# COMPUTATIONAL MODELING REVEALS NEW CONTROL MECHANISMS FOR LIGNIN BIOSYNTHESIS

A Dissertation
Presented to
The Academic Faculty

by

Yun Lee

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philiosophy in the
Interdisciplinary Bioengineering Graduate Program

Georgia Institute of Technology
December 2012

# COMPUTATIONAL MODELING REVEALS NEW CONTROL MECHANISMS FOR LIGNIN BIOSYNTHESIS

Approved by:


Dr. Eberhard O. Voit, Advisor
Wallace H. Coulter Department of Biomedical Engineering
*Georgia Institute of Technology*

Dr. Joshua S. Weitz
School of Biology
*Georgia Institute of Technology*

Dr. Robert J. Butera
School of Electrical and Computer Engineering
*Georgia Institute of Technology*

Dr. Ying Xu
Department of Biochemistry and Molecular Biology
*University of Georgia*

Dr. Melissa L. Kemp
Wallace H. Coulter Department of Biomedical Engineering
*Georgia Institute of Technology*



Date Approved: July 30, 2012

*To Mom, Dad, and Yu-Ting*

# ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Dr. Eberhard Voit, for his exceptional guidance and support throughout my dissertation research. Whenever I encountered difficulties with my research, he would always make time for me and help me focus on the big picture without losing sight of the details. His energy and enthusiasm for research has been an inspiration to me, and I have learned many invaluable lessons from him, both in research and in life. He is the best example of what scientists should be, and I could not have imagined having a better advisor and mentor.

I would like to thank all my collaborators at the Samuel Roberts Noble Foundation for providing me with remarkable data and valuable feedback that greatly enhanced the impact of this dissertation. It has been an honor and a privilege to know and work with Drs. Rick Dixon, Fang Chen, Luis Escamilla-Treviño, and Lina Gallego-Giraldo. I would also like to thank Drs. Robert Butera, Melissa Kemp, Joshua Weitz, and Ying Xu, for their willingness to serve on my committee, for evaluating my dissertation, and for offering constructive criticism that has helped improved this work. Special thanks to Dr. Weitz for his tremendous help and guidance when I was the TA of his course.

My time in the lab would not have been so enjoyable without the company of all my current and previous colleagues: Po-Wei Chen, I-Chun Chou, Sepideh Dolatshahi, Luis Fonseca, Gautam Goel, Xiaoling He, Shinichi Kikuchi, WangHee Lee, Anna Machina, Zhen Qi, Siren Veflingstad, James Wade, Jialiang Wu, and Weiwei Yin. I cannot thank them enough for all the inspiring conversations we have had and for their generosity to share every bit of their knowledge with me.

My sincere gratitude also goes to my friends in Atlanta with whom I have spent countless weekends and holidays. Special thanks to Chien-Chiang Chen, Szu-Yu Huang, Yung-Wen Lee, and Tsun-Yen Wu for all those Friday night talks and memorable trips. I

would also like to thank An-Ting Chien, Cheng-Lin Tsao, Yen-Chu Wang, and Pauline Yu for taking care of me during the summer of 2010.

Last but not least, I would like to thank my mom and dad for their unconditional support ever since I was born. They have always encouraged me to pursue my dream and be the best I can be. Their confidence in me is the anchor that helped me navigate through all the struggles and get where I am today. Of course, this journey would not have been possible without the love and support of my wife, Yu-Ting. No words can adequately express my gratitude for everything she has ever done for me. This dissertation is dedicated to you, my dear.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| 4CL | 4-Coumarate:CoA Ligase |
| 5HG | 5-Hydroxyguiacyl |
| 5-OH-ConifALC | 5-Hydroxyconiferyl Alcohol |
| 5-OH-ConifALD | 5-Hydroxyconiferyl Aldehyde |
| 5-OH-FA | 5-Hydroxyferulic Acid |
| 5-OH-FCoA | 5-Hydroxyferuloyl CoA |
| ABM | Agent-Based Modeling |
| BST | Biochemical System Theory |
| BV | Basis Vector |
| C3H (C3'H) | *p*-Coumaroyl Shikimate 3'-Hydroxylase |
| C4H | Cinnamate 4-Hydroxylase |
| CAD | Cinnamyl Alcohol Dehydrogenase |
| CaffA | Caffeic Acid |
| CaffCoA | Caffeoyl CoA |
| CALD5H | Coniferyl Aldehyde 5-Hydroxylase |
| CCoAOMT | Caffeoyl CoA *O*-Methyltransferase |
| CCR | Cinnamoyl CoA Reductase |
| CinnA | Cinnamic Acid |
| COMT | Caffeic Acid *O*-Methyltransferase |
| ConifALC | Coniferyl Alcohol |
| ConifALD | Coniferyl Aldehyde |
| CoumA | *p*-Coumaric Acid |

| | |
|---|---|
| CoumALD | *p*-Coumaryl Aldehyde |
| CoumALC | *p*-Coumaryl Alcohol |
| CoumCoA | *p*-Coumaroyl CoA |
| EP | Extreme Pathway |
| EM | Elementary Mode |
| F5H | Ferulate 5-Hydroxylase |
| FA | Ferulic Acid |
| FBA | Flux Balance Analysis |
| FCoA | Feruloyl CoA |
| G | Guaiacyl |
| G1P | Glucose 1-Phosphate |
| G6P | Glucose 6-Phosphate |
| $Glc_{ext}$ | External Glucose |
| GMA | Generalized Mass Action |
| H | *p*-Hydroxyphenyl |
| HCT | Hydroxycinnamoyl-CoA:Shikimate Hydroxycinnamoyl Transferase |
| HQT | Hydroxycinnamoyl-CoA:Quinate Hydroxycinnamoyl Transferase |
| iFBA | Integrated Flux Balance Analysis |
| IOM | Indirect Optimization Method |
| IPTG | Isopropyl 1-Thio β-Galactopyranoside |
| MAPK | Mitogen-Activated Protein Kinase |
| MCMC | Method of Controlled Mathematical Comparisons |
| MI | Mutual Information |
| MILP | Mixed-Integer Linear Programming |
| MOMA | Minimization of Metabolic Adjustment |

| | |
|---|---|
| ODE | Ordinary Differential Equation |
| OMT | *O*-Methyltransferase |
| PDE | Partial Differential Equation |
| rFBA | Regulated Flux Balance Analysis |
| S | Syringyl |
| SA | Sinapic Acid/Simulated Annealing |
| SALC | Sinapyl Alcohol |
| SALD | Sinapyl Aldehyde |
| SCoA | Sinapoyl CoA |
| PAL | Phenylalanine Ammonia –Lyase |
| Phe | Phenylalanine |
| PPP | Pentose Phosphate Pathway |
| T6P | Trehalose 6-Phosphate |
| Tre | Trehalose |
| UDGP | Uridine Diphosphate Glucose |

# SUMMARY

Lignin polymers provide natural rigidity to plant cell walls by forming complex molecular networks with polysaccharides such as cellulose and hemicellulose. This evolved strategy equips plants with recalcitrance to biological and chemical degradation. While naturally beneficial, recalcitrance complicates the use of inedible plant materials as feedstocks for biofuel production. Genetically modifying lignin biosynthesis is an effective way to generate varieties of bioenergy crops with reduced recalcitrance, but certain lignin-modified plants display undesirable phenotypes and/or unexplained effects on lignin composition, suggesting that the process and regulation of lignin biosynthesis is not fully understood. Given the intrinsic complexities of metabolic pathways in plants and the technical hurdles in understanding them purely with experimental methods, the objective of this dissertation is to develop novel computational tools combining static, constraint-based, and dynamic, kinetics-based modeling approaches for a systematic analysis of lignin biosynthesis in wild-type and genetically engineered plants. Pathway models are constructed and analyzed, yielding insights that are difficult to obtain with traditional molecular and biochemical approaches and allowing the formulation of new, testable hypotheses with respect to pathway regulation. These model-based insights, once they are verified experimentally, will form a solid foundation for the rational design of genetic modification strategies towards the generation of lignin-modified crops with reduced recalcitrance. More generically, the methods developed in this dissertation are likely to have wide applicability in similar studies of complex, ill-characterized pathways where regulation occurring at the metabolic level is not entirely known.

# CHAPTER 1

# INTRODUCTION

## 1.1 Overview

Petroleum and its derivatives have been the dominant energy source for more than 150 years, leading to high oil prices, an increased dependence on imported oil, and numerous environmental consequences. To reduce our reliance on petroleum, we need renewable alternatives that not only act as viable substitutes for crude oil and coal but also, if possible, carry lower environmental cost. Although natural sources such as the sun and wind have been tapped to generate low-carbon electricity, the reality is that their output is not sufficient and that sustainable transportation requires liquid fuels. Outside fossil-based products, the only currently available alternative for powering the world's millions of motor vehicles appears to be biomass that is converted into liquid *biofuels*.

First-generation biofuels, which are fermented and refined from sugar-rich crops like corn, sugarcane and soybeans, have unfavorably contributed to higher food prices, deforestation, and many other undesirable consequences [2]. Second-generation biofuels, by contrast, are derived from inedible or woody parts of plants, including wheat straw, corn stover and wood shavings that are natural by-products of agricultural and forestry operations. Each year, more than 40 million tons of such materials are produced [3], most of which is simply discarded, because proper uses have not been found. Turning these sustainable, yet minimally utilized, feedstocks into biofuels thus presents a great opportunity for meeting future energy demands, especially for transportation, while minimizing competition with food production.

As promising as this approach is, the extraction of energy from the various woody feedstocks remains a challenging task. One important barrier that prevents biofuels from

living up to their potential is the natural rigidity of plant cell walls. In woody parts of plants, the cell walls are mainly composed of two polysaccharides, cellulose and hemicellulose, and the phenolic polymer lignin. These polymers have evolved to form a complex network that resists biological and chemical degradation and carries out many physiological functions. For biofuel production, however, this natural robustness or "recalcitrance," which can be largely attributed to the cross-linking of lignin with polysaccharides, makes it very difficult to retrieve glucose or xylose molecules that can be subsequently fermented into ethanol or longer-chain alcohols like butanol.

In order to retrieve easily fermentable sugars, the current strategy involves a costly, thermo-chemical pretreatment that breaks up lignin and makes polysaccharides more accessible to specialized enzymes with hydrolytic activity. This process is harsh and energy-intensive, and it has many undesirable side effects, such as the accumulation of chemicals that are known to have an inhibitory effect on the subsequent hydrolysis and fermentation steps [4]. Many different solutions have been proposed to address this and other related issues [5], but the existing technologies are still in a very early stage of development and thus will not be available on a large scale anytime soon.

An alternative approach that has the potential of lowering the processing cost of biofuels is crop engineering, which entails genetic modifications that render the targeted biomass more easily fermentable. For bioenergy crops such as poplar and switchgrass, it has been found that the sugar yields following acid pretreatment are affected by a natural variation in lignin content and monomer composition [6,7]. Moreover, Fu and colleagues [8] showed that certain lignin-modified switchgrass plants produce equivalent ethanol yields as control plants but require a less severe pretreatment. These observations suggest that lignin biosynthesis may be targeted for generating engineered crops with reduced recalcitrance. While conceptually appealing and straightforward, such an approach will require rational design strategies toward such varieties, which in turn presuppose a thorough, multi-level understanding of lignin biosynthesis. However, even after decades

of research, many fundamental aspects of this important pathway, such as its topology [9,10] and regulation [11,12,13], are still unclear.

To enhance our understanding of the process and control of lignin biosynthesis, the **objective of my dissertation research** is to develop novel computational tools for a systematic analysis of lignin biosynthesis in wild-type and genetically engineered plants. Pathway models will be constructed and analyzed, yielding insights that are difficult to obtain with traditional molecular and biochemical approaches and allowing the formulation of new, testable hypotheses with respect to pathway regulation. Once validated with experimental means, these model-based insights will form a solid foundation for the rational design of genetic modification strategies towards the generation of lignin-modified crops with reduced recalcitrance.

## 1.2 Lignin Biosynthesis

In most woody plants, lignin polymers are mainly derived from three hydroxycinnamyl alcohol monomers, namely *p*-coumaryl, coniferyl and sinapyl alcohols. Once synthesized inside the cytoplasm, these *monolignols* are transported into the cell wall and produce *p*-hydroxyphenyl (H), guaiacyl (G) and syringyl (S) units, respectively, that are incorporated into the lignin polymer (Figure 1.1). The lignin content and composition of monomers vary among taxa, cell types and even individual cell wall layers. In potential bioenergy crops like poplar (dicot) and switchgrass (monocot), lignin consists principally of G and S units, while H units are present in low to negligible quantities. However, the natural composition of lignin is susceptible to genetic manipulation, as is evident in some transgenic plants where H units are present at significant levels compared to G and S units [14].

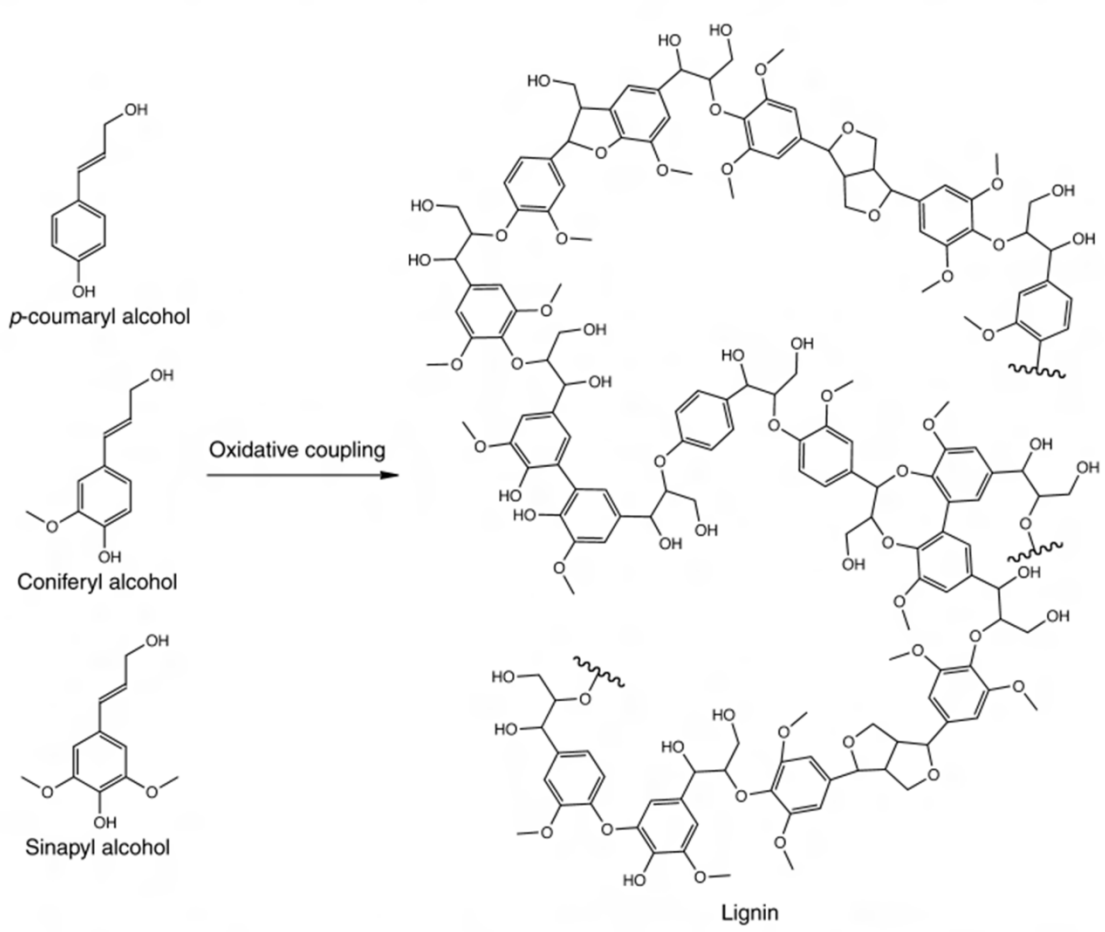


**Figure 1.1: The phenolic polymer lignin is mainly derived from three hydroxycinnamyl alcohols.** Figure is adapted from [15].

Most metabolic genes involved in monolignol biosynthesis have been identified with complete sequence information in model organisms, such as *Arabidopsis thaliana* [16] and the black cottonwood *Populus trichocarpa* [17]. Knowledge obtained from these genomes has been used for homology searches in species where sequencing efforts are ongoing. In most plant species, monolignols are synthesized *de novo* from the amino acid phenylalanine. Phenylalanine ammonia-lyase (PAL) catalyzes the first reaction in which phenylalanine is converted into cinnamic acid; the rest of the pathway involves successive hydroxylation and methylation of the aromatic ring, followed by a conversion of the side-chain carboxyl to an alcohol group. Because ring hydroxylation/methylation can occur at different levels of side chain oxidation, early studies often described

monolignol biosynthesis as a metabolic grid [9]; this view, however, was later modified to reflect substrate preferences of multiple hydroxylases and methyltransferases, and the currently accepted pathway is more or less linear with only a few branch points [18]; genus-specific details about this pathway will be discussed in Chapters 2, 3 and 5.

In addition to a constantly revised topology, it is important to understand the regulatory mechanisms governing monolignol biosynthesis. This regulation occurs in different forms. First, substrate competition may arise because many enzymes in the pathway are multifunctional (*i.e.*, they are active towards multiple substrates). Second, some pathway enzymes have multiple isoforms with distinct kinetics and substrate preferences [19,20,21,22], and these isoforms may be differentially expressed during development or upon environmental cues [20,21,22,23], suggesting that the pathway topology is subject to further revision if context-specific information is available. Third, specific enzymes may assemble into complexes such that the pathway flux can be more flexibly controlled [12,24]. Given the complexity and multitude of regulatory features that characterize monolignol biosynthesis, it is not surprising that genetic perturbations of monolignol biosynthetic genes sometimes yield counterintuitive results that cannot be explained simply by the stoichiometric connectivity of the pathway [25].

### 1.3 Significance of Research

In order to develop a mechanistic understanding of monolignol biosynthesis in bioenergy crops, it seems advantageous to resort to computational modeling, which has become a standard tool for analyzing metabolic pathways. However, partly because of the inherent complexities described above and of technical hurdles, for example in measuring the low intracellular levels of pathway intermediates [26], no modeling effort targeting this specific pathway has been reported in the literature. In this regard, the mathematical model developed in **Specific Aim 1** (see Section 1.6.1) is the first model built. It provides

"proof of concept" that modeling can yield genuine benefits as compared to experimental methods. Building upon this modeling effort, the model analysis in **Specific Aims 2 and 3** (see Sections 1.6.2 and 1.6.3) offers insights that are difficult to obtain with traditional molecular and biochemical approaches and allows the formulation of new, testable hypotheses with respect to the metabolic regulation of lignin biosynthesis.

## 1.4 Mathematical Modeling of Metabolic Systems

The process of constructing mathematical models, as one might imagine, is challenging and typically involves multiple related phases that still await the establishment of standard operating procedures. Among the various challenges of mathematical modeling, the task of parameter estimation is arguably the most complex. This task is required in most modeling efforts, as it converts a purely symbolic representation of the biological system into a numerical model that permits a variety of *in silico* experiments. Once a model is fully parameterized and deemed reliable and appropriate for the target system, one can use it to make predictions, generate testable hypotheses, guide the design of new experiments, or propose manipulation strategies that allow the yield of desired product(s) to be maximized. To a large extent, these potential merits were the drivers that triggered the creation of many parameter estimation methods over the past years (reviewed in [27]), but the consensus among practitioners remains to be that no single method can be declared the best in terms of efficiency, robustness and reliability.

In most cases of parameter estimation, the absence of a clear winner is understandable because the development of new methods is usually tailored to meet the specific requirements of a given modeling framework as well as the availability and types of experimental data. As a pertinent example, ordinary differential equations (ODE) have been widely used to describe systems of interconnected metabolic processes. The

parameterization of such models is typically accomplished in a "bottom-up" approach where individual processes are fitted one at a time, using kinetic parameters of corresponding enzymes. However, the increasing availability of metabolic time series data suggests an entirely different approach in which features of individual processes can be inferred from a comprehensive monitoring of the whole system. Because these "global" data are collected within the same organism and obtained under the same experimental condition, they are more likely to offer an accurate description of what actually happens *in vivo* than the "local" data obtained from traditional experiments. Still, many challenges remain to be addressed (see [27] for a detailed discussion) before the challenge of this inverse or "top-down" approach can be considered solved.

## 1.5 Modeling Frameworks

The choice of a modeling framework is often made on an *ad hoc* basis, depending on the degree of realism a modeler intends to pursue as well as on the type of experimental data against which the parameterized model can be validated. Ideally, a model should be nonlinear and dependent on space and time, with the ability to consider both stochastic and discrete effects. Two possible shortcomings of such a realistic model, if ever feasible, are that simulations would be computationally prohibitive and that high-quality data—maybe at the level of a single cell or single molecule—would be required for model fitting. Although high-throughput methods based on microfluidics, flow cytometry, nuclear magnetic resonance, and mass spectrometry are currently available to obtain such data [28], the number of molecules that can be quantified reliably and simultaneously is still limited [29]. Therefore, it is often desirable to use approximations and/or abstractions that reduce model complexity (and thereby lessen the need for detailed data), but the assumptions on which such simplifications are based must be checked on a case-by-case basis.

The modeling frameworks that have been widely used in the context of metabolic systems analysis consist mainly of the following two classes: (i) static, constraint-based models; and (ii) dynamic, kinetics-based models. Both classes of models, despite their individual differences, are based on ODEs. The generic format for such a representation can be written as

$$\frac{d\mathbf{X}}{dt} = \mathbf{N} \cdot \mathbf{v}. \tag{1.1}$$

In this representation, $\mathbf{X} = [X_1\ X_2\ \cdots\ X_n]^T$ is an *n*-dimensional column vector, with each component $X_i$ denoting the time-dependent concentration of a metabolite, or pool of metabolites; $\mathbf{v} = [v_1\ v_2\ \cdots\ v_l]^T$ is an *l*-dimensional column vector, with each component $v_i$ denoting the flux through the $i^{\text{th}}$ reaction; and **N** is an *n*-by-*l* stoichiometric matrix, with each entry $N_{ij}$ being a positive integer, if the $j^{\text{th}}$ reaction produces $X_i$, a negative integer, if the $j^{\text{th}}$ reaction consumes $X_i$, or zero, if $X_i$ and the $j^{\text{th}}$ reaction are unrelated. It should be noted that each flux $v_i$ depends not only on the metabolic state of the system as characterized by $X_i$, $i = 1,...,n$, but also on a variety of non-metabolic factors such as transcriptional and translational regulators, temperature, or pH. The challenge is thus to identify the functional form of each flux and to find the numerical values for its parameters. In the following sections I will briefly review some particularly relevant implementations from both classes and also discuss how one can take an integrated approach to build a better model.

### 1.5.1 Static, Constraint-Based Modeling

The first class of models focuses on the stoichiometry of a metabolic system. A distinct feature of all models within this class is that they follow a quasi-steady-state assumption: As a metabolic system typically operates on a timescale of seconds, it is assumed to reach equilibrium relatively fast compared to most cellular processes and thus

to reside in a steady state where, for each metabolite pool, the fluxes governing its synthesis and degradation are equal and constant. Under this assumption, the left-hand side of Eq. (1.1) becomes zero, thereby turning the system of differential equations into to a system of linear equations. This linear system is often underdetermined because there are typically fewer equations (metabolites) than unknowns (fluxes). To overcome this issue, methods such as extreme pathways (EP; [30]) and elementary modes (EM; [31]) may be applicable if the task is to find a finite set of flux vectors that characterize all permissible steady-state flux distributions.

Alternatively, flux balance analysis (FBA; [32,33]) is an optimization-based approach that aims to identify a specific flux distribution that satisfies the steady state and optimizes a certain *a priori* objective. Although recent studies have shown that no single objective is suited for all conditions [34,35], it is assumed in most microbial studies that fast-growing microbes tend to maximize their growth rate or biomass production. Given such an assumption, along with various thermodynamic and physicochemical constraints [36], one can formulate the optimization task as a linear program, which is easily solved even for large, genome-scale systems. Examples of constraint-based models can now be found for more than 35 different organisms, including microbes [37,38], higher organisms such as humans [39,40] and plants [41,42].

### 1.5.2 Dynamic, Kinetics-Based Modeling

A significant feature—but also the major weakness—of the FBA approach is that it does not use or generate any information about individual metabolites or enzymes. This may pose a problem, for example, if pathway intermediates are known to affect the activity of certain enzymes in the pathway. Furthermore, FBA focuses exclusively on a given steady state. To address these issues, dynamic, kinetics-based models that center on

metabolite concentrations, protein concentrations, or levels of gene expression are often a better fit.

Traditionally, the formulation of dynamic models of metabolic pathways starts with selecting a functional representation for each reaction that best describes its kinetics *in vitro*. Given the explicit representations of individual reactions, the next step is to integrate them into a system of ODEs where each equation describes the temporal changes in one metabolite as the difference between the sums of rates of its synthesis and degradation. Having determined the initial concentrations, one can solve the ODE to obtain the metabolic concentrations at different time points, which are not necessarily at steady state, and compute fluxes and other results if needed. Although the characterization of dynamic models requires many kinetic details, they eventually offer the ability to predict all metabolite concentrations and fluxes under non-steady state conditions. Furthermore, they permit predictions of the consequences of manipulations that take network regulation into account.

Biochemical Systems Theory

In this dissertation, dynamic models within the framework of *Biochemical Systems Theory* (BST; [43,44]) will be used. BST models have many important advantages, which have been discussed extensively in hundreds of articles since their first inception in the late 1960s. Some distinguished examples with dozens of variables include a red blood cell model with ~100 variables [45], and models of citric acid [46], purine [47], sphingolipid [48,49,50], and dopamine [51] metabolism. BST has also been applied in studies of plant phenomena such as biomass partitioning in growing trees [52,53] and the so-called 3/2 rule of self-thinning in planted forests [54].

There are two major modeling formats within BST, which are Generalized Mass Action (GMA) models and S-system models. The common feature of both formats is a representation of all involved processes as products of power-law functions.

Mathematically, each power-law term can be considered as the result of linearizing a nonlinear process via Taylor expansion around an operating point in logarithmic coordinates. In the S-system format, the derivative of each time-dependent variable is given as the difference between one set of influxes and one set of effluxes, and each set is collectively written as one product of power-law functions. By this definition, the generic S-system format reads

$$\dot{X}_i = \alpha_i \prod_{j=1}^{n+m} X_j^{g_{i,j}} - \beta_i \prod_{j=1}^{n+m} X_j^{h_{i,j}}, \; i = 1, ..., n. \tag{1.2}$$

Here, the first *n* variables are *dependent* variables, which are controlled by the system as well as by the remaining *m* variables (also called *independent* variables) that do not change over time. The non-negative multipliers $\alpha_i$ and $\beta_i$ are *rate constants*, which specify the turnover rate of the collective production and degradation, respectively, and the real-valued exponents $g_{i,j}$ and $h_{i,j}$ are *kinetic orders*, which reflect the strengths of the effects that the corresponding variables $X_j$ have on a given process. For example, a kinetic order $g_{i,j}$ is positive, if $X_j$ has an activating or augmenting effect on $V_i^+$ ; negative, if $X_j$ has an inhibitory effect on $V_i^+$; or zero, if $X_j$ does not have any effect on $V_i^+$.

In the GMA format, instead of aggregating all influxes or effluxes into one term each, every reaction affecting a metabolite is approximated individually with one power-law term such that

$$\dot{X}_i = \sum_{j=1}^{l} N_{ij} \cdot \left( \gamma_j \prod_{k=1}^{n+m} X_k^{f_{j,k}} \right), \; i = 1, ..., n. \tag{1.3}$$

Here, $N_{ij}$ refer to the entries of the stoichiometric matrix **N**, while $\gamma_j$ are non-negative rate constants and $f_{j,k}$ are real-valued kinetic orders as in the S-system format. It should be noted that the only difference between these two formats occurs at branch points in a

pathway system, where more than one flux is involved in the production and/or degradation process. In other words, a linear pathway with no branch points will have only one mathematical formulation regardless of which format is used.

A key question to be considered is thus how to choose between these two formats. On one hand, it is beneficial to use the S-system format because it allows the steady-state equations to be solved very efficiently in a linear fashion. This advantageous property could have a substantial impact on the simulation time if all the available data are taken from different steady states. Furthermore, sensitivity analyses and optimization tasks are much simpler than in other nonlinear models. On the other hand, the GMA format is more intuitive in terms of interpretability since each flux corresponds to a unique power-law term. This exact pairing is also convenient if the task is to compile kinetic parameters for individual enzymes into an integrated model. As I will show in Chapter 5, the two formats may actually be used interchangeably in the same work, depending on the specific requirements of the subtasks.

### 1.5.3 Bridging the Gap between Two Modeling Frameworks

Interestingly, very little effort has been devoted to integrating these two modeling strategies. Covert and colleagues [55] recently proposed an integrated FBA (iFBA) framework and demonstrated its utility by combining a large-scale regulated FBA (rFBA) model of *Escherichia coli* metabolism [56,57] with a small ODE model of *E. coli* carbohydrate uptake [58]. The authors suggested that the integrated model offers several advantages over the original rFBA model, one of them being the feedback regulation by metabolites, which is explicitly considered by the iFBA model and critical in shaping cellular responses, for instance, in the studied example of diauxic growth,. However, it should be noted that the iFBA framework is intrinsically not much different from any

other constraint-based model in that a FBA-type linear programming problem is solved repeatedly to simulate the dynamics.

The iFBA framework offers a good example of how constraint-based models can improve their prediction by imposing constraints derived from kinetics-based models. By the same token, a sensible approach will be to first determine the fluxes at a nominal steady state using constraint-based models and then infer the kinetic parameters based on concentration measurements and functional descriptions of individual fluxes. The proposed approach would have a dual benefit: not only are the flux estimates more accurate, but one also gains information on metabolite concentrations. Notably, such integration is greatly facilitated by using models within the framework of BST because mechanistic assumptions regarding the enzymatic processes can be minimized. Applications of this approach are described in detail in Chapters 2 and 5.

## 1.6 Dissertation Overview

The overall objective of this dissertation (see Section 1.1) is achieved by tackling three specific aims as listed below (see Table 1.1 for an overview of corresponding chapters and appendices):

### 1.6.1 Specific Aim 1: Develop a Dynamic Model of Lignin Biosynthesis in *Populus* Xylem

A dynamic model of lignin biosynthesis in *Populus* xylem is presented in Chapter 2. *Populus* is the genus of choice in this proof-of-concept study not only because the kinetic properties of many enzymes within lignin biosynthesis have been characterized in poplar or aspen, but also because the available data from a number of transgenic poplar and aspen varieties can be used to test the validity of the inferred model. As will be

shown in Chapter 2, the model development involves the following steps: (i) determine the topology and regulation of lignin biosynthetic pathway using literature information; (ii) develop and employ a two-step modeling approach that combines the strengths of FBA and BST; and (iii) optimize and validate the model against the experimental data in several transgenic *Populus* varieties where specific enzymes are genetically perturbed. With the resulting model, I will also demonstrate its applicability in metabolic engineering through an *in silico* case study.

### 1.6.2 Specific Aim 2: Analyze Monolignol Biosynthesis in Various Transgenic Alfalfa (*Medicago sativa* L.) Plants and Different Stem Segments

The dynamic model of monolignol biosynthesis developed in Aim 1, although seemingly consistent with all the published results in *Populus,* does not account for two functionally distinct isoforms of cinnamoyl CoA reductase (CCR) that were identified and characterized in a key position of the same pathway in *Medicago* [59]. This finding, along with the kinetic properties of another enzyme [60], suggests a revision of the pathway topology in alfalfa. Therefore, the goal of this Aim is to evaluate an ensemble of wild-type and lignin-modified alfalfa plants [25] against a revised pathway topology, and the following tasks are achieved in Chapter 3: (i) develop a novel modeling approach that permits an interrogation of monolignol biosynthesis in eight stem segments, called *internodes*, of both wild-type and various transgenic plants; (ii) elucidate or explain mechanisms of metabolic regulation; and (iii) validate postulated regulatory mechanisms by means of *post hoc* experiments.

Pertaining to the goal of this Aim, a computational approach is developed and presented in Chapter 4 that seeks to address the following questions: Why do we observe a specific developmental pattern of fluxes but not *a priori* equally valid alternatives? In what sense might the observed pattern be superior? Answers to these questions will not

only complement the results in this aim, which relate to the mechanism of regulation that emerge between plant *lines* rather than between internodes, but also constitute a major step towards understanding how a hierarchy of transcription factors coordinates the biosynthesis of different monolignols during secondary cell wall thickening.

### 1.6.3 Specific Aim 3: Functional Analysis of Metabolic Channeling and Regulation in Monolignol Biosynthesis

Results from Aim 2 suggest that specific enzymes may co-localize and assemble into independent channels leading to G and S monolignols. The specific hypothesis in this Aim is therefore that these channels may have different modes of operation. These modes correspond to different network configurations and may exhibit different patterns of "crosstalk." Therefore, the objective of this Aim is to refine the channeling postulate by analyzing *every* possible network design, each defined as a specific combination of a topological configuration and a crosstalk pattern. To accomplish this aim, the following tasks are achieved in Chapter 5: (i) construct a library of dynamic models that encompass all possible designs; specifically, I will use a simplified network that only involves the steps pertinent to the channeling mechanism within the monolignol biosynthetic pathway; (ii) identify and analyze the designs with model instantiations whose predictions are consistent with experimental data; (iii) design experiments to validate the hypothesis derived from model predictions.

**Table 1.1: Dissertation overview**

| Chapter | Content | Related Appendices |
|---|---|---|
| **2**[i] | Mathematical Modeling of Monolignol Biosynthesis in *Populus Xylem* | APPENDIX A |
| **3**[ii] | Integrative Analysis of Transgenic Alfalfa (*Medicago sativa* L.) Suggests New Metabolic Control Mechanisms for Monolignol Biosynthesis | APPENDIX B |
| **4**[iii] | Analysis of Operating Principles with S-system Models | |
| **5**[iv] | Functional Analysis of Metabolic Channeling and Regulation in Lignin Biosynthesis: A Computational Approach | APPENDIX C |
| **6** | Conclusions and Future Work | |

i. Adapted from: Lee, Y. and Voit, E.O. (2010) Mathematical Modeling of Monolignol Biosynthesis in Populus Xylem. *Math. Biosci.* 228: 78-89.

ii. Adapted from: Lee, Y., Chen, F., Gallego-Giraldo, L., Dixon, R.A. and Voit, E.O. (2011) Integrative Analysis of Transgenic Alfalfa (*Medicago sativa* L.) Suggests New Metabolic Control Mechanisms for Monolignol Biosynthesis. *PLoS Comput. Biol.* 7(5): e1002047.

iii. Adapted from: Lee, Y.*, Chen, P.-W.* and Voit, E.O. (2011) Analysis of Operating Principles with S-system Models. *Math. Biosci.* 231: 49-60. [*Equal contribution]

iv. Adapted from: Lee, Y., Escamilla-Treviño, L., Dixon, R.A. and Voit, E.O. (*submitted*) Functional Analysis of Metabolic Channeling and Regulation in Lignin Biosynthesis: A Computational Approach.

# CHAPTER 2

# MATHEMATICAL MODELING OF MONOLIGNOL BIOSYNTHESIS IN *POPULUS* XYLEM[1]

## 2.1 Introduction

Extensive and sustained biochemical and physiological research efforts and, especially, numerous insights from investigations of relevant plant genomes, have shed light on the specific roles of most genes involved in the monolignol biosynthetic pathway, which generates the building blocks of lignin. Such genome-based information is very valuable but by itself insufficient for explaining or predicting how the monolignol biosynthetic pathway would respond to untested changes in enzyme activities or gene expression, because at least some of the pathway regulation occurs at the metabolic level in a rather complex fashion.

Recently, metabolite (and specifically phenolic) profiling has been used in various transgenic studies to monitor the *in vivo* concentrations of intermediate phenylpropanoid species in the pathway [61,62]. These studies have generated pertinent information that elucidates the lignin monomer biosynthesis from a different perspective and augments the genomic information from earlier studies in a beneficial fashion. Nevertheless, the application of metabolite profiling, for instance, in the characterization of metabolic phenotypes caused by genetic modification [26], is often limited because the levels of some lignin precursors are low and thus difficult to measure.

Concurrent with the advances in genomic and metabolomic analysis, mathematical and computational techniques from the field of systems biology have

---

[1] Adapted from: Lee, Y. and Voit, E.O. (2010) Mathematical Modeling of Monolignol Biosynthesis in Populus Xylem. *Math. Biosci.* 228: 78-89.

emerged as effective tools to help explain the regulation of complex metabolic networks. Examples from yeast demonstrate that sufficient genome annotation, when augmented with biochemical and physiological information, permits the mathematical reconstruction of essentially the entire metabolic network with reasonable fidelity [63,64]. This reconstructed metabolic network can serve as a solid platform from which one may first infer and investigate the metabolic flux distribution and subsequently derive quantitative relationships between genotype, gene expression and phenotype for the pathway of interest.

Two classes of methods are available to achieve these objectives; they have been reviewed in Chapter 1 and need no further discussion here. Normally, only one of the two classes is used to model a metabolic network, depending on the questions being asked and information available. Given the limited number of concentration measurements in the monolignol biosynthetic pathway, stoichiometric or flux balance analysis appears to be the model of choice. However, understanding the regulatory mechanisms that are not explicitly taken into account by FBA models constitutes an important step toward applying metabolic engineering techniques to improve biofuel production and seems mandatory before genetic alterations are introduced in natural pathways. Thus, in the spirit of a recent study [65], which proposes a discussion of integrating divergent modeling approaches, we use here a combination of FBA and BST models for analyzing the monolignol biosynthetic pathway at the systems level. This novel combination strategy allows us to harness the regulatory aspects of a kinetic model based on the metabolic flux distribution obtained from a flux balance model.

Key features of the new strategy are outlined in the following. First, we begin with a minimal amount of experimental information and construct a stoichiometric flux balance model. In the second step, we augment this model using additional biological information, along with various parameter optimization techniques, and morph the static linear model into a dynamic nonlinear model. The ultimate goal of this two-step approach

is the establishment of a reliable model that can be used to identify target genes and devise effective strategies for generating modified crops with reduced amounts of lignin. So far, we have not reached the goal of absolute numerical reliability because the currently existing information is still rather scarce. Nonetheless, the resulting model structure appears to be qualitatively adequate and has the capacity to serve as the basis for systematically identifying critical system components (enzymes) whose alterations could improve the yield of fermentable sugars by means of genetic engineering.

## 2.2 Materials and Methods

### 2.2.1 Metabolic Mapping

Our main biological target is the xylem in *Populus*, because a rapidly increasing number of transgenic poplar and aspen varieties within this genus has significantly contributed to our understanding of the enzymes driving the monolignol biosynthetic pathway [14]. Focusing on the metabolic processes occurring in the cytoplasm, we start with the biosynthetic pathway leading to the building blocks of lignin (Figure 2.1; also see Section A.1.1 for a detailed discussion of how the pathway structure was determined). The pathway generates four alcohols, three of which—*p*-coumaryl, coniferyl, and sinapyl alcohols—are called monolignols. Once synthesized, the monolignols are transported from the cytoplasm to the cell wall, where they are oxidized and polymerized to form lignin. When incorporated into the lignin polymer, these monolignols produce, respectively, p-hydroxyphenol (H), guaiacyl (G), and syringyl (S) phenylpropanoid units, which are shown at the periphery of the pathway diagram in Figure 2.1. The relative amounts of monolignols, which are affected by a variety of factors [66], determine many of the features of the resulting lignin, such as its structure, toughness and chemical recalcitrance. In dicotyledonous angiosperms, including *Populus*, lignin consists

primarily of G and S monomers, whereas the amount of H is negligible. The ratios of lignin monomers and the total lignin content have been closely monitored in transgenic *Populus* variants because of their important role in lignin extractability, forage digestibility [67] and, most importantly, sugar release by enzymatic hydrolysis [25].

In addition to the topology of the network of all enzymatic reactions, it is necessary to account for regulatory mechanisms that are known, alleged, or hypothesized for the monolignol biosynthetic pathway. Correspondingly, we augmented the pathway model with regulatory features found in the literature, paying special emphasis to *Populus* (Figure 2.1; Table 2.1). It should be mentioned that several of the enzymes in the monolignol biosynthetic pathway have multiple isoforms with slightly different kinetics and substrate preferences, and the genes coding for these isoforms are differentially expressed during development and under different environmental cues and stresses [14]. At this point, this degree of complexity could not be taken into account, due to missing quantitative measurements of the different isozymes in *Populus* xylem, and we focused instead on their collective activity in catalyzing each reaction step. At the same time, if one isoform is known to have a dominant effect over other isoforms, such as Pt4CL1 in aspen xylem [68], the corresponding kinetic constants are assumed to be representative (*cf.* Table A.2).

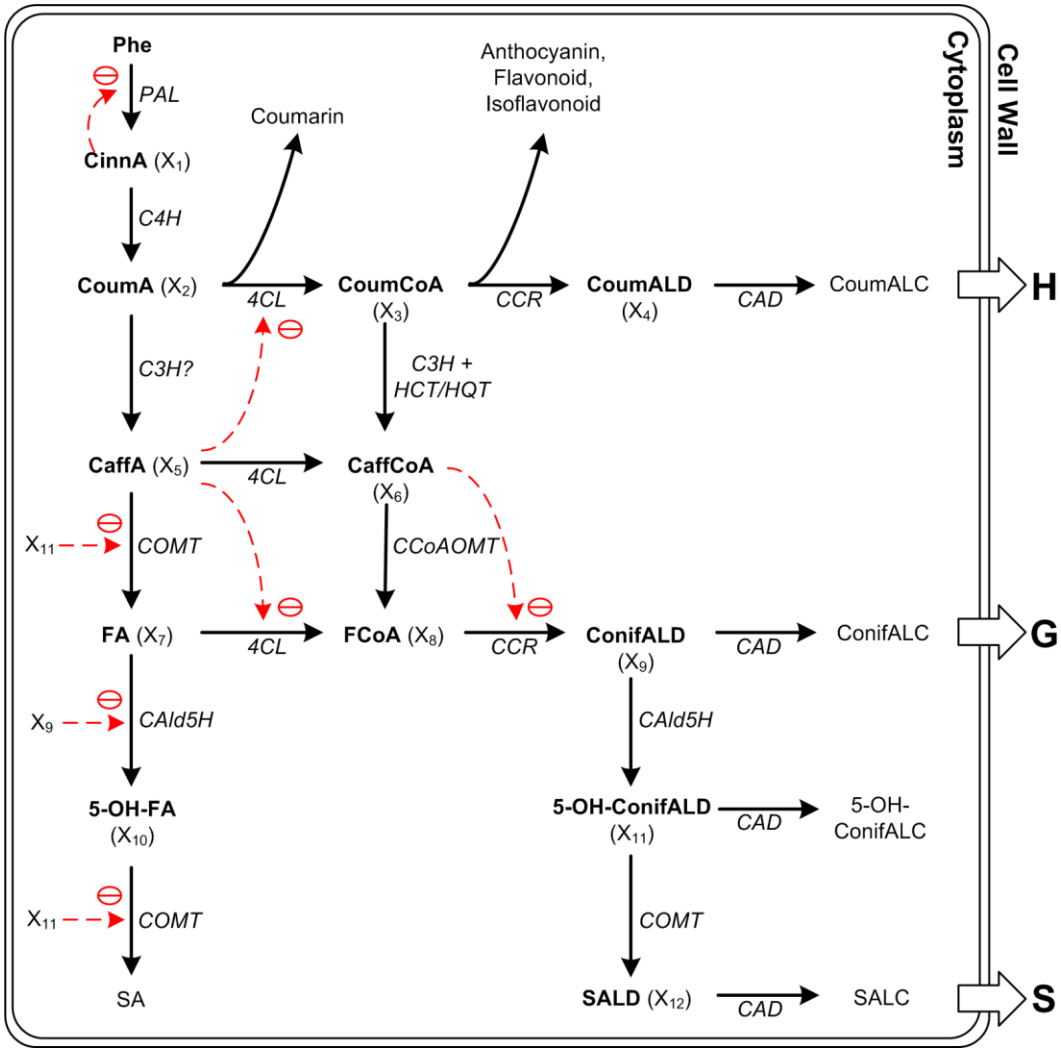


**Figure 2.1: Generic metabolic map of the monolignol biosynthetic pathway in *Populus*.** Metabolites in bold are represented by dependent variables $X_i$, $i = 1,...,12$, while enzymes are shown in italics. Solid black arrows represent material flow, whereas dashed red arrows represent regulatory signals, with negative signs indicating inhibition. Transport processes of monolignols into the cell wall are shown as open arrows.

**Table 2.1: Documented regulatory signals within the monolignol biosynthetic pathway in *Populus***

| Enzymes | Substrate | Regulator | Kinetics (μM) | Reference |
|---|---|---|---|---|
| PAL | Phenylalanine | Cinnamic acid | N/A[c] | [69] |
| 4CL | *p*-coumaric acid | Caffeic acid[a] | $K_I = 4.37$ | [68] |
| | Ferulic acid | | $K_I = 4.17$ | |
| CCR | Feruloyl-CoA | Caffeoyl-CoA[a] | $K_I = 15.3$ | [70] |
| COMT | Caffeic acid | 5-hydroxyconiferyl aldehyde[a] | $K_I = 2.1$ | [71] |
| | 5-hydroxyferulic acid | | $K_I = 0.26$ | |
| CAld5H | Ferulic acid | Coniferyl aldehyde[b] | $K_I = 0.59$[d] | [71,72] |

[a]Competitive inhibitor. [b]Non-competitive inhibitor. [c]No direct evidence has yet been found in *Populus* for this otherwise well-known feedback regulation at the entrance of the pathway. [d]Although this regulation has been experimentally demonstrated in aspen, no quantitative details are known, and the kinetic parameter presented here was measured in the lignifying tissues of sweetgum.

Since the regulatory signals affect several locations within the pathway, their overall effects are difficult to predict and may even be the cause for counterintuitive observations in transgenic plants. For instance, Fang *et al.* [25] recently found lignin monomer compositions that cannot be explained solely by the pathway topology in transgenic alfalfa lines with reduced activities of either cinnamate 4-hydroxylase (C4H) or caffeoyl-CoA *O*-methyltransferase (CCoAOMT).

In conclusion, the complexity and multitude of regulatory features that characterize the monolignol biosynthetic pathway render intuitive assessments problematic and highlight the need for mathematical models capable of explaining the functionality of the pathway system.

### 2.2.2 Experimental Data

The data supporting our modeling effort come in different forms. First, we collected kinetic information and metabolite concentrations from the literature (Tables A.1 and A.2). Secondly, we found pertinent information in five studies of transgenic poplars or aspens, each of which investigated the responses of the pathway to modified protein levels. The investigated proteins were COMT, cinnamyl alcohol dehydrogenase (CAD) [73], 4-coumarate:CoA ligase (4CL), coniferyl aldehyde 5-hydroxylase (CAld5H) [74], and CCoAOMT [75] (Table 2.2). Among these transgenic experiments, three reported an explicit change in the relative proportion of S to G monomers (the so-called *S/G ratio*), as determined by thioacidolysis. Because lignin content [25] and the S/G ratio [6] are related to the degree of recalcitrance, we will use this ratio as a target indicator of the system's response to genetic manipulations.

Several cautionary notes are in order when we interpret the S/G ratio. First, one should bear in mind that only the fraction of monomers connected by β-O-4 linkages, which accounts for only 20-40% of the lignin by weight, can be extracted by thioacidolysis. Second, many of the intervening events, for example, during the transport process or dehydrogenative polymerization, may also contribute to the differences in the observed S/G ratios, but mechanistic details are currently unclear [76]. Third, the composition of lignin monomers is significantly different between two major cell types of xylem tissue, with vessel elements enriched in G monomers and fibers in S monomers [77]. Lastly, genes coding for enzymes like CCoAOMT are expressed in developing vessels but not in fibers, suggesting that different routes to monolignol biosynthesis might be favored in different types of cells[2] [23].

---

[2] Ideally, a comprehensive analysis of the lignin monomer synthesis in xylem should consist of at least two distinct models, representing the two cell types. The numerical results for any physiological feature of

**Table 2.2: Pertinent details of transgenic experiments in *Populus***

| Enzyme | Enzyme activity (in relation to wild-type) | Lignin composition (S/G; in relation to wild-type) | Species |
|---|---|---|---|
| COMT[a] | 32% | 25% | Poplar |
| CAD[a] | 15% | 100% | Poplar |
| 4CL[b] | 10% | 100% | Aspen |
| CAld5H[b] | 280% | 250% | Aspen |
| CCoAOMT[c] | 10%[d] | 111% | Poplar |

The relative proportion of S to G monomers (S/G) was measured by thioacidolysis, which releases the monomers by selectively cleaving the β-O-4 bonds.
[a][73]. [b][74]. [c][75]. [d]This particular quantity refers to 10% of wild-type protein amounts.

### 2.2.3 Mathematical Models

We pursued a two-step approach, using complementary methodologies from flux balance and dynamic-kinetic analysis. An overview of the strategy is shown in Figure 2.2. First, we converted the pathway (Figure 2.1) into a stoichiometric model and used flux balance analysis (FBA) to study phenotypes under different types of constraints [36]. As described in Chapter 1, the central concept of FBA is a balanced flux distribution at steady state, along with numerous physico-chemical constraints and an optimization objective like maximal growth. Fast population growth is a reasonable objective for microbial populations, but it is not pertinent here and must be supplanted with different constraints.

Two types of constraints were used here. First, the capacity of each flux $v_i$ must lie within its physiological range $\alpha_i \le v_i \le \beta_i$, where we allow $\alpha_i = 0$, and where $\beta_i$ may

---

interest, such as the S/G ratio, could then be approximated by combining the two estimates in proportion to their percentage of volume in xylem. While our model could easily be adapted to the two scenarios, currently available data do not allow us to account for such details, and our results therefore reflect averages.

be defined as the maximum rate (*i.e.*, $V_{max}$ as in a conventional rate law like the Michaelis-Menten function). Here, all fluxes are assumed to be unbounded (*i.e.*, $\beta_i$ is defined as $+\infty$), except for the three steps catalyzed by COMT, which are the only reactions for which kinetic constants ($K_M$ and $V_{max}$) have been characterized for *Populus* protein. While the bounds narrow the range of admissible solutions, they are not stringent enough to identify the optimal solution.

The second constraint is based on the assumption that lignified tissue like xylem has evolved to maximize lignin production in a species- and cell type-specific ratio of monolignols. This assumption is at least partially supported by the observation in poplar xylem that two of the three phenolic glucosides—the storage or detoxification products of hydroxycinnamic acids along the metabolic route to the synthesis of sinapic acid—are barely detectable [75]. This finding suggests that the physiological objective of this pathway is to produce its other end products, namely monolignols. However, this piece of evidence must be handled with caution because the same phenolic glucosides, along with other phenolic compounds like flavonoid and chlorogenic acid, can be abundant in leaves or developing stem tissues [78]. If available, measurements such as the relative amounts of phenolic compounds derived from the pathway are essential for defining the physiological objective within a different context.

