitochondrial respiration. Biochim. Biophys. Acta. 1504, 58–69.

Agrafiotis DK, Cede˜no W and Lobanov VS. (2002). On the use of neural network ensembles in QSAR and QSPR, J. Chem. Inf. Comput. Sci., 42, 903–911.

Aluru M, Xu Y, Guo R, Wang Z, Li S, White W, Wang K and Rodermel S. (2008). Generation of transgenic maize with enhanced provitamin A content. J. Exp. Bot, 59, 3551–3562.

Amani A, York P, Chrystyn H, Clark BJ, DO DQ. (2008). Determination of factors controlling the particle size in nanoemulsions using artificial neural networks. Eur J Pharm Sci. 35(1-2):42–51.

Atanassova R, Favet N, Martz F, Chabbert B, Tollier MT, Monties B, Fritig B and Legrand M. (1995) Altered lignin composition in transgenic tobacco expression O-methyltransferase sequence in sense and antisense orientation. Plant J., 8, 465–477.

Bakker BM, Michels PAM, Opperdoes FR, Westerhoff HV. (1999) What controls glycolysis in bloodstream form Trypanosoma brucei. Journal of Biological Chemistry 274 (21), 14551–14559.

Basri M, Rahman A, Ebrahimpour A, Salleh AB, Gunawan ER, Rahmad MB. (2007). Comparison of estimation capabilities of response surface methodology (RSM) with artificial neural network (ANN) in lipase catalyzed synthesis of palm-based wax ester. BMC Biotechnology. 7:53–66.

Basu JK, Bhattacharya D and Kim T. (2010). Use of artificial neural network in pattern recognition, International Journal of Software Engineering and its Applications, 4, 23–33.

Chen F, Dixon RA. (2007). Lignin modification improves fermentable sugar yields for biofuel production, Nature Biotechnology 25, 759.

Choe W, Ersoy OK and Bina M. (2010). Neural network schemes for detecting rare events in human genomic DNA, Bioinformatics, 16, 1062–1072.

Colon AJM, Morgan JA, Dudareva N and Rhodes D. (2009). Application of dynamic flux analysis in plant metabolic networks. In Plant Metabolic Networks (Schwender, J., ed). New York: Springer, pp. 285–305.

Daae EB, Dunhill P, Mitsky TA, Padgette SR, Taylor NB, Valentin HE and Gruys KJ. (1999). Metabolic modeling as a tool for evaluating polyhydroxyalkanoate copolymer production in plants. Metab. Eng. 1, 243–254.

Dauwe R, Morreel K, Goeminne G. et al. (2007). Molecular phenotyping of lignin-modified tobacco reveals associated changes in cell-wall metabolism, primary metabolism, stress metabolism and photorespiration. Plant J. 52, 263–285.

Devillers J. (1996). Neural Networks in QSAR and Drug Design, Academic Press, London.

Dwivedi UN, Campbell WH, Yu J, Datla RS, Bugos RC, Chiang VL and Podila GK. (1994). Modification of lignin biosynthesis in transgenic Nicotiana through expression of an antisense O-methyltransferase gene from Populus. Plant Mol. Biol., 26, 61–71.

Fridlyand LE and Scheibe R. (1999) Regulation of the Calvin cycle for CO_2 fixation as an example for general control mechanisms in metabolic cycles. Biosystems, 51, 79–93.

Fridlyand LE, Backhausen JE and Scheibe R. (1998) Flux control of the malate valve in leaf cells. Arch. Biochem. Biophys. , 349, 290–298.

Hagan MT and Menhaj M (1999)., Training feed-forward networks with the Marquardt algorithm, IEEE Transactions on Neural Networks, Vol. 5, No. 6, pp. 989–993.

Halpin C, Knight ME, Foxon GA, Campbell MM, Boudet AM, Boon JJ, Chabbert B, Tollier M and Schuch W. (1994). Manipulation of lignin quality by down regulation of cinnamyl alcohol dehydrogenase. Plant J., 6, 339–350.

Hamelinck CN, Hooijdonk G, Faaij APC. (2005). Ethanol from lignocellulosic biomass: techno-economic performance in short, middle and long-term. Biomass and Bioenergy 28, 384.

Heinzle E, Matsuda F, Miyagawa H, Wakasa K and Nishioka T. (2007). Estimation of metabolic fluxes, expression levels and metabolite dynamics of a secondary metabolic pathway in potato using label pulse-feeding experiments combined with kinetic network modeling and simulation. Plant J. 50, 176–187.

Izadifar M, ZolghadriJahromi M. (2007). Application of genetic algorithm for optimization of vegetable oil hydrogenation process. J Food Eng. 78(1):1–8.

Jalali-Heravi M. (2009). Neural network in analytical chemistry, in Artificial Neural Networks: Methods and Applications (ed. D. J. Livingstone), Methods in Molecular Biology Series, Vol. 458, Humana Press, Totowa, NJ, 78–118.

Jorjani E, Chehreh CS, Mesroghli SH. (2008). Application of artificial neural networks to predict chemical desulfurization of tabas coal. Fuel. 87(12):2727–2734.

Kasiri MB, Aleboyeh H, Aleboyeh A. (2008). Modeling and optimization of heterogeneous photo-fenton process with response surface methodology and artificial neural networks. Environ Sci Technol. 42(21):7970–7975.

Kebeish R, Niessen M, Thiruveedhi K, Bari R, Hirsch HJ, Rosenkranz R, Stabler N, Schonfeld B, Kreuzaler F and Peterhansel C. (2007). Chloroplastic photorespiratory bypass increases photosynthesis and biomass production in Arabidopsis thaliana. Nat. Biotechnol, 25, 593–599.

Kholodenko B, Yaffe MB and Kolch W. (2012). Computational approaches for analyzing information flow in biological networks. Sci. Signal. 5.

Kitano H. (2007). Towards a theory of biological robustness. Mol Syst Biol 3:137.
Klipp E, Liebermeister W, Helbig A, Kowald A, Schaber J. (2007). Systems Biology standards the community speaks. Nature Biotechnology 25 (4), 390–391.

Krab K. (1995). Kinetic and regulatory aspects of the function of the alternative oxidase in plant respiration. J. Bioenerg. Biomembr. 27,387–396.

Lancashire LJ, Powe DG, Reis-Filho JS, Rakha E, Lemetre C, Weigelt B, Abdel-Fatah TM, Green AR, Mukta R and Blamey R. et al. (2010). A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks, Breast Cancer Research and Treatment, 120, 83–93.

Lee Y and Voit EO. (2010). Mathematical modeling of monolignol biosynthesis in Populus xylem. Mathematical Biosciences 228, 78-89.

Libourel, I.G. and Shachar-Hill, Y. (2008). Metabolic flux analysis in plants: from intelligent design to rational engineering. Annu. Rev. Plant Biol. 59, 625–650.
Marquardt D. (1963).An Algorithm for Least-Squares Estimation of Nonlinear Parameters, SIAM Journal on Applied Mathematics, Vol. 11, No. 2, pp. 431–441.

McNeil SD, Rhodes D, Russell BL, Nuccio ML, Shachar-Hill Y and Hanson AD. (2000b). Metabolic modeling identifies key constraints on an engineered glycine betaine synthesis pathway in tobacco. Plant Physiol. 124, 153–162.

McNeil SD, Nuccio ML, Rhodes D, Shachar-Hill Y and Hanson AD. (2000a). Radiotracer and computer modeling evidence that phosphor-base methylation is the main route of choline synthesis in tobacco. Plant Physiol. 123, 371–380.

McNeil SD, Rhodes D, Russell BL, Nuccio ML, Shachar-Hill Y and Hanson AD. (2000b). Metabolic modeling identifies key constraints on an engineered glycine betaine synthesis pathway in tobacco. Plant Physiol. 124, 153–162.

Morgan JA and Rhodes D. (2002). Mathematical modeling of plant metabolic pathways. Metab. Eng. 4, 80–89.

Napier JA. (2007). The production of unusual fatty acids in transgenic plants. Annu. Rev. Plant Biol. 58, 295–319.

Nuccio ML, McNeil SD, Ziemak MJ, Hanson AD, Jain RK and Selvaraj G. (2000) Choline import into chloroplasts limits glycine betaine synthesis in tobacco: analysis of plants engineered with a chloroplastic or a cytosolic pathway. Metab. Eng. 2,300–311.

Pearcy RW, Gross LJ and He D (1997). An improved dynamic model of photosynthesis for estimation of carbon gain in sunfleck light regimes. Plant Cell Environ. 20, 411–424.

Poolman MG, Fell DA and Thomas S. (2000). Modelling photosynthesis and its control. J. Exp. Bot. 51, 319–328.

Ragauskas AK, Williams CK, Davison BH, Britovsek G, Cairney J, Eckert CA, Frederick WJ, Hallett JP, Leak DJ, Liotta CL, Mielenz JR, Murphy R, Templer R, Tschaplinski R. (2006). The path forward for biofuels and biomaterials, Science 311, 484.

