From Experimental Observations to a Functional Model of the Lignin Pathway: Computational Modeling Reveals New Insights

A Dissertation
Presented to
The Academic Faculty

by

Mojdeh Faraji

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioengineering

Georgia Institute of Technology
May 2018

From Experimental Observations to a Functional Model of the Lignin Pathway: Computational Modeling Reveals New Insights

Approved by:

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

Dr. Andreas S. Bommarius
School of Chemical and Biomolecular Engineering
Georgia Institute of Technology

Dr. Michael J. Leamy
The George W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology

Dr. Pamela Peralta-Yahya
Department of Chemistry and Biochemistry
Georgia Institute of Technology

Dr. Peng Qiu
Wallace H. Coulter Department of Biomedical Engineering
Georgia Institute of Technology

Date Approved: March 13, 2018

To All From Whom I Learned

ACKNOWLEDGMENTS

I would like to express my genuine gratitude to Dr. Eberhard Voit for his constant presence and support throughout my PhD studies. During the most challenging times of my research he provided the greatest guidance and encouragement for me to stay on the right track and move forward. I sincerely thank Dr. Voit for his trust in me and the flexibility in his mentorship that empowered me to become an independent researcher. He has set an excellent example of a research advisor for me, and I am privileged to have had the opportunity to study and research under his supervision and mentorship.

I would like to thank all my collaborators at the Department of Biological Sciences at the University of North Texas and Oak Ridge National Laboratory for providing me with novel data and remarkable insights. My research was tremendously enhanced by valuable feedback from Drs. Richard Dixon, Luis Escamilla-Treviño, Jaime Barros-Rios and Timothy Tschaplinski. I would like to thank my committee, Drs. Andreas Bommarius, Pamela Peralta-Yahya, Michael Leamy, and Peng Qiu, for contributing to my research by criticism they provided during my work, and for evaluating my dissertation.

My transition to the biosciences would not have been possible without the exceptional support that I received from my colleagues: Po-Wei Chen, An Dam, Sepideh Dolatshahi, Luis Fonseca, Anuj Gupta, Zhen Qi, and James Wade. Our constructive discussions in the lab significantly improved my work, and their friendship warmed my every day in Atlanta.

This journey would not have been possible without the love and support of my husband Mostafa. He has inspired me and keeps on inspiring me in my scientific and non-scientific life, and I am and will always be grateful to him for that. Last but not least, I would like to thank my parents for their unconditional love and support during my difficult times far from them and their warm arms.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

LIST OF ABBREVIATIONS

PAL: L-phenylalanine ammonia-lyase

TAL: L-tyrosine ammonia-lyase

C4H: cinnamate 4-hydroxylase

4CL: 4-coumarate:CoA-ligase

CCR1: cinnamoyl CoA reductase

CAD: cinnamyl alcohol dehydrogenase

HCT: hydroxycinnamoyl-CoA:shikimate hydroxycinnamoyl transferase

C3′H: *p*-coumaroyl shikimate 3′-hydroxylase

CSE: caffeoyl shikimate esterase

COMT: caffeic acid *O*-methyltransferase

CCoAOMT: caffeoyl CoA *O*-methyltransferase

F5H: ferulate 5-hydroxylase

ER: endoplasmic reticulum

BST: biochemical systems theory

GMA: generalized mass action

PCA: principal component analysis

FBA: flux balance analysis

SUMMARY

Lignin is a natural polymer that is interwoven with cellulose and hemicellulose within plant cell walls. Due to this molecular arrangement, lignin is a major contributor to the recalcitrance of plant materials with respect to the extraction of sugars and their fermentation into ethanol, butanol, and other potential bioenergy crops. The lignin biosynthetic pathway is similar, but not identical in different plant species. It is in each case comprised of a moderate number of enzymatic steps, but its responses to manipulations, such as gene knock-downs, are complicated by the fact that several of the key enzymes are involved in several reaction steps. This feature poses a challenge to bioenergy production, as it renders it difficult to select the most promising combinations of genetic manipulations for the optimization of lignin composition and amount. Moreover, species specific regulatory features and distinct spatial and topological characteristics hinder accuracy of a unified lignin pathway model. In this dissertation a systems biology approach is used to address these challenges by means of computational modeling. Novel mathematical techniques are employed on different types of experimental data in situ, and shed light on complexities of lignin biosynthesis pathway. The developed methods are

nevertheless general enough to be used in a wide range of metabolic modeling applications.

CHAPTER I

Introduction[1]

1.1 Background

About 440 million years ago plants started to leave the oceans and inhabit land [4, 5]. The emergence of lignin during this time was an adaptation to the new environment and, specifically, a response to gravity and to limitations in accessing water. The new life also demanded plants to store water and develop systems of water transfer. The plant furthermore needed to grow in height in order to have enough access to sunlight and oxygen. Plants ultimately accomplished these multiple tasks through their xylem structures, of which lignin is a key constituent. Lignin is a phenolic polymer that is woven around and between cellulose and hemicellulose within the secondary cell wall; it provides strength and facilitates water transfer in plants. A consequence of these significant benefits for plants is that lignin is very difficult to decompose, because it is an irregular polymer that contains aromatic rings. This resistance against decomposition and digestion is known as *recalcitrance*. It is arguably the most important barrier to industrializing second-

[1] Some of the material in this chapter is excerpted from 1. Faraji, M., L.L. Fonseca, L. Escamilla-Treviño, J. Barros-Rios, N. Engle, Z.K. Yang, T.J. Tschaplinski, R.A. Dixon, and E.O. Voit, *Mathematical models of lignin biosynthesis.* Biotechnology for Biofuels, 2018. **11**(1): p. 34, 2. Faraji, M. and E.O. Voit, *Improving Bioenergy Crops through Dynamic Metabolic Modeling.* Processes, 2017. **5**(4): p. 61, 3. Faraji, M., L.L. Fonseca, L. Escamilla-Treviño, R.A. Dixon, and E.O. Voit, *Computational inference of the structure and regulation of the lignin pathway in Panicum virgatum.* Biotechnology for Biofuels, 2015.

generation biofuels, and in particular the production of ethanol from inedible plant parts as sustainable and affordable biofuels, because recalcitrance necessitates additional treatment steps, such as hot acid or ammonia baths, to loosen the lignin structure [6-8]. These steps require time and expense and therefore reduce feasibility and cost effectiveness. Moreover, most of the pretreatments are not environmentally friendly [9, 10]. Outside the biofuel industry, recalcitrance affects forage digestibility, and progress toward reducing recalcitrance could have a significant impact on the cattle and sheep industry [11]. While lignin is an obstacle to bioenergy production, industries associated with textile production and the synthesis of organic compounds have identified lignin as a valuable starting compound for novel processes and are interested in maximizing it. Thus, both increases and decreases of lignin biosynthesis are of potential importance.

1.2 Reaction System of Lignin Biosynthesis

Monolignols, the end products of the lignin pathway, are the main precursors of the lignin polymer. The traditionally accepted generic lignin biosynthesis pathway is shown in Figure 1.1 by grey arrows. The different end products, *p*-coumaryl alcohol, coniferyl alcohol, and sinapyl alcohol, are precursors of H-, G- and S-lignin, respectively. Once biosynthesized in the cytosol, they are translocated to the cell wall where they polymerize to form the lignin polymer [12]. The pathway branches at *p*-coumaroyl- CoA to provide S and G-lignin precursors, and the downstream compound coniferaldehyde is the branch point where the pathway splits towards G- or S-lignin. The grid-like structure of the pathway system includes several parallel reactions that are catalyzed by the same enzyme; an example is

the conversion of *p*-coumaryl aldehyde to *p*-coumaryl alcohol, conversion coniferaldehyde to coniferyl alcohol and conversion of sinapaldehyde to sinapyl alcohol that are all catalyzed by CAD. However, depending on the plant species there are additional or absent reactions, regulatory mechanisms or spatial characteristics that make the lignin pathway distinct for the particular species or, presumably, group of species. Figure 1.1 illustrates such structural differences in *Medicago*, switchgrass, the model plant *Brachypodium* and *Populus*.

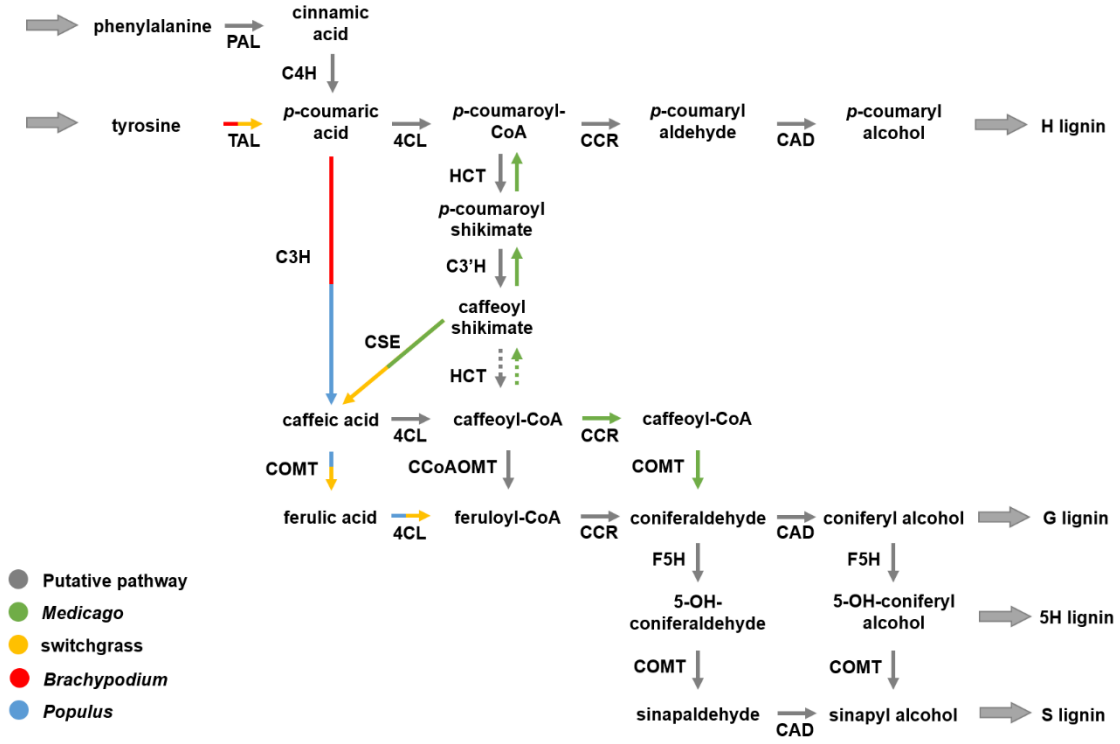

Figure 1.1 Putative lignin biosynthesis pathway with identification of species-specific reactions. Generic reactions, mainly from studies in the model dicot Arabidopsis thaliana, are shown in grey. Other enzymatic reactions are color coded based on the plant species where they were documented. Multicolored arrows represent reactions present in more than one species. PAL phenylalanine ammonia-lyase, TAL tyrosine ammonia-lyase, C4H cinnamate 4-hydroxylase, C3H p-coumarate 3-hydroxylase, C3′H p-coumaroyl shikimate 3-hydroxylase, COMT caffeic acid O-methyltransferase, F5H ferulate 5-hydroxylase, 4CL 4-coumarate:CoA ligase, HCT hydroxycinnamoyl-CoA:shikimate hydroxycinnamoyl transferase, CCoAOMT caffeoyl-CoA O-methyltransferase, CCR cinnamoyl-CoA

reductase, CAD cinnamyl alcohol dehydrogenase, CSE caffeoyl shikimate esterase. Interestingly, some monocots, such as Brachypodium and maize, do not have CSE ortholog genes. Dashed arrows are currently considered less efficient metabolic reactions in vivo. Adapted from [1].

1.3 Significance and Rationale of Research

Numerous attempts have been made in recent times to manipulate the lignin content and composition in candidate plants for biofuel production. Many of these studies relied on the assumption that the lignin biosynthesis pathway was known in sufficient detail. However, this is not necessarily the case, especially in understudied plant species, and the precise pathway structure is often unclear and requires dedicated research for such species. For instance, *Selaginella moellendorffi* and *Medicago truncatula* have basically similar lignin pathways, which however differ in some of their metabolic branch points as well as their enzyme properties [13-15]. Beyond the topological structure of any of these pathways, it is not surprising that different species have evolved distinct regulatory control patterns. The immediate consequence of such discrepancies for the biofuel industry is that the direct extrapolation of knowledge, methods and treatments from one species to another is not always valid. Moreover, it is well known that pathway systems are highly nonlinear and difficult to predict with intuition alone. A feasible strategy is therefore to employ computational approaches of systems biology and metabolic engineering.

Both the amount and composition of lignin are thought to be correlated with the hardness as well as the recalcitrance of structural plant materials. It is therefore important to the production and manipulation of bioenergy crops to understand the details of lignin

synthesis and the deposition and polymerization of monolignols in the plant cell wall. In particular, the question arises whether it is possible and feasible to intervene in the phenylpropanoid pathway of lignin biosynthesis in a targeted and effective manner, for instance, through gene knockdowns. The answer to this question is evidently preconditioned on detailed knowledge of this pathway and its control *in situ*. This knowledge in turn requires different types of biological data and, in cases where these are difficult to understand, the use of computational models that are capable of integrating small or large datasets of the same or different types and explaining observations that are sometimes unexpected.

The design of suitable models for a complex pathway such as lignin biosynthesis is not trivial. First, it is generically unclear which mathematical representations are optimal for describing a natural system. Second, one cannot be sure that information or data from one species are valid in another species, even if the two are closely related. Similarly, it has been shown many times that data obtained *in vitro* are not necessarily applicable *in vivo* [14-18].

At the same time, species-specific experiments are time consuming and expensive. Mechanistic models based on enzyme kinetics seem to be an intriguing choice, but it has been shown that mechanistic models are not always good solutions, for instance, if parameter values and enzymatic rate laws are based on strong assumptions like bulk reactivity that are not necessarily satisfied *in vivo* [16]. An alternative that was recently proposed is the characterization of *in vivo*-like kinetics [17], which however is costly and time consuming and would still require extensive validation, which is seldom truly achieved [16]. An additional challenge for the design of models is the scarcity of high-

quality, quantitative data for model design, testing and validation, which poses a significant obstacle to all analyses, especially of relatively understudied species.

The objective of my dissertation research is therefore to analyze the lignin biosynthesis pathway of different plant species with computational means of systems biology. The analysis is based on different datasets. Most of them come from stem and tiller tissue and contain measurements of the lignin content (H, G and S lignin) and the S/G lignin ratio in wild type and in transgenic lines. Additional information exists in the form of labeling data that were obtained from feeding plants labeled precursors of the lignin pathway. The overall strategy of this dissertation work is to develop computational models that characterize the structure and regulatory control patterns of lignin biosynthesis at a systemic level in select key pant species. Similar to any other modeling effort, the two main steps of this strategy are, first, to explain the experimental results from wild type and transgenic lines and, second, to devise a rational basis for prescribing manipulations of the pathway toward altered lignin content and composition.

1.4 Dissertation Overview

The dissertation includes seven chapters. Below is an overview of each chapter's content.

Chapter I, as already presented, provides the background, motivation for the research and the significance of a computational approach.

Chapter II discusses computational metabolic modeling in bioenergy crops and presents a comprehensive review of the literature on mathematical frameworks and

approaches employed for analyzing plant metabolic pathways. Next, mathematical models of lignin biosynthesis pathway are investigated with a specific focus on data needs for such models. Several case studies illustrate computational analyses of each data type.

Chapter III covers a computational model of the lignin pathway in switchgrass (*Panicum virgatum*), which the Department of Energy considers a most promising species for bioenergy. The model is constructed based on experimental data that include lignin content and composition from wild type and transgenic plants. The focus of this model design are the topology of the pathway and its regulatory mechanisms. Model validation and predictions of lignin profiles in double knockdowns are assessed in the second half of the chapter.

Chapter IV presents a computational model of the lignin pathway in *Brachypodium distachyon*, which has become an important model species in lignin research The model aims to answer intriguing questions regarding ^{13}C-labeling experiments suggesting that two distinct pathways from phenylalanine and tyrosine facilitate the differential incorporation of labeled phenylalanine and tyrosine in different lignin units. The puzzling aspect of this scenario is that the two pathways use some of the same intermediate metabolites. A two-compartment model is proposed that extends the model constructed in Chapter III by including spatial characteristics of the pathway. Labeling data in wild type control plants are used to construct a static model that is able to explain the distinct incorporation of phenylalanine and tyrosine in different lignin units.

Chapter V presents a novel method of stepwise inference of likely dynamic flux distributions from metabolic time series data. Good-quality, high-resolution metabolic data are becoming more readily available, so this method is timely, as it uses minimal

assumptions while offering an effective approach to estimating flux magnitudes in metabolic pathway systems. The method is demonstrated to perform well in a case study targeting an artificial dataset and a pathway that closely resembles the lignin pathway in switchgrass.

Chapter VI discusses an unprecedented nonparametric dynamic method to model metabolic pathways. The rationale for this method is the fact that the choice of the optimal mathematical formalism for representing enzymatic reaction fluxes and the parameterization of the resulting model are among the most challenging and biased steps in any modeling effort. Nonparametric dynamic modeling circumvents these two steps by replacing the closed-form flux formulation in a typical dynamic model with look-up libraries that are established beforehand based on series of time series data. The result is an essentially unbiased modeling approach for nonlinear compartment systems, including metabolic pathway systems.

Chapter VII presents a summary and conclusion of the research results obtained in this dissertation and discusses future work that may be based on the models and methods developed herein.

CHAPTER II

Improving Bioenergy Crops through Metabolic Modeling and Mathematical Models of Lignin Biosynthesis[2]

2.1 Introduction

Crops have been cultivated, bred, and improved for thousands of years, and some successes have been astounding: a modern corn cob weighs between 1 and 1 ½ pounds, whereas its early predecessor, the ancient Latin American grass teosinte, had an average fruit weighing about 35 grams [19]. Achieving this 20-fold increase took about 8000 years.

In contrast to food crops, bioenergy crops have not been investigated for very long, if one ignores the burning of wood and other organic materials. Due to its relative youth, research on bioenergy crops has the immediate advantage of a rich body of genetic and metabolic information. Furthermore, millions of dollars in tax incentives and subsidies reflect the determination of many countries around the world to advance the use of sustainable bioenergy products, wean the world off its dependence from fossil

[2] The material in this chapter has been published as: 1. Faraji, M., L.L. Fonseca, L. Escamilla-Treviño, J. Barros-Rios, N. Engle, Z.K. Yang, T.J. Tschaplinski, R.A. Dixon, and E.O. Voit, *Mathematical models of lignin biosynthesis.* Ibid.2018. **11**(1): p. 34, 2. Faraji, M. and E.O. Voit, *Improving Bioenergy Crops through Dynamic Metabolic Modeling.* Processes, 2017. **5**(4): p. 61.

fuels, and reduce greenhouse emissions. As a consequence, genetic and metabolic engineering have made the production of ethanol, butanol, and fatty acids from corn or sugar cane, competitive. As an example, in September 2014, three plants in Iowa and Kansas started commercial production of cellulosic ethanol with an annual ramp-up capacity of 80 million gallons. While impressive, this amount constitutes only a fraction of the federal call for 1.75 billion gallons per year in the U.S.

The low-hanging fruit of using sources like corn is in direct competition with the supply of food products, and it has become today's challenge to produce "second-generation" bioenergy from other plant materials that do not cause ethical concerns. This new type of biofuel research is sometimes called "advanced," because it relies on lignocellulosic materials which, by and large, correspond to inedible plant parts, such as corn stover, pine bark, grasses, and wood chips. In parallel, algae have been studied extensively, but so far, with little economic success.

The challenges associated with exploiting these plant sources are manifold, but often converge to two overarching issues. First, the energy is stored less in concentrated, easily accessible sugars and more in woody substances that are difficult to ferment; we will discuss this aspect later. Second, plants are enormously complex, which is in part due to large numbers of constituents, and their interactions. As an example, some species of Spruce (*Picea*) are predicted to possess between 50,000 and 60,000 genes [19], which is between two and three times the number of human genes [20]. This large number of genes presumably reflects considerable redundancies, metabolic plasticity, and numerous stress response mechanisms, which are needed to compensate for the plant's lack of motility, and result in distinct differences to animal physiology [21, 22]. Whereas

the human metabolome library (HML) lists slightly more than 1,000 different metabolites in humans [22], the number of metabolites in the plant kingdom is estimated to lie somewhere between 200,000 and 1,000,000 [23-25]; indeed, the width of this range alone indicates how little of plant metabolism we truly understand. Of course, plants also contain uncounted proteins and structural elements, which all contribute to their survival, but also impede attempts of targeted alterations. Collectively, these features render genetic and metabolic engineering of plants very challenging.

Nonetheless, new gene editing techniques, for instance, based on clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated protein 9 (Cas9), have found their way into plant breeding [26]. The first applications targeted *Arabidopsis*, tobacco, rice, and wheat [27-29], but numerous other species have followed (*e.g.*, [30-32]). These experimental advances are directly pertinent for future crop modeling, as they permit modifications that were considered impossible just a few years ago.

A particular, additional challenge associated with plants is that primary and secondary metabolism are tightly linked and regulated by "super-coordinated" gene expression networks [33]. This tight coordination may explain why it is not straightforward to identify and tweak only certain processes of interest, because many processes are possibly affected and robustly compensated. Adding to these complications is the fact that plant cells are highly compartmentalized [34, 35], and that metabolite turnover occurs over a wide range of rates [36].

Another very challenging feature of plants is their polyploidy, that is, the existence of several copies of their chromosomes, which obviously makes targeted alterations cumbersome and complicates essentially all gene manipulations, even if the methods and techniques are routine in microbes. Polyploidy is particularly important for crops, because it offers the opportunity of modifying traits and lineages [37]. In fact, polyploidization is found in many modern crops and wild species, including cotton (*Gossypium hirsutum*), tobacco (*Nicotiana tabacum*), wheat (*Triticum aestivum*), canola (*Brassica napus*), soybean (*Glycine max*), potato (*Solanum tuberosum*), and sugarcane (*Saccharum officinarum*). For instance, bananas are triploid and potatoes tetraploid, while wheat is hexaploid and sugarcane octoploid [38]. Polyploidy can be extreme: members of the genus *Ophioglossum* of adder's-tongue ferns can have very high chromosome counts, with possibly up to 720 chromosomes, due to polyploidy [39]. Further complicating the large numbers of chromosomes is the fact that some plants are alloploid, as they evolved or were bred through the hybridization of different species. An important example is the genus *Brassica*, which contains cabbages, as well as cauliflower, broccoli, turnip and seeds for the production of mustard and for canola oil (*cf.* [40]). Polyploidy can be traced back far within the phylogeny of a species. As just one example, there is strong evidence of polyploidy through breeding that can be seen in the long history of rice cultivation, where massive gene duplications have occurred since ancient times. The result is an estimated count of at least 38,000–40,000 genes in rice, of which only 2–3% are unique to any two rice subspecies, such as *indica* and *japonica* [41]. It is, at this point, unclear what the ramifications of

polyploidy for modeling might be, but it is clear that both experimental and modeling approaches have to grapple with this key issue of crop manipulation.

Faced with the challenges and the enormous diversity of plants, the plant and crop communities have been focusing primarily on a number of model plants. Some of these, notably rice (*Oryza sativa*), maize (*Zea mays*), soybean (*Glycine max*), tobacco (*Nicotiana tabacum*), alfalfa (*Medicago truncatula*), and black cottonwood (*Populus trichocarpa*), are food, feed, or potential energy sources, while others, like *Arabidopsis thaliana* and *Brachypodium distachyon*, have features that greatly facilitate their investigation, such as relatively small genomes, fast growth, and diploidy instead of polyploidy.

Within this context, most improvements in crop production have come from experimental metabolic engineering research, which has been model free in the sense that new plant alterations were guided by biological intuition. This approach has been very successful in many instances, but one should expect it to run into problems as soon as large omics datasets become a standard in crop science. These datasets are very valuable, but they are also so immense in size that they cannot be comprehended by the unaided human mind, and require sophisticated computer algorithms for analysis and interpretation; a review by Yuan *et al.* [22] discusses this need for integrating "big data" with traditional plant systems biology. Yet, even modern machine learning methods of analysis are not sufficient by themselves. These methods are designed to filter information from noise and mine patterns from data that the unaided human mind cannot comprehend, but they rarely suggest mechanisms or provide explanations, especially with respect to the dynamics of a system under study, and if this system is nonlinear,

due to regulation, synergisms, and threshold effects. Thus, if the goal of an investigation is an explanation of why a system behaves the way it does, or a prediction of system responses under untested conditions, intuition and statistical data analysis alone are susceptible to failure in the complex world of plant physiology. They need to be complemented with dynamic systems modeling, which is a rather new subject in bioenergy research.

Among the relatively few recent mathematical modeling efforts in the field of crop research, many studies have focused on photosynthesis (*e.g.*, [42-49]), on pathways of general importance, such as the TCA cycle (*e.g.*, [21, 47, 50-52]), or on specific pathways that are of particular industrial interest, such as flavonoid and isoprenoid metabolism (see reviews [22, 35, 53, 54]). By contrast, metabolic modeling for improved plant biofuels is still relatively scarce. Nonetheless, considering the complexity and variability of plants, the utilization of methods of computational systems biology appears to become an increasingly rational strategy toward realizing economically feasible bioenergy production, and one might expect that computational modeling will become a standard tool of guiding experimentation in the future.

Returning to the specific challenge of difficult access to sugars in inedible plant materials, the focus must shift to lignin, which severely impedes bioenergy extraction from woody substrates. Except for cellulose, lignin is the most abundant terrestrial biopolymer and accounts for roughly 30% of all organic carbon in the biosphere [12]. It is the main constituent of wood, and plays a vital role in terrestrial plant life, as it is the key component of the water transport system in the plant xylem and gives the plant structure and strength to overcome gravity. Chemically, lignin is an irregular phenolic

polymer, whose hydrophobic nature not only facilitates water transfer from the roots, but also blocks surface evaporation from stems and leaves.

Within the cell wall, lignin is physically entangled with cellulose and hemicellulose molecules, and thereby severely limits the production of ethanol and other bioenergy compounds by hindering the access of enzymes to these desirable polysaccharides. It is also resistant to enzymatic digestion and, therefore, difficult to remove from plant materials. The ultimate consequence of lignin in plant walls is *recalcitrance*, which summarily describes the resistance of plant materials to fermentation. Recalcitrance has emerged as a major obstacle in the commercial production of cellulosic ethanol and other bioenergy compounds, and has therefore become a key target for bioenergy research. The complete elimination of lignin is, of course, not desirable, but even a reduction in lignin content and/or certain changes in lignin composition have been shown to improve ethanol yield [6, 7, 55, 56]. As a consequence, the second-generation bioenergy industry has put lignin biosynthesis and degradation into the spotlight. Specifically, one focus area has become the alteration of the quantities and proportions of the three or more types of monolignols, which are the building blocks of the lignin heteropolymer.

As an interesting side note, lignin is not always a problem. In fact, it is a true yin and yang: on the one hand, it is an impediment to bioenergy production, but on the other hand, it is a very intriguing organic compound, and some recent industry efforts actually target the harvesting of lignin as a valuable resource for a variety of chemical syntheses.

Most efforts of altering lignin have been directed toward biotechnological experimentation, and computational modeling efforts are still the exception, although they have emerged with increasing frequency. Examples include computational models by Lee *et al.*, who analyzed the pathways of lignin biosynthesis in poplar [57] and alfalfa [58], based on gene knockdown experiments. Wang *et al.* constructed a kinetic model of the lignin pathway from a large set of *in vivo* and *in vitro* measurements [59]. Faraji *et al.* investigated the lignin biosynthesis pathway in switchgrass (*Panicum virgatum*) [3], which was identified by the U.S. Department of Energy as the most promising monocot plant for biofuel ethanol. Other work in this field includes [60, 61]. A summary of highlights from these studies is provided in the second part of this chapter, as well as in [1].

In the following, we first describe representative mathematical approaches that are currently used for modeling crop metabolism and include a wide variety of techniques. One should note that the physiological attributes of plants often translate into unique mathematical constraints that require reevaluation of the details and underlying principles of popular modeling formalisms. While discussing different approaches, we highlight specifically the metabolic modeling of bioenergy crops. In terms of references, we give preference to articles addressing plant and crop systems, while keeping general references to a minimum.

2.2 Mathematical Modeling Approaches for Metabolic Engineering in Crops

Significant improvements in food and bioenergy crops are very challenging, but the enormous global scale of mobile energy use, and the corresponding potential economic benefits of even minor percent improvements in biofuel yield, are very attractive. As a consequence, many attempts have been made to alter crops with traditional methods of metabolic engineering, where the overriding goal is the targeted alteration of metabolic pathways toward better yields in compounds like ethanol and butanol.

It is only recent that computational biology has begun to partner with experimental biology in advancing and pushing the boundaries of rational crop science [62]. Much of this work has focused on the model plant species discussed earlier, and indeed, some comprehensive, multi-scale models are available that address these model species [56-58, 63]. Also, most of these models have an exclusive focus on steady-state operation, whereas dynamic modeling of bioenergy pathways is still in its infancy. Baghalian *et al.* [53], and Morgan and Rhodes [35], provide excellent reviews on modeling plant metabolism that describe prominent mathematical approaches in the field.

Mathematical models for metabolic pathway analysis are manifold and driven by the availability of data types [64]. They may be classified in two coarse categories. The first uses steady-state approaches, which have the two advantages that they are algebraic, which renders large model sizes possible, and that they are relevant, as many systems operate close to a steady state. The second category contains dynamic models, which are more realistic, and cover transients as well as the steady state, but are mathematically more complicated. Outside these categories, the literature contains a few models that are stochastic, permit spatial considerations, and span multiple scales.

2.2.1 Steady-State Modeling

Any modeling strategy is ruled by the availability of data, and plant metabolic modeling is no exception. For systems operating close to a steady state, ideal data would consist of metabolite concentrations, and of the distribution of fluxes throughout a metabolic system. Unfortunately, such data are seldom available, and the computational estimation of fluxes has become one of the crucial steps in metabolic modeling. One basis for estimation is the technique of ^{13}C labeling, which has become popular for metabolic flux characterizations, and entails computational modeling for metabolic network reconstruction [65, 66]. In particular, stoichiometric modeling [67] has been widely applied to isotopic labeling data [52, 68-72]. In some cases, these methods have allowed the estimation of entire flux maps [34], but most metabolic flux models presently lack sufficient data and are underdetermined. As a consequence, much effort in the field has been dedicated to algorithm development and experimental techniques that try to infer flux values from other data.

The estimation of fluxes falls into two steps. First, one needs to determine which fluxes are likely to exist within a particular metabolic system. This determination is usually performed indirectly, namely through genome sequencing, which permits connecting a genotype to an observable phenotype by means of genome-wide metabolic reconstructions, based on sequence comparisons with better-identified organisms [73, 74]. Once the candidate fluxes and their associations with metabolites are established, the magnitudes of all fluxes are to be determined. The guiding principle is that, at any steady state, the fluxes

entering a metabolite pool must collectively be equal, in total magnitude, to the collection of fluxes exiting this pool [75, 76]. The most prominent implementation of this concept is flux balance analysis (FBA; see below) [73, 77], which computes the flux distribution within a metabolic pathway at a steady state, based on an assumed objective of the system, such as maximum growth or some maximum flux.

Other steady-state approaches at the level of flux distributions are flux variability analysis (FVA) [78], elementary mode analysis (EMA) [79, 80], extreme pathway analysis (EPA) [79, 80], and metabolic flux analysis (MFA) [81]. One might also mention metabolic control analysis (MCA) [82-85] in this category, as it was designed specifically for assessing the control of flux through a pathway at a steady state. Pertinent details of these approaches are presented below.

2.2.1.1 Flux Balance Analysis (FBA)

In typical pathway systems, the number of fluxes is greater than the number of metabolites, because the same metabolite is usually involved in more than one reaction. As a consequence, the stoichiometric matrix of a typical metabolic system is underdetermined and infinitely many solutions are possible. To address this situation, FBA formulates the system as a linear programing problem, where the solution of the underdetermined system is a member of the solution space, and optimizes an objective function of choice, such as maximal growth. The solution space itself is determined by linear constraints of the problem, such as non-negativity and maximal magnitudes of fluxes [72, 86]. FBA is a simple, yet powerful tool that has been widely used to determine steady-state flux

distributions. One caveat of this method is the choice of a suitable objective function. The choice of maximal growth is often suited for microbial populations, but in mammalian systems and in plants, where several pathways simultaneously share metabolites and enzymes, selecting the right objective function is not a straightforward task.

FBA has been used successfully in plant and bioenergy research. For instance, Paez *et al.* analyzed biomass synthesis in *Chlamydomonas reinhardtii* under different CO_2 levels [87]. Chang *et al.* presented a genome-scale metabolic network model of the same organism [88] using FBA and FVA. Employing a variant of flux balance analysis that accounts for dynamics (DFBA), Flassig *et al.* [89] modeled the β-carotene accumulation in *Dunaliella salina* under various light and nutrient conditions. Because it is to be expected that plants must satisfy several objectives, methods of multi-objective optimization have been applied to metabolic plant modeling as an alternative to FBA [90, 91].