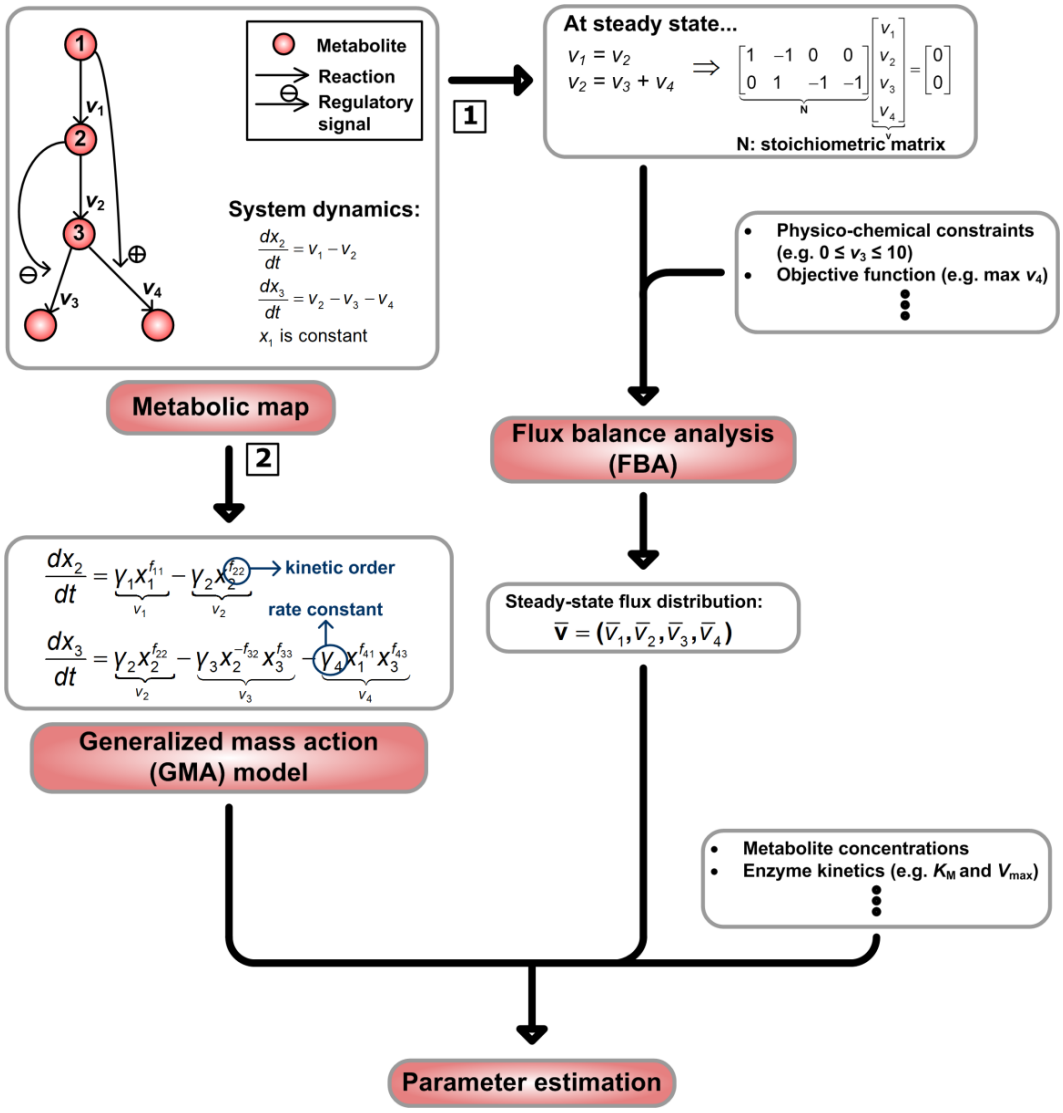


**Figure 2.2: Overview of the two-step modeling approach.**
The two-step modeling approach is illustrated here generically by a system with two dependent variables ($x_2$ and $x_3$) and one independent variable ($x_1$). At steady state, the four fluxes within the system are balanced at both intermediate nodes, resulting in two linear equations with four unknown variables (fluxes). With additional physico-chemical constraints, flux balance analysis (FBA) yields a steady-state flux distribution that satisfies all imposed conditions while optimizing an objective function. Alternately, the same system can be translated into a nonlinear Generalized Mass Action (GMA) model that is characterized by two types of parameters: kinetic orders and rate constants. Collectively, all data, including the steady-state flux distribution, metabolite concentrations, and enzyme kinetic data, are used to estimate the parameters of the ensemble of dynamic models.

The mathematical representation of the physiological objective of monolignol maximization leads to an objective function of the form $\sum_{j \in \Gamma} v_j$, where $\Gamma$ is the set of fluxes representing the production of the three pertinent monolignols, namely, *p*-coumaryl alcohol, coniferyl alcohol, and sinapyl alcohol. These three fluxes are further constrained by equality constraints such that the corresponding lignin monomer composition reflects the thioacidolysis yields from poplar stem ([79]: Table 3). Mathematically, this modeling approach results in a specific formulation that can be solved with methods of linear programming for which a large number of computational routines exist. In the end, this FBA approach reveals an optimal flux distribution at steady state, and the only inputs needed are the pathway stoichiometry, enzyme capacity measurements, and a lignin monomer composition that corresponds to experimental finding.

Various regulatory signals have been identified within the monolignol biosynthetic pathway (Table 2.1). The mechanisms introduce nonlinearities in the system for which steady-state models like FBA are not sufficient. In the second step of our two-step approach, we therefore use Generalized Mass Action (GMA) models within the framework of Biochemical Systems Theory (BST) (see Chapter 1 for detail) to account for the documented regulation of the pathway at the metabolic level. The characteristic feature of BST models is the representation of metabolic fluxes as products of power functions; if an enzyme-catalyzed reaction had been quantified before as a Michaelis-Menten, Hill, or other similar rate law, it is mathematically easy to convert it into a power-law function [44] (see also Section A.1.2).

### 2.2.4 Parameter Estimation

The GMA model for the monolignol biosynthetic pathway consists of 12 dependent variables ($X_1,...,X_{12}$ in Figure 2.1), representing the intermediate metabolites involved in the production of monolignols, and one independent variable, representing the concentration of the initial substrate phenylalanine. As indicated earlier, two types of parameters need to be estimated: kinetic orders $f_{i,j}$ and rate constants $\gamma_k$. Here, 27 kinetic orders and rate constants are unknown. In general, estimation tasks with such a large number of parameters are computationally intensive and time-consuming. Using the GMA formulation, however, confers two advantages. First, it is relatively easy to derive parameter values of GMA models, especially for kinetic orders, if information regarding the kinetic features of enzymes and metabolite concentrations is available (*cf.* Section A.1.2). Second, the steady-state flux distribution estimated per FBA helps us circumvent the problem of determining rate constants in the absence of specific flux measurements.

As an example, consider a Michaelis-Menten process $V(X)=V_{\max}X/(K_M+X)$ where the maximum rate $V_{\max}$ is unknown. Given a steady-state substrate concentration $S$ and the FBA-predicted steady-state flux $V_{FBA}$, the rate constant $\gamma$ for the corresponding power-law term can be represented as:

$$\gamma = V_{FBA}S^{-f}, \tag{2.1}$$

where

$$f = \frac{K_M}{K_M + S} \tag{2.2}$$

is the kinetic order with respect to the substrate. Similar derivations can be applied to conventional rate laws describing competitive or non-competitive inhibition. Details of these types of estimations have been discussed extensively in the literature [27,44] and will not be repeated here.

Once the model is parameterized (that is, all parameters are assigned values), the first priority is to ensure that no parameter affects the pathway unreasonably strongly. Using sensitivity analysis, we confirmed that the system is indeed robust at the steady state we obtain with FBA (data not shown), indicating that only minor fluctuations in metabolite concentrations and fluxes result from slight changes in parameter values. While a favorable outcome, this robustness is no guarantee that the model is correct. In fact, many parameter values derived from the available data might not be reliable because roughly half of the intermediate metabolites, including the CoA esters, have rather low concentrations *in vivo* and are thus difficult to measure with precision (Wout Boerjan, personal communication). Computationally, we can explore this uncertainty by systematically changing all parameter values thousands of times and studying how the system responds to such changes. For validation purposes, the observed changes in the S/G ratio from transgenic experiments in poplar or aspen (Table 2.2) can serve as a quality criterion. To make optimal use of the transgenic experiments for our parameter estimation task, we developed a novel approach consisting of two steps, namely, (1) identification of a subset of significant parameters, and (2) optimization of their values. The steps are summarized in Figure 2.3A and discussed in the following.

First, we need an objective criterion to answer the fundamental question of what constitutes a significant parameter. For any transgenic experiment in our particular context, a parameter is deemed significant if a modest change in its value considerably affects the S/G ratio. To approximate this degree of influence by statistical measures such as Pearson's correlation coefficient or mutual information, we generated a large population of GMA models with different parameter settings, where each parameter (kinetic order) was uniformly sampled from a physiologically realistic range. Given the FBA-derived steady-state flux distribution and the randomly generated values for all kinetic orders, we adjusted each rate constant so that the power-law representation of a flux matched the FBA-derived steady-state value. Typically, the resulting values of

kinetic orders are within the range of 0 and 1, if they are associated with substrates, enzymes, and activators, whereas inhibitors are often associated with kinetic orders within the range of -1 and 0 (see [44]: Chapter 5). The range of 0 and 1 is also consistent with enzyme-catalyzed reactions following a Michaelis-Menten rate law (Figure 2.6A).

With a much reduced number of significant parameters, we gain two important benefits: 1) a reduction—although not total elimination—of the risk of over-fitting; and 2) improved convergence in subsequent parameter optimization tasks, because smaller numbers of parameters are obviously easier to estimate than large numbers. As mentioned earlier, physiological data of the monolignol biosynthetic pathway are available as one-time measurements of the S/G ratio in a number of transgenic experiments. Consequently, our second step—parameter optimization—consists of finding values for those significant parameters that minimize the sum of squared errors (SSE) between the measured and the predicted S/G ratios of all transgenic experiments.

Moreover, we characterize an ensemble of GMA models such that all members have comparable training errors in terms of SSE. This notion of finding not just a single best model, but an entire class of competent fits, is inspired by the argument that inter-individual differences among organisms are reflected in slightly or even moderately different parameter profiles [80]. The search for classes of solutions has also been supported in other scientific domains as diverse as simulations of climate change [81] and models of gene regulatory networks [82] and cell signaling pathways [83].

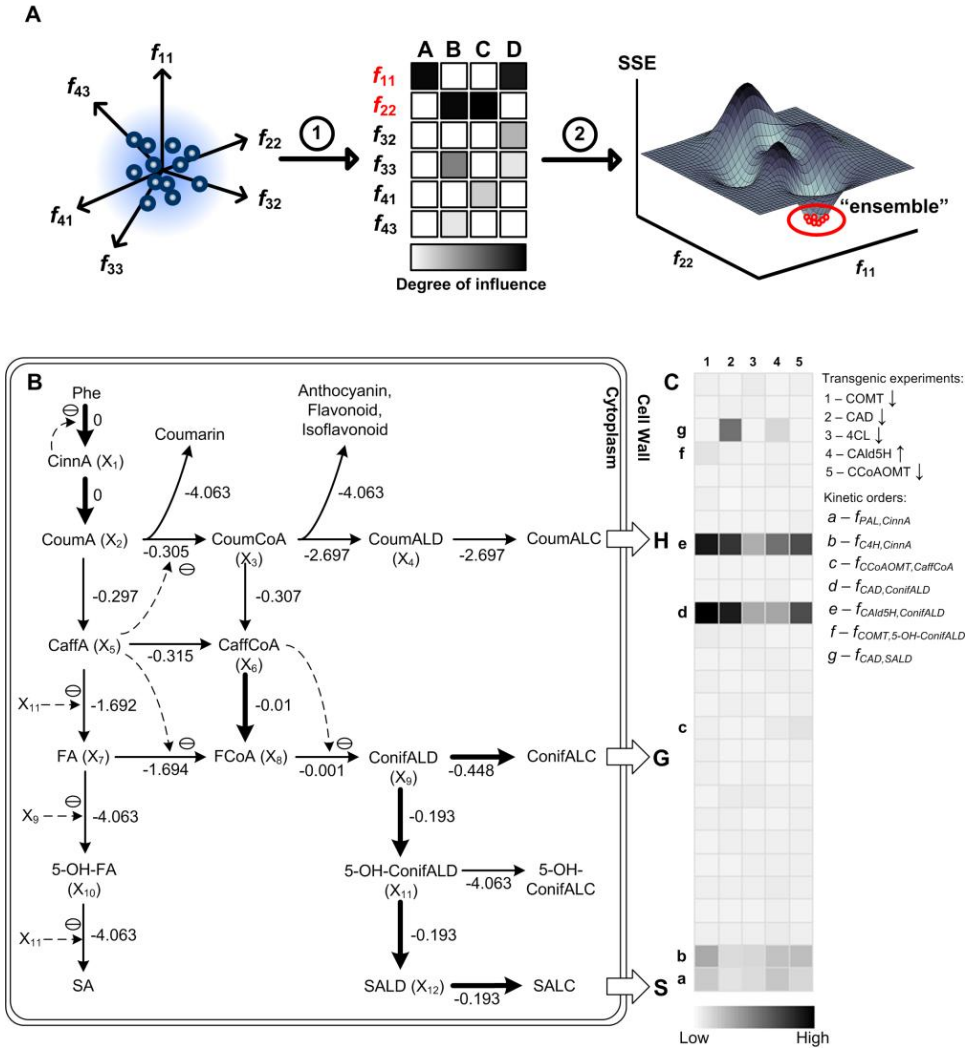


**Figure 2.3: Steps of parameter estimation.**
**(A)** Steps of the parameter estimation process using the system in Figure 2.2: (1) constrain each parameter (kinetic order) to a physiologically realistic range and simulate the transgenic experiments in the training set with thousands of randomly sampled parameter profiles; (2) compute a statistical measure (Pearson's correlation coefficient or mutual information) between each parameter and the S/G ratio for all transgenic experiments (A-D) and select statistically significant parameters; (3) values for significant parameters are further optimized to minimize the SSE, and to find an ensemble of models with comparably low SSE. **(B)** The numerical value next to each reaction represents the magnitude (on a base-10 logarithmic scale) of its steady-state flux, normalized by the input of the pathway, which consists of the reaction converting a constant supply of phenylalanine (Phe) into cinnamic acid (CinnA). **(C)** The intensity of each box reflects the mutual information (MI) between one parameter and the S/G ratio of one transgenic experiment. Kinetic orders with a statistically significant mutual information score (but not always leading to a large value, such as the kinetic order 'c') are listed on the left of their respective rows. Pathway reactions associated with these significant kinetic orders are represented by heavier arrows.

### 2.2.5 Pathway Optimization

Our model of monolignol biosynthesis has the great advantage that it integrates diverse pieces of information from varying experimental conditions. It can be used to address questions like which enzymes should be modified—whether by modulating their expression levels or by improving their turnover activities through directed evolution [84]—to achieve a higher yield of a desired product. Within the context of biofuel production, genetically engineered crops should of course release significant amounts of fermentable sugars that can be converted into ethanol or other biofuel chemicals. In a study on *Populus*, Davison and co-workers [6] indicated that both lignin content and the S/G ratio have significant effects on the yield of xylose, and that a small decrease in S/G ratio alone results in a statistically significant increase in xylose yield. Using our ensemble of GMA models as a framework, we therefore focus on identifying enzymes whose expression levels might allow reductions in the S/G ratio.

GMA models are generally advantageous for modeling the monolignol biosynthetic pathway, but are not trivially optimized with respect to yield because their steady states cannot be computed analytically. This limitation may be overcome with an indirect optimization method (IOM) that permits optimization in an iterative, much simplified manner [85]. Specifically, IOM allows us to transform the nonlinear problem of minimizing the S/G ratio (or the ratio of fluxes producing coniferyl and sinapyl alcohols), into an iterated linear optimization problem that can be solved with various standard methods, including linear programming. Pertinent details about this approach can be found in Section A.1.6.

## 2.3 Results

The FBA analysis resulted in an optimal flux distribution within the metabolic pathway system (Figure 2.3B) that led to the maximal production of three monolignols in the correct composition. Interestingly, this optimal solution shows that several reactions with relatively high steady-state fluxes dominate the activity of the pathway, whereas other reactions are seemingly inactive. If we connect the dominant fluxes whose steady-state values are within one order of magnitude of the phenylalanine consumption, the resulting route is almost identical to the currently alleged structure of the monolignol biosynthetic pathway in angiosperms [71]. Thus, the purely computational result from the FBA analysis reinforces the point that metabolic pathways are seldom fully connected and indeed use sparse connectivity to bring about specific function. This phenomenon has been widely discussed for microbial metabolic networks [86,87], but our results seem to indicate that the same may be true in plant secondary metabolism as well.

Next, we used the optimal steady-state flux distribution from FBA to construct a dynamic GMA model of the pathway. Converting the metabolic map (Figure 2.1) into a symbolic model in GMA format does not take much effort; in fact, this can be done automatically with customized software [88]. The much more difficult step, however, is the numerical identification of parameter values, which is outlined in Figure 2.3A and discussed in detail below.

First, by adapting a grid search method used by Alves and collaborators [89], we uniformly sampled every parameter (kinetic order) from a predetermined range of values and generated thousands of GMA models with the same FBA-derived steady-state flux distribution. For each instantiation, we checked local stability (Section A.1.4) and discarded unstable parameter profiles. Next, we computed the mutual information (Section A.1.5) of each parameter and the output feature of interest, namely the S/G ratio, to evaluate the relative significance of individual parameters (Figure 2.3C). Not

surprisingly, most parameters are not statistically significant, indicating that only a few parameters have an appreciable influence on the S/G ratio in each transgenic experiment.

Notably, two parameters representing the direct influence of coniferyl aldehyde on its own consumption, $f_{CAD,ConifALD}$ and $f_{CAld5H,ConifALD}$, are statistically significant in all five transgenic experiments. Although the identification of significant parameters in our strategy is more or less "biologically blind," this result can easily be interpreted in terms of the logic of the pathway topology: as shown by FBA and also by thioacidolysis yield, the flux leading to the synthesis of 5-hydroxyconiferyl alcohol is negligible, which means that the formation of 5-hydroxyconiferyl aldehyde or coniferyl alcohol from coniferyl aldehyde is arguably the principal branch point where the S/G ratio is determined.

In the second half of the parameter estimation process, we generated an ensemble of GMA models that reproduced a training set of experimental results, using a simulated annealing (SA) algorithm (Section A.1.7) to find optimal values for the significant parameters. For the five transgenic experiments used as training data (Table 2.2), the S/G ratios predicted by the ensemble of models are highly consistent with the experimental measurements (Figure 2.4). The relative errors in two experiments, where either COMT or CCoAOMT is down-regulated, are slightly greater than the corresponding experimental errors (~3%). Considering that only a handful of transgenic experiments are available for training the models, this level of variance is better than one might have expected.

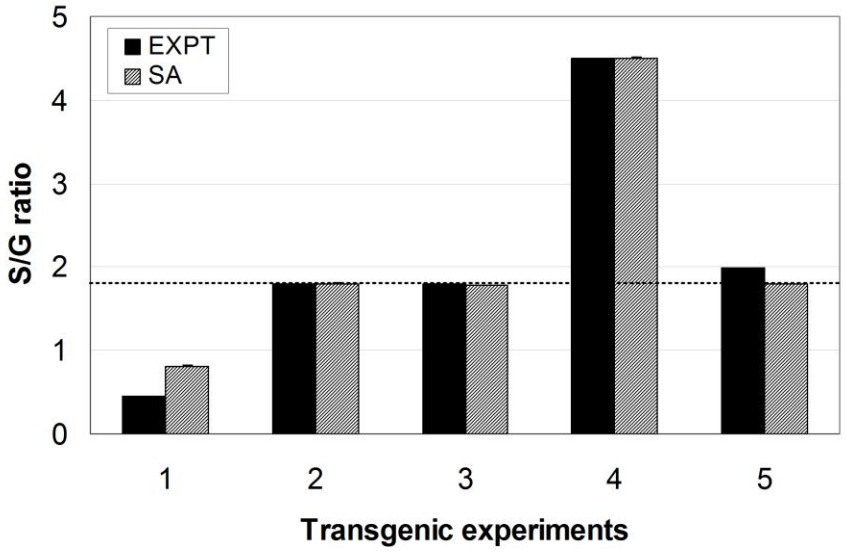


**Figure 2.4: Simulation results of five transgenic experiments used as training data.** Each vertical bar represents either the experimentally observed S/G ratio (EXPT) or the mean of 20 predictions from the ensemble of GMA models fitted by a simulated annealing (SA) algorithm. The transgenic experiments are numbered as in Figure 2.3C, with the dashed line featuring the assumed wild-type value (1.8) and the error bar indicating the 95% confidence interval for the mean.

To assess the reliability of the computed ensemble of models, we used the ensemble to simulate two transgenic experiments not used for training. Specifically, one of the experiments studied a multi-gene co-transformation where the 4CL enzyme activity was reduced by 80% and the CAld5H enzyme activity increased by 2.1-fold [74]. As shown in Figure 2.5, the predicted S/G ratio follows the same upward trend and even falls within ~20% of the observed value. In the second transgenic experiment, the CCR transcript levels were severely decreased to < 5% of the wild-type levels [61]. Again, the observed S/G ratio was predicted accurately by the ensemble of models.

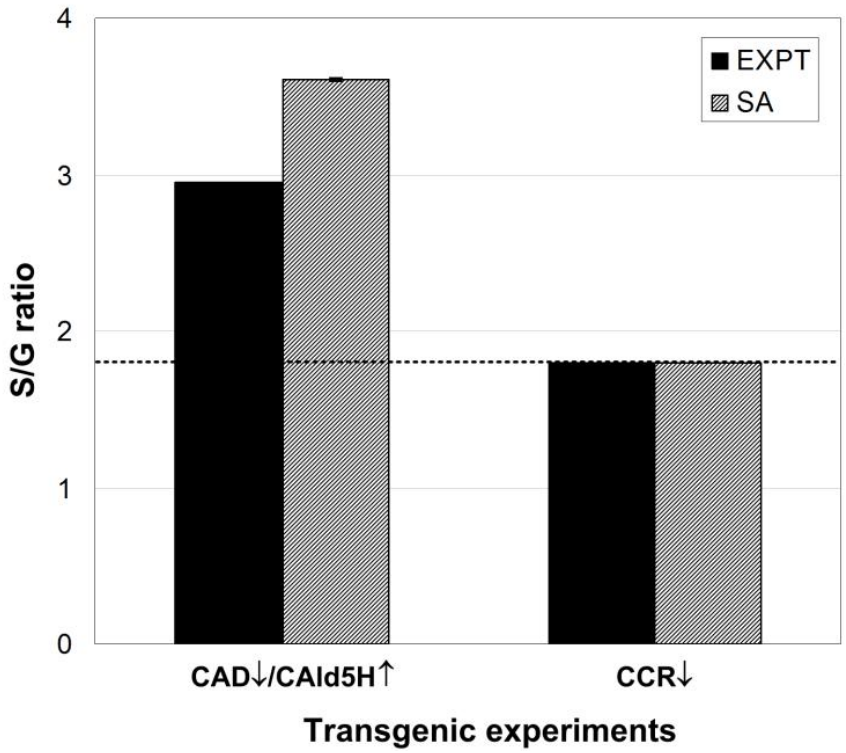


**Figure 2.5: Simulation results of two transgenic experiments not used for model training.**
As in Figure 2.4, each vertical bar represents either the experimentally observed S/G ratio (EXPT), or the mean of 20 predictions from the ensemble of GMA models fitted by SA. Also, the dashed line features the assumed wide-type value (1.8), and the error bar indicates the 95% confidence interval for the mean. For the CCR down-regulation experiment, the confidence interval is so small (~$10^{-5}$) that it is nearly invisible.

Beyond its good agreement with the experimental results, the ensemble of GMA models permits further mechanistic insights. For instance, most of the significant parameters with positive values (which are thus associated with substrates or activators) have optimal values between 0.4 and 0.7, a typical range for kinetic orders estimated from Michaelis-Menten reactions operating close to the $K_M$ (Figure 2.6; see also [44]: Chapter 5). By contrast, both $f_{COMT,5-OH-ConifALD}$ and $f_{CAD,SALD}$ take on very small values within the ensemble of models, which according to the theory behind GMA models suggests that both the *O*-methylation of 5-hydroxyconiferyl aldehyde and the reduction of sinapyl aldehyde to sinapyl alcohol operate at an essentially constant rate that is almost

independent of fluctuations in their substrate concentrations. Although there has not yet been direct evidence for this predicted operation close to saturation, one notices that the nominal concentration of sinapyl aldehyde in wild-type poplar is much greater than the reported Michaelis constant of its CAD-catalyzed reduction to alcohol (see Tables A.1 and A.2 for specific values), which is directly consistent with our model deduction.

Interestingly, the distributions of optimal parameter values reveal a linear relationship between $f_{CAD,ConifALD}$ and $f_{CAld5H,ConifALD}$ (Figure 2.7). As discussed in more detail in Section A.1.3, this co-linearity implies that the ratio between the corresponding fluxes remains unchanged over time and is thus equal to the steady-state value obtained from FBA. More importantly, a constant ratio between these two fluxes suggests that the S/G ratio might be insulated from any genetic modulation prior to the reactions involving coniferyl aldehyde, provided that the synthesis of 5-hydroxyconiferyl alcohol is negligible. In fact, this is exactly what happens in transgenic experiments where 4CL (Figure 2.4) or CCR (Figure 2.5) is down-regulated. Even if the situation is not as expected in the CCoAOMT down-regulation experiment (Figure 2.4), the observed S/G ratio is raised only by ~11% despite a 90% decrease in the CCoAOMT protein level.

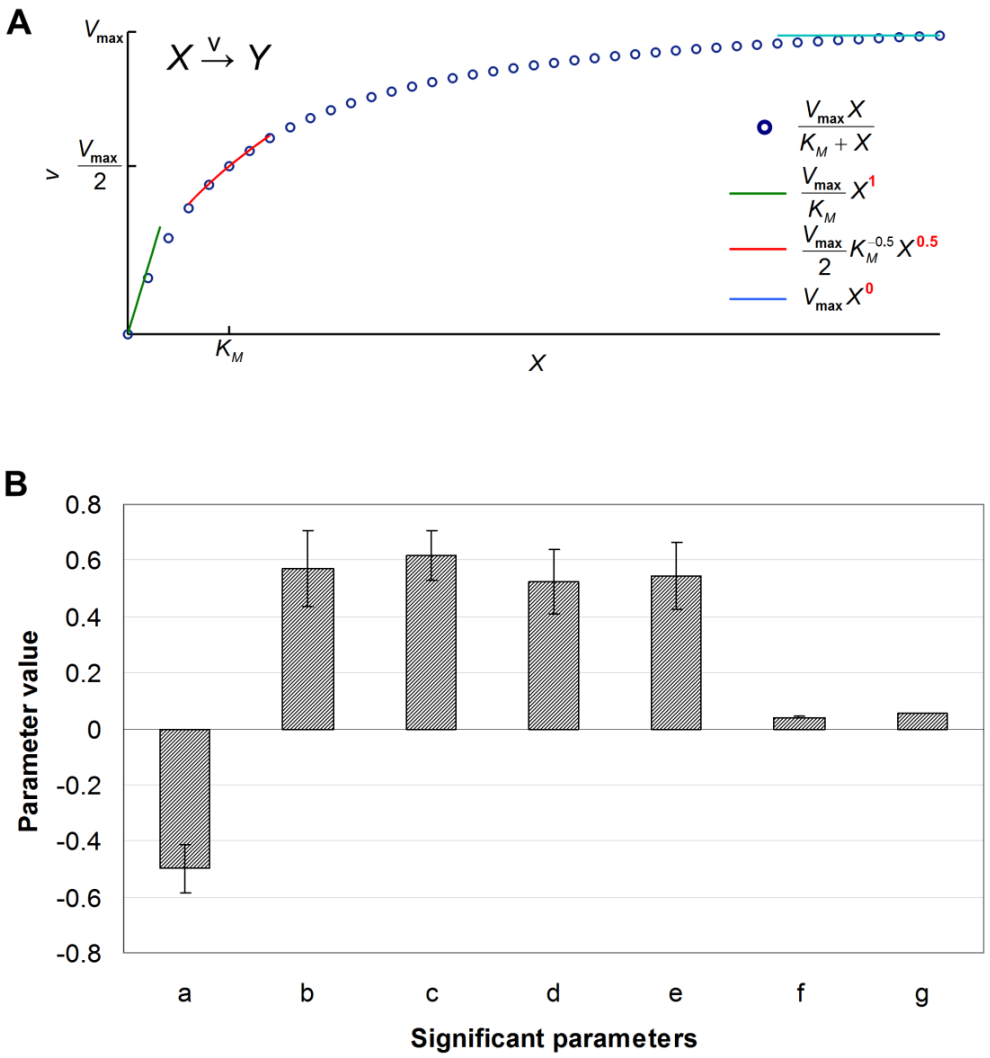


**Figure 2.6: Illustration of kinetic orders derived from a Michaelis-Menten function and distributions of values for seven significant parameters within the ensemble of GMA models.**

**(A)** The kinetic order (red number) in each power-law representation of a Michaelis-Menten function is within the range of 0 and 1, with the specific value depending on the assumed *in vivo* concentration of *X*. If the reaction operates at a point where the concentration of *X* is much greater than $K_M$, the corresponding power-law representation has a kinetic order close to zero, implying that the reaction rate is almost saturated and therefore unaffected by the concentration of *X*. **(B)** As in Figure 2.4, the height of a vertical bar is proportional to the mean value of a significant parameter within the ensemble of models fitted by SA, with the error bar representing the 95% confidence interval for the mean. See Figure 2.3C for the identity of each parameter.

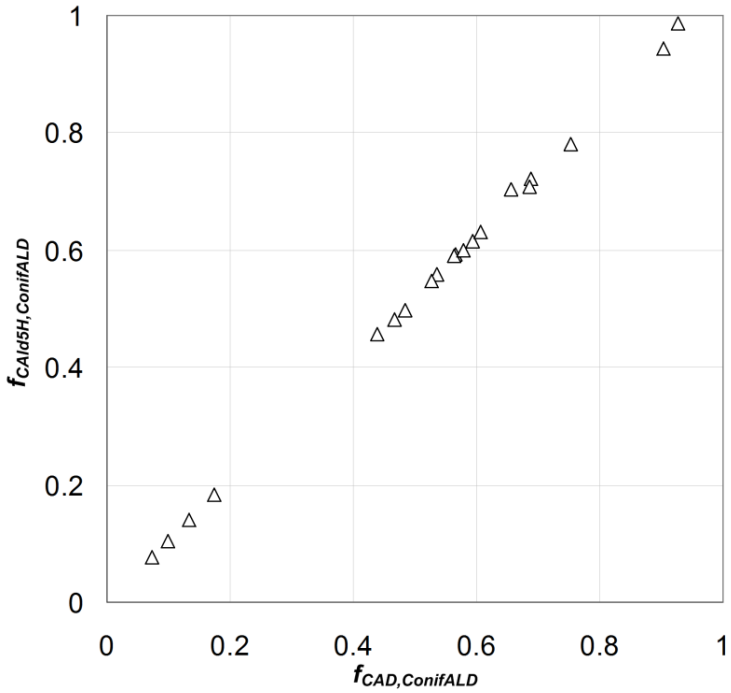


**Figure 2.7: Plot of $f_{CAld5H,ConifALD}$ against $f_{CAD,ConifALD}$.**
Each point represents the pair ( $f_{CAD,ConifALD}$ , $f_{CAld5H,ConifALD}$ ) found in one GMA model within the ensemble obtained through simulated annealing.

With an ensemble of models that seems to be qualitatively adequate, we can now apply the IOM approach to minimize the S/G ratio of the monolignol biosynthetic pathway toward a higher yield of xylose. Normally, IOM can be implemented in many different ways. The most common scenario is that all enzymes (genes) involved in the pathway are accessible to manipulations, which unfortunately is not feasible with current biotechnological techniques in plants [90]. Instead, we mimic the current state of the art (Fang Chen, personal communication) by allowing only one, two, or three enzyme activities to be altered between 5% and 5 times the basal levels. Furthermore, we enforce physiological constraints that are necessary for plant viability and that are discussed in Section A.1.6.

The optimization results (Table 2.3) indicate that by altering the activity levels of three enzymes in prescribed amounts, the S/G ratio predicted by the ensemble of models

can be reduced from about 1.8 to about 1.11—a significant decrease that far exceeds the natural variation observed in poplar [6]. Moreover, by modulating just one enzyme (CAld5H), we can already achieve ~60% of the maximal reduction that is obtained when three enzymes are manipulated. In other words, the S/G ratio is predicted to decrease from 1.8 to about 1.39 if one down-regulates the enzyme activity of CAld5H by one quarter. Overall, the optimized solutions require only a moderate degree of modulation of the selected enzymes (from approximately 70% to 4.3 times the wild-type activity levels), which are well within the range of modern recombinant DNA techniques.

## 2.4 Discussion and Conclusions

The application of mathematical modeling to studies of the monolignol biosynthetic pathway, or of plant secondary metabolism in general, has not yet attracted much attention, especially when compared with central metabolism in microorganisms. One reason is that the *in vivo* concentrations of secondary metabolites are often low and difficult to measure, which makes quantitative modeling difficult.

**Table 2.3: Minimization of the S/G ratio using the IOM approach[a]**

| No. of enzymes | Modified enzymes[a] | IOM solution[c] (S/G) |
|---|---|---|
| 1 | CAld5H (0.76) | 1.3886 |
| 2 | COMT (0.96) CAld5H (0.71) | 1.29 |
| 3 | C4H (4.31) CAD (1.67) CAld5H (1.34) | 1.1133 |

[a]Baseline S/G ratio is 1.8.
[b]Numbers in parentheses represent the optimized ratio of change in enzyme activities related to the wild-type levels.
[c]Average values of the ensemble of model

In this work, we used diverse types of data to pursue a two-step model analysis of the monolignol pathway, using both Flux Balance Analysis (FBA) and Biochemical Systems Theory (BST). These two approaches had so far not been combined in the construction of a dynamic model. Thus, we first constructed an initial, coarse FBA model and used it in a second phase as a constraint for developing fully parameterized nonlinear BST models. The result of this dual procedure was an ensemble of models that yield interesting qualitative insights into the topological and regulatory properties of monolignol biosynthesis. These models also lead to simulation results and predictions that are quantitatively consistent with experimental measurements that were either used for model training or validation. This concordance is quite striking, because the data and information supporting the models are rather scarce and involve a number of assumptions. Two reasons seem to be responsible for the good performance of the model in predicting the outcomes of validation experiments. The first is the proven robustness of BST models, which is manifest in low model sensitivity with respect to most parameters, as long as the connectivity and regulatory structure of a system is adequately captured by the model equations. The second reason is our strategic, severe model reduction, which effectively eliminated many parameters which we had proven to be relatively inconsequential.

Because we used all available metabolite concentrations and S/G ratios in transgenic experiments, either to estimate unknown parameters or to validate our models, it is presently not feasible to try improving the model further with purely computational means. To construct a "crisper" mathematical model in the future, specific data of the following types will be very helpful. At the metabolic level, intracellular metabolite concentrations, *in vitro* assays of individual enzymes, and perhaps intracellular flux measurements from dynamic labeling experiments [91] are in dire need. As demonstrated in our parameter estimation approach, these data should ideally be accompanied by

measurements of lignin monomers from transgenic plants with various genetic modulations of monolignol biosynthesis.

Another source of relevant information will come from gene expression data and specifically from microarray analyses, which have already revealed distinct transcriptional regulation patterns in genes encoding lignin biosynthetic enzymes at different developmental stages [92]. At present, the growth periods in different transgenic experiments span from several months to years, but it is implicitly assumed that enzyme activities are more or less constant. Future experiments and models should account for (slowly) changing levels of enzyme activities over the course of xylem formation during primary and secondary growth. Furthermore, since most reactions within the pathway are catalyzed by several isozymes, changes in gene expression should be confirmed with measurements of changes in enzyme activities. As a first approximation, the number of mRNA copies for each corresponding gene may be an indication of enzyme activity, but direct measurements would eliminate uncertainties associated with different splice variants and posttranslational modifications. Experiments and models should also focus on the dynamics of transcription factors, such as MYB and LIM, that have been found to coordinate the regulation of the expression of genes encoding lignin biosynthetic enzymes [93,94].

The proposed ensemble of models is clearly preliminary. Nevertheless, the models appear to be robust to modest variations in parameter values, are qualitatively consistent with five training experiments, and are even capable of semi-quantitatively reproducing the results of two validation experiments that had not been used for model construction. These initial successes are grounds for cautious optimism that the model might serve as a basis from which future developments may be launched.

As an illustration, we demonstrated one of its potential applications in genetic engineering, namely the optimization of the pathway toward a reduced S/G ratio and a higher yield of xylose. The results of this optimization seem to be reasonable in a sense

that all proposed changes in enzyme activities are modest and therefore implementable. The estimated improvements in the optimized system are actually very conservative compared with the 75% decrease in the S/G ratio observed in the COMT down-regulation experiment (Table 2.2). The reason for this discrepancy is that we imposed much more stringent bounds on metabolites than what is observed in the COMT down-regulation experiment. While wider bounds are clearly implementable in optimizations with the computational model and would result in much stronger reductions in the S/G ratio, large metabolite variations *in vivo* might lead to toxicity or reduced viability. Two explanations are possible for the observed 75% decrease in the S/G ratio. First, evidence indicates that metabolites that might be expected to accumulate in the cytoplasm are instead being transported to the cell wall and incorporated into lignin by so far unknown mechanisms [14], thereby precluding toxicity. Second, the observed variation in the S/G ratio may result from a change in the subcellular structure of pathway enzymes—or alleged "metabolic channeling" [24]—that is currently outside the scope of our GMA models. Taken together, the observed physiological response seems to suggest that our optimization settings might be overly cautious and that the S/G ratio could be reduced further than predicted.

As new data are being generated in the emerging field of plant systems biology, the next goal will be to integrate a wider variety of "omics" data from different organizational levels into the construction of multi-scale models that will be capable of predicting the physiological consequences of hypothetical transgenic experiments. Models of this capability will be particularly helpful as the corresponding experiments in actual trees are slow and laborious. The need to test model predictions, as well as proposed genetic engineering strategies, will not abate. However, once a model is sufficiently reliable, it may be able to screen out experiments that are unlikely to lead to improved outcomes.

# CHAPTER 3

# INTEGRATIVE ANALYSIS OF TRANSGENIC ALFALFA (*MEDICAGO SATIVA* L.) SUGGESTS NEW METABOLIC CONTROL MECHANISMS FOR MONOLIGNOL BIOSYNTHESIS[3]

## 3.1 Introduction

Although the generic sequences of metabolic reactions within the monolignol pathway have been identified, it is becoming increasingly clear that critical details of the pathway structure and its regulation are not entirely understood. As a case in point, Chen *et al.* [25] recently introduced systematic, transgenic alterations in alfalfa (*Medicago sativa* L.) plants by independently modifying the activities of seven key enzymes of monolignol biosynthesis. While many of the results were easily explained, down-regulation of caffeoyl coenzyme A 3-*O*-methyltransferase (CCoAOMT) had little effect on S lignin, an observation that is conceptually inconsistent with the commonly accepted pathway structure (Figure 3.1; black colored arrows). A recent study identified two isoforms of cinnamoyl CoA reductase (CCR), MtCCR1 and MtCCR2, in *Medicago truncatula* [59]. Furthermore, an earlier finding had suggested that caffeyl aldehyde is one of the preferred substrates for caffeic acid 3-*O*-methyltransferase (COMT) in alfalfa [60]. Taken together, these findings could imply an alternative route for S lignin synthesis (Figure 3.1; red colored arrows) upon CCoAOMT down-regulation [9,60].

---

[3] Adapted from: Lee, Y., Chen, F., Gallego-Giraldo, L., Dixon, R.A. and Voit, E.O. (2011) Integrative Analysis of Transgenic Alfalfa (*Medicago sativa* L.) Suggests New Metabolic Control Mechanisms for Monolignol Biosynthesis. *PLoS Comput. Biol.* 7(5): e1002047.

However, they cannot explain why only G lignin is decreased because feruloyl-CoA is a common precursor of both G and S lignin.

In dicotyledonous plants like alfalfa, the stem consists of many segments, called *internodes*. During maturation, all internodes grow asynchronously and thus independently represent different developmental stages. This phenomenon suggests a customized modeling approach: Instead of studying the pathway within a single developmental context, it seems advantageous to launch a systematic investigation that simultaneously encompasses dozens of internodes from seven wild-type or transgenic plants. This comprehensive approach circumvents the potential problem that regulatory mechanisms might escape discovery during an analysis based on singular phenotypic datasets, such as lignin content and monomer composition, if only one internode or one transgenic line is studied at a time. This potential failure to detect regulatory signals is exacerbated in the lignin system by the fact that several enzymes in the pathway catalyze multiple steps, which makes intuitive analyses difficult.

With a comprehensive analysis of several datasets as the target, we propose here a novel modeling approach that integrates the data in a semi-dynamic fashion. First, flux balance analysis (FBA) ([33]; Section 1.4.2) is applied independently in each individual internode of the wild-type plant. In contrast to microbial systems, where maximization of the growth rate is usually assumed to be the species' overall objective, we use the monolignol production as the objective function for FBA. Second, for every internode of a lignin-modified line, we use the method of minimization of metabolic adjustment (MOMA) [95] to characterize the altered flux distribution in relation to the corresponding FBA solution for the same wild-type internode. Specifically, the relative proportions of the fluxes leading to three lignin monomers are constrained at experimentally-observed values to improve the prediction. Finally, we perform a Monte Carlo-like simulation of randomly parameterized kinetic models in cases where the results arising from the static models may have alternative, kinetics-based explanations.

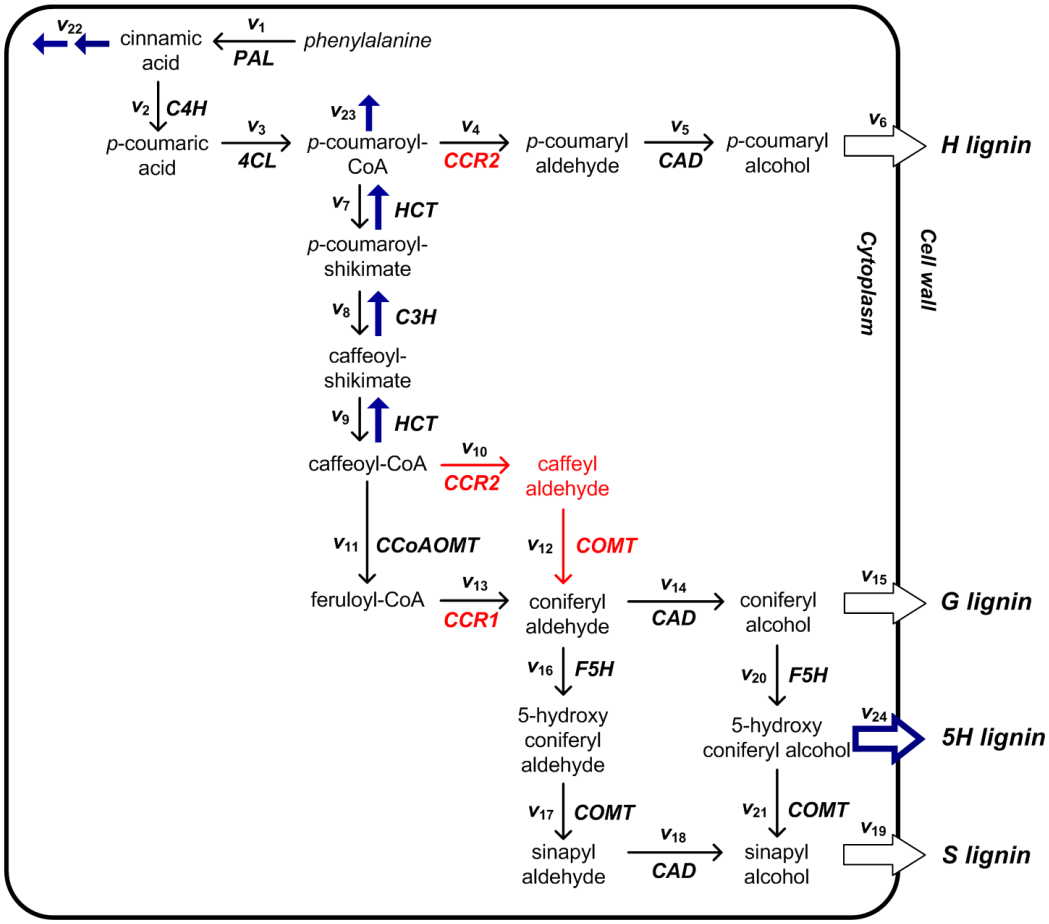


**Figure 3.1: Successive amendments of the metabolic pathways and transport processes leading to four hydroxycinnamyl alcohols in *Medicago*.**
The commonly accepted pathway of monolignol biosynthesis, which produces *p*-hydroxyphenyl (H), guaiacyl (G), 5-hydroxyconiferyl (5H), and syringyl (S) lignin monomers, is presented in black, with solid arrows representing metabolic conversions and open arrows collectively representing all events during the transport of monolignol precursors into the cell wall. Important revisions suggested by the recent identification of two CCR isoforms—CCR1 and CCR2—are colored in red and discussed in the text. Arrows colored in blue represent additional reactions and transport processes that are probably negligible in wild-type plants but found to become significant in some transgenic strains.

This combined modeling approach represents, to the best of our knowledge, the first computational study of lignin biosynthesis in angiosperm stem tissues and, more generally, of secondary plant metabolism in angiosperms. As we will discuss later, the model analysis resulted in six postulates concerning the metabolic control of monolignol biosynthesis that had not been considered at all or at least not in detail. These postulates address the reversibility of some enzymatic reactions, shed light on the hypothesis of independent pathways for the synthesis of G and S monolignols, and suggest a novel feedforward regulatory mechanism exerted by a cinnamic acid-derived compound. Of note is the fact that evidence in support of this last postulate has subsequently been obtained in laboratory experiments. By critically evaluating the transgenic data against a revised pathway structure in alfalfa, we hope these postulates will not only serve as guidelines for directing future experiments, but also provide mechanistic insights that will aid the design of combined genetic modification strategies toward the generation of bioenergy crops with reduced recalcitrance.

## 3.2 Results

### 3.2.1 FBA-Guided Elucidation of Three Principal Branch Points

Accounting for recent experimental observations, we adopted a revised pathway structure of monolignol biosynthesis in alfalfa stems that includes the CCR2-catalyzed reduction of caffeoyl-CoA to caffeyl aldehyde and the subsequent synthesis of coniferyl aldehyde by COMT (Figure 3.1: black and red colored reactions), as explained earlier. The pathway of monolignol biosynthesis contains a fairly small number of branch points, and it is known that flux partitioning at these branch points determines the ultimate transport fluxes $v_6$, $v_{15}$ and $v_{19}$ and thus the relative amounts of lignin monomers (*cf.* [96]). The FBA-derived steady-state flux analysis for wild-type plants supports this

argument. It suggests that variation in lignin composition from young to mature internodes is accomplished by modulating the flux partitioning at three principal branch points: *p*-coumaroyl CoA, coniferyl aldehyde, and coniferyl alcohol. As a paradigm illustration, the proportion of H lignin declines from 7% of the total monomer yields in the first two internodes to 1% in the eighth internode. This decline is singularly achieved through a monotonic decrease in $v_4$ (Figure 3.2A). A parallel increase in the ratio of S to G lignin—commonly termed the S/G ratio—from 0.09 in the first two internodes to 0.64 in the eighth internode requires a combined effort of flux adjustments at coniferyl aldehyde and coniferyl alcohol (Figure 3.2B). Since F5H controls the first committed steps (*i.e.*, $v_{16}$ and $v_{20}$) towards the synthesis of S lignin, one would expect to see its expression being up-regulated in mature versus young internodes, which has recently been validated by microarray analysis (Table 4 of [97]).

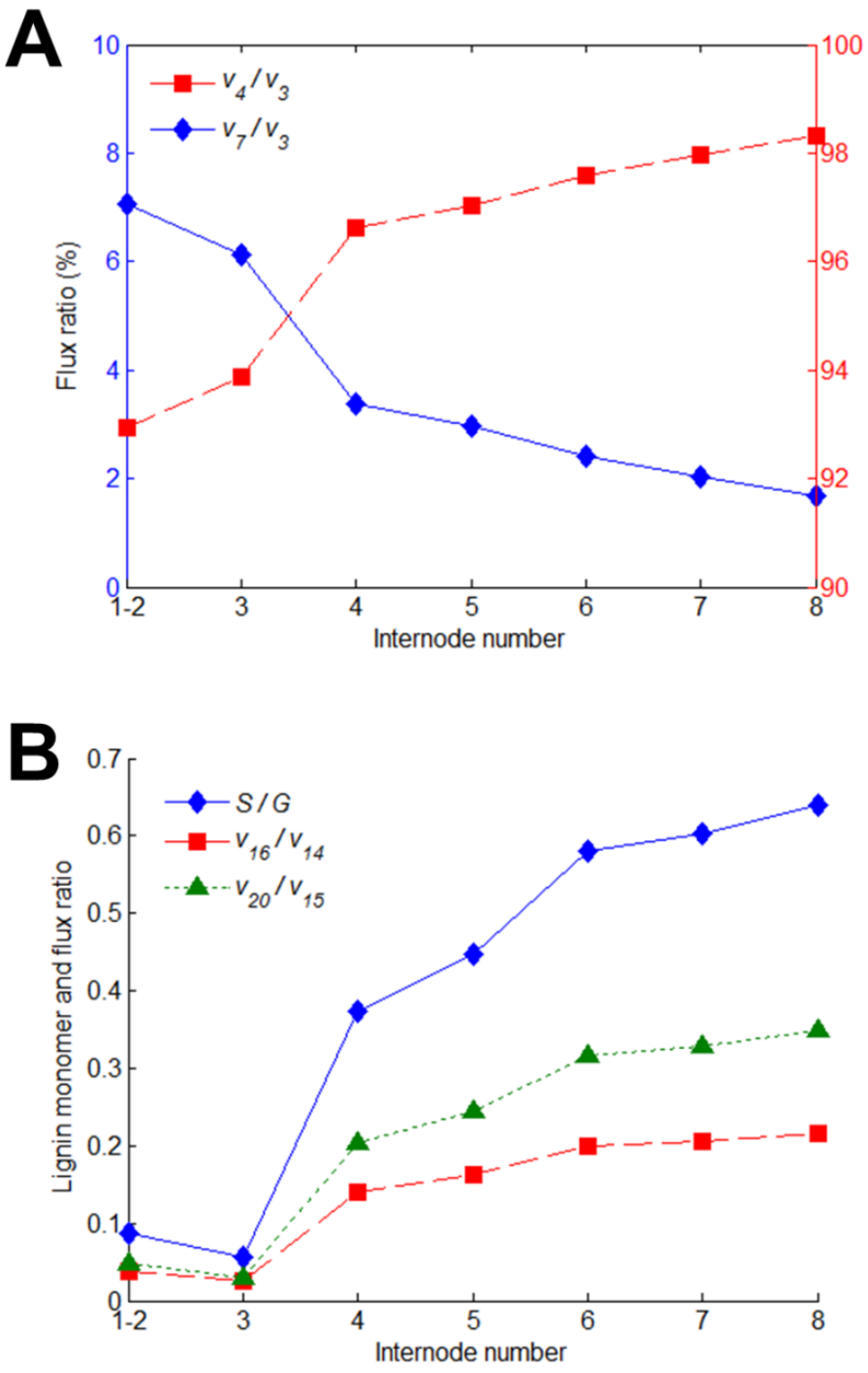


**Figure 3.2: Flux partitioning at principal branch points in different internodes.**
**(A)** Developmental patterns of flux partitioning at *p*-coumaroyl CoA branch point in wild-type plants, given as percentage of $v_3$. **(B)** Comparison of flux partitioning at coniferyl aldehyde ($v_{16}/v_{14}$) and coniferyl alcohol ($v_{20}/v_{15}$) branch points with the ratio of S to G lignin (S/G) in individual internodes of wild-type plants.

### 3.2.2 Minor Extension of the Pathway Structure

For a systemic analysis of the pathway we used the results of a gene modification study in alfalfa where genes encoding for PAL, C4H, HCT, C3H, CCoAOMT, F5H, and COMT were independently down-regulated. With the exception of F5H-modified lines, which did not permit measurements of the targeted enzyme activity, we applied MOMA to each strain and each internode and predicted the new steady-state flux distribution (see Section 3.4).

A very interesting result is the fact that no feasible solution exists for four of the six transgenic plants, if the revised metabolic map is correct (Figure 3.1; black and red colored arrows). For example, if C4H activity is down-regulated to 45% of its wild-type level, it is analytically impossible to derive a set of fluxes that satisfies the mass balance at cinnamic acid as well as the observed lignin composition, if the supply of phenylalanine is constant. To remedy this situation, it seems to be necessary to add to the pathway structure three "overflow" fluxes counteracting the potential accumulation of the intermediate metabolites cinnamic acid, p-coumaryl aldehyde, and 5-hydroxyconiferyl alcohol (blue arrows $v_{22}$, $v_{23}$, $v_{24}$ in Figure 3.1). This proposed amendment is at least partially supported by observations. First, salicylic acid (SA), an essential signaling molecule for systemic acquired resistance against pathogen attack, can be formed from cinnamic acid [98,99,100], although it may also originate from the shikimate pathway via isochorismate [101]. Second, the biosynthesis of all flavonoids begins with the condensation of p-coumaroyl CoA and three molecules of malonyl CoA by the enzyme chalcone synthase [102]. And third, incorporation of 5-hydroxyconiferyl alcohol into lignin polymer is found in COMT-deficient alfalfa [103]. Thus, we included these additional effluxes, and the expanded system (Figure 3.1; $v_1$ to $v_{24}$) permitted feasible solutions in all cases tested.