Rios-Estepa R, Turner GW, Lee JM, Croteau RB and Lange BM. (2008). A systems biology approach identifies the biochemical mechanisms regulating monoterpenoid essential oil composition in peppermint. Proc. Natl Acad. Sci. USA, 105, 2818–2823.

Rohwer JM. (2012). Kinetic modelling of plant metabolic pathways. Journal of Experimental Botany, 63, 2275-2292.

Schwender J. (2009). Kinetic properties of metabolic networks. In Plant Metabolic Networks (Schwender, J., ed). New York: Springer, pp. 307–322.

Sewalt V, Ni W, Blount JW, Jung HG, Masoud SA, Howles PA, Lamb C. and Dixon RA. (1994). Reduced lignin content and altered lignin composition in transgenic tobacco down-regulated in expression of l-phenylalanine ammonia-lyase or cinnamate 4-hydroxylase. Plant Physiol, 115, 41–50.

Somerville C and Somerville S. (1999). Plant functional genomics, Science 285, 380.
Song X, Mitnitski A, Macknight C, Rockwood K. (2004). Assessment of individual risk of death using self-report data: an artificial neural network compared with a frailty index. J Am Geriatr Soc. 52(7):1180–1184.

Sticklen M. (2006). Plant genetic engineering to improve biomass characteristics for biofuels. Current Opinion in Biotechnology 17,315.

Thomas S, Mooney PJF, Burrell MM and Fell DA. (1997a). Finite change analysis of glycolytic intermediates in tuber tissue of lines of transgenic potato (Solanum tuberosum) overexpression phophofructokinase. Biochem. J. 322, 111–117.

Thomas S, Mooney PJF, Burrell MM and Fell DA. (1997b). Metabolic control analysis of glycolysis in tuber tissue of potato (Solanum tuberosum): explanation for the low control coefficient of phosphofructokinase over respiratory flux. Biochem. J. 322,119–127.

Urda D, Subirats J, Franco L and Jerez JM. (2010). Constructive neural networks to predict breast cancer outcome by using gene expression profiles, in Trends in Applied Intelligent Systems: 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2010, Cordoba, Spain, June 1-4, 2010, Proceedings, Part I , Lecture Notes in Computer Science, 6096, Springer-Verlag, Berlin-Heidelberg, 317–326.

Van Doorsselaere J, Baucher M, Chognot E et al. (1995). A novel lignin in poplar trees with a reduced caffeic acid/5-hydroxyferulic acid O-methyltransferase activity. Plant J., 8, 855–864.

Von Caemmerer S. (2000). Biochemical Models of Photosynthesis. Collingwood, Victoria, Australia: Commonwealth Scientific and Industrial Research Organization.
Wang L, Yang B, Wang R, DU X. (2008). Extraction of pepsin-soluble collagen from grass carp (Ctenopharyngodonidella) skin using an artificial neural network. Food Chem. 111(3):683–686.

Wu S, Schalk M, Clark A, Miles RB, Coates R and Chappell J. (2006). Redirection of cytosolic or plastidic isoprenoid precursors elevates terpene production in plants. Nat. Biotech. 24, 1359–1361.

Ye X, Al-Babili S, Kloti A, Zhang J, Lucca P, Beyer P and Potrykus I. (2000). Engineering the provitamin A (beta-carotene) biosynthetic pathway into (carotenoid-free) rice endosperm. Science, 287, 303–305.

CHAPTER VI

A JOINT REGULATORY METABOLIC MODEL FOR UNDERSTANDING MONOLIGNOL BIOSYNTHESIS IN *POPULUS TRICHOCARPA*

Abstract:

Lignin is an important component of wood and supports the structure of the secondary cell wall by forming covalent and non-covalent bonds to cellulose and hemicelluloses. Lignin is a complex phenylpropanoid polymer, which is not degraded by chemical, physical or biological means. The binding to cellulose and hemicellulose makes lignin difficult to extract and hinders the utilization of biomass from the plant. The transcription factors that control regulation of monolignol biosynthesis are key to understand and manipulate the monolignol biosynthesis genes that are important for the modification of the secondary cell walls. Gene regulation and metabolic reactions are two primary activities that regulate the biosynthesis process. Although significant prior research has been dedicated to study each system, the coupling between gene regulation and metabolism is less well understood. To overcome this knowledge gap, we developed a joint model that incorporates gene regulation and metabolic reactions. We integrated regulatory and metabolic networks using the Hill equation to quantify pairwise interactions between Transcription Factors (TFs), transcripts, and proteins concentrations. The interactions between TFs and transcripts were obtained from analyzing the RNA-Seq data. The transcription factors that control regulation of monolignol biosynthesis are key to manipulating the monolignol biosynthesis gene expressions that modify secondary cell wall structures. We quantified the effect of downregulating the TFs on the lignin composition using a previously developed mass action kinetics model. The results suggest that the MYB221 TFs play a significant role in the regulation of the lignin biosynthesis as well the composition and structure.

6.1 Introduction:

High throughput post genomic technologies such as microarray and proteomics analyses have provided powerful tools to study gene expression and regulation (Horak and Snyder 2002; Smith et al. 2002). In order to build a predictive model of complex biological systems, all aspects of the systems need to be quantified or understood. Gene expression is achieved through a multi-step process involving transcription, translation, and protein synthesis. Understanding a biochemical pathway requires the knowledge of

quantitative relationships between genes, transcripts, proteins, and metabolites. Previous studies based on the analyses of gene expression data have shown that the correlation between mRNA and protein abundance was typically moderate (Futcher et al., 1999; Gygi et al., 1999; Ideker et al., 2001; Greenbaum et al., 2003; Washburn et al., 2003). It has been proposed that the three primary reasons for the weak correlation between transcript abundance and protein expression levels are: (i) translational regulation, (ii) differences in *in vivo* half-lives of proteins, and (iii) experimental error that includes the differences in experimental test conditions (Greenbaum et al., 2003; Beyer et al., 2004). The steady state concentration of protein depends on the post transcriptional, translational, and protein degradation (Schwanhäusser et al., 2011).

Lignin is an abundant polymer that is synthesized in the secondary cell walls of all vascular plants. It enables conduction of water through the stem, mechanical strength and protects against pathogens. In addition, lignin hinders the utilization of the cellulosic cell walls of plants in the production of pulp and paper and as forage. The removal of lignin for the above-mentioned uses typically requires chemical pretreatment (Chiang, 2002 and Ragauskas et al., 2006). Lignin is a hydrophobic polymer that is formed from three hydroxycinnamyl alcohols, p-coumaryl, coniferyl and sinapyl alcohol (pCA, CA and SA, respectively), which differ in their methoxylation degree on the aromatic ring (Koutaniemi, 2007). They give rise to p-hydroxyphenyl (H), guaiacyl (G) and syringyl (S) units in lignin. Gymnosperm lignin is composed mainly of G units, whereas in angiosperm lignin, both G and S units predominate. H units are a minor component in both gymnosperms and angiosperms (Higuchi, 1997). Monolignols are biosynthesized through

the phenylpropanoid pathway. A total of 21 enzymes catalyze various reactants into products through a series of reactions (Figure 6.1). The first step in lignin biosynthesis is the deamination of *phenylalanine* to *cinnamic acid* by PtrPAL (Higuchi, 1997). The next step is the conversion of cinnamic acid to *P-hydroxy cinnamic acid* catalyzed by PtrC4H. *P-hydroxy cinnamic acid* is the last common precursor for all monolignols (Dixon et al., 2006). In *P. trichocharpa*, there are two PtrC4H present, PtrC4H1 and PtrC4H2, respectively (Lu et al., 2006). The next step is the formation of CoA esters from hydroxy cinnamic acid, caffeic acid, ferulic acid, 5-hydroxy ferulic acid and sinapic acid in the presence of Ptr4CL (Ptr4CL3 and Ptr4CL5). PtrHCT catalyzes the conversion of *P-Coumaryl CoA* to *shikimic acid*. The C3H protein converts *P-coumaryl shikimic acid* to *caffeoyl shikimic acid*. In the next set of reactions, CCoAOMT catalyzes the methylation of *caffeoyl CoA* to *feruloyl CoA*, eventually leading to the biosynthesis of *coniferyl alcohol* and *sinapyl alcohol*, through the reduction of the CoA derivatives to the corresponding aldehydes and alcohols.