A somewhat problematic aspect of FBA is the omission of nonlinearities, such as regulatory signals, which clearly operate in actual cells. While the FBA solution itself is unaffected by regulation, any extrapolations to new situations, such as gene knockouts, can be significantly influenced by regulatory signals, thereby rendering FBA predictions questionable. A second issue is the fact that plant cells are highly compartmentalized, which complicates any type of modeling. In particular, one must question whether it is admissible to merge "parallel" fluxes, using the same substrates, which are, however, proceeding in different compartments. Experimental studies have shown that even within the cytosol, spatial channeling of multiple enzymes can mimic pseudo-compartmental behavior, without which some aspects of the dynamics of a plant cell cannot be explained

[92]. Finally, it is unclear to what degree plant cells truly operate under (quasi-) steady-state conditions.

An interesting variation of FBA is the method of *minimization of metabolic adjustment* (MOMA) [93], which characterizes a flux distribution that is altered due to a mutation or intervention in relation to the corresponding FBA solution for the same wild type internode. Expressed differently, MOMA focuses on the admissible solution within the solution simplex that most closely mimics the wild type. Lee and coworkers used MOMA to analyze data from knockdown experiments with genes associated with lignin biosynthesis in alfalfa [58].

2.2.1.2 Flux Variability Analysis (FVA)

FVA is a constraint-based modeling variant of FBA. It addresses the well-known situation in linear programming (LP) that a problem has infinitely many solutions, because the optimal solution is not one of the vertices of the solution simplex. This situation arises when the objective function is parallel to one of the LP constraints. For such a case, FVA determines the variability in each of the fluxes in the proximity of equivalent, admissible solutions [78]. An unexpected merit of the method is that biological systems do not necessarily operate truly optimally, and that it is hence important to explore flux distributions in slightly suboptimal solutions as well. As a more conservative method, FVA may appear to be a better fit for the complex biology of plants than FBA.

Hay and Schwender employed flux variability analysis (FVA) to reconstruct the seed storage metabolism pathway in oilseed rape (canola; *Brassica napus*) and to characterize the changes in this pathway during seed development [94, 95]. As one of the

largest sources of edible vegetable oil in the world, oilseed rape is also a favored biofuel crop. The authors were able to identify the differential roles of fluxes and their variability under different nutritional conditions. Their results provide an interesting computational validation of how metabolic redundancies can play a crucial role during the important phase of seed development within the rapeseed life cycle.

In a different application of FVA, combined with FBA, Chang *et al.* developed a genome-scale metabolic network model for the microalga *Chlamydomonas reinhardtii* [88], to investigate the effect of light on metabolism. Their specific goal was to create a predictive tool for an optimal light source design.

2.2.1.3 Extreme Pathway Analysis (EPA) and Elementary Mode Analysis (EMA)

Extreme pathways represent the structure of a pathway network as a linear combination of flux pathways that act as the vector basis, in the sense of linear algebra [79, 80]. With this set-up, any steady-state vector of the system can be written as a linear combination of this basis. In a geometric interpretation, the extreme pathways are the lateral edges of the admissible cone of solutions that is anchored at the origin. The extreme pathways are a subset of the so-called elementary modes of the pathway system. In EMA, non-decomposability constraints ensure these elementary modes are genetically independent. As a consequence, they can explain the links between the genotypes and the corresponding phenotypes. Steuer *et al.* applied elementary modes in their analysis of the mitochondrial TCA cycle in plants [96].

Extreme pathways are unique and irreducible sets of elementary modes. As such, an important drawback arises when a pathway has many degrees of freedom, because the

number of the elementary modes is equal to the degrees of freedom in the pathway. In such cases, analysis of the system through the assessment of elementary modes becomes cumbersome due to the combinatorial explosion of admissible routes in the system [97]. A second limitation of the method is that extreme pathways cannot always convert an input into a desired product, although the elementary modes of the system allow such a conversion [80]. Also, typical EPAs assume a predominant reaction for every reaction, which is often, but not always a given. If the system contains reversible pathways, the extreme rays of the solution cone may lose this property, and it may be preferable to work with *extreme currents*, or to define different classes for reversible and irreversible fluxes [98].

The main advantage of EMA/EPA is the following: in a metabolic engineering problem, diagnostics of elementary modes and extreme pathways will assist in designing a scheme of multiple genetic alterations in a targeted manner. Specifically, an optimized solution derived from techniques such as linear programing might provide a more desirable numerical value for the objective of the problem when suboptimal solutions derived from EMA provide more biologically meaningful solutions, due to the synergy in the regulation of the genes involved in the chain of reactions in each elementary mode.

2.2.1.4 Metabolic Flux Analysis (MFA)

MFA relies on labeling data, which are usually generated with an experiment where ^{13}C labeled substrate is given to the system. After some while, the label distributes among the metabolites according to the magnitudes of fluxes within the system. ^{13}C is a stable isotope of carbon which contains one extra neutron relative to ^{12}C, the most abundant isotope of

carbon. Hence, methods such as mass spectrometry are able to detect the level of isotope abundance in different metabolites, which in turn, assists in the elucidation of the fluxes in a metabolic pathway. The idea behind MFA is that measuring sufficiently many fluxes leads to a substantial reduction of the degrees of freedom of a pathway, possibly to zero, in which case, a unique solution is achievable [80]. Although conceptually straightforward, MFA is technically quite difficult, and measuring internal fluxes is still a challenge. However, new experimental techniques are expected to provide us with the desired information in the foreseeable future [99]. Roscher *et al.* discussed applications of metabolic flux analysis in photosynthetic and non-photosynthetic plant tissues [100]. The comprehensive review by Dieuaide-Noubhani and Alonso [101] covers MFA in plants, and describes both experimental and mathematical modeling steps.

2.2.1.5 Metabolic Control Analysis (MCA)

MCA [82-85] was proposed specifically for assessing the control of flux through a pathway at a steady state. Before MCA was developed, it was assumed that every pathway has a rate-limiting step, which controls the flux through the pathway. The proponents of MCA showed convincingly that there is seldom a single rate-limiting step in a metabolic pathway. Instead, the control of the flux is distributed, with different degrees of importance, among many or all reaction steps. MCA addresses this issue by computing flux and metabolite control coefficients, and elasticity coefficients, which coarsely correspond to sensitivities, and may be derived from alleged functional forms of rate laws or direct experimental measurements [102]. A review by Rees and Hill discusses MCA specifically

in the context of plant metabolism [103]. Giersch *et al.* [104] applied MCA to the system of photosynthetic carbon fixation.

2.2.1.6 Limitations of Steady-State Approaches

Steady-state modeling has the advantage of relative mathematical simplicity, and in particular, the fact that no differential equations are involved. However, the restriction to steady-state operation must be considered with some caution in plant and crop modeling, as plants seldom truly operate at the same steady state throughout the day [35, 36, 53, 100]. In particular, the dynamics of the light–dark cycle is an important reminder of the non-steady-state operation of plants [52]. Parallel pathways, often occurring in several compartments, add to the complexity of plant metabolism [34]. Finally, the large range of turnover times of metabolite pools may affect the validity of pure steady-state models [43].

2.2.2 Dynamic Modeling

Dynamic modeling has the potential of capturing the complex physiology of plants more accurately. Specially, kinetic models are, at least in principle, capable of simulating time course data and permit a variety of dynamical analyses of metabolic pathways. However, in comparison to steady-state models, dynamic models are more difficult to analyze, and require much more data support.

2.2.2.1 Explicit Kinetic Models

The law of mass action is the basis for the earliest quantitative modeling of a chemical reaction rate law [105]. Kinetic mass action models are widely used in metabolic modeling. In plant metabolic modeling, an example is the work by Farre *et al.* [106], who developed a model of carotenoid biosynthesis in maize to identify effective genetic intervention points. Bai *et al.* [107] used a mass action kinetic model to investigate the carotenoid pathway in rice embryonic callus, and presented model-driven metabolic engineering strategies.

Rooted in the law of mass action, the first mechanistic kinetic models of metabolism were based on the concept of the Henri–Michaelis–Menten mechanism and its generalizations [108, 109]. The mathematical representations of the reaction steps, according to these concepts, contain physical properties that are represented by measurable parameters, such as V_{max}, K_M, and K_i [108, 110]. Although these mechanistic kinetic models are still predominant [59], their underlying assumptions are seldom justified in a living cell. For instance, these models implicitly rely on the homogeneity of the medium in which the reactions occur, which doesn't hold true in the *in vivo* environment of a cell. For larger systems, the parameterization of mechanistic kinetic models becomes laborious, expensive, and time consuming [53], and the resulting measurements are often obtained *in vitro*, and may not be representative of enzyme kinetics *in vivo* [43, 111]. An additional problem with mechanistic models is that it may become impossible to infer their structure if multiple regulatory mechanisms are involved in a reaction [112].

The simplest mechanistic models of metabolism are over a century old, and describe the kinetics of individual enzyme catalyzed reactions [108, 109, 113]. A modern example of their use within the context of bioenergy crops is the work of Nag *et al.* [114], which elucidates the carbon flow in plant cells, using a mechanistic kinetic model of starch degradation. Starch is of great interest in the biofuel industry, as it is readily fermentable into alcohol or other energy products. Wang *et al.* [59] constructed a kinetic metabolic model of lignin synthesis in black cottonwood (*Populus trichocarpa*), based on a large array of *in vivo* and *in vitro* measurements.

Uncounted variations and alternatives of the original Michaelis–Menten concept were developed over the years to represent more complicated enzymatic processes and their regulation in an appropriate manner. These variants have been reviewed numerous times [112, 115, 116], and are not described here much. Instead, we focus on more global approaches that permit streamlined representations of entire pathway systems. As alternatives to the original mechanistic models, several modeling frameworks have been proposed as semi-mechanistic strategies that represent which variables affect which fluxes, but do not dictate specific mechanisms. Examples include biochemical systems theory (BST) [3, 18, 57, 58], structural kinetic modeling [117], dynamic flux estimation [118] and nonparametric dynamic modeling [119, 120]. They are briefly described below.

2.2.2.2 Biochemical Systems Theory (BST)

BST is a kinetic modeling approach that uses power-law functions to model all fluxes [121, 122]. The core idea behind BST is that, in logarithmic space and close to an operating point, a rate law is well represented by a linear function of the substrate(s) and regulator(s)

of the reaction [123]. Therefore, a multivariate Taylor linearization in logarithmic space about the biological operating point approximates the often unknown kinetic process with reasonable accuracy. In Cartesian space, the result of this approximation is a term consisting of a rate constant and a product of power-law functions of all contributing variables, each raised to an exponent, called the kinetic order. Each exponent can have any real value; it is positive for substrates and activators, negative for inhibitors, and zero for variables without direct effect on the flux. The power-law functions are easy to adjust for any number of substrates or regulators, and BST has been widely used in a variety of organisms. Lee *et al.* used BST in a steady-state and dynamic flux characterization of the lignin biosynthesis pathway in *Medicago* [18]. Other examples of BST framework in plant modeling are [3, 57, 58, 124-131].

2.2.2.3 Other Dynamic Modeling Approaches in a Predefined Format

The *saturable and cooperative formalism* has its roots in BST, but instead of presenting the variables in each term with simple power-law functions, uses Hill-type functions [132]. Thus, every process is guaranteed to saturate, and the accuracy of models in this formalism is often higher than in BST. However, this improvement is paid for with a considerable increase in the number of parameters.

The *linear-logarithmic* (lin-log) representation of enzyme-catalyzed reactions is closely aligned with the concepts of MCA and can be seen as the dynamic arm of this formalism [133, 134]. It represents all variables, reaction rates, and fluxes in relation to their steady-state analogues. For very large substrate concentrations, the accuracy of these

models is superior to those in BST [135], but for very small concentrations, the lin-log rates become negative and approach $-\infty$ when the substrate converges to zero [123, 136, 137].

Structural kinetic modeling recognizes the disadvantages of rate laws whose mathematical formats cannot be justified on biological grounds, and assigns a local linear representation at each point of a simulation. The system is first written in terms of the Jacobian of the system, and then the Jacobian is reconstructed such that its components are either directly measurable or estimated. The resulting model is free of explicit functional forms [117]. Steuer *et al.* presented a structural kinetic model of Calvin cycle in chloroplast stroma [117].

Dynamic flux estimation (DFE). Given the co-existence of very many, very different mathematical representations for metabolic processes, and the fact that none of these are mathematically guaranteed to be correct, with the exception of a small range of validity about an operating point, one might ask whether one can obtain a glimpse of true representations directly from data. A related question is whether it is absolutely necessary to specify functional formats before one starts modeling. DFE offers some answers to these questions.

DFE is a dynamic modeling approach that requires good time series of metabolite concentrations, and uses these to circumvent the initial need for selecting suitable functional forms and their parameterizations [118]. DFE does this by algebraically isolating each flux, and deriving graphical and numerical flux–substrate relationships, in the following manner. First, the time series data are smoothed, and the slopes of the time courses are numerically estimated at many time points. These slope values are substituted

for the derivatives on the left-hand sides of the differential equations of the system. The result is a large algebraic system of equations that represent the pathway at many time points. This system is solved, and the result is a set of arrays that assign flux values to time points or to metabolites on which they depend. These results can be plotted, which reveals flux representations that are presumably very close to the truth. In an independent second step, one attempts to represent the numerical flux representations with parametric functions. In addition to being close to the truth, DFE minimizes compensation errors that commonly arise in a simultaneous parametrization of systems. A drawback of DFE is that its direct implementation would require a square stoichiometric matrix. However, several procedures have been proposed [120, 138-141] to relax this assumption, which is seldom true. Chapter 5 will detail one such method.

Nonparametric dynamic modeling. When it comes to selecting the functional format of rate laws in DFE, there is no silver bullet. Whether mechanistic or non-mechanistic, any rate law needs to be mathematically specified, and then parameterized. A rare exception is the recently proposed method of nonparametric dynamic modeling, which circumvents the need to select functional forms by deriving and utilizing their shapes directly from time series data [119]. Chapter 6 will describe this method in detail. This method is a direct variant of DFE that uses the same initial steps, but then replaces the choice and fitting of functional forms with look-up tables that were derived from the data.

Specifically, the look-up tables are assembled from the flux–substrate relationships that were established in the first phase of DFE. The information in these look-up tables consists of discrete points on curves or surfaces representing flux values throughout the ranges of the experimental time series data. The numerical solver for the otherwise typical

ODEs is discretized, and uses the look-up tables instead the closed-form rate laws to calculate flux values at each iteration. Because the method depends so strongly on available data, it is numerically valid only over the given experimental ranges and close-by. Although the nonparametric character of the method might appear to be a limiting factor, this type of dynamic modeling is surprisingly accurate and powerful. For instance, it is possible to perform stability and sensitivity analysis, and to compute steady states from non-steady-state data. By its definition, nonparametric dynamic modeling is an essentially unbiased approach that is almost free of assumptions.

2.2.3 Other Approaches: Stochastic, Spatial, and Multi-Scale Models

Stochastic models. Models containing randomness have been studied for a long time within the realm of statistical analyses of stochastic processes. In the context of plants and crops, Hartmann and Schreiber analyzed sucrose degradation using various formalisms, including stochastic Petri net (SPN) simulations in potato (*Solanum tuberosum*) [142]. Wu and Tian developed a stochastic multistep modeling framework to improve the accuracy of delayed reactions [143], and applied their method to the aliphatic glucosinolate biosynthesis pathway. One notable phenomenon in plants is the circadian clock, which has been studied in detail in the bread mold *Neurospora crassa* [144, 145]. The review by Guerriero *et al.* [146] presents examples of stochastic models that investigate the effects of intrinsic noise in these circadian rhythms (see also [147-149]).

Spatial models. The importance of spatial assumptions in a plant metabolic model was discussed earlier. To capture the highly compartmentalized environment of plant cells

demands a more detailed approach than is possible with the models discussed so far. A good example in this category is the work by Bogart and Myers [63], who constructed a spatial model of a maize leave to explain its metabolic state in response to a developmental gradient observed between base and tip tissue. The review by Sweetlove and Fernie [150] gives an overview of spatial modeling in plants, and identifies experimental and computational challenges that must be overcome before a realistic spatial or compartmental model for plants can be achieved.

Multi-scale models. Multi-scale modeling is challenging because different aspects of a system in space, time and biological organization require different degrees of granularity. For instance, a model that includes a wide range of time scales suffers from necessary compromises between high temporal resolution in detailed modules, and coarseness at a higher level; it may also have problems with stiffness. By contrast, a model focusing on a very narrow time scale might not capture the essence of a system's dynamics. Therefore, the ideal may be a hierarchical hybrid modeling scheme, with an ensemble of modules where each module covers a certain time scale, which often aligns with corresponding spatial and organizational scales.

In spite of these challenges, prominent examples of multi-scale modeling have been elaborated in the form of multi-organ, and even whole-plant models. A multi-organ FBA model by Grafahrend-Belau *et al.* [151], was developed to investigate the metabolic behavior of source and sink organs during the generative phase in barley (*Hordeum vulgare*). The SOYSIM project is a whole-plant model developed by University of Nebraska at Lincoln, that simulates the soybean growth from emergence to maturity [58].

WIMOVAC is a simulation model of vegetation responses to environmental changes; it focuses specifically on the carbon balance in plants [57].

2.3 Models of Lignin Biosynthesis: Data Needs for Different Modeling Approaches and Uses of Model Output

2.3.1 An Ideal Dataset

In an ideal modeling world, experimental teams would be able to measure every piece of information needed to create a comprehensive model. The data would be of high quality, obtained *in situ,* from the same species and from multiple organisms. Obviously, this high bar cannot often be reached, and one must ask instead what compromises are still sufficient for modeling. We discuss this issue in the following.

To design and explore a model with computational methods, one needs to choose proper functional forms for the fluxes and determine their parameters. In a true mechanistic model, the mathematical format of a flux corresponds directly to the alleged biophysical or chemical mechanism, and typical parameters may be pH and temperature, and more specifically for metabolic models, may include quantities such as $V_{\max}$, K_M, K_{cat}, or K_i, which correspond to rates and affinities in conceptual frameworks like the Michaelis–Menten mechanism.

In an idealized modeling situation, two scenarios can lead to a full model. First, knowledge of all metabolite concentrations and of all mechanisms, including input to the

system, along with a complete set of physical and kinetic parameters, measured *in vivo*, can quite easily be converted into a comprehensive model. However, even in this quite unrealistic case, the model would ignore the spatial distribution of processes and stochastic events, which could, for instance, be due to environmental randomness or to very low numbers of enzyme or substrate molecules. Second, knowledge of all fluxes of the system and a complete set of measured physical parameters would allow the design of the model, again with the same limitation as before. At present, neither scenario is realistic, and missing information must be obtained from other sources, such as *in vitro* measurements, or inferred through computational means.

At this point, many modeling approaches and methods are readily available that could create functioning models out of such data, if they were available. However, they are not, and the more important point therefore is to realign the existing modeling techniques with the realities of data acquisition in a field where some of the key metabolic intermediates are below the level of solid quantification.

As a premier example, flux balance analysis (FBA) [86] and its extensions are based on a mathematical framework that allows assessments of the distribution of fluxes within a metabolic pathway at a steady state under the assumption of an alleged objective of the cell or organism, such as maximal growth, the maximal efflux of some metabolite, or the production of a compound like lignin. FBA formulates the operation of the pathway system as a so-called "linear programing problem" that optimizes the chosen objective, while satisfying biological constraints, such as non-negativity and maximal magnitudes of fluxes.

FBA is a computationally simple, yet powerful tool that has been widely used in many contexts, including plant systems. For instance, in a plant context, Paez et al. [87] analyzed biomass synthesis in *Chlamydomonas reinhardtii* under different CO_2 conditions, and Chang *et al.* [88] presented a genome-scale metabolic network model of the same organism. An interesting variation of FBA is the method of minimization of metabolic adjustment (MOMA) [93], which in a mutated organism tries to emulate a flux distribution that most closely mimics the wild type. Lee *et al.* [58] used MOMA to analyze data from knock-down experiments with genes associated with lignin biosynthesis in alfalfa.

While FBA and MOMA focus on the important distribution of fluxes at a steady state, dynamic modeling attempts to capture time-dependent changes in metabolites following any sort of perturbation. The hope is not only to understand short-term responses better, but also to capture regulatory features of the pathway system that are likely to become critical when the system is mutated. Expressed differently, FBA by and large assumes that everything in the organism remains the same, except for the mutated process and its direct derivatives, although it is to be expected that the organism will attempt to regain normalcy upon such a perturbation by evoking compensatory mechanisms. Thus, dynamic modeling is in principle more powerful but requires much more data support.

In the following, we describe case studies addressing lignin biosynthesis in different plants and with different methods. As stated before, we will focus primarily on data needs and different model uses.

2.3.2 Use of in vitro Data

At present, metabolic modeling is far from having access to ideal comprehensive data obtained *in vivo*. To overcome this challenge, a common approach is the use of *in vitro* equivalents. An excellent example of this strategy in the context of lignin modeling is the work by Wang *et al.* [59], who constructed a dynamic model based on kinetic reaction and inhibition parameters of pathway enzymes in the black cottonwood, *Populus trichocarpa*. The authors derived 189 kinetic parameters associated with generalized Michaelis–Menten mechanisms, primarily in the form of K_{cat}, K_m, and K_i of the 21 enzymes involved in monolignol biosynthesis. They also measured absolute enzyme quantities using mass spectrometry. Furthermore, the authors used a measured S/G ratio to quantify the input flux with a customized optimization algorithm. Such optimization methods are often needed in large-scale metabolic modeling, because the number of fluxes is typically greater than the number of metabolites, which creates a mathematical situation that cannot be directly solved. The information from their experiments allowed Wang's team to construct a fully parameterized model with estimated input flux, which they formulated as ordinary differential equations (ODEs). They were able to obtain the steady-state flux distribution and to investigate the effects of enzyme perturbations on lignin content and composition.

In principle, the well-established strategy used by Wang's team is excellent, as it leads to a fully dynamic model that permits explanations and predictions. The somewhat disconcerting issue is the use of *in vitro* data, which at present seems unavoidable, but leads to the following questions: (1) To what extent are *in vitro* data accurate and representative of the pathway behavior *in vivo*, and does enough *in vivo* information exist to validate the results of such models? In other words, it is unclear how to assess the reliability of these

models. (2) It is clear that no biomathematical modeling effort can presently claim to have taken all components and modulators of a pathway into account. Thus, is it possible to ensure that all relevant information is present quantitatively to reproduce and explain in vivo observations? Or is it simply not feasible to reconstruct the complex *in vivo* cell environment with sufficient reliability from in vitro information? For example, Wang *et al.* did not include the enzyme caffeoyl shikimate esterase [152] in their poplar lignin model [59]; this enzyme was discovered as a new component in the lignin pathway while their studies were ongoing.

These concerns are not exaggerated and can even be found in a very detailed microbial investigation by Teusink *et al.* [111], which provides a good perspective in this regard based on the much simpler pathway system of glycolysis in baker's yeast, *Saccharomyces cerevisiae*. Specifically, these authors compared *in vivo* flux and concentration profiles with the results of a computational model that had been constructed based on the best available kinetic parameters obtained *in vitro*. Despite the authors' dedicated efforts to use the same yeast source and obtain measurements under the same assay conditions, the discrepancies between the model results and the observed *in vivo* behavior were alarming. For possible explanations, Teusink *et al.* pointed to potential factors that may be active *in vivo* and cause uncertainties that are almost impossible to implement in *in vitro* models. Some of these uncertainties are apparently not adjustable by tuning of rate constants or through modifications in the model structure, but may be due to complicated combinations of molecular interactions between the pathway metabolites and enzymes or agents outside the investigated metabolic pathway. The authors proffered that

these small details might have caused drastic differences during the integration of *in vivo* information into systemic models.

Similar concerns about in *vitro–in vivo* extrapolations were voiced some while ago [153, 154], while others came to the conclusion that such an extrapolation is, by and large, justified in many cases [155, 156]. In any case, these problems are disconcerting, as *in vivo* measurements are incomparably more difficult to perform than experiments *in vitro*. Then again, if it is only possible to obtain *in vitro* data, are there means of *in vivo* validation? It appears that a direct validation of individual process representations will be difficult. Thus, one must hope that different types of *in vivo* data may fill the gap as they are combined with models designed from *in vitro* data for more reliable results.

2.3.3 Use of Limited in vivo Data

2.3.3.1 Lignin Synthesis in Poplar

At present, the total body of *in vivo* data is dwarfed by information obtained *in vitro*, and this situation is not likely to change any time soon. As a case in point here, the lignin pathway simply does not permit many concentration or flux measurements *in vivo*. Instead, the typical dataset that can reasonably be expected today consists of lignin content and composition under different conditions, possibly augmented with a few metabolite concentrations. Although this situation might seem to be quite dire for modeling, mathematical and computational approaches can still offer interesting results.

As a pertinent example, Lee and Voit [57] investigated the lignin biosynthesis pathway in *Populus* xylem based on a relatively small set of data consisting of the S/G

ratios and down-regulation levels of enzymes in five transgenic plants. In addition, the authors used information regarding the pathway stoichiometry, regulatory information of five enzymes of the pathway, and an enzyme capacity measurement for COMT. Utilizing a predetermined lignin monomer composition as input and maximum lignin production as the cell's alleged objective, the authors were able to generate steady-state flux distributions through FBA methods. Furthermore, to convert this information into a dynamic model, they employed a strategy derived from biochemical systems theory (BST) [121, 123, 157, 158]. In this modeling framework, all fluxes are represented with power-law functions, so that the parameters can be coarsely estimated without knowledge of direct measurements.

Once the model was fully parameterized, the authors were able to run simulations that ultimately reproduced the lignin composition in all measured transgenics. In fact, an entire ensemble of models was generated, rather than a single model with a unique set of parameter values. This ensemble of models was validated against two transgenics that had not been used to set up the model. Upon validation, an indirect optimization method [159] was implemented to propose enzyme profiles that were expected to lead to a minimal S/G ratio in order to minimize recalcitrance. Single, double, and triple enzyme alterations were conducted to give insights and to determine the most effective perturbations. An interesting detail to note is that the best triple mutation did not contain the double mutation plus an additional mutation, but a different set. Specifically, in comparison to the wild-type S/G ratio of about 1.8, the model predicted a minimal S/G ratio of about 1.3 for two modifications, namely reduction of COMT and CAld5H activities, but a minimum of about 1.1 for three modifications in which the activities of C4H, CAD, and CAld5H were somewhat increased. These computational predictions have not been tested in actual plants.

The fact that a model is able to predict the results of perturbations is intriguing, especially because the power-law representation does not explicitly model specific reaction mechanisms, but only the overall effect of a metabolite or regulator on a given process. Then again, the *in vivo* data used to formulate and instantiate the model encapsulate in some sense everything occurring in the plant, which is not the case for *in vitro* models. Encouragingly, the estimated parameters in Lee's analysis are in agreement with biochemical knowledge of the pathway and provide new insights into the dynamics of the pathway (see results in [57]). Similarly, the predictive capacity of the model to characterize the best candidates for gene alterations is interesting, but it remains to be seen whether explanations and hypotheses obtained with the model are comparable with those obtained with a model like Wang's [59], which was based on experimentally laborious in vitro data.

2.3.3.2 Lignin Synthesis in Alfalfa

The structure of the lignin biosynthesis pathway and its regulation in alfalfa (*Medicago sativa* L.) are fairly well known, but some observations on transgenics were confusing as they seemed to contradict the pathway structure. In particular, some gene knock-downs led to different S/G ratios even though they occurred before the branch point where the pathways toward S- and G-monolignols diverge. Lee *et al.* [58] set out to investigate this situation, using an *in vivo* dataset of lignin content and composition in eight stem internodes in wild-type and seven transgenic lines (with reduced PAL, C4H, HCT, C3H, CCoAOMT, F5H, or COMT activity). The internode classification in this case provided the opportunity to characterize the differential biosynthesis of lignin during the maturation of stem tissue.

Without formal computation, an analysis of the logic of the pathway topology mandated the reversibility of the enzymatic steps catalyzed by HCT and C3H (Figure 1.1), which had not been considered before. Taking this reversibility into account did not resolve the puzzle regarding S/G ratios though. Thus, the authors constructed a computational model of the pathway by first using FBA to compute the steady-state flux distribution in wild type, and then applying the method of MOMA [93] to analyze the redistribution of fluxes in transgenics. This analysis revealed that the results regarding S/G ratios in transgenics could not be explained unless functional channels were active to partition the pathway flux into dedicated S- and G-pathways.

Using statistical analysis, the authors showed that there was a strong correlation between the flux catalyzed by CCR1 and the flux of the consecutive reaction catalyzed by CAD in all strains except for the CCoAOMT-deficient line. This curious result indicated a lack of product exchange between coniferyl aldehyde produced by either COMT or by CCR1. To examine this situation more carefully, the authors tested the possibility of kinetic regulation by the CCR2-COMT and CCoAOMT-CCR1 routes (Figure 1.1), but extensive Monte-Carlo simulations indicated only a very remote possibility of kinetic regulation by substrate/product interactions. Instead, the analysis suggested regulation by one or more distant metabolites. The authors proposed that salicylic acid (SA) could act as the potential regulator of the pathway leading to S-lignin synthesis. Indeed, experimental data characterizing the correlation between SA and lignin content supported the computational hypothesis. Moreover, additional *in vivo* data, demonstrating the co-localization of COMT and F5H [160, 161], provided further evidence supporting the channeling hypothesis.

Wang *et al.* [59] criticized Lee's approach on grounds that the method was rather indirect and, in particular, suggested that a complete kinetic model would be able to capture the experimental data without the need for channeling. While the existence of channels awaits further validation with direct experimental means, it is unclear whether a bottom-up kinetic approach would have led to the crisply targeted hypothesis of differentially regulated channels directing flux toward either S- or G-lignin.

In a different study, Lee *et al.* [18] investigated the channeling hypothesis in *Medicago* by setting up an ensemble of dynamic kinetic models in 19 pathway configuration variants. Each of these variants preserved mass conservation, while allowing alternative routes including one or two metabolic channels across coniferaldehyde (Figure 2.1). The models also examined the presence or absence of putative regulatory mechanisms. Extensive Monte-Carlo simulations over a biologically meaningful range of kinetic values identified only 6 among the 19 plausible configurations as feasible and demonstrated that only 4 out of 16 combinations of plausible regulatory mechanisms could match the experimental data. A graph analysis of these six configurations showed that they were topologically closely related and corresponded to a closed network, if closeness between two configurations was defined as a difference in only one enzymatic reaction. Interestingly, all six feasible configurations in the analysis included one or both proposed metabolic channels.

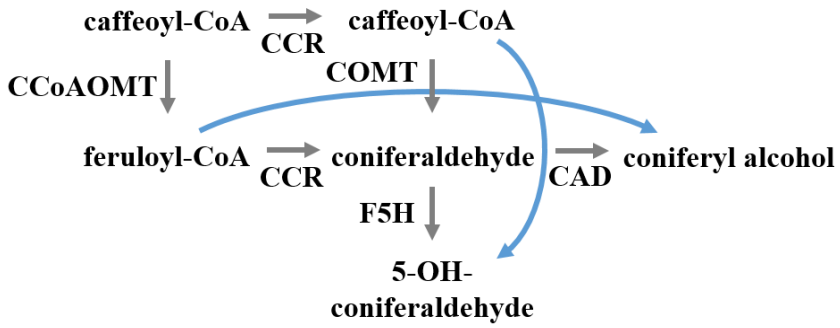

Figure 2.1 Metabolic channeling in *Medicago* proposed by Lee *et al.* [57]. The two crossing channels are associated with coniferaldehyde (see Figure 1.1)

While the computational results strongly suggest the existence of channels, and independent experimental evidence supports these results [58, 160, 161], it is of course imaginable that other explanations could be found for the counterintuitive data in alfalfa, because even the best model fit to data can never offer a guarantee that the model is in some sense correct or that there could not be other models satisfying the same data in a similar manner. It is interesting though that the computational results were inferred directly from actual data from these same species and with a minimum of assumptions, whereas models based on *in vitro* data, obtained from bacteria, should be validated in the target species *in situ*, before they can be considered true. Furthermore, while the power-law formulation used by Lee is mathematically guaranteed to be correct at an operating point of choice, there is no such guarantee for Michaelis–Menten functions; in fact, it is clear that their underlying assumptions and prerequisites are violated in situ [116, 153, 162].