In wild-type plants, the FBA-derived steady-state values of the three added fluxes are minimized to prevent lignin precursors from being channeled into peripheral pathways producing SA or flavonoids. In the transgenic plants, these auxiliary fluxes are no longer restricted to small values and thus can be raised to substantial levels to facilitate the re-distribution of fluxes. However, the assumption that the peripheral fluxes are minimized in wild-type plants must be handled with caution: although the phenylpropanoid pathway in cells undergoing secondary wall thickening may evolve towards maximizing the synthesis of lignin precursors, this is apparently not the case when biosynthesis of flavonoid-derived products, which may function as floral pigments or as anti-microbial agents, becomes the plant's top priority.

### 3.2.3 Trends in Flux Patterns

The MOMA analysis revealed flux distributions for all transgenic lines and their individual internodes. Figure 3.3 shows the developmental evolution of flux patterns in CCoAOMT-deficient plants. Of note is that all computed fluxes exhibit strong and essentially monotonic trends: for each transgenic line, the flux partitioning at important branch points follows clear trends throughout the internodes rather than jumping in value from one internode to the next. This result is surprising and encouraging, because MOMA simply assumes that the fluxes undergo a minimal re-distribution when the pathway system is perturbed. Because these perturbations occur independently for each internode, there is no mathematical guarantee that individual fluxes would follow any smooth trend from internode to internode. In other words, the collective results, while fitting into the context of a gradual change in lignification pattern during stem development, are by no means "automatic," because no external constraints or conditions were imposed or enforced on the transition from one internode to the next. The computed trends are summarized in Table 3.1.

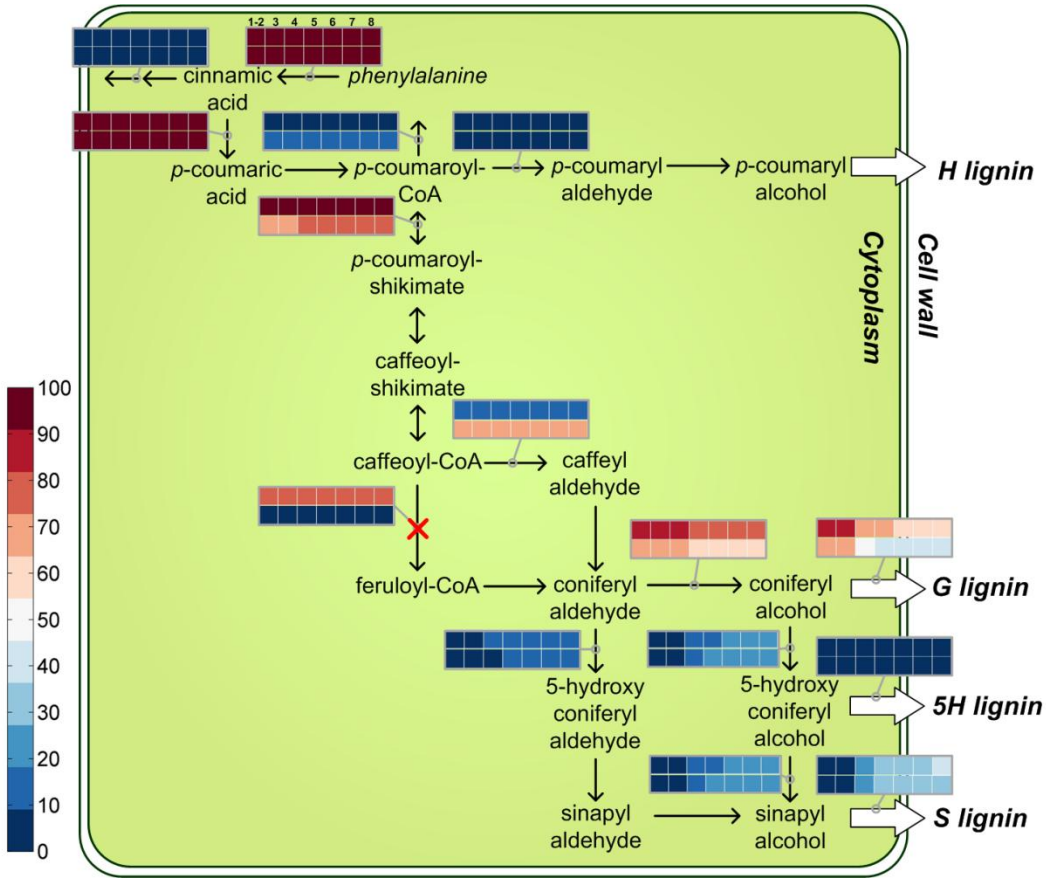


**Figure 3.3: Developmental evolution of the steady-state flux distribution in CCoAOMT-deficient plants versus the wild-type plants.**
The reaction crossed out in red is dysfunctional in this particular transgenic strain. Two rows of colored boxes are placed either above horizontally plotted fluxes, or to the left of vertically plotted fluxes. The first row represents wild-type plants, whereas the second row refers to transgenic line (here a CCoAOMT-deficient plant). Each row contains seven colored boxes, which represent the seven stem internodes (with internodes 1 and 2 merged). In FBA and MOMA, all fluxes are normalized to the initial step in the pathway, namely the conversion of phenylalanine to cinnamic acid. Therefore, the color of each box shows the normalized steady-state value of the corresponding flux in one specific internode: low values are dark blue, intermediate values are white, and high values are dark red. Because all the reactions along a linear pathway have the same flux values at steady state, only the first one is shown.

**Table 3.1: Developmental trends in flux partitioning between successive internodes.** The developmental evolution of fluxes diverging at the intermediate metabolite listed in the first column, when normalized by the total flux entering the branch point, can be described as monotonically increasing (↑↑), increasing with minor variations (↑), essentially unchanged (─), decreasing with minor variations (↓), or monotonically decreasing (↓↓).

| Branch Point | Flux | Transgenic Strain | | | | | |
|---|---|---|---|---|---|---|---|
| | | **PAL↓** | **C4H↓** | **HCT↓** | **C3H↓** | **CCoAOMT↓** | **COMT↓** |
| Cinnamic acid | $v_2$ | ─ | ─ | ↑ | ↑↑ | ↓↓ | ↑↑ |
| | $v_{22}$ | ─ | ─ | ↓ | ↓↓ | ↑↑ | ↓↓ |
| *p*-coumaroyl CoA | $v_4$ | ↓↓ | ↓↓ | ↑ | ↑↑ | ↓↓ | ─ |
| | $v_7$ | ↑↑ | ↑↑ | ↑↑ | ↑↑ | ↑↑ | ↑↑ |
| | $v_{23}$ | ─ | ─ | ↓ | ↓↓ | ↑↑ | ↓↓ |
| Caffeoyl CoA | $v_{10}$ | ─ | ─ | ─ | ─ | ─ | ─ |
| | $v_{11}$ | ─ | ─ | ─ | ─ | ─ | ─ |
| Coniferyl aldehyde | $v_{14}$ | ↓↓ | ↓↓ | ─ | ─ | ↓↓ | ─ |
| | $v_{16}$ | ↑↑ | ↑↑ | ─ | ─ | ↑↑ | ─ |
| Coniferyl alcohol | $v_{15}$ | ↓↓ | ↓ | ↓↓ | ↓↓ | ↓↓ | ↑ |
| | $v_{20}$ | ↑↑ | ↑ | ↑↑ | ↑↑ | ↑↑ | ↓ |
| 5-hydroxy-coniferyl alcohol | $v_{21}$ | ─ | ─ | ↓↓ | ─ | ─ | ─ |
| | $v_{24}$ | ─ | ─ | ↑↑ | ─ | ─ | ─ |

The following paragraphs are structured as follows. First, we re-evaluate the gene knock-down data in a systematic way across different stages of growth and formulate four postulates that actually do not require a full model analysis, but emerge from the "logic" of the pathway. Second, we discuss two postulates regarding novel mechanisms of metabolic regulation that result from our comprehensive model analysis. Third, we present new experimental results that directly support one of the model-based postulates.

### 3.2.4 Availability of Phenylalanine Drives Lignin Production

The total lignin production is driven by the availability of phenylalanine rather than by enzymatic limitations. This conclusion results from the observation that the down-regulation of PAL has much less effect on total lignin content and/or lignin composition in young internodes with small amounts of lignin than in mature internodes with high lignin production (Table B.3; [25]). Expressed differently, PAL is not acting at capacity when the demand for lignin is relatively low, as is the case in young internodes. This conclusion is also supported by the observation that lignin production is not enhanced proportionately when PAL enzyme is over-expressed in transgenic plants [104].

### 3.2.5 HCT Is Reversible

In transgenic plants where C3H is down-regulated, the proportion of H lignin among total monomer yields is significantly increased over control plants, especially in mature internodes (Figure 3.4A). This finding is at first puzzling, because it is unlikely that the cell can detect changes in C3H activity and adapt accordingly by exerting appropriate flux control at an earlier branch point (*i.e.*, *p*-coumaroyl CoA) within the network. Arguably the simplest explanation is that HCT (possibly along with other plant acyltransferases) is reversible [105]. If so, the following scenario is possible: as *p*-coumaroyl shikimate accumulates due to a reduced C3H activity, HCT converts it back to *p*-coumaroyl CoA in the presence of free CoA, thereby allowing the cell to escalate the production of H lignin beyond the wild-type level. The catalytic efficiency of HCT acting on *p*-coumaroyl shikimate as substrate remains to be experimentally determined, along with the possible competition for CoA between two shikimate esters (*i.e.*, *p*-coumaroyl shikimate and caffeoyl shikimate).

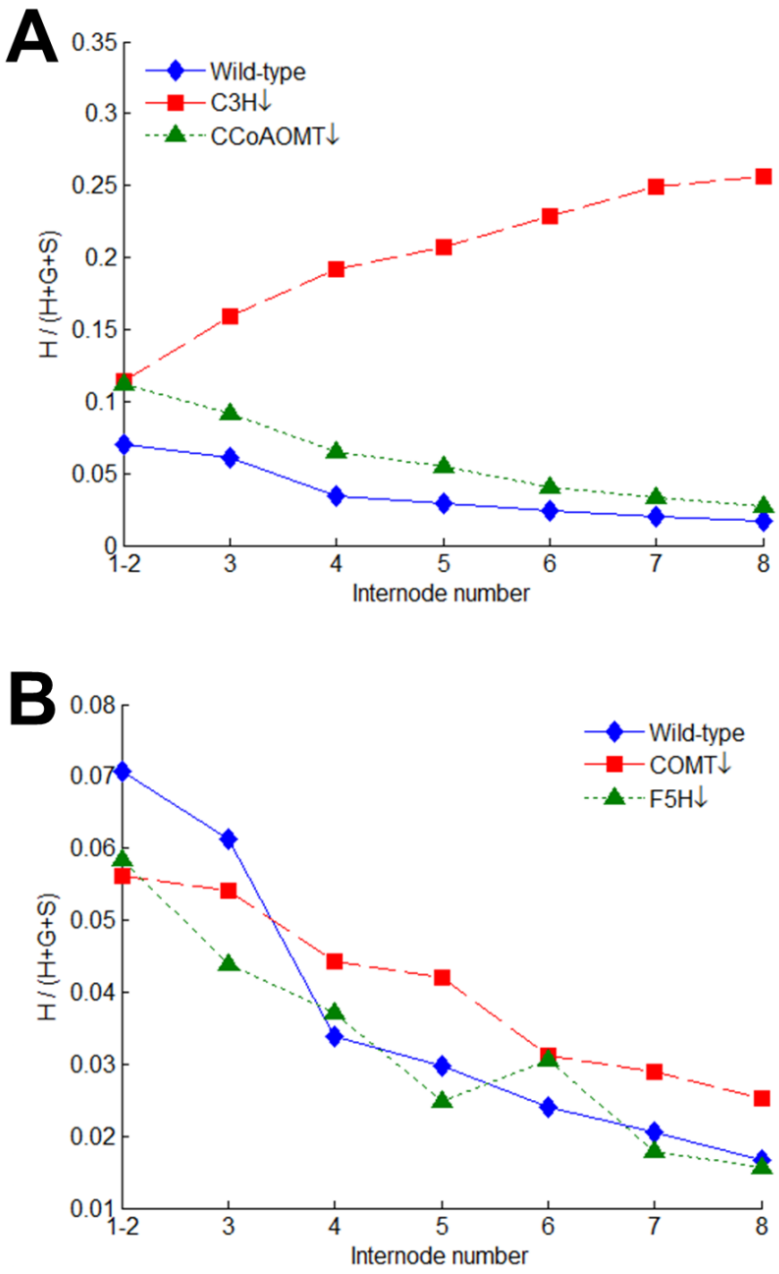


**Figure 3.4: Developmental patterns of the proportion of H lignin in control and transgenic alfalfa plants.**
**(A)** The proportion of H lignin in total monomer yields (H+G+S) is substantially or slightly increased in transgenic plants with reduced activities of C3H or CCoAOMT, respectively. **(B)** Down-regulation of COMT or F5H has essentially no effect on the proportion of H lignin in total monomer yields: the amounts of H lignin are very small and the trends do not differ significantly from wild type.

### 3.2.6 Is C3H Mildly Reversible?

The hypothesis of HCT being reversible prompts us to investigate whether C3H, which controls the material flow between two HCT-catalyzed steps, also permits catalysis in both directions. A slightly increased proportion of H lignin in CCoAOMT-deficient plants (Figure 3.4A) seems to suggest that C3H is mildly reversible and that part of the accumulated caffeoyl CoA is therefore converted back to *p*-coumaroyl CoA and subsequently channeled towards H lignin, a scenario which seems unlikely based on the known catalysis by cytochrome P450 enzymes. However, the amounts of H lignin determined by thioacidolysis appear to be unaffected by the low CCoAOMT activity despite a noticeable decrease in total lignin content (Table B.3; [25]). One plausible explanation is that thioacidolysis yields are highly correlated with the *in vivo* abundance of S lignin [14], which might suggest that plants may in effect produce more H lignin than was measured against the down-regulation of CCoAOMT.

### 3.2.7 Two CCR-Catalyzed Reactions Are Essentially Irreversible

If both HCT and C3H are reversible, the two CCR-catalyzed reactions—$v_{10}$ and $v_{13}$—can be regarded as the "committed" steps (*i.e.*, they are essentially irreversible), because manipulation of any downstream enzyme, such as COMT and F5H, has no substantial effect on H lignin (Figure 3.4B). Interestingly, the postulate seems to echo the conclusion from a previous enzyme assay [106]: CCR purified from poplar stems was able to catalyze the conversion of coniferyl aldehyde into feruloyl CoA in the presence of other co-factors but preferentially reduced CoA-esters, as judged by the calculated equilibrium constants.

### 3.2.8 The Pathway Contains Crossing Channels towards G and S Lignin

In addition to a modest increase in H lignin, down-regulation of CCoAOMT leads to a noticeable increase in the S/G ratio of all internodes except for internodes 1 and 2 (Figure 3.5A). This finding is puzzling because coniferyl aldehyde is a common precursor to both S and G lignin and one would therefore expect a similar effect on both. The analogous situation arises in COMT-deficient plants, where the S/G ratio is reduced (Figure 3.5A). This case, however, is not quite as clear-cut because COMT also shows activities towards downstream intermediates like 5-hydroxyconiferyl aldehyde and 5-hydroxyconiferyl alcohol. Thus, in this case of COMT deficiency, the S/G ratio might not be a good indicator of the flux partitioning at coniferyl aldehyde towards G and S lignin.

As an explanation for the altered S/G ratios in cases of CCoAOMT or COMT down-regulation, we postulate that the enzymes controlling $v_{12}$ and $v_{16}$ (and maybe even $v_{10}$ and $v_{17}$) are organized into a functional complex (each) through which the intermediates are channeled without much leakage. Similarly, we postulate that $v_{13}$ and $v_{14}$ form a corresponding complex without much leakage. This dual postulate for crossing channels is supported indirectly by literature information and by findings from our flux analysis, as outlined below.

First, an analysis of mature stems (internodes 6-9) collected from CCoAOMT down-regulated transgenic lines indicated that the levels of G lignin were greatly reduced, whereas those of S lignin were nearly unaffected (cf. CCOMT antisense line ACC305 in Table 1 of [107]). Similarly, down-regulation of CCR1, which actively catalyzes the subsequent reduction of feruloyl-CoA to coniferyl aldehyde, also resulted in an increased S/G ratio in mature internodes of alfalfa stems [108], again with G lignin being more strongly reduced than S lignin. Although the existence of the CCR2-COMT pathway helps sustain the lignin content in either CCoAOMT or CCR1 down-regulated lines, the findings do not explain why S lignin is synthesized at the expense of G lignin

upon genetic modifications of the CCoAOMT-CCR1 pathway. Nevertheless, the findings are entirely consistent with the postulate of crossing channels.

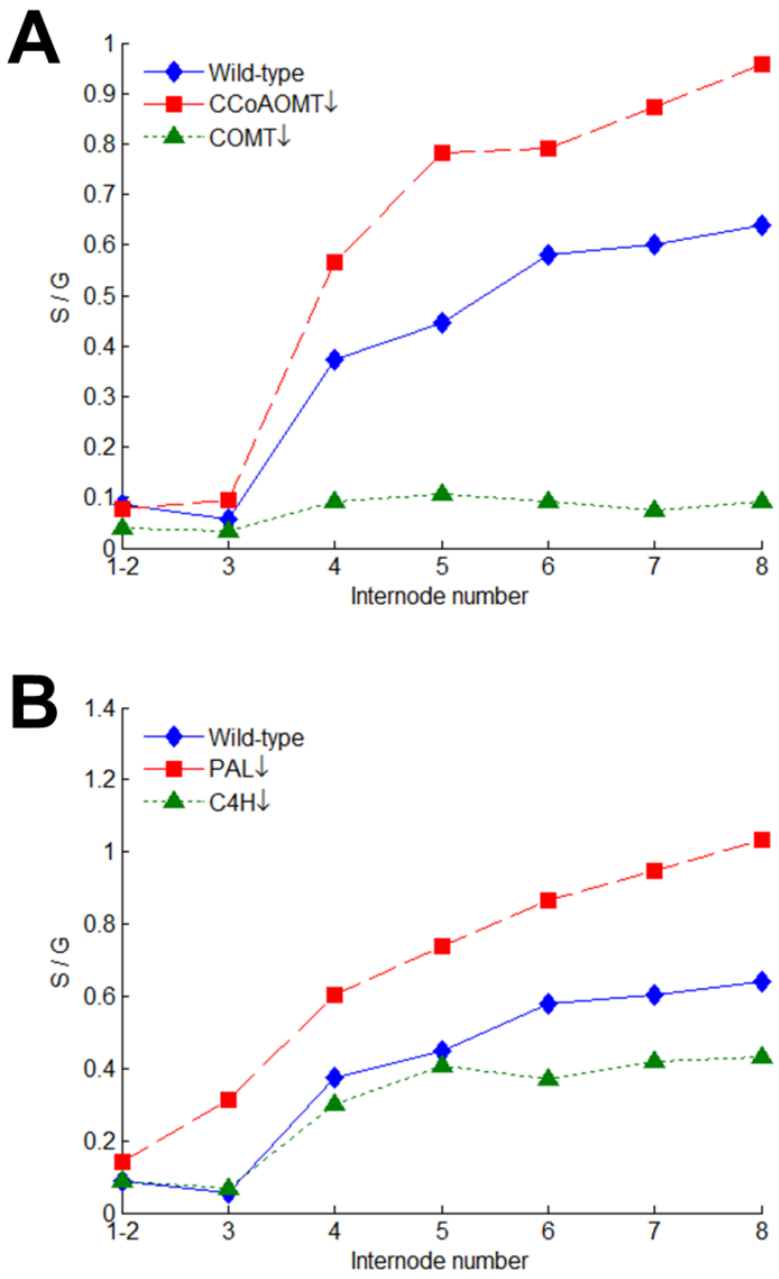


**Figure 3.5: Developmental patterns of the S/G ratio in control and transgenic plants.** **(A)** In comparison with control plants, the S/G ratio is increased in CCoAOMT-deficient plants but drastically decreased in COMT-deficient plants. **(B)** Similarly intriguing, the S/G ratio is increased in PAL-deficient plants but decreased in C4H-deficient plants.

Second, one of the constituent enzymes, F5H, is localized to the external surface of the endoplasmic reticulum [109], so that the proposed channel may exist in the form of an enzyme complex anchored in the endomembrane. Indeed, a labeling experiment in microsomes extracted from lignifying alfalfa stems suggested such a co-localization of COMT and F5H [110]. It showed that caffeyl aldehyde, when incubated with [methyl-$^{14}$C]-labeled S-adenosyl L-methionine (a co-substrate necessary for COMT-mediated *O*-methylation) and NADPH (the reducing agent for F5H), is converted to coniferyl aldehyde, 5-hydroxyconiferyl aldehyde, and a small amount of sinapyl aldehyde.

Finally, our flux distribution analysis reveals a strong correlation between the computed flux values of $v_{13}$ and $v_{14}$ for all but the CCoAOMT-deficient plants (Pearson correlation coefficient $\rho = 0.9952$; $p$-value $< 0.001$) (Figure 3.6). This correlation suggests that there is normally almost no exchange of products between $v_{12}$ and $v_{13}$, and that most of the coniferyl aldehydes produced through the CCR2-COMT shunt are directly utilized by F5H without having the opportunity of diversion into G lignin biosynthesis. A notable exception seems to be the situation where CCoAOMT is significantly down-regulated. In this case, caffeoyl CoA tends to accumulate at least in the short term, thus providing the CCR2-COMT pathway and the associated metabolic channel with an abundance of substrate. The predicted flux distribution (Figure 3.3) and the observed lignin composition (Table B.3) indicate that CCoAOMT-deficient plants produce a considerable amount of G lignin, although the levels of S lignin are comparable to those in the controls, which implies that only some of the extra caffeoyl CoA can be converted efficiently into S lignin through the proposed channel. Overall, the proposed functional channels seem to be consistent with results of the flux analysis as well as with earlier discussions in the literature [9,60]. The correlation between $v_{12}$ and $v_{16}$ is less pronounced, which is presumably due to the fact that F5H and COMT catalyze parallel pathways, with the latter ($v_{20}$ and $v_{21}$) buffering changes in earlier precursors.

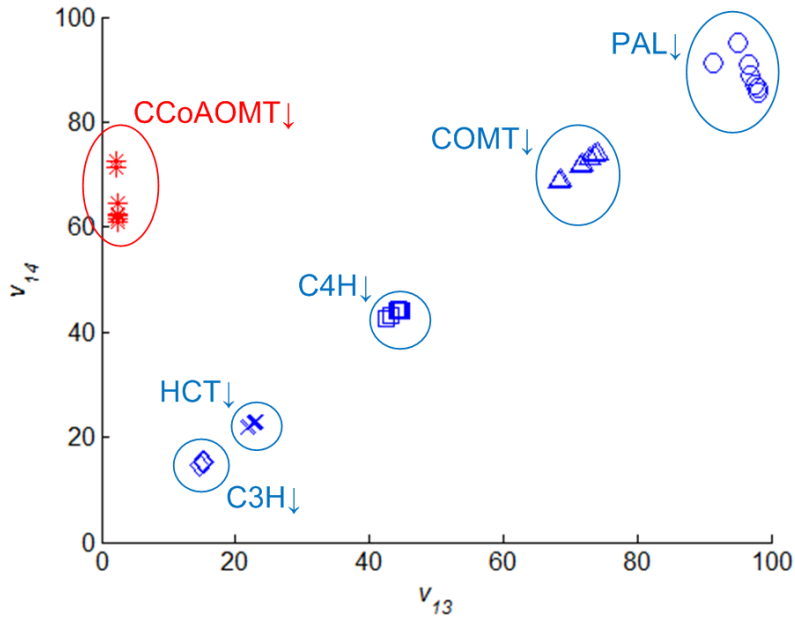


**Figure 3.6: Plot of $v_{14}$ versus $v_{13}$ in transgenic alfalfa plants.**
Expression of PAL, C4H, HCT, C3H, CCoAOMT, or COMT was independently down-regulated. Symbols within each ellipse represent different internodes. With the exception of CCoAOMT, the two fluxes are very strongly and linearly correlated.

An alternative explanation for an increased S/G ratio upon modifications of the CCoAOMT-CCR1 pathway could be that the kinetic features of the enzymes that catalyze coniferyl aldehyde and coniferyl alcohol are fine-tuned such that they could permit the adjustment of fluxes leading to G and S lignin and thus change the S/G ratio. For instance, given that down-regulation of CCoAOMT or CCR1 may alter the intracellular level of coniferyl aldehyde, the relative values of $v_{14}$ and $v_{16}$ at steady state could depend on whether the respective enzyme works within the linear or saturation region of its kinetic profile.

To investigate this alternate hypothesis, we designed and analyzed a kinetic Michaelis-Menten model that contains the two alternative pathways from caffeoyl CoA to coniferyl aldehyde as well as the two principal branch points where the fluxes leading to G and S lignin diverge (see Section B.3). The model was simulated 10,000 times with randomly sampled kinetic parameter values, as described in Sections 3.4.2 and B.3, and

we recorded the percentage of admissible parameter sets that yielded a significantly increased S/G ratio in response to a 80% reduced CCoAOMT or CCR1 activity.

We first examined the case where CCoAOMT is down-regulated. Only ~5% of all admissible systems (see Section B.3 for definition) yielded a significantly increased S/G ratio, whereas nearly half of all systems resulted in an S/G ratio that differed by less than 5%. The few cases of significant increases in the S/G ratio did not reveal particular patterns, which may not be too surprising because the system involves 16 kinetic parameters that affect each other in a nonlinear fashion. Intriguingly, for the scenario of CCR1 down-regulation, none of the admissible systems showed a significant increase in S/G ratio; in fact, all changes in S/G ratios were less than 0.5%. Replacing the Michaelis-Menten kinetics with cooperative Hill kinetics allowed more flexibility. Still, only ~3% of all admissible systems exhibited an increase in S/G ratio upon CCR1 down-regulation. Taken together, it seems that, theoretically, some precisely tuned sets of kinetic parameters could lead to the observed effects on the S/G ratio. However, these sets are extremely rare and do not seem to be robust enough to render the kinetics-based hypothesis viable.

### 3.2.9 Feedforward Regulation by a Compound Derived from Cinnamic Acid

One of the most paradoxical findings among the collective results from the transgenic plants is the opposite effect on lignin composition (and specifically the S/G ratio) when either PAL or C4H is down-regulated. It seems that these alterations should not differentially affect monolignol biosynthesis, because both occur before the first branch point, but they do. Closer inspection of the data from different internodes reveals that the S/G ratio is consistently increased in PAL-deficient plants but decreased in C4H-deficient plants (Figure 3.5B). While experiments with tobacco have suggested that the differential co-localization of PAL isoforms and C4H might be the underlying cause of

such observations [11], there is as yet no direct evidence for this intracellular association in alfalfa or other related legume species.

In accordance with the proposition of separate metabolic channels for G and S lignin, we postulate that the different effects of PAL or C4H down-regulation on the S/G ratio are due to feedforward regulation. Specifically, we suggest that this regulation is mediated by a downstream product of the cinnamic acid degradation pathway, which is represented collectively as $v_{22}$ in Figure 3.1. Notice that this feedforward regulation had not been recognized by the scientific community and was postulated by the model analysis purely with computational means.

Consistent with the observation of all transgenic experiments, an appropriate control strategy by this unknown compound "X" is summarized in Figure 3.7 and discussed below. In the case of PAL-deficiency, where the biosynthesis of cinnamic acid from phenylalanine declines, a diminished pool of X could directly or indirectly reduce the expression of CCoAOMT/CCR1/CAD and/or activate the expression of CCR2/COMT/F5H, thereby altering the channeling towards G and S lignin and increasing the S/G ratio. Intriguingly, this proposed inhibition of CCoAOMT expression following PAL down-regulation is supported by a strong correlation of the proportion of G and S lignin in total monomer yields in internodes 4-8 of the PAL- and CCoAOMT-deficient plants (Figure 3.8).

In the case of C4H deficiency, however, the production of X through $v_{22}$ is likely to increase because the consumption of cinnamic acid through a competing branch $v_2$ is not as effective as in wild-type plants. Thus, an accumulation of X could in turn activate the expression of CCoAOMT/CCR1/CAD and/or reduce the expression of CCR2/COMT/F5H, leading to a smaller S/G ratio.

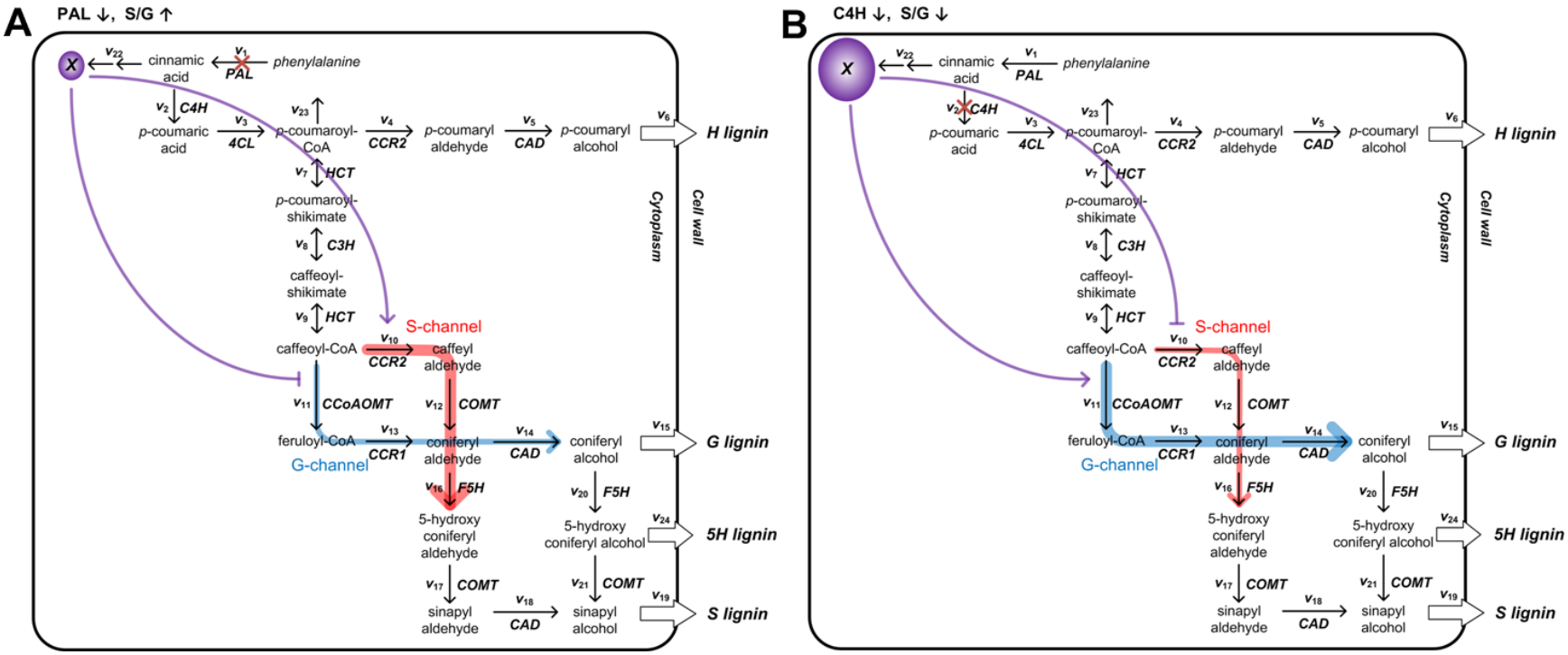


**Figure 3.7: Conjectured effects of PAL (A) or C4H (B) down-regulation on the postulated channels.**
The postulated G lignin- and S lignin-specific channels are colored in blue and red, respectively, with their widths representing the relative capacity in the designated transgenic plants. The size of the circle with the unknown compound *X* correlates symbolically with its intracellular pool size. The blocked purple line indicates repression, whereas the purple arrow indicates activation.

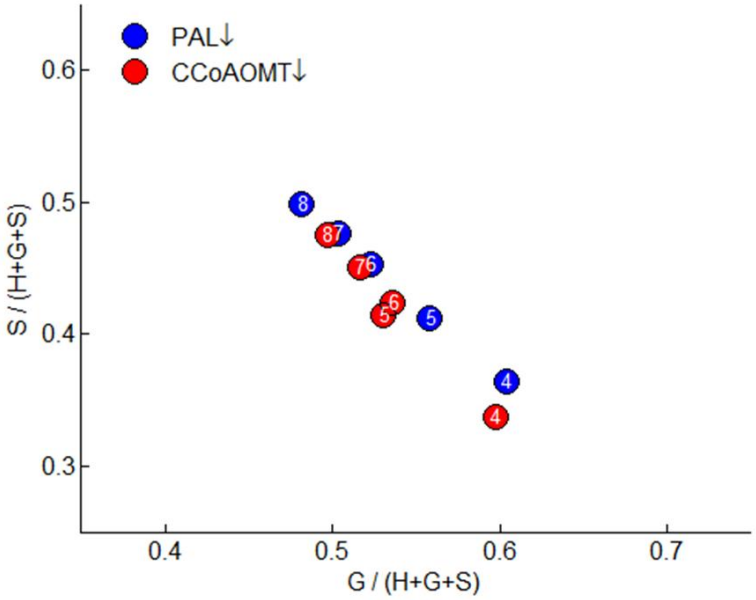


**Figure 3.8: Plot of the proportion of S versus the proportion of G in total monomer yields.**
The colored circles represent internodes 4-8 of stems collected from transgenic plants where PAL or CCoAOMT is down-regulated. Numbers in the symbols refer to the specific internodes.

### 3.2.10 Salicylic Acid Is a Signaling Molecule for Monolignol Biosynthesis

Salicylic acid (SA) is a notable endogenous signaling molecule that is known to be derived from cinnamic acid [111]. Down-regulation of one pathway enzyme other than C4H (*e.g.* HCT [112]) had recently been shown to lead to elevated levels of SA. To investigate whether SA is the postulated signaling compound X, we measured its intracellular levels in many independent transgenic alfalfa lines in which different monolignol biosynthesis genes had been down-regulated. Indeed, the results show that the intracellular levels of SA are highly proportional to the extent of lignin reduction (Figure 3.9). Based on our postulated feedforward regulation, this effect can be explained through the participation of SA in the inhibition of the metabolic channel committed to S lignin biosynthesis, thus reducing the total lignin content.

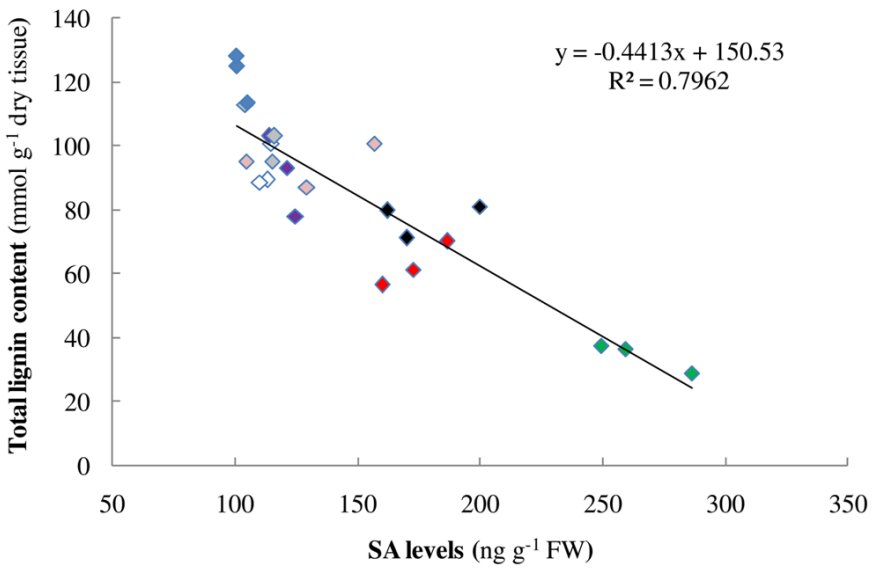


**Figure 3.9: Relationship between lignin content and salicylic acid accumulation in different alfalfa antisense lignin down-regulated lines.**
Red: 4CL, Black: C3H, Green: HCT, White: CCoAOMT, Violet: CCR, Pink: CAD, Gray: COMT and Blue: Wild-type.

## 3.3 Discussion

Functional genomics is a premier tool for identifying metabolic pathways in sequenced model species and for pinpointing genes involved in them [113]. However, it is known that many enzymes coexist in multiple isoforms with unique expression patterns and substrate specificities. A pertinent example seems to be the recent discovery of two CCR isoforms with distinct catalytic properties towards major CoA-esters in *Medicago* [59]. Steady-state flux analysis of an extended pathway system that accounts for the isoforms reveals that the alternative path is dispensable in wild-type plants, but that it may rise to significant levels in specific transgenic lines. Indeed, CCoAOMT-deficient plants support a much higher lignin production than lines where HCT or C3H is down-regulated (Table B.3; [25]). The intricate differences in pathway operation among otherwise very similar transgenic lines point to the need of investigating flux patterns not only in different plants, but also in different strains, lines and even different internodes and tissues. The results shown here furthermore demonstrate that subtle variances among

tissues and lines are difficult to discern with intuition alone, but that computational analyses can serve as objective and rigorous tools for explaining such differences.

Specifically, the new integrative modeling approach proposed here combines static flux-based models and a Monte Carlo simulation of randomly parameterized kinetic models. This approach has the advantage that it allows the collective analysis of many experimental results and sheds light on pathway features that are particularly important for functionality under normal and altered conditions. The analysis here revealed a quantitative trend of flux patterns during development, which in turn allowed the identification of principal branch-point metabolites at which internode-specific flux partitioning patterns control the observed mode of lignification. While it is relatively easy to single out principal metabolites in linear or slightly branched pathways, the system studied here is confounded by the plant's employment of the same enzymes, such as CCR and CAD, in different key positions. Due to this multiple use, manipulating the flux partitioning pattern towards a desired mode of lignification may incur undesired "side effects."

The computational analysis indicates that a single flux analysis just for wild-type plants is insufficient for understanding pathway functionality because even a seemingly simple pathway like monolignol biosynthesis requires relatively minor, yet important, extensions to account for the overflow of some intermediate metabolites that only occurs in transgenic plants. At the same time, the analysis also demonstrates that the simultaneous analysis of several independent datasets, in this case transgenic lines and sequential internodes, can lead to insights that otherwise would have been difficult to obtain. Here, it led to several postulates that are specific enough for experimental validation or refutation.

Some model-free postulates refer to the need for reversibility or committedness of key reactions, which might not be too surprising. Two further postulates are more intriguing. They refer to the functional channeling within the pathway and its mechanistic

control. Based on the observation of an increased S/G ratio in CCoAOMT or CCR1 down-regulated lines, the computational results suggest an S lignin-specific channel capable of converting caffeyl aldehyde directly into 5-hydroxyconiferyl aldehyde or sinapyl aldehyde. Different experiments in the literature suggested the co-localization of COMT and F5H in lignifying alfalfa stems [110] and the localization of F5H to the external surface of the endoplasmic reticulum [109]. These and our findings would imply the likely location for a functional S-channel complex to be associated with the endomembrane.

While the proposed membrane-bound channel for synthesizing S lignin could constitute an important control mechanism, it may only have comparatively limited capacity because even in CCoAOMT down-regulated lines G lignin is generated in a higher proportion of total monomer yields than S lignin (Table B.3; [25]). One likely cause is that different *O*-methyltransferases (OMTs) are involved in converting caffeyl aldehyde to coniferyl aldehyde. These OMTs may have distinct sub-cellular localization (to cytoplasm or endomembrane) and therefore a different affinity to F5H. Thus, it could be that the cytosolic OMT in the transgenic lines with reduced CCoAOMT expression is up-regulated and helps consume extra caffeyl aldehyde outside the proposed channel. A corresponding labeling experiment in alfalfa [110] confirmed that only a small proportion of total cellular COMT activity against caffeyl aldehyde is associated with the microsomal membrane, and that adding excess recombinant COMT has little effect on the metabolism of caffeyl aldehyde by microsomes.

To examine whether the observed increase in the S/G ratio upon modifications of the CCoAOMT-CCR1 pathway could be explained alternatively by a kinetically-controlled mechanism, we generated 10,000 ODE model instantiations for a reduced pathway system (Section B.3) and simulated both down-regulation schemes. Among all sampled parameter sets, only a minute percentage of systems had the ability to increase their S/G ratio significantly in either case. Although the results neither reject the

possibility of a kinetically-controlled S/G ratio nor directly corroborate our channeling postulate, they do suggest that purely kinetic control might be unlikely, because it would require rather precise implementations of specific parameters in different tissues, which seems to compromise the robustness of the system. As shown in a structural study of alfalfa COMT [114], mutations of some key residues lining the active site result in significantly different substrate binding and/or turnover rate. Moreover, it is likely that the kinetic properties of other enzymes may also exhibit a similar, if not more severe, susceptibility to genetic perturbations (*e.g.*, [115,116]). Since the variation in the S/G ratio is typically small (s.d.≈0.03 in two control lines; [25]), the proposed functional channeling mechanisms seem to offer a more robust option to help maintain a physiologically proper S/G ratio.

The observed decrease in the S/G ratio of COMT down-regulated lines alone is not sufficient to prove the existence of a G lignin-specific channel, because a reduced COMT activity affects all fluxes that are specific for the synthesis of S lignin, thus leading to a smaller S/G ratio. Nevertheless, the strong correlation between $v_{13}$ and $v_{14}$ that emerged from our computations for most transgenic experiments lends further credence to such an inference. This correlation not only supports the operation of a G lignin-specific channel, but also hints at the possibility of CCR1 and CAD (and maybe CCoAOMT) being complexed or co-localized on internal membranes.

One option for testing this postulate would be to down-regulate CCR2 and record if the strain exhibits a greater decrease in S lignin than in G lignin, giving a smaller S/G ratio. Surprisingly, knocking out CCR2 in *M. truncatula*, a species closely related to alfalfa, leads to an increased S/G ratio, whereas *M. truncatula* CCR1 knock-out mutants show a reduction in the S/G ratio [59]. However, in spite of their close taxonomic relatedness, the operation and control of monolignol biosynthesis might be quite different in tetraploid alfalfa (*M. sativa* L.) and diploid *M. truncatula*. For instance, the S/G ratio in wild-type alfalfa stems (0.62; internodes 1-8) is approximately twice as large as that in

wild-type *M. truncatula* stems (0.29; internodes 1-7). Consequently, further experimental work is required to validate or reject the postulate that a G lignin-specific channel is operational in alfalfa.

If the postulates of specific channels towards the synthesis of G and S lignin are valid, one may further surmise that the opposite effects of PAL or C4H down-regulation on lignin composition are the results of differential gene or enzyme expression, which could be mediated by a cinnamic acid derivative. However, the model could not identify this molecule, leading us to call it *Compound X*. Supporting this hypothesis, the transgenic experiments used here have shown that down-regulation of CCoAOMT, which we postulate to be involved in the G lignin-specific channel, yields similar proportions of G and S lignin among total monomers as does the down-regulation of PAL, which is postulated to inhibit and/or activate the functioning of the G lignin- and S lignin-specific channels, respectively (Figure 3.8).

Salicylic acid (SA), a phenolic phytohormone derived from phenylalanine, was proposed as a potential candidate for this unknown Compound X. Intriguingly, post-hoc experiments showed that the intracellular levels of SA are indeed highly proportional to the extent of lignin reduction in transgenics where different pathway genes are down-regulated (Figure 3.9). This result fits directly into the context of our feedforward control postulate. At the same time, it makes us wonder why putting a block on monolignol biosynthesis could affect the homeostasis of SA, especially if the blockage is located away from the pathway entrance. Based on previous findings that SA can be derived both from cinnamic acid and from isochorismate via the shikimate pathway [111], and that HCT uses shikimate as a preferred cofactor (Figure 3.10), we propose the following scenario: when the flux going through the pathway is decreased due to some genetic manipulation, fewer shikimate molecules will be trapped in shikimate esters (*p*-coumaroyl shikimate and caffeoyl shikimate) and thus become available to make SA. In other words, the shikimate recycling facilitated by HCT enables the shikimate pool to

work as a sensor of the flux into lignin. Future in-depth studies, whether they are experimental or computational, are required to justify this hypothesis. It is noteworthy, however, that the reason why plants shuttle monolignol pathway intermediates between Coenzyme A and shikimate esters has yet to be explained.

In conclusion, our analysis shows that a combined modeling effort can uniquely and effectively complement experimental studies of the type used here. In contrast to analyzing one dataset at a time, it allowed us to integrate all results from a comprehensive experimental investigation of various transgenic lines and internodes. This integration, in turn, revealed dynamic, developmental patterns and their dependence on key enzymes. Together, the analyses uncovered elusive control mechanisms of monolignol biosynthesis and led to testable hypotheses regarding various pathway aspects that should be clarified before one attempts to generate and optimize viable, productive "designer" crops with minimal recalcitrance.

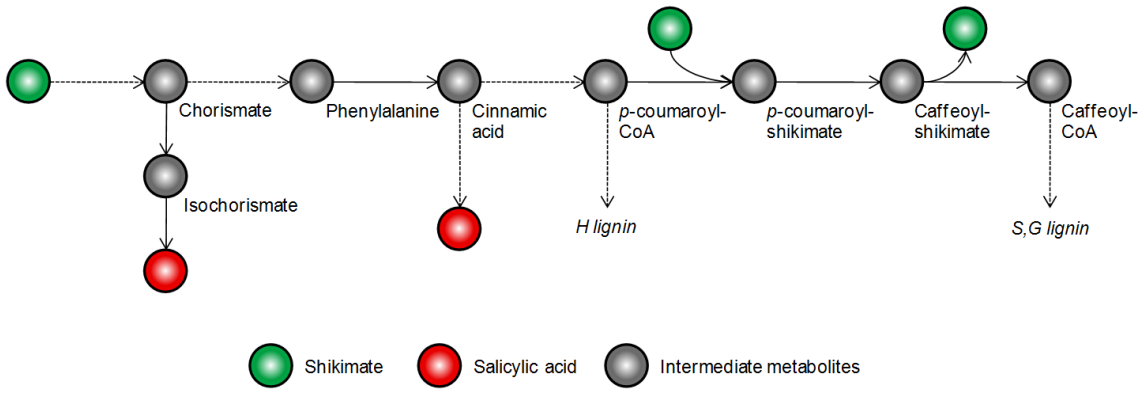


**Figure 3.10: Alternative routes of salicylic acid biosynthesis and shikimate recycling.** Steps likely to involve more than one enzyme and intermediate are shown with dashed arrows. In addition, pathways such as the tyrosine and flavonoid branches are not shown for the sake of clarity.

## 3.4 Materials and Methods

### 3.4.1 Experimental Data[4]

In a previous study [25], lignin content and composition were analyzed in transgenic alfalfa plants in which seven enzymes were independently down-regulated (*cf.* Figure 3.1). These enzymes were: L-phenylalanine ammonia-lyase (PAL), cinnamate 4-hydroxylase (C4H), hydroxycinnamoyl CoA:quinate/shikimate hydroxycinnamoyl transferase (HCT), coumarate 3-hydroxylase (C3H), caffeoyl coenzyme A 3-*O*-methyltransferase (CCoAOMT), ferulate 5-hydroxylase (F5H), and caffeic acid 3-*O*-methyltransferase (COMT). Each transgenic plant was cultivated to early flowering stage, and the mature stem consisting of eight internodes was harvested and divided into individual segments; all internodes were numbered according to their maturity, with internodes 1-2 representing the pooling of the two uppermost stem segments. The lignin content and monomer composition for each internode were determined for each transgenic line via established protocols [25]; the results are summarized in Table B.3. The activities of all targeted enzymes were also measured and summarized elsewhere, with the exception of F5H, which showed no activity towards any documented substrates when assayed in crude alfalfa extracts *in vitro* (Table 2c of [25]). Thus, the F5H-deficient line is excluded from the following analysis.

Salicylic Acid Determination

Salicylic acid levels in stems from the same plant lines (excepting PAL down-regulated plants) as well as from CAD down-regulated lines [108] were determined using the biosensor *Acinetobacter sp.* ADPWH_lux as described previously [117,118]. Samples

---

[4]The experiments reported in this Chapter were conducted by our collaborators at the Noble Foundation.

consisted of detached stems consisting of six internodes. SA was extracted by grinding stems (100 mg fresh weight) in fresh LB liquid medium (2.5 ml LB per 1 g of stem) by vortexing for 30 sec and sonicating for 5 min on ice, after which the homogenates were centrifuged at 12,000 g for 15 min. The supernatants were used for SA measurement and an equivalent volume of LB medium was used to make a SA standard curve (SA final concentrations of 0, 0.05, 0.25, 0.5, 1.6, 8.3, 20, 40, 83, 166 and 200 µM). An overnight culture of *Acinetobacter sp.* ADPWH_lux was diluted in LB (1:20) and grown at 37°C for ~2 hrs to an $OD_{600}$ of 0.4. Sixty µl of LB medium, 50 µl of salicylate biosensor culture and 20 µl of each crude extract were mixed in a 96-well cell culture plate. The plate was incubated at 37°C for 1 h without shaking and bioluminescence and $OD_{600}$ of negative controls (LB alone or water) were read using a Glomax Multi detection system (Promega Corporation, Sunnyvale, CA). SA standard and negative controls were read in parallel with the experimental samples and every sample was replicated five times. Relative bioluminescence was obtained by subtracting bioluminescence $OD_{600}$ of negative controls, and SA concentration was estimate according to the SA standard curve.

Lignin Content

Lignin content was determined by the thioacidolysis method as described previously [112].

### 3.4.2 Modeling Approach

FBA and MOMA

As described in Chapter 1, static flux balance models build on the assumption that a metabolic pathway system is in a quasi-steady state where, for any metabolite pool, fluxes governing its synthesis and degradation are equal. Mathematically, such a mass balance constraint can be represented as

$$\mathbf{Nv} = \mathbf{0}, \tag{3.1}$$

where **N** is the stoichiometric matrix of the pathway system and **v** is a vector of fluxes. Other commonly used constraints are upper and lower bounds on individual fluxes,

$$l_i \le v_i \le u_i, \tag{3.2}$$

that define the possibility of reversibility and the maximal reaction rate, respectively. Here, we assume that all the metabolic reactions and transport processes are irreversible and therefore set $l_i = 0$ for all *i*. The only exceptions are the three overflow fluxes ($v_{22-24}$; defined in Section 3.2.2), for which we arbitrarily choose 0.01 as the lower bound to prevent their values from becoming too small in the subsequent optimization step. The maximal reaction rates used for defining the upper bounds are currently unavailable because most enzymes within the pathway have not been characterized in *Medicago*, according to the enzyme databases like BRENDA [119]. Therefore, we normalized all fluxes to the value of $v_1$ as a means of standardization. This normalization, which is achieved by introducing an extra constraint $v_1 = 1$, works to ensure that all fluxes are less than or equal to one.

Specific to this work, the measured, relative amounts of lignin monomers can be reformulated as "proportionality constraints" on the three fluxes $v_6$, $v_{15}$ and $v_{19}$ that represent the transport of monolignols into the cell wall (Figure 3.1). As an illustration, if the first two stem internodes consist of 7.1% H lignin, 85.5% G lignin, and 7.4% S lignin, we can define the following equality constraints:

$$\begin{aligned} 7.1v_{15} - 85.5v_6 &= 0 \\ 7.1v_{19} - 7.4v_6 &= 0 \end{aligned}. \tag{3.3}$$

For a specific internode of wild-type alfalfa plants, all constraints taken together define the feasible space of all permissible flux distribution, which is denoted by $U^{wt}$. In FBA, the optimal solution is identified within $U^{wt}$ by solving a linear programming problem where an appropriate objective function *f* is maximized or minimized. In the case of

lignin biosynthesis, we assume that the lignified stem tissues of *wild-type* alfalfa plants have evolved to maximize the production of lignin monomers, which translates into the following objective function

$$f(\mathbf{v}) = v_6 + v_{15} + v_{19}. \tag{3.4}$$

The optimal flux distribution in wild-type alfalfa plants, resulting from this maximization, is denoted as $\mathbf{v}^{wt}$.

The issue of multiple solutions giving the same optimal value of the objective function has been widely discussed [30,120,121]. In contrast to genome-scale models, we have the opportunity here to enumerate all equivalent flux distributions for a moderately-sized metabolic pathway system like monolignol biosynthesis, for instance, using Gauss-Jordan elimination. This advantage in turn enables us to identify a unique, physiologically relevant flux distribution for wild-type plants (see Section B.2 for further details).