Cell wall formation is regulated primarily by two main group of transcription factors. The first group is the Mining Yeast Binding site (MYB) transcription factors. MYB protein is a direct target of a NAC transcription factor controlling regulation of secondary cell wall biosynthesis (Nakano et al., 2010). The second group is the LIM group, which is a transcription factor that consists of two LIM domains in its protein (Arnaud et al., 2007). LIM1 regulates the monolignol biosynthesis in tobacco (Kawaoka et al., 2001). LIM1 downregulation in eucalyptus results in the reduction in expression levels of PAL, C4H, and 4CL (Kawaoka et al., 2006). In *P. trichocarpa,* however, the monolignol biosynthesis

pathway is regulated by 12 LIM transcription factors (Arnaud et al.,2012). In differentiating xylem from tension wood and normal differentiating xylem tissue, the levels of LIM transcription factors are highly expressed (Arnaud et al., 2012). The MYB transcription factor family is characterized by three domains, a DNA binding domain, DNA activation domain, and a repression domain. In *Arabidopsis*, downregulation of MYB4 results in overexpression of C4H and decreased expression levels in CCoAOMT (Jin et al., 2000). Overexpression of MYB4 results in an increased expression of CCoAOMT and reduction in levels of 4CL and C4H. In *P. trichocarpa,* MYB156 and MYB221 are homologs of MYB4. In Arabidopsis, it has been reported that MYB52 controls the biosynthesis of the secondary cell wall (Zhong et al., 2008). Similarly MYB90 and MYB167 are homologs of MYB52. Studies on tobacco has shown that the upregulation of MYB transcription factors result in a decrease in the total lignin content (Tamagnone et al, 1998).

Clearly there are a number of studies that have shown different phenotypic responses to changes in MYB and LIM transcription factor groups. Yet no study has used modeling to uncover why these responses occur from analyzing the regulatory pathway in isolation. This study sets out to use a combine regulatory-metabolic pathway model that seeks to trace the regulation of the lignin biosynthesis pathway from TF to overall lignin content and structure. The model will help elucidate how enzyme protein concentrations, metabolite concentrations, and metabolite pathway fluxes change to produce specific phenotypes

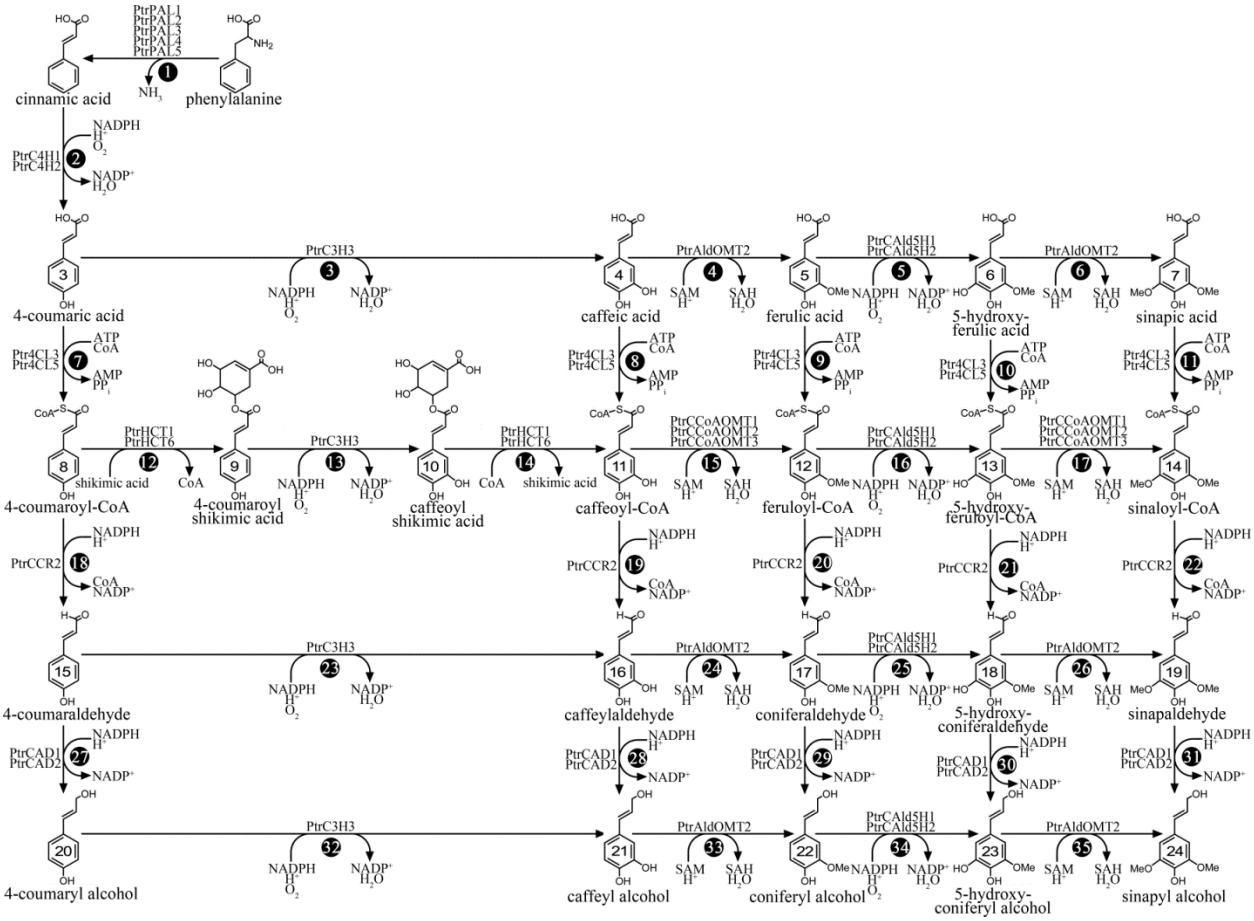

Figure 6.1: The monolignol biosynthetic pathway showing the various reactions and the enzymes that catalyze the reactions. The circled number indicates the reaction flux of individual reactions.

6.2 Materials and Methods:

6.2.1 Modeling Approach for Gene Regulatory Network:

The protein concentrations of Transcription Factors (TF), Transcript Abundance (TA) and absolute protein quantity (p) were estimated from RNA-Seq analysis (Sermsawat et al, 2015). A curve fitting procedure was performed to obtain a mathematical description of the relationship between the TF - transcript abundance and Transcript abundance-protein quantity. We used a Hill equation and non-linear regression

to fit the experimental data. The advantage of using the Hill equation is in its simplicity in biochemical modeling when the kinetic parameters are not readily available.

6.2.2 Parameter Estimation:

The parameters of the Hill equation were estimated using the experimental data available from RNA-Seq experiments, by performing an optimization procedure. The goal of the optimization problem was to minimize the objective function MSE (Mean Squared Error) shown in Eq. 6.1, where K is a vector of 3 parameters that will be estimated and n, is the number of data points. Because the objective function is highly non-linear; we used a hybrid optimization method to minimize the objective function. The hybrid method uses a gradient-based approach in conjunction with a heuristic based method (Ebessen et al., 2007). We used Fmincon (Mathworks. Optimization Toolbox, Version 3, User's Guide 2007) for gradient based optimization and Particle Swarm Optimization (PSO) for heuristic optimization (Kennedy and Eberhart, 1995). The hybrid optimization procedure was repeated 100 times with random initial conditions over the range n = [0.01-5], p_{max}= [0.01-150], k = [0-100]. A 95% confidence interval for each parameter was also computed.

$$\min_{k \in R} \left(\frac{(\mathrm{y}_{\exp} - y_{\mathrm{model}}(K))^2}{n} \right) \tag{6.1}$$

6.2.3 Modeling Approach for Metabolic Network

The metabolic flux for all the reactions involved in the monolignol biosynthesis pathway were quantified using Michaelis Menten kinetics. The Michaelis Menten

activation and inhibition parameters were experimentally identified and obtained from Wang et al (2014). Utilizing the conservation of mass principle, the rate of change of metabolite concentrations was then expressed as a function of the reaction fluxes resulting in the production and consumption of individual metabolites. The resulting set of Ordinary Differential Equations (ODE) are solved to steady state using MATLAB® ode15s function. The steady state metabolite concentrations are then plugged back into the reaction flux equations and the resulting flux for the formation of S and G monolignols were calculated.

6.2.4 Monte Carlo Simulation:

Because the overall goal of the predictive model is to assess the role of perturbing various TFs on the lignin composition, we performed a Monte Carlo simulation to quantify the effect of perturbing the TFs on the S/G ratio. The TF protein concentrations were varied from 0.5 – 1.5 times the WT concentration, which was obtained from the RNA-Seq data. We employed a Latin Hypercube Sampling (LHS) procedure to randomly sample from the distribution of the TF concentrations. The TF concentrations were assumed to be uniformly distributed. Using LHS procedure, 1000 different concentrations of TFs were sampled and for each iteration, the model was simulated to steady state and the resulting flux of S and G monolignol formation was calculated. The input flux (V_1) to the metabolic network was calculated such that, at WT TF levels, the S/G ratio corresponds to 2:1.