2.3.3.3 Lignin Synthesis in Switchgrass

Similar to the investigations on poplar and alfalfa, a limited dataset characterizing lignin content and composition was available for switchgrass (*Panicum virgatum*) [3], one of the most promising plants in bioenergy research. This dataset was used to set up a model of lignin biosynthesis and to examine for this species the hypothesis of channeling at a diverging branch point, leading to either S- or G-lignin. Specifically, wild-type and four transgenic (4CL, CCR, CAD, and COMT) lignin profiles were analyzed with FBA methods to compute steady-state flux distributions. The stoichiometric model included three variants permitting alternative, slightly differing pathways with and without a hypothetical metabolic channel comprising CCR and CAD. Extensive Monte-Carlo simulations generated thousands of random kinetic parameters to test whether any of the three configurations could reproduce the experimental data in a dynamical manner. Surprisingly, none of the configurations was able to capture the increase in H-lignin in 4CL-transgenics. Instead, the computational results suggested the necessity to include product inhibition by downstream pathway metabolites, as well as substrate competition between CCR substrates. These computational suggestions identified *p*-coumaroyl-CoA and feruloyl-CoA as possible regulators that were arguably necessary to reproduce the observed increases in H-lignin. The model also revealed that the reaction catalyzed by 4CL, which converts ferulic acid into feruloyl-CoA, constitutes an impediment for explaining the counterintuitive accumulation of ferulic acid in COMT transgenics.

Further computational analysis suggested the accumulation of some so-far unidentified metabolite as an inhibitor of 4CL and as the mechanism by which ferulic acid is increased. Revisiting the experimental data indicated a slight accumulation of *p*-

coumaric acid and caffeic acid, which was shown to suffice to support the model-based hypothesis. Taken together, the pathway configuration including both the CCR-CAD channel and two independent CCR and CAD reactions, along with the deduced regulatory mechanisms, turned out to be the only structure capable of matching the *in vivo* data. The authors validated the model to some degree by testing the responses to an enzyme expression profile in an independent transgenic PvMYB4 line that had not been used at all to set up the model. Overall, the analysis produced satisfactory results with respect to lignin content and composition, as well as the concentration profiles of several of the pathway intermediates [2, 3]. The switchgrass lignin model is discussed in detail in Chapter III.

2.3.4 Use of Pathway Data and ^{13}C-labeling Data in *Brachypodium Distachyon*

Brachypodium distachyon uses both phenylalanine and tyrosine to produce lignin, with the source affecting the ultimate proportions of different monolignols [92]. Phenylalanine and tyrosine contribute as substrates almost equally to lignin production. However, ^{13}C-labeling experiments reveal phenylalanine is preferentially incorporated into G-lignin, and tyrosine into S-lignin. In the putative lignin pathway, input from phenylalanine and tyrosine merge at the *p*-coumaric acid node. Furthermore, the pathways of G- and S-lignin split at coniferaldehyde and share their precursor. Therefore preferential incorporation of phenylalanine and tyrosine in different lignin units cannot be explained by the putative structure of lignin pathway.

Experimental reports show that three enzymes of the lignin biosynthesis pathway in *B. distachyon*, namely C4H, C3'H and F5H, are bound to the outer surface of the ER, and the remaining enzymes are located freely in the cytosol. To examine whether this

spatial localization of enzymes is key to explain the unintuitive labeling experiments results, a two-compartment model of lignin biosynthesis was designed [1]. As three of the enzymes are bound to the outer membrane of ER facing the cytosol, namely C4H, C3'H and F5H, it is reasonable to consider a gradient, at least, in the concentration of the substrates and the products of the reactions catalyzed by such enzymes. This gradient is the motivation to consider a compartmental model. One compartment is a portion of the volume of cytosol that surrounds ER, in which the aforementioned reactions catalyzed by C4H, C3'H and F5H are taking place. The other compartment is the rest of the volume of cytosol, which C4H, C3'H and F5H are absent in. Although there is no physical barrier between the two compartments, the spatial localization of reactions is modeled by this simplified compartmental model and avoiding a highly complicated nonhomogeneous model, which would require applying 3-D partial differential equations.

Computational results showed that a single-compartment model cannot capture the distinct roles of the phenylalanine and tyrosine pathways. Spatial localization of enzymes and metabolic channeling are critical for explaining the monolignol composition in *B. distachyon*. In addition, partial activity of some of the free cytosol enzymes is necessary at the outer ER surface to explain the data. The *Brachypodium* lignin model is discussed in detail in Chapter IV.

2.4 Discussion and Conclusion

Mathematical modeling in biology is still in its infancy. Especially within the realm of plant and crop science, the number of modeling articles is negligible in comparison to

experimental papers. As a consequence, the collective experience with plant and crop modeling approaches is still limited, and much more practice and many more case studies will be needed to gain a glimpse into the systemic responses of plants to interventions and manipulations. It may even be, as some experts claim (Leroy Hood, *pers. comm.*), that a "new math" is needed that allows us to combine different data and heterogeneous information in a more efficacious manner than is possible today. Ultimately, a deeper understanding of such responses would allow us to answer questions like "how does 'a' plant react to natural or artificial changes?" or "why does plant (or plant species) A respond differently to a perturbation than plant (or plant species) B?"

To obtain more practice and experience of this type, experimentalists and modelers should collaborate more closely. On the one hand, modelers will need experiments specifically performed for some modeling aspects. At present, many data are available, and the data flow from -omics experiments can be overwhelming. However, not all data are useful for the type of modeling outlined in this article, and modelers will be dependent on experimentalists to perform other types of experiments [64]. On the other hand, experimentalists will want to see genuinely new results coming out of models, especially if they had contributed data to the modeling effort. They will benefit from new, integrative interpretations of their data and from reliable modeling results and computationally achieved hypotheses guiding the "next steps" in their research programs. The generic differences between laboratory or field experiments and computational approaches render it evident that this type of collaboration has true potential, but that it will take time and patience on both sides to make progress toward reaching some of this potential.

As a tangible target, experimentalists and modelers should explore together to what degree metabolic responses can be predicted (qualitatively or quantitatively) from the existence of genes and enzymes (as, for instance, TAL in the case of *Brachypodium*) or from quantitative transcriptomics, where one would expect to find similarities between gene expression and changes in enzyme activities, which however do not always materialize in reality, due to post-transcriptional alterations. It would also benefit both sides to obtain and computationally analyze data describing the same process in different species, as we demonstrated here with the different models for lignin biosynthesis. At first, these comparative analyses could shed light on questions such as: whether apparent differences between species are experimental or modeling errors; whether different designs have evolved in line with the general phylogeny of these species or whether they are due to other factors; and whether different natural designs are dictated by different environmental needs or demands. Together, these combined analyses would have the potential of revealing design principles that govern these processes and could provide deep explanations for why certain species solve a task in the observed fashion and not in a different fashion.

CHAPTER III

Lignin Synthesis in Switchgrass[3]

3.1 Introduction

Switchgrass is a prime target for biofuel production from inedible plant parts and has been the subject of numerous investigations in recent years. Yet, one of the main obstacles to effective biofuel production remains to be the major problem of recalcitrance caused by entanglement of cellulose and hemicellulose content of the cell by lignin. The biosynthetic pathway leading to monolignols, the lignin monomer precursors, in switchgrass is not completely known, and difficulties associated with in vivo measurements of these intermediates pose a challenge for a true understanding of the functioning of the pathway.

In this chapter, a systems biological modeling approach is used to address this challenge and to elucidate the structure and regulation of the lignin pathway through a computational characterization of alternate candidate topologies. The analysis is based on experimental data characterizing stem and tiller tissue of four transgenic lines (knock-

[3] The material in this chapter has been published as: 2. Faraji, M. and E.O. Voit, *Improving Bioenergy Crops through Dynamic Metabolic Modeling.* Processes, 2017. **5**(4): p. 61, 3. Faraji, M., L.L. Fonseca, L. Escamilla-Treviño, R.A. Dixon, and E.O. Voit, *Computational inference of the structure and regulation of the lignin pathway in Panicum virgatum.* Biotechnology for Biofuels, 2015.

downs of genes coding for key enzymes in the pathway) as well as wild-type switchgrass plants. These data consist of the observed content and composition of monolignols.

3.2 Results

The results are described in a sequence that follows the step-by-step model design and conveys the rationale for utilizing the observations to remediate discrepancies with the data and for suggesting the investigation of new features to the model in the next step of the analysis. We begin by assessing the pathway structure in switchgrass as it is alleged in the current literature. Next, we examine possible channeling of CCR/CAD, which has been reported for the lignin pathway in alfalfa [8, 18], but not in switchgrass. Even accounting for the possibility of channeling, the experimental data regarding H lignin cannot be captured at this point. Thus, we investigate the effects of product inhibition and competitive inhibition. In the next phase, 4CL inhibition is added as a potential explanation for the accumulation of 4CL substrates, along with a simultaneous decrease in coniferaldehyde in the COMT knockdown. Finally, principal component analysis is performed to investigate the distribution of parameters within the high-dimensional parameter space and to reduce the feasible subspace of parameter values. The results section ends with a validation of the model.

3.2.1 Lignin Biosynthesis in Switchgrass

The traditionally accepted lignin biosynthesis pathway branches at *p*-coumaroyl CoA to provide S and G-lignin precursors (Figure 3.1). The hexagon in this figure shows the details

of this branch point. It was also previously assumed, based on studies in the dicots *A. thaliana* and *N. benthamiana*, that *p*-coumaroyl CoA is converted to *p*-coumaroyl

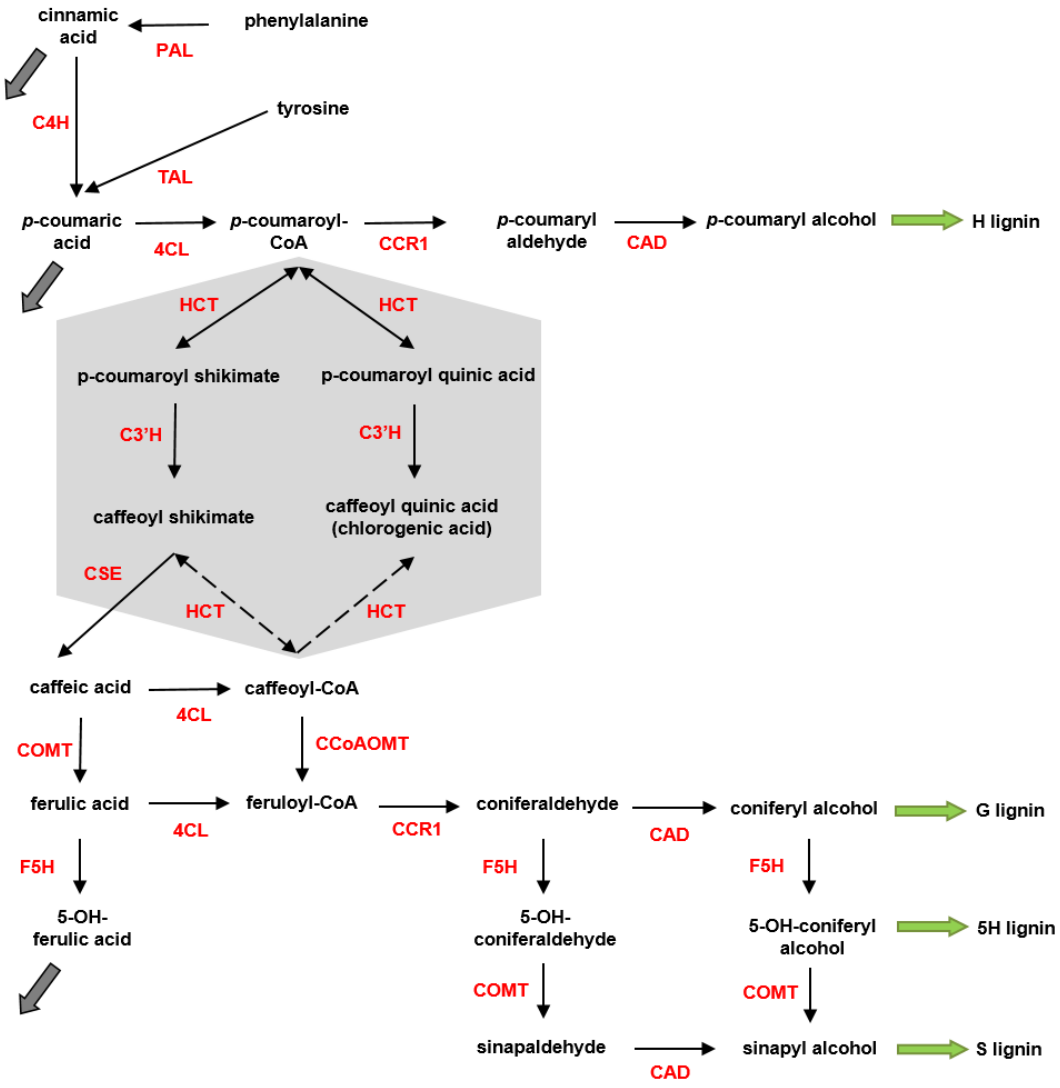

Figure 3.1 Lignin biosynthesis pathway. *Dashed arrows* represent the traditionally accepted pathway of lignin biosynthesis, while the *arrow* from caffeoyl shikimate to caffeic acid captures a newly discovered enzymatic activity [152] now known to be present in switchgrass. Caffeoyl shikimate esterase turns caffeoyl shikimate into caffeic acid and circumvents the previously accepted route. 4CL has recently been shown to exhibit activity towards caffeic acid and ferulic acid in switchgrass by which a new network topology is introduced for switchgrass lignin biosynthesis. Note that tyrosine is shown here, but not included in the model.

shikimate and *p*-coumaroyl quinic acid by HCT. Subsequently, both products, *p*-coumaroyl shikimate and *p*-coumaroyl quinic acid, were shown to be converted to caffeoyl shikimate and caffeoyl quinic acid, respectively [163]. The enzyme for these unidirectional reactions is C3′H. Downstream, HCT was proposed to operate in the reverse direction to convert caffeoyl shikimate and caffeoyl quinic acid into caffeoyl-CoA.

A recent study demonstrated that this pathway organization is unlikely to occur in switchgrass [164]. Based on kinetic measurements of PvHCT1a, PvHCT2a and PvHCT-Like1, it was shown that caffeoyl shikimate is not converted to caffeoyl-CoA by the reverse HCT reaction, but is more likely converted into caffeic acid through caffeoyl shikimate esterase, and that this step is actually the main route of mass transfer into the pathway towards S and G monolignols. As indicated with dashed arrows in Figure 3.1, HCT is not active in the formation of caffeoyl-CoA. This new information helps us reduce the steps in Figure 3.1. It has furthermore been suggested that cinnamic acid is a precursor for salicylic acid; this process is represented by the thick grey arrow [8]. Similarly, a considerable portion of ferulic acid leaves the pathway [165]. Finally, the efflux out of *p*-coumaric acid acts to avoid accumulation of the metabolite in the 4CL knockdown strain (Figure 3.1). These simplifications yield the pathway diagram in Figure 3.2.

At this point, it is not entirely clear whether the lignin pathway in switchgrass contains caffeyl aldehyde. It appears that this is not the case, and the following analysis assumes that caffeyl aldehyde is indeed not produced. Nonetheless, since other species do generate this intermediate, Appendix A: Text A.1 analyzes this case.

Large-scale simulation studies with this pathway structure lead to irreconcilable differences between the experimental data and the model results, which indicate that the model has genuine flaws. In particular, the dynamics of the different lignin species cannot be explained for the various transgenics (data not shown).

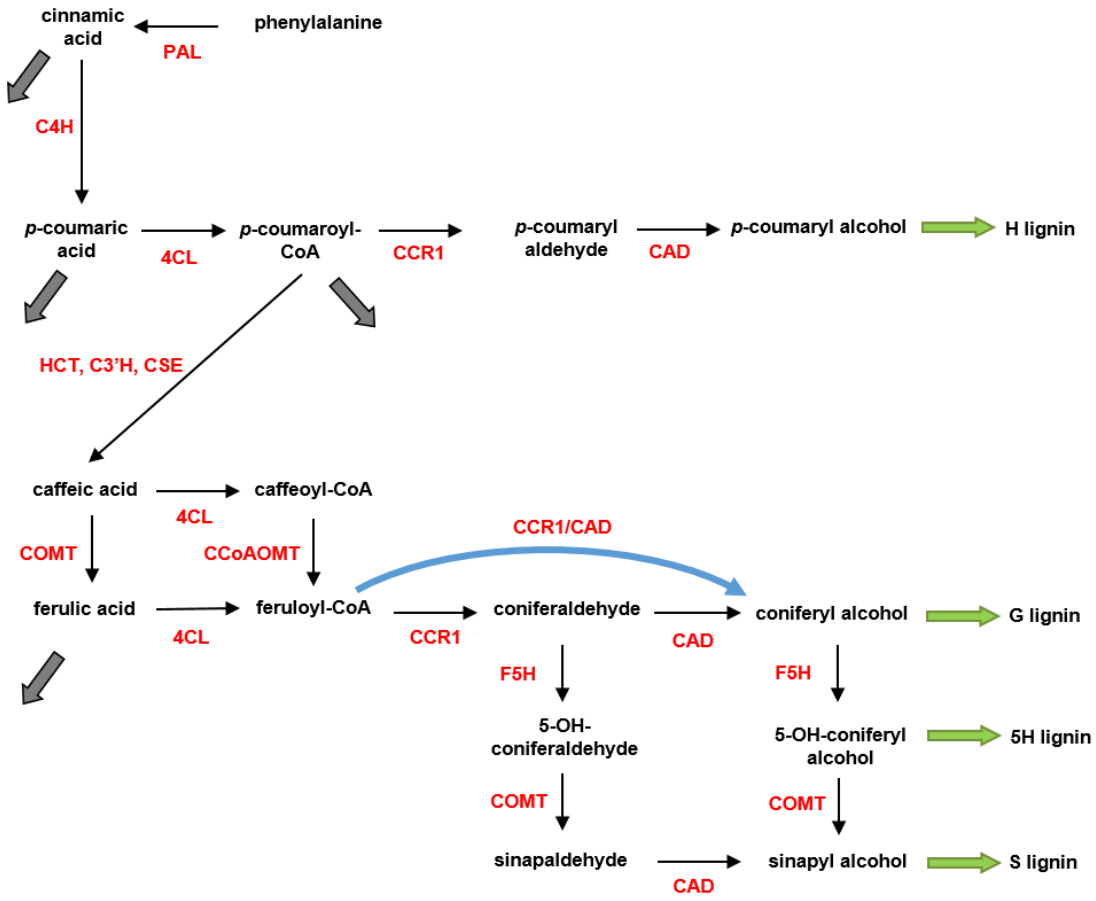

Figure 3.2 Revised and simplified pathway in switchgrass. By eliminating HCT from the diagram in Figure 3.1 and adding CSE, the pathway system becomes simpler. The *right* branch in the *grey box* in Figure 3.1 is merged into an efflux and the *left* branch is simplified to a one-step process. It is hypothesized that a specific functional channel could facilitate the conversion of feruloyl-CoA into coniferyl alcohol. Such a channel could be the result of co-localization of the involved pathway enzymes.

3.2.2 Channeling

Experimental and theoretical work in alfalfa has suggested that functional enzymatic channeling likely occurs at the coniferaldehyde node [8, 18]. According to this suggestion, the "G-channel" facilitates the use of feruloyl-CoA for the production of coniferyl alcohol, which is the precursor of the G monolignol (Figure 3.2). We investigate the same channeling hypothesis here as a possibility. Specifically, we use pertinent experimental data from switchgrass to analyze the feasibility of different hypothetical pathway topologies. The potential existence of a functional complex consisting of CCR1/CAD leads to three possible pathway topologies that satisfy the requirement of mass conservation (Figure 3.3).

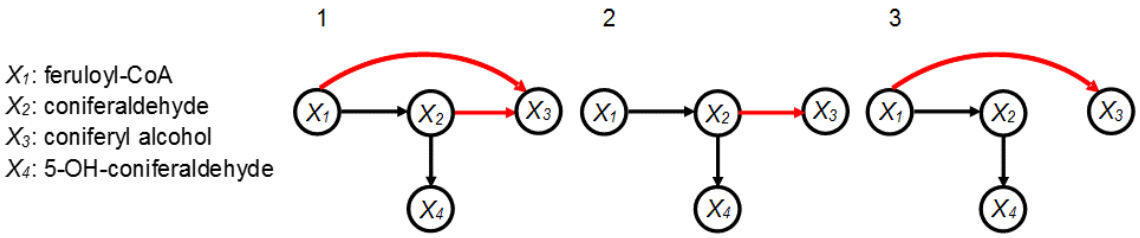

Figure 3.3 Topological Configurations. Three pathway structures are plausible when a CCR1/CAD channel is considered. Configuration 2 lacks the channel, while the other two configurations represent alternatives involving the channel.

Each of these so-far unregulated topologies was modeled as a generalized mass action (GMA) model, whose parameter values were obtained with a sophisticated large-scale sampling scheme (see 3.4 Methods). Although all topologies were found to be consistent with most of the experimental results, no topology was compatible with the

accumulation of H lignin in 4CL knockdown transgenics (Table 3.1); this situation could not be simulated by any of the candidate models, regardless of the presence or absence of the channel. This strong result suggests the existence of regulatory mechanisms, and considering the structure of the pathway and the branch toward H lignin in particular, we decided to analyze the possible role of product inhibition, which is frequently found in pathway systems *in vivo*.

Table 3.1 Fold change in lignin monomers, total lignin, and S/G in transgenic plants relative to wild-type plants

Knockdown gene	**4CL 40 % [6]**	**CCR 50 % [166]**	**COMT 30 % [8]**	**CAD 30 % [56]**
Down-regulation	27–95 %	Up to 75 %	Up to 90 %	55–86 %
H lignin	1.82	NR	NR	NR
G lignin	0.53	~0.75	0.76–0.98	0.67–0.83
S lignin	1.00	~0.75	0.42–0.96	0.58–0.87
Total lignin	0.78	~0.75	0.84–0.96	0.78–0.86
S/G	Increased	Increased	Decreased	Decreased

NR not reported

3.2.3 Product Inhibition

Experimental results from transgenic plants have demonstrated that H lignin accumulates when the enzyme 4CL is down-regulated [6]. Analyzing this initially counterintuitive

observation closer suggests that there might be a wave of accumulation in the metabolites preceding H lignin. Such a wave can be explained with product inhibition (Figure 3.4). When an enzyme is down-regulated, the corresponding substrate accumulates. The secondary effect is that the accumulated substrate is by itself a product of a previous reaction whose increased concentration decreases its own rate of production. This backward cascade has an upstream domino effect along the pathway and, depending on the kinetics of the reactions, can lead to the accumulation of upstream metabolites. This observation can be explained by the following chain of events: Down-regulating 4CL leads to a decrease in the products of this enzyme, i.e., *p*-coumaroyl-CoA, caffeoyl-CoA, and feruloyl-CoA. At the same time, product inhibition leads to a backward accumulation in upstream metabolites, which compensates, at least partially, for the initial decrease in *p*-coumaroyl-CoA. Product inhibition is easily incorporated into the GMA model (see 3.4 Methods). Thus, in a new round of simulations, a new set of 100,000 randomly sampled parameter values was generated as before, this time accounting for product inhibition. Again, the configurations satisfying the experimental results were recorded.

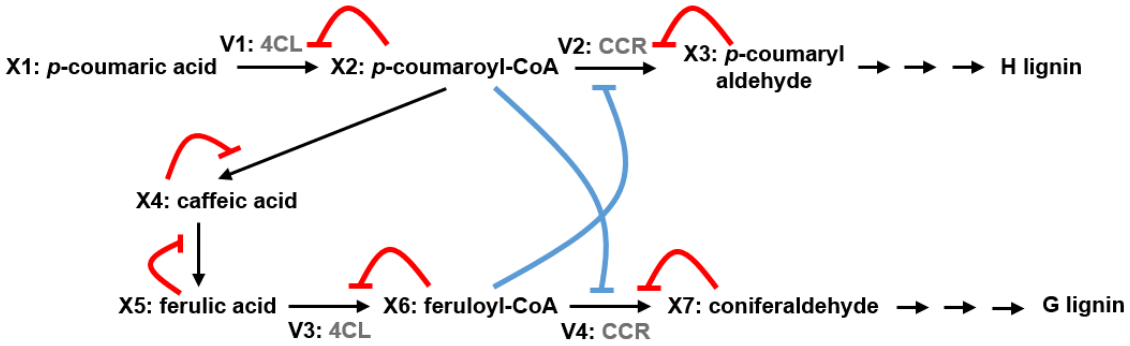

Figure 3.4 Substrate competition for a shared enzyme, combined with product inhibition. The accumulation of H lignin in the 4CL transgenic line calls for a regulatory mechanism that guides the flow towards the upper branch of the pathway. Direct activation or an inhibited inhibitor can achieve this result. Simulation results support the second option.

Although the simulations showed an improvement regarding the H lignin accumulation in the 4CL knockdown, no topology reached the twofold increase that was reported in the literature [6].

3.2.4 Substrate Competition for Shared Enzymes

Several enzymes in the lignin pathway catalyze multiple reactions with slightly different substrates, and it is reasonable to assume substrate competition for an enzyme among the multiple substrates. This competition can play an important role in altering the flow of mass in a mutant plant.

We explored the consequences of substrate competition with respect to the pertinent enzyme CCR. The analysis yielded the following result. If CCR favors *p*-coumaroyl-CoA over feruloyl-CoA, due to substrate competition, the flux towards H lignin is increased. In

fact, simulation analysis shows that the increase in H lignin is strong enough to match the experimental data.

It could be possible that substrate competition alone would be sufficient for increased H lignin production. We tested this conjecture with a corresponding simulation, which revealed that only the combined model with product inhibition and substrate competition matches the experimental observations. The strength of inhibition is *a priori* unknown, but simply becomes a parameter value in the GMA model (see 3.4 Methods section). For instance, consider the pathway in Figure 3.4, where X_2 and X_6 share the same enzyme for fluxes V_2 and V_4. Blue arrows represent the competition between the substrates, while red arrows represent product inhibition. In this case the equation for V_2 becomes

$$V_2 = \alpha_2 X_2^{g_{2,2}} X_3^{-g_{3,2}} X_8^{-g_{8,2}} Y_2 \tag{3.1}$$

where Y_2 is the enzyme catalyzing the reaction (CCR).

3.2.5 Inhibition of 4CL in COMT Knockdown Transgenics

Although product inhibition and substrate competition improve the consistency between the experimental data and numerical results in CCR1 transgenic plants, the model does not match COMT knockdown data sufficiently well. Specifically, the model does not capture the observed 30% increase in ferulic acid in COMT knockdowns [7]. This observation becomes even more difficult to explain if one considers the simultaneous 20% decrease in coniferyl aldehyde. One could speculate that the high accumulation in 5-OH-ferulic acid

might trigger a cascade of product inhibition that leads to the accumulation of ferulic acid, but computational results did not support the idea.

Further analysis with the model revealed that the reaction from ferulic acid to feruloyl-CoA, which is catalyzed by 4CL, is the bottleneck. Indeed, the computational results show that this reaction has a flux that is 10 times as large as the efflux from ferulic acid towards 5-OH-ferulic acid. Thus, if the flux towards ferulic acid decreases, any substantial accumulation is impossible unless the 4CL reaction is inhibited. This model-based deduction is indirectly supported by experimental data from one of our collaborators' labs that exhibit a slight accumulation in the distant *p*-coumaric acid and caffeic acid, which is explained by 4CL inhibition as well (data not shown).

Accounting for the deduced 4CL inhibition in the model leads to simulations that faithfully capture all experimental data associated with the COMT knockdown; in particular, the 4CL substrates accumulate and the concentration of coniferaldehyde decreases, as observed. From a biochemical point of view, one might be interested in identifying the inhibiting agent. As it was mentioned earlier, the 5-OH-ferulic acid concentration increases by 70% in COMT knockdown plants. While the metabolite has not been identified as a substrate for 4CL, it might be reasonable to assume that it binds to 4CL in high concentrations, due to its molecular similarity, and thereby inhibit the enzyme competitively (Figure 3.5). While this hypothesis remains to be experimentally validated, the same type of substrate competition with respect to 4CL has recently been proposed by others [167]. To implement 4CL inhibition in the model in the most generic manner, we simply lowered the corresponding rate constants.

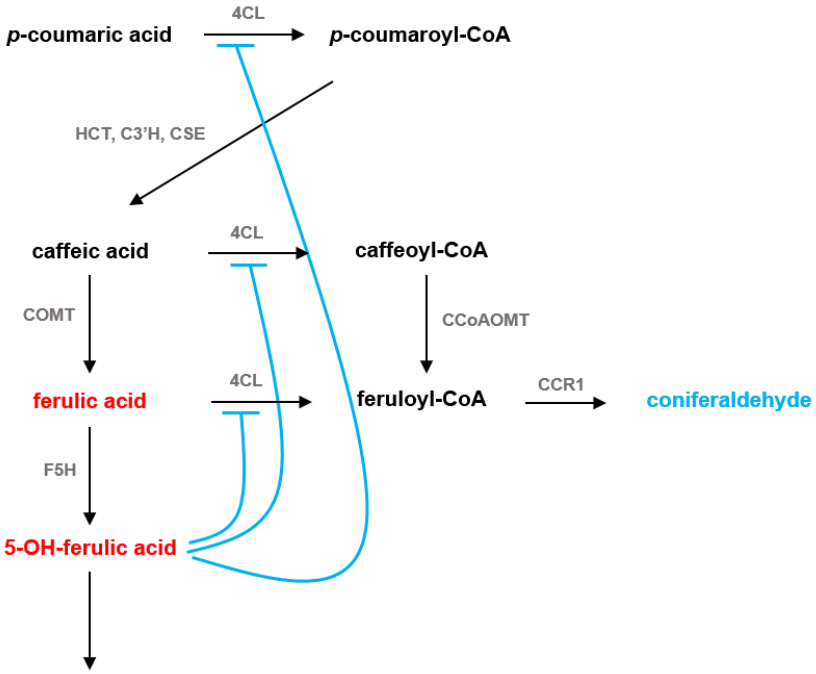

Figure 3.5 Parallel reactions catalyzed by 4CL. The observed simultaneous accumulation of 4CL substrates and decrease in coniferaldehyde in COMT transgenic lines can be explained with the assumption of an inhibitory effect on the reactions catalyzed by 4CL. 5-OH-ferulic acid could be a candidate for this role. Although 5-OH-ferulic acid is not a substrate for 4CL in switchgrass, it has a similar molecular shape as ferulic acid, so that high concentrations of 5-OH-ferulic acid might exert competitive inhibition that is comparable to the inhibitory effects of ferulic acid.

3.2.6 Compatible Configurations

The mathematical model with universal product inhibition, substrate competition for CCR1, inhibition of 4CL, and the possibility of a metabolic channel was subjected to large-scale simulations aimed at inferring the most likely topology of the lignin pathway (recall Figure 3.3). Similar to previous simulations, a sample of 100,000 parameter sets was generated to test model consistency with the experimental data and to provide likely kinetic

orders for the model (see 3.4 Methods). Intriguingly, the only pathway configuration that is compatible with all available data is Configuration 1 of Figure 3.3. Note that the speculated coniferaldehyde channel is indeed present. In fact, no parameter set, using Configurations 2 and 3, could reproduce the experimental data which eliminates the chance to compare the relative performance of the configurations.

3.2.7 Principal Component Analysis

To gain a better understanding of the parameter space of the system, principal component analysis (PCA) was performed on the parameter sets that had been filtered by the model criteria. Once the principal components of the parameter space were identified, a new round of simulations was executed. Specifically, a sample of 100,000 parameter sets was generated along the principal directions and within the reduced space. The set was then transformed back to the original coordinates. The successful parameter sets were recorded and are depicted in Appendix A: Figure A.8. Ultimately, principal components 1 through 4 collectively account for 88 % of the variance.

3.2.8 Model Uniqueness

It is theoretically impossible to proof the uniqueness of a model for such complex nonlinear problem, because it is always possible to evoke additional processes in such a fashion that the original model could be subsumed as a simpler special case. In our case, one should note that our large-scale simulation approach led to a structurally and numerically compact ensemble of similar solutions within the high-dimensional parameter space of the system.

Given that we determined the ensemble with Monte Carlo simulations that cast a very wide net over the parameter space, it is difficult to imagine entirely different parameterizations that would capture all data as well as our ensemble and perform well in the validation studies we performed.

Moreover, considering that the available data were obtained from several independent transgenics, and that the stoichiometric system of the system is underdetermined, the likelihood of significantly other solutions appears to be rather small. Also, our simulations show that the system converges to the same steady-state starting from a wide array of initial conditions. Some arbitrary initial conditions actually lead to steady-state values outside of the defined physiological bounds; however, among the initial conditions that lead to admissible steady-states, several rounds of screening showed identical results.

In summary, it is well understood that model design is an iterative procedure, and while our logical analysis of numerical results suggested the step-wise addition or elimination of new features, there is no mathematical proof that the model ensemble is truly unique.

Outside these purely mathematical arguments, we might also look at the biological reasonableness of the model. For instance, one could ask why only CCR was subjected to substrate competition, while there are other shared enzymes. The answer is a matter of simplicity, as suggested by Ockham's razor. Namely, we demonstrate that the substrate competition of CCR is needed to match the available data, while additional mechanisms are not necessary to explain the experimental data. Thus, we cannot exclude that additional

regulatory mechanisms might exist, but we would need additional, independent data to confirm or refute such a hypothesis.