For lignin-modified lines, where a particular enzyme is genetically down-regulated, we use the *method of minimization of metabolic adjustment* (MOMA) [95] to predict their altered flux distributions. In its original application to gene knockout studies in bacteria, MOMA posited that a mutant strain tries to function as similarly to the wild type as possible within the limitations imposed by the mutation. In mathematical terms, the effect of a gene knockdown on the metabolic pathway system is mimicked by imposing an extra inequality constraint $v_j \leq \delta_j \cdot v_j^{wt}$ on reaction *j*: If $v_j^{wt}$ is the wild-type flux, then the activity of the mutated enzyme catalyzing this reaction is down-regulated to at most $\left(100 \cdot \delta_j\right)\%$ of the wild-type activity. The feasible space consisting of all flux distributions in mutants is thus defined by these inequality constraints along with all balance constraints and upper and lower bounds for the same wild-type internode, as discussed above. The notable difference for the mutant is that the specific values of the lignin monomer composition can now be significantly different (*cf.* Table B.3). Within

this reduced feasible space, which is denoted by $U^j$, the MOMA solution $\mathbf{v}^j$ is the point that is closest to the reference point $\mathbf{v}^{wt}$ in terms of the Euclidean distance

$$\mathbf{v}^j = \arg\min_{\mathbf{v}\in U^j} \left\| \mathbf{v} - \mathbf{v}^{wt} \right\|. \tag{3.5}$$

Monte Carlo Analysis of Kinetic Parameters

A major surprise emerging from the experiments with lignin-modified alfalfa lines was the differential effect of CCoAOMT down-regulation on G and S lignin production. It is conceivable that these results could be due to the kinetic features of the participating enzymes. To analyze this possibility, we designed a kinetic model of the involved reactions and tested this model with a large-scale simulation. The details of this model can be found in Section B.3 and the Monte Carlo techniques *per se* are straightforward. Importantly, this Monte Carlo-type simulation allowed us to examine thousands of combinations of kinetic parameters without limiting ourselves to a few particular cases of manually tuned, ill-characterized parameter values.

# CHAPTER 4

# ANALYSIS OF OPERATING PRINCIPLES WITH S-SYSTEM MODELS[5]

The results presented in Chapter 3 have greatly improved our knowledge of the regulation of monolignol biosynthesis. Nonetheless, they also provoke new questions regarding the organization of the pathway and its regulation that have not been answered; neither specifically, nor in generality. For example, why do we observe a specific developmental pattern of fluxes but not other alternative patterns that could seem equally valid? And by which criteria, if any, are the observed patterns chosen? To establish the theoretical foundation on which these types of *operating principles* can be deciphered, two distinct, yet complementary methods will be developed in this Chapter. While we are specifically interested in metabolically channeled system as we encountered them in monolignol biosynthesis, and which will be revisited in Chapter 5, the theoretical and computational characterization of alternative strategies within a space of admissible solutions, which is developed here, is much more general and, indeed, applicable to a wide range of biological systems. In particular, we will address the very common scenario where a biological system must shift its operation from a normal steady state to a new target steady state, in response to physiological or environmental demands. This situation is quite common and will be investigated with artificial pathways and illustrated with an analysis of the heat stress response in yeast, which is rather well characterized

---

[5] Adapted from: Lee, Y.*, Chen, P.-W.* and Voit, E.O. (2011) Analysis of Operating Principles with S-system Models. *Math. Biosci.* 231: 49-60. [*Equal contribution]

experimentally. Chapter 5 will use similar techniques to assess the reprogramming of lignin biosynthesis during plant development.

## 4.1 Introduction

Biological design principles refer to structural or regulatory features of biological systems that are observed more often than expected. They are thought to have survived evolution, thereby making them apparently superior to hypothesized alternative structures that *a priori* might seem equally reasonable and valid [43,122]. The typical question in the investigation of design principles is: What is the advantage of a particular structural or regulatory feature over an otherwise equivalent design that lacks this feature?

Design principles are identified and investigated through comparisons with reference cases. In static network analysis, a candidate structure is declared a *motif* [87,123,124,125] if it is found significantly more often than in random graphs, as they were originally proposed over fifty years ago by Erdös and Rényi [126]. Within Biochemical Systems Theory (BST; [43,44,127,128]), which was discussed in Chapter 1 of this dissertation, the role of a design feature is analyzed by comparing two systems that have exactly the same structure except for the feature or interest. The approach of choice for such an analysis has been the *Method of Controlled Mathematical Comparisons* (MCMC) [122,129]. A key component of this method is the establishment of objective criteria of *functional effectiveness* [129,130]. These criteria, which are formulated before the comparison of two system structures is performed and interpreted, serve as a metric according to which either the system of interest or some alternative is deemed superior. Typical criteria are stability, robustness, a short response time to stimuli, adequate responsiveness to external demands, and maybe a transient response profile that does not deviate too far from the nominal profile.

MCMC originally focused on algebraic analyses, but was subsequently augmented with computational and statistical methods [1,43,130,131,132]. Dynamic biological systems that were successfully analyzed with respect to design principles include pathway topologies [43,130,131,133], immune cascades [130], gene regulatory circuits [134,135,136], signaling systems [137], and riboswitches [138].

While design principles have become a fashionable topic of investigation in recent years, their dynamic counterparts, *operating principles*, have received only a small fraction of the attention. Operating principles address questions regarding the dynamics of a response as we observe or hypothesize it, in comparison to *a priori* equally valid alternatives [139,140,141]. Like in the case of design principles, operating principles may be investigated in natural systems, where the goal is to discover an objective explanation for the suitability or optimality of an observed set of procedures, or in synthetic, engineered systems, where the goal is the optimization of a procedure with respect to some target objective.

An example for an investigation of natural operating principles is the following question: If a system is forced by the environment to move to a new steady state, and if this state may be achieved either by drastically changing a few control variables or by slightly changing many control variables, which strategy is preferable? Alvarez and colleagues [142] analyzed this question heuristically for changes in yeast metabolism during the diauxic shift and determined that many genes in the living yeast cell were changed by a modest degree. A different aspect of natural operating principles was investigated in the response of yeast cells exposed to heat stress [1,143,144,145,146]. In this case, the lead questions were: Which genes are actually up-regulated in expression and by how much? What are the metabolic consequences of this up-regulation? Could there be alternative up-regulation scenarios that might perform better? Can we find objective criteria explaining the emergence or natural selection of the strategy that is actually observed in yeast? Yet another example concerned the question of how bacteria

using a PTS system for energy production can restart glycolysis after starvation, when one would expect the initial phosphate donor, phosphoenolpyruvate, to be depleted [147,148].

Questions regarding the operation of synthetic systems are of the following types. Which sets of process manipulations or alterations will cause the system to reach a target objective? What is the advantage of utilizing or altering a particular sequence of processes instead of an alternative sequence? Is one set "better" than another? Is one of them optimal with respect to objective criteria? As a specific example of this situation, the task was posed to optimize the product yield of a feedback-regulated pathway with two successive branches by selecting and altering a small, fixed number of genes or enzymes. The results, which were not easy to predict without a quantitative analysis, demonstrated that the locations and magnitudes of optimal manipulations depended not so much on the topological structure of the pathway as on the locations of its regulatory signals [139].

One might ask whether operating principles are truly different from design principles, because the possible space of dynamic responses is clearly constrained, if not determined, by the physical and regulatory structure of a system. While design and operation are coupled to some degree, their distinction is both reasonable and necessary, because a cell or organism could theoretically respond to the same demand in different ways, even within exactly the same structural confines, as the diauxic shift study [142] demonstrates. Furthermore, cells can be exposed to drastically different demands, which require appropriate responses within exactly the same structural design. A good example is the blue-green alga *Synechocystis,* which generates energy either autotrophically per photosynthesis, heterotrophically per consumption of carbohydrates, or through a mixture of the two. It has been shown that the distribution of flux rates within its metabolic pathway system, and thus the operation of the system, shifts dramatically between these three modes [149]. In a different example, it was shown that plant cells use the same

metabolic pathway system, but with distinctly different, dynamically changing flux distributions, to produce woody materials during their development or in different transgenic strains [150,151].

As in the case of design principles, it is impossible to study operating principles in exhaustive generality. The analysis described here therefore focuses exclusively on one pertinent special case, namely, where a biological system must shift from its normal steady state to a new steady state, a response that is typical in the face of persistent changes in a cell's environment. While the two steady states will be at the center of the present analysis, features of transients will also be discussed. In first approximation it may even be possible to consider slow-changing, longer-term trends as a series of different "almost-steady-states" [152].

Most analyses of design principles in the past had the benefit of clear reference systems that were topologically very similar to the system of interest. For instance, a system with feedback was compared to a system without this particular feedback signal. In the case of operating principles, it is not always *a priori* clear what the alternatives are. For instance, we cannot simply compare up-regulation of one process against unaltered operation, because the two would lead to different transients and presumably to different steady states. Instead, the approach toward a new steady state will almost always require alterations in larger sets of independent variables. Thus, the first important step in the analysis of operating principles is an exhaustive exploration of the *admissible set of operating strategies*. Once this set is characterized, the true discovery of *operating principles* consists of the selection of the one strategy that is superior to all others under the chosen criteria of functional effectiveness and optimality.

## 4.2 Methods and Theoretical Results

Canonical models, and in particular S-systems within BST [43,127], are especially well suited for analyzing operating principles. As in the case of design principles, the primary reasons are twofold. First, these systems have a fixed structure, where each component has a well-defined meaning and where system features are mapped onto parameters in a one-to-one fashion [43,44]. Secondly, S-systems permit a linear representation of their steady states within the language of linear algebra, upon a logarithmic transformation of all variables [153].

The generic situation to be addressed here concerns a biological system, represented by S-system equations (Eq. (1.2)), that needs to respond to a changed environmental demand by assuming a new steady state. It is not difficult to imagine that this task usually has many solutions and that distinctly different settings of independent variables may lead to the same steady state with respect to the dependent variables. This multiplicity of possibilities is due to the fact that most systems contain many more processes than variables. Because these processes are usually under the control of independent variables, different choices of independent variables correspond to distinct solution strategies.

The non-trivial steady state of an S-system model can be formulated in matrix notation as [154]

$$\mathbf{A}_D \cdot \mathbf{y}_D + \mathbf{A}_I \cdot \mathbf{y}_I = \mathbf{b}, \tag{4.1}$$

where $\mathbf{y}_D$ denotes the vector of the logarithms of the dependent variables at steady state, $\mathbf{y}_I$ is the corresponding vector of independent variables, the elements of the matrices $\mathbf{A}_D$ and $\mathbf{A}_I$ are $a_{ij} = g_{ij} - h_{ij}$ for all $i$ and $j$, separated into dependent and independent variables, and $b_i = \log(\beta_i/\alpha_i)$ for $i = 1, \ldots, n$.

In a typical analysis, all parameter values are known and one computes the non-trivial steady state, which may then be used for other diagnostics like stability, sensitivity, and gain analysis [43,44,153]. This steady state can be expressed explicitly as

$$\mathbf{y}_D = \mathbf{S} \cdot \mathbf{b} + \mathbf{L} \cdot \mathbf{y}_I, \tag{4.2}$$

where $\mathbf{S} = \mathbf{A}_D^{-1}$ and $\mathbf{L} = -\mathbf{A}_D^{-1}\mathbf{A}_I$ are the so-called *sensitivity* and *logarithmic gain matrices*, respectively [154].

For our purposes here, we must turn the task around. We assume that the system has to switch from some initial steady state to a target steady state $\mathbf{y}_D$ that is mandated by new environmental demands. We furthermore suppose that we know the numerical values of the dependent variables at this target steady state. The question thus becomes how the independent variables should be changed to achieve this state (*cf.* [152,155]). Again using stress as an example, we might observe an altered metabolic steady state and ask which enzymes would have to be altered in activity to reach the stress state.

For ease of representation, we rewrite Eq. (4.2) as

$$-\mathbf{A}_D^{-1}\mathbf{A}_I\mathbf{y}_I = \mathbf{y}_D - \mathbf{A}_D^{-1}\mathbf{b}\,. \tag{4.3}$$

Since $\mathbf{A}_D^{-1}\mathbf{b}$ is constant and $\mathbf{A}_D$ and $\mathbf{b}$ are known, we define

$$\mathbf{y}'_D = \mathbf{y}_D - \mathbf{A}_D^{-1}\mathbf{b}\,, \tag{4.4}$$

which yields the simplified representation

$$\mathbf{L}\mathbf{y}_I = \mathbf{y}'_D\,. \tag{4.5}$$

For the special case where $m = n$ and $\mathbf{L}$ has full rank, we can invert the system of equations and express each independent variable as a unique linear function of the new variables that constitute $\mathbf{y}'_D$; namely we obtain

$$\mathbf{y}_I = \mathbf{L}^{-1}\mathbf{y}'_D\,. \tag{4.6}$$

Expressed in words, we can demand numerical values for the dependent variables of a particular target steady state, and Eq. (4.6) determines how the independent variables

have to be set for the system to reach this state. If the new state is stable, and if the system starts within its basin of attraction, one may actually reach this state by starting the system at the original steady state and resetting the independent variables according to Eq. (4.6). Of course, we do not know how much time the dynamic system will require to come sufficiently close to the target.

For cases where $m < n$, the matrix **L** is "tall," which reflects an over-determined system that generally permits no solution. Nevertheless, for practical purposes we can compute a least-squares solution, which minimizes the deviation from the target state and is given as the regression equation

$$\mathbf{y}_{I\_LS} = \mathbf{L}^{+}\mathbf{y}'_{D}, \tag{4.7}$$

where $\mathbf{L}^{+}$ is the pseudo-inverse of **L** [156].

In the most pertinent case, the number of independent variables is larger than the number of dependent variables ($m > n$). This relationship is not always true in actual systems, but it usually holds, because most systems contain more processes than pools and each process normally involves at least one independent variable. The matrix equation (4.5) now can no longer be inverted directly, and if the rank of **L** is $r$, the solution consists of an $m - r$ dimensional space. Even though an inversion is not directly possible, the solution space may be characterized with methods of linear algebra, where the starting point is the pseudo-inverse. Specifically, the solution space, which consists of every admissible $\mathbf{y}_I$, can be spanned through the following steps. First, find a particular solution $\mathbf{y}_{I\_PS}$. Then use $\mathbf{y}_{I\_PS}$ and the span of the null space of **L** to describe the entire solution space as

$$\mathbf{y}_{I\_PS} = \mathbf{L}^{+}\mathbf{y}'_{D}, \tag{4.8}$$

$$\mathbf{y}_{I} = \mathbf{y}_{I\_PS} + \mathbf{B}\cdot\boldsymbol{\lambda}. \tag{4.9}$$

Here, $\boldsymbol{\lambda}$ is any given real-valued ($m$–$n$)-dimensional vector, rank($\mathbf{L}$) = $n$, and $\mathbf{B}$ is a matrix in which each column is a basis vector. Together, these column vectors constitute a basis of the null space of $\mathbf{L}$.

### 4.3 Illustration Examples

It is useful to demonstrate the theoretical results with simple didactic examples. The first representative case is a cascaded system (Figure 4.1), where the numbers of precursors and state variables are the same ($n$ = $m$ = 4) and the system has a unique solution. The cascade could describe the expression of a formerly inactive gene $X_5$, which becomes activated ($X_1$) and is subsequently transcribed; $X_6$ could model nucleotides that are assembled into mRNA ($X_2$); $X_7$ could represent amino acids, which are assembled into an enzyme ($X_3$), which subsequently catalyzes the conversion of a metabolic substrate $X_8$ into a product $X_4$. The final product could directly or indirectly repress the expression of the gene. The generic S-system representation of the model is

$$\begin{aligned}
\dot{X}_1 &= \alpha_1 X_4^{g_{14}} X_5^{g_{15}} - \beta_1 X_1^{h_{11}} \\
\dot{X}_2 &= \alpha_2 X_1^{g_{21}} X_6^{g_{26}} - \beta_2 X_2^{h_{22}} \\
\dot{X}_3 &= \alpha_3 X_2^{g_{32}} X_7^{g_{37}} - \beta_3 X_3^{h_{33}} \\
\dot{X}_4 &= \alpha_4 X_3^{g_{43}} X_8^{g_{48}} - \beta_4 X_4^{h_{44}}
\end{aligned} \quad . \tag{4.10}$$

Without loss of generality in this and the later illustration examples, all rate constants $\alpha_i$ and $\beta_i$ are arbitrarily set to 1 and the independent variables are initially defined as 1.2. By this definition we know that $\mathbf{y}'_D = \mathbf{y}_D$ because $\mathbf{b}$ = $\mathbf{0}$. The values of the kinetic order parameters in this and other systems are given in Table 4.1.

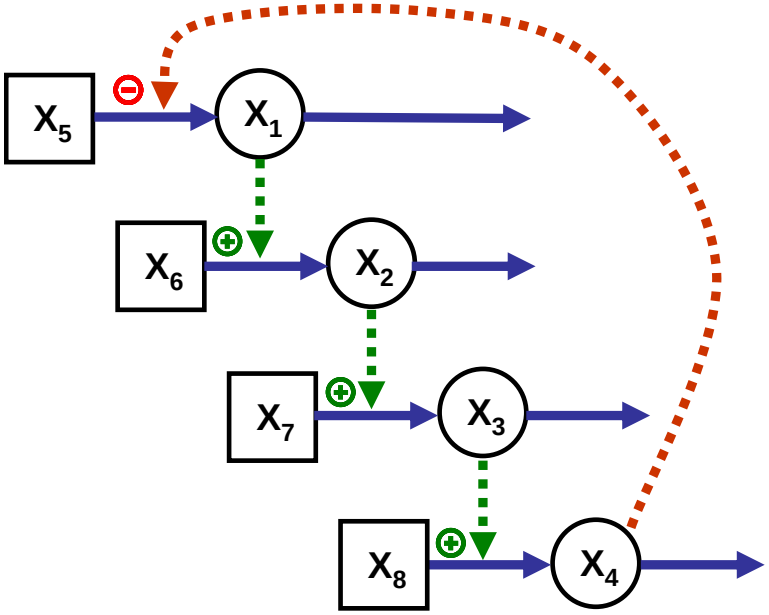


**Figure 4.1: A cascaded system with as many dependent (circles) as independent (squares) variables.**
The cascade could represent, from top to bottom, gene expression, transcription into mRNA, translation into protein, and a metabolic process catalyzed by enzyme $X_3$.

**Table 4.1: Numerical values of kinetic parameters for all illustration examples***

| Cascade 1 Figure 4.1 | | Linear Pathway Figure 4.2 | | Cascade 2 Figure 4.4 | | Branched Pathway Figure 4.7 | | | |
|---|---|---|---|---|---|---|---|---|---|
| $g_{14}$ | -0.8 | $g_{15}$ | 0.5 | $g_{14}$ | -0.8 | $\alpha_1$ | 1.755 | $\beta_1$ | 1 |
| $g_{15}$ | 0.25 | $g_{21}$ | 0.2 | $g_{15}$ | 0.24 | $\alpha_2$ | 1 | $\beta_2$ | 2 |
| $g_{21}$ | 0.4 | $g_{24}$ | -0.25 | $g_{21}$ | 0.4 | $\alpha_3$ | 1 | $\beta_3$ | 2 |
| $g_{26}$ | 0.3 | $g_{32}$ | 0.8 | $g_{26}$ | 0.3 | $\alpha_4$ | 1 | $\beta_4$ | 1 |
| $g_{32}$ | 0.5 | $g_{36}$ | 0.35 | $g_{32}$ | 0.5 | $g_{13}$ | 0.05 | $h_{11}$ | 1 |
| $g_{37}$ | 0.3 | $g_{43}$ | 0.4 | $g_{37}$ | 0.3 | $g_{15}$ | 0.75 | $h_{16}$ | 1 |
| $g_{43}$ | 0.1 | $g_{47}$ | 0.1 | $g_{43}$ | 0.4 | $g_{1,11}$ | 0.125 | $h_{22}$ | 0.5 |
| $g_{48}$ | 0.2 | $h_{11}$ | 0.2 | $h_{11}$ | 0.2 | $g_{21}$ | 1 | $h_{27}$ | 0.5 |
| $h_{11}$ | 0.2 | $h_{14}$ | -0.25 | $h_{22}$ | 1 | $g_{26}$ | 1 | $h_{29}$ | 0.5 |
| $h_{22}$ | 1 | $h_{22}$ | 0.8 | $h_{33}$ | 0.8 | $g_{32}$ | 0.5 | $h_{33}$ | 0.2 |
| $h_{33}$ | 0.4 | $h_{26}$ | 0.35 | $h_{44}$ | 0.9 | $g_{39}$ | 1 | $h_{3,10}$ | 0.25 |
| $h_{44}$ | 0.2 | $h_{33}$ | 0.4 | | | $g_{42}$ | 0.5 | $h_{3,11}$ | 0.25 |
| | | $h_{37}$ | 0.1 | | | $g_{47}$ | 1 | $h_{44}$ | 0.5 |
| | | $h_{44}$ | 0.2 | | | | | $h_{48}$ | 1 |
| | | $h_{48}$ | 0.25 | | | | | | |

* The rate constants for the linear and the two cascaded pathways were set equal to 1.

The second example is a simple linear pathway with feedback and an exogenous demand for product (Figure 4.2). This example was chosen in contrast to the cascaded system, because it involves several precursor-product relationships, which constrain the parameters of the corresponding effluxes and influxes. While one may initially wonder what the effects of these constraints may be, we will see that these constraints have no real bearing on the characterization of a set of independent variables that moves the system to the target steady state. The enzymes for the conversions of $X_2$ into $X_3$ and $X_4$ are explicitly modeled, as are the input to the pathway and the demand for $X_4$, such that $n = m = 4$. The generic S-system model is

$$\begin{aligned}
\dot{X}_1 &= \alpha_1 X_5^{g_{15}} - \beta_1 X_1^{h_{11}} X_4^{h_{14}} \\
\dot{X}_2 &= \beta_1 X_1^{h_{11}} X_4^{h_{14}} - \beta_2 X_2^{h_{22}} X_6^{h_{26}} \\
\dot{X}_3 &= \beta_2 X_2^{h_{22}} X_6^{h_{26}} - \beta_3 X_3^{h_{33}} X_7^{h_{37}} \\
\dot{X}_4 &= \beta_3 X_3^{h_{33}} X_7^{h_{37}} - \beta_4 X_4^{h_{44}} X_8^{h_{48}}
\end{aligned} \qquad (4.11)$$

Again, all rate constants $\alpha_i$ and $\beta_i$ are arbitrarily set to 1 and the independent variables to 1.2. The values of the kinetic orders are given in Table 4.1.

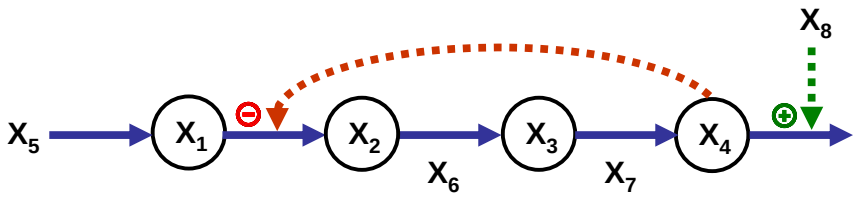


**Figure 4.2: Linear pathway with feedback and an exogenous demand for product.** The task of moving the system to a new steady state has a unique solution.

For our illustration, we start both systems arbitrarily at (1, 1, 1, 1) and let them reach their nominal steady states. While at the steady state, the environmental demand changes at time $t = 60$ or $t = 150$, respectively, and we assume that all variables in the cascade and the linear pathway must move to a new target value of 2. Because $n = m = 4$,

the solutions are in both cases unique. They are given as $\mathbf{X}_I = [16.0 \;\; 4.0 \;\; 0.7937 \;\; 1.4142]^T$ and $\mathbf{X}_I = [0.933 \;\; 0.1857 \;\; 0.0442 \;\; 0.5]^T$, respectively. Numerical simulation demonstrates that the systems indeed respond by moving to the desired target states (Figure 4.3). The vectors $\mathbf{X}_I$ in the inverse solutions do not convey anything about the transients.

The third and fourth introductory examples are cascaded and linear pathways with fewer independent than dependent variables (Figure 4.4). S-systems models were constructed according to well-documented guidelines, and the values of the kinetic orders for the cascaded system were defined as presented in Table 4.1. The target values were defined as 3. It could seem that these scenarios are rather unrealistic, but they do occur in cases like the ones shown here as well as in cases of strongly connected pathways where not all genes or enzymes are accessible to manipulations. If it is infeasible or impossible to alter some of the independent variables, *m* is in effect decreased and may become lower than *n*.

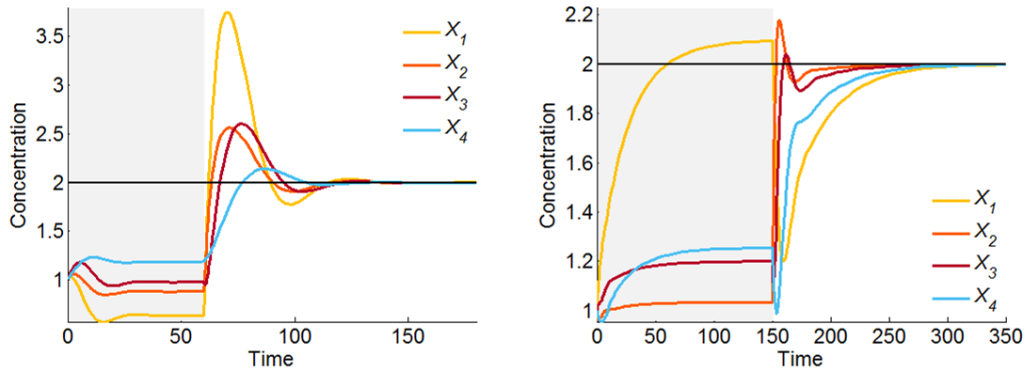


**Figure 4.3: Resetting the independent variables according to the computed unique solutions moves the cascaded (left) and linear (right) pathway systems to the desired target (2, 2, 2, 2).**
During the initial phase (shaded light grey), the systems move from their arbitrary initial values (1, 1, 1, 1) to their nominal steady states. At time $t$ = 60 or $t$ = 150, respectively, the environment changes, requiring all variables to reach the target value 2.

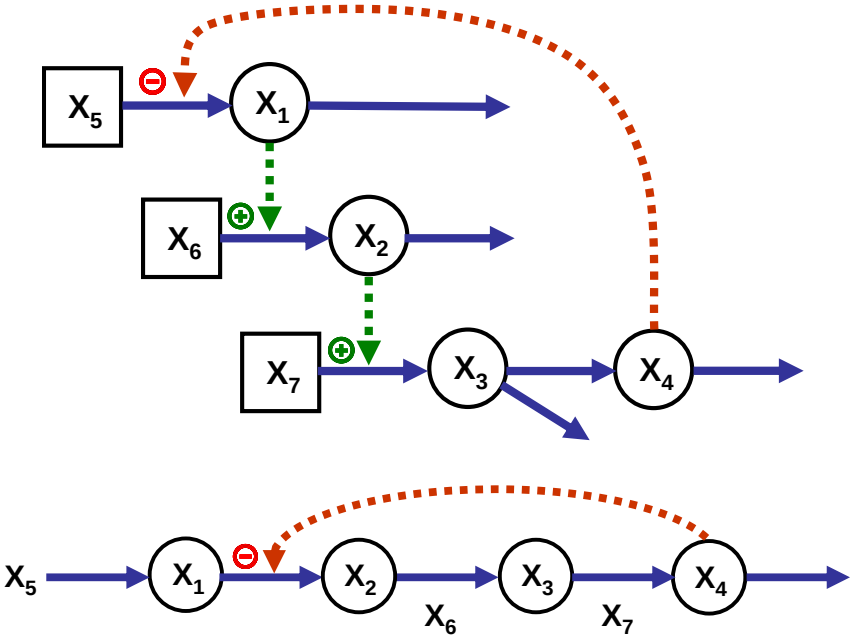


**Figure 4.4: Over-determined cascaded and linear pathway systems with $n$ = 4, $m$ = 3.**
In the example of a linear pathway, the reaction between $X_1$ and $X_2$ may not be accessible to alterations.

This "unsolvable" situation may be addressed in different ways. First, instead of searching for an exact solution, one may solve the corresponding regression problem (see Eq. (4.7)) and find a set of independent variables that moves the system to a steady state that is as close as possible to the target state (Figure 4.5). In the numerical example here, the solution vector is $\mathbf{X}_I = [17.7905 \;\; 9 \;\; 5.4885]^T$, and we see that $X_3$ and $X_4$ are not quite on target.

As a variation on this theme, closeness to the target state may be defined differently for each dependent variable, through the use of appropriate weights. This strategy allows for the option that some "important" dependent variables can be selected to come as closely as possible to their target values, while others are possibly not. Finally, one may ignore some of the dependent variables, whose specific values are not considered as important as those of other variables, and restrict the optimization to a subset of important dependent variables, thereby in effect reducing $n$. Examples for less important variables might be intermediates in linear pathways.

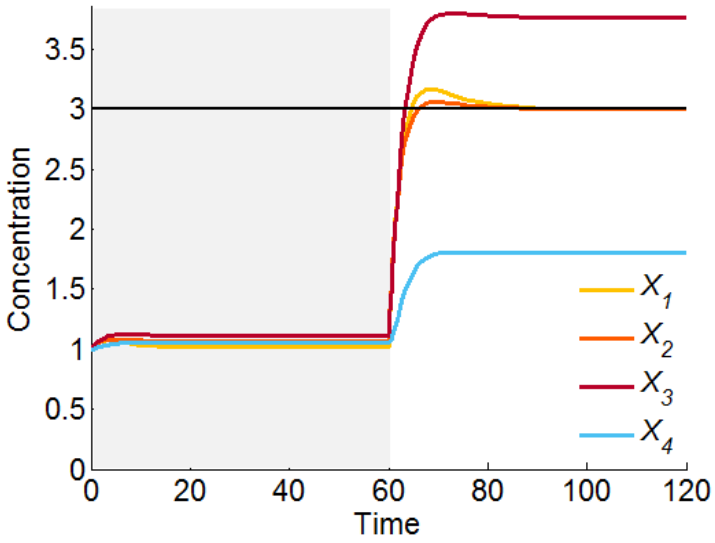


**Figure 4.5: Least squares solution for the over-determined cascaded system in Figure 4.4.**

To be specific, suppose it is most important that variable $X_4$ of the cascaded pathway attain the target value, while other variables are of secondary importance. The original task can be written as

$$\mathbf{y}_D = \mathbf{L}\mathbf{y}_I \quad \text{where} \quad y_{\mathrm{D}} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}, \ \mathbf{L} = \begin{bmatrix} L_{11} & L_{12} & L_{13} \\ L_{21} & L_{22} & L_{23} \\ L_{31} & L_{32} & L_{33} \\ L_{41} & L_{42} & L_{43} \end{bmatrix}, \text{ and } \ y_{\mathrm{I}} = \begin{bmatrix} y_5 \\ y_6 \\ y_7 \end{bmatrix}. \tag{4.12}$$

To enforce that $X_4$ moves to the target, presumably at the cost of other variables, we separate the equation for $X_4$ in Eq. (4.12) from the rest, which yields

$$y_4 = \begin{bmatrix} \mathrm{L}_{41} & \mathrm{L}_{42} & \mathrm{L}_{43} \end{bmatrix} \begin{bmatrix} y_5 \\ y_6 \\ y_7 \end{bmatrix}. \tag{4.13}$$

Using the notation $\mathbf{L}_{123} = \begin{bmatrix} \mathrm{L}_{11} & \mathrm{L}_{12} & \mathrm{L}_{13} \\ \mathrm{L}_{21} & \mathrm{L}_{22} & \mathrm{L}_{23} \\ \mathrm{L}_{31} & \mathrm{L}_{32} & \mathrm{L}_{33} \end{bmatrix}$ and $\mathbf{L}_4 = [\mathrm{L}_{41} \quad \mathrm{L}_{42} \quad \mathrm{L}_{43}]$, the particular solution of $\mathbf{y}_I$ based on this separated equation is now given as

$$\mathbf{y}_I = \mathbf{y}_{I\_PS} + \mathbf{B}_4 \cdot \boldsymbol{\lambda}, \tag{4.14}$$

where

$$\mathbf{y}_{I\_PS} = \mathbf{L}_4^{+} y_4, \tag{4.15}$$

$\mathbf{B}_4$ is a 3×2 matrix where each column is a basis vector of the null space of $\mathbf{L}_4$, and $\boldsymbol{\lambda}$ is any real-valued 2-dimensional vector. Having enforced that the fourth variable will reach the target value, we still have options for the remaining independent variables. Namely, the equation

$$\begin{aligned} \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix}^T &= \mathbf{L}_{123}\mathbf{y}_I \\ &= \mathbf{L}_{123}(\mathbf{y}_{I\_PS} + \mathbf{B}_4 \cdot \boldsymbol{\lambda}) \end{aligned} \tag{4.16}$$

allows us to define criteria such as a least-squares error for the remaining variables, which correspond to different choices for $\boldsymbol{\lambda}$. For instance, we can use the pseudo-inverse to define

$$\boldsymbol{\lambda} = (\mathbf{L}_{123} \cdot \mathbf{B}_4)^{+} \left( \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix}^T - \mathbf{L}_{123} \cdot \mathbf{y}_{I\_PS} \right), \tag{4.17}$$

which yields the solution as

$$\mathbf{y}_I = \mathbf{y}_{I\_PS} + \mathbf{B}_4 \cdot (\mathbf{L}_{123} \cdot \mathbf{B}_4)^{+} \left( \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix}^T - \mathbf{L}_{123} \cdot \mathbf{y}_{I\_PS} \right). \tag{4.18}$$

The result of this operation is shown in Figure 4.6. In comparison with Figure 4.5, $X_4$ now reaches the target value 3 exactly, while the remaining variables approach the value 3 only approximately. In particular, the improvement in $X_4$ is "paid for" with an inferior performance of $X_3$. The solution vector of independent variables in this case is $\mathbf{X}_I$ = [97.2759 9.0 116.8222]$^T$. If $X_3$ is most important in the same system, the solution vector is $\mathbf{X}_I$ = [12.7188 9.0 3.0]$^T$ and $X_4$ overshoots the target (plot not shown).

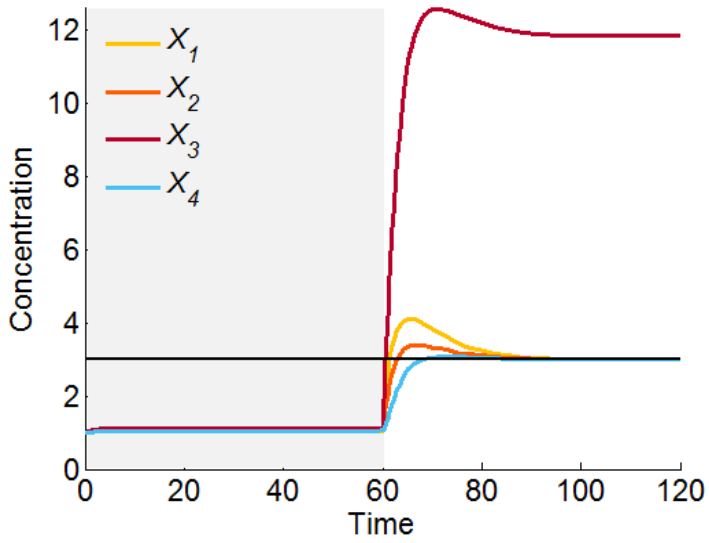


**Figure 4.6: Solution for the over-determined cascaded system in Figure 4.4, where $X_4$ is forced to reach the target state 3.**

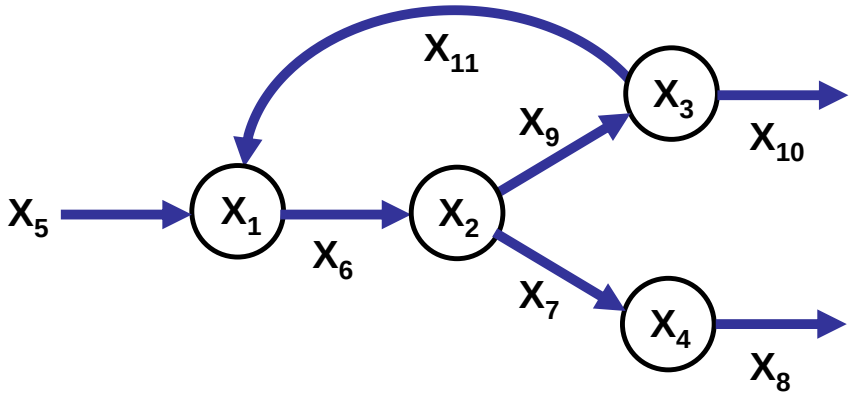


**Figure. 4.7: Branched pathway with a substrate cycle.**
The system contains four dependent variables (circles) and seven independent variables, which model the system input ($X_5$) and catalyzing enzymes ($X_6$, …, $X_{11}$). The system is representative of the most prevalent situation where $n < m$.

The most pertinent case is $n < m$. A representative example is the pathway shown in Figure 4.7, which has four dependent and seven independent variables. The S-system was constructed according to usual guidelines (see Table 4.1 for parameter values), and as before, we set all independent variables arbitrarily set to 1.2 and solved the system from (1, 1, 1, 1) to its normal steady state. Subsequently, we assumed that the environmental demand changed by requiring all target values for the dependent variables to assume the value of 2.

Because $n < m$, the solution consists of a space that can be expressed by a particular solution plus a linear span of a basis of the null space of $\mathbf{L} = -\mathbf{A}_D^{-1}\mathbf{A}_I$. The particular solution is computed as

$$\mathbf{y}_D = \mathbf{L}\mathbf{y}_I \tag{4.19}$$

$$\mathbf{y}_{I\_PS} = \mathbf{L}^{+}\mathbf{y}_D \tag{4.20}$$

and any feasible solution can be characterized by the particular solution plus an arbitrary vector in the null space of $\mathbf{L}$:

$$\mathbf{y}_I = \mathbf{B}\cdot\boldsymbol{\lambda} + \mathbf{y}_{I\_PS}, \tag{4.21}$$

where $\boldsymbol{\lambda}$ may be any 3-dimensional real-valued vector and **B** is a matrix in which each column is a basis vector of the null space of **L**.

Choosing *any* $\mathbf{y}_I$ inside this solution space is guaranteed to lead the system to the target steady state. Two examples of admissible solutions in Cartesian space are $\mathbf{X}_I$ = [1.9363 1.4646 0.7293 0.7293 1.4705 0.7755 0.8658]$^T$ and $\mathbf{X}_I$ = [4.3355 2.8505 2.1973 2.1973 1.8490 1.1861 1.4149]$^T$. The former of these solutions is the least-squares solution, while the latter is the least-squares solution plus the first basis vector of the null space. These and other solutions within the admissible space move the system to the target steady state of (2, 2, 2, 2) as expected, but the transient behaviors of these systems are different, and it is not *a priori* clear how to manipulate them.

Interestingly, it is possible to alter any solution to some degree in a targeted fashion by controlling the basis vectors of the three-dimensional null space of **L**. In the given numerical case, the basis vectors are

$$\begin{aligned}
\mathbf{B}_1 &= [0.403 \;\; 0.333 \;\; 0.5514 \;\; 0.5514 \;\; 0.1145 \;\; 0.2124 \;\; 0.2456]^T, \\
\mathbf{B}_2 &= [0.0226 \;\; 0.0091 \;\; -0.1991 \;\; -0.1991 \;\; 0.2173 \;\; 0.9322 \;\; -0.063]^T, \\
\mathbf{B}_3 &= [-0.1285 \;\; 0.0201 \;\; -0.1776 \;\; -0.1776 \;\; 0.2178 \;\; -0.0607 \;\; 0.9321]^T.
\end{aligned} \tag{4.22}$$

These basis vectors can be computed directly in Matlab with the *Null* command, which applies singular value decomposition to obtain an orthogonal basis set. Different effects are observed when any of these basis vectors is altered. For instance, increasing $\mathbf{B}_1$ by a positive factor causes all responses to speed up (Figure 4.8), while increasing $\mathbf{B}_2$ or $\mathbf{B}_3$ causes $X_1$, $X_2$ and $X_3$ to accelerate but $X_4$ to slow down (data not shown). Thus, the transient behavior can be controlled to some degree through the basis vectors. However, the effects of such manipulations are difficult to predict, and it is more straightforward to use direct optimization methods as we will discuss them next.

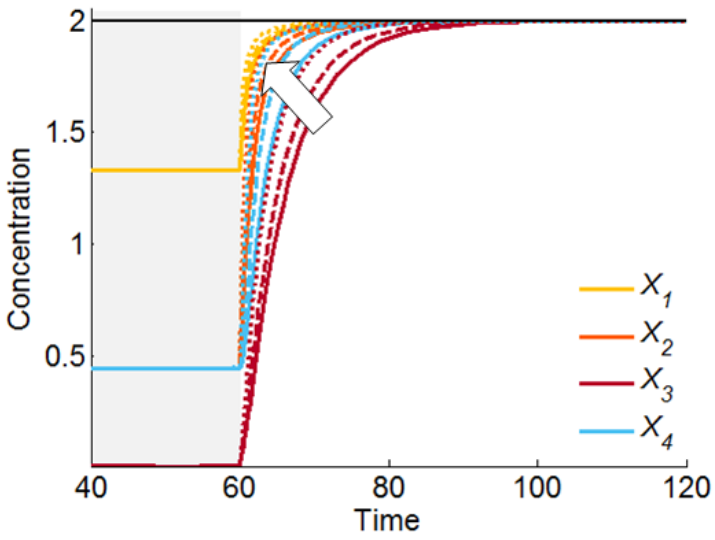


**Figure 4.8: Manipulation of the basis vectors permits modest changes in transient speed.**
Here, increasing $B_1$ causes all transients to accelerate (arrow). Solid lines: no acceleration; dashed lines: acceleration by increasing $B_1$ twofold; dotted lines: acceleration by increasing $B_1$ fourfold.

### 4.4 Optimal Operating Strategies

The computation of the pseudo-inverse in the steady-state equations of the S-system, along with the characterization of the null space, results in the space of all possible solutions. Within this space, any computed resetting of the independent variables leads to a desired steady state in terms of the dependent variables. While it is mathematically and practically satisfying to have a concise representation of this solution space, one will wonder whether some admissible solutions within this space are "better" than others. Clearly, the answer requires optimization, which, interestingly, does not need an explicit characterization of the solution space *per se*. The optimization does require an objective function, which is to be selected according to the chosen criteria of functional effectiveness.

Operating principles have not yet been analyzed often enough to permit a listing of "typical" criteria of functional effectiveness, and judging by the exploration of design

principles, one might expect them to change from one application to another. Among likely, generic criteria one will often establish similar metrics as for design principles, which often include local stability, modest gains and sensitivities, and tolerance of the steady state to perturbations. Also as in the case of design principles, one might prefer fast response times and bounded transients. Another typical criterion in superior designs is a minimal accumulation of intermediates. Here, this criterion is automatically satisfied when a complete target profile of steady-state values is mandated, but if no target values for intermediates are specified, it may indeed serve as a criterion.

In addition to these criteria gleaned from design principle analysis, operating strategies are distinguishable in other respects. In the work presented here, we focus primarily on two aspects that appear to be particularly pertinent: the *collective deviation* of independent variables from their nominal levels, and the *number of independent variables* that are to be changed. These criteria are important to a cell, because they are directly related to the effort that has to be expended in terms of gene expression and the dynamics of RNAs and proteins [157], and to the degree of possible side effects from such changes. Secondarily, we will look into the profiles of transients between steady states. One could presumably study a variety of additional criteria, such as a favorable dynamic sensitivity profile [158,159].

To formalize the deviation from normal operation, we introduce a vector $\mathbf{d}$ that represents the change in the vector of independent variables such that the system reaches the target steady state $\tilde{\mathbf{y}}_D$, which is assumed to be known. With these definitions, we can formulate the target state as

$$\tilde{\mathbf{y}}_D = \mathbf{A}_D^{-1}\mathbf{b} - \mathbf{A}_D^{-1}\mathbf{A}_I(\mathbf{y}_I + \mathbf{d}), \tag{4.23}$$

and this expression can be rearranged as a linear constraint on $\mathbf{d}$. Namely, we can write

$$\mathbf{A}_D^{-1}\mathbf{A}_I\mathbf{d} = \mathbf{A}_D^{-1}(\mathbf{b} - \mathbf{A}_I\mathbf{y}_I) - \tilde{\mathbf{y}}_D. \tag{4.24}$$

Now let

$$z_i = \begin{cases} 1 & \text{if the catalytic step coded by } d_i \text{ is induced to reach the steady state} \\ 0 & \text{otherwise} \end{cases}. \quad (4.25)$$

If all $z_i$ are set to 1, the optimization task allows every independent variable to change as long as the linear constraints are satisfied, but the identification of specific solutions still depends on the dimension and rank of $\mathbf{A}_D^{-1}\mathbf{A}_I$ as well as the chosen criteria of functional effectiveness.

One of the most commonly used criteria for finding a particular solution is the total squared error *E*, which in this case can be written as

$$E = \left\| \mathbf{A}_D^{-1}(\mathbf{b} - \mathbf{A}_I \mathbf{y}_I) - \tilde{\mathbf{y}}_D - \mathbf{A}_D^{-1}\mathbf{A}_I \mathbf{d} \right\|^2, \quad (4.26)$$

where $\|\bullet\|^2$ is the 2-norm. The solution $\hat{\mathbf{d}}$ with the lowest *E* corresponds to an optimal operating strategy where the first criterion, *i.e.*, the collective deviation in independent variables from their nominal values, is minimized.

The second criterion requires finding a minimum set of independent variables whose alteration is necessary for reaching the target steady state. This task is equivalent to solving the following Mixed Integer Linear Programming (MILP) problem:

$$\begin{aligned} &\min \sum_{i=1}^{m} z_i \text{ subject to} \\ &\mathbf{A}_D^{-1}\mathbf{A}_I \mathbf{d} = \mathbf{A}_D^{-1}(\mathbf{b} - \mathbf{A}_I \mathbf{y}_I) - \tilde{\mathbf{y}}_D \\ &d_i \geq 0 \\ &d_i \leq z_i D \text{ (D is an arbitrarily large positive number)} \\ &\mathrm{z}_i : 0/1 \\ &\forall i = 1, \ldots, m \end{aligned} \quad (4.27)$$

The CPLEX solver in AMPL can be used to solve this type of MILP.

Similar to optimization tasks in the field of biotechnology, where the typical objective is the maximization of a metabolite pool or flux, it is here also possible to account for constraints on concentrations and fluxes [85,160,161,162,163] as well as more complex limitations such as metabolic burden [128] or the feasibility of parameter

regions that correspond to admissible physiological states [144]. In particular, the metabolic burden, which is associated with the total mass of all proteins (*cf.* discussion in [128]), can be an important issue of cellular protein economy because it was shown for the case of recombinant bacteria that the growth rate decreased monotonically with increasing numbers of introduced plasmid copies (*e.g.*, [164,165,166]).

Moreover, one should expect that it is easier to up- or down-regulate some genes or enzyme activities than others. In fact, it might not be practically feasible to change some enzyme activities at all. If so, the corresponding independent variables in the model are off limits in the selection of any viable operating strategies. Other processes might be accessible to manipulations but limited in the degree of alteration. We will discuss some of these concepts in Section 4.5.

## 4.5 Case Study

As a specific case study, we consider the response of yeast cells to heat stress. The first indications of such a response are observable within minutes of the initiation of heat stress: transcription factors are mobilized and translocated [167], and numerous genes respond with strong changes in expression [168,169,170]. At the proteomic level, heat shock proteins emerge in high numbers [171,172,173]. At the metabolic level, a significantly altered profile of sphingolipids guides the expression of some key genes [174], and, most important for the following illustration, the protective disaccharide trehalose is produced in huge amounts [157,175].

Several modeling studies have investigated the dynamics of trehalose upon heat shock in recent years [1,144,145,146,157,176,177], which allows us to keep the discussion of background information to a minimum. In a nutshell, material is siphoned off glycolysis at the level of glucose 6-phosphate and channeled toward the production of glucose 1-phosphate, UDPG, glycogen, trehalose 6-phospate and trehalose, with trehalose accumulating in large quantities. The enzyme trehalase splits trehalose into two

glucose molecules and thereby completes the trehalose cycle (see Figure 4.9). Because the present study is focused on methodological advances rather than new biological insights, we take the S-system model of the trehalose cycle in [1] at face value and analyze alternative operating strategies.

The S-system equations describing the system were taken directly from [1]. They are

$$\text{Glucose:} \quad \dot{X}_1 = 31.912 X_0^{0.968} X_2^{-0.194} X_7^{0.00968} X_8^{0.968} X_{19}^{0.0323} - 89.935 X_1^{0.75} X_6^{-0.4} X_9$$

$$\text{G6P:} \quad \dot{X}_2 = 142.72 X_1^{0.517} X_2^{-0.179} X_3^{0.183} X_6^{-0.276} X_9^{0.689} X_{12r}^{0.311}$$
$$- 30.120 X_1^{-0.00333} X_2^{0.575} X_3^{-0.17} X_4^{0.00333} X_{10}^{0.5111} X_{11}^{0.0667} X_{12f}^{0.411} X_{17}^{0.0111}$$

$$\text{G1P:} \quad \dot{X}_3 = 7.8819 X_2^{0.394} X_3^{-0.392} X_4^{-0.010} X_5^{0.0128} X_{12f}^{0.949} X_{15r}^{0.0513}$$
$$- 76.434 X_2^{-0.412} X_3^{0.593} X_{12r}^{0.718} X_{13}^{0.180} X_{15f}^{0.103} \tag{4.28}$$

$$\text{UDPG:} \quad \dot{X}_4 = 11.070 X_3^{0.5} X_{13} - 3.4556 X_1^{-0.0429} X_2^{0.214} X_4^{0.386} X_{14}^{0.857} X_{17}^{0.143}$$

$$\text{Glycogen:} \quad \dot{X}_5 = 11.060 X_2^{0.040} X_3^{0.320} X_4^{0.160} X_{14}^{0.600} X_{15f}^{0.400}$$
$$- 4.9290 X_2^{-0.04} X_4^{-0.04} X_5^{0.25} X_{15r}^{0.200} X_{16}^{0.800}$$

$$\text{T6P:} \quad \dot{X}_6 = 0.19424 X_1^{-0.300} X_2^{0.300} X_4^{0.300} X_{17} - 1.0939 X_6^{0.200} X_{18}$$

$$\text{Trehalose:} \quad \dot{X}_7 = 1.0939 X_6^{0.200} X_{18} - 1.2288 X_7^{0.300} X_{19}$$

Of primary interest here is the response of yeast to heat stress, which affects most of the reactions steps in the pathway. According to literature studies (cited in [1]), the alterations among the dependent and independent variables under heat stress are distinctly different, with some variables and steps changing substantially and others not as much or not at all (Tables 4.2 and 4.3).

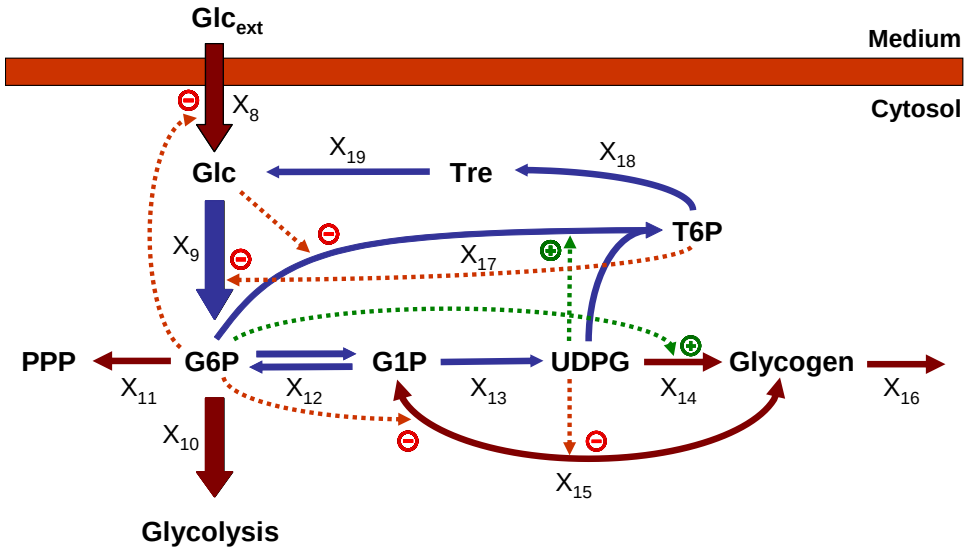


**Figure 4.9: Diagram of the trehalose cycle (solid blue arrows) in yeast.**
Solid brown arrows represent other pertinent reactions. The main glycolytic flux is presented with heavy arrows. Red dotted arrows with associated minus signs indicate inhibition, while green dotted arrows associated with plus signs indicate activation. Abbreviations: $Glc_{ext}$: external glucose; Glc: internal glucose ($X_1$); G6P: glucose 6-phosphate ($X_2$); G1P: glucose 1-phosphate ($X_3$); UDGP: uridine diphosphate glucose ($X_4$); glycogen ($X_5$); T6P: trehalose 6-phosphate ($X_6$); Tre: trehalose ($X_7$); PPP: pentose phosphate pathway. $X_8$, …, $X_{19}$ represent independent variables (see Table 4.4).