6.2.5 Gene Network Inference

The levels of gene products that influence cellular function are controlled via Gene Regulatory Networks (GRNs). Several methods have been used to reconstruct GRNs in systems biology including: Differential Equations (Novak et al, 1995; Goss and Peccoud, 1998, Chen et al, 1999; Kyoda and Kitano et al 1999) Stochastic Petri Net (Goss and Peccoud, 1999), Boolean Network (Liang et al, 1999), regression method (Gardner et al, 2003), linear programming (Wang et al, 2006) and Bayesian Network (Friedman, 2000 and Marcbach et al, 2010). Each method has its own advantages and limitations and it has been shown that inferring GRNs from gene expression data is a Non deterministic Polynomial time (NP)-hard problem. As a consequence, there is still a great potential to improve the current approaches for the inference of GRNs (Marcbach et al, 2010). Inferring GRNs from gene expression data based on Bayesian Network models has been previously explored (Friedman et al, 2000 and Sprites et al, 2000). Structural learning of Bayesian networks is also an NP-hard problem. Hence, there exists many methods for structural learning. There are basically three methods for learning the structure of Bayesian networks from data; 1) constraint-based methods (Sprites et al, 2000; Pearl, 2000 and Zhang et al, 2012), 2) score and search methods (Imoto et al, 2002; De Campos, 2006 and Faulkner, 2007) and 3) Hybrid methods (Acid and De Campos, 2001 and Tsamardinos et al, 2006). The constraint-based methods uses Conditional Independent (CI) to determine the degree of dependency between random variables. The score and search methods identifies a set of candidate networks and returns the network with the highest –log likelihood score. The hybrid method is a combination of these two methods. In this study, we present a hybrid method, which is a

combination of Peter and Clarke (PC) -algorithm-based approaches such as a PC algorithm based on Conditional Mutual Information (PCA-CMI) (Zhang et al, 2012) and score and search method (Hill Climbing (HC) algorithm based on MIT (Mutual Information Tests) score (De Campos, 2006)).

The PC algorithm tries to find the causal relationships between random variables in a network. It assumes a Bayesian causal network model and it makes use of valid statistical testing to produce a Directed Acyclic Graph (DAG) as output (Spirtes and Glymour, 1991). It is comprised of three steps. In the first step, it applies the conditional independent test to discover relationships between variables. In the other steps, it tries to orientate these relationships without creating cyclic structures. More information on the PC algorithm can be found in Harris and Drton (2013). The gene network inference was performed using the package *pacalg* in R. The following constraints were applied on the interactions (1) Edges can be present between TFs and Transcript Abundance (TA)'s but not the other way round. (2) Each TA edge exits only with its product (ex t4CL3 -> Ptr4CL3) (3) No feedback interactions between proteins and TFs and TAs were allowed.

6.2.6 Discrete to Continuous Boolean Transformation:

As described previously, the Boolean approach has been used significantly in recent years to build GRNs, based on the interactions between the various components within the network. Assuming that a fully connected network (gene or metabolic) has been developed, the first step is to define a discrete Boolean function for each interaction and then convert it into a continuous function using a Hill equation

approximation for each component in the network. Hill equation is a three-parameter equation based on a non-linear linear relationship between the variables. The Hill equation for a single variable can be defined as

$$y = \frac{y_{\max} x^n}{x^n + k^n} \tag{6.2}$$

where, y_{max} is the maximum concentration of y, n is the measure of cooperativity between y and x, k is the association constant. The methodology is shown in a simple network in Figure 6.2 showing interaction of 3 genes A, B and C. In Figure 6.2, gene C is activated by A and B. The Boolean function for C in terms of A and B is shown in Eq. (6.3) and the discrete Boolean function can be then converted into a continuous function using a hill cube approximation (Wittman et al., 2008) shown in Equation (6.4).

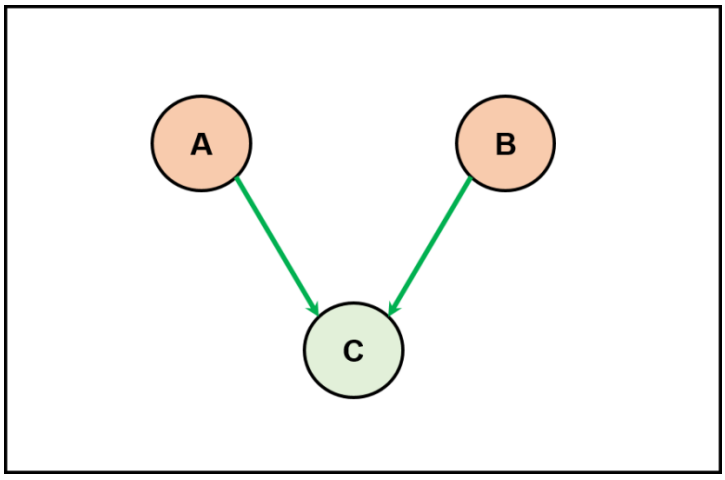

Figure 6.2: Depiction of a toy network that includes three genes A, B and C, with A and B influencing C

Discrete Boolean Logic C = $f(A \cap B)$ (6.3)

Continuous Function

$$C = C_{\max}\left[\frac{A^{n_A}}{A^{n_A} + k^{n_A}}\right]\left[\frac{B^{n_B}}{B^{n_B} + k^{n_B}}\right] \quad (6.4)$$

The above approach was used to quantify every pairwise relationship between the TFs and TA's as well as the relationship between TA and proteins.

6.3 Results:

Although several of the TFs responsible for regulating the lignin biosynthesis have been identified (Zhong et al, 2009), their effect on lignin biosynthesis has not been quantified. Researchers have shown that tMYB002 and tMYB021 are MYB homologs in the *P. trichocarpa* genome (Li et al, 2012). The variation of tMYB021 as a function of tMYB002 is shown in Figure 6.3. In Figure 6.3, a strong nonlinear relationship is displayed between tMYB021 and tMYB002. As seen in the Figure, the levels of tMYB021 increases as the level of tMYB002 increases. Similarly the interactions between other TFs were quantified and the hill parameters associated with the pairwise interactions are shown in Table 6.S.1

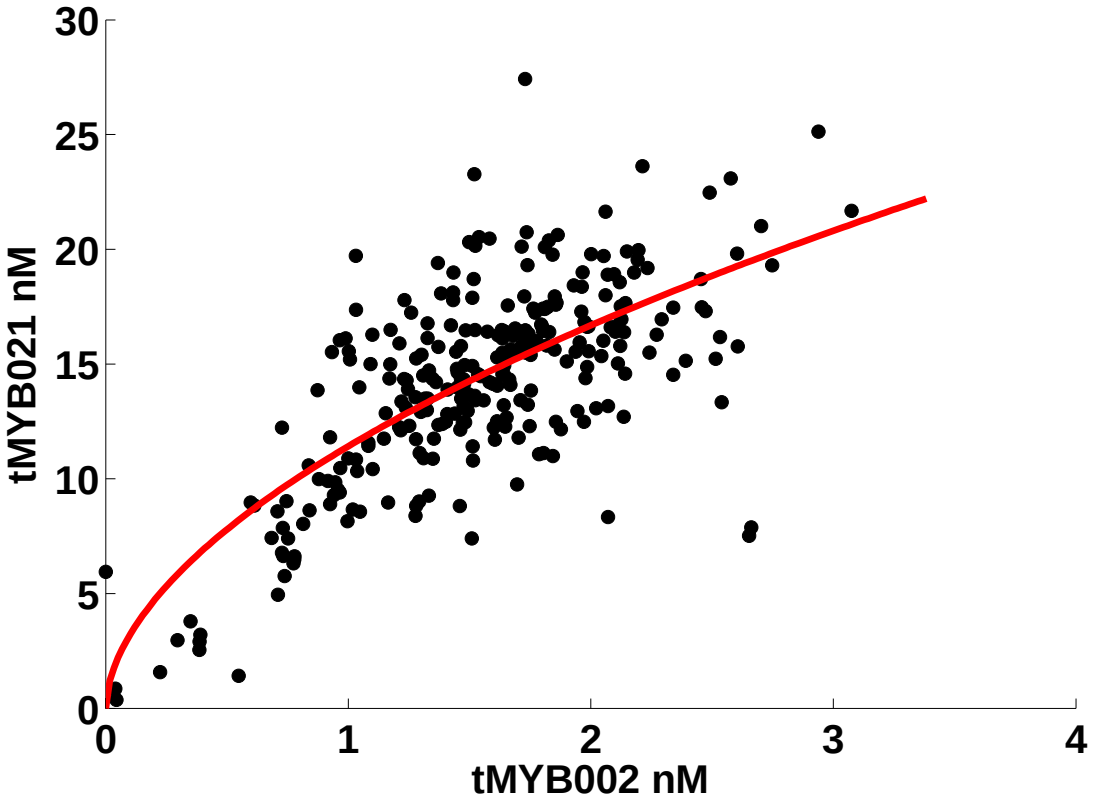

Figure 6.3: Variation of tMYB021 as a function of tMYB021. The red line shows the variation of tMYB021 predicted by Hill equation and the dots show the experimentally determined tMYB021 levels.