We also note that, although the model design progressed iteratively, we carefully investigated the necessity of including each individual mechanism *a posteriori*. For example, upon discovering that competitive inhibition over CCR improves H-lignin accumulation, we asked whether product inhibition was still vital for the model to explain the observations. We examined this hypothesis and determined that H-lignin accumulation could not be captured anymore. We therefore concluded that both mechanisms, product inhibition and CCR competition, are necessary. We found this conclusion reasonable, as both product inhibition and substrate competition are common in metabolic pathway systems.

3.2.9 Model Validation

The model with parameter values described above was constructed based on experimental data from wild-type switchgrass and four transgenic lines (4CL, CCR1, CAD and COMT knock-downs). To validate the model, experimental data from a separate transgenic plant, which had not been used in any way during the model design, were used to investigate how well the system performs under untested conditions. Namely, in a recent study, the transcription inhibitor PvMYB4 was over-expressed in order to reduce enzyme expression in the lignin pathway [168]. While metabolite concentrations were not measured for any of the pathway intermediates, the published data contain H, G and S lignin levels, as well as comparisons of enzyme activities between the wild type and PvMYB4 plants. The overall

result of the study is a global reduction in the expression of the enzymes of the pathway, which in turn leads to 40–70% decreases in total lignin.

We tested our model against the profile of observed enzyme expression under overexpression of PvMYB4. We started with the already parameterized model without introducing any alterations or adjustments, except for resetting the appropriate enzyme activities, and tested how the system responded to the inhibition in comparison to the *in vivo* experiments [168]. Encouragingly, the altered G- and S-lignin amounts and their ratio, reported in the experimental study, are captured by the model with the compatible topological configuration quite well. The H-lignin was essentially unchanged in the experiment, while it slightly decreases in our model, in accordance with the data we used. However, H-lignin constitutes only about 3 % of the total lignin so that this difference is of no particular pertinence. Results are shown in Figures 3.6 and 3.7. Figure 3.6 compares the fold change in lignin monomers between the experimental data and model results. The first row shows the fold change in G, S, the total lignin, and the S/G ratio comparing the wild type and PvMYB4 lines from the experiment; the second row corresponds to the computed configuration. As can be seen, the model results are quite consistent with experimental data.

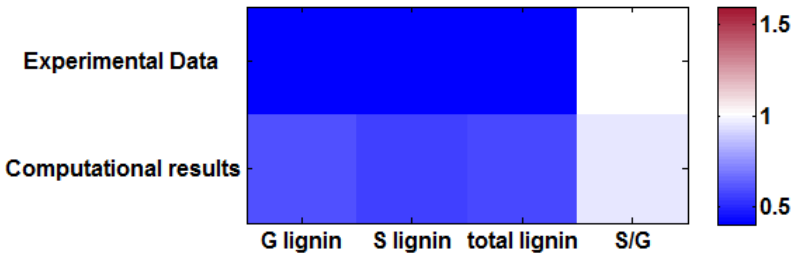

Figure 3.6 Fold changes in lignin monomer concentrations in PvMYB4 transgenic plants. The *top row* represents the average of PvMYB4 plants experimental data normalized with respect to the average of the control plants. The *second row* represent the results of the model with settings corresponding to the PvMYB4 experiment in [168], normalized with respect to wild-type model results. Wild type is set to 1, which corresponds to *white* in the *color bar*. H lignin only counts for 3 % of total lignin and is not shown in here.

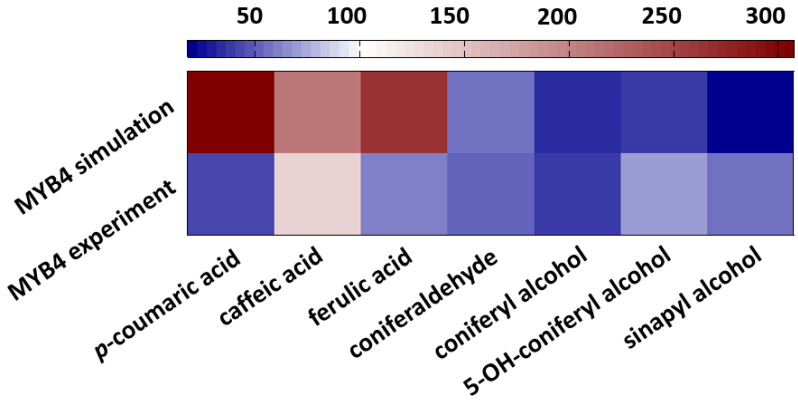

Figure 3.7 Steady-state profiles of key pathway metabolites in PvMYB4 overexpression as predicted by the model. Concentrations are normalized and the base value is set to 100, which corresponds to *white* in the *color bar*. Any increases with respect to the wild-type steady state are reflected in the *red* spectrum and any decreases in the *blue* spectrum.

This independent validation is very reassuring, especially with respect to future attempts to use metabolic engineering techniques to alter the S/G ratio in switchgrass. For instance, if further model predictions prove similarly reliable, the model could be used to simulate

and optimize the outcome of combinatorial knockdowns, whose outcomes are not necessarily predictable with intuition alone. Such predictions would be very valuable, as a comprehensive combinatorial screening of double and triple knock-downs would neither be economical nor experimentally feasible.

While the published PvMYB4 data used for the first validation do not contain intermediate metabolite concentrations, a more recent study provides steady-state data for several of the pathway metabolites [55]. Comparing the published data in [55] with those in our model, we find that seven metabolites are represented in both, namely, caffeic acid, 5-OH-coniferyl alcohol, ferulic acid, sinapyl alcohol, coniferaldehyde, *p*-coumaric acid and coniferyl alcohol.

Figure 3.7 exhibits a comparison of the steady-state profiles. The top row shows the simulation results, while the bottom row represents experimentally measured steady-state concentrations in PvMYB4 normalized to wild type from [55]. The wild-type value for each concentration is set to 100 (white), and the red-blue spectrum represents increases or decreases in steady-state values of knockdowns. For five of these seven metabolites, our computational results of PvMYB4 conditions show the same semi-quantitative behavior in steady-state concentrations compared to the wild type; these are caffeic acid, 5-OH-coniferyl alcohol, sinapyl alcohol, coniferaldehyde and coniferyl alcohol. Discrepancies are seen in ferulic acid and *p*-coumaric acid. Here, the experimental data show a decrease in the steady-state concentrations, while our computational results predict an accumulation. Interestingly, these differences occur for metabolites whose effluxes out of the lignin pathway are ill defined, because their characteristics were not documented in the literature. It is therefore likely that they are not optimally parameterized in the model.

3.2.10 A Library of Virtual Strains

While reducing lignin content is one of the targets in bioenergy science, there is uncertainty in the literature as to whether the total lignin content plays a more important role for recalcitrance than the S/G ratio. It is not even entirely clear whether a higher or lower S/G ratio would benefit the ethanol yield. In fact, there have been contradictory reports in the literature [169, 170]. The computational model is capable of simulating both scenarios, and allows optimization toward either objective.

The developed model is, in fact, an ensemble of model variants that are equally capable of capturing the available experimental data in the wild type and four knockdowns, namely in COMT, CAD, CCR, and 4CL. Here, we use the ensemble to simulate the pathway over a wide range of perturbations, to determine the responses of the system to single or multiple increases or decreases in enzyme activities, and the consequent changes in lignin content and composition.

3.2.10.1 Single Perturbations

In the first set of computational experiments, one enzyme at a time was perturbed up to ±5-fold relative to its wild type activity. Since every scenario is simulated for the entire ensemble of models; the analysis yields many results for each scenario. Using all these results collectively, the median of total lignin content, and the median of the S/G ratio in the perturbed systems are recorded. The medians are normalized with respect to the wild type value, so that value 1 represents the wild type (see section 3.4.3).

The results of this analysis are shown in Figures 3.8 and 3.9. The X-axis shows the $\log_2$-fold change in the amount of a given enzyme, and the Y-axis indicates the specifically perturbed enzyme. The grouping of HCT, C3'H, and CSE represents the flux from *p*-coumaroyl-CoA to caffeic acid, as the corresponding reactions have been merged into one in our model. The color code represents the relative change in total lignin or S/G ratio, for which white depicts no change from the wild type phenotype. The green spectrum (in the panel for total lignin) and the blue spectrum (in the panel for the S/G ratio) represent reductions, while the red spectrum (in both panels) represents an increase relative to wild type. The color bar indicates the intensity of fold change in the enzymes. The greatest lignin reduction achieved is close to 50%, which is the predicted result of an 80% CAD knockdown, with 20% activity remaining. The most significant reduction and increase in S/G ratio is predicted for perturbations in F5H. Thus, the different criteria for total lignin and lignin composition point to different knockdowns.

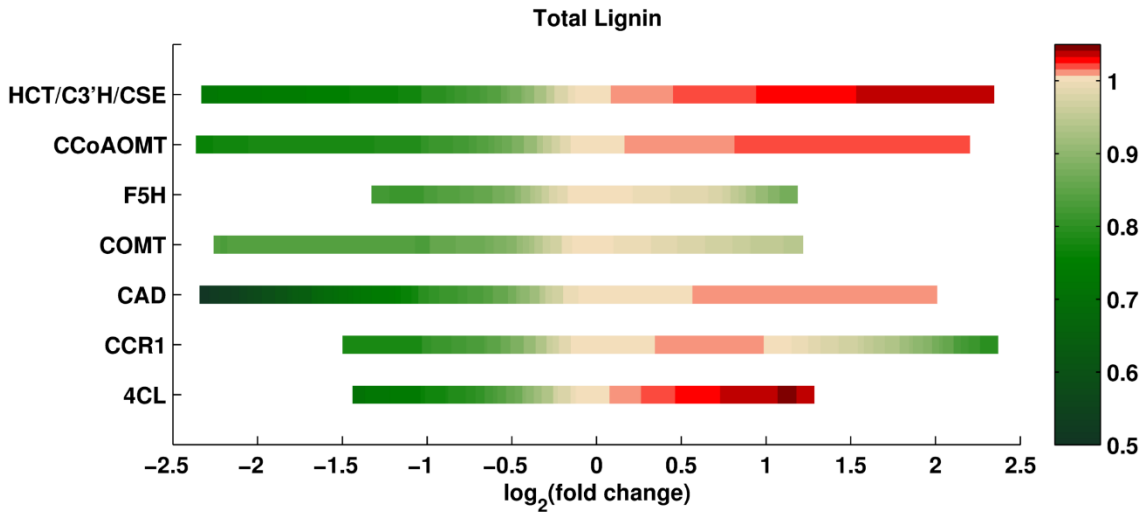

Figure 3.8 Total lignin in response to single enzyme perturbations. The total lignin level is color-coded, where green represents decreases in total lignin, red represents increases, while white corresponds to the wild type level. CAD seems to be the most effective enzyme in reducing lignin. Note that, surprisingly, the change in lignin content is not always monotonic. Simulations show that lignin can be reduced by knocking down or overexpressing the activities of F5H, COMT and CCR.

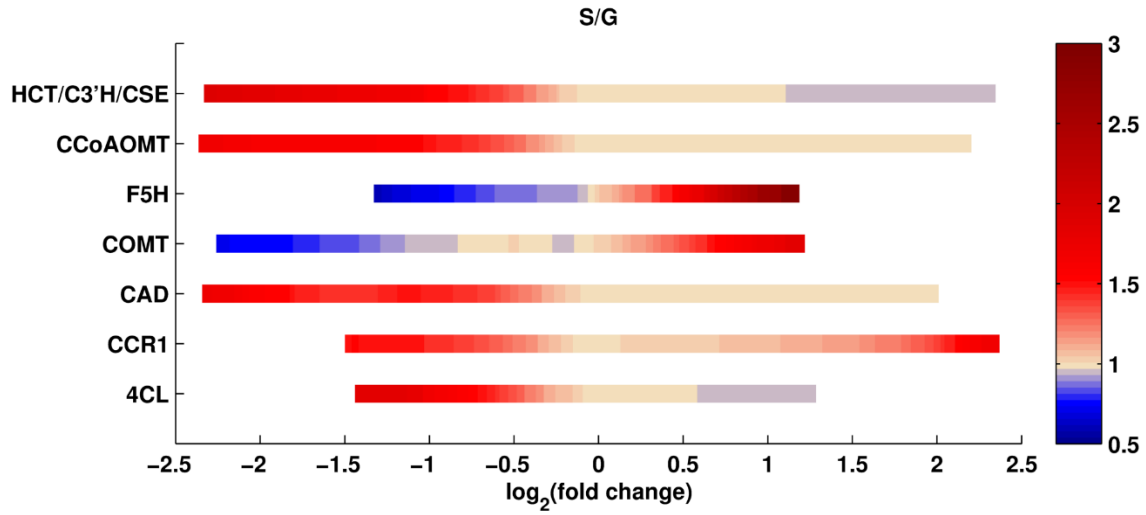

Figure 3.9 S/G ratios in response to single enzyme perturbations. Blue represents decreases in S/G ratios, red represents increases, while white is the wild type base level. F5H seems to be the most effective enzyme for altering the S/G ratio, both toward increases and decreases. Similar to the total lignin response, changes in S/G ratios are not necessarily monotonic: the S/G ratio increases in both knocked down and overexpressed CCR1.

3.2.10.2 Double Perturbations

It seems reasonable to surmise that simultaneous changes in two enzymes might be more effective in altering total lignin and/or the S/G ratio. Thus, we analyzed simultaneous perturbations in pairs of enzymes. The perturbations were again restricted to magnitudes of ±5-fold relative to the corresponding wild type activities. Again, every scenario was simulated with the ensemble of models, and the medians of total lignin content and of the S/G ratios were computed and normalized with respect to the wild type values. The results are shown in Figures 3.10 and 3.11. The X- and Y-axes represent the $\log_2$-fold changes in the perturbed enzyme activities.

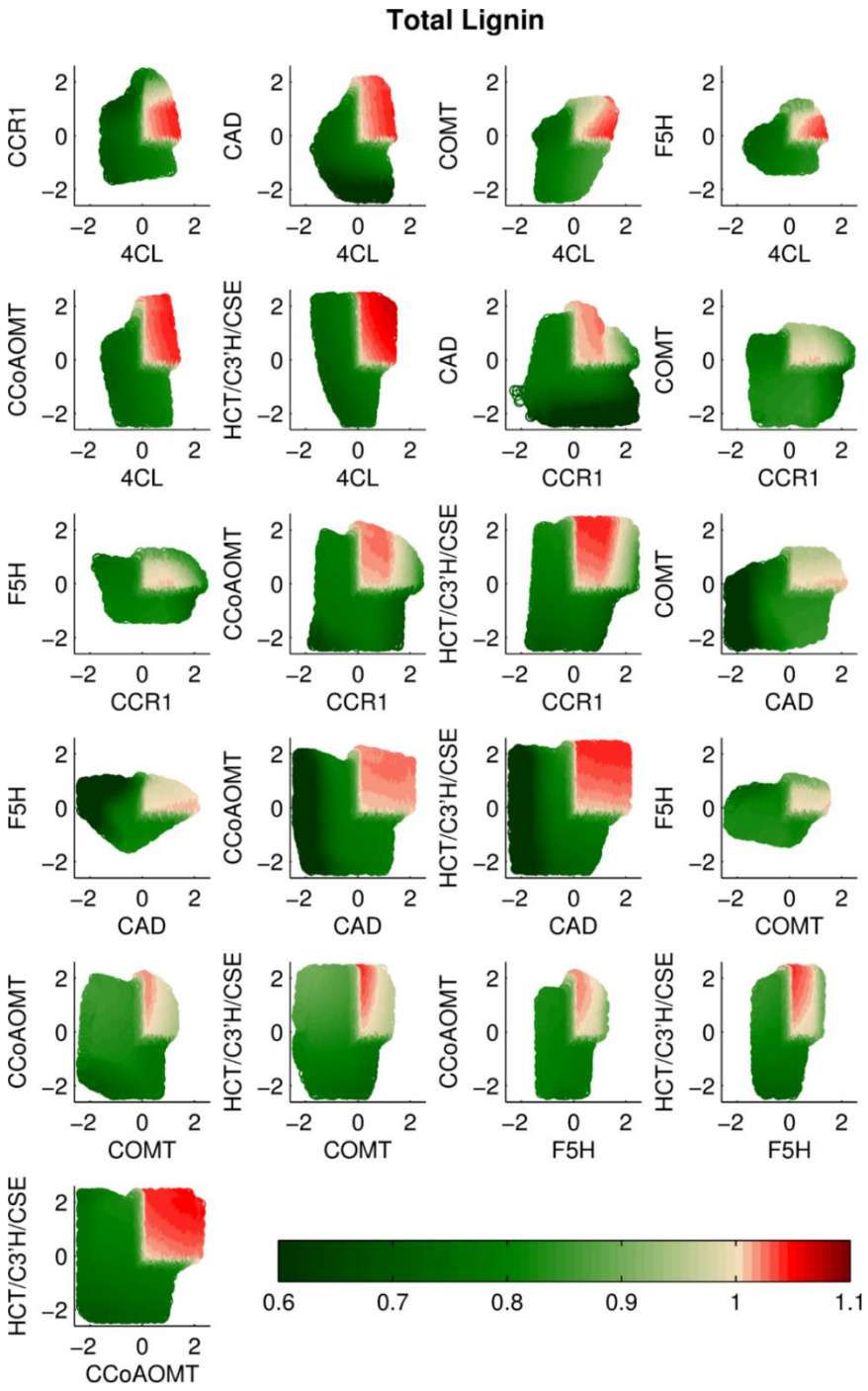

Figure 3.10 Total lignin in double enzyme perturbations. The color code is the same as in Figure 3.8. Pairs of CAD/4CL, CAD/CCR, and CAD/F5H are predicted as the most effective combinations; in particular, the pair of CAD/F5H shows strong synergism: an increase in F5H, combined with a small reduction in CAD, reduces total lignin dramatically. The nonlinear behavior of the pathway is evident in the dual-overexpression scenarios, especially in pairs, including CCR1, COMT, or F5H.

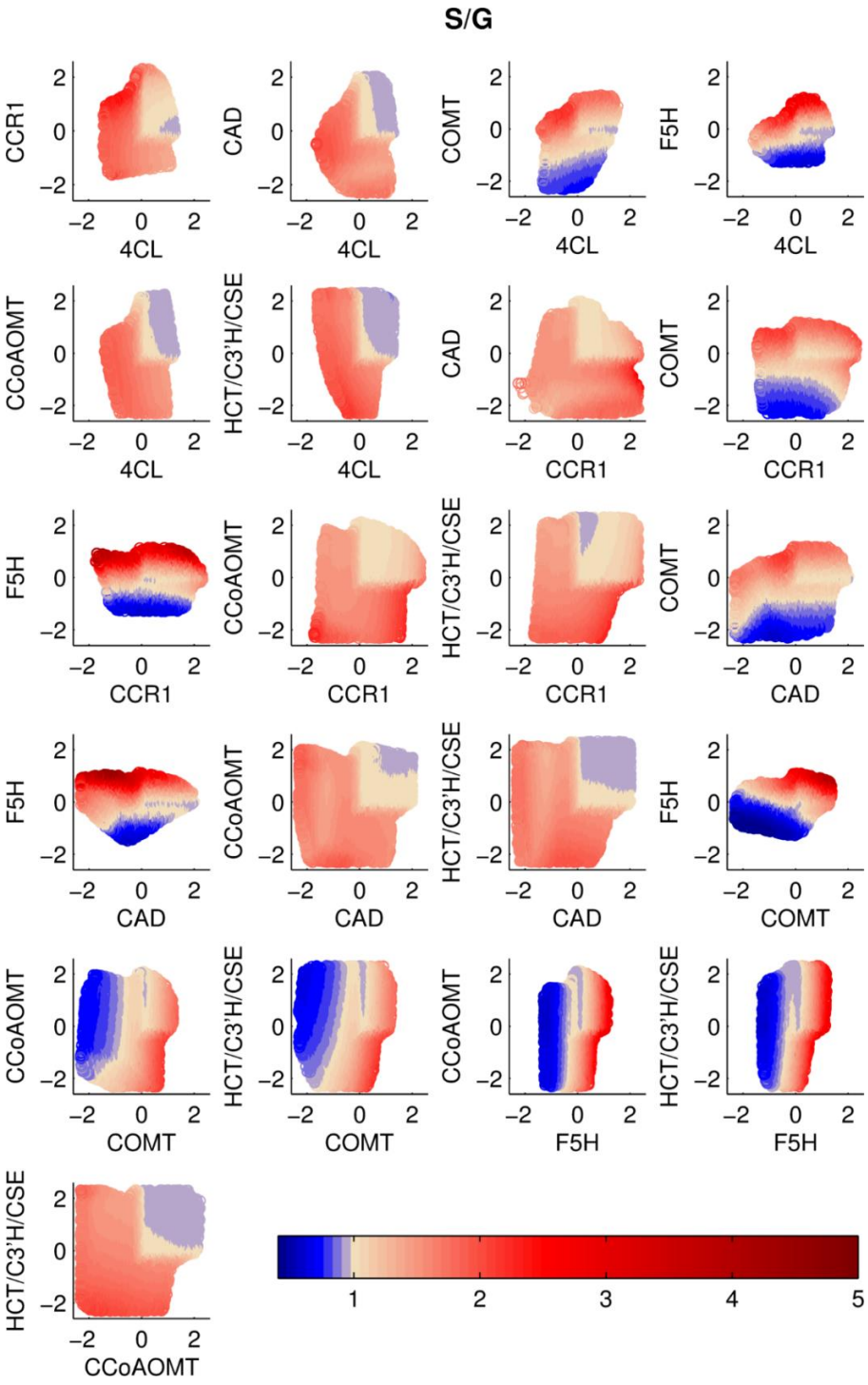

Figure 3.11 S/G ratios in response to two simultaneous enzyme perturbations. The color code is the same as in Figure 3.9. Pairs including F5H and COMT (F5H/4CL, F5H/CCR, F5H/CAD, F5H/COMT, COMT/4CL, COMT/CCR) show the highest changes in S/G ratio. In particular, F5H and COMT work well synergistically, even for moderate perturbations.

It is evident from the results that some perturbations are more effective in altering lignin content and S/G ratio, whereas the system response is more robust to others. It is also clear that many solutions reveal compromises between alterations in total lignin and the S/G ratio. If the S/G ratio is to be altered, F5H seems again to be the key enzyme, and pairs like F5H/4CL, F5H/CCR, F5H/CAD, and F5H/COMT are predicted to be most successful. At the same time, if the goal is solely to reduce lignin, irrespective of the S/G ratio, other solutions exist, including the pairs of CAD/4CL, CAD/CCR, and CAD/F5H.

3.2.10.3 Single Perturbations in a PvMYB4 Overexpression Strain

As it was discussed in 3.2.9 Model Validation, a recent study [55, 168] analyzed the consequences of overexpressing the inhibitor, PvMYB4, of a transcription factor in switchgrass. The main result was an altered expression profile of many of the enzymes involved in lignin biosynthesis. Consequently, the lignin content was reduced to 40–70%, while the S/G ratio remained the same as for wild type. So far, the PvMYB4 strain has been the most effective transgenic line in reducing recalcitrance in switchgrass. To build upon this success, we combined this scenario with additional single enzyme perturbations, and investigated whether it could be possible to improve the results from the PvMYB4 transgenic strain further. As before, each enzyme was perturbed up to ±5-fold relative to the wild type level. Simulation results are shown in Figures 3.12 and 3.13. The color code is the same as for previous figures. The black square in each row represents the enzyme activity in the reference PvMYB4 perturbation. An interesting observation is that the lignin content is predicted to decrease even more than in the reference PvMYB4 experiment if

CCR1 is overexpressed in this background. Additional simulations show that lignin content could be reduced further if CCoAOMT, CAD, or 4CL are reduced to even lower levels relative to the PvMYB4 background.

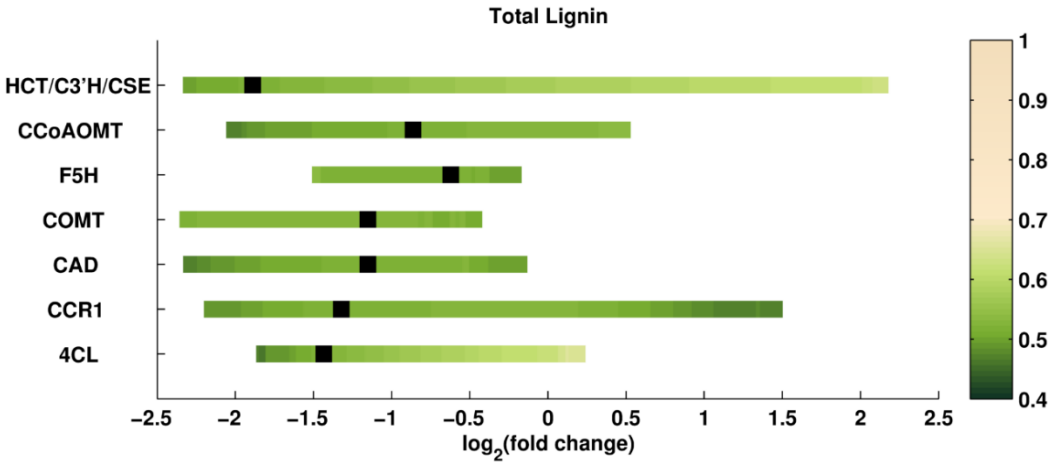

Figure 3.12 Total lignin in overexpressed PvMYB4 plus a single enzyme perturbation. The color code is the same as in Figure 3.8. The black squares represent the original amount of the corresponding enzyme in the PvMYB4 experiment. Additional overexpression of CCR is predicted to improve the total lignin results. Decreasing the level of CAD and CCoAOMT, relative to the reference PvMyb4 experiment, can reduce the total lignin further.

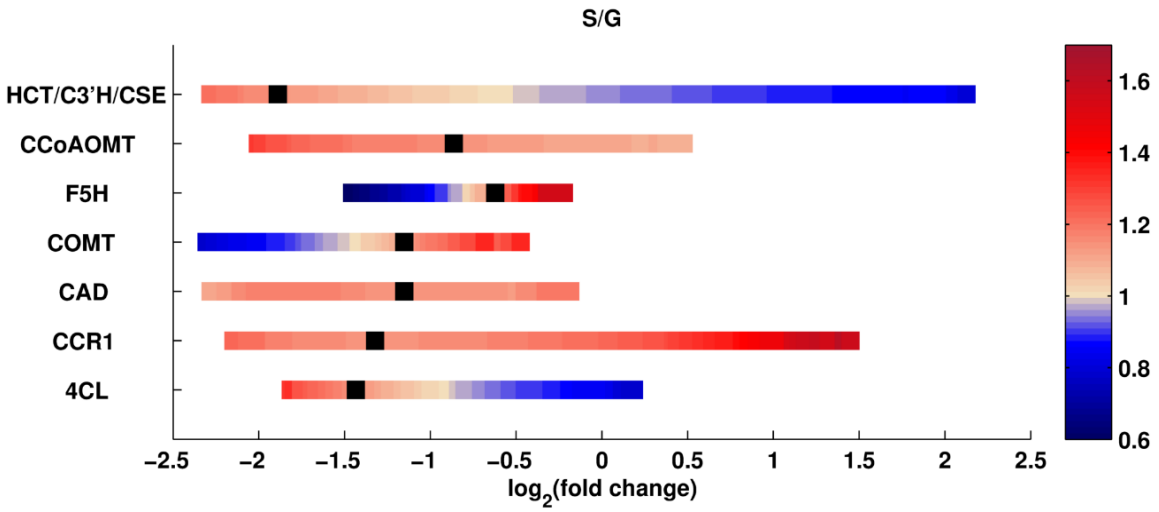

Figure 3.13 S/G ratio in overexpressed PvMYB4 plus an additional single enzyme perturbation. The color code is the same as in Figure 3.9. The black squares represent the reference amount of the corresponding enzyme in the PvMYB4 experiment. The S/G ratio can be significantly changed compared to the background PvMYB4 experiment. A change in F5H can alter the S/G ratio dramatically in a narrow perturbation interval.

3.2.10.4 System Optimization through Global Perturbations

So far, all perturbation profiles were determined by Monte Carlo sampling, where the goal was to find a desired combination of lignin content and composition. Now, we pursue a somewhat similar goal, except that it represents a different intent. Namely, a desired target combination of lignin content and S/G ratio is chosen *a priori* as the criterion for an optimization, where the goal is to determine those admissible combinatorial perturbation profiles that satisfy the criteria in an optimal manner. The main difference between this approach, and the results in Figures 3.8 - 3.11, is that the earlier approach tries to keep the number of enzymes to be manipulated as low as possible. Hence, the reduction in lignin content and composition is certainly not necessarily optimized. Furthermore, the changes were restricted by physiological limits, because dramatic changes in enzyme activities may

lead to instability of the system, which might translate into the accumulation of toxic intermediate metabolites, or the emergence of undesired phenotypes. Therefore, the level of perturbations should be accordingly limited for at least some enzymes of the pathway.

A different approach toward identifying desirable strains is the optimization of all enzymes within physiological bounds. Although the optimized combinations might not be experimentally implementable at present, they do indicate what changes are theoretically achievable, and in which specific directions novel alterations should be pursued. Thus, this part of the project aims to compute ensembles of optimized enzyme activity patterns within physiological constraints.

As it is not clear which combination of lignin content and composition (S/G ratio) is considered optimal for a particular purpose, all enzymes in the model were simultaneously perturbed randomly up to ±5-fold, using Monte Carlo simulations. Admissible system responses were defined as stable scenarios that reached a steady state and led to an accumulation of metabolites, of at most, 6 times their normal levels, or fell at most, to 5% of the normal level. The admissible solutions were recorded and categorized based on lignin content and S/G ratio.

The results are shown in Figure 3.14. The three rows of subpanels represent the degree of reduction in lignin, categorized in three intervals, and the four columns of subpanels represent intervals for changes in the S/G ratio. Thus, the top row indicates the strongest reduction in lignin, and the left-most subpanel exhibits the lowest S/G ratios. Due to the randomized nature of the Monte Carlo method, we obtain many perturbation profiles that satisfy the constraints of each subpanel. All such profiles are grouped into 3 to 10

clusters, denoted by c_i. The default number of clusters is three, and each cluster contains, at most, 100 profiles. Some subpanels include more clusters, which is an indication of the abundance of admissible profiles for that range of constraints. The median of each set of profiles is computed for each cluster, and the clusters are then sorted based on the total fold change in all enzymes collectively, shown as the bottom row in each subpanel. The darker a box in the bottom row is, the more distant the strain is from wild type switchgrass. This total fold change is a measure of how distant or close a mutant strain is to the wild type. This strategy accounts for the observation that profiles closer to wild type are probably to be favored metabolically, and is in line with the concept of the minimization of metabolic adjustment [58].

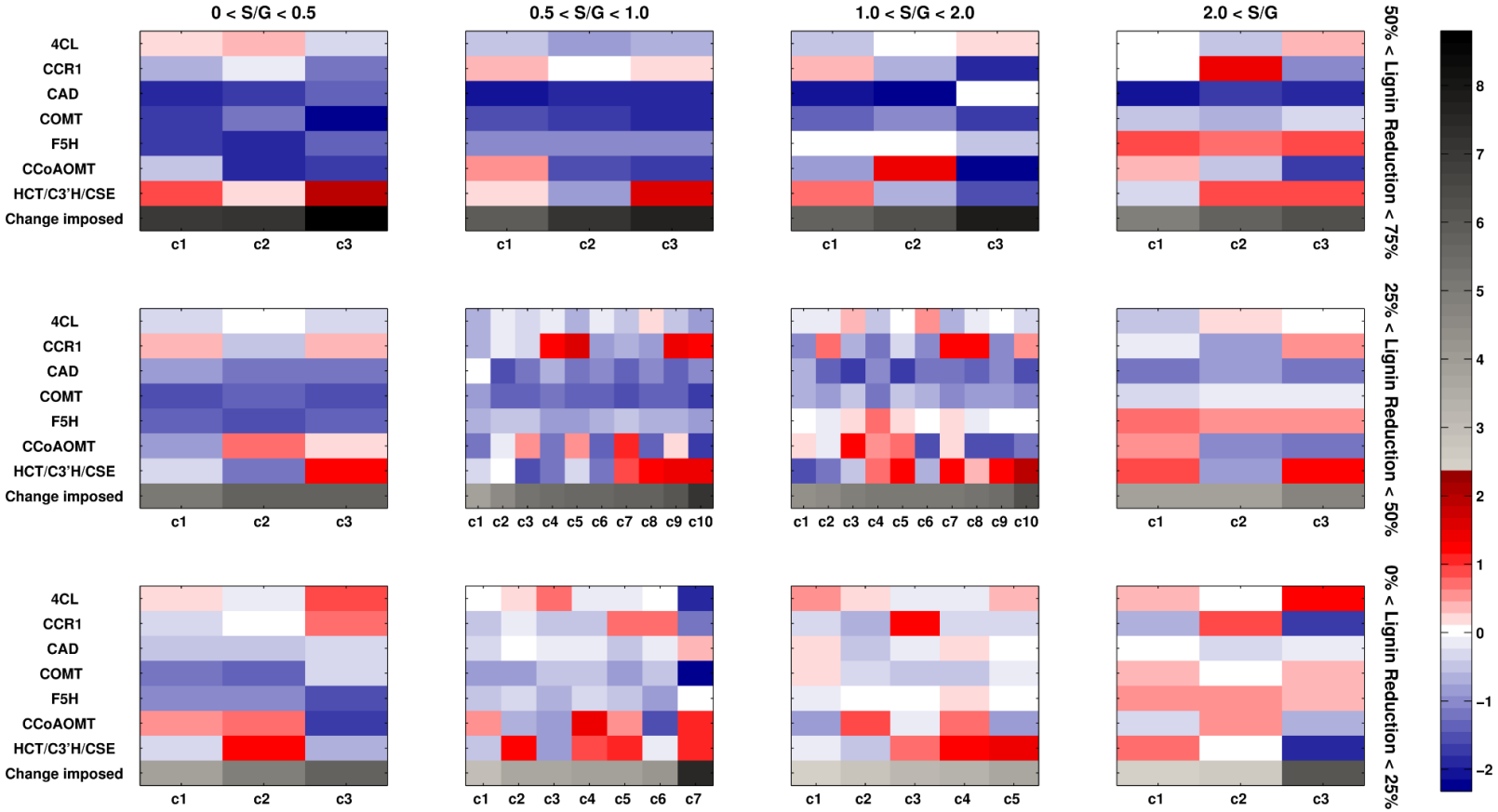

Figure 3.14 Global perturbation scenarios. All seven enzymes are perturbed simultaneously. The results are broken into 12 subpanels. The subpanels in the same columns share the same S/G ratio, and the subpanels in the same row share the same total lignin. Each of the subpanels includes several columns, where each column represents a cluster of perturbation vectors that are sorted based on the distance from the wild type. Cluster c1 in each subpanel is the closest perturbation scenario to the wild type. The grey scale represents the distance from the wild type, and the red/blue spectrum shows the increase/decrease in pathway enzyme. White represents the wild type, therefore, no change in the enzyme.