**Table 4.2: Dependent variables of the S-system model (Eq. (4.28)) of the trehalose cycle.** Steady-state values under optimal temperature conditions were collected from the literature [1]; heat-stress values (scaled by optimal steady-state values) computed with the S-system model upon changes in independent variables as shown in Table 4.4.

| Metabolite | Variable Name | Steady-State Concentration [mM] under Optimal Temperature Conditions (from the Literature) | Computed Fold Change in Steady-State Concentration during Heat Stress (Scaled by Normal Steady State) |
|---|---|---|---|
| **Glucose** | $X_1$ | 0.03 | 1.46 |
| **G6P** | $X_2$ | 1 | 5.54 |
| **G1P** | $X_3$ | 0.1 | 3.99 |
| **UDGP** | $X_4$ | 0.7 | 2.69 |
| **Glycogen** | $X_5$ | 1 | 55.8 |
| **T6P** | $X_6$ | 0.02 | 4.28 |
| **Trehalose** | $X_7$ | 0.05 | 103 |

In this case, $n$ = 7 and $m$ = 12, which indicates quite a bit of flexibility among different solutions. Application of the pseudo-inverse method reveals the space of all admissible solutions; an example of a possible solution is $\mathbf{X}_I$ = [5.4169 5.2877 0.9723 2.3770 2.3004 3.0197 2.8855 2.8574 2.1650 2.9739 4.4620 1.4873]$^T$. $\mathbf{X}_I$ is computed using the pseudo-inverse of **L** and the original basis of the null space of **L**, which was obtained through singular value decomposition in MATLAB, and $\boldsymbol{\lambda}$ = [1 1 1 1 1]$^T$. As to be expected, this vector of independent variables moves the system to the target steady state. However, the solution is much slower than the observed solution (Figure 4.10).

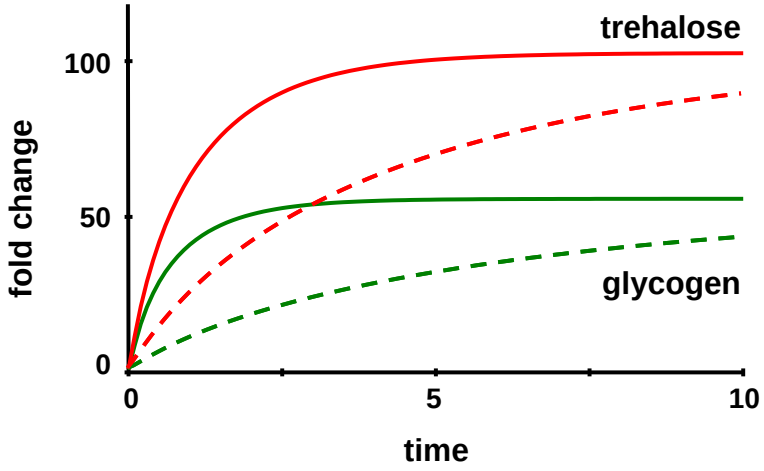


**Figure 4.10: A possible solution within the space characterized by the pseudo-inverse method (dashed), in comparison with the nominal solution discussed in [1].** While both solutions eventually reach the same steady state, the transient of the solution computed here is comparatively slow (see text for details).

The solution space obtained with the pseudo-inverse method is 5-dimensional, and a basis is

$\mathbf{B}_1$ = [0.1635 0.1582 0.2144 -0.1562 0.0986 -0.1328 -0.1543 0.9026 0.1099 -0.0034 -0.0034 -0.0034]$^T$

$\mathbf{B}_2$ = [-0.2260 -0.2188 -0.3985 0.2716 0.3480 0.3708 0.4330 0.2639 0.3907 -0.0017 -0.0017 -0.0017]$^T$

$\mathbf{B}_3$ = [-0.2260 -0.2087 -0.1967 -0.7547 0.0363 0.0466 0.0021 0.0002

$0.0016 \;\; 0.3128 \;\; 0.3128 \;\; 0.3128]^T$ (4.29)

$\mathbf{B}_4 = [0.3576 \;\; 0.3568 \;\; 0.4235 \;\; 0.1207 \;\; 0.1771 \;\; 0.2723 \;\; 0.2627 \;\; -0.1569$

$0.1578 \;\; 0.3301 \;\; 0.3301 \;\; 0.3301]^T$

$\mathbf{B}_5 = [-0.1451 \;\; -0.1291 \;\; -0.2276 \;\; 0.5556 \;\; -0.1360 \;\; -0.2193 \;\; -0.3147$

$0.1558 \;\; -0.1971 \;\; 0.3527 \;\; 0.3527 \; 0.3527]^T$

As in the illustrative example of a branched pathway, it is to some degree possible to affect the transient speed by manipulating the basis vectors. Tuning $\mathbf{B}_1$ or $\mathbf{B}_2$ causes the glycogen concentration to speed up but has almost no effect on trehalose or the other variables. Increasing $\mathbf{B}_3$ accelerates trehalose and no other variables, increasing $\mathbf{B}_4$ speeds up both trehalose and glycogen, while increasing $\mathbf{B}_5$ speeds up trehalose but slows down glycogen production (Figure 4.11).

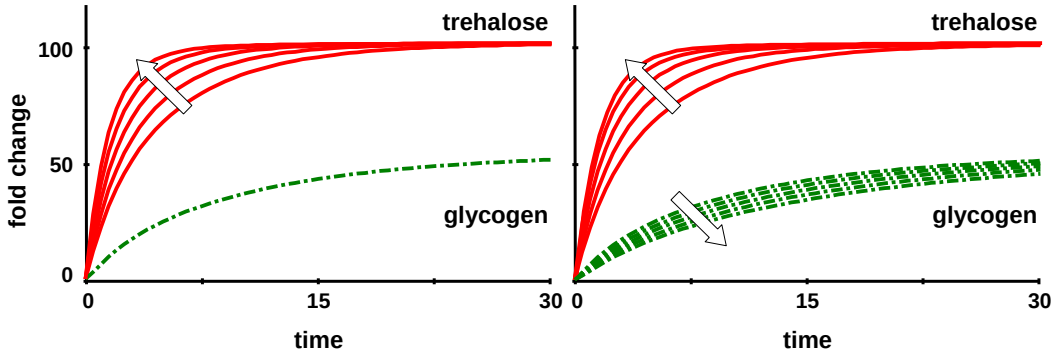


**Figure 4.11: The solutions obtained with the pseudo-inverse method can be manipulated by modifying the basis vectors.**
In the left panel, basis vector $\mathbf{B}_3$ was multiplied with factors 1, …, 5 (in direction of the arrow); this action did not affect the glycogen profile. In the right panel, basis vector $\mathbf{B}_5$ was multiplied with factors 1, …, 5, in direction of the arrows. All solutions eventually reach the same target steady state.

In contrast to exploring the entire solution space, the direct optimization method allows us to select criteria of functional effectiveness a priori and to optimize the solution toward these criteria under the constraint that the target steady state is reached. As the first example, suppose the overriding criterion is to alter the independent variables as little as possible in magnitude. Least-squares optimization toward this criterion yields a solution that not only reaches the target steady state but also exhibits only modest variations in enzyme activities (Table 4.3; column 4).

As a second example, we mandate to keep the number of altered independent variables to a minimum. MILP optimization reveals that this minimum number is 7, and the steady state is reached upon quite strong alterations in this minimum set (Table 4.3; column 5).

Both results are interesting. First, the constrained least-squares solution turns out to be very similar to the nominal solution, which indicates a similar strategy as in the case of the diauxic shift (see introduction and [142]). Second, the minimum-set solution shows drastically different values than the nominal solution and identifies glycogen phosphorylase as the most dispensable reaction step. In an entirely different study [157], this same step was also identified as only modestly relevant for the trehalose response. The question of which strategy is superior depends on the criteria of functional effectiveness. One might say that the least-squares solution should be the preferred means of operation because all variables remain as close to their normal operating points as possible and the strategy produces glycogen faster. Yet, if the cost of gene expression and the production of transcription factors and mRNAs is a major concern, then the minimum-set solution might be superior because its sum of independent variables is smaller. In short, the superiority is context-dependent rather than universal.

**Table 4.3: Different implementations of computed heat stress responses, which all lead to exactly the same target steady state.**

| Catalytic or Transport Step | Variable Name | Nominal* | Least Squares | Minimum Set | Least Squares ($X_8$, $X_{10}$, $V_5^+$, and $V_7^+$ fixed) | Minimum Set ($X_8$, $X_{10}$, $V_5^+$, and $V_7^+$ fixed) |
|---|---|---|---|---|---|---|
| **Glucose transport** | $X_8$ | 8 | 2.0096 | 4.6155 | 8 (fixed) | 8 (fixed) |
| **Hexokinase/Glucokinase** | $X_9$ | 8 | 1.9440 | 4.4334 | 8 | 8 |
| **Phosphofructokinase** | $X_{10}$ | 1 | 0.3577 | 1 | 1 (fixed) | 1 (fixed) |
| **G6P dehydrogenase** | $X_{11}$ | 6 | 0.8745 | 1 | 6.2371 | 1.6377 |
| **Phosphoglucomutase** | $X_{12}$ | 16 | 0.8435 | 1.1046 | 15.5916 | 38.0406 |
| **UDPG pyrophosphorylase** | $X_{13}$ | 16 | 1.1110 | 1.5932 | 14.9673 | 149.5541 |
| **Glycogen synthase** | $X_{14}$ | 16 | 1.0616 | 1.4620 | 14.8016 | 217.1534 |
| **Glycogen phosphorylase** | $X_{15}$ | 50 | 1.0512 | 1 | 56.1937 | 1 |
| **Glycogen use** | $X_{16}$ | 16 | 0.7965 | 1 | 15.5396 | 42.5464 |
| **α, α-T6P synthase** | $X_{17}$ | 12 | 1.0942 | 2 | 12 | 12 |
| **α, α-T6P phosphatase** | $X_{18}$ | 18 | 1.6413 | 3 | 18 | 18 |
| **Trehalase** | $X_{19}$ | 6 | 0.5471 | 1 | 6 | 6 |

*Heat-induced fold-increase in activity used in the model (Eq. (4.28))

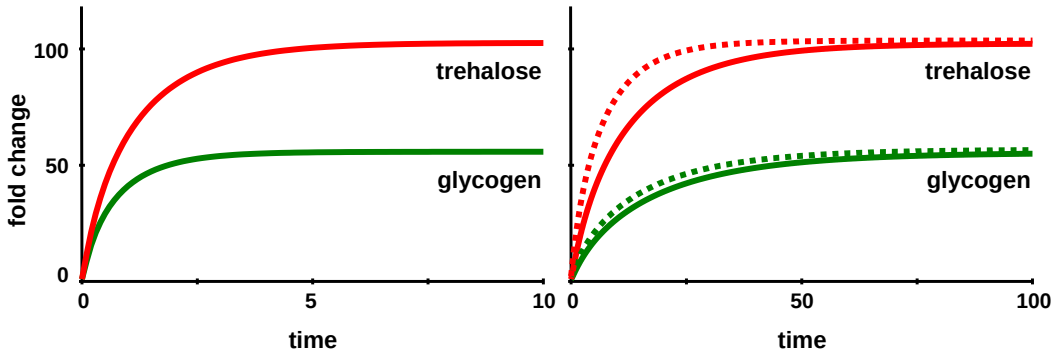


**Figure 4.12: All solutions eventually reach the exact same steady state and the transients have similar shapes, but the timing is quite different.**
While glycogen and trehalose in the nominal solution come close to their steady state values within about 5 time minutes (left panel), reaching the same levels takes ten or more times as long in the least-squares (right panel; solid lines) and minimum-set (right panel; dotted lines) solutions (note different time scales). Other variables respond on a time scale that is more similar to the nominal solution (not shown).

Table 4.3 seems to indicate that much "cheaper" solutions than the nominal solution can be found, which raises the question of why the nominal solution employs alterations in independent variables that are so much more dramatic than the least squares or minimum set solutions. At least one answer can be found in the response time: although all solutions reach exactly the same steady state, the nominal solution is more than ten times faster than the least squares and minimum-set solutions (Figure 4.12; note different time scales).

The issue of drastically different transient speeds begs the question of whether and how the least-squares and minimum-set solutions could be accelerated. The most direct way of accomplishing acceleration arises if every flux contains its own independent variable. For instance, if every flux is governed by an enzyme which enters the flux with a kinetic order of 1, then multiplication by the same factor $\varphi > 1$ will speed up the dynamics of the entire system by $\varphi$. This advance does not come for free though, because the cost of the solution with respect to the chosen criterion increases and the result may no longer be optimal. For instance, the metabolic burden, which roughly corresponds to the sum of independent variables, increases $\varphi$-fold. An increased

metabolic burden can be a disadvantage because it puts additional stress on the cell due to higher levels of transcription and translation [164]. If minimal metabolic burden is indeed a pertinent criterion of functional effectiveness, the totality of changes in independent variables should be kept as small as possible.

If the independent variables have different kinetic orders or appear in several equations, a systemic speed-up may still be possible. Specifically, one has to solve the equations

$$\begin{aligned} \prod_{j=n+1}^{n+m} \tilde{X}_j^{g_{ij}} &= \varphi \prod_{j=n+1}^{n+m} X_j^{g_{ij}} \\ \prod_{j=n+1}^{n+m} \tilde{X}_j^{h_{ij}} &= \varphi \prod_{j=n+1}^{n+m} X_j^{h_{ij}} \end{aligned} \tag{4.30}$$

for all $i = 1, \ldots, n$. In the trehalose case, these conditions result in a set of 14 linear equations with 12 unknowns (in log coordinates), which has no algebraic solution. Nevertheless, one can obtain a solution in a least-squares sense, which indeed leads to an acceleration of the transients and approximately reaches the target steady state. The required changes in independent variables are presented in Table 4.4, where $\varphi_{LS} = 11.19$ and $\varphi_{MS} = 6.29$ are the acceleration factors for the least-squares and the minimum-set solutions, respectively. These factors are computed based on the settling time $\tau$, which here is the amount of time needed for trehalose to reach and stay within 95% of its nominal heat stress value. While the resulting trehalose profiles are essentially the same as in the nominal scenario, the glycogen trends are still slower (Figure 4.13). Interestingly, the steps directly associated with the dynamics of trehalose are very similar to the nominal solution, and the glycogen phosphorylase step is again much lower (Table 4.4).

Distinctly different solutions to speeding up the transients could possibly be reached in two ways. First, the cell could initiate a fast transient toward a steady state with more extreme values than needed, and in a second phase relax these values toward

the true target state. This strategy is expected to incur overshoots before the true target steady state is reached [152]. Second, it is possible to compute settings in independent variables that reach states that are not steady states. These computations require methods of nonlinear control theory, which were demonstrated for S-systems elsewhere [178].

## 4.6 Discussion

Deciphering how nature solves problems has been the dream of scientists for a long time. Consequently, enormous effort has been devoted to shining light on operating procedures in nature, dissecting systems, and identifying and characterizing processes that cells employ to solve specific problems. Given the seemingly unlimited variability and complexity of tasks that need to be addressed, a comprehensive understanding of operating procedures, let alone operating strategies or even operating principles, will not be gained in the foreseeable future. Nonetheless, the overwhelming magnitude of the challenge does not suggest that we give up, but that even small advances might be beneficial on our long journey.

**Table 4.4: Accelerated least squares and minimum set solutions for the trehalose cycle**

| Catalytic or Transport Step | Nominal | Least Squares (accelerated) | Minimum Set (accelerated) |
|---|---|---|---|
| **Glucose transport** | 8 | 22.4731 | 28.8238 |
| **Hexokinase/Glucokinase** | 8 | 21.7616 | 27.7065 |
| **Phosphofructokinase** | 1 | 5.2507 | 7.9468 |
| **G6P dehydrogenase** | 6 | 1.2416 | 1 |
| **Phosphoglucomutase** | 16 | 9.4252 | 6.8941 |
| **UDPG pyrophosphorylase** | 16 | 12.4311 | 9.9536 |
| **Glycogen synthase** | 16 | 11.8831 | 9.1365 |
| **Glycogen phosphorylase** | 50 | 11.7559 | 6.2449 |
| **Glycogen use** | 16 | 8.9158 | 6.2494 |
| **α, α-T6P synthase** | 12 | 12.2458 | 12.4968 |
| **α, α-T6P phosphatase** | 18 | 18.3691 | 18.7455 |
| **Trehalase** | 6 | 6.1230 | 6.2485 |

Thanks to high-throughput techniques of molecular biology, the availability of large datasets has grown immensely and will continue to increase. Along with this increase will be a more and more pressing need to find means of interpretation and of comparing similar, yet structurally different solution strategies. Similar to the investigation of design principles, the study of operating principles is expected to lead to the discovery of motifs, which will provide explanations of naturally evolved systems as well as guidance regarding the design of new systems within the field of synthetic biology.

We have shown in this Chapter that a small sub-class of cellular tasks can be addressed quite efficiently with mathematical and computational tools. Namely, we propose methods for investigating the situation where a biological system is forced to move to a new steady state, which we assume to be known. For example, in the heat stress scenario discussed here, the cell must accumulate sufficient amounts of trehalose and possibly glycogen, while internal glucose and trehalose 6-phosphate need to be carefully controlled, because they cause adverse effects in high concentrations [175,179,180]. Thus, some pools in a pathway need to be altered substantially, while others must remain more or less at their nominal level. We show here that such tasks can be formulated rigorously in the language of linear algebra and constrained optimization.

The analysis yields two main results. First, it defines the entire solution space of the problem, and second, it allows a direct system optimization toward given criteria of functional effectiveness. The elegance of these solutions is primarily due to the special structure of S-system models, whose steady states are characterized by systems of linear equations. With the exception of Lotka-Volterra [181,182,183] and lin-log models [184,185], whose steady states are also governed by linear equations, it seems very difficult to obtain similarly general results with *ad hoc* models, such as pathway systems that are represented with Michaelis-Menten rate laws and their generalizations.

Interestingly, Generalized Mass Action (GMA) representations within BST ([44,186]; Eq. (1.3)), as well as other model structures, may permit numerical solutions under favorable conditions, although these solutions are not as general as in the case of S-systems. Namely, consider the important special case where each flux representation contains at most one independent variable, which enters the flux in a linear fashion, as it is typical for most enzymes. If all parameter values and the target steady state are known, all terms in the steady-state equations either become linear functions of one independent variable, or they do not contain an independent variable at all. Furthermore, outside the independent variables, all other components of each term combine to a single numerical value, so that the entire system of steady-state equations is linear in the independent variables. As in the cases shown here, this system may have a unique solution or be over- or underdetermined, and it can be analyzed in each case with methods of linear algebra and optimization. The condition of linearity with respect to independent variables can actually be further relaxed, for instance, to the requirement that the same independent variable, if it appears in different terms, always has the same kinetic order.

The tasks and solutions proposed here are reminiscent of optimization problems that have been analyzed in the field for two decades [85,128,144,160,161,163]. However, the two lines of investigation represent different aspects of targeted alterations in pathways. In the typical optimization tasks in biotechnology or metabolic engineering, a metabolite pool or flux is to be maximized, while other features of the steady state profile are rather irrelevant as long as they remain within general physiological constraints. As a consequence, the task typically has a clearly defined, single optimal solution, although in some cases alternative optima with the same value of the objective function occur, and it is furthermore possible to investigate multi-objective optimization tasks [128,187]. In the analysis here, the primary requirement is that the system must reach a specified steady-state profile. This task often admits an entire solution space, within which the system must operate. Within this space, questions of superiority of one solution over another

with respect to selected criteria can be explored. Functional effectiveness is not usually considered in biotechnological optimization, but in the case analyzed here provides the metric for comparing alternative strategies and declaring one solution superior to another.

An unresolved issue is the definition of criteria for functional effectiveness, which are not necessarily known *a priori*. Is it advantageous to up-regulate just a few genes substantially, or is it better to up-regulate many genes by a small amount? We do not yet have answers to such questions, but we have taken a first step by asking these questions and by suggesting that it might be advisable to observe how nature solves tasks in order for us to develop ideas for what types of operating strategies might be candidates for optimality. Moreover, the work presented here suggests tools for comparing different solutions with objectivity and for declaring superiority of different alternatives once criteria are established.

# CHAPTER 5

# FUNCTIONAL ANALYSIS OF METABOLIC CHANNELING AND REGULATION IN LIGNIN BIOSYNTHESIS: A COMPUTATIONAL APPROACH[6]

Chapter 4 presented two methods for characterizing alternative strategies employed by biological systems to shift operation from a normal steady state to a new target steady state. Such transitions are quite common and include not only stress responses or other adaptations to external perturbations, but also normal, physiological processes, such as the reprogramming of lignin biosynthesis during plant development. A direct application of the methods developed in Chapter 4 to the case of lignin biosynthesis, however, is complicated due to the fact that the design and operation of the G- and S-channels are not yet sufficiently characterized. Nonetheless, similar in concept to Chapter 4, the goal here is to gain deeper insights into the functional role of metabolic channeling and its associated regulation. Thus, this Chapter will focus on exploring the design space (see below) and comparing alternative operating solutions that seem *a priori* equally valid. The methods are somewhat different in approach and implementation from those in Chapter 4, but the philosophy of comparatively assessing design and operating principles is the same.

---

[6] Adapted from: Lee, Y., Escamilla-Treviño, L., Dixon, R.A. and Voit, E.O. (*submitted*) Functional Analysis of Metabolic Channeling and Regulation in Lignin Biosynthesis: A Computational Approach.

## 5.1 Introduction

The metabolic scaffold for the biosynthesis of the three building blocks of lignin was originally seen as a grid-like structure [9], but this initial structure has been revised and refined and is now understood as an essentially linear pathway with only a few branch points (Figure 5.1). Although this generic pathway structure is now widely accepted, it has become clear that different lineages of vascular plants have evolved variants that engage distinct biosynthetic strategies. An interesting example is the model legume *Medicago truncatula*, where the characterization of two distinct cinnamoyl CoA reductases, CCR1 and CCR2, has suggested parallel routes from caffeoyl CoA to coniferyl aldehyde (Figure 5.1) [59]. A more unusual case is the lycophyte *Selaginella moellendorffi*. Functional analyses of two enzymes recently discovered in this species, *Sm*F5H and *Sm*COMT, support the notion that *S. moellendorffi* may have adopted a non-canonical pathway to synthesize coniferyl and sinapyl alcohol, which differs from that in angiosperms (Figure 5.1) [188,189,190].

Given such variations, it would appear reasonable to consider genus- or species-specific similarities and differences. However, such data are seldom available, and even if a customized pathway structure can be established, its regulation often remains obscure. This shortcoming tends to become evident with new, precise data. For instance, experiments using genetically modified *M. truncatula* lines with reduced CCR1 activity exhibited an unexplainable decrease in the ratio of S to G lignin over wild type [59]. Such discrepancies between expectation and observation suggest that the currently accepted pathway diagrams may require further revisions that include regulatory mechanisms affecting the physiological outcome when the pathway is perturbed.

The focus of this Chapter is an assessment of such a regulatory system associated with lignin biosynthesis in *Medicago*. This genus includes model species like *M. truncatula*, as well as alfalfa (*Medicago sativa* L.), an important forage legume. *Medicago* is particularly suited for these studies, because comparatively extensive

information is available. For instance, as described in Chapter 3, a detailed dataset was established that characterized different lines in which seven lignin biosynthetic enzymes were independently down-regulated, and the resulting lignin content and monomer composition were determined in several stem segments [25]. In a recent study, we demonstrated that these types of data contain substantial, although hidden, information. In particular, we used these data to show that certain enzymes may co-localize and/or assemble into two independent channels for the synthesis of G and S lignin, and that salicylic acid acts as a potential regulatory molecule for the lignin biosynthetic pathway (Chapter 3 and [150]).

Although these earlier results provided significant insights into the mechanisms of regulation in this pathway, several critical questions, especially regarding the biological function as well as the operating mode of the channels, remain unanswered: For instance, are these channels always active *in vivo*? Are they sufficient to explain all available data in *Medicago*? Is there crosstalk between them, and if so, how is it organized? Exploring all pertinent scenarios associated with such questions would be experimentally intractable because they are simply too numerous.

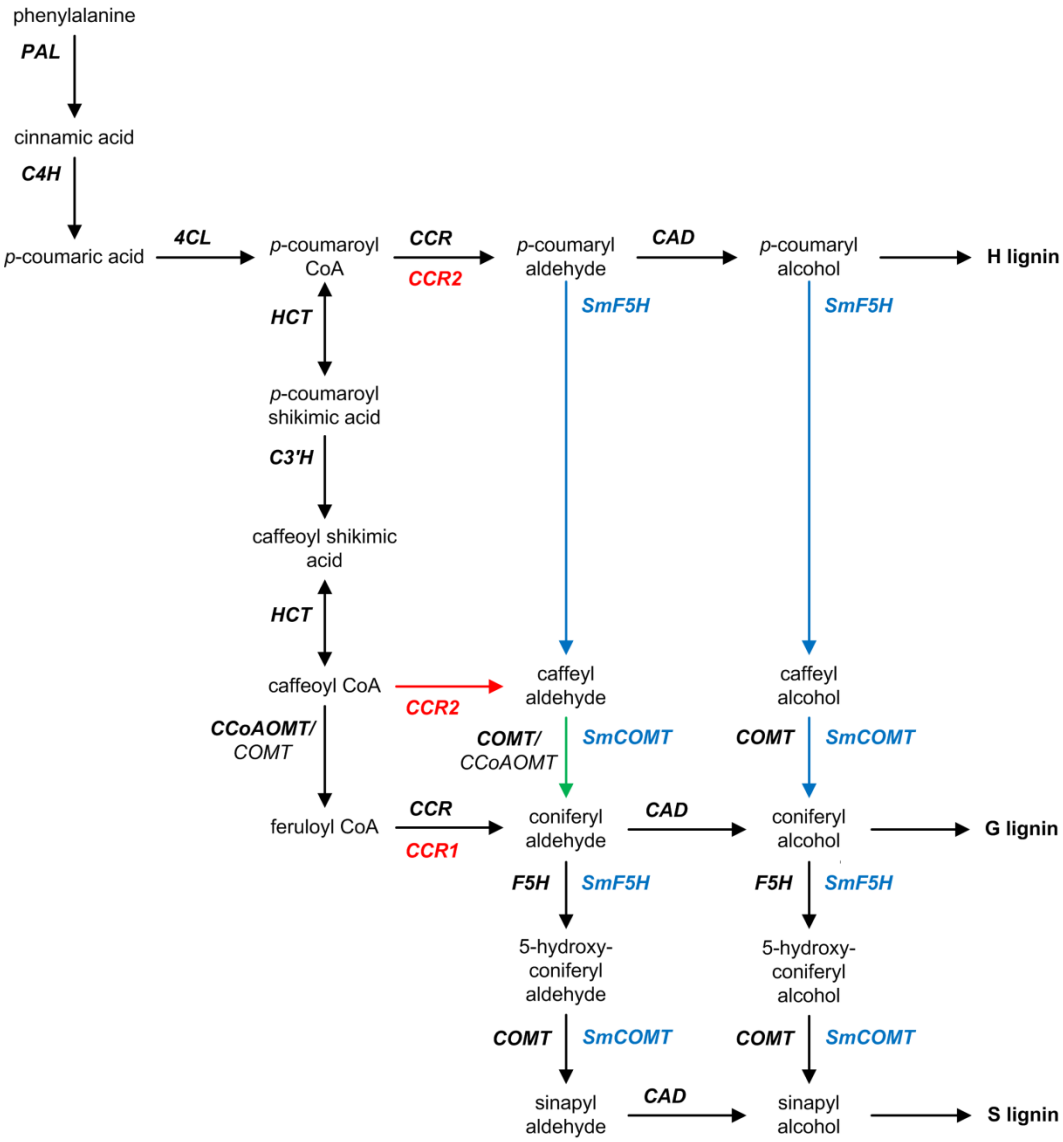


**Figure 5.1: Generic pathway diagram of lignin biosynthesis with species-specific extensions.**

The widely accepted generic pathway consists of black arrows. It leads to the biosynthesis of three hydroxycinnamyl alcohol monomers that in turn give rise to *p*-hydroxyphenyl (H), guaiacyl (G), and syringyl (S) subunits of lignin. Listed next to each reaction arrow are the catalyzing enzymes, which are highlighted in bold if considered major. The lycophyte *Selaginella moellendorffi* contains two bi-functional enzymes, *Sm*F5H and *Sm*COMT, which are shown in blue. Co-expression of these two enzymes would permit *S. moellendorffi* to synthesize coniferyl and sinapyl alcohol directly from *p*-coumaryl aldehyde and *p*-coumaryl alcohol. By contrast, *Medicago truncatula* has two functionally distinct isoforms of CCR, which are shown in red. The green arrow connecting caffeyl aldehyde to coniferyl aldehyde denotes the only non-canonical reaction that is likely to be functional in both *S. moellendorffi* and *M. truncatula*.

Instead, we present here a novel computational approach to investigate exhaustively all regulatory schemes involving the key reactions associated with G and S channels in the lignin biosynthetic pathway (Figure 5.2). The specific hypothesis is that the formerly postulated and validated channels may have two different modes of operation. Either they are *permanent* in a sense that the component enzymes are persistently assembled into a complex; such a complex could be realized through membrane co-localization, thereby ensuring that the corresponding alcohol is always synthesized. As an alternative, the channels could be *facultative*, thereby displaying a functionality that depends on the sub-cellular localization of the component enzymes and the metabolic milieu. This hypothesis, in turn, leads to 19 possible topological configurations (Figure 5.3A). For each of these topologies, we consider an additional level of regulation, involving individual or combined regulatory mechanisms that may serve as a means of "crosstalk" between the two channels (Figure 5.3B). The emphasis of this approach is on mechanisms at the metabolic level, but one must not forget that the transcriptional network governing the system could be involved in the regulation of the pathway as well [191].

The goal is thus to assess and compare the functionality of all given combinations of topological configurations and crosstalk patterns, each of which we call a *design*. To obtain insights that are independent of parameter choices, we constructed for each design a library of 100,000 loosely constrained dynamic models and tested each of them against the observed ratios of S to G lignin in four lignin-modified *Medicago* lines. The resulting analysis of hundreds of designs and millions of models led to the intriguing hypothesis that either a single activation mechanism or a dual-inhibition mechanism lies at the core of all experimentally supported designs. The former mechanism was not supported by an *in vitro* enzyme assay, while the latter is consistent with several lines of evidence from *Medicago* and other species. As an added insight, the analysis suggested that functionality of the G lignin channel is more important than that of the S lignin channel.

Overall, these findings not only enrich our current understanding of how lignin biosynthesis is regulated, but they also demonstrate the possible application of the proposed approach in entirely different biological scenarios where the task is to identify true regulatory circuit among many theoretically feasible designs that depend on the functionality and localization of interacting molecules.

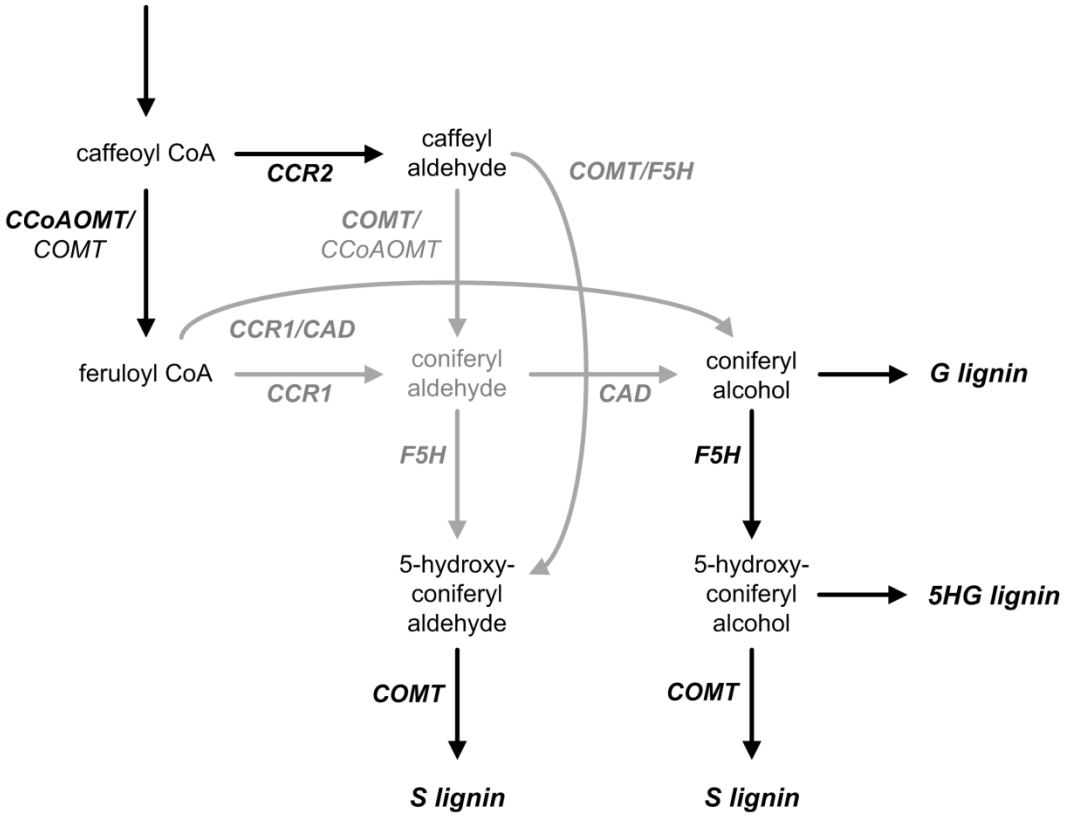


**Figure 5.2: Scaffold of topological configurations.**
The relevant metabolites and enzymatic reactions (arrows) for the biosynthesis of guaiacyl (G), 5-hydroxyguaiacyl (5HG), and syringyl (S) lignin monomers are shown in black, if they are included in all topological configurations, or gray, if they are included in only some specific configurations. Notice that 5-hydroxyconiferyl alcohol is allowed to be incorporated into lignin polymer as 5HG subunit because this actually occurs when COMT is down-regulated [103]. Enzymes and lignin monomers are highlighted in bold and italics.

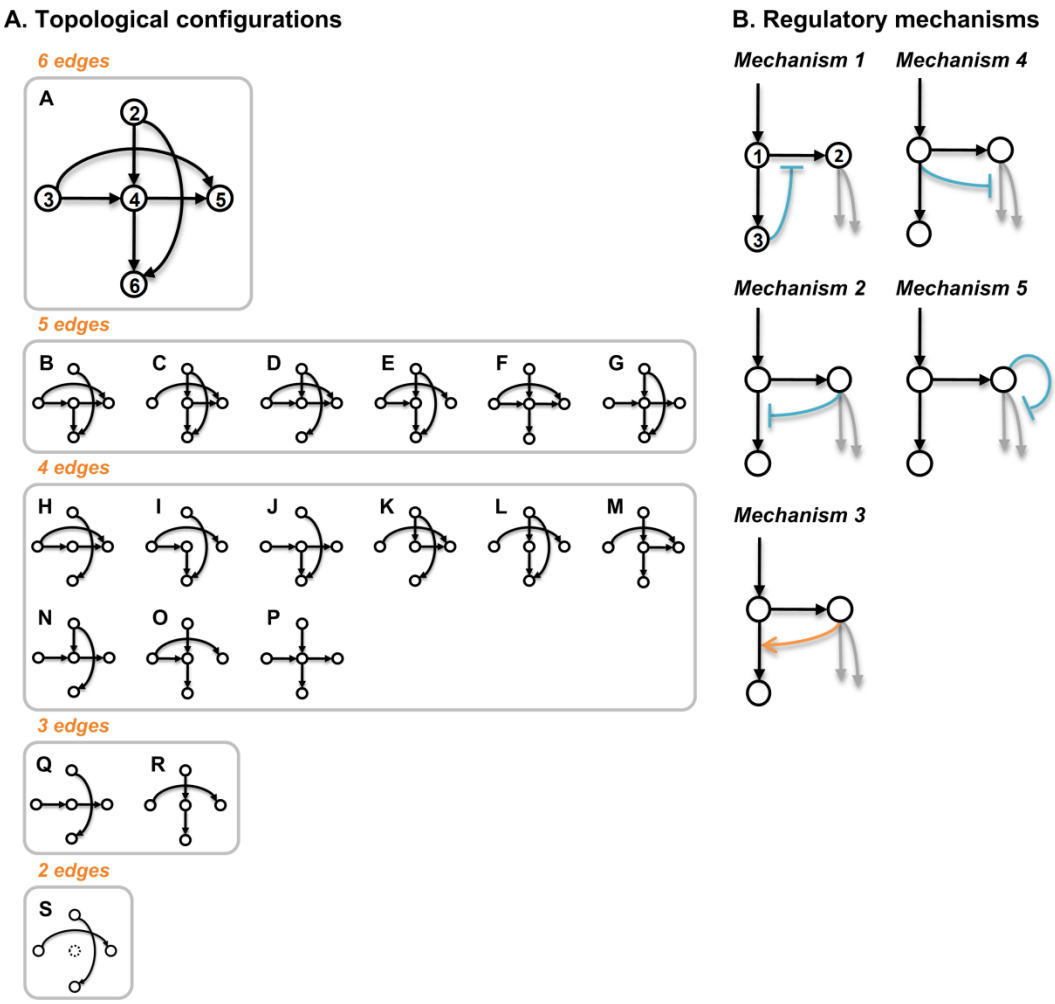


**Figure 5.3: Lists of topological configurations and regulatory mechanisms.**
Panel A**:** The topological configurations differ in their numbers of edges. Panel B: The orange arrows represent activation processes, whereas the blocked lines (aqua) represent inhibition processes. Arrows colored in gray are reactions included in only some specific topological configurations. Metabolite names: ① caffeoyl CoA; ②, caffeyl aldehyde; ③ feruloyl CoA; ④ coniferyl aldehyde; ⑤ coniferyl alcohol; ⑥ 5-hydroxyconiferyl aldehyde.

## 5.2 Results

### 5.2.1 Enumeration of Circuit Designs

The base scaffold on which the different topological variants were built is shown in Figure 5.2. It consists of all relevant steps in the lignin biosynthetic pathway that possibly affect the relative amounts of G and S lignin. The G and S lignin channels are represented as directed edges linking feruloyl CoA and coniferyl alcohol, or linking caffeyl aldehyde and 5-hydroxyconiferyl aldehyde, respectively. The experimentally validated channeling hypothesis [150] permits 19 different topological configurations (Figure 5.3A) that satisfy the following constraints. First, at least one edge must be leaving caffeyl aldehyde and feruloyl CoA, and at least one edge must be entering coniferyl alcohol and 5-hydroxyconiferyl aldehyde; otherwise mass would unduly accumulate in intermediate pools. Second, if coniferyl aldehyde can be produced by a free CCR1 and/or caffeic acid *O*-methyltransferase (COMT), it must also be consumed by a free enzyme, thereby decreasing the metabolic burden that would otherwise be imposed on the cell. For reasons that will be explained below, we also consider for each topological configuration various crosstalk patterns between the CCR2/COMT and CCoAOMT/CCR1 pathways. Each pattern is composed of documented or postulated mechanisms of metabolic regulation (activation or inhibition) (Figure 5.3B). The specific combinations of topological configurations and crosstalk patterns lead to hundreds of different designs, which were analyzed and compared.

For each design, we first constructed 100,000 Generalized Mass Action (GMA) models (see Chapter 1 for definition) by randomly sampling loosely-constrained parameter combinations from a parameter space that was deemed biologically realistic. A notable feature of this approach was that the parameter space was not only constrained at the level of individual parameters (*e.g.*, kinetic orders), but also at the level of steady-

state fluxes. For instance, the ratio of fluxes leading to S and G lignin was fixed at a value observed in the wild-type *Medicago* species (see Sections 5.4 and C.1 for details). Once all parameters for a given GMA model instantiation were specified, we determined steady-state fluxes under conditions that mimic CCoAOMT and COMT down-regulated alfalfa lines as well as *ccr1* and *ccr2 M. truncatula* mutant lines and computed the S/G ratios for which we had experimental data. We declared a model as valid if it yielded quantitatively and qualitatively correct results for both transgenic alfalfa and *M. truncatula* plants (see Materials and Methods). To assess the robustness of a design to parametric perturbations, we defined *Q* as the total number of valid model instantiations.

### 5.2.2 Channels Are Necessary but Not Sufficient

As a reasonable baseline, we first assumed the absence of crosstalk between the CCR2/COMT and CCoAOMT/CCR1 pathways (Figure 5.4). Of all possible topological configurations lacking crosstalk, only six had at least one parameter combination that yielded quantitatively correct predictions of S/G ratios for CCoAOMT and COMT down-regulated alfalfa plants. Supporting our previous findings [150], all six configurations include either one or both channels, suggesting that the channels are necessary. In other words, the pathway models are consistent with the observed changes in the S/G ratios of CCoAOMT and COMT down-regulated alfalfa plants only if at least one channel is present.

To assess these initially feasible parameter combinations further, we used the models with these parameter values to predict the S/G ratios for *ccr1* and *ccr2* knockout mutants. The *M. truncatula* lines harboring transposon insertions in *CCR1* and *CCR2* show a corresponding reduction in CCR1 and CCR2 activity, and their S/G ratio is decreased or increased, respectively, compared to the wild-type level [59]. Moreover, the activities of CCR1 and CCoAOMT, as well as their mRNA transcripts and proteins, are

increased in the *ccr2* knockout mutant, indicating that part of their activation might be processed through a hierarchical control of gene expression [192].

Figure 5.4 shows simulation results for those topological configurations where at least one out of 100,000 randomly parameterized models yielded quantitatively correct predictions of S/G ratios for both CCoAOMT and COMT down-regulated alfalfa plants. In these plots, a model is valid only if its predicted S/G ratios for *ccr1* and *ccr2* knockout mutants fall within the northwest quadrant.

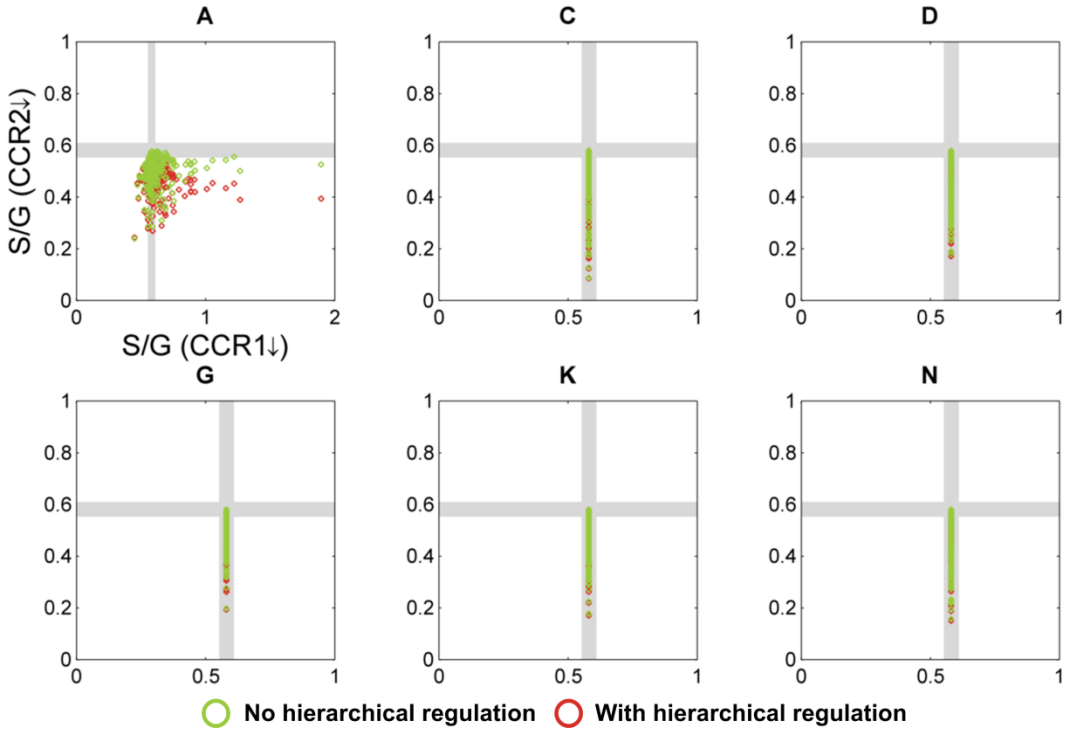


**Figure 5.4: Simulation results for pathway designs without crosstalk.**
Each of the 6 panels corresponds to one topological configuration with at least one randomly parameterized S-system model that yields quantitatively correct predictions of S/G ratios for both CCoAOMT and COMT down-regulated alfalfa plants. Each open circle refers to the S/G ratio of CCR1 and CCR2 in a *M. truncatula* knockout mutant, as predicted by one randomly parameterized S-system model; its color indicates the type of regulation. The gray strips denote regions within 5% of the wild-type level; model predictions within these strips are considered essentially the same as wild-type. Qualitatively correct predictions should fall into the northwest quadrant. It is evident that not a single model instantiation is admissible. The total number of randomly parameterized model instantiations per panel was $10^5$.

In the case of no hierarchical regulation, *i.e.*, the *ccr2* mutant exhibits only reduced CCR2 activity, some model instantiations from configuration A showed a decreased S/G ratio for the *ccr1* knockout mutant, but not a single case exhibited an increased S/G ratio for the *ccr2* knockout mutant. This outcome did not improve much when hierarchical regulation was considered: not one of the 1.9 million model instantiations from the 19 possible configurations yielded qualitatively acceptable predictions for both *ccr1* and *ccr2* knockout mutants. These findings indicate that the S and G channels alone are not sufficient to explain all available transgenic data, and that some type of crosstalk is highly likely to occur between the CCR2/COMT and CCoAOMT/CCR1 pathways.

### 5.2.3 Crosstalk between the CCR2/COMT and CCoAOMT/CCR1 Pathways

One potential source of crosstalk between the CCR2/COMT and CCoAOMT/CCR1 pathways is substrate competition. CCR1/2 converts hydroxycinnamoyl CoA esters to their corresponding cinnamyl aldehydes, whereas CCoAOMT and COMT together complete the methylation of the aromatic $C_3$ and $C_5$ positions of the aldehydes and alcohols (Figure 5.1). All these enzymes are known to be multi-functional, acting upon multiple substrates with distinct catalytic efficiency. Because of their promiscuous nature, different substrates compete with each other if the supply of enzyme is limited. As a consequence, the enzymatic conversion of one substrate is effectively subjected to competitive inhibition by another substrate, and *vice versa*. This type of cross-inhibition is not necessarily equally strong in both directions because a promiscuous enzyme often displays preference for some substrates over others.

In the case of lignin biosynthesis, two regulatory mechanisms could arise from substrate competition. First, recombinant *Medicago* CCR2 exhibits similar $k_{cat}/K_M$ values for caffeoyl CoA (0.49 $\mu M^{-1} \cdot min^{-1}$) and feruloyl CoA (0.40 $\mu M^{-1} \cdot min^{-1}$) [59], suggesting

that the CCR2-mediated conversion of caffeoyl CoA to caffeyl aldehyde in *Medicago* might be competitively inhibited by feruloyl CoA (Figure 5.3B; Mechanism 1). Furthermore, CCR2 is inhibited by feruloyl CoA at a concentration above 20 μM [59]. Conversely, it is highly unlikely that the CCR1-mediated conversion of feruloyl CoA to coniferyl aldehyde is significantly affected by caffeoyl CoA, because CCR1 has a $k_{cat}/K_M$ value for caffeoyl CoA (0.019 $\mu M^{-1} \cdot min^{-1}$) that is 60 times lower than that for feruloyl CoA (1.14 $\mu M^{-1} \cdot min^{-1}$) [59].

Second, the methylation of caffeoyl CoA by the combined activity of COMT and CCoAOMT may be subject to weak competitive inhibition by caffeyl aldehyde (Figure 5.3B; Mechanism 2). This assumption is based on the following observation. Although the combined *O*-methyltransferase (OMT) activity against caffeoyl CoA in extracts from internodes 6 to 8 of CCoAOMT-down-regulated alfalfa was reduced by 4.2-fold compared with the empty vector control line, about ~25% of OMT activity remained [60]. This activity is presumably associated with COMT, for which caffeyl aldehyde is the preferred substrate. Notably, both mechanisms are independent of each other and may work individually or collaboratively to establish crosstalk between the two channels, thereby leading to three different crosstalk patterns and 57 different designs.

In the case where only Mechanism 1 (Figure 5.3B) was incorporated in the design, we observed a substantial increase in the number of model instantiations showing a decreased S/G ratio for the *ccr1* knockout mutant (Figure C.1). Yet, even when we accounted for the effect of hierarchical regulation, none of the models was capable of delivering a qualitatively correct change in the S/G ratio for the *ccr2* knockout mutant. This finding indicates that the experimentally inferred inhibition evidently exists but is not sufficient. Similarly, we found no valid models when Mechanism 2, either by itself or coupled with Mechanism 1, was employed (Figures C.2 and C.3). An explanation may be that, with caffeyl aldehyde inhibiting the 3-*O*-methylation of caffeoyl CoA, knocking

down CCR2 activity will consistently lead to a deregulation of CCoAOMT by caffeyl aldehyde, thereby increasing the flux to G lignin and reducing the S/G ratio.

### 5.2.4 Is Caffeyl Aldehyde an Activator of CCoAOMT?

One could surmise that the 3-*O*-methylation of caffeoyl CoA, for which CCoAOMT is the primary enzyme, is actually *activated* by caffeyl aldehyde. This conjecture is based on the following argument. When the production of S lignin is compromised due to a knockout of *ccr2*, the only way of raising the S/G ratio beyond its wild-type level appears to be a further reduction of the flux through the CCoAOMT/CCR1 pathway, which can be accomplished if CCoAOMT is activated by caffeyl aldehyde. The simulation results using this type of postulated mechanism, either by itself (Figure 5.5) or coupled with the documented inhibition of CCR2 by feruloyl CoA (Figure 5.6), are very intriguing: For each crosstalk pattern where millions of randomly parameterized models were generated, we found thousands of valid instantiations that yielded quantitatively and qualitatively correct predictions for both transgenic alfalfa and *M. truncatula* plants. Perhaps more surprisingly, only six topological configurations (A, B, E, F, I, O) had at least one valid model ($Q > 0$; see Section 5.4). To ensure that this result was not due to the use of overly restrictive thresholds, we relaxed the criteria and found more parameter combinations that qualified. Nevertheless, the same six topological configurations always passed the screening test by a wide margin (Table C.1). Collectively, these findings suggested that this activation mechanism, acting alone or with the inhibition of CCR2 by feruloyl CoA, is necessary for consistency with the *ccr1* and *ccr2* knockout data. This conclusion immediately translated into a targeted hypothesis that was independent of specific parameter choices and readily testable by experiment.