Using the hill equation, we were able to quantify the relationship between the absolute protein concentration and transcript abundance. The variation of some of the key proteins as a function of their respective transcript abundance will be discussed in this section. Figure 6.4 displays the variation of Ptr4CL3 as a function of its transcript abundance. The variation of the Ptr4CL3 with the transcript t4CL3 follows a sigmoidal shaped curve, where at low levels of t4CL3 the concentration of Ptr4CL3 is also low, as the concentration of t4CL3 is increased, Ptr4CL3 levels rises linearly and then levels off for very high levels of t4CL3. Since the data used to fit the relationship between Ptr4CL3 and Ptr4CL5 was from all the constructs, some of the Ptr4CL3 concentrations

may be increased due to the indirect effect of perturbations of other proteins in the pathway.

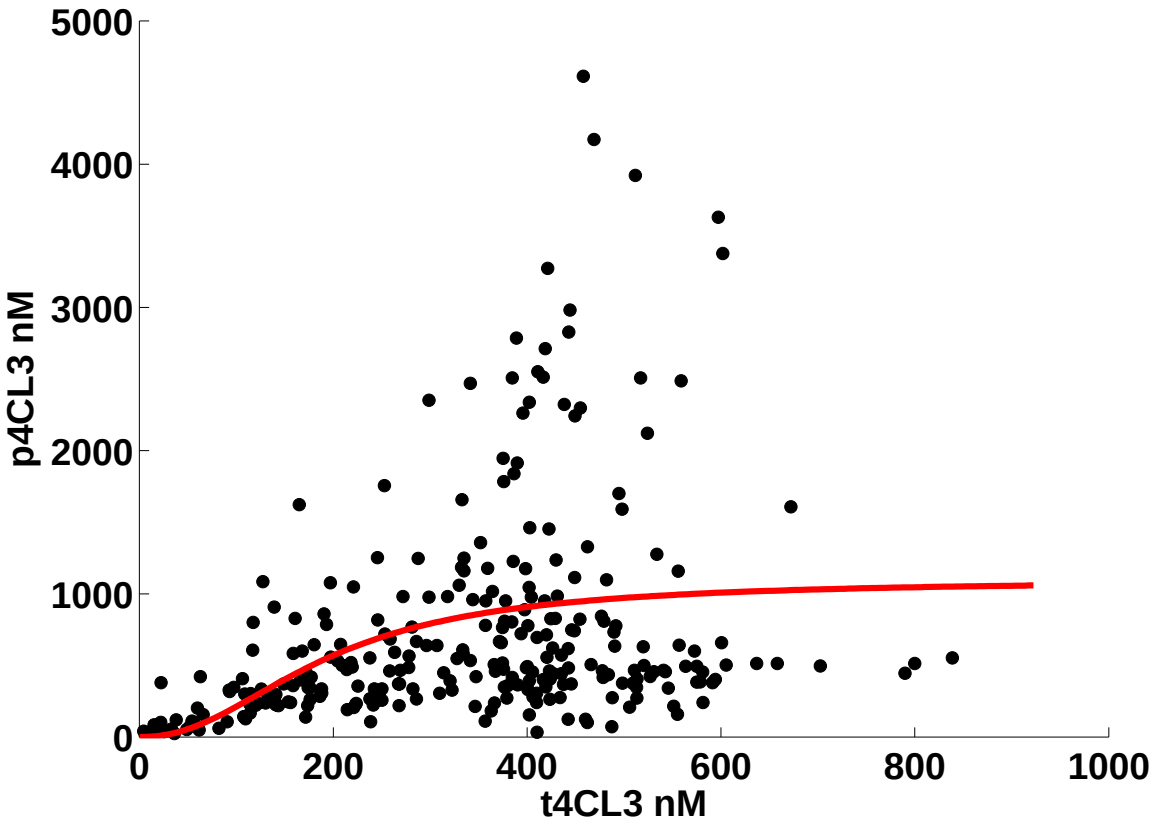

Figure 6.4: Variation of Ptr4CL3 as a function of t4CL3. The red line shows the variation of Ptr4CL3 predicted by Hill equation and the dots show the experimentally determined Ptr4CL3 levels.

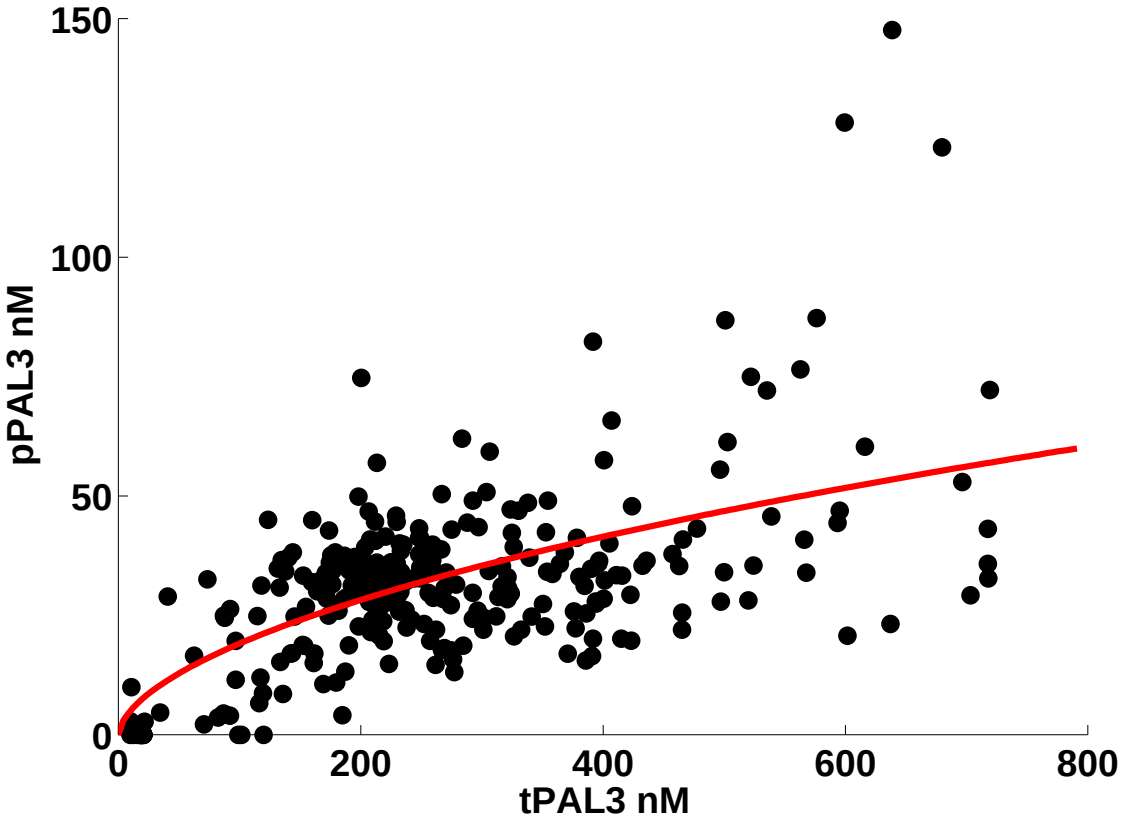

Figure 6.5: Variation of PtrPAL3 as a function of tPAL3. The red line shows the variation of PtrPAL3 predicted by Hill equation and the dots show the experimentally determined PtrPAL3 levels.

The variation of protein abundance as a function of transcript abundance for PAL is shown in Figure 6.5. PAL catalyzes the first step in the monolignol biosynthesis pathway. The lignin content and composition is a strong function of the activity of PtrPAL genes. The PtrPAL genes are functionally redundant and xylem specific (Shi et al., 2010). The data for the protein abundance as a function of transcript abundance shows that a strong correlation exits at lower transcript abundance and as it increases to its wildtype concentration, the protein quantity tends to level off. From this result, we can infer that PtrPAL, being a crucial gene in monolignol biosynthesis, is robust to small perturbations.

These results are consistent with the results of perturbation analyses performed on PtrPAL enzymes in metabolic network, which required a very high level of down regulation of PtrPAL enzyme concentration to observe changes in the lignin content and composition (Sewalt et al., 1977; Elkind et al., 1990; Bate at al., 1994 and Wang et al, 2014).