Since the S/G ratio in wild type switchgrass is about 1, the central subpanels are closer to the wild type. If an experimentalist is interested in a strain with the strongest possible reduction in lignin, and the highest possible increase in the S/G ratio, the subpanel in the top right corner exhibits perturbation profiles that are predicted to achieve these criteria. Among these, cluster c_1 is the closer to wild type than c_2 and c_3 in the same subpanel. If the c_1 column is the chosen profile, the enzyme perturbation scenario is indicated by the color code. White represents the wild type, the blue spectrum represents reduced expression of the enzyme, and the red spectrum shows overexpression. The intensity of the color represents the degree of perturbation needed. The $\log_2$-fold change is indicated in the color bar.

As a specific example, suppose that a high increase in S/G ratio and a moderate decrease in total lignin is desired, which leads us to the top right subpanel. If a medium total change in enzyme profile is allowed, we choose column c_2 as the perturbation scenario. Then, CCR1 must be overexpressed 2.8-fold, F5H must be overexpressed to 1.6-fold, the flux from the group of HCT/C3'H/CSE must be increased 1.9-fold, while 4CL, COMT, and CCoAOMT must be knocked down, as indicated in the blue range of the color scale.

Although the lignin profile in some pathway enzyme knockdowns has been measured before *in vivo* [6, 8, 56, 168, 171], our results cover seven pathway enzyme knockdowns in single, double, and combinatorial perturbations. The possibly strong ±5-fold perturbations presumably cover the realistic range of behaviors of the lignin biosynthetic pathway in response to gene knockdowns. Determining the total lignin and

S/G ratio, simultaneously, provides a powerful tool for lignin researchers to choose the desired knockdown scenario based on the lignin characteristics of choice.

Figures 3.8 - 3.14 make it clear that the response of the system is nonlinear in some perturbation scenarios. For instance, in single enzyme perturbations (Figure 3.8), F5H, COMT, and CCR1 exhibit non-monotonic changes in total lignin, which means that reducing the enzyme concentration is not the only way to reduce lignin content, but that targeted overexpression may lead to the same result. In the specific case of CCR, an increase in total lignin with small degrees of enzyme overexpression, followed by a decrease in total lignin at higher levels of enzyme overexpression, is a good example of the occasional counterintuitive behavior of the pathway. The same pattern is even clearer in double knockdowns, such as the pairs of 4CL–CCR, 4CL–F5H, CCR–CAD, CCR–CCoAOMT, and COMT–CCoAOMT (Figure 3.9). Another interesting result is that choosing a specific perturbation scenario can retain the same amount of total lignin, while leaving room to adjust the desired S/G ratio, as it is the case for the pair 4CL–F5H: there is no substantial difference in total lignin in the left half of the figure, but there is a drastic change in S/G ratio based on the fold change in F5H. The same applies to other pairs including F5H, and also, for the combination of COMT–CCoAOMT.

The computational model turned out to be helpful for an investigation of the transgenic strain of overexpressed PvMYB4 (Figure 3.11). It is interesting to note that, similar to single enzyme perturbations, combinations of the profile of pathway enzymes in PvMYB4 line with overexpressed CCR can further improve the reduction in total lignin. Again, the S/G ratio is easy to manipulate, while keeping the total lignin almost unchanged; see for instance, combinations of the profile with F5H or COMT.

An added benefit of the developed library is that our results could be complemented with a record of the ethanol yield in the transgenic plants containing different total lignin contents and different S/G ratios. This record could provide the desired lignin content and S/G ratio, and with this target, one could use our results in Figures 3.8 - 3.14 to choose the perturbation scenario needed to achieve the target characteristics in the transgenic plant. In other words, this combination of computational results and literature information could be of value and assistance for the targeted design of transgenic plants.

While, from a technical point of view, growing transgenic plants with more than two knocked down enzymes does not seem to be practical at present, fast and inexpensive computational modeling is not really limited. Thus, it offers the opportunity to investigate more complex perturbation scenarios that could shed light onto virtually optimized transgenics. For example, one could restrict the number of enzymes to be perturbed, or permit higher levels of perturbation to achieve a more significant change in total lignin or S/G ratio (*cf.* [157]).

Of course, caution is advised, and it will be necessary to validate the predictions with correspondingly manipulated strains. As an example of possibly wrong predictions, a drastic change in an enzyme concentration is not a problem, theoretically, but physiological restrictions might not allow it. It could also happen that a significant change in one or two enzymes might lead to intolerable changes in fluxes, or an accumulation of toxic intermediates within, or outside, the lignin pathway. By contrast, a combination of small changes at multiple locations of the pathway is experimentally more challenging, but might avoid such issues. Thus, there is range of options for reaching the same result. Among these admissible perturbation scenarios, which give the same combination of lignin content and

S/G ratio, one should presumably focus on the optimized scenario that reaches the target but deviates the least from the wild type. This overall deviation is reminiscent of the philosophy of the method of minimization of metabolic adjustment (MOMA) [58, 93], which we discussed before. It may be assessed with a metric, like the Euclidian distance between the enzyme profiles in the virtual transgenic and the wild type. As one might expect, our results show that the more the lignin content is to be reduced, the further the optimized enzyme perturbation profile deviates from the wild type. Interestingly, the S/G ratio is not particularly sensitive to the distance from the wild type, and for the same lignin content, the optimized enzyme perturbation profiles for different S/G ratios are quite close to each other. In other words, it seems that altering the S/G ratio does not introduce plants with severely altered characteristics.

3.3 Discussion and Conclusions

In this work, we developed an ensemble of models of lignin biosynthesis in stem and tiller tissue of switchgrass, *P. virgatum*. The model reflects the consequences of various enzyme knock-downs quite well and performed satisfactorily in two validation studies with experimental data that had not been used in the model design or implementation. We used as the modeling framework the generalized mass action (GMA) format within biochemical systems theory (BST) [122, 123, 157, 158, 172]. The power-law representation, which is the hallmark of this type of model, is arguably the least biased default formulation and by its mathematical nature avoids problems due to possibly invalid assumptions that may cast doubt on traditional Michaelis–Menten models *in vivo* [153]. Parameter values were, as

always, difficult to obtain in a direct manner. We used for this purpose experimental knock-down data and a sophisticated Monte Carlo sampling strategy that has been used very successfully for similar systems before [18]. As a particular sub-goal, we investigated the regulatory mechanism of the pathway and the possible co-localization or coupling of the pair of enzymes, CCR1/CAD that was previously suggested for *Medicago* [8].

To elucidate the co-localization or coupling of these enzymes in switchgrass, we studied multiple configurations that seemed a priori plausible and identified those natural designs that were consistent with the experimental data. The consistent designs were further examined under different regulation scenarios. The main result from this study is a very robust model of lignin biosynthesis in switchgrass that is consistent with all available data. The model was, at least to some degree, validated with a formerly unused dataset. If this validation can be confirmed and expanded experimentally, the model proposed here may be used to predict responses of the natural pathway system to alterations that are difficult to assess with experimental means. For instance, a further validated model will allow the prediction of responses to combinatorial knockdowns that could be the basis for future designs of more sophisticated transgenic lines than are currently available.

The computational analysis suggests the co-localization or functional coupling of the two enzymes CCR1 and CAD. Metabolic channeling and compartmentalization in plants have been identified in many biochemical pathways [173]. Of importance here, it has been suggested that enzymes catalyzing early reactions in the monolignol pathway may be co-localized in their binding to the ER. For instance, a multi-protein complex has been identified between PAL and C4H, and it seems that most of the substrates use these channels, but that some substrate undergoes the metabolic conversion in two steps [174-

176]. C4H can also form a complex with C3′H [177], and it has been suggested that different forms of 4CL form a complex in poplar [178]. Independent computational work on alfalfa came to a similar conclusion for channeling of enzymes associated with coniferaldehyde, which were proposed to form a metabolic channel [18]. Our results on switchgrass, presented in this article, are in line with the latter result and suggest moreover that channeling around coniferaldehyde is necessary to capture the available data.

The comparative study of different configurations revealed that consistency with the available experimental data was most difficult to achieve for transgenic 4CL down-regulated lines, in which, surprisingly, the H lignin concentration is increased. This observation is at first counterintuitive because 4CL is located directly upstream of the H lignin precursors, which would lead to the *a priori* expectation of a decrease in H lignin. The combination of two postulated types of regulatory mechanisms was able to explain this observation. The first is product inhibition, which is observed quite frequently in biochemical systems. While improving the data compatibility, this mechanism turned out to be insufficient, thus requiring additional signaling. Arguably the simplest explanation is a regulatory structure that works in either of the mechanisms below:

- An intermediate in the pathway is increased in response to the 4CL knockdown and activates the precursors of H lignin synthesis. The most likely candidates for this scenario appear to be *p*-coumaric acid, caffeic acid, and ferulic acid (Figure 3.15a).

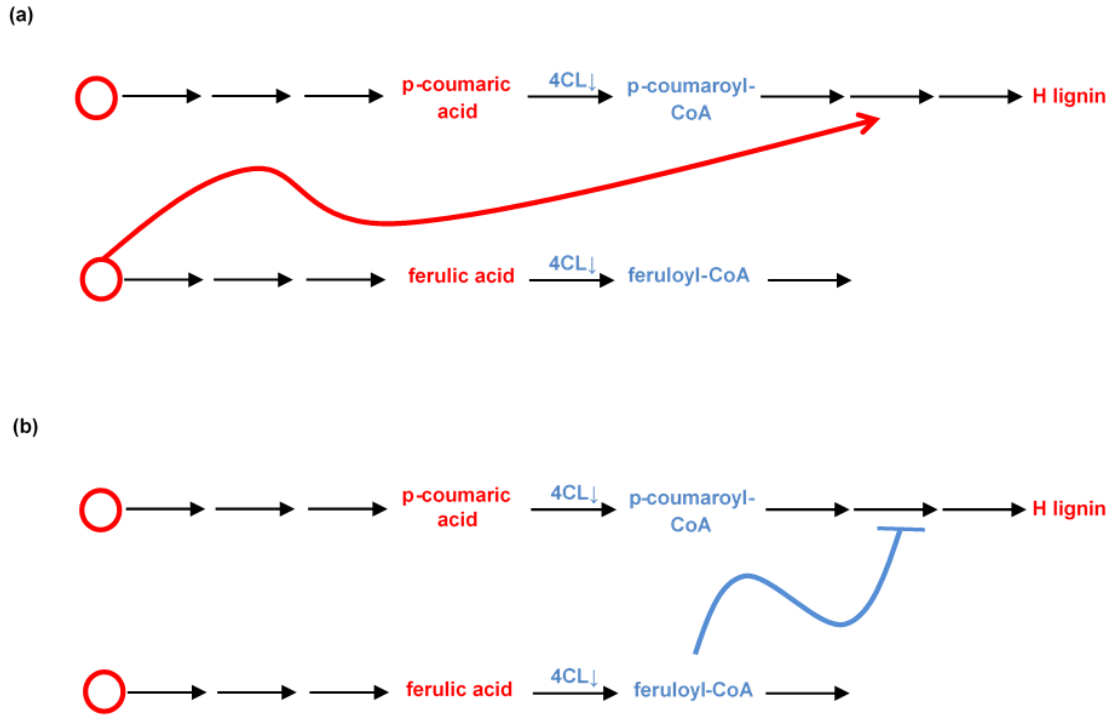

Figure 3.15 Two plausible explanations for an increase in the H lignin concentration in 4CL transgenic lines. (a) represents a putative increase in an activator located *upstream* of the enzyme 4CL, whereas **(b)** shows a putative decrease in an inhibitor located *downstream* of 4CL.

- There exists an inhibitor for the H lignin branch. This metabolite would have to be located such that its concentration is decreased due to the 4CL knockdown, which means that the inhibitor activity is inhibited and therefore exerts a net positive effect on the system (Figure 3.15b). Feruloyl-CoA could be a good candidate for this scenario.

The current literature does not support the first hypothesis. By contrast, multiple candidates are available for the second scenario. A reasonable scenario arises from the fact that the

lignin pathway in switchgrass includes parallel fluxes that share the same enzymes. Indeed, 4CL, CAD, COMT, F5H and CCR1 all catalyze multiple reactions, and it is likely that the substrates exert competitive inhibition for the shared enzyme, as it was also suggested in [59]. Supporting this scenario, a targeted numerical analysis demonstrated that competition over CCR1 perfectly matches the results of the 4CL knockdown line in the model with product inhibition. One could surmise that the latter mechanism would suffice to represent the increase in H lignin concentration. To test this hypothesis, we simulated the model with enzyme competition but without product inhibition. The results showed that competitive inhibition by itself could not satisfactorily resolve the issue. By contrast, the combined model containing product inhibition and competitive inhibition matches the experimental results very well. One should also recall that the product inhibition and substrate competition mechanisms only work properly if the proposed metabolic channel is present (Figure 3.3, Configuration 1).

Another aspect of the experimental data that was not captured well by the original model, even when product inhibition and substrate competition over CCR1 were taken into account, is the accumulation of 4CL substrates in COMT transgenic plants. Particularly counterintuitive appears to be the accumulation of ferulic acid as a product of a reaction catalyzed by COMT. The observed concomitant decrease in the steady-state concentration of coniferaldehyde supports the possible explanation that the observation is due to regulation that begins to inhibit the conversion of ferulic acid into coniferaldehyde, when 4CL substrates are in excess. The simultaneous accumulation of *p*-coumaric acid and caffeic acid provides additional evidence that reactions catalyzed by 4CL are inhibited in COMT knockdown plants. Accounting for this feature to our model, all experimental data

are represented well. The mechanism of the regulation remains a subject of further experimental investigations. Figure 3.16 shows the pathway including all inferred regulatory signals.

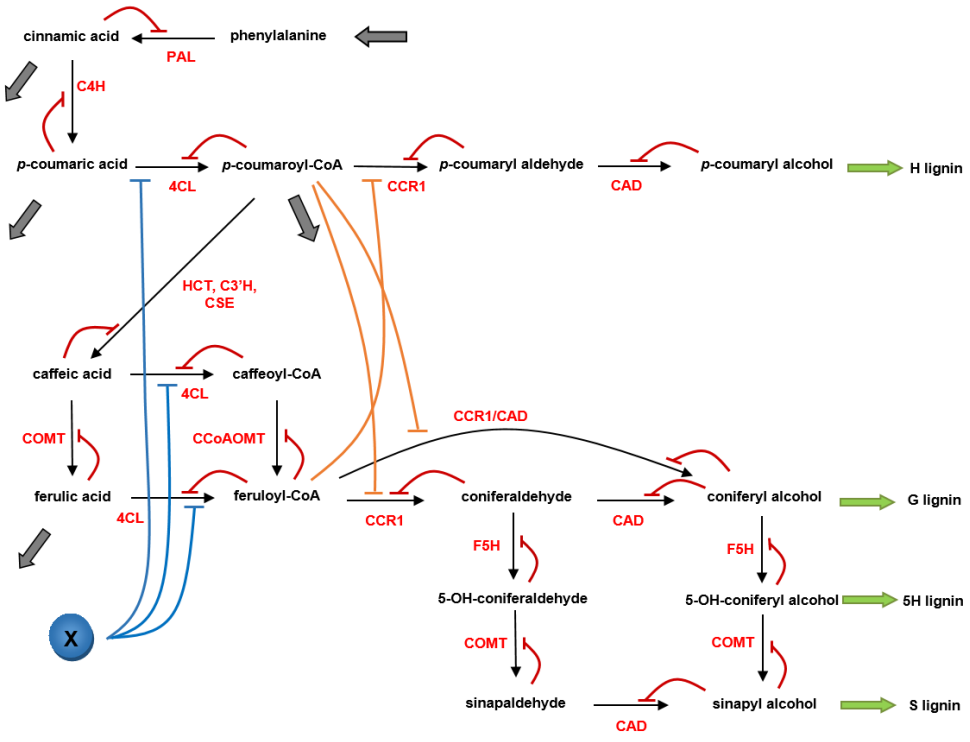

Figure 3.16 Full scheme of the lignin biosynthetic pathway in switchgrass suggested by the computational results of this study. All regulatory signals, *i.e.*, universal product inhibition, substrate competition over CCR1, and 4CL inhibition are shown. The 4CL inhibiting agent is unknown and therefore denoted with X. 5-OH-ferulic acid might be a candidate for this role.

The model proposed in this article captures all available data and performed well in independent PvMYB4 validation experiments. This good match with data is reason for cautious optimism, which however is to be supported with further experimental confirmation. Indeed, work is in progress to generate and analyze additional transgenic

switchgrass lines and to incorporate further lignin compositional and enzyme activity and kinetic data into the model. If the model fares well in these additional validation studies, the results from the present study suggest that one might use the model for predictions, for instance, with respect to double knock-downs, and for optimization studies that could potentially affect the lignin-based recalcitrance in switchgrass in a favorable manner.

3.4 Methods

3.4.1 Model Construction

Much of the analysis in this article consists of comparisons and simulations with different models. Each of these models consists of a system of differential equations that represent the rate of change in metabolite concentrations, which are represented as dependent variables. The right-hand side of each equation contains a set of fluxes which enter (influxes) or leave (effluxes) the metabolite pool. Enzymes are included in the model as independent variables; that is, they do not change in activity during any given computational experiment. The generic formulation of each equation is

$$\frac{dX_i}{dt} = \sum_{j=1}^{k} s_{i,j} V_j \tag{3.2}$$

where each X_i is a metabolite, V_j are fluxes associated with X_i, and the quantities $s_{i,j}$ are stoichiometric coefficients, which here are simply 0, 1 or −1 and determine whether flux V_j affects X_i as influx or efflux or not at all. Each V_j is a function of some or potentially all of the X_i. At the steady state, the left-hand side is equal to zero, and fluxes

can be assessed with methods of linear algebra [86]. Because the system in our case is underdetermined, infinitely many solutions satisfy the steady-state condition. Following the tenets of Flux Balance Analysis (FBA), an objective function is chosen and the problem is solved as a linear programming problem [86]. In the present study, maximizing the total amount of lignin is set as the objective of the system. The optimization problem is solved using MATLAB (version R2014a, The MathWorks, Natick, MA, USA) function *linpro*. The output is the set of fluxes at the steady state that maximizes the defined objective.

The fluxes themselves are formulated as general mass action (GMA) models of the type

$$V_j = \alpha_j \prod_{r=1}^{n} X_r^{g_{r,j}} \prod_{r=n+1}^{n+m} X_r^{h_{r,j}} \tag{3.3}$$

within the modeling framework of BST [121-123, 158]. Here, α_j is the rate constant, each X_r, for $1 < r < n$, is a metabolite or, for $n + 1 < r < n + m$, an enzyme involved in the reaction. Thus, n is the number of metabolites and m is the number of enzymes in the pathway. The exponents $g_{r,j}$ are kinetic orders that quantify the effect of X_r on V_j. Similarly, $h_{r,j}$ describes the effect of the enzyme on the reaction. It is customary to set each $h_{r,j}$ to 0 or 1, thus merely reflecting absence or presence of an enzyme in a specific flux. This setting of $h_{r,j} = 1$ is consistent with the underpinnings of Michaelis–Menten, mass-action, and other traditional models, where a reaction is assumed to be a linear function of enzyme activity. All other kinetic orders $g_{r,j}$ are sampled from the range between 0 and 1 if X_r is a substrate or activator of the flux, or from the range between −1 and 0 if X_r is an inhibitor.

Due to the nature of the present experimental data for switchgrass, the real concentrations of metabolites and enzyme activities *in vivo* are unknown. As a remedy, we normalize these quantities with respect to the steady state and set all base values to 100. Thus, we set

$$Z_i = \frac{100 X_i}{X_{SS,i}} \tag{3.4}$$

and express Equation 3.4 as

$$\frac{dZ_i}{dt} = \frac{100}{X_{SS,i}} \frac{dX_i}{dt} = \frac{100}{X_{SS,i}} \sum_{j=1}^{k} s_{i,j} V_j. \tag{3.5}$$

Since the constant $X_{ss,i}$ refers to the steady state, simple algebra adjusts the rate constants to this steady state. Thus, we obtain

$$\frac{dZ_i}{dt} = \sum_{j=1}^{k} s_{i,j} \alpha_j \prod_{r=1}^{n} \left(\frac{100 X_r}{X_{SS,r}} \right)^{g_{r,j}} \prod_{r=n+1}^{n+m} \left(\frac{X_r}{X_{SS,r}} \right)^{h_{r,j}}. \tag{3.6}$$

The enzymes are independent variables and therefore constant for each experiment. Therefore, $X_r = X_{ss}$ for $n + 1 < r < n + m$ for wild type, whereas for a transgenic line it takes a value between 0 and 1, according to the level of knockdown. At the steady state we have

$$\begin{aligned} 0 &= \sum_{j=1}^{k} s_{i,j} \alpha_j \prod_{r=1}^{n} \left(\frac{100 X_r}{X_{SS,r}} \right)^{g_{r,j}}, \\ 0 &= \sum_{j=1}^{k} s_{i,j} \alpha_j \prod_{r=1}^{n} 100^{g_{r,j}}. \end{aligned} \tag{3.7}$$

With this setting, each steady-state flux is given as

$$V_j = \alpha_j \prod_{r=1}^{n} 100^{g_{r,j}} = \alpha_j 100^{\sum_{r=1}^{n} g_{r,j}}. \tag{3.8}$$

If the flux is known, the rate constant can be computed as

$$\alpha_j = V_j / 100^{\sum_{r=1}^{n} g_{r,j}}. \tag{3.9}$$

With these settings, the set of the differential equations for the model takes the form below

$$\begin{aligned}
\frac{dZ_1}{dt} &= I_1 - V_1 & \frac{dZ_9}{dt} &= V_{12} - V_{14} - V_{18} \\
\frac{dZ_2}{dt} &= V_1 - V_2 - V_3 & \frac{dZ_{10}}{dt} &= V_{13} + V_{14} - V_{15} - V_{26} \\
\frac{dZ_3}{dt} &= I_2 + V_3 - V_4 - V_8 & \frac{dZ_{11}}{dt} &= V_{15} - V_{16} - V_{19} \\
\frac{dZ_4}{dt} &= V_4 - V_5 - V_9 - V_{10} & \frac{dZ_{12}}{dt} &= V_{16} + V_{26} - V_{17} - V_{20} \\
\frac{dZ_5}{dt} &= V_5 - V_6 & \frac{dZ_{13}}{dt} &= V_{19} - V_{22} \\
\frac{dZ_6}{dt} &= V_6 - V_7 & \frac{dZ_{14}}{dt} &= V_{20} - V_{21} - V_{23} \\
\frac{dZ_7}{dt} &= V_9 - V_{11} - V_{12} & \frac{dZ_{15}}{dt} &= V_{22} - V_{24} \\
\frac{dZ_8}{dt} &= V_{11} - V_{13} & \frac{dZ_{16}}{dt} &= V_{23} + V_{24} - V_{25}
\end{aligned} \tag{3.10}$$

where the quantities I_i include the influxes into the pathway and the fluxes, V_i, are defined as follows:

$$
\begin{aligned}
V_1 &= \alpha_1 Z_1^{g_{1,1}} Z_2^{g_{2,1}} Z_{17} & V_{14} &= \alpha_{14} Z_9^{g_{9,14}} Z_{10}^{g_{10,14}} Z_{20} \\
V_2 &= \alpha_2 Z_2^{g_{2,2}} Z_{18} & V_{15} &= \alpha_{15} Z_{10}^{g_{10,15}} Z_{11}^{g_{11,15}} Z_4^{g_{4,15}} Z_{21} \\
V_3 &= \alpha_3 Z_2^{g_{2,3}} Z_3^{g_{3,3}} Z_{19} & V_{16} &= \alpha_{16} Z_{11}^{g_{11,16}} Z_{12}^{g_{12,16}} Z_{22} \\
V_4 &= \alpha_4 Z_3^{g_{3,4}} Z_4^{g_{4,4}} Z_{20} & V_{17} &= \alpha_{17} Z_{12}^{g_{12,17}} Z_{29} \\
V_5 &= \alpha_5 Z_4^{g_{4,5}} Z_5^{g_{5,5}} Z_{10}^{g_{10,5}} Z_{21} & V_{18} &= \alpha_{18} Z_9^{g_{9,18}} Z_{30} \\
V_6 &= \alpha_6 Z_5^{g_{5,6}} Z_6^{g_{6,6}} Z_{22} & V_{19} &= \alpha_{19} Z_{11}^{g_{11,19}} Z_{13}^{g_{13,19}} Z_{31} \\
V_7 &= \alpha_7 Z_6^{g_{6,7}} Z_{23} & V_{20} &= \alpha_{20} Z_{12}^{g_{12,20}} Z_{14}^{g_{14,20}} Z_{31} \\
V_8 &= \alpha_8 Z_3^{g_{3,8}} Z_{24} & V_{21} &= \alpha_{21} Z_{14}^{g_{14,21}} Z_{32} \\
V_9 &= \alpha_9 Z_4^{g_{4,9}} Z_7^{g_{7,9}} Z_{25} & V_{22} &= \alpha_{22} Z_{13}^{g_{13,22}} Z_{15}^{g_{15,22}} Z_{27} \\
V_{10} &= \alpha_{10} Z_4^{g_{4,10}} Z_{26} & V_{23} &= \alpha_{23} Z_{14}^{g_{14,23}} Z_{16}^{16,23} Z_{27} \\
V_{11} &= \alpha_{11} Z_7^{g_{7,11}} Z_8^{g_{8,11}} Z_{20} & V_{24} &= \alpha_{24} Z_{15}^{g_{15,24}} Z_{16}^{g_{16,24}} Z_{22} \\
V_{12} &= \alpha_{12} Z_7^{g_{7,12}} Z_9^{g_{9,12}} Z_{27} & V_{25} &= \alpha_{25} Z_{16}^{g_{16,25}} Z_{33} \\
V_{13} &= \alpha_{13} Z_8^{g_{8,13}} Z_{10}^{g_{10,13}} Z_{28} & V_{26} &= \alpha_{26} Z_{10}^{g_{10,26}} Z_{12}^{g_{12,26}} Z_4^{g_{4,26}} Z_{34}
\end{aligned}
\tag{3.11}
$$

The metabolites of the pathway are

$$
\begin{aligned}
Z_1 &: \text{phenylalanine} & Z_9 &: \text{ferulic acid} \\
Z_2 &: \text{cinnamic acid} & Z_{10} &: \text{feruloyl-CoA} \\
Z_3 &: \text{p-coumaric acid} & Z_{11} &: \text{coniferaldehyde} \\
Z_4 &: \text{p-coumaroyl CoA} & Z_{12} &: \text{coniferyl alcohol} \\
Z_5 &: \text{p-coumaryl aldehyde} & Z_{13} &: \text{5-OH-coniferaldehyde} \\
Z_6 &: \text{p-coumaryl alcohol} & Z_{14} &: \text{5-OH-coniferyl alcohol} \\
Z_7 &: \text{caffeic acid} & Z_{15} &: \text{sinapaldehyde} \\
Z_8 &: \text{caffeoyl CoA} & Z_{16} &: \text{sinapyl alcohol}
\end{aligned}
\tag{3.12}
$$

while the enzymes of the pathway are

Z_{17} : PAL, L-phenylalanine ammonia-lyase
Z_{19} : C4H, cinnamate 4-hydroxylase
Z_{20} : 4CL, 4-coumarate:CoA ligase
Z_{21} : CCR1, cinnamoyl CoA reductase
Z_{22} : CAD, cinnamyl alcohol dehydrogenase
Z_{25} : HCT, hydroxycinnamoyl-CoA:shikimate-
hydroxycinnamoyl transferase/
C3'H, p-coumaroyl shikimate 3'-hydroxylase/
CSE, caffeoyl shikimate esterase
Z_{27} : COMT, caffeic acid O-methyltransferase
Z_{28} : CCoAOMT, caffeoyl CoA O-methyltransferase
Z_{31} : F5H, ferulate 5-hydroxylase

(3.13)

Note that the model does not account for the dynamics of tyrosine, which we consider constant here. The model scheme is shown in Figure 3.17.

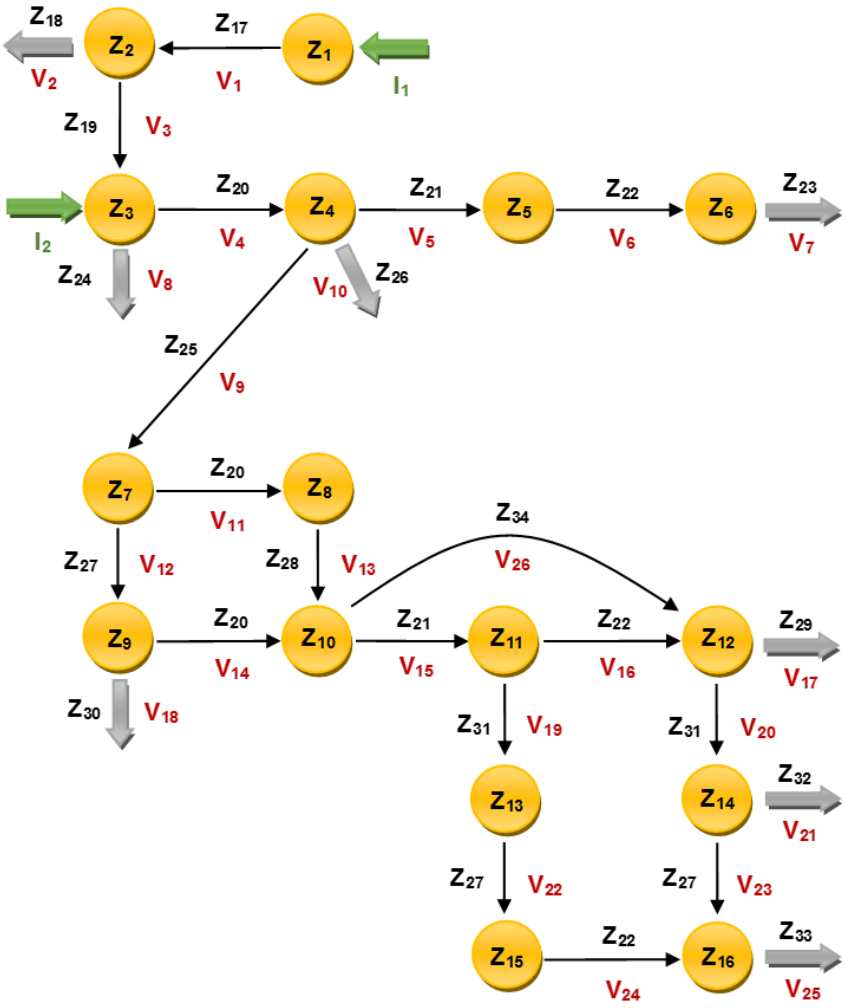

Figure 3.17 Lignin pathway in the notation of the model. Redundancy of enzymes, i.e., 4CL, CCR1, CAD, COMT and F5H in parallel fluxes reduces the dimension of state space. The enzymes HCT, C3′H and CSE in flux V_9 are merged into one independent variable, Z_{25}. Note that the presence of the G-channel, V_{26}, is an inference from the computational simulations results.