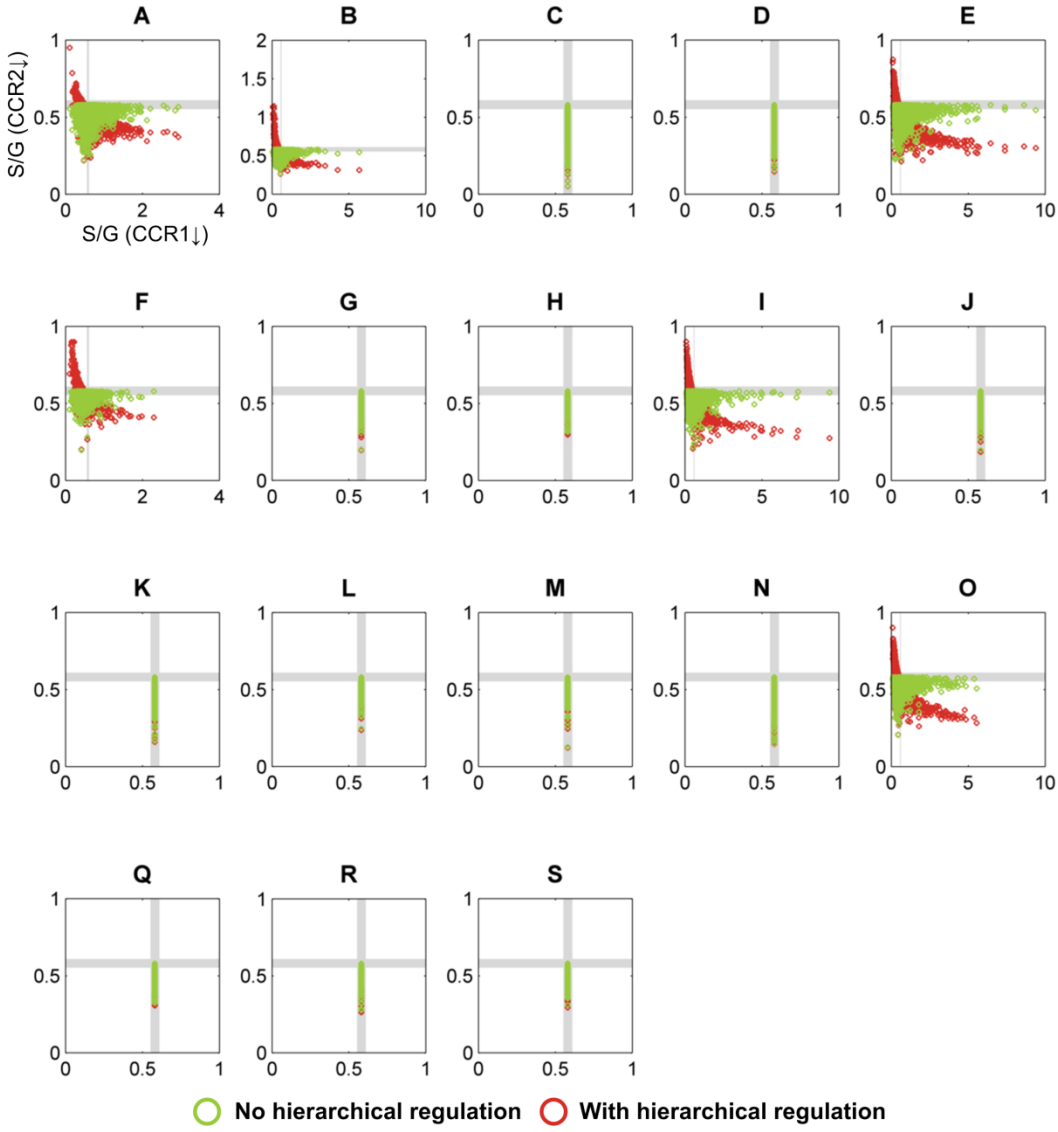


**Figure 5.5: Simulation results for pathway designs using only Mechanism 3.**
See Figure 5.3B for the structure of this mechanism and the legend of Figure 5.4 for more information on details shown. In contrast to the results in Figure 5.4, the pathway designs analyzed here permit numerous admissible model instantiations (topologies A, B, E, F, I, and O), which fall into the northwest quadrant. The total number of randomly parameter zed model instantiations per panel was $10^5$.

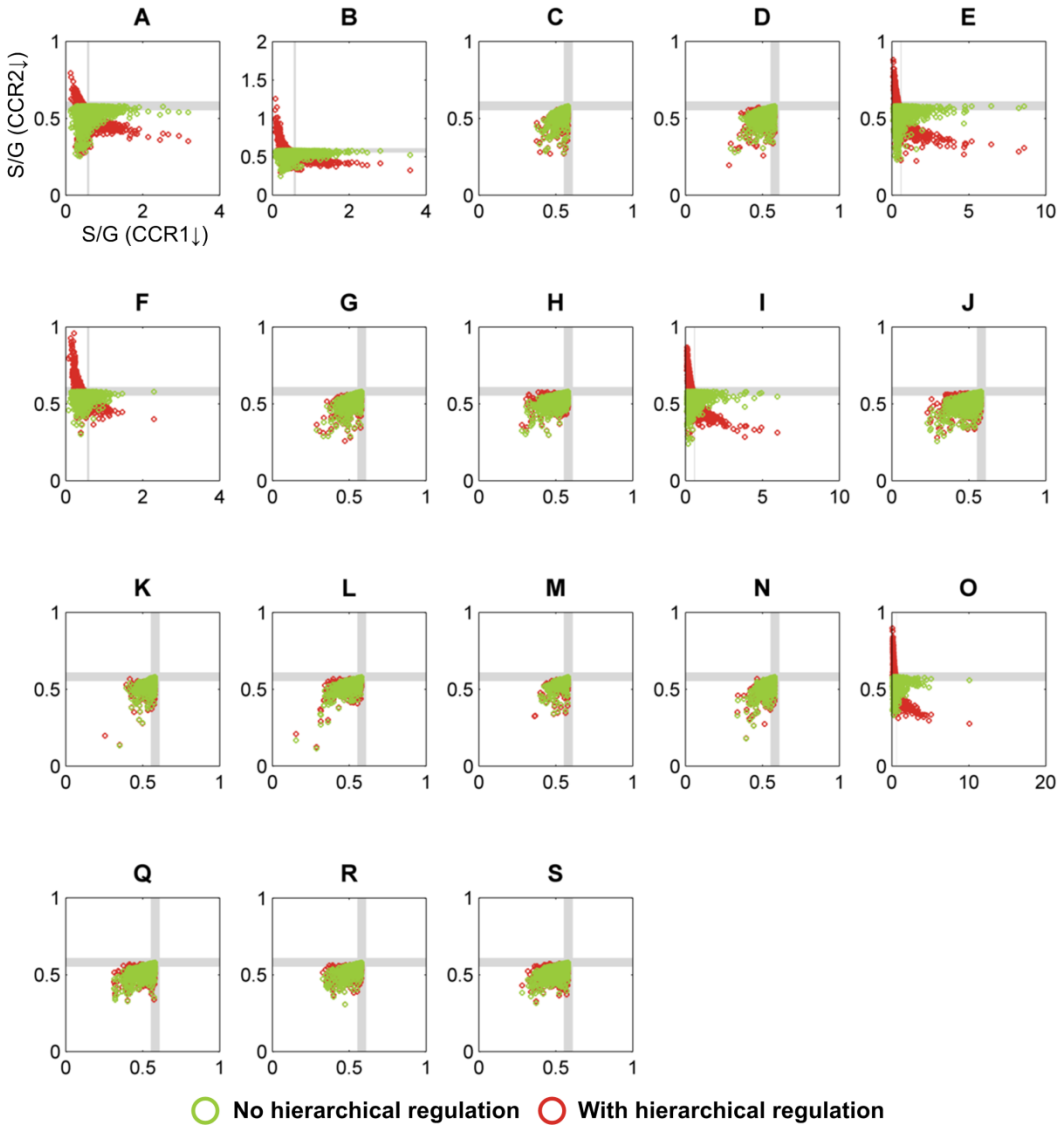


**Figure 5.6: Simulation results for pathway designs that contain Mechanisms 1 and 3 simultaneously.**
See Figure 5.3B for the structure of these mechanisms and the legend of Figure 5.4 for more information on details shown. Similar to the results in Figure 5.5, the pathway designs analyzed here permit numerous admissible model instantiations (topologies A, B, E, F, I, and O), which fall into the northwest quadrant. The total number of randomly parameterized model instantiations per panel was $10^5$.

### 5.2.5 The Hypothesized Activation Is Not Supported by Experimental Data

To examine whether caffeyl aldehyde indeed activates CCoAOMT, our collaborators expressed alfalfa CCoAOMT in *Escherichia coli* and assayed the purified recombinant enzyme with caffeoyl CoA as substrate and caffeyl aldehyde as the putative activator. As shown in Figure 5.7, the CCoAOMT activity increased by 16% at 2 μM of caffeyl aldehyde and 20 μM of caffeoyl CoA; at higher substrate concentrations (*i.e.*, 30 and 40 μM of caffeoyl CoA), the increase in mean CCoAOMT activity became less. Assays using lower concentrations of the substrate caffeoyl CoA (2, 4, 5 and 10 μM) and the putative activator caffeyl aldehyde (0.5, 1, 2 and 4 μM) showed no increase in CCoAOMT activity compared to the reaction without caffeyl aldehyde (data no shown). The maximal activation achieved *in vitro* was only 16 %, which was statistically significant but may not be biologically significant.

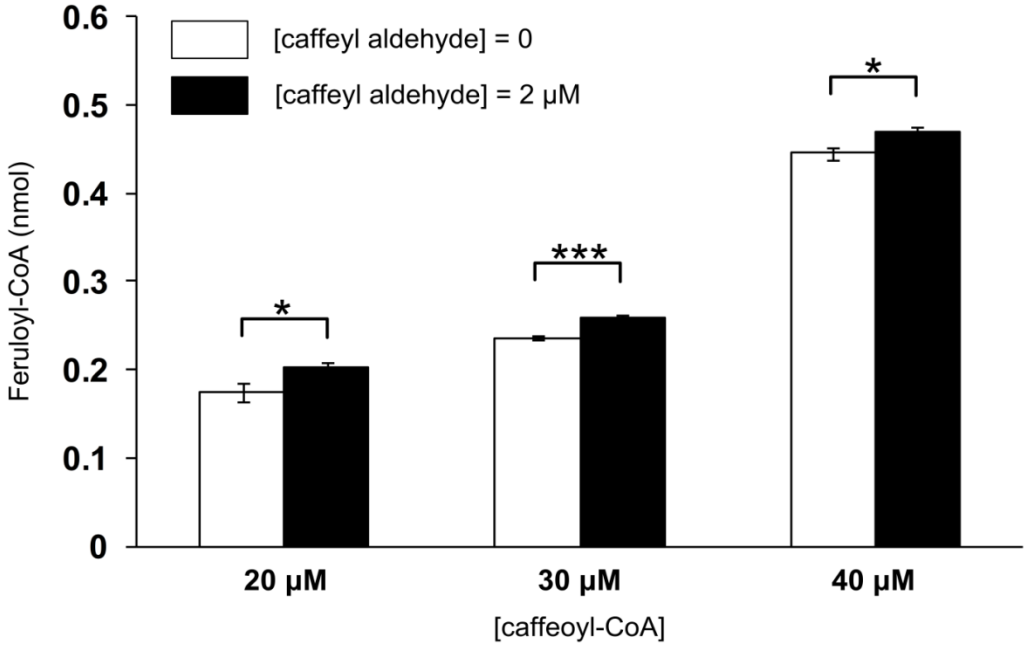


**Figure 5.7: 2 μM caffeyl aldehyde activates CCoAOMT-mediated methylation of caffeoyl CoA *in vitro*.**
Error bars, mean ± s.d.; ***$p < 0.001$, *$p < 0.05$ by Student's *t*-test; $n = 3$.

### 5.2.6 Analysis of Caffeyl Aldehyde as a Dual Inhibitor of Two 3-*O*-Methylation Reactions

Since a direct activation of CCoAOMT by caffeyl aldehyde was not observed in experiments with recombinant enzymes, we tested other regulatory mechanisms by themselves and in combination with known mechanisms. According to one possible mechanism, based again on the concept of substrate competition, caffeoyl CoA could be a competitive inhibitor for the 3-*O*-methylation of caffeyl aldehyde (Figure 5.3B; Mechanism 4). This proposal agrees with the fact that CCoAOMT may contribute up to ~10% of the methylation reaction in alfalfa [60]. In addition, evidence in ryegrass (*Lolium perenne*) points to the possibility of COMT being inhibited by different substrates, such as caffeyl aldehyde and 5-hydroxyconiferyl aldehyde [193]. Interestingly, substrate inhibition by caffeyl alcohol and 5-hydroxyconiferyl alcohol has also been observed in *Selagniella moellendorffii* COMT [189]. Thus, we hypothesized that COMT might be inhibited by caffeyl aldehyde (Figure 5.3B; Mechanism 5) in *Medicago* as well; direct evidence supporting this hypothesis in *Medicago* remains to be determined.

In total, there are $2^4 = 16$ different crosstalk patterns that can result from the combination of four independent regulatory mechanisms (Figure 5.3B; Mechanisms 1, 2, 4 and 5). However, only four of them, when combined with the same six topological configurations (A, B, E, F, I and O) that were identified previously (*cf.* Figures 5.5 and 5.6), gave rise to designs with at least one valid model instantiation (Figure 5.8). Interestingly, all these crosstalk patterns require that caffeyl aldehyde is an inhibitor of the 3-*O*-methylation of both caffeoyl CoA and itself (Figure 5.3B; Mechanisms 2 and 5), providing computational evidence that this synergy between the two seemingly unrelated mechanisms is necessary for consistency with the *ccr1* and *ccr2* knockout data. Indeed, with respect to the *ccr2* knockout, such a combination of two inhibition mechanisms

appears to have a similar ultimate effect as a single activation mechanism (see Section 5.3).

Inspecting the crosstalk patterns giving rise to at least one design with valid model instantiations (rows colored in red in Figure 5.8), one might surmise that caffeyl aldehyde would accumulate to an unduly high level, because Mechanism 5, which is employed in all these patterns, reflects substrate inhibition of COMT by caffeyl aldehyde. To examine the validity of this inference, we checked, for all designs with valid model instantiations, the predicted changes in caffeyl aldehyde under conditions that mimic the down-regulation of four lignin biosynthetic enzymes. As shown in Figure 5.9, it appears that down-regulation of CCoAOMT or COMT is consistently associated with a lower caffeyl aldehyde level compared with wild type, regardless of the crosstalk pattern being considered. Similarly, knocking out *ccr2* consistently raises the caffeyl aldehyde level in all crosstalk patterns examined. However, in the case of the *ccr1* knockout mutant, the results are mixed in a sense that some crosstalk patterns are associated with significantly higher caffeyl aldehyde levels, whereas others are associated with only modest changes. Interestingly, both crosstalk patterns suffering from an undue accumulation of caffeyl aldehyde contain Mechanism 1. By contrast, this mechanism is absent from other patterns, which maintain a relatively stable caffeyl aldehyde level. This finding suggests that the control pattern in Mechanism 1 may disrupt the metabolic homeostasis via accumulation of caffeyl aldehyde when CCR1 drops below its normal level. As any cellular system is constantly afflicted by a variety of intrinsic and extrinsic noises, this type of fluctuation must be expected to occur frequently and spontaneously, suggesting that Mechanism 1 is disadvantageous.

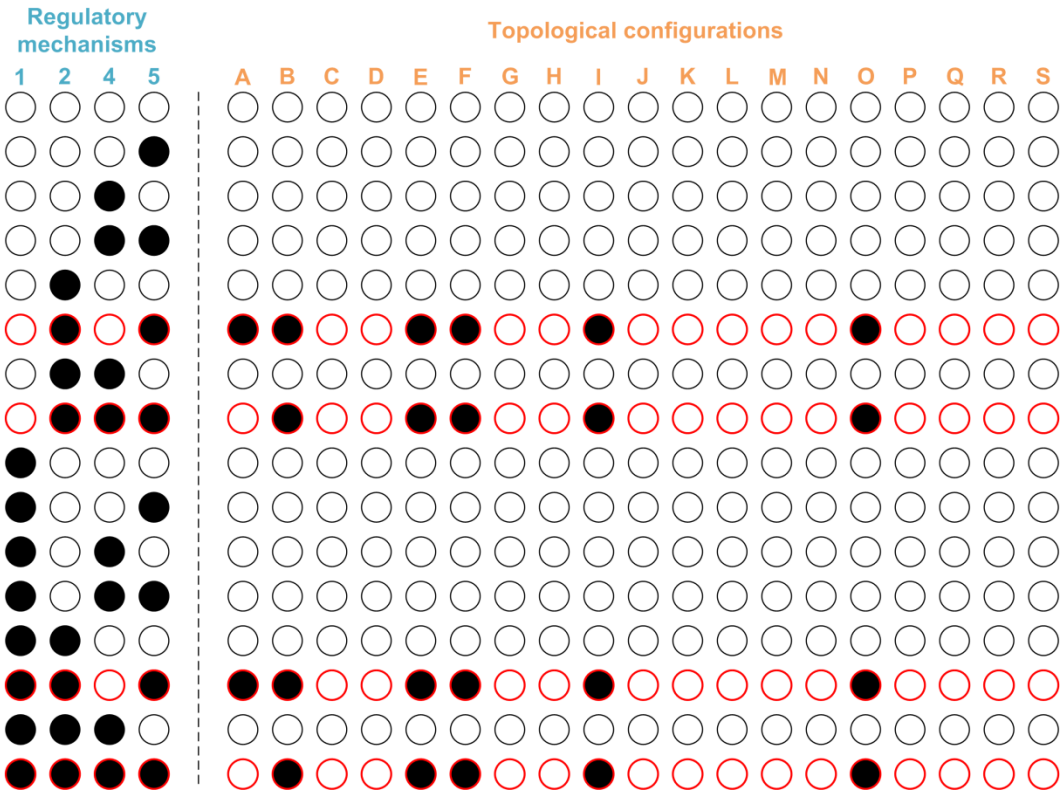


**Figure 5.8: Summary of simulation results from 304 designs.**
Each row corresponds to one crosstalk pattern, whereas each column in the table to the right of dashed line corresponds to one topological configuration. A design is represented by a filled circle if at least one of the $10^5$ randomly parameterized model instantiations is valid. Empty circles thus refer to designs that are incongruent with observations. Rows highlighted in red contain at least one topological configuration with valid model instantiations.

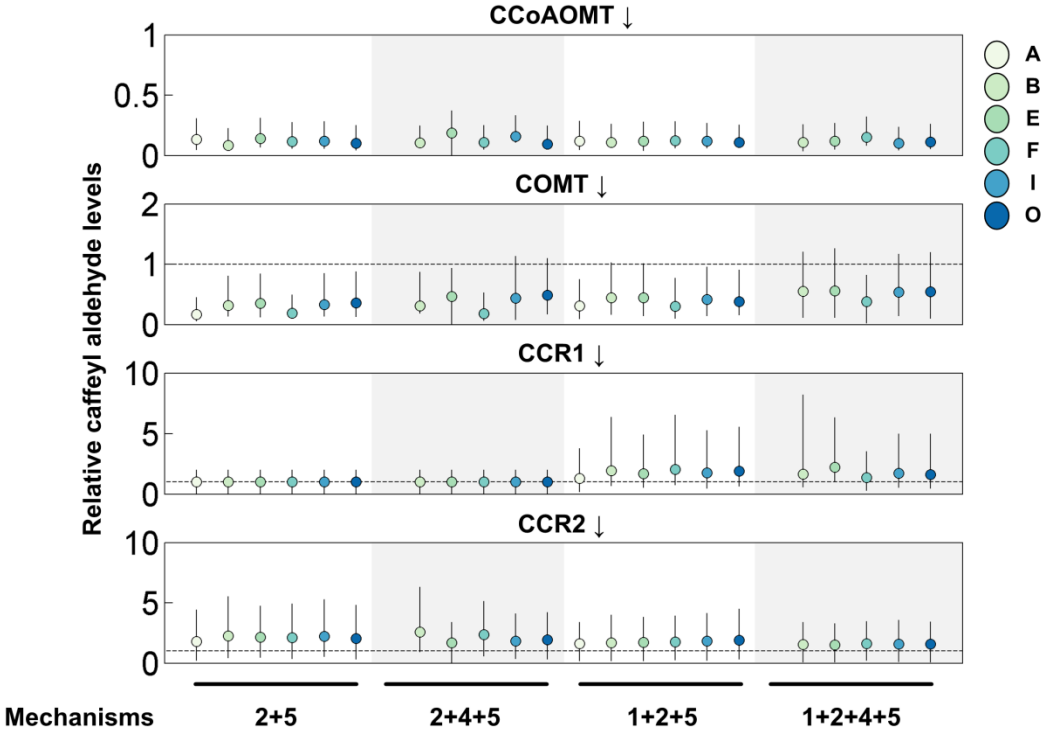


**Figure 5.9: Relative levels of caffeyl aldehyde (compared to wild-type values) in simulations of four down-regulated lines.**
Each panel is shaded to highlight the results from four different crosstalk patterns. These patterns, when combined with specific topological configurations, give rise to designs with valid model instantiations (*cf.* rows with red circles in Figure 5.8). In contrast to the other three perturbation schemes, where all four crosstalk patterns (and their corresponding designs) exhibit similar responses regarding the caffeyl aldehyde level, knocking out *ccr1* is associated with a higher caffeyl aldehyde level only for the two crosstalk patterns including Mechanism 1. The circles, colored according to topological configuration, are the medians, and the error bars represent interquartile ranges. The dashed line in each panel, if present, denotes the wild-type level of caffeyl aldehyde.

### 5.2.7 Robust Designs Are Evolutionarily Connected

Investigation of the six robust topological configurations, which contain at least one valid model instantiation, revealed interesting structural features of the pathway. In particular, the G lignin channel is common to all robust designs and thus may be considered critical for the proper functioning of the pathway, at least for the cases studied. The evolutionary conservation of such a feature, one may argue further, is not due to the fact that it cannot possibly be altered, but that this particular design can sustain maximally tolerable changes and variability in other features [194]. These arguments lead to an interesting follow-up question, namely: Are the robust topological configurations related in an evolutionary sense?

To address this question, we constructed a "topology graph" where each node corresponds to a topological configuration. Two nodes are connected if the corresponding topological configurations differ only by one edge. For instance, configurations A and B are directly linked to each other because the only difference between them is whether caffeyl aldehyde can be converted, via free COMT, to coniferyl aldehyde. In other words, moving from a node to its neighbor may be considered a singular evolutionary event where an enzyme's preferred mode of action is changed.

Two outcomes are possible for the structure of such a topology graph. First, the graph may be disconnected, that is, there exist pairs of topological configurations such that no evolutionary path (defined as a series of evolutionary events) connects one to the other. In the most extreme case, the graph would consist exclusively of isolated nodes. Second, the graph is fully connected, so that any pair of topological configurations is connected by at least one evolutionary path. As shown in Figure 5.10, the actual topology graph of the six robust configurations of lignin biosynthesis is indeed connected, and so is the graph of all configurations, except for design S. This interconnectedness can be interpreted as facilitating the evolvability of the system [194], because the gain or loss of

specific features that are needed to produce phenotypically novel traits will be tolerated and survive during evolution if robustness is preserved. Of course, this evolutionary aspect, which was derived purely with computational means, will require additional analysis.

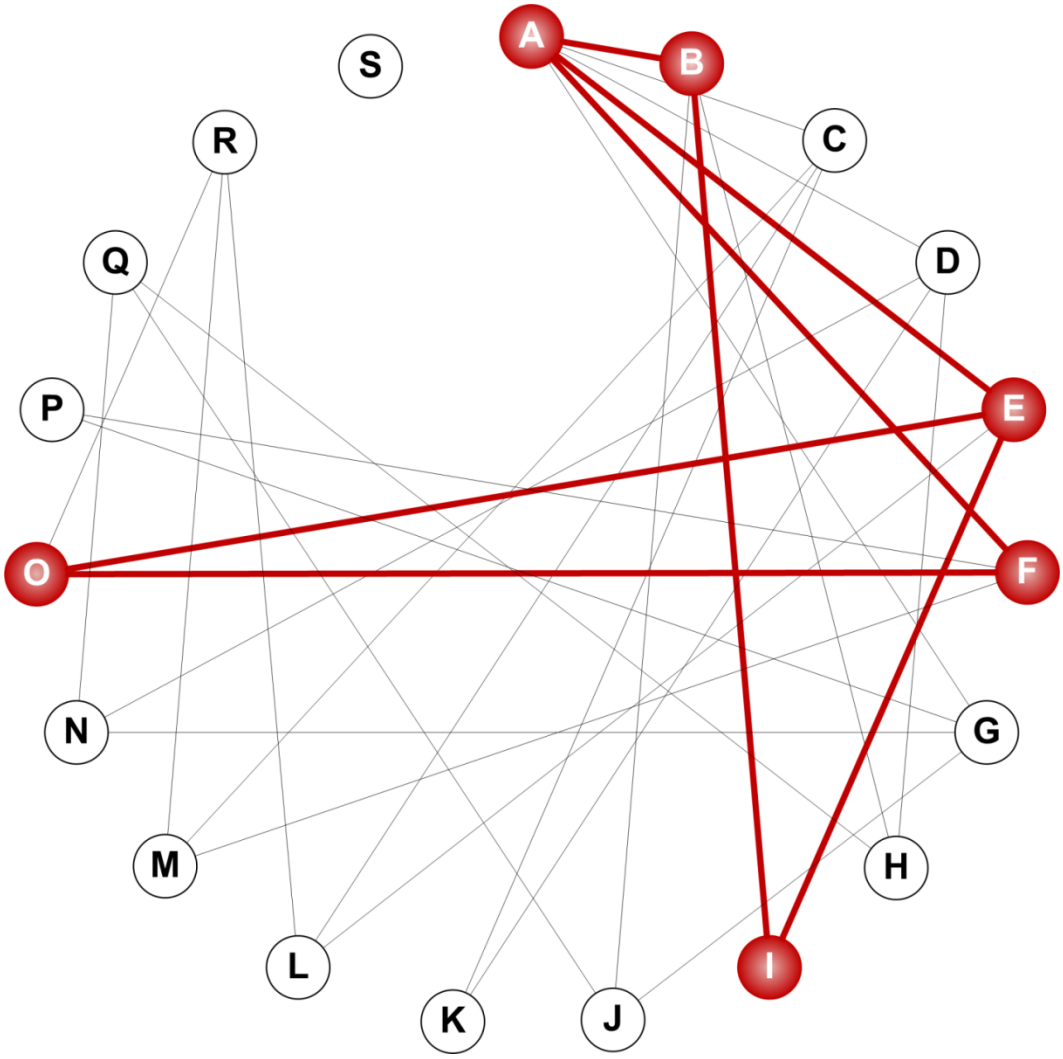


**Figure 5.10: Robust configurations are evolutionarily connected.**
Each node represents a specific topological configuration (see Figure 5.3); two nodes are connected if the corresponding configurations differ only by one edge. The subgraph of all the robust configurations, colored in red, is connected, thereby indicating the potential of direct evolvability.

## 5.3 Discussion

The spatial organization of cooperating enzymes, known as metabolic channeling, has long been recognized as an effective means of regulation in primary and secondary plant metabolism [12,24,195]. This channeling phenomenon involves the organization of enzymes into complexes and/or the co-localization of enzymes at the plasma membrane or on the surfaces of organelles, as was demonstrated for the two initial enzymes, L-phenylalanine ammonia-lyase (PAL) and cinnamate 4-hydroxylase (C4H), in the phenylpropanoid pathway [11,196]. Interestingly, some complexes or interactions are persistent, while others are temporary. In fact, many of the component enzymes such as PAL may be operationally soluble and are therefore only facultatively channeled. Such short-lived or dynamic complexes, while being readily responsive to the metabolic status of the cell, are inherently difficult to study with existing or emerging experimental models.

Using the lignin biosynthetic pathway as a model system, we propose here a novel strategy for studying metabolic channeling in unprecedented detail. Specifically, we consider all possible modes of action for both the G lignin and S lignin channels, and these can be mapped into 19 different topological configurations (Figure 5.3A). Metabolic channeling is clearly not the only process that affects the functionality of this system, and it is therefore necessary to study control processes affecting a channeled system. In the present case, this control is potentially exerted by individual or combined mechanisms of crosstalk between the CCR2/COMT and CCoAOMT/CCR1 pathways (Figure 5.3B). Some of these were documented in the literature, while others were hypothesized. Taken together, a topological configuration and a specific crosstalk pattern constitute a design. We evaluated each design with or without consideration of non-allosteric or hierarchical regulation which could involve transcription, as well as a variety of non-transcriptional processes such as phosphorylation, methylation, and targeted degradation of proteins and mRNA.

Ideally, the comparative assessment of design features would be entirely symbolic and independent of specific parameter values. However, systems of a realistic size are rarely analyzable in such fashion. As a reasonable alternative, we analyzed the possible design space comprehensively with widely varying parameter values, which resulted in a computational analysis of millions of models from hundreds of designs. This analysis yielded several interesting findings.

Importantly, it predicted that CCoAOMT is directly or indirectly activated by caffeyl aldehyde. This piece of information by itself is essentially unbiased, but insufficient to explain the exact mechanism of regulation. Nevertheless, it offered a specifically targeted hypothesis and was therefore experimentally testable. However, the hypothesis of a direct activation was refuted by subsequent experiments using the recombinant *Medicago* CCoAOMT, which failed to provide evidence confirming the putative role of caffeyl aldehyde as an allosteric activator. It might still be possible that activation exists *in vivo*, but it seems more likely that the activation is indirect rather than direct.

As a possible mechanism, the design analysis suggested that caffeyl aldehyde inhibits the 3-*O*-methylation of both caffeoyl CoA and itself. Several lines of evidence, although not exclusively from *Medicago*, support this computational prediction. Most importantly, the same six topological configurations were identified in the indirect design analysis and in the initial analysis of a putative activation mechanism. However, the two most parsimonious mechanisms differ in their proposed control strategies. The original analysis suggested just one activation mechanism, while the second analysis proposed two inhibition mechanisms. To some degree, these two mechanisms have the same ultimate effect. If *ccr2* is knocked out, the flux entering the CCR2/COMT pathway and the subsequent synthesis of S lignin decline. The only possibility to increase the S/G ratio is to reduce the flux entering the CCoAOMT/CCR1 pathway. This task can be accomplished either through a diminished activation, as suggested for the single

activation mechanism, or through an enhanced inhibition, as suggested for the dual-inhibition mechanism. The latter mechanism seems sufficient to restore consistency with the data, but it is of course possible that more complicated control patterns are present.

The computational analysis suggests that the G lignin channel is necessary for the system to respond correctly and robustly to certain genetic perturbations. By contrast, the S lignin channel appears to be dispensable. This theoretical deduction is indirectly in line with the fact that S lignin has arisen much later in the evolution of higher plants than G lignin [15]. It is also consistent with the observation that its formation, which in many plant species is dictated by ferulate 5-hydroxylase (F5H) expression [197,198,199], is directly regulated by a secondary cell wall master switch NST1/SND1 and not by MYB58, a SND1-regulated transcription factor that can activate other lignin biosynthetic genes [200]. It could also be possible that S lignin, which is specifically involved in the pathogen defense of some plants [201], was relatively recently recruited for lignin biosynthesis and thus may not be essential for plant growth. Evidence supporting this postulate includes an *Arabidopsis* NST1/SND1 double knockout mutant that shows a complete suppression of secondary cell wall thickening in woody tissues, including interfascicular fibers and secondary xylem, but otherwise grows quite well as compared to the wild-type plants [202].

Within an evolutionary context, the multiplicity of robust solutions can be represented with a graph representation that connects any two (robust) topological configurations differing by a single edge. This graph is reminiscent of the "neutral network" concept that was initially proposed in genotype-phenotype models for RNA secondary structures [203] and protein folds [204], but also more recently extended to Boolean models for gene regulatory networks [205]. In the case of proteins, neutral networks are defined as sets of amino acid sequences that are connected by single-mutation neighbors and that map into the same tertiary structure. Such degeneracy of the mapping from genotype to phenotype allows a neutral drift in genotypic space, which is

critical for accessing adjacent neutral networks with novel phenotypes that may confer higher fitness to the cells. As of yet, it is unclear whether individual plants within a *Medicago* population use the same or different designs, or whether the response to selected perturbations is an adequate phenotypic feature. Further investigation of the protein-protein interactions between lignin biosynthetic enzymes is thus necessary to confirm that a G lignin channel is indeed necessary for optimal functioning.

The work in this Chapter describes a novel computational approach that shows promise in deciphering the principles of channel assembly in a biosynthetic pathway when relevant information is limited. It also provides a clear direction in which to proceed with more targeted experiments. Beyond the application described here, the proposed strategy might be beneficial in entirely different biological contexts, such as gene regulatory and signaling networks, where the task is to analyze how information flow is controlled by the spatial organization of molecules in the cell.

## 5.4 Materials and Methods

### 5.4.1 Model Equations in GMA Format

Since the two metabolic channels of interest are assumed to affect only the relative amounts of G and S lignin, the analysis is restricted to those critical steps within the lignin biosynthetic pathway system that govern the flow of material either toward G or S (Figure 5.2). For each possible design, we first formulate the corresponding generalized mass action (GMA) model [43,44] in a symbolic format (Eq. (1.3)). The model contains either six or seven dependent variables, depending on whether coniferyl aldehyde is explicitly included, and 10 to 16 distinct power-law terms, depending on the topology in a specific design. Also, there are six independent variables, each of them representing the extractable activity of an enzyme.

As a typical example, the differential equation for caffeoyl CoA, defined as $X_1$, takes the following form:

$$\frac{dX_1}{dt} = \gamma_1 - \gamma_2 X_1^{f_{2,1}} X_3^{f_{2,3}} X_{n+1} - \gamma_3 X_1^{f_{3,1}} X_2^{f_{3,2}} X_{n+2}. \quad (5.1)$$

In addition to $X_1$ itself, two other dependent variables are included in the equation; they are $X_2$ and $X_3$ and refer to caffeyl aldehyde and feruloyl CoA, respectively. They are included because they are candidates of modulating the consumption of $X_1$. Applying the rules for kinetic orders described in Chapter 1, we can immediately impose bounds on the values of $f_{2,3}$ and $f_{3,2}$ for different regulatory mechanisms (Figure 5.3B). For instance, modeling Mechanism 1 requires the following constraints,

$$f_{2,3} < 0 \;\&\; f_{3,2} = 0, \quad (5.2)$$

because $X_3$ is considered an inhibitor, so that $f_{2,3} < 0$, while $X_2$ has no influence on the degradation of $X_1$ through reaction 3 in this design, so that $f_{3,2} = 0$, which in effect eliminates the factor $X_3^{f_{3,2}}$ from the term on the far right. The two independent variables $X_{n+1}$ and $X_{n+2}$ represent CCoAOMT and CCR2, respectively, where $n$ is the number of dependent variables. By convention, all independent variables have a kinetic order of 1.

### 5.4.2 Sampling of Steady-State Fluxes

Determination of all parameters in a GMA model, including kinetic orders and rate constants, is required prior to most simulation tasks. For the lignin pathway in *Medicago*, very little information is available on exact concentrations of intermediates or fluxes through the pathway; in fact, many metabolites *in vivo* are below detection level with standard HPLC [26]. To address this issue of insufficient data, we sample parameter values from relatively wide, biologically realistic ranges. The procedure involves the following steps. First, we sample uniformly from a set $P$ of steady-state reaction rates in $m$-dimensional space, where $m$ equals the number of reactions and $P$ is bounded by many

linear equality or inequality constraints with physiological meaning, such as the reaction stoichiometry, the ratio of S to G lignin in a wild-type *Medicago* species, and the degree of reversibility of individual reactions (Section C.1). For all designs studied, the resulting set $P$ is a bounded polyhedron (or polytope) and therefore has a concise parametric description

$$P = \left\{ \sum_{i=1}^{k} \alpha_i u_i \;\middle|\; u_i \in \mathbf{R}^m,\ \alpha_i \in \mathbf{R},\ \alpha_i \geq 0,\ \sum_{i=1}^{k} \alpha_i = 1 \right\}, \tag{5.3}$$

where the vectors $u_i$ can be identified using first principles [206]; in a different context, the vectors $u_i$ have been called "extreme pathways" [30].

### 5.4.3 Steady-State Equations in S-System Format

Once a set of steady-state reaction rates is randomly generated, we sample kinetic orders ($f_{i,j}$) from their respective ranges (Table C.2), which are chosen based on the unique role of each kinetic order. Even with this information, the lack of concentration data from a wild-type Medicago species remains an issue that needs to be solved. To this end, we perform two transformations. First, we define a normalization of variables by replacing $X_i$ with $Y_i \equiv X_i/X_{iS}$, where $X_{iS}$ are the unknown steady-state levels of $X_i$ in wild type. As an example, the differential equation for caffeoyl CoA assumes the form

$$\frac{dY_1}{dt} = \left( V_{1S} - V_{2S} Y_1^{f_{2,1}} Y_3^{f_{2,3}} Y_{n+1} - V_{3S} Y_1^{f_{3,1}} Y_2^{f_{3,2}} Y_{n+2} \right) \cdot \frac{1}{X_{1S}}, \tag{5.4}$$

where $V_{iS}$ are the steady-state reaction rates sampled from a set $P$ that is representative of a wild-type *Medicago* species. This representation is well suited for the current analysis because the exact values of $X_{iS}$ become irrelevant once all the equations are set to zero, that is, at a wild-type or perturbed steady state. Second, after all parameters for a given GMA model instantiation are specified, we derive the corresponding S-system equations with straightforward mathematical manipulations that do not require any additional

biological information ([44]: Chapter 3). At the steady state, GMA and S-system models are equivalent, but they offer different advantages for further analyses. In particular, S-system differential equations, despite being intrinsically nonlinear, become linear at the steady state after a logarithmic transformation, thereby facilitating the computation of secondary steady-state features and bypassing the time-consuming numerical integration that is otherwise required for assessing nonlinear models. Given this convenient feature, we are able to obtain, in a very efficient manner, estimates of steady-state fluxes under conditions that mimic the two transgenic alfalfa lines and two *M. truncatula* mutant lines; we can also easily compute the S/G ratios for which we had experimental data.

### 5.4.4 Simulation of Knock-Down Experiments

Down-regulation of specific lignin biosynthetic enzymes is simulated by setting the corresponding $Y_i$ to values between 0 and 1 that represent the degree of down-regulation, and solving the steady-state equations. In cases where knocking down the activity of one enzyme (*e.g.*, CCR2) somehow increases the activities of other enzymes (*e.g.*, CCR1 and CCoAOMT), all affected $Y_i$ are given values that mirror the specific changes in activities. The Parallel Computing Toolbox™ in MATLAB (version R2009b, The MathWorks, Natick, MA) was used to divide the simulation job among multiple cores for speedup.

Not all models behaved properly during simulation, and some ill-behaved models were excluded from further analysis. These were defined, arbitrarily, as models that showed a more than 1000-fold increase or decrease in any dependent variable during any simulation. Further, a properly behaved parameter set was deemed valid if the following criteria were met:

1. Quantitative correctness for simulations of CCoAOMT and COMT down-regulation, which was defined as a mean squared difference between the predicted and observed S/G ratios of less than 0.01 in these two cases.
2. Qualitative correctness for the simulations of *ccr1* and *ccr2* knockout mutants. Specifically, the predicted S/G ratio must show a decrease of more than 5% for *ccr1* (or an increase of more than 5% for *ccr2*), compared to the wild-type value.

### 5.4.5 Expression of Alfalfa CCoAOMT in *E.coli*[7]

The cloning of the alfalfa CCoAOMT cDNA into the expression vector pET15b was as described previously [60]. *E. coli* Rosetta strains containing the constructed plasmid were cultured at 37 ºC until $OD_{600}$ reached 0.6-0.7, and protein expression was then induced by adding isopropyl 1-thio β-galactopyranoside (IPTG) at a final concentration of 0.5 mM, followed by 3 h incubation at the same temperature. Cell pellets from 25 ml induced medium were harvested and frozen at -80ºC for further use. Induced cell pellets were thawed at room temperature, resuspended in 1.2 ml of extraction-washing buffer (10 mM imidazole, 50 mM Tris-HCl pH 8.0, 500 mM NaCl, 10% glycerol and 10 mM β-mercaptoethanol), and sonicated three times for 20 s. Supernatants were recovered after centrifugation (16,000 x *g*), and incubated at 4 ºC for 30 min with equilibrated Ni-NTA beads (Qiagen, Germantown, MD) under constant inversion to allow the His-tag protein to bind to the beads. The beads were washed three times with 1 ml of extraction-washing buffer, and the target protein was eluted with 250 µl of elution solution (250 mM imidazole, 50 mM Tris-HCl buffer pH 8.0, 500 mM NaCl, 10% glycerol and 10 mM β-mercaptoethanol). The concentration of the eluted

[7] This part was done by our collaborators at the Noble Foundation.

target protein was determined using the BioRad protein assay (BioRad, Hercules, CA) and its purity was verified by SDS-PAGE.

### 5.4.6 Materials and Enzyme Activity Assays

Caffeoyl CoA for the enzyme assays, and feruloyl CoA for the calibration curve, were synthesized as described previously [207]. Caffeyl aldehyde was synthesized as described by Chen et al. [208]. Pure recombinant CCoAOMT enzyme (100 ng) was incubated at 30 ºC for 20 min with 60 mM sodium phosphate buffer pH 7.5, 200 µM S-adenosyl methionine, 600 µM $MgCl_2$ and 2 mM dithiothreitol. The substrate (caffeoyl CoA) concentration was 20, 30 or 40 µM and the putative activator (caffeyl aldehyde) concentration was 0, 2, 5 or 10 µM. Since caffeyl aldehyde was in dimethyl sulfoxide solution, the final concentration of dimethyl sulfoxide in the reaction was 4% and the final volume of the reaction was 50 µl. The reactions were stopped by adding 10 µl of 24 % w/v trichloroacetic acid. Reaction products were analyzed by reverse-phase HPLC on a C18 column (Spherisorb 5µ ODS2, Waters, Milford, MA) in a step gradient using 1% phosphoric acid in water as solvent A and acetonitrile as solvent B. Calibration curves were constructed with authentic standard of the product feruloyl CoA. Activity assays using lower concentrations of the substrate caffeoyl CoA (2, 4, 5 and 10 µM) and the putative activator caffeyl aldehyde (0.5, 1, 2 and 4 µM) were performed using a sensitive radioactive assay method as described previously [60].

# CHAPTER 6

# CONCLUSIONS AND FUTURE WORK

## 6.1 Conclusions

The contributions of this dissertation are two-fold. First, novel methods were proposed to: (i) integrate kinetic models with fluxes predicted by static, constraint-based models; (ii) simultaneously investigate various wild-type and transgenic lines and their different developmental stages; and (iii) assess all experimentally supported pathway designs via computational enumeration. Second, mechanistic insights that were derived from model predictions and later confirmed experimentally have advanced our knowledge of how lignin biosynthesis is regulated in bioenergy crops. These contributions were made in three projects, corresponding to stated Specific Aims, which will be discussed in detail.

As proof of concept, the goal of Aim 1 was to develop a dynamic model of monolignol biosynthetic pathway in *Populus* xylem. The target genus *Populus* includes poplar, the first tree and potential bioenergy crop to have its genome sequenced [17], and aspen. Both species have been well-characterized with many *in vitro* assays of individual pathway enzymes as well as with transgenic variants modified in monolignol biosynthesis. As revealed by these biochemical and phenotypic data, the pathway is controlled by various levels of metabolic regulation. To address such complexity, we developed a novel modeling approach that combines the strengths of both static, constraint-based and dynamic, kinetic-based models. As demonstrated in Chapter 2, the resulting dynamic model not only allowed the prediction of S/G ratio in response to genetic perturbations in the pathway, but also assisted in the design of gene modification strategies towards the maximum release of sugars from *Populus* plants. Given its intuitive

structure, the model offers a solid foundation and starting point for future analytical efforts when new information becomes available.

A severe limitation to the model developed in Aim 1 is that the composition of lignin, which is an important feature of its structure, varies among taxa, cell types, and individual cell wall layers and is influenced by developmental and environmental cues. For this reason, Aim 2 of this dissertation was devoted to analyzing a compilation of wild-type and transgenic alfalfa lines where measurements of lignin content and composition are available for eight stem internodes. To analyze several internodes simultaneously, we developed a FBA or MOMA model for each internode in a wild-type or transgenic plant and integrated the data in a semi-dynamic fashion (Figure 3.3). By evaluating the transgenic data in such a systematic way across different stages of growth, we were able to elucidate regulatory mechanisms that may have remained elusive in traditional approaches where only one internode or one transgenic line is studied at a time.

The result of this comprehensive analysis was formulated into six postulates, and two of them are especially intriguing. The first suggests that certain pathway enzymes may assemble into functionally independent channels towards the synthesis of different monolignols, while the second proposes a novel feedforward regulation by an unknown cinnamic acid-derivative. Interestingly, this latter postulate was verified in a *post hoc* experiment where salicylic acid, a notable endogenous signaling molecule known to be derived from cinnamic acid, was identified as a candidate for carrying out the postulated regulation. Together, these model-based findings not only direct new, targeted experiments towards a better understanding of monolignol biosynthesis, but also highlight the importance of context when it comes to analyzing this pathway.

While the results presented in Chapter 3 have greatly improved our knowledge of how monolignol biosynthesis is regulated, they seem to provoke more questions than they have answered. For example, why do we observe a specific developmental pattern of

fluxes but not other alternatives that may seem equally valid? And by which criteria, if any, is it chosen? To address these questions, we presented in Chapter 4 two methods for characterizing alternative operating strategies for metabolic pathways. Specifically, we studied the frequently required transition of a biological system from its normal steady state to a new target steady state. This situation is quite common and includes the heat stress response in yeast as well as developmental reprogramming of lignin biosynthesis. In a very generic fashion, the two methods yield distinct yet complementary insights about the set of solutions that are *a priori* equally valid: One defines the entire space of admissible solutions, whereas the other identifies an optimal solution based on given criteria of functional effectiveness. Although we do not yet have answers to the original questions, the work presented in Chapter 4 suggests tools for comparing different solutions with objectivity and for selecting the fittest among different alternatives once criteria are established.

Another type of questions that could arise from inspecting the results in Chapter 3, especially regarding the channeling postulate, includes the following: What is the biological function of the postulated channels and how do they operate *in vivo*? Are they constitutively or conditionally active? Is there crosstalk between them, and if so, how is it organized? Indeed, answers to these questions are critical for dissecting the regulatory role of metabolic channeling in monolignol biosynthesis, but they are difficult, if not impossible, to obtain exclusively with experimental means. Therefore, we proposed in Chapter 5 a novel computational approach that permits an expedient and exhaustive assessment of hundreds of scenarios (here called *designs*) that could occur *in vivo*. Interestingly, this comparative analysis not only helped distinguish two most parsimonious mechanisms of crosstalk between the two channels by formulating a targeted and readily testable hypothesis, but also suggested that the G lignin-specific channel is more important for proper functioning than the S lignin-specific channel. Although the strategy of analysis presented in Chapter 5 is tightly focused on monolignol

biosynthesis, it is likely to be of similar utility in extracting unbiased information in a variety of situations, where the spatial organization of molecular components is critical for coordinating the flow of cellular information, and where initially different variant designs seem equally valid.

## 6.2 Future Work

The following directions are proposed for future research:

- In Aim 1 of this dissertation, we developed a dynamic model of lignin biosynthesis in *Populus* xylem and demonstrated its predictive power. With the genus-specific findings from Aims 2 and 3, it seems to be a natural next step to convert the currently static model of lignin biosynthesis in *Medicago* into an integrated, dynamic model. Once this step is accomplished, such a model will become an invaluable tool for: (i) guiding the rational design of engineered crops with reduced recalcitrance, based on model optimization; and (ii) investigating the operating principles that govern the developmental re-programming of lignin biosynthesis.
- In Chapter 5 of this dissertation, we found dozens of designs with at least one valid model instantiation but did not seek to figure out which design might be overall the *best*. One reason is that the criteria for the best design are not necessarily known *a priori*. Typical performance criteria for functionally effective systems include stability, robustness and responsiveness [209], and it seems that most, if not all candidate designs in Chapter 5 indeed satisfy these criteria. However, one could explore a metric such as the Bayes factor [210], which offers an objective tool for evaluating the evidence given by the data in favor of one design as opposed to another. It would be interesting to assign such a design score to each of the aforementioned criterion and test whether the collective results can

be explained by a Pareto front, as it was shown in recent studies of evolutionary trade-offs [35,211].

- The models developed in this dissertation are limited in scope due to the fact that they only consider lignin biosynthesis. There are many other pathways implicated in cell wall synthesis, including the biosynthesis of cellulose, hemicellulose, and pectin. Future work in this area could include the development of a specific model for each individual pathway and later integrate all pathway models into a comprehensive cell wall model. A complex issue is that these cell wall components are synthesized at different cellular locations: cellulose at the plasma membrane, monolignols in cytosol, and most other hemicellulosic polysaccharides in the Golgi apparatus [212]. If one cannot validly argue that the spatial effects can be ignored, then other modeling frameworks, capable of incorporating spatial information, will need to be developed. The typical approaches for such tasks are partial differential equation (PDE) models or, more likely, methods of agent-based modeling (ABM) [213].
- An interesting expansion of the models presented in this dissertation could involve a detailed characterization of molecular species that are also utilized in a variety of other cellular processes that are or are not associated with cell wall synthesis. One such species is shikimate, which is not only a co-substrate of HCT in monolignol biosynthesis but also a precursor for the three aromatic amino acids phenylalanine, tyrosine, and tryptophan [214]. More intriguingly, it has been shown that salicylic acid, the candidate molecule for carrying out the postulated feedforward regulation (Chapter 3), can be synthesized from isochorismate via the shikimate pathway (*cf.* Figure 3.10; [111]), suggesting that there might be extensive crosstalk between the biosynthesis of aromatic amino acids and cell wall components.

- The methods developed here reach beyond the lignin pathway. In fact, neither the spatial organization of molecules into complexes nor the proposed substrate competition is unique to monolignol biosynthesis. In signal transduction pathways, mixtures of kinases, phosphatases, and other signaling proteins may form transient, nanoclusters on the plasma membrane that operate as temporary signaling platforms or reaction chambers [215]. Additionally, it was recently found in *Drosophila* embryo that substrate competition plays an important role in mitogen-activated protein kinase (MAPK) signaling [216]. As these mechanisms are often studied independently but most likely operate concurrently, the computational approaches developed in Chapter 5 may be readily applicable to signaling systems and help us understand how distinct control mechanisms are coordinated to carry out specific biological functions in the cell.

# APPENDIX A

# SUPPLEMENARY MATERIALS FOR CHAPTER 2

## A.1 Supplementary Methods

### A.1.1 Determination of pathway structure

As a first step, we start with a detailed description of the monolignol biosynthetic pathway as shown in Figure A.1. Some of the constituent reactions/processes, however, require further consideration because the corresponding genetic or biochemical evidence has only been found in genera other than *Populus*, and thus may be affected by species-to-species variations. In this regard, the next paragraphs will discuss three simplifying assumptions; the resulting pathway structure (or metabolic map) of monolignol biosynthesis is shown in Figure 2.1.

First, the conversion of *p*-coumaroyl CoA ($X_3$) to caffeoyl CoA ($X_6$) in effect represents the collective effort of *p*-coumaroyl shikimate 3-hydroxylase (C3H) and two acyltransferases, namely hydroxycinnamoyl CoA:shikimate hydroxycinnamoyl transferase (HCT) and hydroxycinnamoyl CoA:quinate hydroxycinnamoyl transferase (HQT). The major reason for this simplification is that no reports have truly quantified the accumulation of intermediate products [105], so that the entire conversion process can only be regarded as a single step.