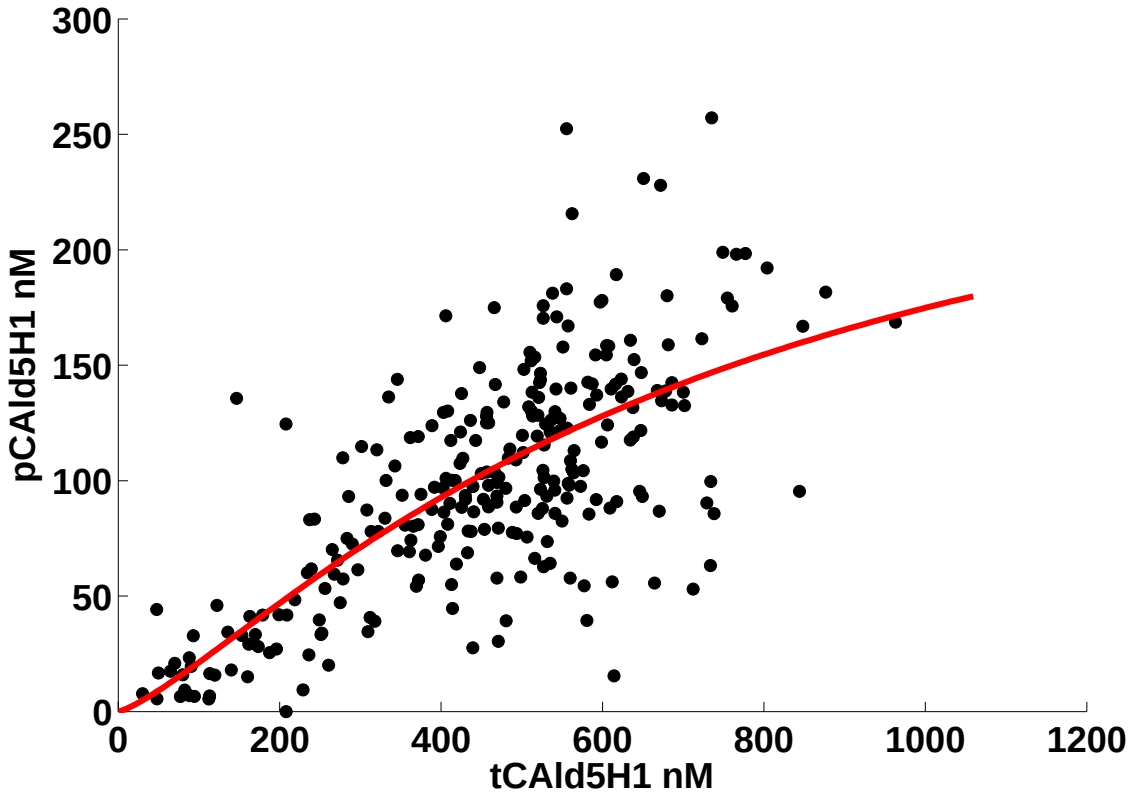

Figure 6.6: Variation of PtrCAld5H1 as a function of tCAld5H1. The red line shows the variation of PtrCAld5H1 predicted by Hill equation and the dots show the experimentally determined PtrCAld5H1 levels.

The variation in PtrCAld5H1 concentrations as a function of tCAld5H1 is shown in Figure 6.6. The concertation of PtrCAld5H varies linearly as a function of tCAld5H1 at

low concentrations of tCAld5H1. As the concentration of tCAld5H1 increases, the concentration of pCAld5H1 levels off.

In poplar, there are only a few studies that provided information on how transcription factors regulate the monolignol pathway (Sermsawat, 2015). The transgenic plants that down regulate the monolignol pathway can be a great resource to study the genes that control the monolignol pathway. The whole transcriptome can reveal the relationship of all transcription factors and the monolignol enzymes. The interactions between the TFs, transcripts, and the proteins involved in the monolignol biosynthesis pathway is shown in Figure 6.7. Since the roles of many specific TFs on the lignin biosynthesis pathway are not known, we used only the known MYB and LIM as the TFs in the network. The results suggest that MYB090 is a key regulator, which controls the expression levels of other genes. Each pairwise interaction was quantified using the Hill equation. For the case of more than 1 regulator, we used Boolean logic to quantify the interactions using Hill equations. Once all the interactions were quantified, the next step was to incorporate this information into the PKMF model to quantify the effect of perturbing TFs on the lignin content and structure. The regulatory information is incorporated into the PKMF model by substituting the enzyme concentrations in the Michaelis Menten kinetics rate equation used to quantify the metabolic reactions in the metabolic network, with the Hill equations, which quantifies protein concentration as a function of transcript levels.

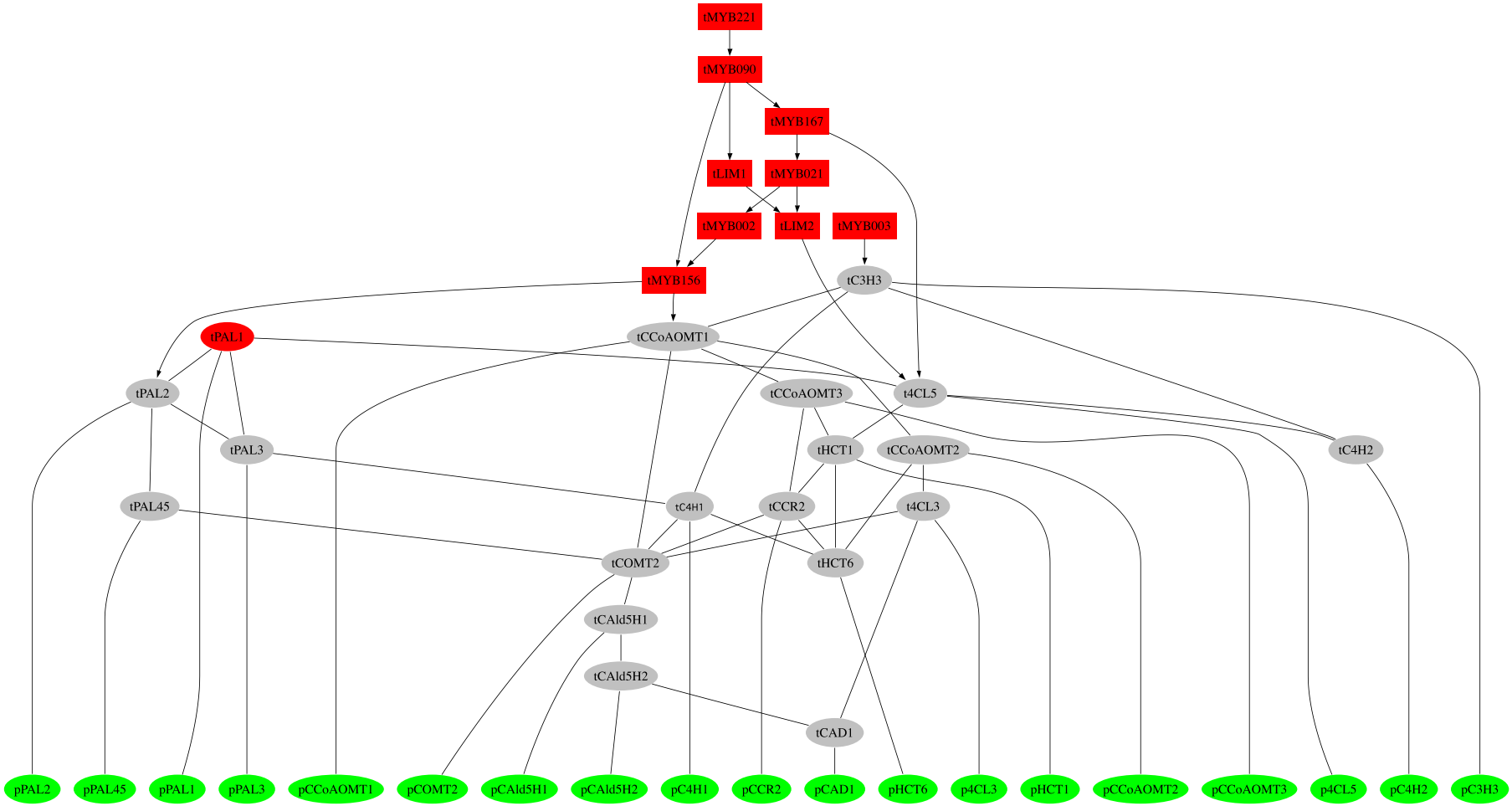

Figure 6.7: A gene regulation network for monolignol biosynthetic pathway. The boxes represent the TF's, the grey ellipses represent the transcripts of the monolignol biosynthesis genes and the proteins are denoted in green ellipses.

The ODE model was simulated to steady state and the steady state metabolic flux are computed for all the reactions in the pathway. The input flux to the PKMF model was fixed such that the S/G ratio was 2:1. The concentrations of TF's were varied from 50% to 150% of their WT levels. Assuming that the concentrations of TF's were uniformly distributed, we sampled the concentrations of the TF's using LHS procedure. 10000 different concentrations of the TF's were sampled and for each concentration of the TF, the PKMF model was simulated to steady state and the resulting metabolic flux was calculated.

The effect of varying MYB221 levels on the lignin composition was quantified by performing LHS simulation. The concentration of MYB221 was varied from a specified range. The resulting steady state S/G values were then plotted as a function of tMYB221, which is shown in the Figure 6.8.a. As seen in the figure, as the MYB221 levels are down regulated, the S/G levels decrease, which is in agreement with the results obtained by Tamagnone et al. (1998) for tobacco transgenics. The histogram plot showing the distribution of S/G plot is shown in Figure 6.8.b.