3.4.2 Parameter space and sampling

Similar to earlier work [18, 57, 58], flux rates are computed with FBA. Next, the parameters to be estimated are the kinetic orders and rate constants are in turn estimated from the FBA

results and randomly sampled kinetic orders through the steps mentioned above. The kinetic order of a metabolite is positive if the metabolite is a substrate or activator of the flux and negative if it acts as an inhibitor. The kinetic order of each enzyme has a default value of 1, which is in line with traditional enzyme kinetics, because it is customary to assume that a flux has a linear relationship with the enzyme. This assumption is explicitly or implicitly made in essentially all traditional models of enzyme kinetics as, for instance, in the Michaelis–Menten formalism, where V_{max} equals k_{cat} times the enzyme concentration.

The down-regulation of an enzyme is modeled through the enzyme concentration, not the kinetic order. Since the concentrations of metabolites and enzymes are normalized, the concentration of an enzyme in the wild type has the default value of 1. In transgenics, the concentration of the corresponding enzyme is set to a value less than one if it is down-regulated. For example, to represent the 4CL knockdown, the concentration of the enzyme is set to 0.6 as the enzyme is down-regulated by 40%.

To account for product inhibition, the inhibiting product is represented in each reaction by a factor consisting of its concentration, raised to a negative power. The result is as follows:

$$V = \alpha S^{g_S} P^{g_I} \quad , \quad -1 < \frac{g_I}{g_S} < 0 \tag{3.14}$$

Here, S is the substrate, P is the product, g_I is the kinetic order of the inhibiting product and g_S is the kinetic order of the substrate. The ratio of kinetic orders could be derived

directly [158] from the corresponding expression for a product-inhibited Michaelis–Menten reaction, which takes the form

$$V = \frac{V_{\max}\dfrac{S}{K_m}}{1+\dfrac{S}{K_m}+\dfrac{P}{K_I}} \tag{3.15}$$

The power-law form of Equation 3.15 can directly be computed from the tenets of Biological Systems Theory (BST), which defines the kinetic orders as

$$\begin{aligned} g_S &= \frac{\partial V}{\partial S}\cdot\frac{S}{V}\bigg|_{OP} = \left.\frac{1+\dfrac{P}{K_I}}{1+\dfrac{S}{K_m}+\dfrac{P}{K_I}}\right|_{OP} \\ g_I &= \frac{\partial V}{\partial P}\cdot\frac{P}{V}\bigg|_{OP} = \left.\frac{-\dfrac{P}{K_I}}{1+\dfrac{S}{K_m}+\dfrac{P}{K_I}}\right|_{OP} \end{aligned} \tag{3.16}$$

Rearrangement of these equations gives the ratio of kinetic orders as follows:

$$-1 < \frac{g_I}{g_S} = \left.\frac{-P}{P+K_I}\right|_{OP} < 0 \tag{3.17}$$

The bounded ratio of kinetic orders provides a valuable constraint for the Monte Carlo simulations, because a fixed ratio does not affect the dimension of the parameter space.

For the initial set of simulations, the sampling space is chosen as a unit hypercube in $\mathbb{R}^n$ where n is number of kinetic orders to be estimated. A set of 100,000 parameter sets is generated for each scenario simulation. 10,000 sets are randomly generated from the sampling space using Latin Hypercube Sampling to assure a homogeneous coverage of the

space, while 90,000 sets are generated by the MATLAB (version R2014a, The MathWorks, Natick, MA, USA) function *rand*. Each parameter set is simulated to examine whether the model with this set can match the experimental results for the wild type and transgenics. The model is deemed a match for the experimental results if:

- The model returns proper lignin contents and S/G ratios for the wild type and different transgenics, with down-regulation of 4CL (40 %), CCR1 (50 %), COMT (30 %), and CAD (30 %).
- The model returns the proper decrease in lignin content in the case of knockdowns in 4CL, CCR1, COMT, and CAD.
- The model demonstrates an increase in H lignin in 4CL transgenics.
- The model matches the altered metabolite concentrations in the COMT transgenic.

If a parameter profile satisfies the above conditions, it is recorded along with the corresponding topological configuration.

While our model approach emphasizes ensembles of feasible models, the parameter values in Tables 3.2, 3.3, and 3.4 represent one implementation, which we used for further numerical exploration. This specific parameter set corresponds to the minimum error in the comparison of the model results in PvMYB4 and the experimental data.

Table 3.2 A sample of rate constants from the ensemble of rate constants

α_1	0.5233	α_8	0.0058	α_{15}	0.0771	α_{22}	0.0392
α_2	0.1053	α_9	0.2265	α_{16}	0.0881	α_{23}	0.1573
α_3	0.15	α_{10}	0.0024	α_{17}	0.0168	α_{24}	0.0712
α_4	0.2711	α_{11}	0.1054	α_{18}	0.002	α_{25}	0.1154
α_5	0.1832	α_{12}	0.1095	α_{19}	0.2212	α_{26}	0.0814
α_6	0.003	α_{13}	0.1452	α_{20}	0.0402		
α_7	0.0042	α_{14}	0.2681	α_{21}	0.0002		

Table 3.3 A sample of kinetic orders from the ensemble of kinetic orders

$g_{1,1}$	0.2813	$g_{6,7}$	0.4040	$g_{10,14}$	-0.1023	$g_{14,21}$	0.8535
$g_{2,1}$	-0.1406	$g_{3,8}$	0.5759	$g_{10,15}$	0.9009	$g_{13,22}$	0.7673
$g_{2,2}$	0.0846	$g_{4,9}$	0.6118	$g_{11,15}$	-0.4505	$g_{15,22}$	-0.3836
$g_{2,3}$	0.8240	$g_{7,9}$	-0.3059	$g_{4,15}$	-0.0355	$g_{14,23}$	0.1043
$g_{3,3}$	-0.4120	$g_{4,10}$	0.6398	$g_{11,16}$	0.5198	$g_{16,23}$	-0.0521
$g_{3,4}$	0.5669	$g_{7,11}$	0.6277	$g_{12,16}$	-0.2599	$g_{15,24}$	0.5080
$g_{4,4}$	-0.2835	$g_{8,11}$	-0.3138	$g_{12,17}$	0.7121	$g_{16,24}$	-0.2540
$g_{4,5}$	0.0710	$g_{7,12}$	0.6414	$g_{9,18}$	0.6982	$g_{16,25}$	0.2855
$g_{5,5}$	-0.0355	$g_{9,12}$	-0.3207	$g_{11,19}$	0.0160	$g_{10,26}$	0.7116
$g_{10,5}$	-0.4505	$g_{8,13}$	0.4885	$g_{13,19}$	-0.0080	$g_{12,26}$	-0.3558
$g_{5,6}$	0.9630	$g_{10,13}$	-0.2442	$g_{12,20}$	0.6973	$g_{4,26}$	-0.0355
$g_{6,6}$	-0.4815	$g_{9,14}$	0.2046	$g_{14,20}$	-0.3487		

Table 3.4 Initial values

$Z_{0,1}$	100	$Z_{0,5}$	100	$Z_{0,9}$	100	$Z_{0,13}$	100
$Z_{0,2}$	100	$Z_{0,6}$	100	$Z_{0,10}$	100	$Z_{0,14}$	100
$Z_{0,3}$	100	$Z_{0,7}$	100	$Z_{0,11}$	100	$Z_{0,15}$	100
$Z_{0,4}$	100	$Z_{0,8}$	100	$Z_{0,12}$	100	$Z_{0,16}$	100

3.4.3 A Library of Virtual Mutant Strains

As a consequence of the normalization of the X_r variables, the values for the enzymes are equal to 1 when a wild type strain is modeled. If a knockdown strain is modeled, the corresponding enzyme X_r will have a value less than 1, and if a strain with an upregulated enzyme is modeled, the corresponding enzyme X_r value will be greater than 1. Once a perturbation by down- or upregulation of an enzyme has been introduced, the system rearranges itself and typically achieves a new steady state. At this state, at least some of the fluxes and metabolites typically assume new values. Thus, the affected flux becomes

$$V_j^{'} = \alpha_j \prod_{r=1}^{n} X_r^{'\, g_{r,j}} \prod_{r=n+1}^{n+m} X_r^{'} \tag{3.18}$$

To generate a library of virtual mutants, enzyme concentrations, X_r ($n+1 < r < n+m$), are perturbed up to ±5-fold. If the number of enzymes to be perturbed is *k*, using an extensive Monte Carlo simulation, a hypercube in $\mathbb{R}^k$ is randomly sampled, and the generated set of arrays of random values is fed into the system. Since α_j's and $g_{r,j}$'s are already known, the differential equations can be immediately simulated for single or double alterations, or for the PvMYB4 overexpression strain combined with an additional perturbation.

3.4.3.1 Admissible Results

Each scenario corresponds to an array of perturbed enzymes. The total lignin content and S/G ratio for the scenario are recorded, if the perturbed system satisfies the following criteria:

1. The system is stable at the steady state, and reaches this state after a perturbation.
2. The steady-state values of the metabolites do not exceed a value of 6 times the wild type concentration.
3. The steady-state values of the metabolites do not fall below 5% of the wild type value.

Since an ensemble of models is used to simulate each scenario, multiple lignin profiles exist for each perturbation scenario. For visualization purposes, the median of the total lignin and the corresponding S/G ratio of the ensemble for each scenario is plotted against the perturbed enzyme(s).

3.4.3.2 Global Perturbations and Optimized Virtual Mutant Strains

All seven enzymes are perturbed simultaneously within a range of up to ±5-fold about the wild type value. Perturbation scenarios that satisfy the criteria described in the previous section are recorded. The recorded results are arranged in a matrix based on the total lignin content and S/G ratio. The total lignin range is subdivided into three intervals, namely for 25–50%, 50–75%, and 75–100% reduction in total lignin relative to the wild type, while the S/G ratio is subdivided into four intervals, namely 0–0.5, 0.5–1, 1–2, and >2. These

intervals group the results into 12 sets with different total lignin and S/G ratio characteristics.

Each square in the results matrix contains a set with numerous scenarios satisfying specific interval criteria for total lignin and S/G ratio. To facilitate the interpreting of the results, the scenarios are clustered, and the clusters are sorted based on the distance from the wild type, which reflects the overall change imposed upon the system (Figure 3.14). This distance is defined as the Euclidian distance between the vector of perturbed enzymes and the vector of the corresponding wild type value. The smaller the distance is, the closer is the virtual mutant is to the wild type (Figure 3.18).

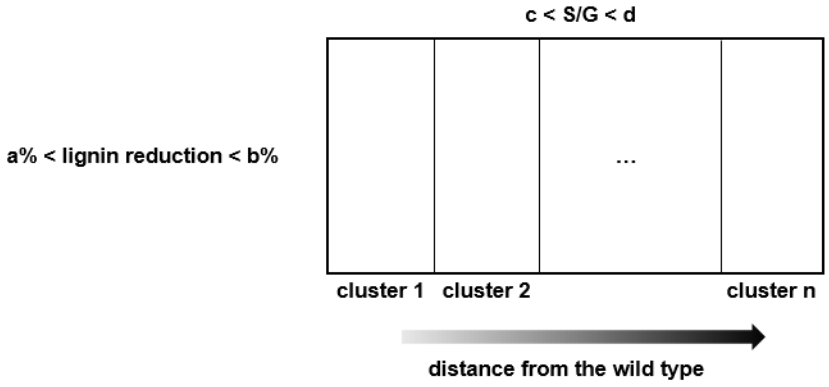

Figure 3.18 Clusters of results in matrix subpanels of Figure 3.14.

CHAPTER IV

Lignin Synthesis in *Brachypodium*[4]

4.1 Introduction

While the results of analyzing *in vivo* alfalfa and switchgrass transgenics data in a somewhat indirect manner were interesting and could be validated to some degree, the data themselves constitute a rather thin base for model development. This base becomes more solid if it is combined with other types of data. An example for such a merging of heterogeneous data types is the lignin biosynthetic pathway in *Brachypodium distachyon*. In contrast to dicots, monocot grasses use both phenylalanine and tyrosine as the initial substrate for monolignol production (Figure 4.1). One puzzling aspect of this apparent redundancy is that, despite the nearly equal contribution of both precursors to the total lignin content, phenylalanine is preferentially incorporated into G-lignin, and tyrosine into S-lignin, although both pathways converge at the same intermediate metabolite, *p*-coumaric acid [92]. This result is surprising and cannot easily be explained with the putative structure of the lignin pathway in *Brachypodium*. Beyond the existence of this

[4] The material in this chapter has been published as: 1. Faraji, M., L.L. Fonseca, L. Escamilla-Treviño, J. Barros-Rios, N. Engle, Z.K. Yang, T.J. Tschaplinski, R.A. Dixon, and E.O. Voit, *Mathematical models of lignin biosynthesis.* Ibid.2018. **11**(1): p. 34.

intermediate, where the two pathways converge, the G- and S-lignin pathways appear to be the same until they split at the coniferaldehyde node.

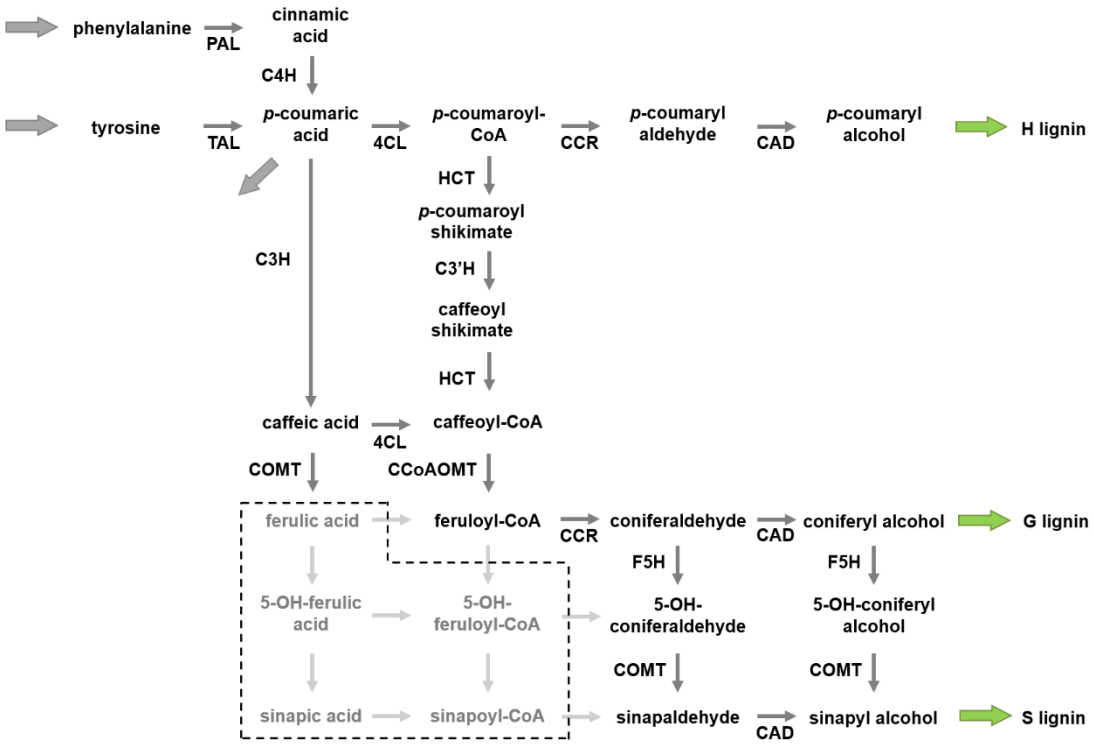

Figure 4.1 Putative lignin biosynthesis pathway in *Brachypodium distachyon*. *Brachypodium* can use both phenylalanine and tyrosine as substrates for lignin biosynthesis. At this point, the direct conversion of *p*-coumaric acid into caffeic acid and the existence of C3H in this organism are speculative. Putative reactions shown in the shaded box have not been fully explored in the current literature.

4.2 Results

A computational model directly corresponding to the alleged pathway structure (Figure 4.1) confirms the logic-based analysis: the pathway, as currently alleged, cannot reproduce

key observations, such as the differential channeling of phenylalanine and tyrosine toward G- and S-lignin. Specifically, model simulations demonstrate that the pathway scheme in Figure 4.1 is unable simultaneously to satisfy the following observed requirements:

- Match the amount of ^{13}C-labeled H-lignin in experiments with [U-$^{13}C_9$]phenylalanine;
- Match the observed ^{13}C incorporation into ER-bound ferulic acid in the same experiment;
- Capture the differential ^{13}C incorporation levels from [U-$^{13}C_9$]phenylalanine and [U-$^{13}C_9$]tyrosine in lignin units.

One great advantage of a modeling approach is the relative ease with which it is possible to test different hypotheses and variations of the pathway structure in order to obtain possible explanations. As a specific example, it was reported that the three enzymes C4H, C3′H and F5H of the lignin biosynthesis pathway in *B. distachyon* are bound to the outer surface of the ER, while the remaining enzymes are located freely in the cytosol ([92]; unpubl. data). This finding led to the hypothesis that the spatial localization of enzymes might be a reason for the preferential incorporation of phenylalanine and tyrosine into different monolignols. This hypothesis was readily tested with a computational model that distinguishes the two locations. These two locations, or compartments, are physically not strictly separated, but allow the handing over of metabolites through diffusion.

To test the hypothesis of two distinct locations, we set up a refined model scheme by assigning the reactions catalyzed by the ER-bound enzymes, C4H, C3′H and F5H, to the ER compartment, and all others to the cytosol compartment (Figure 4.2). While there

is no strict spatial separation between ER and cytosol, we assumed preferential enzyme activity within each compartment and slower diffusion between compartments. As a note, only the net diffusion fluxes are shown in the pathway model, but both forward and reverse diffusions are considered explicitly in the computational model (see later section). Specifically, we took the following steps for our model design. In the current scheme (Figure 4.2), the only means for incorporation of ^{13}C into H-lignin is through the diffusion flux D_2, and this flux is diluted with the influx from unlabeled tyrosine. To increase ^{13}C incorporation into H-lignin, a second diffusion flux, D_9, is added between the ER compartment downstream of D_2, and this flux compensates for the dilution of tyrosine (Figure 4.3). This diffusion flux D_9 can be interpreted as partial activity of 4CL in the ER compartment.

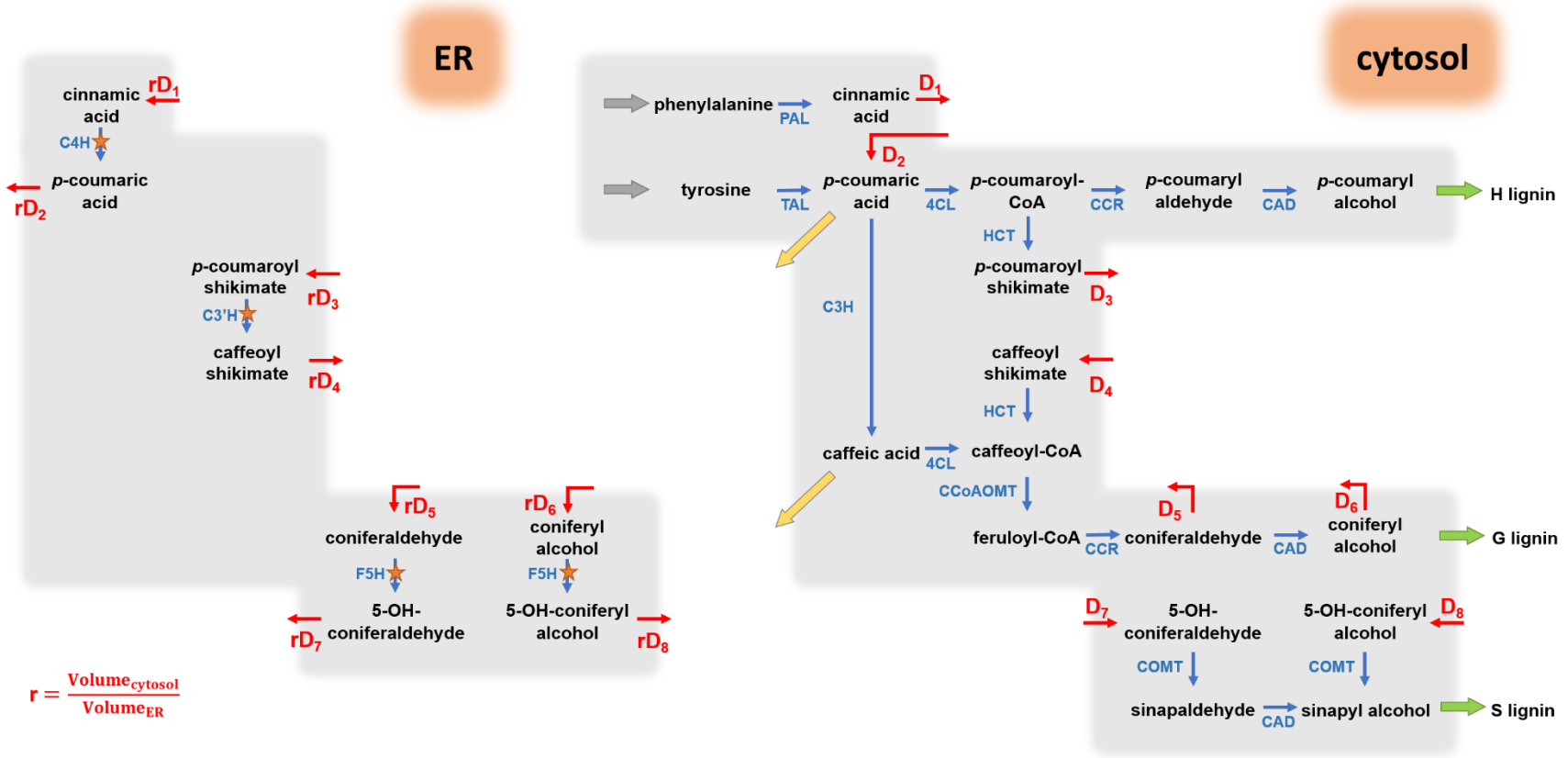

Figure 4.2 Proposed compartmentalized pathway of lignin biosynthesis in *B. distachyon*. The blue arrows represent enzymatic reactions within each compartment. The blue arrows marked by orange stars depict reactions whose catalytic enzymes are bound to the outer ER surface. The red arrows show diffusion fluxes between the compartments. The two yellow arrows are effluxes. The quantity r is a compensation constant to address the different volumes of the compartment.

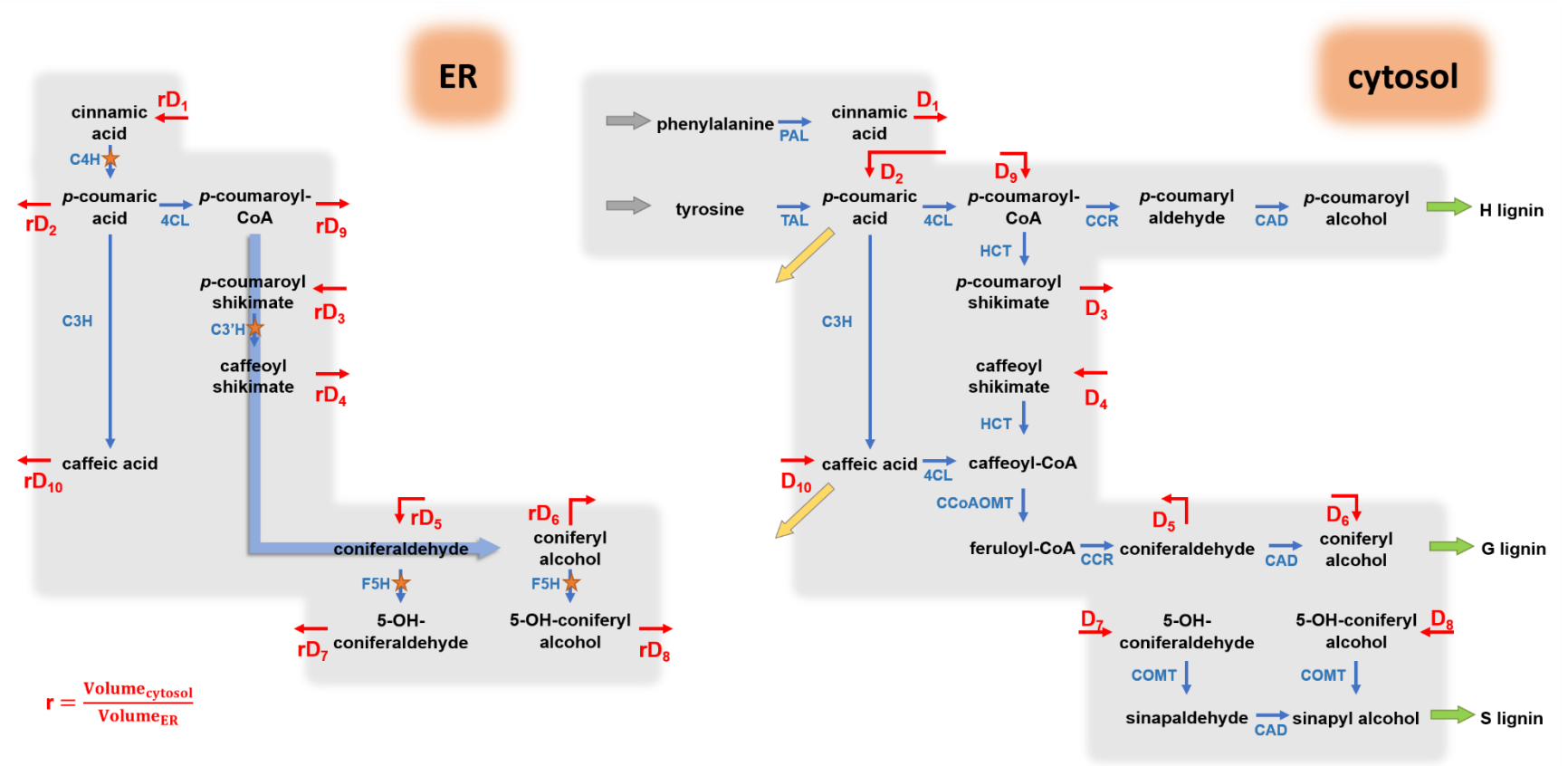

Figure 4.3 Extended compartmentalized lignin pathway model in *B. distachyon*. Conversion of *p*-coumaric acid to *p*-coumaroyl CoA by 4CL and diffusion flux D_9 are necessary to explain label incorporation into H-lignin in experiments with labeled phenylalanine. Conversion of *p*-coumaric acid to caffeic acid by C3H and the diffusion flux D_{10} are necessary to explain label incorporation into ferulic acid in the same labeling experiments with phenylalanine. The metabolic channel in the ER compartment keeps some of the ^{13}C-label from being diluted by the cytosol diffusion fluxes and permits preferential incorporation of phenylalanine and tyrosine in S- and G-lignin.

Beyond the inconsistent amount of H-lignin, low incorporation of ^{13}C in ferulic acid in the [U-$^{13}C_9$]phenylalanine labeling experiment is an indication for dilution by unlabeled tyrosine through caffeic acid. Therefore, a downstream influx, D_{10}, from the ER compartment is postulated to compensate for tyrosine dilution and to increase ^{13}C incorporation in wall-bound ferulic acid. Again, this flux corresponds to partial activity of C3H in the ER compartment (Figure 4.3).

Closer inspection of the pathway reveals that the key site for preferential incorporation of [U-$^{13}C_9$]phenylalanine and [U-$^{13}C_9$]tyrosine into different lignin units is the branch point where the pathways toward G- and S-lignin diverge; this divergence happens at the coniferaldehyde node. The original scheme in Figure 4.2 dictates the same level of ^{13}C-labeling into both G and S units, due to dilution in both compartments at the coniferaldehyde node into the free cytosol. To explain the actually observed higher incorporation of ^{13}C into G-lignin in the phenylalanine labeling experiment, an undiluted upstream flux from the ER is necessary to compensate for the dilution from the cytosol influx (D_5 and D_6) into the immediate G-lignin precursors coniferaldehyde and/or coniferyl alcohol. We first modeled this hypothesis by simply adding a suspected direct flux from *p*-coumaroyl-CoA into coniferyl alcohol (Figure 4.3, thick blue arrow).

Simulations with this amended model showed that the scheme in Figure 4.3 is able to capture the levels ^{13}C incorporation in H-lignin and ferulic acid from [U-$^{13}C_9$]phenylalanine experiments. Also, by acting as a metabolic channel, the direct flux from *p*-coumaroyl-CoA into coniferyl alcohol shields the flow within the ER compartment from strong dilution by diffusion from the cytosol compartment, and thereby enables the

preferential incorporation of phenylalanine and tyrosine–born carbons into different monolignols units.

While the long metabolic channel in Figure 4.3 is able to simulate the preferential incorporation of precursors into lignin units, it is intriguing to determine whether fewer enzymes in such channel could still reproduce the data. Therefore, we examined the scheme in Figure 4.3 toward the shortest channel possible (Figure 4.4). This analysis suggested that the critical point to shield the ER compartment from strong dilution by cytosolic diffusion fluxes is coniferaldehyde. Without this compound protected, G- and S-lignins cannot attain different $^{13}C_9$-labeling levels. If this conjecture can be validated, the simplest scheme consists merely of a CCR/CAD channel.

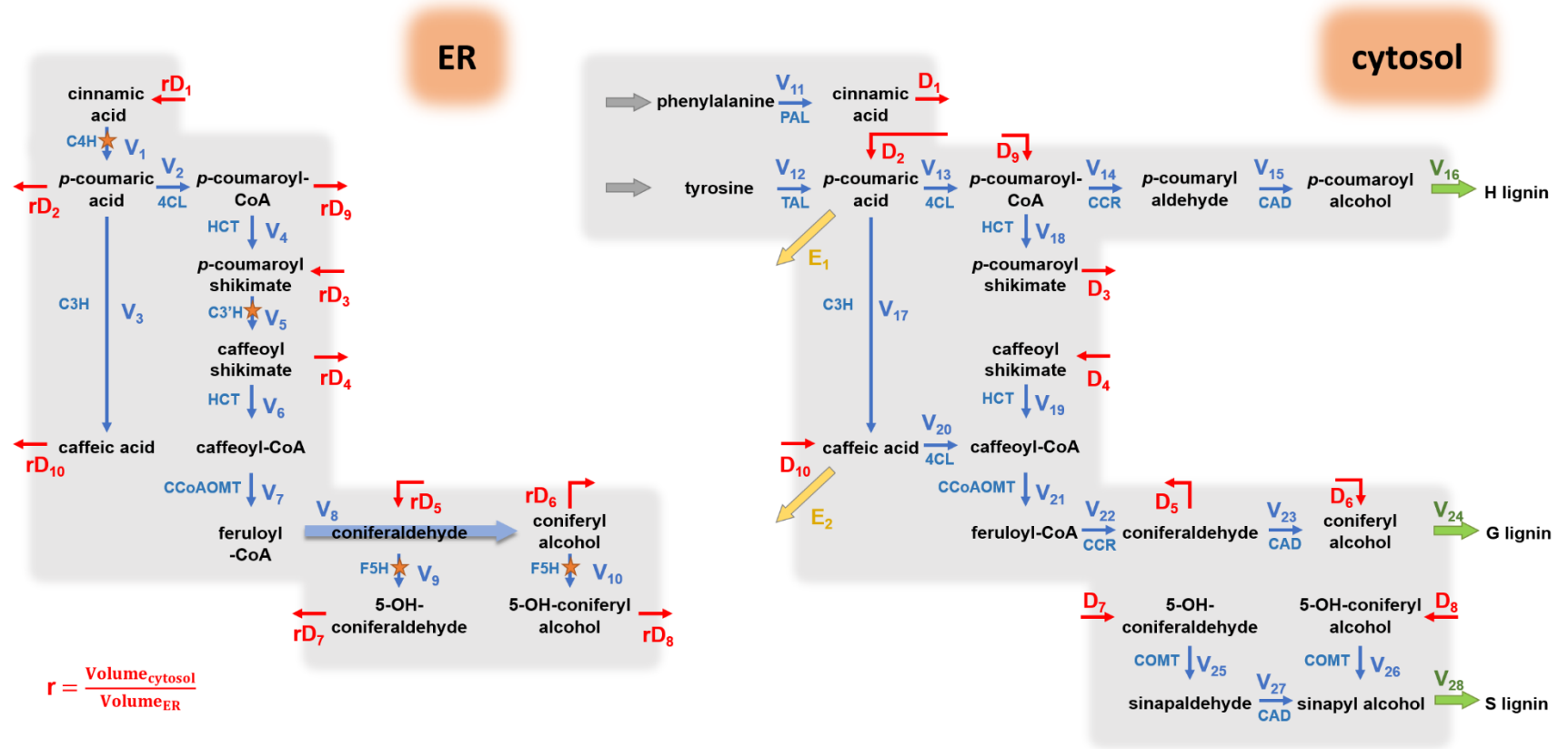

Figure 4.4 Revisited compartmental model of lignin pathway with the shortest feasible metabolic channel. The CCR/CAD channel (V_8) appears to be the shortest path that is able to preserve the flow in the ER compartment from complete dilution by cytosol diffusion fluxes.