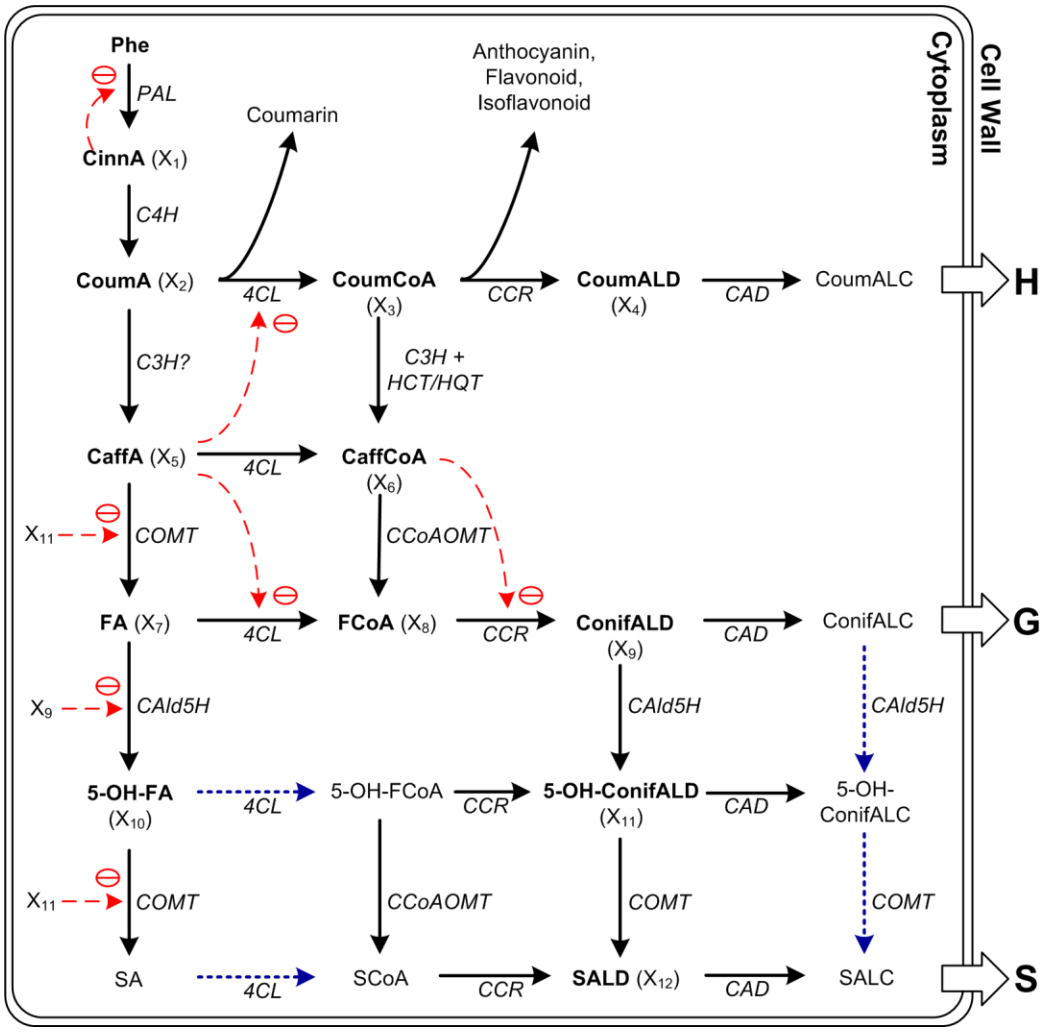


**Figure A.1: Generic metabolic map of the monolignol biosynthetic pathway.** Metabolites in bold are represented by dependent variables $X_i$, $i = 1,...,12$, whereas enzymes are shown in italics. Solid black arrows represent material flow, whereas dashed red arrows represent regulatory signals, with negative signs indicating inhibition. As a reference, blue dotted arrows refer to putative enzymatic reactions that have been validated in other species but not yet in *Populus*. Transport processes of monolignols into the cell wall are shown as open arrows. For the definition of abbreviations, please refer to the legend of Figure 2.1.

Second, previous studies have shown that the hydroxylation of the aromatic ring at the $C_5$ position and the subsequent phenolic *O*-methylation could occur at the alcohol level [14]. We decided not to include the putative pathway from coniferyl alcohol to sinapyl alcohol through 5-hydroxyconiferyl alcohol for the following reason. Although there is evidence for the conversion between these alcohols (or monolignols) in

sweetgum [72] and *Arabidopsis* [10], the findings in [71] clearly demonstrate that the recombinant caffeic acid *O*-methyltransferase (COMT)—the principal enzyme responsible for *O*-methylation in the pathway—as well as protein extracts from aspen xylem tissues show little activity towards 5-hydroxyconiferyl alcohol. Further experimental data will be needed to assess the validity of this simplification with greater rigor.

Third, while 4-coumarate:CoA ligase (4CL) had been shown to mediate the *in vitro* CoA ligation of 5-hydroxyferulic acid with minor efficiency [217], this activity was later found to be severely compromised by the presence of other substrates such as *p*-coumaric acid, caffeic acid and ferulic acid [68]. In their presence, 5-hydroxyferuloyl CoA is present in insignificant amounts, which is consistent with its scarcity *in vivo* (Wout Boerjan, personal communication). Hence, we exclude both 5-hydroxyferuloyl-CoA and its downstream derivative, sinapoyl CoA, from our considerations until new evidence demonstrates their importance.

### A.1.2 Derivation of parameter values

For any biochemical reaction assuming a Michaelis-Menten rate law, we need the following information to characterize its corresponding power-law representation: kinetic features of the enzyme (*e.g.*, $V_{max}$ and $K_M$) and the substrate concentration at the chosen operating point, which often corresponds to the metabolite concentration at the system's nominal steady state. The kinetic order $g$ for such a rate law with steady-state substrate concentration $S$ is given as

$$g = \frac{K_M}{K_M + S}. \tag{A.1}$$

By equating the power-law representation and the original rate law at the chosen operating point, we can solve for the rate constant $\alpha$:

$$\alpha = \frac{V_{\max} S}{K_M + S} S^{-g} \quad \text{at the operating point.} \tag{A.2}$$

If an appropriate rate law is not available for a specific biochemical reaction, other types of information are required for the derivation of parameter values; Table 2.2 shows some examples. A detailed description of estimation techniques for kinetic orders and rate constants from different rate laws can be found in Chapter 5 of [44].

One issue regarding the use of experimental data collected from different cells, tissues, or even organisms is that the unit of a given quantity (*e.g.*, concentration, $V_{max}$, catalytic activity, etc.) often appears in many distinct variants. In our case, the unit of concentration such as pmol/mg DW is not directly comparable with that of a $K_M$ documented in μM, which implies that further efforts are needed, including the search for pertinent biological information that permits the conversion. As an example, consider the unit for the concentration of ferulic acid (FA). Fisher [218] published the density of air-dry *Populus* as approximately 0.45 g/ml. Since the air-dry wood still contains about 15 percent of its weight in water [218], we can compute the density of the supposedly desiccated *Populus* as 0.38 g/ml. Next, we need to compute the percentage of water in fresh *Populus* by volume. Assuming that the volumes taken up by dry matter and water are $V_1$ and $V_2$, respectively, and that water constitutes 90% of fresh plant material by weight, we can determine the percentage of water in fresh *Populus* by volume $\rho_W$ as follows:

$$\begin{aligned} V_2 &= 9 \times 0.38 \times V_1 = 3.42 V_1, \\ \rho_W &= \frac{V_2}{V_1 + V_2} \approx 0.77. \end{aligned} \tag{A.3}$$

With these two quantities, the concentration of FA in μM can be approximated as

$$\frac{75.5\,pmol/mg \times 0.38 g/ml}{0.77} \approx 37.26 \mu M. \tag{A.4}$$

Although it seems reasonable to perform the conversion of units for all metabolite concentrations in the same fashion, this is not always easy. For instance, we are still in need of new concentration measurements, especially for the CoA esters, to confirm the accuracy of their estimated concentration values. As for those metabolites known to be lowly abundant *in vivo* or below the detection limit, we can presently only use a small number, such as 0.1 μM, as the nominal value. The steady-state concentration of cinnamic acid, on the other hand, is set to a larger value (1 μM) as implied by experimental findings in *Pinus taeda* cells [219].

As shown in Table A.2, we do not possess a complete kinetic description—$K_M$, $K_I$, and $V_{max}$—for every reaction within the pathway. In fact, some of the measurements that are available to date can only be found in organisms other than aspen or poplar. However, since the only type of parameter that needs to be defined within the current context is the kinetic order, we might be able to estimate a small number of parameters with our *a priori* knowledge of $K_M$ and the steady-state substrate concentrations. For example, given that all CoA esters are known to be lowly abundant *in vivo* (Table A.1), we may assume from the first equation that a kinetic order with any of the CoA esters as substrate is close to (or equal to) 1. Similarly, the high $K_M$ values as observed in reactions catalyzed by *p*-coumarate 3-hydroxylase (C3H) or coniferyl aldehyde 5-hydroxylase (CAld5H) with ferulic acid (FA) as substrate (Table A.2) would have the same effect on the corresponding kinetic orders. As for the parameters lacking any useful information, we accept the widely used default hypothesis that substrate concentrations are similar to the values of $K_M$ *in vivo*, and use 0.5 as an initial guess for the corresponding kinetic order [44].

**Table A.1: Metabolite concentrations.**

| Dependent variable ($X_i$) | Metabolite | Concentration[a] (pmol/mg DW) | Concentration (μM) |
|---|---|---|---|
| **1** | cinnamic acid | N/A | |
| **2** | *p*-coumaric acid | < 0.5 | < 0.25 |
| **3** | *p*-coumaroyl CoA | Low[b] | |
| **4** | *p*-coumaryl aldehyde | 0.5 | 0.25 |
| **5** | caffeic acid | < 0.5 | < 0.25 |
| **6** | caffeoyl CoA | Low[b] | |
| **7** | ferulic acid | 75.5 ± 16 | 37.26 |
| **8** | feruloyl CoA | Low[b] | |
| **9** | coniferyl aldehyde | 28.4 ± 4 | 14.02 |
| **10** | 5-hydroxyferulic acid | < 0.5 | < 0.25 |
| **11** | 5-hydroxyconiferyl aldehyde | < 0.5 | < 0.25 |
| **12** | sinapyl aldehyde | 89.5 ± 14 | 44.17 |

[a]The *in vivo* concentrations of these hydroxycinnamic acids and hydroxycinnamyl aldehydes are from [62]. We later discovered that the concentration of ferulic acid (FA) is apparently lower than measured in another study [61], which used exactly the same analysis. Consequently, we decided to replace the previous value (147 ± 70) with a new measurement (75.5 ± 16) for the concentration of FA. DW = dry weight.
[b]Wout Boerjan, personal communication

**Table A.2: Enzyme kinetic constants.**

| Enzyme (EC number) | Gene (GenBank ID) | Substrate | $V_{max}$[a] | $K_M$ (µM) | Organism | Reference |
|---|---|---|---|---|---|---|
| Phenylalanine ammonia-lyase (EC 4.3.1.24) | AtPAL2[b] (AY303129) | Phenylalanine | 10.5 pmol/s/µg | 64 | *Arabidopsis thaliana* | [220], Table 2 |
| Cinnamate 4-hydroxylase (EC 1.14.13.11) | | CinnA | | 0.7 | *A. thaliana* | [221], Table 1 |
| 4-coumarate:CoA ligase (4CL) (EC 6.2.1.12) | Pt4CL1[c] (AF041049) | CoumA | 27.41 µM/min/µg | **55.64**[d] | Aspen | [68], Table I |
| | | CaffA | 17.49 µM/min/µg | **34.68** | | |
| | | FA | 15.8 µM/min/µg | **112.05** | | |
| Cinnamoyl CoA reductase (EC 1.2.1.44) | | CoumCoA | 1.3 nmol/s/mg | 4.27 | Aspen | [106], Table 5 |
| | PtCCR (AF217958) | FCoA | 158.6 µM/min | **13.7** | Poplar | [70], Table 1 |
| Cinnamyl alcohol dehydrogenase (EC 1.1.1.195) | PtCAD[e] (AF217957) | CoumALD | | **6.2** | Aspen | [222], Table 1 |
| | | ConifALD | | **2.3** | | |

**Table A.2 continued.**

| | | | | | | |
|---|---|---|---|---|---|---|
| | | 5-OH-ConifALD | | **17.5** | | |
| | | SALD | | **9.1** | | |
| *p*-coumarate 3-hydroxylase (N/A) | CYP98A3 (NM_180006) | CoumA | | High | *A. thaliana* | [223], p.39-40 |
| Caffeic acid *O*-methyltransferase (EC 2.1.1.68) | | CaffA | **1.85** µM/min | **75.1** | Aspen | [71], Table I |
| | | 5-OH-FA | **2.2** µM/min | **15** | | |
| | | 5-OH-ConifALD | **2** µM/min | **2.6** | | |
| Caffeoyl-CoA *O*-methyltransferase (EC 2.1.1.104) | | CaffCoA | | 27.5 | Tobacco | [116], p.36836 |
| Coniferyl aldehyde 5-hydroxylase (N/A) | | FA | 46.5 nM/min | 286.05 | Sweetgum | [72], Table 2 |
| | | ConifALD | 64.58 nM/min | 2.77 | | |

[a]Units other than µM/min or nM/min are not valid because the corresponding enzyme concentrations are missing/lacking.
[b]The PAL family is known to constitute four isoforms in *Arabidopsis*, but only AtPAL2 is listed here because its gene product is the most catalytically effective, as is indicated by the highest $V_{max}/K_M$ among the isoforms.
[c]Previous findings [20,217] suggested that the gene product of Pt4CL1, but not the other isoform Pt4CL2, dominates in the developing xylem tissue of aspen.
[d]Values in bold are used in the model
[e]A homolog to CAD, sinapyl alcohol dehydrogenase (SAD), has been identified in aspen [222]. Nevertheless, the implication of SAD being the enzyme responsible for the conversion of sinapyl aldehyde to S monolignol was later proved to be inconsequential in poplar [224] and is thus not listed.

### A.1.3 Co-linearity between two kinetic orders

The distributions of parameter values within the ensemble of models indicate that the parameters $f_{CAD,ConifALD}$ and $f_{CAld5H,ConifALD}$ are linearly dependent with a slope close to 1. If we denote by $v_{CAD}$ and $v_{CAld5H}$ the fluxes catalyzed by enzymes CAD and CAld5H, with coniferyl aldehyde (ConifALD) as a common substrate, the ratio between these two fluxes in power-law representation can be characterized as

$$\frac{v_{CAD}}{v_{CAld5H}} = \frac{\gamma_{CAD}}{\gamma_{CAld5H}}[ConifALD]^{f_{CAD,ConifALD} - f_{CAld5H,ConifALD}} . \quad \text{(A.5)}$$

Assuming that $f_{CAD,ConifALD}$ and $f_{CAld5H,ConifALD}$ are exactly identical in value, we can reduce the equation to

$$\frac{v_{CAD}}{v_{CAld5H}} = \frac{\gamma_{CAD}}{\gamma_{CAld5H}} = \frac{v_{CAD}^0}{v_{CAld5H}^0} , \quad \text{(A.6)}$$

where $v_{CAD}^0$ and $v_{CAld5H}^0$ are the corresponding steady-state values estimated by FBA. Consequently, the ratio between these two fluxes is equal to the ratio between their respective rate constants, which in turn corresponds to the ratio of these two fluxes at steady state.

### A.1.4 Local stability analysis

For each randomly sampled generalized mass action (GMA) model, we must ensure that the system behavior is robust to small fluctuations in metabolite concentrations. In other words, the system has to return to its original FBA-based steady state when confronted with minor perturbations in the dependent variables. To determine local stability, we computed the eigenvalues of the Jacobian matrix derived from the GMA model and evaluated it at the steady state. The necessary condition for local stability demands that the largest real part of all eigenvalues must be less than zero. For a

GMA model with a stoichiometric matrix **N**, we can use elementary calculus and specify every entry $J_{ij}$ of the corresponding Jacobian matrix as

$$J_{ij} = \sum_{k=1}^{Z} N_{ik} f_{kj} \frac{v_k^0}{X_j^0}, \tag{A.7}$$

where $Z$ is the total number of fluxes within the pathway and $f_{kj}$ is the kinetic order of metabolite $X_j$ with respect to metabolic flux $v_k$; the superscript 0 indicates that these are steady-state values. Clearly, different parameter profiles will lead to different Jacobian matrices, and probably different dynamics in the vicinity of the steady state. Although the computational efforts for carrying out the local stability analysis are considerable as the number of dependent variable grows, it is nonetheless a significant test to prove the system's ability to maintain its homeostatic behavior in the face of spurious perturbations.

### A.1.5 Mutual information and its numerical estimation

By definition, the mutual information $I(X;Y)$ between two discrete random variables $X$ and $Y$ with joint distribution $p(x,y)$ and marginal distributions $p(x)$ and $p(y)$ can be written as

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{A.8}$$

It is noteworthy that $I(X;Y)$ becomes zero if and only if two variables are statistically independent, *i.e.*, $p(x,y) = p(x)p(y)$. In order to estimate the required probability distributions from a population of stable GMA models, we first replaced all numerical values—whether they refer to a kinetic order or the S/G ratio of one transgenic experiment—by their respective rank order, and then divided them into *M* discrete bins, as if there were *M* different states. Next, we computed the mutual information between

any two variables using the naïve algorithm [225] as a function of *M*. While the choice of *M* has been shown to affect the absolute value of the estimated mutual information [225], we are in fact more concerned with whether the two variables are significantly correlated. Consequently, we generated another 99 datasets $\{X^i, Y^i\}$, $i = 1,...,99$, through a random permutation of the original data $\{X, Y\}$. By assuming that $X^i$ and $Y^i$ are independent for all *i*, we may claim with a $(100-1)\%$ level of significance that $X$ and $Y$ are truly correlated if $I(X;Y)$ is distinct from $I(X^i;Y^i)$ [226]. Given that the identification of significant parameters seems invariant to our choice of *M*, we simply choose *M* = 10 for all cases.

### A.1.6 Indirect Optimization Method

Following the original proponents of the Indirect Optimization Method (IOM) [85], the first step involves the reformulation of the original GMA model as an S-system model. The GMA and S-system variants within BST, while both using products of power-law functions in a similar fashion, differ in one key aspect: while each flux in a GMA model is represented as an individual power-law term, S-system models aggregate all production fluxes of a (metabolite) pool into one power-law term and all degradation fluxes into a second power-law term. Previous studies have shown that in many cases the differences between these two representations are negligible [47,227,228,229,230], and that sometimes the accuracy of modeling biochemical reactions with known kinetics is improved in the S-system representation [230]. While both formats have their own advantages and disadvantages, the clear winner for optimization tasks is the S-system format, because the optimization problem becomes strictly linear (in logarithmic space), which permits application of the well-developed theory of linear programming and access to many readily available software packages. Like the core equations of the optimization

task, the objective function and relevant constraints on variables and fluxes also become linear when the S-system format is used, so that the entire constrained optimization task becomes linear.

A complete definition of a linear optimization problem demands more details than can be explained here, specifically with respect to constraints on variables and fluxes, and the interested reader is referred to [128]. Variables representing metabolites and enzymes are allowed to vary within a certain range determined from professional experience, technological capacity, or just educated guesses. Here, we allow the enzyme activities to change between 5% and 5 times the basal levels, which is in line with past expert experience. Furthermore, we assume that metabolite concentrations may vary by a factor of 10, without being physiologically detrimental. Obviously, it is easy to reset these numerical values if it is deemed appropriate. Constraints on fluxes are determined by the steady-state definition of the system and, in particular, insights from the prior flux balance analysis. Finally, the objective function is given as the ratio between two fluxes representing the production of S and G. This function can again be represented by a single linear equation in logarithmic coordinates.

The optimization with an ensemble of models, which were fitted to data with a simulated annealing algorithm (see the following section), is executed with the function *linprog* in MATLAB (Mathworks Inc.). For an assessment of the approximation errors in the IOM approach, we apply the optimized enzyme profile obtained from the intermediate S-system model as an input to the original GMA models and solve for the steady state.

### A.1.7 Model-fitting algorithm

We used a simulated annealing algorithm [231] to find the values of the significant parameters that minimize the sum of squared errors (SSE) between the

measured and the predicted S/G ratios of the five transgenic experiments used as the training data (see Table 2.2). Applying this algorithm to the search of "global" solutions, but at the expense of computational efforts, we executed 20 runs to obtain an ensemble of GMA models, using the *simulannealbnd* function in MATLAB (Mathworks Inc.). In brief, the SA algorithm begins with a randomly selected point in the parameter space that defines a locally stable GMA model. In each of the following iterations, we generate a new point by perturbing the old one in a randomly chosen coordinate (or parameter) with a scale proportional to the variable *T*, which is called the system temperature. One interesting feature of the SA algorithm is that it allows the SSE to temporarily increase in each iteration (which potentially avoids abundant local optima), but only with a probability controlled by *T*. In general, the probability takes the form $1/(1+e^{\Delta E/T})$, where $\Delta E$ is the difference in the SSE between the current and the previous iteration. As the algorithm proceeds, the temperature decreases on a certain "cooling" schedule, which terminates with $T = 0$, and so does the probability to take an adventurous path that raises the SSE.

# APPENDIX B

# SUPPLEMENARY MATERIALS FOR CHAPTER 3

## B.1 Overview

This Appendix includes two main sections. In Section B.2, we present the model formulation, and identify equivalent pathways that underlie the occurrence of alternate flux balance analysis (FBA) solutions. In Section B.3, we present a kinetic model for the analysis of pathway operation at the critical branch point of coniferyl aldehyde.

## B.2 Use of Flux Balance Analysis (FBA) and Minimization of Metabolic Adjustment (MOMA) for Modeling Monolignol Biosynthesis

### B.2.1 Model formulation

We constructed steady-state flux-based models for wild-type and transgenic plants based on the revised pathway structure (Fig. 3.1). The model comprises 24 flux variables; Table B.1 shows the corresponding metabolic reaction or transport process for each flux. If a reaction is associated with a specific isozyme, as in the case of CCR1 and CCR2, the encoding *Medicago* gene (represented by its tentative consensus TC number) is also listed.

**Table B.1: List of flux variables and their corresponding metabolic reaction.**

| Flux | Enzyme (TC#) | Reaction |
|---|---|---|
| $v_1$ | PAL | L-phenylalanine → cinnamic acid + $NH_3$ |
| $v_2$ | C4H | cinnamic acid + NADPH + $O_2$ → *p*-coumaric acid + $NADP^+$ + $H_2O$ |
| $v_3$ | 4CL | *p*-coumaric acid + CoA + ATP → *p*-coumaroyl CoA + Pi + AMP |
| $v_4$ | CCR2 (TC100678) | *p*-coumaroyl CoA + NADPH → *p*-coumaryl aldehyde + $NADP^+$ + CoA |
| $v_5$ | CAD | *p*-coumaryl aldehyde + NADPH → *p*-coumaryl alcohol + $NADP^+$ |
| $v_6$ | Tr* | *p*-coumaryl alcohol → Ø |
| $v_7$ | HCT | *p*-coumaroyl CoA + shikimate → *p*-coumaroyl shikimate + CoA |
| $v_8$ | C3H | *p*-coumaroyl shikimate + NADPH + $O_2$ → caffeoyl shikimate + $NADP^+$ + $H_2O$ |
| $v_9$ | HCT | caffeoyl shikimate + CoA → caffeoyl CoA + shikimate |
| $v_{10}$ | CCR2 (TC100678) | caffeoyl CoA + NADPH → caffeoyl aldehyde + $NADP^+$ + CoA |
| $v_{11}$ | CCoAOMT | caffeoyl CoA + *S*-adenosyl L-methionine → feruloyl-CoA + S-adenosyl homocysteine |
| $v_{12}$ | COMT | caffeoyl aldehyde + *S*-adenosyl L-methionine → coniferyl aldehyde + S-adenosyl homocysteine |
| $v_{13}$ | CCR1 (TC106830) | feruloyl CoA + NADPH → coniferyl aldehyde + $NADP^+$ + CoA |
| $v_{14}$ | CAD | coniferyl aldehyde + NADPH → coniferyl alcohol + $NADP^+$ |
| $v_{15}$ | Tr | coniferyl alcohol → Ø |
| $v_{16}$ | F5H | coniferyl aldehyde + NADPH + $O_2$ → 5-hydroxyconiferyl aldehyde + $NADP^+$ + $H_2O$ |
| $v_{17}$ | COMT | 5-hydroxyconiferyl aldehyde + *S*-adenosyl L-methionine → sinapyl aldehyde + S-adenosyl homocysteine |
| $v_{18}$ | CAD | sinapyl aldehyde + NADPH → sinapyl alcohol + $NADP^+$ |
| $v_{19}$ | Tr | sinapyl alcohol → Ø |
| $v_{20}$ | F5H | coniferyl alcohol + NADPH + $O_2$ → 5-hydroxyconiferyl alcohol + $NADP^+$ + $H_2O$ |

**Table B.1 continued.**

| | | |
|---|---|---|
| $v_{21}$ | COMT | 5-hydroxyconiferyl alcohol + NADPH → sinapyl alcohol + $NADP^+$ |
| $v_{22}$ | N/A[†] | cinnamic acid →→ salicylic acid |
| $v_{23}$ | N/A[†] | *p*-coumaroyl-CoA →→ anthocyanin, flavonoid, isoflavonoid,… |
| $v_{24}$ | Tr | 5-hydroxyconiferyl alcohol → Ø |

*Tr represents collectively all biochemical events during the transport of alcohol precursors into the cell wall, *i.e.*, outside the cytoplasm (Ø).
[†]$v_{22}$ and $v_{23}$ refer to the sequence of reactions that leads to the synthesis of salicylic acid and flavonoid derivatives, respectively. Thus, they are not associated with a single enzyme.

Typically, two classes of constraints are employed for steady-state flux balance models. The first is conservation of mass, which can be characterized mathematically by Eq. (3.1). Instead of presenting the constraint as the product of a stoichiometric matrix and a column vector of fluxes, we list the mass balance equation for each of the 16 intermediate metabolites in Table B.2. Details of the second class of constraints, which concerns the reversibility and maximal reaction rates of individual fluxes, have been discussed in Chapter 3 and will not be repeated here.

**Table B.2: Mass balance equations.**

| Metabolite | Balance Equation of Influxes and Effluxes |
|---|---|
| cinnamic acid | $v_1 - v_2 - v_{22}^* = 0$ |
| *p*-coumaric acid | $v_2 - v_3 = 0$ |
| *p*-coumaroyl-CoA | $v_3 - v_4 - v_7 - v_{23} = 0$ |
| *p*-coumaryl aldehyde | $v_4 - v_5 = 0$ |
| *p*-coumaryl alcohol | $v_5 - v_6 = 0$ |
| p-coumaroyl-shikimate | $v_7 - v_8 = 0$ |
| caffeoyl-shikimate | $v_8 - v_9 = 0$ |
| caffeoyl-CoA | $v_9 - v_{10} - v_{11} = 0$ |
| caffeoyl aldehyde | $v_{10} - v_{12} = 0$ |
| feruloyl-CoA | $v_{11} - v_{13} = 0$ |
| coniferyl aldehyde | $v_{12} + v_{13} - v_{14} - v_{16} = 0$ |
| coniferyl alcohol | $v_{14} - v_{15} - v_{20} = 0$ |
| 5-hydroxyconiferyl aldehyde | $v_{16} - v_{17} = 0$ |
| sinapyl aldehyde | $v_{17} - v_{18} = 0$ |
| 5-hydroxyconiferyl alcohol | $v_{20} - v_{21} - v_{24} = 0$ |
| sinapyl alcohol | $v_{18} + v_{21} - v_{19} = 0$ |

*Variables in red indicate "overflow" fluxes (*cf.* red arrows in Figure 3.1).

Constraints on lignin composition along with numerical values are presented in Table B.3. It is straightforward to translate them into a set of equality constraints in the form of Eq. (3.3). To implement MOMA, we further define $\delta_i$ (see definition in Chapter 3) in the following way: find the flux $v_i$ whose catalyzing enzyme is modified, identify the percentage of the residual enzyme activity related to its wild-type level, and set $\delta_i$ to this number; unaffected fluxes have $\delta_i = 1$.

We used *linprog* and *quadprog* routines in MATLAB to solve the linear and quadratic programming problems in FBA and MOMA, respectively.

### B.2.2 Identification of equivalent pathways

Given the constraints in Eqs. (3.1)-(3.3), we first perform an FBA for wild-type plants and then use this FBA-optimum as a reference in MOMA to infer the flux distribution for transgenic plants. A key issue that may arise from this approach is the existence of alternate optimal FBA solutions that give the same objective function value but with different flux distributions [120,121]. To address this issue, we define an (16+2+1) × 24 matrix **A** and a (16+2+1)-dimensional vector **b** such that

$$\mathbf{A}\mathbf{v} = \mathbf{b} \tag{B.1}$$

collectively represents Eqs. (3.1) and (3.3), as well as the normalization constraint $v_1 = 1$. By this definition, we know that $\mathbf{v}^{wt}$ is a solution for the following problem:

$$\begin{aligned} & z^* = \mathbf{c}^T \mathbf{v} \\ & \mathbf{A}\mathbf{v} = \mathbf{b} \\ & \mathbf{l} \le \mathbf{v} \le \mathbf{u} \end{aligned} \tag{B.2}$$

where $z^* = \mathbf{c}^T \mathbf{v}^{wt}$ is the optimal objective function value, and **l** and **u** are vectors of the lower and upper bounds on individual fluxes, respectively.

**Table B.3: Lignin content and monomer composition in wild-type and transgenic plants.**

| Internode | Lignin content and monomer composition | Control | PAL (55%)* | C4H (46%) | HCT (24%) | C3H (16%) | CCoAOMT (3%) | F5H (N/A) | COMT (3%) |
|---|---|---|---|---|---|---|---|---|---|
| 1-2 | H/T† | 7.06% | 8.71% | 6.50% | 50.17% | 11.40% | 11.19% | 5.84% | 5.63% |
| | G/T | 85.52% | 79.97% | 85.86% | 45.46% | 81.19% | 82.32% | 90.33% | 90.80% |
| | S/T | 7.42% | 11.32% | 7.63% | 4.37% | 7.41% | 6.49% | 3.83% | 3.58% |
| | AcBr Lignin (mg) | 93.13 | 71.63 | 82.09 | 62.4 | 58.8 | 83.49 | 72.56 | 80.03 |
| 3 | H/T | 6.13% | 4.85% | 5.08% | 51.12% | 15.91% | 9.20% | 4.38% | 5.42% |
| | G/T | 88.90% | 72.51% | 88.78% | 33.11% | 76.06% | 82.92% | 92.38% | 91.63% |
| | S/T | 4.97% | 22.64% | 6.14% | 15.77% | 8.04% | 7.88% | 3.24% | 2.95% |
| | AcBr Lignin (mg) | 80.86 | 80.75 | 70.8 | 64.95 | 52.73 | 71.92 | 79.48 | 76.15 |
| 4 | H/T | 3.39% | 3.29% | 3.31% | 51.29% | 19.18% | 6.53% | 3.71% | 4.43% |
| | G/T | 70.33% | 60.36% | 74.37% | 29.34% | 59.95% | 59.73% | 80.88% | 87.52% |
| | S/T | 26.28% | 36.35% | 22.32% | 19.37% | 20.87% | 33.74% | 15.41% | 8.05% |
| | AcBr Lignin (mg) | 130.2 | 106.7 | 76.92 | 77.42 | 82.09 | 99.45 | 214.9 | 117.2 |
| 5 | H/T | 2.97% | 3.01% | 3.11% | 55.93% | 20.77% | 5.48% | 2.48% | 4.21% |
| | G/T | 67.07% | 55.81% | 68.80% | 24.17% | 55.80% | 53.04% | 82.59% | 86.60% |
| | S/T | 29.97% | 41.18% | 28.09% | 19.90% | 23.43% | 41.48% | 14.93% | 9.19% |
| | AcBr Lignin (mg) | 190.6 | 109 | 150.3 | 78.07 | 113.4 | 138.2 | 235.2 | 149.2 |
| 6 | H/T | 2.40% | 2.38% | 2.40% | 64.51% | 22.88% | 4.09% | 3.06% | 3.13% |
| | G/T | 61.74% | 52.30% | 71.32% | 18.10% | 50.22% | 53.59% | 86.54% | 88.60% |
| | S/T | 35.86% | 45.32% | 26.28% | 17.40% | 26.91% | 42.32% | 10.41% | 8.28% |
| | AcBr Lignin (mg) | 225.7 | 124 | 172.1 | 81.22 | 128.9 | 169.8 | 239.8 | 182 |
| 7 | H/T | 2.04% | 2.07% | 1.96% | 68.52% | 24.96% | 3.36% | 1.79% | 2.90% |
| | G/T | 61.14% | 50.30% | 68.98% | 15.31% | 46.86% | 51.61% | 76.01% | 90.38% |
| | S/T | 36.81% | 47.63% | 29.06% | 16.17% | 28.18% | 45.03% | 22.20% | 6.72% |
| | AcBr Lignin (mg) | 248.8 | 119 | 182.2 | 78.23 | 131.4 | 172.3 | 246.4 | 189.4 |
| 8 | H/T | 1.67% | 1.97% | 1.79% | 66.56% | 25.63% | 2.75% | 1.56% | 2.53% |
| | G/T | 59.96% | 48.16% | 68.65% | 16.61% | 45.80% | 49.69% | 75.49% | 89.14% |
| | S/T | 38.38% | 49.87% | 29.56% | 16.83% | 28.57% | 47.56% | 22.95% | 8.33% |
| | AcBr Lignin (mg) | 251 | 126.9 | 182.6 | 89.33 | 130 | 186.8 | 260.4 | 199.3 |

*Percentages within the parentheses are the residual enzyme activity related to the wild-type level.
†T = H+G+S

Apparently, alternate optima occur if there are solutions for Eq. (B.2) other than $\mathbf{v}^{wt}$. If this is the case, the difference between an alternate solution and $\mathbf{v}^{wt}$, defined as $\mathbf{w}$, must also be a solution for the following sub-problem:

$$\begin{aligned} &\mathbf{c}^T\mathbf{w} = 0 \\ &\mathbf{A}\mathbf{w} = \mathbf{0} \\ &\mathbf{l} \le \mathbf{v}^{wt} + \mathbf{w} \le \mathbf{u} \end{aligned} \tag{B.3}$$

because

$$\mathbf{c}^T\mathbf{w} = \mathbf{c}^T(\mathbf{v}^{wt} + \mathbf{w}) - \mathbf{c}^T\mathbf{v} = z^* - z^* = 0 \tag{B.4}$$

and

$$\mathbf{A}\mathbf{w} = \mathbf{A}(\mathbf{v}^{wt} + \mathbf{w}) - \mathbf{A}\mathbf{v}^{wt} = \mathbf{b} - \mathbf{b} = \mathbf{0}. \tag{B.5}$$

If we define an (16+4) × 24 matrix $\mathbf{B} = \begin{bmatrix} \mathbf{c}^T \\ \mathbf{A} \end{bmatrix}$, then it is clear that $\mathbf{w}$ lies in the null space of $\mathbf{B}$, *i.e.*, $\mathbf{B}\mathbf{w} = \mathbf{0}$. Identification of the equivalent pathways, in this respect, is thus related to finding a meaningful basis of the null space of $\mathbf{B}$. By applying the Gauss-Jordan elimination to $\mathbf{B}$, we identified a basis for the pathway shown in Figure 3.1; the vectors that constitute the basis are listed in Table B.4 and also illustrated in Figure B.1.

**Table B.4: Basis vectors (*BV*) for the pathway shown in Figure 3.1.**

| | ***BV*1** | ***BV*2** | ***BV*3** | ***BV*4** |
|---|---|---|---|---|
| $v_1$ | 0 | 0 | 0 | 0 |
| $v_2$ | 0 | 0 | 1 | 1 |
| $v_3$ | 0 | 0 | 1 | 1 |
| $v_4$ | 0 | 0 | 0 | 0 |
| $v_5$ | 0 | 0 | 0 | 0 |
| $v_6$ | 0 | 0 | 0 | 0 |
| $v_7$ | 0 | 0 | 0 | 1 |
| $v_8$ | 0 | 0 | 0 | 1 |
| $v_9$ | 0 | 0 | 0 | 1 |
| $v_{10}$ | -1 | 0 | 0 | 1 |
| $v_{11}$ | 1 | 0 | 0 | 0 |
| $v_{12}$ | -1 | 0 | 0 | 1 |
| $v_{13}$ | 1 | 0 | 0 | 0 |
| $v_{14}$ | 0 | 1 | 0 | 1 |

**Table B.4 continued.**

| | | | | |
|---|---|---|---|---|
| $v_{15}$ | 0 | 0 | 0 | 0 |
| $v_{16}$ | 0 | -1 | 0 | 0 |
| $v_{17}$ | 0 | -1 | 0 | 0 |
| $v_{18}$ | 0 | -1 | 0 | 0 |
| $v_{19}$ | 0 | 0 | 0 | 0 |
| $v_{20}$ | 0 | 1 | 0 | 1 |
| $v_{21}$ | 0 | 1 | 0 | 0 |
| $v_{22}$ | 0 | 0 | -1 | -1 |
| $v_{23}$ | 0 | 0 | 1 | 0 |
| $v_{24}$ | 0 | 0 | 0 | 1 |

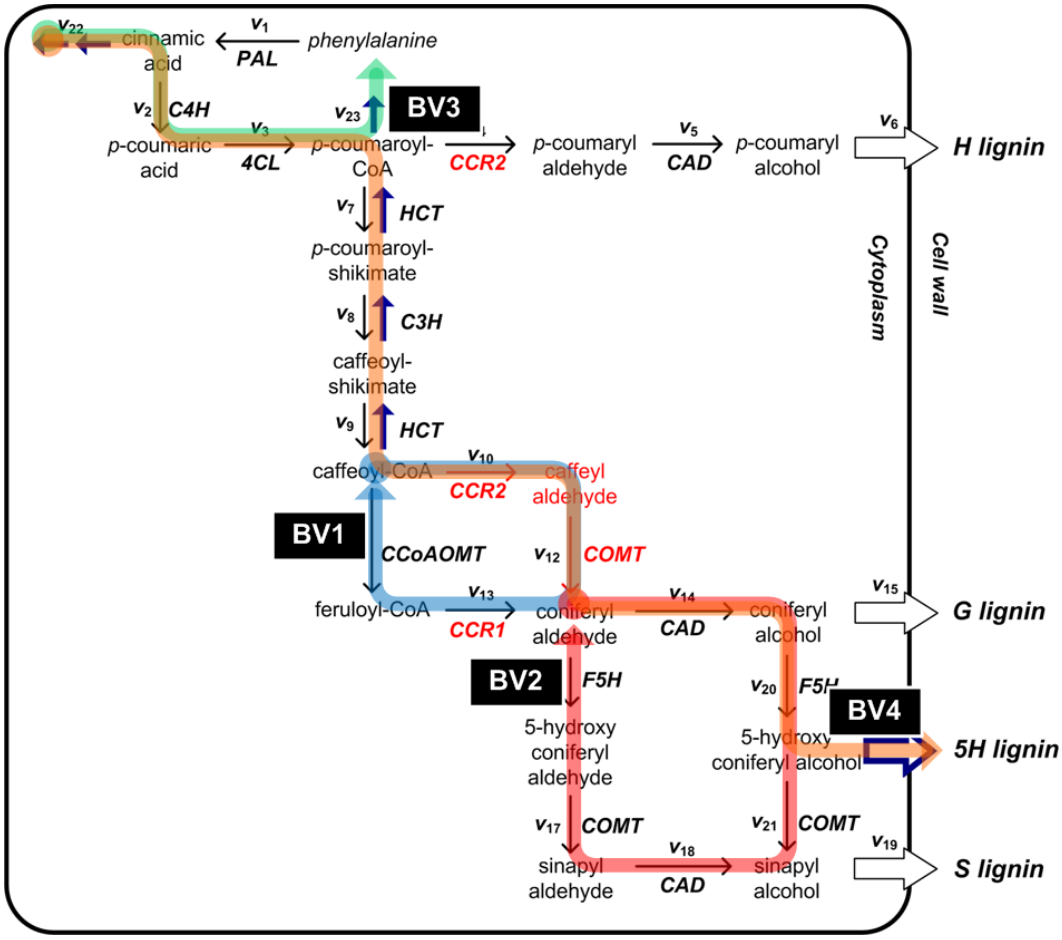


**Figure B.1: Illustration of the four basis vectors.**

Two observations are made from the identified basis. First, *BV*1 and *BV*2 correspond to the two inner loops within the pathway. Second, both *BV*3 and *BV*4 have non-zero components corresponding to two overflow fluxes, with one being positive and

the other one negative. Since the three overflow fluxes are presumably minimized in wild-type plants and thus set to a small positive number in the original FBA-derived optimum $\mathbf{v}^{wt}$, any perturbation $\mathbf{w}$ involving a non-trivial linear combination of *BV*3 and *BV*4 cannot be a solution for the system described in (B.3) because adding a negative value to one of these overflow fluxes would make it smaller than the lower bound. Thus, a valid perturbation $\mathbf{w}$ can be represented as:

$$\begin{aligned} &\mathbf{w} = \alpha_1 BV1 + \alpha_2 BV2 \\ &\text{subject to } \mathbf{1} \le \mathbf{v}^{wt} + \mathbf{w} \le \mathbf{u} \text{ and } \alpha_1, \alpha_2 \in \mathbf{R} \end{aligned} \tag{B.6}$$

The two sets of equivalent pathways as specified by *BV*1 and *BV*2 are ($v_{10} \rightarrow v_{12}$, $v_{11} \rightarrow v_{13}$) and ($v_{14} \rightarrow v_{20} \rightarrow v_{21}$, $v_{16} \rightarrow v_{17} \rightarrow v_{18}$). To identify a unique, physiologically relevant flux distribution for wild-type plants, we used the maximum activities of two *Medicago* CCR isoforms (Table B.5) to constrain the first two equivalent pathways with the following constraint: $v_{10} / v_{13} = 0.35 / 1.64$. The constraint is justified because, assuming that the two CCR-catalyzed reactions are described by Michaelis-Menten kinetics and that the levels of both CoA esters are well below the corresponding Michaelis constraints (54.5 μM for feruloyl CoA and 23.4 μM for caffeoyl CoA; [59]), the ratio between the two CoA esters is approximately

$$\frac{[\text{Caffeoyl CoA}]}{[\text{Feruloyl CoA}]} \cong \frac{v_{10}}{v_{13}} \times \frac{1.64}{0.35} \times \frac{23.4}{54.5} \cong 0.43, \tag{B.7}$$

which is consistent with the prediction in potato tubers that feruloyl CoA is more abundant than caffeoyl CoA [232].

Since the enzymes implicated in the other two equivalent pathways have not yet been characterized for *Medicago*, we instead used the maximum activities of *Arabidopsis* F5H to set up the constraint: $v_{16} / v_{20} = 5 / 6$. Notice that this approximation is not an important issue because all the main results and postulates still hold whether or not the later constraint is applied (data not shown).

**Table B.5: Documented enzyme kinetic constants for CCR and F5H.**

| Enzyme | Gene | Substrate | $V_{max}$ | Reference |
|---|---|---|---|---|
| Cinnamoyl CoA reductase (CCR) | MtCCR1 | Feruloyl CoA | 1.64[a] | [59] |
| | MtCCR2 | Caffeoyl CoA | 0.35[a] | |
| Ferulate 5-hydroxylase (F5H) | FAH1[c] | Coniferyl aldehyde | 5[b] | [10] |
| | | Coniferyl alcohol | 6[b] | |

[a]Unit in μmol/min
[b]Unit in pkat/mg; kat = mol/s
[c]The gene encoding ferulate 5-hydroxulase was cloned in *Arabidopsis*

Interestingly, the three major monolignols (H, G, and S) are not involved in the basis vectors. A possible reason is the following: The three fluxes $v_6$, $v_{15}$, and $v_{19}$ are more or less fixed by the normalization ($v_1 = 1$) and the two "proportion" constraints in Eq. (3.3), if the task is to maximize their sum (or equivalently, to minimize the sum of three "overflow" fluxes). As a result, their values would not be influenced by the different weighting of equivalent pathways, whereas values of some other intermediate fluxes would.

## B.3 Kinetic Analysis of a Reduced Model

In order to validate the results from the flux-based analysis in some independent fashion, we generated an ensemble of ordinary differential equation (ODE) models for the core of the pathway (Figure B.2) that controls the relative proportion of G and S lignin. Using a standard formulation with simplified variable names and Michaelis-Menten functions for each enzymatic step, we defined

$$
\begin{aligned}
\frac{dX_1}{dt} &= \frac{V_1 I}{I + K_1} - \frac{V_3 X_1}{X_1 + K_3} \\
\frac{dX_2}{dt} &= \frac{V_2 I}{I + K_2} - \frac{V_4 X_2}{X_2 + K_4} \\
\frac{dX_3}{dt} &= \frac{V_3 X_1}{X_1 + K_3} + \frac{V_4 X_2}{X_2 + K_4} - \frac{V_5 X_3}{X_3 + K_5} - \frac{V_7 X_3}{X_3 + K_7} \\
\frac{dX_4}{dt} &= \frac{V_5 X_3}{X_3 + K_5} - \frac{V_6 X_4}{X_4 + K_6} - \frac{V_8 X_4}{X_4 + K_8}
\end{aligned}
\tag{B.8}
$$

where $K_i$'s are Michaelis constants and $V_i$'s are maximum rates. To ensure that the search was representative of the parameter space, we sampled 10,000 sets of kinetic parameters uniformly over logarithmic scales, using the Latin hypercube sampling method. The sampling ranges were $V_i$ ~0.1-10 and $K_i$ ~0.1-10. Furthermore, in order to account for the possibility of cooperative binding, we replaced $V_i S/(S+K_i)$ in Eq. (B.8) with Hill functions of the type $V_i S^n/(S^n+K_i^n)$ and sampled the Hill coefficient $n$ from the range 1-4.

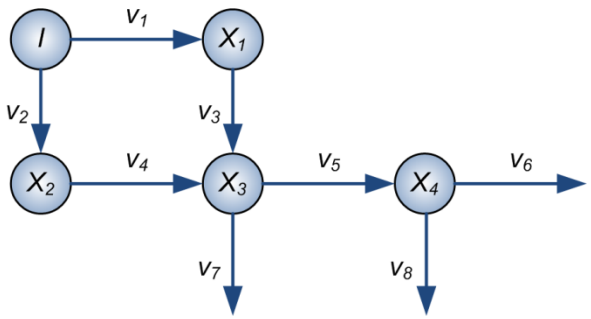


**Figure B.2: Simplified network with one fixed input (*I*) and four metabolites ($X_1$-$X_4$), which was used as a reduced model for studying the roles of CCR1 and CCR2 in the monolignol pathway.**
Metabolic fluxes, denoted as $v_1$-$v_8$, are represented by arrows that connect metabolites or leave the system. Each kinetic parameter in Eq. (B.8) is numbered by the corresponding flux. Reactions $v_1,\ldots,v_5$ correspond to CCR2, CCoAOMT, COMT, CCR1, and CAD, respectively. $v_6$ represents transport into the cell wall, and $v_7$ and $v_8$ represent F5H. Pools $I$, $X_1$, …, $X_4$ correspond to caffeoyl-CoA, caffeoyl aldehyde, feruloyl CoA, coniferyl aldehyde, and conferyl alcohol, respectively.

Each sampled parameter set defines a kinetic model with which we can simulate different cases of genetic modifications and monitor how the S/G ratio responds. First, we numerically determined a steady state by solving the ODEs with all dependent variables in the network, as well as the input *I*, set to a concentration of 1. Gene modifications were modeled by decreasing the $V_i$ of the targeted enzyme (*e.g.*, $V_2$ for CCoAOMT). With this adjustment, we solved the ODEs again and then computed the S/G ratio as

$$\mathrm{S/G}\ ratio \approx \frac{\bar{v}_7 + \bar{v}_8}{\bar{v}_6} = \frac{\dfrac{V_7 \bar{X}_3}{\bar{X}_3 + K_7} + \dfrac{V_8 \bar{X}_4}{\bar{X}_4 + K_8}}{\dfrac{V_6 \bar{X}_4}{\bar{X}_4 + K_6}}, \tag{B.9}$$

where variables with bars indicate steady-state values. The further analysis excluded ill-behaved models, which were defined as systems spending an unduly large amount of time approaching the post-modification steady state, or systems in which one or more metabolites were depleted during the transition. The remaining admissible models were evaluated for their ability to change the S/G ratio; an increase in the S/G ratio was deemed significant if it was greater than 50%.

# APPENDIX C

# SUPPLEMENARY MATERIALS FOR CHAPTER 5

## C.1 Supplementary Text

### C.1.1 Selection of target tissue in a wild-type *Medicago* species

The parameter values for each model instantiation were selected in such a way that the nominal steady state is representative of wild-type *Medicago*. In this study, we chose alfalfa (*Medicago sativa* L.) as the model organism because of its extensive depository of perturbation-response data, including the results of experiments in which seven lignin biosynthetic enzymes were genetically down-regulated and the lignin content and composition in several stem internodes of each down-regulated line were determined [25]. Of note, this list of down-regulated genes does not include CCR1 and CCR2, which have only recently been analyzed with *Medicago truncatula* lines harboring transposon insertions in CCR1 and CCR2 [59]. In order to minimize the discrepancy in our biological context, we thus chose the sixth internode (numbered from top to bottom) of stem as the target tissue because this is where the lignin content and composition were determined for the *ccr1* and *ccr2* mutants [59].

### C.1.2 Physiochemical constraints on steady-state fluxes

Two pieces of information for this specific stem internode in a wild-type alfalfa plant can be exploited, along with other stoichiometric and thermodynamic constraints, to define a biologically realistic set of reaction rates (or fluxes) at the nominal steady state. First, wild-type alfalfa is known to contain principally S and G lignin, while the incorporation of 5-hydroxyconiferyl alcohol into lignin polymer only occurs in COMT-

deficient plants [103,107]. Thus, it is reasonable to assume that the target tissue in a wild-type alfalfa plant has evolved to maximize the production of G and S lignin at the expense of 5-hydroxyguaiacyl (5HG) lignin. Second, the S/G ratio, that is, the ratio of sinapyl (S) to guaiacyl (G) lignin monomers, is equal to 0.58 [25]. As in our previous work [150], this information can be translated into a "proportionality constraint" on the fluxes leading to G and S lignin. Combining this information, we can represent the set *P* of steady-state fluxes, defined as *m*-dimensional real vectors, in the following mathematical format

$$P = \{v \mid \mathbf{c}^T\mathbf{v} = f^*,\ \mathbf{b}^T\mathbf{v} = 0,\ \mathbf{Nv} = \mathbf{0},\ v_1 = 1,\ v_i \geq l_i,\ i = 1,...,m\}. \tag{C.1}$$

The definition for each of the five conditions is listed below:

1. $\mathbf{c}^T\mathbf{v} = f^*$: This condition states that the sum of fluxes leading to G and S lignin ($\mathbf{c}^T\mathbf{v}$) should be fixed at a value $f^*$, which is obtained by solving the following linear programming problem:

$$\begin{aligned} f^* = \max\ & \mathbf{c}^T\mathbf{v} \\ \text{subject to } & \mathbf{b}^T\mathbf{v} = 0 \\ & \mathbf{Nv} = 0 \\ & v_1 = 1 \\ & v_i \geq l_i \\ & i = 1,...,m \end{aligned} \tag{C.2}$$

2. $\mathbf{b}^T\mathbf{v} = 0$: This equation defines the proportionality constraint on the fluxes leading to G and S lignin as described above. Elements in **b** are determined by the specific value of the S/G ratio.
3. $\mathbf{Nv} = \mathbf{0}$: This condition describes the conservation of mass, or mass balance. **N** is an $n \times m$ stoichiometric matrix for a given design with *n* dependent variables and *m* reactions.

4. $v_1 = 1$: As no reaction in the model is known to be reversible, with the exception of HCT, setting the input flux ($v_1$) to 1 ensures that all fluxes are less than or equal to one. In other words, this condition works as a means of standardization.
5. $v_i \geq l_i$: This condition defines the lower bounds on individual reactions. For the $i^{th}$ flux $v_i$, it is bounded from below by $l_i$. Here, we assume that all the enzymatic reactions and transport processes are irreversible and thus have a lower bound of zero. The only exception is the process that represents the transport of 5-hydroxyconiferyl alcohol into the cell wall, for which we arbitrarily choose 0.01 as the lower bound to prevent its value from becoming too small when solving for $f^*$.

## C.2 Supplementary Tables and Figures

**Table C.1: Number of valid model instantiations as judged by two different robustness measures (*Q* and *Q*').** Statistics with a non-zero value of *Q* or *Q*' are marked in bold. The first number is the result of a simulation with Mechanism 3 only, whereas the second number stems from a simulation with both Mechanisms, 1 and 3.

| Configuration # | *Q* | *Q*' |
|---|---|---|
| A | **96/100** | **218/222** |
| B | **284/303** | **461/486** |
| C | 0/0 | 0/0 |
| D | 0/0 | 0/0 |
| E | **307/361** | **489/555** |
| F | **176/180** | **278/273** |
| G | 0/0 | 0/0 |
| H | 0/0 | 0/**4** |
| I | **431/422** | **626/649** |
| J | 0/0 | 0/0 |
| K | 0/0 | 0/0 |
| L | 0/0 | 0/0 |
| M | 0/0 | 0/0 |
| N | 0/0 | 0/0 |
| O | **381/447** | **619/662** |
| P | 0/0 | 0/0 |
| Q | 0/0 | 0/0 |
| R | 0/0 | 0/0 |
| S | 0/0 | 0/0 |

$Q$ = # models showing a decrease of more than 5% and an increase of more than 5%, compared to the wild-type level, in simulations of CCR1 and CCR2 down-regulation, respectively.
$Q$' = # models showing a decreased and an increased S/G ratio, compared to the wild-type level, in simulations of CCR1 and CCR2 down-regulation, respectively.