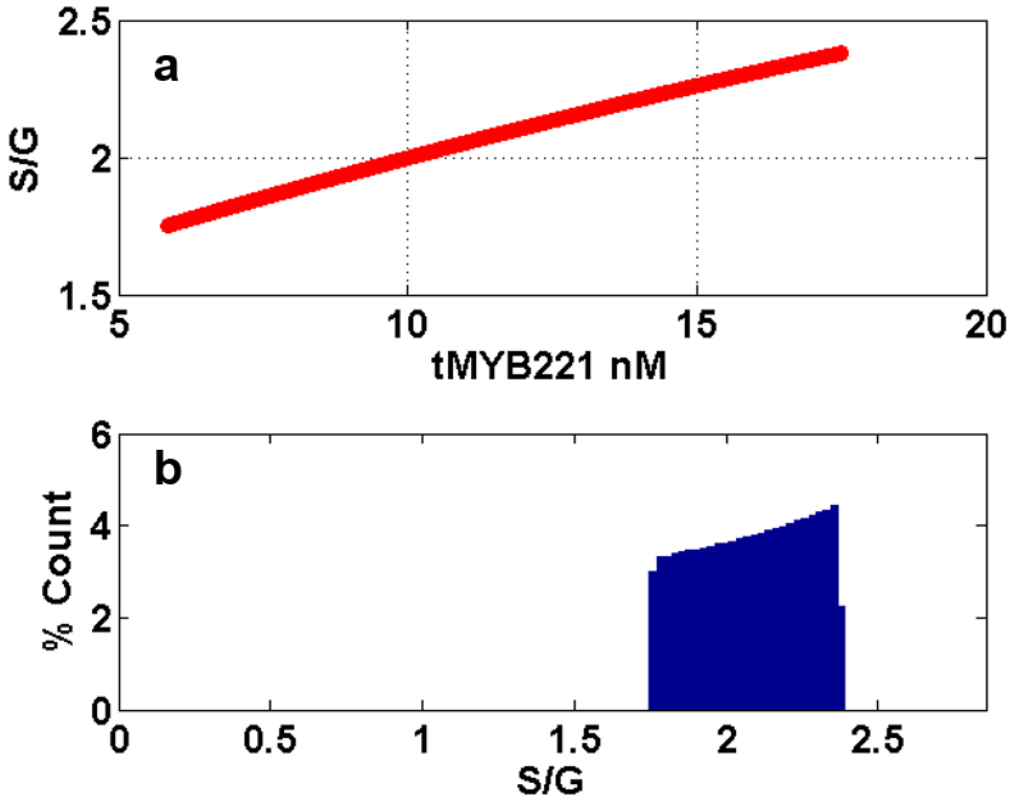

Figure 6.8: (a) Variation of S/G ratio as a function of MYB221. (b) Histogram showing the distribution of S/G ratios.

Similarly the effect of perturbing MYB003 on the lignin composition was quantified. The variation of S/G ratio as a function of MYB003 is shown in Figure 6.9. The S/G variation follows a similar trend as observed for MYB021, however for this case, it is non-linear.

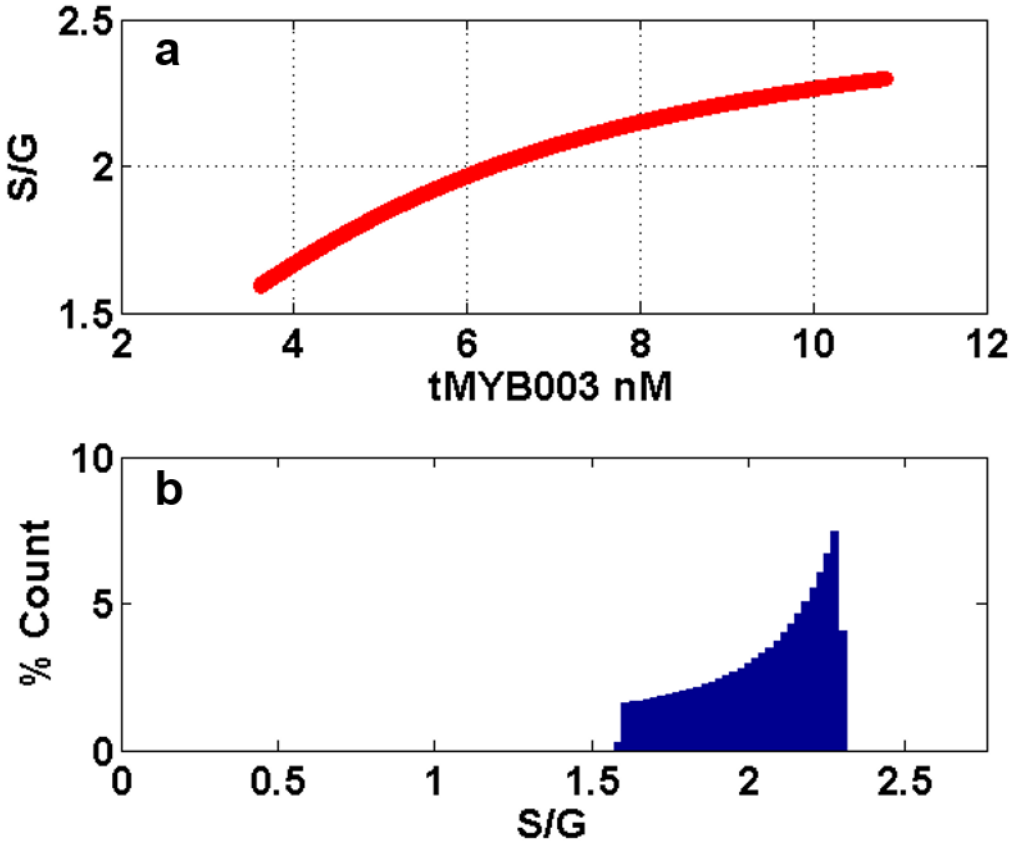

Figure 6.9: (a) Variation of S/G ratio as a function of MYB003. (b) Histogram showing the distribution of S/G ratios.

6.4 Discussion:

The transgenic data obtained by downregulating various genes involved in the monolignol biosynthesis can be promising in inferring the gene regulatory network of monolignol biosynthesis. The results from the gene regulatory network suggest that MYB090 plays a crucial role in regulating the biosynthetic pathway along with the LIM Transcription factors. The MYB transcription factors are one of the largest families of transcription factors in plants (Martin and Paz-Ares, 1997). In poplar, many of the MYBs (PtrMYB2, PtrMYB3, PtrMYB20, PtrMYB21, PtrMYB103, PtrMYB90, PtrMYB167,

PtrMYB0, PtrMYB128, PtrMYB92, and PtrMYB125) have previously been shown to control monolignol biosynthetic genes (McCarthy et al., 2010; Chai et al., 2014; Wang et al., 2014).The relationship between transcript abundance and proteins provide crucial information about the regulatory interactions that exists between the transcript's and proteins. The methodology used to develop a joint regulatory and metabolic network can be crucial to infer the role of TFs on the lignin content and composition.

References:

Acid S, De Campos LM (2001).A hybrid methodology for learning belief networks: Benedict, International Journal of Approximate Reasoning 27 (3) 235–262.

Beyer,A. et al. (2004). Post-transcriptional expression regulation in the yeast Saccharomyces cerevisiae on a genomic scale. Mol. Cell. Proteomics, 3, 1083-1092.

Chen T, Filkov V and Skiena SS (1999). Identifying gene regulatory networks from experimental data. In Proc. 3rd Ann. Int. Conf. Comp. Mol. Biol. (RECOMB'99), 94–103, New York, ACM Press.

Chiang V (2002). From rags to riches. Nat. Biotechnol. 20:557-558 .

De Campos LM (2006). A scoring function for learning Bayesian networks based on mutual information and conditional independence tests, The Journal of Machine Learning Research 7 2149–2187.

Dixon R, Achnine L, Deavours BE and Naoumkina M (2006). Metabolomics and gene identification in plant natural product pathways. In Biotechnology in Agriculture and Forestry, Volume 57: Plant Metabolomics. Edited by Saito, K. , Dixon , R.A. and Willmitzer , L. pp. 243 – 259 . Springer-Verlag, Berlin

Faulkner E (2007). K2ga: heuristically guided evolution of Bayesian network structures from data, in: Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on, IEEE, pp. 18–25.

Friedman N (2004). Inferring cellular networks using probabilistic graphical models. Science 303, 799-805. 10.1126/science.1094068

Futcher B, Latter GI, Monardo P, McLaughlin CS and Garrels JI (1999). A sampling of the yeast proteome. Mol Cell Biol 19: 7357-7368

Gardner TS, Cantor CR and Collins JJ (2000). Construction of a genetic toggle switch in Escherichia coli. Nature 403, 339-342.

Goss PJE and Peccoud J (1998). Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets. Proc. Natl. Acad. Sci. USA 95, 6750–6755.

Greenbaum D, Colangelo C, Williams K, Gerstein M (2003). Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol 4: 117.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 17: 994-999.

Higuchi T (1997). Biochemistry and Molecular Biology of Wood. pp. 131-181. Springer, New York.