Simulations of the scheme in Figure 4.4 resulted in steady-state flux distributions that capture the experimental ^{13}C-labeling data (Figure 4.5). Phenylalanine and tyrosine contribute nearly equally to the resulting lignin content: in the [U-^{13}C]phenylalanine experiment, 35% of phenylalanine is labeled and tyrosine is unlabeled (natural abundance), while in the [U-$^{13}C_9$]tyrosine experiment, 35% of tyrosine is labeled and phenylalanine is unlabeled (natural abundance). The labeled fluxes in Figure 4.5 compare the contributions of phenylalanine and tyrosine in each pathway flux. Figure 4.6 exhibits the total flux values, which combine the values of labeled and unlabeled fluxes. Since the magnitude of the input flux is unknown, we normalized the input to a base value of 100 units of mass per unit of time.

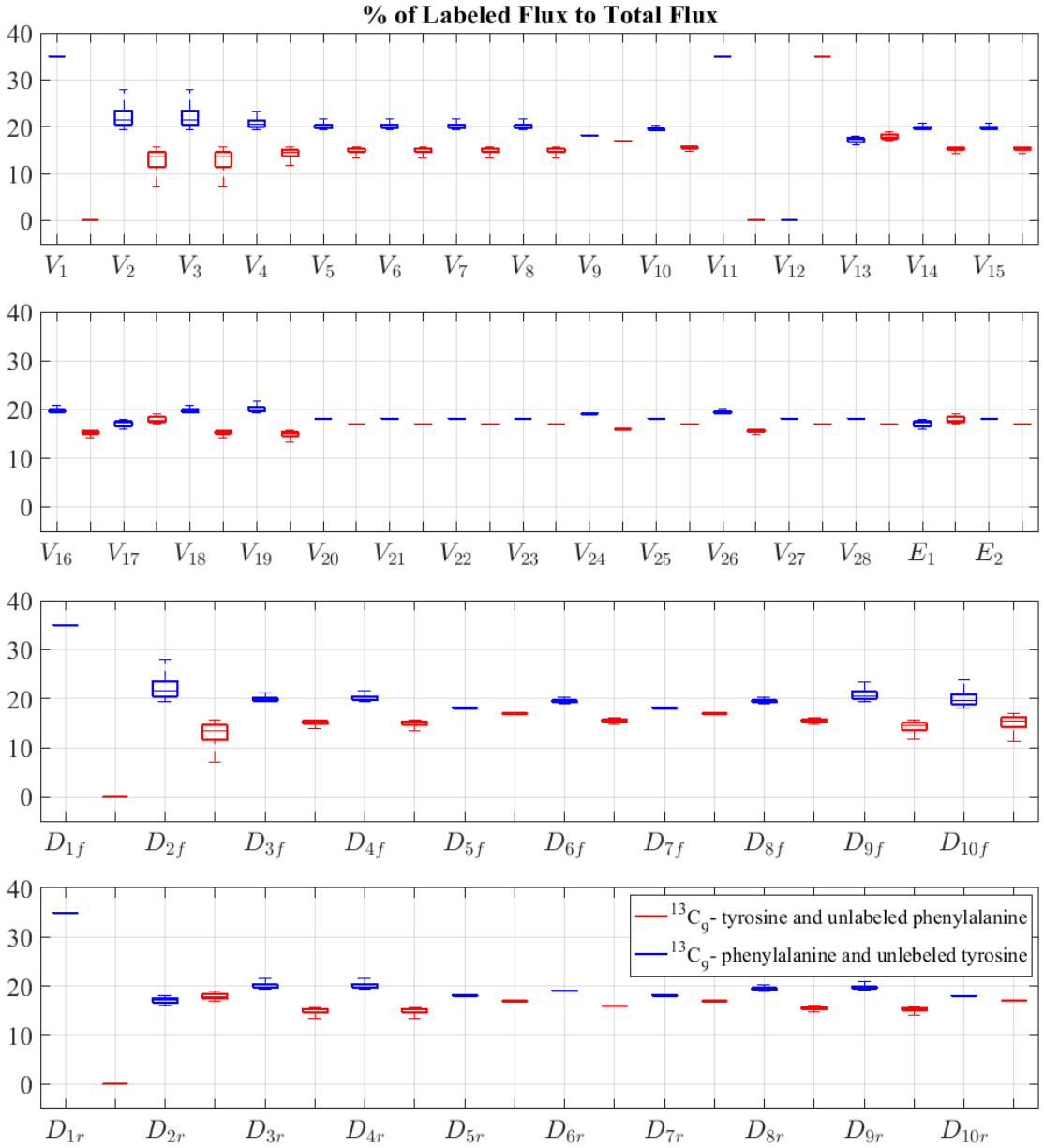

Figure 4.5 Steady-state flux distribution of labeled fluxes in *Brachypodium*. The results compare the percentage of steady-state labeled flow within the steady-state total flux in [U-$^{13}C_9$]phenylalanine and [U-$^{13}C_9$]tyrosine experiments; they correspond to the pathway scheme in Figure 4.4. Both directions of diffusion for each diffusion flux are shown: D_{if} aligns with the direction of D_i in Figure 4.4 and D_{ir} with the opposite direction (see 4.3.3 Modeling ^{13}C-labeling experiments).

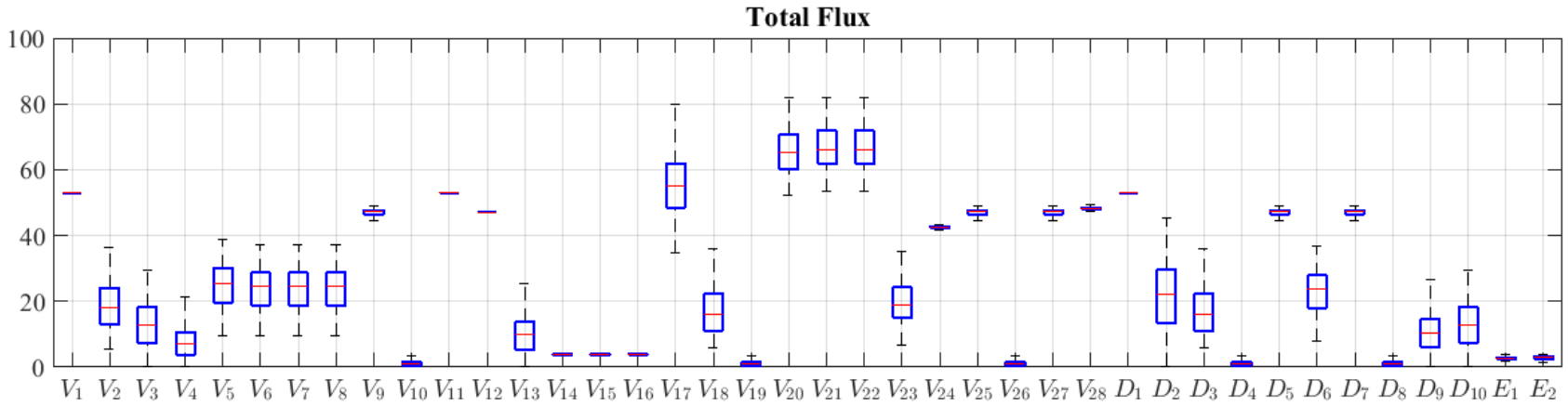

Figure 4.6 Total steady-state flux distribution in *Brachypodium*. The total flux includes both labeled and unlabeled fluxes. The results correspond to the scheme in Figure 4.4.

Because the system is mathematically underdetermined, its steady-state solution is not unique (see 4.3.2 Steady-state analysis). Therefore, a range of admissible steady-state values is possible for each flux. It is worth emphasizing in this context that all solutions in the resulting ensemble are consistent with all pertinent observations; namely:

- Each model in the ensemble captures the experimental data with respect to the label distribution in steady-state fluxes. For instance, V_{24} shows a higher labeled portion than V_{28} when phenylalanine contains the feeding label;
- The lignin compositions and S/G ratios in all scenarios are compatible with experimental data;
- The labeled lignin composition is compatible with $^{13}C_9$-phenylalanine and $^{13}C_9$-tyrosine experimental data; and
- The labeled ferulic acid and *p*-coumaric acid match with experimental data.

Further details are presented in Table 4.1.

Table 4.1 Computational model results compared to experimental data. The model results demonstrate a good match with experimental data in terms of total lignin (A) and the incorporation of label (B).

A				
	H/total lignin (%)	**G/total lignin (%)**	**S/total lignin (%)**	**S/G**
Experimental data	4	41	55	1.09
Model result	4.1	45	51	1.13

B						
	H-lignin*(%)	**G-lignin (%)**	**S-lignin (%)**	**Total lignin (%)**	***p*-Coumaric acid (%)**	**Ferulic acid (%)**
Label incorporation in **[U-$^{13}C_9$]phenylalanine** feeding experiment						
Experimental data	36	22.3	21	22.2	21	23
Model result	19.6	19.1	18.1	18.6	17.2	18
Label incorporation in **[U-$^{13}C_9$]tyrosine** feeding experiment						
Experimental data	24.6	16.5	18.1	18.6	17	13
Model result	15.4	15.9	16.9	16.4	17.8	17

*Label incorporation in H-lignin was not considered as a criterion during the model calibration. The recorded experimental value in the [U-$^{13}C_9$]phenylalanine feeding experiment is greater than the reported label level in phe, which is 35% [92]. As a consequence, we deemed the measurement unreliable and did not use labeled H-lignin measurements.

The boxplots in Figures 4.5 and 4.6 reflect the distributions of admissible values. As can be seen, V_8 admits small values in comparison to its parallel reactions in cytosol compartment, i.e., V_{22} and V_{23}. This result demonstrates that, while the main pathway for the reactions catalyzed by CCR and CAD resides in the cytosol, a relatively small and undisturbed flux through CCR/CAD at the ER is sufficient to establish the metabolic channel necessary for preferential incorporation. In fact, considering the wrinkled environment of the ER surface, it is not hard to imagine that localized pools would keep a small fraction of the pathway undisturbed from exchanges of metabolites with the cytosol.

4.3 Methods

4.3.1 Generic Model Formulation

In a kinetic systems model, the dynamics of the pathway is represented by a system of ordinary differential equations (ODEs) in which the metabolites are the states. The rate of change in each metabolite is determined by sums and differences of all fluxes that directly affect this metabolite. Each flux is a mathematical function of the metabolites and other variables of the system that needs to be selected. Although the fluxes are usually nonlinear functions, the collection of fluxes itself forms a linear system, which can be represented as a matrix equation of the type

$$\dot{\mathbf{X}} = \mathbf{S} \cdot \mathbf{V} \tag{4.1}$$

Here, $\mathbf{X}$ is the vector of metabolites, $\mathbf{S}$ is the stoichiometric matrix, and $\mathbf{V}$ is the vector of fluxes. The stoichiometric matrix § defines the pathway structure. An element $s_{i,j}$ of this matrix equals 1 if flux V_j is directed toward metabolite X_i. It is -1, if flux V_j removes material from metabolite X_i, and it is equal to 0, if flux V_j has no direct effect on metabolite X_i. In long form, the matrix equation can be rewritten for each equation as

$$\dot{X}_i = \sum_{j=1}^{n} s_{i,j} V_j \tag{4.2}$$

where *n* is the total number of fluxes.

4.3.2 Steady-state Analysis

The steady state of a system is important for two reasons. First, many biological systems tend to operate close to such a state, where the overall concentrations of metabolites do not change, even though flux is running through the system. Second, from a mathematical point of view, many analyses at a steady state are much simpler than for the differential equations themselves, because now one has, by definition, $\dot{\mathbf{X}} = 0$, so that all differential equations become explicit algebraic equations that can be analyzed with methods of linear algebra. If all fluxes are known, it is usually not difficult to compute the steady-state of a system. However, the reverse is not true: if only the metabolite concentrations at the steady state are known, it is not easy to compute the corresponding flux distribution, because metabolic systems almost always contain more reactions than variables. In this case, optimization methods like FBA or MOMA need to be employed.

In the *Brachypodium* study, we chose an alternative to FBA and MOMA. Namely, we intended to obtain the most likely solution without specifying an objective function for the FBA optimization. Because the degrees of freedom of a solution to our system are directly associated with diverging branch points, we focused on the flux split ratios (FSRs) at these points. In cases where these FSRs were known, we used their values; otherwise, we performed large-scale Monte-Carlo simulations with thousands of combinations of FSRs and retained only those solutions where all fluxes were positive at all time points of an experiment. This strategy led to the most likely flux profiles. Details of this method are discussed in [120] and Chapter V.

As a simplified example, consider the hypothetical pathway in Figure 4.7, which has two FSRs, and hence two degrees of freedom.

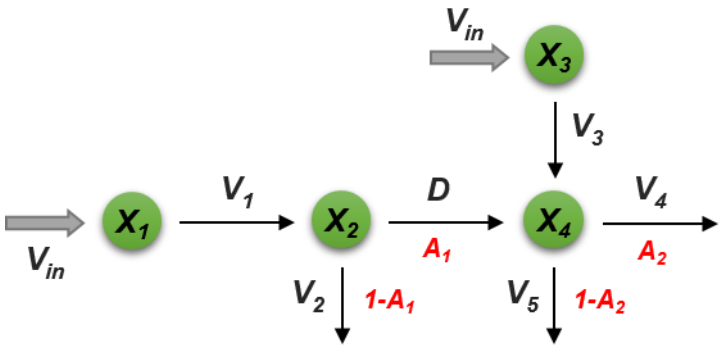

Figure 4.7 Material flow through net fluxes in an illustration example. Without labeling, it is sufficient to model diffusion fluxes as net fluxes. However, this is not the case for labeling experiments (Figure 4.8).

The system of differential equations corresponding to pathway in Figure 4.7 is

$$
\begin{aligned}
\dot{X}_1 &= V_{in} - V_1, \\
\dot{X}_2 &= V_1 - V_2 - D, \\
\dot{X}_3 &= V_{in} - V_3, \\
\dot{X}_4 &= V_3 + D - V_4 - V_5.
\end{aligned}
\tag{4.3}
$$

Given a set of metabolite concentrations over time, the pathway can be driven by infinitely many flux distributions [139]. To determine the most likely, the system in Equation (4.3) is first rewritten in terms of FSRs of the system at the steady state ($\dot{\mathbf{X}} = 0$).

$$
\begin{aligned}
V_1 &= V_{in}, & V_3 &= V_{in,} \\
V_2 &= (1 - A_1) \cdot V_1, & V_4 &= A_2 \cdot (V_3 + D), \\
D &= A_1 \cdot V_1 & V_5 &= (1 - A_2) \cdot (V_3 + D).
\end{aligned}
\tag{4.4}
$$

Now, thousands of pairs (A_1, A_2) of FSRs are randomly generated by Monte-Carlo sampling with $A_i \in [0,1]$. Each pair, entered into the model, yields steady-state values of the fluxes V_1,…, V_5 and D. These are filtered to retain only desired fluxes. For instance, in the actual case study of *Brachypodium*, only those flux profiles are retained that satisfy the following criteria:

- Fluxes take only non-negative values at all time points;
- The lignin composition and S/G ratio are compatible with experimental data.

It is theoretically possible that the estimation strategy based solely on split ratios does not converge to an acceptable solution, and we have discussed means of addressing this situation elsewhere [120]. Here the split-ratio method succeeded without the need for alternative methods.

4.3.3 Modeling ^{13}C-labeling Experiments

The diffusion flux between two pools of the same metabolite in different locations is comprised of two directions (Figure 4.8). Although the two opposing fluxes have a net value, as shown in Figures 4.4 and 4.7, it is necessary to consider them individually when modeling a labeling experiment. The reason is that the labeling content of each pool affects the flow of label, but the net diffusion alone would not reflect the free passing of label in both directions. For instance, the illustration scheme in Figure 4.7 does not allow flow of label from X_4 to X_2 through D when labeled metabolite is fed to the pathway through X_3, but due to the bidirectional nature of diffusion fluxes, it is evident that flow would happen in reality. As Figure 4.8 illustrates, these bidirectional diffusion fluxes form cycles and don't allow the direct computation of steady-state fluxes.

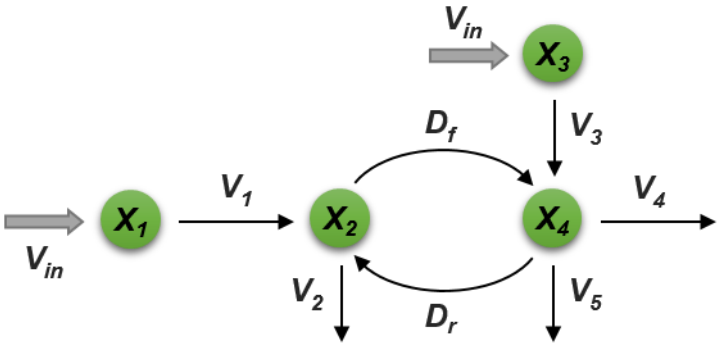

Figure 4.8 Illustration of the flow of label in the same example as Figure 4.7, but with explicit flux directions. In contrast to the scenario in Figure 4.7, labeling experiments mandate the modeling of diffusion fluxes in both directions. Specifically, the simpler model in Figure 4.7 does not allow flow of label from X_4 to X_2 through D when labeled metabolite is fed to the pathway through X_3, but the figure here demonstrates that such flow is clearly possible.

To tackle this issue, we first consider only the net diffusion flux as shown in Figure 4.7 and compute the steady states. Then we consider the bidirectional model in Figure 4.8, employ Equation 4.2, and use conservation of mass for labeled and total fluxes at each metabolite. For a given labeling percentage L_i, where L_i represents the labeled portion of the pool of metabolite X_i, we obtain for the pathway system in Figure 4.8:

$$\begin{aligned} &V_1 + D_r = V_2 + D_f, \\ &L_1V_1 + L_4D_r = L_2V_2 + L_2D_f, \end{aligned} \tag{4.5}$$

which can be rewritten as

$$\begin{aligned} D_r &= \frac{(L_2 - L_1)V_1}{L_4 - L_2}, \\ D_f &= \frac{(L_4 - L_1)V_1}{L_4 - L_2} - V_2. \end{aligned} \tag{4.6}$$

Similar calculations for pool X_4 yield

$$\begin{aligned} &V_3 + D_f = V_4 + V_5 + D_r, \\ &L_3V_3 + L_2D_f = L_4V_4 + L_4V_5 + L_4D_r, \\ D_r &= \frac{(L_2 - L_3)V_3}{L_2 - L_4} - V_4 - V_5, \\ D_f &= \frac{(L_4 - L_3)V_3}{L_2 - L_4}. \end{aligned} \tag{4.7}$$

By equating D_f from Equations 4.6 and 4.7 one obtains

$$L_4 = \frac{L_2V_2 - L_1V_1 - L_3V_3}{V_2 - V_1 - V_3}. \tag{4.8}$$

Assuming that L_1 and L_3 are known from inputs of the pathway, V_{in}, we can compute L_4 from the computed steady-state fluxes in the previous step and an estimated L_2.

Therefore, L_2 is the only unknown to be estimated, and L_4, D_f and D_r can consequently be computed. In fact, we only need to estimate the label level of one of the parallel metabolite pools in the cytosol and ER compartments.

Similar to the use of split ratios, vector L is generated randomly by Monte-Carlo sampling, and the labeled fluxes can then be computed. The labeled fluxes corresponding to Figure 4.8 are

$$\begin{aligned}
V_{1,L} &= L_1 \cdot V_1, & V_{1,U} &= (1-L_1)\cdot V_1,\\
V_{2,L} &= L_2 \cdot V_2, & V_{2,L} &= (1-L_2)\cdot V_2,\\
V_{3,L} &= L_3 \cdot V_3, & V_{3,U} &= (1-L_3)\cdot V_3,\\
V_{4,L} &= L_4 \cdot V_4, & V_{4,U} &= (1-L_4)\cdot V_4,\\
V_{5,L} &= L_4 \cdot V_5, & V_{5,U} &= (1-L_4)\cdot V_5,\\
D_{f,L} &= L_2 \cdot D_f, & D_{f,U} &= (1-L_2)\cdot D_f,\\
D_{r,L} &= L_4 \cdot D_r, & D_{r,U} &= (1-L_4)\cdot D_r.
\end{aligned} \tag{4.9}$$

Closer inspection demonstrates that the model in Figure 4.7, which considers only net diffusion, computes the labeled portion of D as L_2D, which is equal to $L_2(D_f - D_r)$, whereas Equation (4.9) quantifies the net labeled flux as $L_2D_f - L_4D_r$.

The fluxes for the *Brachypodium* example were defined in this manner. Labeled fluxes that satisfied the model criteria for labeling experiments were recorded. The criteria were

- The labeled lignin composition is compatible with $^{13}C_9$-phenylalanine and $^{13}C_9$-tyrosine experimental data; and
- The labeled ferulic acid and *p*-coumaric acid levels match the experimental data.

The recorded flux vectors were plotted using boxplots, which offer a visual representation of the distribution of most likely flux values within their admissible ranges.

4.4 Discussion and Conclusions

The puzzling results of ^{13}C-labeling experiments on lignin pathway in *Brachypodium* have raised the hypothesis of the possibility of distinct phenylalanine and tyrosine pathways. In this chapter we presented a computational model that agreed with the hypothesis in a step by step manner. In fact, we showed that it is not possible for the lignin pathway in *Brachypodium* to have distinct incorporation of phenylalanine and tyrosine in different lignin units without spatial compartmentalization of the pathway. We showed the necessity of this compartmentalization in keeping the fluxes from a uniform dilution. Furthermore, similar to our previous model in switchgrass [3], and the models developed by Lee *et al.* in *Medicago* [18, 58], channeling of downstream enzymes CCR and CAD is a key feature to explain labeling data.

These computational results yield new insights into the dynamics of the pathway, and at the same time pose a question on the correctness of the previously developed lignin models that do not include compartmentalization of the pathway. One should note that in a systemic approach, and in contrast to a mechanistic approach, the overall dynamic behavior of the system is the measure for designing a model. In other words, necessity of including a feature of a system defines the structure of a model, rather than the actual existence of that feature. Therefore, while compartmentalization might be also present in

switchgrass and *Medicago*, the models did not need to include them in order to be able to capture the experimental data.

CHAPTER V

Stepwise Inference of Likely Dynamic Flux Distributions from Metabolic Time Series Data[5]

5.1 Introduction

A key step of any computational modeling is the identification of an adequate mathematical representation of the phenomenon under study. Only with an appropriate representation of all processes involved in the phenomenon will the model have sufficient predictive capacity with respect to new experiments and data. While the importance of an adequate model choice is quite obvious, most modeling efforts begin by simply assuming mathematical formats for the process representations, even though these are often unproven and may be substantially wrong. For instance, many metabolic systems are modeled with Michaelis-Menten rate laws and their generalizations, even though it is not known to what degree these functions, which were developed for analyses *in vitro,* are valid *in vivo* [153, 162, 179].

[5] The material in this chapter has been published as: 120. Faraji, M. and E.O. Voit, *Stepwise Inference of Likely Dynamic Flux Distributions from Metabolic Time Series Data.* Bioinformatics, 2017.

A method that attempts to infer unbiased process representations from metabolic time series data is dynamic flux estimation (DFE), which was introduced a few years ago [118]. DFE consists of two phases: a model-free and a model-based estimation. In phase 1, the dynamic flux profiles are estimated. This occurs through smoothing the time series of metabolite concentrations, for which numerous techniques are available (*e.g.*, [180-184]). The slopes at many time points are then substituted for the time derivatives on the left-hand sides of the differential equations (ODEs) of the model, so that each original ODE is replaced with a system of algebraic equations, which are linear in the fluxes. If the stoichiometric matrix of the system is square and has full rank, these equations can be solved directly with methods of linear algebra. The result is a collection of flux profiles, which can be plotted against the appropriate variables and reveal the shape of each flux, although not its mathematical format. In phase 2, functional formats of the fluxes are chosen, based on their shapes, and parameterization provides a fully characterized kinetic model. By separating the procedures of flux profile estimation and parameterizing the flux profiles, the method minimizes compensation errors which are commonplace in simultaneous parameterizations of the entire system [118, 172, 185].

The drawback of DFE in phase 1 is that most stoichiometric matrices are 'wide,' because the number of fluxes exceeds the number of metabolites in most pathways. Thus, the system is underdetermined, and linear algebra tells us that infinitely many solutions exist for such a system. As a consequence, a given time series of metabolic concentration data is theoretically consistent with infinitely many different dynamic flux distributions. Expressed differently, a wrong model may not only be obtained from a given dataset due to invalid biological assumptions or omissions of important features, but also due to the

fact that numerous equivalent models may exist that represent the dynamics of this dataset with the same, often very reasonable accuracy. However, these alternative models may diverge substantially for other datasets. This divergence implies that a model may be representative of the system within a limited vicinity of the provided dataset, but easily fails to explain the pathway system in an expanded space.

Several approaches have been proposed to address this stoichiometric under-determinedness, ranging from *ad hoc* strategies to generic mathematical methods. Biologically the most straightforward solution is the use of additional, independent information regarding a flux, which may come from knowledge of its kinetics. Also intuitive is the merging of fluxes where, for instance, a pair of forward and reverse reactions is replaced with a single net reaction, which reduces the degrees of freedom by one. Along the same lines, a cycle of metabolites may be condensed into a single pool [186]. From a mathematical point of view, one could try to address the underdetermined systems with pseudoinverse techniques [187-189]. However, using something like the Moore-Penrose pseudoinverse typically leads to negative fluxes, which are often not feasible biologically [139]. Because we pretend to know the directionality of all fluxes, we will show that we can filter out solutions that include negative fluxes, which we consider infeasible. The generic, interesting problem we have is not that there is no solution, but that there are too many solutions. Furthermore, we know that if the problem is set up correctly, the true solution should be among them. Thus, appropriate constraints allow us to eliminate wrong solutions and to restrict the space of feasible solution, which should still contain the true solution. Computational approaches may introduce upper and lower bounds for some or all of the fluxes, which constrain the feasible solution space. One could also limit the space of

potential flux profiles by eliminating oscillatory solutions that might not be consistent with the biological system. A different strategy for systems at steady state is the optimization of the solution within the feasible space. As a prominent example, stoichiometric and flux balance analyses (FBA) often define maximal growth as an objective function for microbial systems and determine the flux distribution that optimizes this objective function [75-77, 86]. For other systems, the choice of a suitable objective function is sometimes unclear.

In this work, we propose an essentially unbiased method for inferring dynamic flux distributions in underdetermined metabolic models that are statistically most likely. The approach reduces the degrees of freedom of the stoichiometric matrix in a stepwise manner. In favorable situations, the method is able to eliminate all degrees of freedom in a small number of steps and quickly yields a statistically likely distribution of fluxes.

5.2 Methods

The dynamics of a metabolic pathway system is typically represented by the stoichiometric equation

$$\dot{\mathbf{X}} = \mathbf{N} \cdot \mathbf{V} \tag{5.1}$$

In this representation, the state variables, X_i, are the metabolites and $\dot{\mathbf{X}}$ is the corresponding vector of the rates of change. $\mathbf{V}$ is the vector of fluxes and $\mathbf{N}$ is the stoichiometric matrix, which describes the connectivity between the fluxes and the pools of metabolites. We assume that time series measurements of the metabolite concentrations $\mathbf{X}_{t_1}, \mathbf{X}_{t_2}, ..., \mathbf{X}_{t_K}$ are available at K time points. It is to be expected that these

data are spread out relatively far and contain moderate noise which, at present, is the typical situation. Nevertheless, continuing improvements in experimental techniques render it likely that future time series will be generated in replicates and are denser and less noisy than what is feasible today. Even moderately noisy data contain rich information, and numerous good smoothers have been developed over the past decades. As long as the data represent the time trends adequately, these smoothers can be employed not only to approximate the true trends, but also to obtain estimates $\mathbf{S}_{t_1}, \mathbf{S}_{t_2}, ..., \mathbf{S}_{t_K}$ of the slopes $\dot{\mathbf{X}}_{t_1}, \dot{\mathbf{X}}_{t_2}, ..., \dot{\mathbf{X}}_{t_K}$ the *K* time points of measurements. These slopes may be positive or negative, depending on the situation. For instance, it may be possible that a substrate is being used up, which will lead to a decreasing trend. In terms of locally negative slopes that are entirely due to noise in the data, effective modern smoothers provide for the option of predetermining a desired balance between data fit and roughness of the resulting trend curve (*e.g.*, [180, 183, 190, 191]). In our case, it is important to capture the trend rather smoothly. Of course, there will always be cases where the data are so noisy that they do not reflect the true trends. If so, this method, as well as many others, will not yield reliable results and, in fact, the data themselves seem to be of little value in such cases.

Substituting the estimated slopes for the time derivatives at the *K* time points, the stoichiometric equation (5.1) becomes

$$\mathbf{S} = \mathbf{N} \cdot \mathbf{V} \tag{5.2}$$

which, at every time point, is a set of linear algebraic equations in the flux values **V**, because the left-hand side now contains numerical values [192-195]. Of course, the fluxes are functions of metabolites and thus of time, but at each time point, Equation 5.2 is a

regular algebraic matrix equation in flux values. If it is possible to solve these equations at every time point, the collective result is a complete set of flux values, for each process and for every time point. Expressed differently, a set of metabolic time series data can be converted into a complete set of fluxes over the measured time horizon, and this conversion is essentially assumption free and unbiased. The fluxes are numerically determined, and their functional formats are unknown. Appropriate formats may be inspired by plots where the values of a flux are plotted against the metabolites that affect this flux, one time point at a time. This procedure was introduced in this journal as Dynamic Flux Estimation (DFE) [118] and is also the basis for a novel manner of nonparametric dynamic modeling [119].

DFE works very well if Equation 5.2 can be solved uniquely. However, the number of fluxes in metabolic pathway systems is typically greater than the number of metabolites, which causes the system to be underdetermined. In other words, given the slopes at a set of specific time points, the system has infinitely many solutions for the vector V at these time points, which all match the data perfectly. This infinite set of flux solutions corresponds to an infinite set of models of the system, which all perfectly coincide in the observed dataset. One could surmise that more data points or time series would solve the problem, but that is not necessarily the case, because the redundancy is a structural feature of the stoichiometric matrix and, thus, the connectivity of the pathway system. As it was mentioned earlier, numerous approaches have been proposed to address this issue, including general techniques and *ad hoc* strategies. All of these approaches require either additional biological information about the system or assumptions that may or may not be justified. Here, we propose a mathematical and computational tool for identifying flux distributions that are in a statistical sense most likely.

5.2.1 Split Ratios at Branch Points

Most metabolic pathways contain branch points where a compound is used as the starting substrate for two or more pathways with different end products. The amounts of mass entering these different pathways are characterized by the flux split ratio and increase the degrees of freedom in Equation 5.2. Figure 5.1 depicts a very simple hypothetical pathway with two branches, at X_1 and X_3. The three metabolites and five fluxes lead to two degrees of freedom. At the steady state, the first split ratio, A, is defined as the ratio of V_1 to V_{in}, and B is correspondingly given as the ratio of V_4 to V_3. In general, A and B can have any real values between 0 and 1. Due to the conservation of mass, the ratio of V_2 to V_{in} is $1-A$, and the ratio of V_5 to V_3 is $1-B$. Analogous considerations hold for more than two pathways diverging from a branch point.

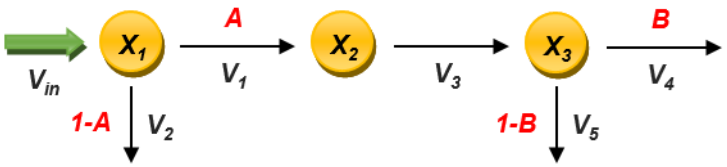

Figure 5.1 A hypothetical pathway with two degrees of freedom. Split ratios determine the percentage of each flux leaving a metabolite pool toward different pathways, compared with the total influx to the pool.

At a transient state, the rate of change in the metabolites needs to be taken into account as well. For the system in Figure 5.1, system equations are

$$\begin{aligned}\dot{X}_1 &= V_{in} - V_1 - V_2,\\ \dot{X}_2 &= V_1 - V_3,\\ \dot{X}_3 &= V_3 - V_4 - V_5.\end{aligned} \tag{5.3}$$

Given a known input, V_{in}, and the rates of change in the metabolites, we can rewrite Equation 5.3 such that the fluxes are functions of the input V_{in} and the split ratios:

$$\begin{aligned}V_1 &= A\cdot\left(V_{in} - \dot{X}_1\right), & V_4 &= B\cdot\left(V_3 - \dot{X}_3\right),\\ V_2 &= (1-A)\cdot\left(V_{in} - \dot{X}_1\right), & V_5 &= (1-B)\cdot\left(V_3 - \dot{X}_3\right).\\ V_3 &= V_1 - \dot{X}_2,\end{aligned} \tag{5.4}$$

In general, the numerical values of split ratios are unknown, although coarse information is available for many branch points. For instance, the amount of material branching off glycolysis into the pentose phosphate pathway is often 10% or less (*e.g.*, [196-198]).