**Table C.2: Upper and lower bounds for kinetic orders.**

| Kinetic order ($f_{\text{enzyme, substrate/regulator}}$) | Lower bound | Upper bound |
|---|---|---|
| $f_{\text{CCR2, caffeoyl CoA}}$ | 0 | 2[a] |
| $f_{\text{CCoAOMT, caffeoyl CoA}}$ | 0 | 1[b] |
| $f_{\text{COMT, caffeyl aldehyde}}$ | 0 | 1 |
| $f_{\text{COMT/F5H, caffeyl aldehyde}}$ | 0 | 1 |
| $f_{\text{CCR1, feruloyl CoA}}$ | 0 | 1 |
| $f_{\text{CCR1/CAD, feruloyl CoA}}$ | 0 | 1 |
| $f_{\text{CAD, coniferyl aldehyde}}$ | 0 | 1 |
| $f_{\text{F5H, coniferyl aldehyde}}$ | 0 | 1 |
| $f_{\text{Tr, coniferyl alcohol}}$ | 1[c] | 1 |
| $f_{\text{F5H, coniferyl alcohol}}$ | 0 | 1 |
| $f_{\text{COMT, 5-hydroxy coniferyl aldehyde}}$ | 0 | 1 |
| $f_{\text{Tr, 5-hydroxy coniferyl alcohol}}$ | 1[c] | 1 |
| $f_{\text{COMT, 5-hydroxy coniferyl alcohol}}$ | 0 | 1 |
| all kinetic orders for activators | 0 | 2 |
| all kinetic orders for inhibitors | -2 | 0 |

[a]CCR2 shows positive cooperativity towards caffeoyl-CoA [59]
[b]A kinetic order of 0 corresponds to a Michaelis-Menten process where the enzyme is saturated, and a kinetic order of 1 describes the situation in which the substrate concentration is negligibly small compared to the Michaelis constant $K_M$ [44].
[c]The transport process (Tr) is assumed to be first order

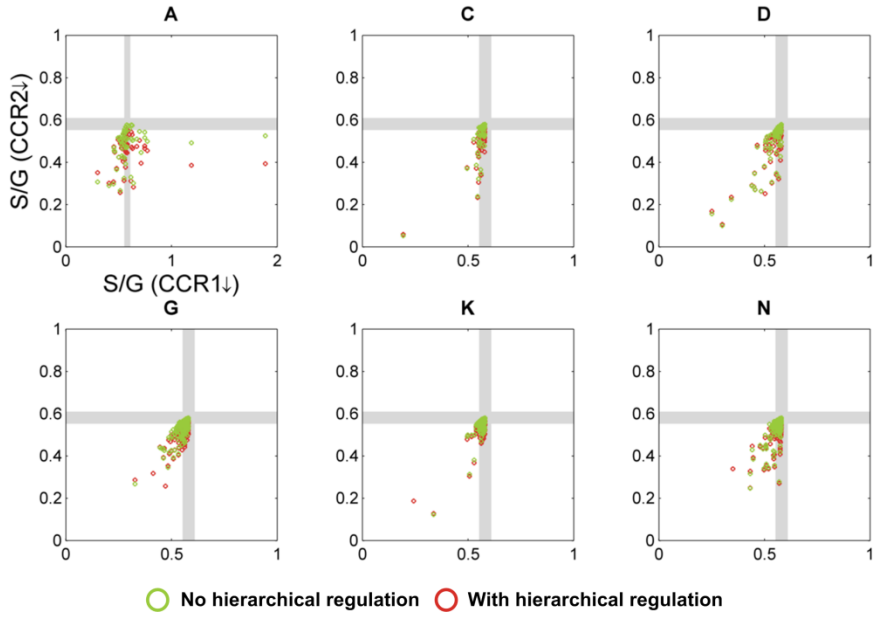


**Figure C.1: Simulation results for CCR1 and CCR2 down-regulation using only Mechanism 1.**
As with Figures 5.3 and 5.4, only topological configurations with at least one model showing quantitatively correct predictions for both CCoAOMT and COMT down-regulation are plotted.

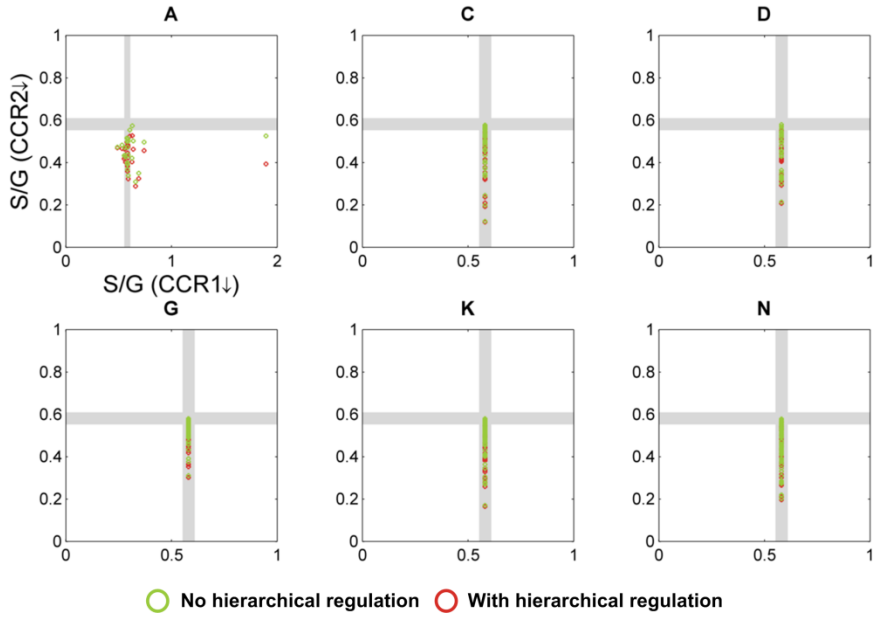


**Figure C.2: Simulation results for CCR1 and CCR2 down-regulation using only Mechanism 2.**
See legend of Figure C.1 for more details.

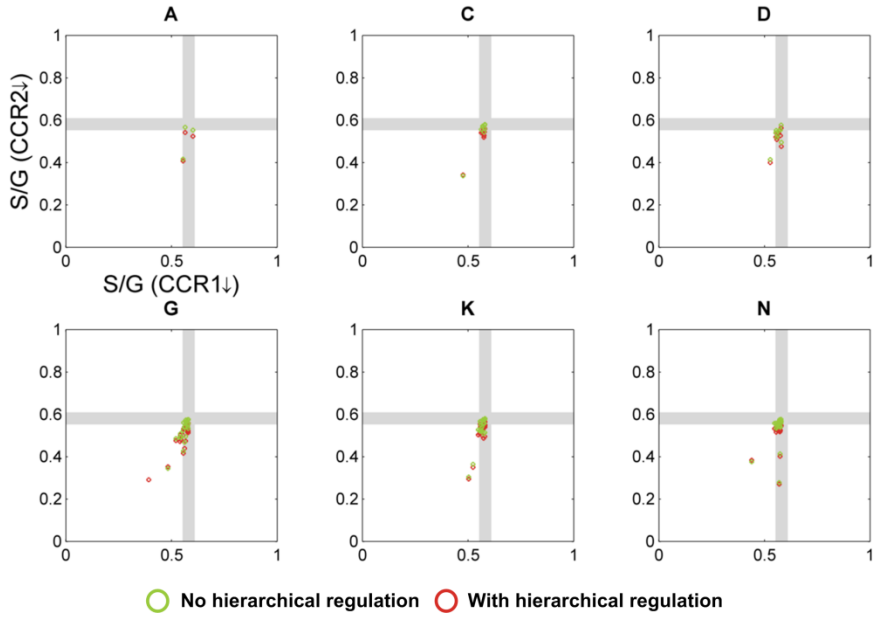


**Figure C.3: Simulation results for CCR1 and CCR2 down-regulation using Mechanisms 1 and 2.**
See legend of Figure C.1 for more details.

# REFERENCES


1. Voit EO (2003) Biochemical and genomic regulation of the trehalose cycle in yeast: review of observations and canonical model analysis. *J Theor Biol* **223**: 55-78.

2. Fairley P (2011) Introduction: Next generation biofuels. *Nature* **474**: S2-S5.

3. Sanderson K (2011) Lignocellulose: A chewy problem. *Nature* **474**: S12-S14.

4. Palmqvist E, Hahn-Hägerdal B, Galbe M, Zacchi G (1996) The effect of water-soluble inhibitors from steam-pretreated willow on enzymatic hydrolysis and ethanol fermentation. *Enzyme Microb Technol* **19**: 470-476.

5. Mosier N, Wyman C, Dale B, Elander R, Lee YY, et al. (2005) Features of promising technologies for pretreatment of lignocellulosic biomass. *Bioresour Technol* **96**: 673-686.

6. Davison BH, Drescher SR, Tuskan GA, Davis MF, Nghiem NP (2006) Variation of S/G ratio and lignin content in a *Populus* family influences the release of xylose by dilute acid hydrolysis. *Appl Biochem Biotechnol* **130**: 427-435.

7. Dien BS, Jung H-JG, Vogel KP, Casler MD, Lamb JAFS, et al. (2006) Chemical composition and response to dilute-acid pretreatment and enzymatic saccharification of alfalfa, reed canarygrass, and switchgrass. *Biomass Bioenergy* **30**: 880-891.

8. Fu C, Mielenz JR, Xiao X, Ge Y, Hamilton CY, et al. (2011) Genetic manipulation of lignin reduces recalcitrance and improves ethanol production from switchgrass. *Proc Natl Acad Sci U S A* **108**: 3803-3808.

9. Dixon RA, Chen F, Guo D, Parvathi K (2001) The biosynthesis of monolignols: a "metabolic grid", or independent pathways to guaiacyl and syringyl units? *Phytochemistry* **57**: 1069-1084.

10. Humphreys JM, Hemm MR, 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.

11. Achnine L, Blancaflor EB, Rasmussen S, Dixon RA (2004) Colocalization of L-phenylalanine ammonia-lyase and cinnamate 4-hydroxylase for metabolic channeling in phenylpropanoid biosynthesis. *Plant Cell* **16**: 3098-3109.

12. Winkel BSJ (2004) Metabolic channeling in plants. *Annu Rev Plant Biol* **55**: 85-107.

13. Zhong R, Ye Z-H (2007) Regulation of cell wall biosynthesis. *Curr Opin Plant Biol* **10**: 564-572.

14. Boerjan W, Ralph J, Baucher M (2003) Lignin Biosynthesis. *Annu Rev Plant Biol* **54**: 519-546.

15. Weng J-K, Chapple C (2010) The origin and evolution of lignin biosynthesis. *New Phytol* **187**: 273-285.

16. The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796-815.

17. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, et al. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596-1604.

18. Humphreys JM, Chapple C (2002) Rewriting the lignin roadmap. *Curr Opin Plant Biol* **5**: 224-229.

19. Ehlting J, Büttner D, Wang Q, Douglas CJ, Somssich IE, et al. (1999) Three 4 coumarate:coenzyme A ligases in *Arabidopsis thaliana* represent two evolutionarily divergent classes in angiosperms. *Plant J* **19**: 9-20.

20. Hu W-J, Kawaoka A, Tsai C-J, Lung J, Osakabe K, et al. (1998) Compartmentalized expression of two structurally and functionally distinct 4-coumarate: CoA ligase genes in aspen (*Populus tremuloides*). *Proc Natl Acad Sci U S A* **95**: 5407-5412.

21. Lauvergeat V, Lacomme C, Lacombe E, Lasserre E, Roby D, et al. (2001) Two cinnamoyl-CoA reductase (CCR) genes from *Arabidopsis thaliana* are differentially expressed during development and in response to infection with pathogenic bacteria. *Phytochemistry* **57**: 1187-1195.

22. Lindermayr C, Möllers B, Fliegmann J, Uhlmann A, Lottspeich F, et al. (2002) Divergent members of a soybean (*Glycine max* L.) 4 coumarate:coenzyme A ligase gene family. *Eur J Biochem* **269**: 1304-1315.

23. Chen C, Meyermans H, Burggraeve B, De Rycke RM, Inoue K, et al. (2000) Cell-specific and conditional expression of caffeoyl-coenzyme A-3-*O*-methyltransferase in poplar. *Plant Physiol* **123**: 853-868.

24. Jørgensen K, Rasmussen AV, Morant M, Nielsen AH, Bjarnholt N, et al. (2005) Metabolon formation and metabolic channeling in the biosynthesis of plant natural products. *Curr Opin Plant Biol* **8**: 280-291.

25. Chen F, Srinivasa Reddy MS, Temple S, Jackson L, Shadle G, et al. (2006) Multi-site genetic modulation of monolignol biosynthesis suggests new routes for formation of syringyl lignin and wall-bound ferulic acid in alfalfa (*Medicago sativa L.*). *Plant J* **48**: 113-124.

26. Chen F, Duran AL, Blount JW, Sumner LW, Dixon RA (2003) Profiling phenolic metabolites in transgenic alfalfa modified in lignin biosynthesis. *Phytochemistry* **64**: 1013-1021.

27. Chou I-C, Voit EO (2009) Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Math Biosci* **219**: 57-83.

28. Schubert C (2011) Single-cell analysis: the deepest differences. *Nature* **480**: 133-137.

29. Bendall SC, Simonds EF, Qiu P, Amir ED, Krutzik PO, et al. (2011) Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**: 687-696.

30. Schilling CH, Letscher D, Palsson BØ (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J Theor Biol* **203**: 229-248.

31. Schuster S, Dandekar T, Fell DA (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol* **17**: 53-60.

32. Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? *Nat Biotechnol* **28**: 245-248.

33. Palsson BØ (2006) Systems biology: properties of reconstructed networks. Cambridge, UK: Cambridge Univ Press.

34. Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol Syst Biol* **3**: 119.

35. Schuetz R, Zamboni N, Zampieri M, Heinemann M, Sauer U (2012) Multidimensional optimality of microbial metabolism. *Science* **336**: 601-604.

36. Price ND, Reed JL, Palsson BØ (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* **2**: 886-897.

37. Feist AM, Palsson BØ (2008) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat Biotechnol* **26**: 659-667.

38. Oberhardt MA, Palsson BØ, Papin JA (2009) Applications of genome-scale metabolic reconstructions. *Mol Syst Biol* **5**: 320.

39. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, et al. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A* **104**: 1777-1782.

40. Ma H, Sorokin A, Mazein A, Selkov A, Selkov E, et al. (2007) The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol* **3**: 135.

41. de Oliveira Dal'Molin CG, Quek L-E, Palfreyman RW, Brumbley SM, Nielsen LK (2010) C4GEM, a genome-scale metabolic model to study $C_4$ plant metabolism. *Plant Physiol* **154**: 1871-1885.

42. de Oliveira Dal'Molin CG, Quek L-E, Palfreyman RW, Brumbley SM, Nielsen LK (2010) AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. *Plant Physiol* **152**: 579-589.

43. Savageau MA (1976) Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology. Reading, MA: Addison Wesley Publishing Company.

44. Voit EO (2000) Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists. Cambridge, UK: Cambridge Univ Press.

45. Ni T-C, Savageau MA (1996) Model assessment and refinement using strategies from biochemical systems theory: application to metabolism in human red blood cells. *J Theor Biol* **179**: 329-368.

46. Alvarez-Vasquez F, Cánovas M, Iborra JL, Torres NV (2002) Modeling, optimization and experimental assessment of continuous L-(-)-carnitine production by *Escherichia coli* cultures. *Biotechnol Bioeng* **80**: 794-805.

47. Curto R, Voit EO, Sorribas A, Cascante M (1998) Mathematical models of purine metabolism in man. *Math Biosci* **151**: 1-49.

48. Alvarez-Vasquez F, Riezman H, Hannun YA, Voit EO (2011) Mathematical modeling and validation of the ergosterol pathway in *Saccharomyces cerevisiae*. *PLoS One* **6**: e28344.

49. Alvarez-Vasquez F, Sims KJ, Cowart LA, Okamoto Y, Voit EO, et al. (2005) Simulation and validation of modelled sphingolipid metabolism in *Saccharomyces cerevisiae*. *Nature* **433**: 425-430.

50. Alvarez-Vasquez F, Sims KJ, Hannun YA, Voit EO (2004) Integration of kinetic information on yeast sphingolipid metabolism in dynamical pathway models. *J Theor Biol* **226**: 265-291.

51. Qi Z, Miller GW, Voit EO (2008) Computational systems analysis of dopamine metabolism. *PLoS One* **3**: e2444.

52. Voit EO, Sands PJ (1996) Modeling forest growth. I. Canonical approach. *Ecol Modell* **86**: 51-71.

53. Voit EO, Sands PJ (1996) Modeling forest growth. II. Biomass partitioning in Scots pine. *Ecol Modell* **86**: 73-89.

54. Voit EO (1988) Dynamics of self-thinning plant stands. *Ann Bot* **62**: 67-78.

55. Covert MW, Xiao N, Chen TJ, Karr JR (2008) Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. *Bioinformatics* **24**: 2044-2050.

56. Covert MW, Palsson BØ (2002) Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *J Biol Chem* **277**: 28058-28064.

57. Covert MW, Schilling CH, Palsson B (2001) Regulation of gene expression in flux balance models of metabolism. *J Theor Biol* **213**: 73-88.

58. Kremling A, Bettenbrock K, Gilles E (2007) Analysis of global control of *Escherichia coli* carbohydrate uptake. *BMC Syst Biol* **1**: 42.

59. Zhou R, Jackson L, Shadle G, Nakashima J, Temple S, et al. (2010) Distinct cinnamoyl CoA reductases involved in parallel routes to lignin in *Medicago truncatula*. *Proc Natl Acad Sci USA* **107**: 17803-17808.

60. Parvathi K, Chen F, Guo D, Blount JW, Dixon RA (2001) Substrate preferences of *O*-methyltransferases in alfalfa suggest new pathways for 3-*O*-methylation of monolignols. *Plant J* **25**: 193-202.

61. Leplé JC, Dauwe R, Morreel K, Storme V, Lapierre C, et al. (2007) Downregulation of cinnamoyl-coenzyme A reductase in poplar: multiple-level phenotyping reveals effects on cell wall polymer metabolism and structure. *Plant Cell* **19**: 3669-3691.

62. Morreel K, Ralph J, Lu F, Goeminne G, Busson R, et al. (2004) Phenolic profiling of caffeic acid O-methyltransferase-deficient poplar reveals novel benzodioxane oligolignols. *Plant Physiol* **136**: 4023-4036.

63. Famili I, Forster J, Nielsen J, Palsson BØ (2003) *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc Natl Acad Sci U S A* **100**: 13134-13139.

64. Förster J, Famili I, Fu P, Palsson BØ, Nielsen J (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res* **13**: 244-253.

65. Smallbone K, Simeonidis E, Broomhead DS, Kell DB (2007) Something from nothing - bridging the gap between constraint-based and kinetic modelling. *FEBS J* **274**: 5576-5585.

66. Campbell MM, Sederoff RR (1996) Variation in lignin content and composition (mechanisms of control and implications for the genetic improvement of plants). *Plant Physiol* **110**: 3-13.

67. Baucher M, Chabbert B, Pilate G, Van Doorsselaere J, Tollier MT, et al. (1996) Red xylem and higher lignin extractability by down-regulating a cinnamyl alcohol dehydrogenase in poplar. *Plant Physiol* **112**: 1479-1490.

68. Harding SA, Leshkevich J, Chiang VL, Tsai C-J (2002) Differential substrate inhibition couples kinetically distinct 4-coumarate: coenzyme A ligases with spatially distinct metabolic roles in quaking aspen. *Plant Physiol* **128**: 428-438.

69. Blount JW, Korth KL, Masoud SA, Rasmussen S, Lamb C, et al. (2000) Altering expression of cinnamic acid 4-hydroxylase in transgenic plants provides evidence for a feedback loop at the entry point into the phenylpropanoid pathway. *Plant Physiol* **122**: 107-116.

70. Li L, Cheng X, Lu S, Nakatsubo T, Umezawa T, et al. (2005) Clarification of cinnamoyl co-enzyme A reductase catalysis in monolignol biosynthesis of aspen. *Plant Cell Physiol* **46**: 1073-1082.

71. Li L, Popko JL, Umezawa T, Chiang VL (2000) 5-hydroxyconiferyl aldehyde modulates enzymatic methylation for syringyl monolignol formation, a new view of monolignol biosynthesis in angiosperms. *J Biol Chem* **275**: 6537-6545.

72. Osakabe K, Tsao CC, Li L, Popko JL, Umezawa T, et al. (1999) Coniferyl aldehyde 5-hydroxylation and methylation direct syringyl lignin biosynthesis in angiosperms. *Proc Natl Acad Sci U S A* **96**: 8955-8960.

73. Pilate G, Guiney E, Holt K, Petit-Conil M, Lapierre C, et al. (2002) Field and pulping performances of transgenic trees with altered lignification. *Nat Biotechnol* **20**: 607-612.

74. Li L, Zhou Y, Cheng X, Sun J, Marita JM, et al. (2003) Combinatorial modification of multiple lignin traits in trees through multigene cotransformation. *Proc Natl Acad Sci U S A* **100**: 4939-4944.

75. Meyermans H, Morreel K, Lapierre C, Pollet B, De Bruyn A, et al. (2000) Modifications in lignin and accumulation of phenolic glucosides in poplar xylem upon down-regulation of caffeoyl-coenzyme A O-methyltransferase, an enzyme involved in lignin biosynthesis. *J Biol Chem* **275**: 36899-36909.

76. Vanholme R, Morreel K, Ralph J, Boerjan W (2008) Lignin engineering. *Curr Opin Plant Biol* **11**: 278-285.

77. Chapple CCS, Vogt T, Ellis BE, Somerville CR (1992) An *Arabidopsis* mutant defective in the general phenylpropanoid pathway. *Plant Cell* **4**: 1413-1424.

78. Lindroth RL, Hwang SY. Diversity, redundancy and multiplicity in chemical defense systems of aspen. In: Romeo JT, Saunders JA, Barbosa P, editors; 1996; New York, NY, USA. Plenum Press. pp. 25-54.

79. Coleman HD, Park J-Y, Nair R, Chapple CCS, Mansfield SD (2008) RNAi-mediated suppression of *p*-coumaroyl-CoA 3'-hydroxylase in hybrid poplar impacts lignin deposition and soluble secondary metabolism. *Proc Natl Acad Sci U S A* **105**: 4501-4506.

80. Vilela M, Chou I-C, Vinga S, Vasconcelos ATR, Voit EO, et al. (2008) Parameter optimization in S-system models. *BMC Syst Biol* **2**: 35.

81. Murphy JM, Sexton DMH, Barnett DN, Jones GS, Webb MJ, et al. (2004) Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature* **430**: 768-772.

82. Battogtokh D, Asch DK, Case ME, Arnold J, Schüttler HB (2002) An ensemble method for identifying regulatory circuits with special reference to the qa gene cluster of *Neurospora crassa*. *Proc Natl Acad Sci U S A* **99**: 16904-16909.

83. Kuepfer L, Peter M, Sauer U (2007) Ensemble modeling for analysis of cell signaling dynamics. *Nat Biotechnol* **25**: 1001-1006.

84. Bloom JD, Meyer MM, Meinhold P, Otey CR, MacMillan D, et al. (2005) Evolving strategies for enzyme engineering. *Curr Opin Struct Biol* **15**: 447-452.

85. Torres NV, Voit EO, Glez-Alcón C, Rodriguez F (1997) An indirect optimization method for biochemical systems: description of method and application to the maximization of the rate of ethanol, glycerol, and carbohydrate production in *Saccharomyces cerevisiae*. *Biotechnol Bioeng* **55**: 758-772.

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

87. Barabási A-L, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**: 101-113.

88. Goel G, Chou I-C, Voit EO (2006) Biological systems modeling and analysis: a biomolecular technique of the twenty-first century. *J Biomol Tech* **17**: 252-269.

89. Alves R, Herrero E, Sorribas A (2004) Predictive reconstruction of the mitochondrial iron-sulfur cluster assembly metabolism: I. The role of the protein pair ferredoxin-ferredoxin reductase (Yah1-Arh1). *Proteins* **56**: 354-366.

90. Halpin C (2005) Gene stacking in transgenic plants - the challenge for 21st century plant biotechnology. *Plant Biotechnol J* **3**: 141-155.

91. Ratcliffe RG, Shachar-Hill Y (2006) Measuring multiple fluxes through plant metabolic networks. *Plant J* **45**: 490-511.

92. Hertzberg M, Aspeborg H, Schrader J, Andersson A, Erlandsson R, et al. (2001) A transcriptional roadmap to wood formation. *Proc Natl Acad Sci U S A* **98**: 14732-14737.

93. Karpinska B, Karlsson M, Srivastava M, Stenberg A, Schrader J, et al. (2004) MYB transcription factors are differentially expressed and regulated during secondary vascular tissue development in hybrid aspen. *Plant Mol Biol* **56**: 255-270.

94. Kawaoka A, Kaothien P, Yoshida K, Endo S, Yamada K, et al. (2000) Functional analysis of tobacco LIM protein Ntlim1 involved in lignin biosynthesis. *Plant J* **22**: 289-301.

95. Segrè D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A* **99**: 15112-15117.

96. Stephanopoulos GN, Vallino JJ (1991) Network rigidity and metabolic engineering in metabolite overproduction. *Science* **252**: 1675-1681.

97. Tesfaye M, Yang SS, Lamb JFS, Jung H-JG, Samac DA, et al. (2009) *Medicago truncatula* as a model for dicot cell wall development. *Bioenergy Res* **2**: 59-76.

98. León J, Shulaev V, Yalpani N, Lawton MA, Raskin I (1995) Benzoic acid 2-hydroxylase, a soluble oxygenase from tobacco, catalyzes salicylic acid biosynthesis. *Proc Natl Acad Sci U S A* **92**: 10413-10417.

99. Mauch-Mani B, Slusarenko AJ (1996) Production of salicylic acid precursors is a major function of phenylalanine ammonia-lyase in the resistance of *Arabidopsis* to *Peronospora parasitica*. *Plant Cell* **8**: 203-212.

100. Yalpani N, León J, Lawton MA, Raskin I (1993) Pathway of salicylic acid biosynthesis in healthy and virus-inoculated tobacco. *Plant Physiol* **103**: 315-321.

101. Wildermuth MC, Dewdney J, Wu G, Ausubel FM (2001) Isochorismate synthase is required to synthesize salicylic acid for plant defence. *Nature* **414**: 562-565.

102. Dixon RA, Paiva NL (1995) Stress-induced phenylpropanoid metabolism. *Plant Cell* **7**: 1085-1097.

103. Marita JM, Ralph J, Hatfield RD, Guo D, Chen F, et al. (2003) Structural and compositional modifications in lignin of transgenic alfalfa down-regulated in caffeic acid 3-*O*-methyltransferase and caffeoyl coenzyme A 3-*O*-methyltransferase. *Phytochemistry* **62**: 53-65.

104. Howles PA, Sewalt VJH, Paiva NL, Elkind Y, Bate NJ, et al. (1996) Overexpression of L-phenylalanine ammonia-lyase in transgenic tobacco plants reveals control points for flux into phenylpropanoid biosynthesis. *Plant Physiol* **112**: 1617-1624.

105. Hoffmann L, Maury S, Martz F, Geoffroy P, Legrand M (2003) Purification, cloning, and properties of an acyltransferase controlling shikimate and quinate ester intermediates in phenylpropanoid metabolism. *J Biol Chem* **278**: 95-103.

106. Sarni F, Grand C, Boudet AM (1984) Purification and properties of cinnamoyl-CoA reductase and cinnamyl alcohol dehydrogenase from poplar stems (*Populus* X *euramericana*). *Eur J Biochem* **139**: 259-265.

107. Guo D, Chen F, Inoue K, Blount JW, Dixon RA (2001) Downregulation of caffeic acid 3-O-methyltransferase and caffeoyl CoA 3-*O*-methyltransferase in transgenic alfalfa: Impacts on lignin structure and implications for the biosynthesis of G and S lignin. *Plant Cell* **13**: 73-88.

108. Jackson LA, Shadle GL, Zhou R, Nakashima J, Chen F, et al. (2008) Improving saccharification efficiency of alfalfa stems through modification of the terminal stages of monolignol biosynthesis. *Bioenergy Res* **1**: 180-192.

109. Chapple C (1998) Molecular-genetic analysis of plant cytochrome P450-dependent monooxygenases. *Annu Rev Plant Biol* **49**: 311-343.

110. Guo D, Chen F, Dixon RA (2002) Monolignol biosynthesis in microsomal preparations from lignifying stems of alfalfa (*Medicago sativa* L.). *Phytochemistry* **61**: 657-667.

111. Chen Z, Zheng Z, Huang J, Lai Z, Fan B (2009) Biosynthesis of salicylic acid in plants. *Plant Signal Behav* **4**: 493-496.

112. Gallego-Giraldo L, Jikumaru Y, Kamiya Y, Tang Y, Dixon RA (2011) Selective lignin downregulation leads to constitutive defense response expression in alfalfa (*Medicago sativa* L.). *New Phytol* **190**: 627-639.

113. Mao F, Wu H, Dam P, Chou I-C, Voit EO, et al. (2008) Prediction of biological pathways through data mining and information fusion. In: Xu Y, Gogarten JP, editors. Computational Methods for Understanding Bacterial and Archaeal Genomes. London, UK: Imperial College Press. pp. 281-314.

114. Zubieta C, Kota P, Ferrer J-L, Dixon RA, Noel JP (2002) Structural basis for the modulation of lignin monomer methylation by caffeic acid/5-hydroxyferulic acid 3/5-*O*-methyltransferase. *Plant Cell* **14**: 1265-1277.

115. Bomati EK, Noel JP (2005) Structural and kinetic basis for substrate selectivity in Populus tremuloides sinapyl alcohol dehydrogenase. *Plant Cell* **17**: 1598-1611.

116. Hoffmann L, Maury S, Bergdoll M, Thion L, Erard M, et al. (2001) Identification of the enzymatic active site of tobacco caffeoyl-coenzyme A *O*-methyltransferase by site-directed mutagenesis. *J Biol Chem* **276**: 36831-36838.

117. Huang WE, Huang L, Preston GM, Naylor M, Carr JP, et al. (2006) Quantitative *in situ* assay of salicylic acid in tobacco leaves using a genetically modified biosensor strain of *Acinetobacter* sp. ADP1. *Plant J* **46**: 1073-1083.

118. Huang WE, Wang H, Zheng H, Huang L, Singer AC, et al. (2005) Chromosomally located gene fusions constructed in *Acinetobacter* sp. ADP1 for the detection of salicylate. *Environ Microbiol* **7**: 1339-1348.

119. Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, et al. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res* **32**: D431-D433.

120. Lee S, Phalakornkule C, Domach MM, Grossmann IE (2000) Recursive MILP model for finding all the alternate optima in LP models for metabolic networks. *Comput Chem Eng* **24**: 711-716.

121. Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng* **5**: 264-276.

122. Savageau MA (1985) A theory of alternative designs for biochemical control systems. *Biomed Biochim Acta* **44**: 875-880.

123. Alon U (2006) An Introduction to Systems Biology: Design Principles of Biological Circuits. Boca Raton, FL: Chapman & Hall/CRC.

124. Lipshtat A, Purushothaman SP, Iyengar R, Ma'ayan A (2008) Functions of bifans in context of multiple regulatory motifs in signaling networks. *Biophys J* **94**: 2566-2579.

125. Ma'ayan A, Cecchi GA, Wagner J, Rao AR, Iyengar R, et al. (2008) Ordered cyclic motifs contribute to dynamic stability in biological and engineered networks. *Proc Natl Acad Sci USA* **105**: 19235-19249.

126. Erdős P, Rényi A (1959) On random graphs. *Publicationes Mathematicae* **6**: 290–297.

127. Savageau MA (1969) Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions. *J Theor Biol* **25**: 365-369.

128. Torres NV, Voit EO (2002) Pathway Analysis and Optimization in Metabolic Engineering: Cambridge University Press.

129. Irvine DH (1991) The method of controlled mathematical comparison. In: Voit EO, editor. Canonical Nonlinear Modeling. New York: Van Nostrand Reinhold. pp. 90-109.

130. Irvine DH, Savageau MA (1985) Network regulation of the immune response: Alternative control points for suppressor modulation of effector lymphocytes. *J Immunol* **134**: 2100-2116.

131. Alves R, Savageau MA (2000) Effect of overall feedback inhibition in unbranched biosynthetic pathways. *Biophys J* **79**: 2290-2304.

132. Schwacke JH, Voit EO (2004) Improved methods for the mathematically controlled comparison of biochemical systems. *Theor Biol Med Model* **1**: 1.

133. Savageau MA, Coelho PM, Fasani RA, Tolla DA, Salvador A (2009) Phenotypes and tolerances in the design space of biochemical systems. *Proc Natl Acad Sci U S A* **106**: 6435-6440.

134. Hlavacek WS, Savageau MA (1996) Rules for coupled expression of regulator and effector genes in inducible circuits. *J Mol Biol* **255**: 121-139.

135. Savageau MA (1998) Demand theory of gene regulation. I. Quantitative development of the theory. *Genetics* **149**: 1665-1676.

136. Savageau MA (2001) Design principles for elementary gene circuits: Elements, methods, and examples. *Chaos* **11**: 142-159.

137. Igoshin OA, Alves R, Savageau MA (2008) Hysteretic and graded responses in bacterial two-component signal transduction. *Mol Microbiol* **68**.

138. Beisel CL, Smolke CD (2009) Design principles for riboswitch function. *PLoS Comput Biol* **5**: e1000363.

139. Voit EO (2003) Design principles and operating principles: the yin and yang of optimal functioning. *Math Biosci* **182**: 81-92.

140. Voit EO (2004) Design and operation: Keys to understanding biological systems. In: Deutsch A, Howard J, Falcke M, Zimmermann W, editors. Function and Regulation of Cellular Systems: Experiments and Models. Bael: Birkhäuser-Verlag.

141. Voit EO (2004) The dawn of a new era of metabolic systems analysis. *Drug Discov Today BioSilico* **2**: 182-189.

142. Alvarez-Vasquez F, Sims KJ, Voit EO, Hannun YA (2007) Coordination of the dynamics of yeast sphingolipid metabolism during the diauxic shift. *Theor Biol Med Model* **4**: 42.

143. Alves R, Vilaprinyo E, Hernández-Bermejo B, Sorribas A (2008) Mathematical formalisms based on approximated kinetic representations for modeling genetic and metabolic pathways. *Biotechnol Genet Eng Rev* **25**: 1-40.

144. Guillén-Gosálbez G, Sorribas A (2009) Identifying quantitative operation principles in metabolic pathways: a systematic method for searching feasible enzyme activity patterns leading to cellular adaptive responses. *BMC Bioinformatics* **10**.

145. Vilaprinyo E, Alves R, Sorribas A (2006) Use of physiological constraints to identify quantitative design principles for gene expression in yeast adaptation to heat shock. *BMC Bioinformatics* **7**: 184.

146. Voit EO, Radivoyevitch T (2000) Biochemical systems analysis of genome-wide expression data. *Bioinformatics* **16**: 1023-1037.

147. Voit EO, Almeida JS, Marino S, Lall R, Goel G, et al. (2006) Regulation of glycolysis in *Lactococcus lactis*: An unfinished systems biological case study. *Syst Biol (Stevenage)* **153**: 286-298.

148. Voit EO, Neves AR, Santos H (2006) The intricate side of systems biology. *Proc Natl Acad Sci U S A* **103**: 9452-9457.

149. Navarro E, Montagud A, Fernández de Córdoba P, Urchueguía JF (2009) Metabolic flux analysis of the hydrogen production potential in *Synechocystis sp.* PCC6803. *Int J Hydrogen Energy* **34**: 8828-8838.

150. 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.

151. Lee Y, Voit EO (2010) Mathematical modeling of monolignol biosynthesis in *Populus* xylem. *Math Biosci* **228**: 78-89.

152. Voit EO, Alvarez-Vasquez F, Hannun YA (2009) Computational Analysis of Sphingolipid Pathway Systems. In: Chalfant C, Del Poeta M, editors. Sphingolipids as Signaling and Regulatory Molecules. Austin, TX: Landes Bioscience.

153. Savageau MA (1969) Biochemical systems analysis. II. The steady-state solutions for an n-pool system using a power-law approximation. *J Theor Biol* **25**: 370-379.

154. Savageau MA (1972) The behavior of intact biochemical control systems. *Curr Top Cell Regul* **6**: 63-129.

155. Voit EO (2009) A systems-theoretical framework for health and disease: inflammation and preconditioning from an abstract modeling point of view. *Math Biosci* **217**: 11-18.

156. Penrose R (1955) A generalized inverse for matrices. *Proc Camb Philol Soc* **51**: 406-413.

157. Fonseca LL, Sánchez C, Santos H, Voit EO (2011) Complex coordination of multi-scale cellular responses to environmental stress. *Mol BioSyst* **7**: 731 – 741.

158. Schwacke JH, Voit EO (2005) Computation and analysis of time-dependent sensitivities in Generalized Mass Action systems. *J Theor Biol* **236**: 21-38.

159. Shiraishi F, Hatoh Y, Irie T (2005) An efficient method for calculation of dynamic logarithmic gains in biochemical systems theory. *J Theor Biol* **234**: 79–85.

160. Hatzimanikatis V, Floudas CA, Bailey JE (1996) Analysis and design of metabolic reaction networks via mixed-integer linear optimization. *AIChE J* **42**: 1277-1292.

161. Polisetty PK, Gatzke EP, Voit EO (2008) Yield optimization of regulated metabolic systems using deterministic branch-and-reduce methods. *Biotechnol Bioeng* **99**: 1154-1169.

162. Sands PJ, Voit EO (1996) Flux-based estimation of parameters in S-systems. *Ecol Modeling* **93**: 75-88.

163. Voit EO (1992) Optimization in integrated biochemical systems. *Biotechnol Bioeng* **40**: 572-582.

164. Bentley WE, Mirjalili N, Andersen DC, Davis RH, Kompala DS (1990) Plasmid-encoded protein: the principal factor in the "metabolic burden" associated with recombinant bacteria. *Biotechnol Bioeng* **35**: 668-681.

165. Smits HP, Hauf J, Müller S, T.J. H, Zimmermann FK, et al. (2000) Simultaneous overepression of enzymes of the lower part of glycolysis can enhance the fermentative capacity of *Saccharomyces cerevisiae*. *Yeast* **16**: 1325-1334.

166. Snoep JL, Yomano LP, Westerhoff HV, Ingram LO (1995) Protein burden in *Zymomonas mobilis*: negative flux and growth control due to overproduction of glycolytic enzymes. *Microbiol* **141**: 2329-2337.

167. Görner W, Durchschlag E, Martinez-Pastor MT, Estruch F, Ammerer G, et al. (1998) Nuclear localization of the C2H2 zinc finger protein Msn2p is regulated by stress and protein kinase A activity. *Genes Dev* **12**: 586-597.

168. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **11**: 4241-4257.

169. Postmus J, Canelas AB, Bouwman J, Bakker BM, van Gulik W, et al. (2008) Quantitative analysis of the high temperature-induced glycolytic flux increase in *Saccharomyces cerevisiae* reveals dominant metabolic regulation. *J Biol Chem* **283**: 23524-23532.

170. Ye Y, Zhu Y, Pan L, Li L, Wang X, et al. (2009) Gaining insight into the response logic of *Saccharomyces cerevisiae* to heat shock by combining expression profiles with metabolic pathways. *Biochem Biophys Res Commun* **385**: 357-362.

171. Davidson JF, Whyte B, Bissinger PH, Schiestl RH (1996) Oxidative stress is involved in heat-induced cell death in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* **93**: 5116-5121.

172. Estruch F (2000) Stress-controlled transcription factors, stress-induced genes and stress tolerance in budding yeast. *FEMS Microbiol Rev* **24**: 469-486.

173. Sanchez Y, Taulien J, Borkovich KA, Lindquist S (1992) Hsp104 is required for tolerance to many forms of stress. *EMBO J* **11**: 2357-2364.

174. Cowart LA, Shotwell M, Worley ML, Richards AJ, Montefusco DJ, et al. (2010) Revealing a signaling role of phytosphingosine-1-phosphate in yeast. *Mol Syst Biol* **6**: 349.

175. Hottiger T, Schmutz P, Wiemken A (1987) Heat-induced accumulation and futile cycling of trehalose in *Saccharomyces cerevisiae*. *J Bacteriol* **169**: 5518-5522.

176. Aranda JS, Salgado E, Taillandier P (2004) Trehalose accumulation in *Saccharomyces cerevisiae* cells: experimental data and structured modeling. *Biochem Eng J* **17**: 119-140.

177. Vilaprinyo E, Alves R, Sorribas A (2010) Minimization of biosynthetic costs in adaptive gene expression responses of yeast to environmental changes. *PLoS Comput Biol* **6**: e1000674.

178. Ervadi-Radhakrishnan A, Voit EO (2005) Controllability of non-linear biochemical systems. *Math Biosci* **196**: 99-123.

179. Entian KD, Fröhlich KU, Mecke D (1984) Regulation of enzymes and isoenzymes of carbohydrate metabolism in the yeast *Saccharomyces cerevisiae*. *Biochim Biophys Acta* **799**: 181-186.

180. Thevelein JM, Hohmann S (1995) Trehalose synthase: guard to the gate of glycolysis in yeast? *Trends Biochem Sci* **20**: 3-10.

181. Lotka A (1924) Elements of Physical Biology. Baltimore: Williams and Wilkins; reprinted as 'Elements of Mathematical Biology'. Dover, New York, 1956.

182. Peschel M, Mende W (1986) The Predator-Prey Model: Do we Live in a Volterra World? Berlin: Akademie-Verlag.

183. Volterra V (1926) Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. *Mem R Accad dei Lincei* **2**: 31-113.

184. Hatzimanikatis V, Bailey JE (1996) MCA has more to say. *J Theor Biol* **182**: 233-242.

185. Wu L, Wang W, van Winden WA, van Gulik WM, Heijnen JJ (2004) A new framework for the estimation of control parameters in metabolic pathways using lin-log kinetics. *Eur J Biochem* **271**: 3348-3359.

186. Shiraishi F, Savageau MA (1992) The tricarboxylic-acid cycle in *Dictyostelium discoideum*. 1. Formulation of alternative kinetic representations. *J Biol Chem* **267**: 22912-22918.

187. Vera J, de Atauri P, Cascante M, Torres NV (2003) Multicriteria optimization of biochemical systems by linear programming: application to production of ethanol by *Saccharomyces cerevisiae*. *Biotechnol Bioeng* **83**: 335-343.

188. Weng J-K, Akiyama T, Bonawitz ND, Li X, Ralph J, et al. (2010) Convergent evolution of syringyl lignin biosynthesis via distinct pathways in the lycophyte *Selaginella* and flowering plants. *Plant Cell* **22**: 1033-1045.

189. Weng J-K, Akiyama T, Ralph J, Chapple C (2011) Independent recruitment of an *O*-methyltransferase for syringyl lignin biosynthesis in *Selaginella moellendorffii*. *Plant Cell* **23**: 2708-2724.

190. Weng J-K, Li X, Stout J, Chapple C (2008) Independent origins of syringyl lignin in vascular plants. *Proc Natl Acad Sci USA* **105**: 7887-7892.

191. Zhao Q, Dixon RA (2011) Transcriptional networks for lignin biosynthesis: more complex than we thought? *Trends Plant Sci* **16**: 227-233.

192. ter Kuile BH, Westerhoff HV (2001) Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett* **500**: 169-171.

193. Louie GV, Bowman ME, Tu Y, Mouradov A, Spangenberg G, et al. (2010) Structure-function analyses of a caffeic acid *O*-methyltransferase from perennial ryegrass reveal the molecular basis for substrate preference. *Plant Cell* **22**: 4114-4127.

194. Kirschner M, Gerhart J (1998) Evolvability. *Proc Natl Acad Sci USA* **95**: 8420-8427.

195. Graham JWA, Williams TCR, Morgan M, Fernie AR, Ratcliffe RG, et al. (2007) Glycolytic enzymes associate dynamically with mitochondria in response to respiratory demand and support substrate channeling. *Plant Cell* **19**: 3723-3738.

196. Rasmussen S, Dixon RA (1999) Transgene-mediated and elicitor-induced perturbation of metabolic channeling at the entry point into the phenylpropanoid pathway. *Plant Cell* **11**: 1537-1552.

197. Franke R, McMichael CM, Meyer K, Shirley AM, Cusumano JC, et al. (2000) Modified lignin in tobacco and poplar plants over-expressing the *Arabidopsis* gene encoding ferulate 5-hydroxylase. *Plant J* **22**: 223-234.

198. Meyer K, Shirley AM, Cusumano JC, Bell-Lelong DA, Chapple C (1998) Lignin monomer composition is determined by the expression of a cytochrome P450-dependent monooxygenase in *Arabidopsis*. *Proc Natl Acad Sci USA* **95**: 6619-6623.

199. Reddy MSS, Chen F, Shadle G, Jackson L, Aljoe H, et al. (2005) Targeted down-regulation of cytochrome P450 enzymes for forage quality improvement in alfalfa (*Medicago sativa* L.). *Proc Natl Acad Sci U S A* **102**: 16573-16578.

200. Zhao Q, Wang H, Yin Y, Xu Y, Chen F, et al. (2010) Syringyl lignin biosynthesis is directly regulated by a secondary cell wall master switch. *Proc Natl Acad Sci USA* **107**: 14496-14501.

201. Menden B, Kohlhoff M, Moerschbacher BM (2007) Wheat cells accumulate a syringyl-rich lignin during the hypersensitive resistance response. *Phytochemistry* **68**: 513-520.

202. Mitsuda N, Iwase A, Yamamoto H, Yoshida M, Seki M, et al. (2007) NAC transcription factors, NST1 and NST3, are key regulators of the formation of secondary walls in woody tissues of *Arabidopsis*. *Plant Cell* **19**: 270-280.

203. Schuster P, Fontana W, Stadler PF, Hofacker IL (1994) From sequences to shapes and back: a case study in RNA secondary structures. *Proc R Soc London B* **255**: 279-284.

204. Babajide A, Hofacker IL, Sippl MJ, Stadler PF (1997) Neutral networks in protein space: a computational study based on knowledge-based potentials of mean force. *Fold Des* **2**: 261-269.

205. Ciliberti S, Martin OC, Wagner A (2007) Robustness can evolve gradually in complex regulatory gene networks with varying topology. *PLoS Comput Biol* **3**: e15.

206. Schrijver A (1998) Theory of Linear and Integer Programming: John Wiley & Sons.

207. Stöckigt J, Zenk MH (1975) Chemical syntheses and properties of hydroxycinnamoyl-coenzyme A derivatives. *Z Naturforsch C* **30**: 352-358.

208. Chen F, Kota P, Blount JW, Dixon RA (2001) Chemical syntheses of caffeoyl and 5-OH coniferyl aldehydes and alcohols and determination of lignin *O*-methyltransferase activities in dicot and monocot species. *Phytochemistry* **58**: 1035-1042.

209. Wall ME, Hlavacek WS, Savageau MA (2004) Design of gene circuits: lessons from bacteria. *Nat Rev Genet* **5**: 34-42.

210. Kass RE, Raftery AE (1995) Bayes factors. *J Am Stat Assoc* **90**: 773-795.

211. Shoval O, Sheftel H, Shinar G, Hart Y, Ramote O, et al. (2012) Evolutionary trade-offs, pareto optimality, and the geometry of phenotype space. *Science* **336**: 1157-1160.

212. Lerouxel O, Cavalier DM, Liepman AH, Keegstra K (2006) Biosynthesis of plant cell wall polysaccharides--a complex process. *Curr Opin Plant Biol* **9**: 621-630.

213. Vodovotz Y, Constantine G, Rubin J, Csete M, Voit EO, et al. (2009) Mechanistic simulations of inflammation: current state and future prospects. *Math Biosci* **217**: 1-10.

214. Herrmann KM (1995) The shikimate pathway: early steps in the biosynthesis of aromatic compounds. *Plant Cell* **7**: 907-919.

215. Kholodenko BN, Hancock JF, Kolch W (2010) Signalling ballet in space and time. *Nat Rev Mol Cell Biol* **11**: 414-426.

216. Kim Y, Andreu MJ, Lim B, Chung K, Terayama M, et al. (2011) Gene regulation by MAPK substrate competition. *Dev Cell* **20**: 880-887.

217. Hu W-J, Harding SA, Lung J, Popko JL, Ralph J, et al. (1999) Repression of lignin biosynthesis promotes cellulose accumulation and growth in transgenic trees. *Nat Biotechnol* **17**: 808-812.

218. Fisher WR (1908) Manual of Forestry: Forest Utilization. London: Bradbury, Agnew, & Co.

219. Anterola AM, van Rensburg H, van Heerden PS, Davin LB, Lewis NG (1999) Multi-site modulation of flux during monolignol formation in loblolly pine (*Pinus taeda*). *Biochem Biophys Res Commun* **261**: 652-657.

220. Cochrane FC, Davin LB, Lewis NG (2004) The *Arabidopsis* phenylalanine ammonia lyase gene family: kinetic characterization of the four PAL isoforms. *Phytochemistry* **65**: 1557-1564.

221. Humphreys JM, Chapple CCS (2004) Immunodetection and quantification of cytochromes P450 using epitope tagging: immunological, spectroscopic, and kinetic analysis of cinnamate 4-hydroxylase. *J Immunol Methods* **292**: 97-107.

222. Li L, Cheng XF, Leshkevich J, Umezawa T, Harding SA, et al. (2001) The last step of syringyl monolignol biosynthesis in angiosperms is regulated by a novel gene encoding sinapyl alcohol dehydrogenase. *Plant Cell* **13**: 1567-1585.

223. Franke R, Humphreys JM, Hemm MR, Denault JW, Ruegger MO, et al. (2002) The *Arabidopsis* REF8 gene encodes the 3-hydroxylase of phenylpropanoid metabolism. *Plant J* **30**: 33-45.

224. Lapierre C, Pilate G, Pollet B, Mila I, Leplé J-C, et al. (2004) Signatures of cinnamyl alcohol dehydrogenase deficiency in poplar lignins. *Phytochemistry* **65**: 313-321.

225. Steuer R, Kurths J, Daub CO, Weise J, Selbig J (2002) The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* **18**: 231-240.

226. Schreiber T, Schmitz A (2000) Surrogate time series. *Physica D: Nonlinear Phenomena* **142**: 346-382.

227. Sorribas A, Savageau MA (1989) A comparison of variant theories of intact biochemical systems. I. Enzyme-enzyme interactions and biochemical systems theory. *Math Biosci* **94**: 161-193.

228. Sorribas A, Savageau MA (1989) A comparison of variant theories of intact biochemical systems. II. Flux-oriented and metabolic control theories. *Math Biosci* **94**: 195-238.

229. Sorribas A, Savageau MA (1989) Strategies for representing metabolic pathways within biochemical systems theory: reversible pathways. *Math Biosci* **94**: 239-269.

230. Voit EO, Savageau MA (1987) Accuracy of alternative representations for integrated biochemical systems. *Biochemistry* **26**: 6869-6880.

231. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science* **220**: 671-680.

232. Heinzle E, Matsuda F, Miyagawa H, Wakasa K, 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 modelling and simulation. *Plant J* **50**: 176-187.

# VITA

## YUN LEE

Yun Lee was born in Taipei City, Taiwan on May 19, 1981. He grew up in Jingmei, a lovely neighborhood situated at the south of Taipei City. Lee received a B.S. in June 2003 and a M.S. in July 2005, both in Electrical Engineering, from National Tsing Hua University. In October 2005, Lee was conscripted to fulfill the compulsory military service and spent the next 15 months training and serving as a platoon leader for Coast Guard Administration in Northern Taiwan. From February to June 2007, Lee took an interim position as a research assistant at Academia Sinica where he implemented machine learning algorithms for identifying predictive marker genes from microarray data. In August 2007, Lee came to Georgia Institute of Technology to pursue a Ph.D. in Bioengineering and joined the Voit lab in October 2007. During his Ph.D. research, Lee developed novel computational tools for analyzing a biofuel-related metabolic pathway in wild-type and genetically engineered plants. He also collaborated with researchers at the Samuel Roberts Noble Foundation to validate experimentally the model-driven hypotheses. Lee is author/co-author of three peer-reviewed articles in journals such as PLoS Computational Biology and Mathematical Biosciences. He has also given over 15 contributed talks and poster presentations at international and domestic conferences. In 2011, Lee received the Martin Keller Award for the best student poster presentation at the BioEnergy Science Center 5$^{th}$ annual retreat, and served as a Co-PI for the first Georgia Research Alliance Challenge Program for Research Fellows.