Horak CE and Snyder M (2002). Global analysis of gene expression in yeast. Funct. Integr. Genomics, 2, 171-180.

Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L (2001b). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292: 929-934

Imoto S, Goto T, Miyano S. et al. (2002). Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. Pacific Symposium on Biocomputing, Vol. 7, World Scientific, pp. 175–186.

Kyoda KM, Muraki M and Kitano H (2000). Construction of a generalized simulator for multi-cellular organisms and its application to Smad signal transduction. In R.B. Altman, K. Lauderdale, A.K. Dunker, L. Hunter, and T.E. Klein, eds. Proc. Pac. Symp. Biocomput. (PSB 2000), vol. 5, 314–325, Singapore, World Scientific Publishing.

Liang S, Fuhrman S and Somogyi R (1998). REVEAL: A general reverse engineering algorithm for inference of genetic network architectures. In R.B. Altman, A.K. Dunker, L. Hunter, and T.E. Klein, eds. Proc. Pac. Symp. Biocomput. (PSB'98), vol. 3, 18–29, Singapore, 1998, World Scientifc Publishing.

Marbach D, Costello JC, KÙffner R, Vega NM, Prill RJ, Camacho DM (2012). Wisdom of crowds for robust gene network inference. Nat. Methods 9, 796–804. 10.1038/nmeth.2016

Novak B and Tyson JJ (1995). Quantitative analysis of a molecular model of mitotic control in yeast. J. Theor. Biol. 173, 283–305.

Pearl J (2000). Causality: models, reasoning and inference, Vol. 29, Cambridge Univ Press.

Ragauskas A, Williams C, Davison B, Britovsek G, Cairney J, Eckert CA (2006). The path forward for biofuels and biomaterials. Science 311: 484 -489

Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W and Selbach M (2011). Global quantification of mammalian gene expression control. Nature 473, 337-342

Smith RD. et al. (2002).The use of accurate mass tags for high-throughput microbial proteomics. OMICS, 6, 61-90.

Spirtes P, Glymour C, Scheines R (2000). Causation, prediction, and search, Vol. 81, The MIT Press.

Tsamardinos I, Brown LE, Aliferis CF (2006). The max-min hill-climbing Bayesian network structure learning algorithm, Machine learning 65 (1) 31–78.

Walhout AJM. What does biologically meaningful mean? A perspective on gene regulatory network validation. Genome Biol.2011:12:109.

Washburn MP, Koller A, Oshiro G, Ulaszek RR, Plouffe D, Deciu C, Winzeler E, Yates JR III (2003). Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae*. Proc Natl Acad Sci USA 100: 3107-3112

Zhang X, Zhao XM, He K, Lu L, Cao Y, Liu J, Hao JK , Liu ZP, Chen L (2012). Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information, Bioinformatics 28 (1) 98-104.

CHAPTER VII

CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

In this dissertation, mathematical and statistical techniques were used to: (i) Identify the role of Ptr4CL Protein-Protein complex formation on the steady state flux distribution in the monolignol biosynthetic pathway which also included the quantification of metabolite concentrations, lignin content, and composition under perturbation; (ii) The identification of protein communities and their role on the overall monolignol biosynthesis pathway under perturbations; and (iii) Validate the PKMF model using the experimentally determined phenotype data The results from model predictions, have enhanced our knowledge about the regulation of the monolignol biosynthesis pathway. The main goal of Aim 1 was to use a previously developed kinetic-based dynamic model of monolignol biosynthetic pathway to understand the role of protein complex formation on the monolignol biosynthetic pathway, as well as their roles on the lignin composition and structure. As shown in Chapter 3, the kinetic model not only facilitated in the changes in S/G ratio in response to enzymatic perturbations in the pathway, but also provided insights about the role of the protein complex on the robustness/resilience of the pathway. Using the results from chapter III we were also able to confirm that Ptr4CL5 enzyme dominates the complex and is primarily responsible for the changes in the lignin composition and structure. One of the simplified assumptions that was made in simulating the PKMF model was that the

changes in proteins concentrations are independent of other protein families. To incorporate the associative relationship which exists between proteins, Aim 2 of this dissertation was devoted towards identifying the modular structure of the proteins in wild-type and transgenic lines where measurements of absolute protein concentrations of enzymes in the monolignol biosynthesis pathway are available. By identifying the community structure, we were able to analyze the role of perturbing a particular module on the lignin biosynthesis as well as the lignin composition and structure. By evaluating the protein modules in a systematic way, we were able to elucidate regulatory mechanisms that may have remained elusive in traditional approaches. The analysis of the model incorporating the protein modules suggest that the presence of modules improve the resiliency of the pathway. The perturbation analysis also suggest that the lignin content and structure can be fine-tuned by target a group of enzymes rather than targeting individual enzymes. These findings based on the model simulation would result in formulation of testable hypothesis towards a better understanding of monolignol biosynthesis. While the results presented in Chapter 4 have greatly improved our knowledge of how monolignol biosynthesis is regulated, we wanted to assess the predictable power of the PKMF model. In order to assess the predictability of the PKMF model, we compared the S/G ratios predicted by the model for different perturbations with the transgenic S/G data in Chapter 5. The model was able to predict majority of the variations in the S/G ratios as of result of enzymatic perturbations. While, the PKMF model is a valuable tool to understand the effect of perturbations on the biosynthetic pathway and quantify the role of perturbations on the lignin composition and structure, one main drawback is that developing such a comprehensive model is a tedious and

time consuming process. If the primary purpose of a model is to predict the changes in lignin composition as a function of changes in protein concentrations, one can use statistical machine learning approaches to attain that goal. In Chapter 5, we also developed an ANN model to predict the changes in S/G ratios resulting from enzymatic perturbations. The key results from each chapter is summarized in the table 7.1.

Table 7.1: Table outlining the key results accomplished in this research.

Chapter	Key Results
III	• The pathway is unaffected by the presence of complex under WT conditions • The Ptr4CL5 is the dominant enzyme and responsible for the changes in the steady state flux as well as the lignin composition and structure under pertrubations • Presence of Ptr4CL3-Ptr4CL5 results in alternate route for monolignol biosynthesis • The Ptr4CL3-Ptr4CL5 complex increases the resiliency/robustness of the pathway by almost 20%
IV	• The proteins involved in monolignol biosynthesis pathway are oriented into 3 distinct modules • The modules provided resiliency to perturbations • Protein modules can be targeted to tailor lignin composition and structure
V	• Validate the results from the PKMF model using experimental S:G data. Developed a data driven ANN model to predict the variations in S:G ratio and total lignin content as a function of protein concentrations • PKMF model was able to predict 72% of the variation in S:G ratios as the function of protein concentrations • The ANN was able to account for 82% of the variation in S:G ratios as a function of protein concentrations • The ANN model accounted for variations in 74% of the experimentally determined (S+G) values as a function of protein concentrations
VI	• The interaction network developed using RNA-Seq data suggest that the regulation of monolignol biosynthesis is controlled by MYB221 TF.

7.2 Future Work

In chapter 3 we modified the PKMF model to incorporate the Ptr4CL complex into the model. Recent studies have shown that several other proteins involved in the monolignol biosynthesis interact with each other to form complexes (Chen et al, 2012). The next step in the model development would be to incorporate such complex interactions into the predictive model and quantify the role of the protein complexes on the lignin composition and structure. One of the assumptions that we made in predicting the lignin content and composition is that there exists a 1:1 relationship between the flux of formation of the monolignol subunits and the concentration of subunits. This assumption was primarily because the mechanism of polymerization of monolignol was not clear. Future modeling efforts should be focused on developing a computational framework to model the polymerization process in the monolignol biosynthetic pathway. The other improvement that can be made to the PKMF model is to validate the steady state concentrations of the metabolites with the experimentally determined steady state concentrations both under WT and transgenic conditions. This would enable us to fine tune the input concentrations as well as the input flux to the pathway and provide an additional step in validating the model.

The next step should be focused on incorporating the feedback inhibitions that exists between the metabolite and regulatory network level. Once this step is accomplished, such a model will become an invaluable tool for: (i) designing genetically engineered crops with reduced recalcitrance (ii) gain a deeper understanding of the

regulatory mechanisms that govern the monolignol biosynthetic pathway. The computational framework developed in this dissertation only considers lignin biosynthesis. It is well known that other pathways such as biosynthesis of cellulose, hemicellulose, and pectin are associated with cell wall synthesis. Future modeling efforts should develop strategies that would integrate the pathway models into a comprehensive cell wall model. One complex issue that needs to be addressed is that, lignin, cellulose and hemicellulose are biosynthesized at different locations within the cell. Cellulose undergoes synthesis in the plasma membrane, monolignols in cytoplasm and most other hemicellulose in the Golgi apparatus (Wolf et al, 2012). The new modeling frameworks should incorporate spatial information, partial differential equation (PDE) or some other data driven method like agent-based modeling (ABM) can be used to incorporate such spatial information.