If reliable information is available, it can be used to define the numerical range for a split ratio. If not, the range is taken to be [0, 1]. Split ratios collectively form a vector in $\mathbb{R}^n$, where n denotes the degrees of freedom. To explore the repertoire of behaviors of a given system, a large-scale Monte Carlo simulation may be conducted where the set of split ratios forms a hypercube in $\mathbb{R}^n$. For each time point, the randomly generated split ratios are plugged into Equation 5.4, and a set of vectors of fluxes is computed. Some of these are likely infeasible due to biological constraints; for instance, they could contain negative rates for other fluxes in the system. Other constraints, such as upper or lower bounds for some of the fluxes, or any other *a priori* knowledge regarding the split ratios, are very helpful, as they reduce the feasible solution set. For small numbers of branch points, grid sampling may be preferred over a Monte-Carlo simulation.

5.2.2 Metabolic Energy Assumption

Even if various constraints can be imposed, the solution set typically contains infinitely many solutions with diverse dynamic qualities, which all are equivalent in a sense that they fit the metabolic time series data. However, these equivalent solutions often differ quite substantially in the total amount of metabolic energy they incur, because some flux distributions contain overall much higher flux values than others, even though they result in exactly the same metabolic profiles. A measure for this total energy is the Euclidean norm, which we examine at each time point (Figure 5.2).

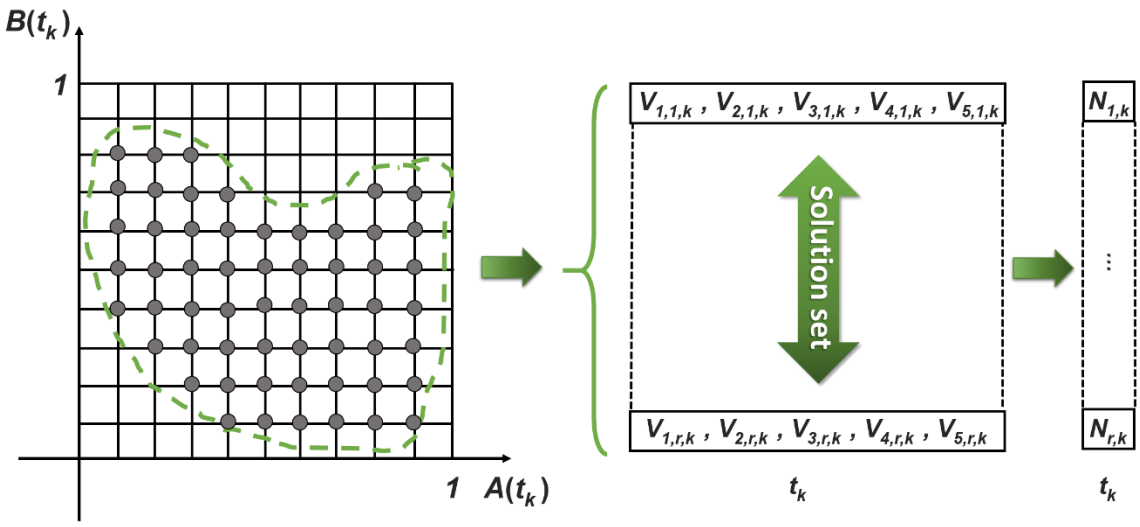

Figure 5.2 Admissible solutions, matrix of the admissible fluxes and array of admissible flux norms. Using a grid search (here for an illustration pathway with two branch points) leads to a matrix representing the entire set of admissible fluxes of the pathway at every time point. To capture the entire time horizon, these matrices are stacked up. At any given time point t_k, some grid points may be inadmissible, for instance, because they contain negative flux values. As a consequence, only a subset of grid points (gray circles) is admissible. Each grid point corresponds to a vector of admissible fluxes that constitute the row of the matrix at each time slice, and different rows are admissible

solutions that correspond to different grid points. It should be noted that different sets of grid points may be admissible at different time points. For each row in the matrix of the fluxes, the Euclidean norm is computed. This collection of norms forms a matrix where the columns correspond to different time points and the rows to admissible solutions at each time point.

Collectively, for a Monte-Carlo or grid sample of flux distributions, the values of these norms form a frequency distribution at each time point.

Thus, the initially uniform distribution of admissible split ratios is nonlinearly transformed into a non-uniform distribution of norms. The benefit of this transformation is that we can now zoom onto intervals with the most likely flux distributions, namely those intervals that result from many more combinations or split ratios than others, and disregard solutions outside this interval. Specifically, we can estimate the most likely flux distribution norm y^* by minimizing the sum of the distance functions, weighted by the probability distribution of all admissible solutions at each time point. Thus, the objective function takes the form

$$\min_{y^*} \sum_{i=1}^{r} \Pr(y_i) \cdot \left(y_i - y^*\right)^2, \tag{5.5}$$

where the y_i's are the norms of admissible solutions and $\Pr(y_i)$ is the probability of y_i according to the distribution of the *r* admissible solutions at each time point. Setting the first derivative in Equation 5.5 equal to zero yields the estimator

$$y^* = \sum_{i=1}^{r} \Pr(y_i) \cdot y_i, \tag{6}$$

which is the expected value of the norm distribution, and thus its mean, μ. Using this mean, as well as the standard deviation σ at each time point, we select those norms that fall within the range $\mu \pm \sigma$. In a distribution resembling the normal distribution, the range $\mu \pm \sigma$ roughly covers 68% of the data. In cases of highly skewed or multimodal distributions, other selection strategies for likely ranges might be preferable. For instance, one could take the smallest range that corresponds to 50% of the mass of the distribution.

5.2.3 Reducing the Degrees of Freedom

It is sometimes technically feasible to measure influxes and effluxes, and maybe even interior fluxes within the systems. Such measurements automatically reduce the degrees of freedom [99, 141]. For instance, if flux V_3 can be measured, the i^{th} equation in (5.1) becomes

$$\frac{dX_i}{dt} - N_{i,3} \cdot V_3 = \sum_{j=1, j\neq 3}^{m} N_{i,j} V_j \tag{5.7}$$

so that the entire system effectively contains one variable less. Nonetheless, it is rare that the model can be reduced to a full-ranked system. Thus, even though it is possible to constrain the underdetermined solution, the resulting system typically still permits infinitely many solutions.

To achieve further reduction, the sets of fluxes and split ratios are investigated individually. The generic argument is the following: If the vast majority of simulations identifies a particular split ratio that always falls within the same narrow range, then this range is assumed to be most likely, and the split ratio is subsequently fixed at the mean or

median of this range. This setting reduces the degrees of freedom among fluxes by one. With the split ratio fixed, a new round of simulations is initiated, and further split ratios within narrow ranges are again fixed. This procedure eventually leads to a unique flux distribution that is in some sense most likely (Figure 5.3).

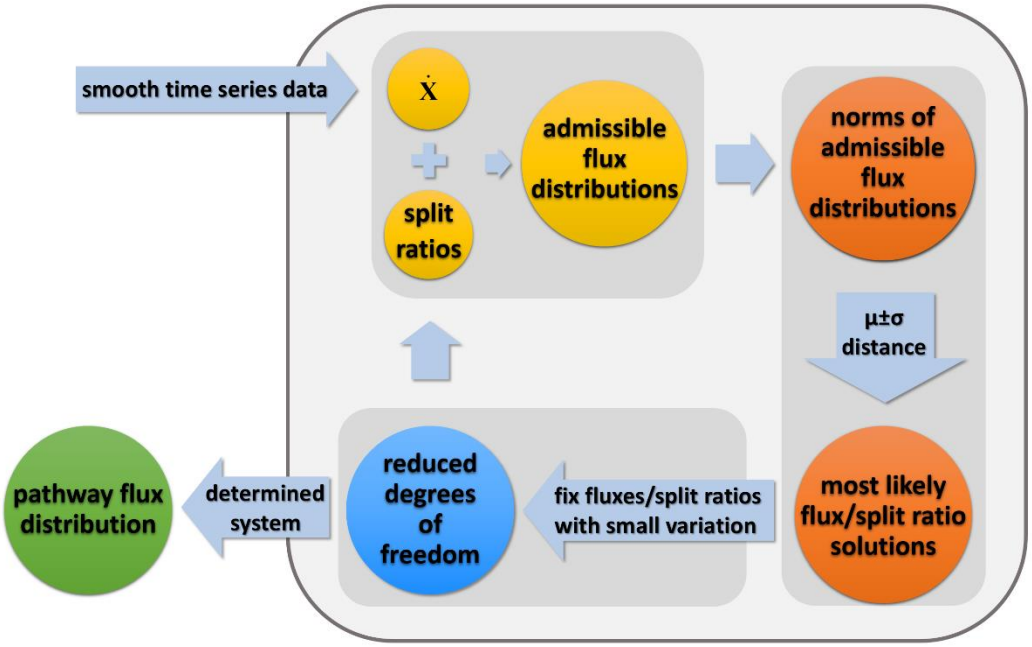

Figure 5.3 Flowchart of the proposed inference method for flux distributions. The input to the method consists of smoothed time series of metabolite concentrations, from which slopes are computed. A Monte-Carlo or grid simulation generates very many flux distributions corresponding to sampled split ratios. Non-negative (and possibly otherwise constrained) flux vectors are retained. Euclidean norms of the retained flux vectors are calculated at each time point, and all solutions within the range $\mu \pm \sigma$ are kept. The time series of split ratios and the flux distributions of the selected solutions are plotted, and solutions with relatively small variations are numerically fixed at their means or medians in order to decrease the degrees of freedom. The process is iterated until it yields a unique solution.

A challenging situation occurs when all fluxes hold wide variations, which would prevent the algorithm from proceeding. This case is addressed in the discussion.

5.3 Results

Case study: Lignin biosynthesis in switchgrass

Switchgrass (*Panicum virgatum*) has been identified by the U.S. Department of Energy as a target source for bioethanol production (BioEnergy Science Center; http://bioenergycenter.org/besc/research/biomass.cfm). In previous work (Chapter 3 and [3], we analyzed the structure and regulation of its lignin biosynthesis pathway, with the ultimate goal of prescribing gene or enzyme modulations leading to more favorable lignin production. Specifically, we constructed a computational model of the pathway using steady-state data of the wild type and four transgenic strains of switchgrass (Figure 5.4). The data included measured levels of H-, S- and G-lignin, which are the building blocks of the lignin heteropolymer. Here, we use this model as an illustration where we assume to have full knowledge of the system and its features. Specifically, we generate virtual time series of concentrations of all metabolites in the pathway and pretend that they were smoothed experimental data. The aim is to infer the most likely flux distribution using the method proposed in the previous section.

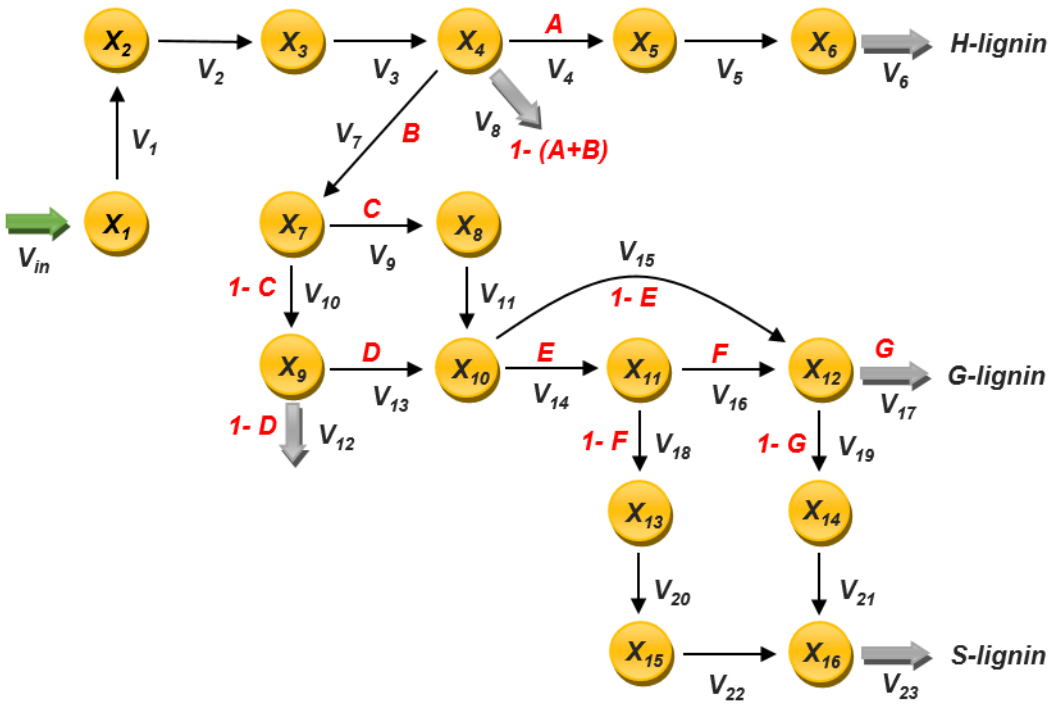

Figure 5.4 Lignin biosynthesis pathway in switchgrass. Seven branch points give rise to as many degrees of freedom. At each branch point, the total efflux splits into two or three fluxes. The parametric split ratios are shown with capital letters. Conservation of mass dictates the sum of the split ratios at each branch point to be 1. Adapted from [3].

The lignin biosynthesis pathway consists of 16 metabolites and 23 fluxes, which result in seven branches; in Figure 5.4, the split ratios are marked with capital letters. The variables are

$$
\begin{aligned}
&X_1: \text{phenylalanine}, && X_9: \text{ferulic acid},\\
&X_2: \text{cinnamic acid}, && X_{10}: \text{feruloyl-CoA},\\
&X_3: \text{p-coumaric acid}, && X_{11}: \text{coniferaldehyde},\\
&X_4: \text{p-coumaroyl CoA}, && X_{12}: \text{coniferyl alcohol},\\
&X_5: \text{p-coumaryl aldehyde}, && X_{13}: \text{5-OH-coniferaldehyde},\\
&X_6: \text{p-coumaryl alcohol}, && X_{14}: \text{5-OH-coniferyl alcohol},\\
&X_7: \text{caffeic acid}, && X_{15}: \text{sinapaldehyde},\\
&X_8: \text{caffeoyl CoA}, && X_{16}: \text{sinapyl alcohol}.
\end{aligned}
\tag{5.8}
$$

The system equations can be rewritten for the split ratio analysis as follows:

$$
\begin{aligned}
V_1 &= V_{in} - \dot{X}_1, & V_{13} &= D\cdot\left(V_{10} - \dot{X}_9\right),\\
V_2 &= V_1 - \dot{X}_2, & V_{14} &= E\cdot\left(V_{11} + V_{13} - \dot{X}_{10}\right),\\
V_3 &= V_2 - \dot{X}_3, & V_{15} &= \left(1-E\right)\cdot\left(V_{11} + V_{13} - \dot{X}_{10}\right),\\
V_4 &= A\cdot\left(V_3 - \dot{X}_4\right), & V_{16} &= F\cdot\left(V_{14} - \dot{X}_{11}\right),\\
V_5 &= V_4 - \dot{X}_5, & V_{17} &= G\cdot\left(V_{15} + V_{16} - \dot{X}_{12}\right),\\
V_6 &= V_5 - \dot{X}_6, & V_{18} &= \left(1-F\right)\cdot\left(V_{14} - \dot{X}_{11}\right),\\
V_7 &= B\cdot\left(V_3 - \dot{X}_4\right), & V_{19} &= \left(1-G\right)\cdot\left(V_{15} + V_{16} - \dot{X}_{12}\right),\\
V_8 &= \left(1-A-B\right)\cdot\left(V_3 - \dot{X}_4\right), & V_{20} &= V_{18} - \dot{X}_{13},\\
V_9 &= C\cdot\left(V_7 - \dot{X}_7\right), & V_{21} &= V_{19} - \dot{X}_{14},\\
V_{10} &= \left(1-C\right)\cdot\left(V_7 - \dot{X}_7\right), & V_{22} &= V_{20} - \dot{X}_{15},\\
V_{11} &= V_9 - \dot{X}_8, & V_{23} &= V_{21} + V_{22} - \dot{X}_{16},\\
V_{12} &= \left(1-D\right)\cdot\left(V_{10} - \dot{X}_9\right).
\end{aligned}
\tag{5.9}
$$

Here V_{in} is the input, and $[A,B,C,D,E,F,G]$ is the vector of split ratios. For normalized comparisons, V_{in} is set as 100%, so that the input of the system is a unit flux which is distributed throughout the pathway.

The seven degrees of freedom correspond to an initial sampling space in $\mathbb{R}^7$. *A priori* knowledge about the pathway helps us to reduce the sampling space from a unit hypercube to a smaller volume. Namely, biological information about the pathway attests that the ratio between the production of S-lignin and G-lignin in switchgrass is in the vicinity of 1, which allows us to define a constraint for the ratio of V_{23}/V_{17}. It is also known that H-lignin accounts for only about 3% of the total lignin. We use this information to constrain A loosely to a value smaller than 10% of the flux that leaves the pool X_4.

Furthermore, V_8 consumes at most 10% of the efflux from X_4. Thus, along with the constraint for A, we may set a loose lower bound of 80% for B. Similarly, it is reasonable to assume that V_{12} accounts for roughly 10% of the efflux from X_9, which allows us to constrain D. Taken together, the sampling space for a grid search in the first round is

$$\begin{aligned} &A \in [0.01, 0.1], \\ &B, D \in [0.8, 0.99], \\ &C, E, F, G \in [0.01, 0.99]. \end{aligned} \tag{5.10}$$

Regarding the S/G ratio, we set the range

$$\left| V_{23} / V_{17} - 1 \right| < 0.2, \tag{5.11}$$

which allows for 20% deviation from 1.

Using the settings in Equations 5.10 and 5.11, we generated a grid in $\mathbb{R}^7$. Each grid point corresponds to a seven-dimensional vector that contains the values for the seven split ratios. We substituted these vectors back into Equation 5.9 and computed the flux distributions for all grid points. Intriguingly, only 5% of these flux distributions had entirely positive values at all time points and were retained.

As the next step we computed the norms of the flux distributions. Figure 5.5 is a visualization of the time-array of the norms. Each vertical slice corresponds to a column in the matrix of the norms and depicts the distribution of the norms of the admissible fluxes at a specific time point. The gray band exhibits the range $\mu \pm \sigma$. As it was described in the Methods section, we record the solutions within the gray band and discarded the rest.

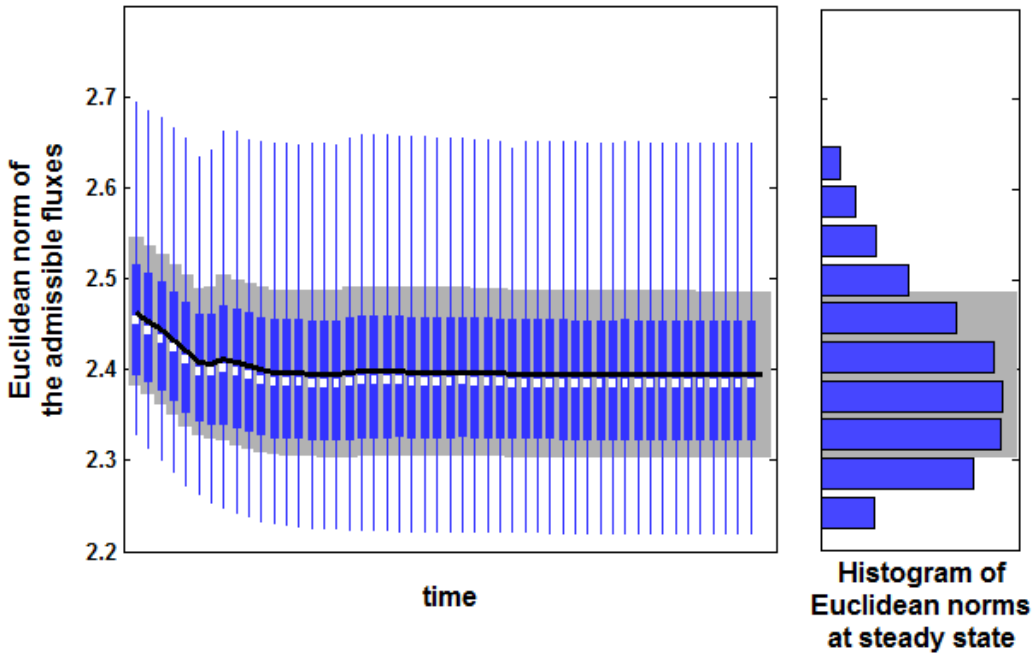

Figure 5.5 Distribution of flux norms in iteration 1. Left panel: The gray band depicts the range $\mu \pm \sigma$, which contains about two thirds of the solutions, at each time point; the mean is shown in black. The thick dark blue boxes represent the second and third quartiles, and the white line is the median, which is similar to the mean. The thin blue lines are the first and fourth quartiles. Right panel: Histogram of norms in the left panel at the steady state, and range $\mu \pm \sigma$ (grey).

All split ratios and flux distributions corresponding to the retained solutions (gray band in Figure 5.5) are plotted in Figure 5.6. The plots indicate that both A and B show small variations. The computed mean of A is 0.04. Therefore, for the next iteration we fix $A = 0.04$, which decreases the degrees of freedom by one. At this point, we could have fixed B as well, but for our illustration we will not take this shortcut. The same results, restricted to the steady state, are shown in Figure 5.7i for comparisons with further iterations.

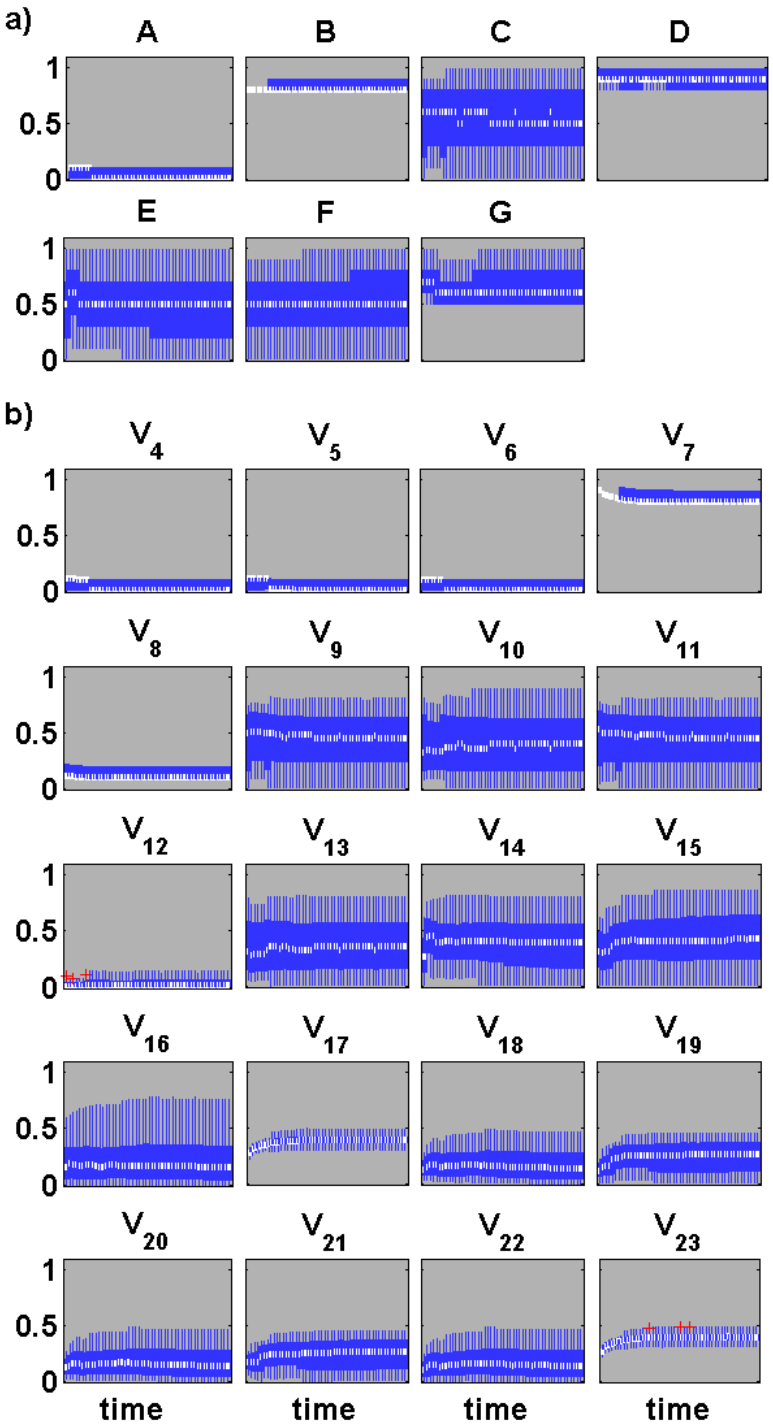

Figure 5.6 Split ratios and flux distributions within the range $\mu\pm\sigma$ of admissible solutions in iteration 1. Similar to Figure 5.5, the boxplots at each time point reflect the quartiles. Panel (**a**) depicts the split ratios and panel (**b**) the flux distributions. Note that four of the fluxes already exhibit rather narrow ranges. The first three fluxes are independent of the split ratios, and therefore not shown.

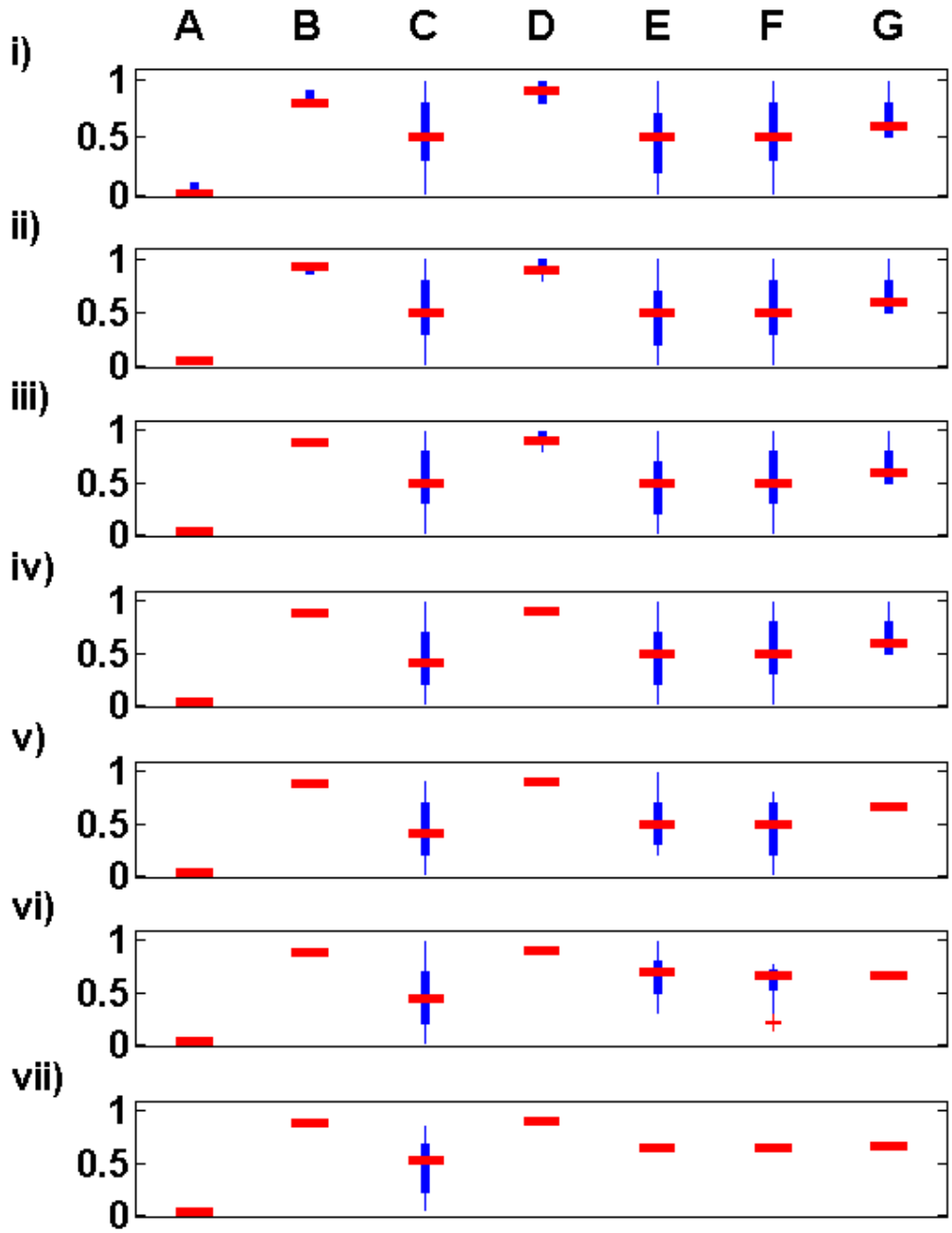

Figure 5.7 Steady-state split ratios within the range $\mu \pm \sigma$ of admissible solutions. The subpanels correspond to the last time point (steady state) of each iteration. Each horizontal bar shows the median at the given iteration, and the boxplots represent the quartiles of split ratios.

Fixing A reduces the sampling space for the second iteration to $\mathbb{R}^6$. Since $A = 0.04$, and because we considered an upper bound of 10% for V_8, we can refine the sampling interval of B and set its lower bound to 86%. Computing the flux distributions using the refined sampling space and screening for nonnegative solutions determines the admissible solutions of iteration two. Similar to iteration 1, the norms of the admissible fluxes are computed and the solutions within the range $\mu \pm \sigma$ are retained. The time array of split ratios and flux distributions corresponding to the retained admissible solutions are exhibited in Appendix B: Figure B.1. Figure 5.7ii exhibits the split ratios corresponding to the retained admissible solutions at steady state. Split ratio B exhibits the smallest variation, and therefore is fixed at its mean value, 0.89, for the third iteration.

For the third iteration we sample the array of $[C, D, E, F, G]$ from $\mathbb{R}^5$, and compute the flux distributions. The admissible fluxes within the range $\mu \pm \sigma$ are retained (see Appendix B: Figure B.2b). The corresponding split ratios at steady state are illustrated in Figure 5.7iii. The plot suggests that D is the best candidate to be fixed in this iteration. We assign $D = 0.91$, the mean value, and proceed to the fourth iteration.

Following the same procedure as in previous iterations, we achieve the split ratios and flux distributions of the retained admissible solutions with four degrees of freedom, associated with C, E, F and G (see Appendix B: Figure B.3). Although the width of the G distribution is not all that small, the standard deviation of G is relatively small and the distribution is dense at the lower bound. Therefore, for the next iteration, we fix G on its mean value, equal to 0.66 (Figure 5.7iv).

Figure 5.7v indicates that the distributions of the remaining split ratios in iteration five are relatively wider than before so that the next choice is ambiguous. However, the plot of V_{18} in Appendix B: Figure B.4b indicates a very narrow range, independent of the split ratios. Thus, we can fix V_{18} and compute V_{16} for any given value of V_{14} such that conservation of mass is preserved at X_{11}. The split ratio F is then automatically determined by V_{16}/V_{14} (see Figure 5.4) rather than by sampling from the grid, which reduces the degrees of freedom to two. Furthermore, fixing V_{18} leads to the direct identification of V_{20} and V_{22}. One could simultaneously fix V_{17} and V_{19}. However, this strategy would not reduce the degrees of freedom further, but add an extra constraint, which would make the system overdetermined (see Figure 5.4). A regression model could then determine the optimal flux distribution. We do not pursue this solution here.

In iteration six, only the split ratios C and E remain to be unknown (Figure 5.7vi). Variation in the split ratio F is explained by the fact that although V_{18} is fixed, V_{16} is dependent on V_{14}, where V_{14} itself is a function of the unknown split ratio E (see Appendix B: Figure B.4). Therefore, sampling E from the grid space leads to variation in V_{14}, V_{16} and consequently F as the ratio of the two. Thus, we set E to its mean value of 0.65 and proceed to the next iteration.

Figure 5.7vii shows that in this iteration the range $\mu \pm \sigma$ of the norms of the admissible solutions (corresponding to the gray band) is not sensitive to the value of split ratio C, and the retained flux distributions (Appendix B: Figure B.5a) are such that C can have values almost throughout the entire interval of (0, 1). This observation leads to the conclusion that any error in C does not affect the flux norms distribution much. At the

same time, choosing the mean value statistically minimizes the error of the estimation, as we argued in the Methods section. Thus, we set C to is mean, 0.45.

With this setting, all degrees of freedom are eliminated, and the system is fully determined by the end of iteration seven. Figure 8 exhibits the likely split ratios and flux distributions, superimposed on the corresponding fluxes in the reference model. The match between the true and the estimated solution is very good. Appendix B: Figure B.7 depicts the time-trend of the norms of the likely flux distribution.

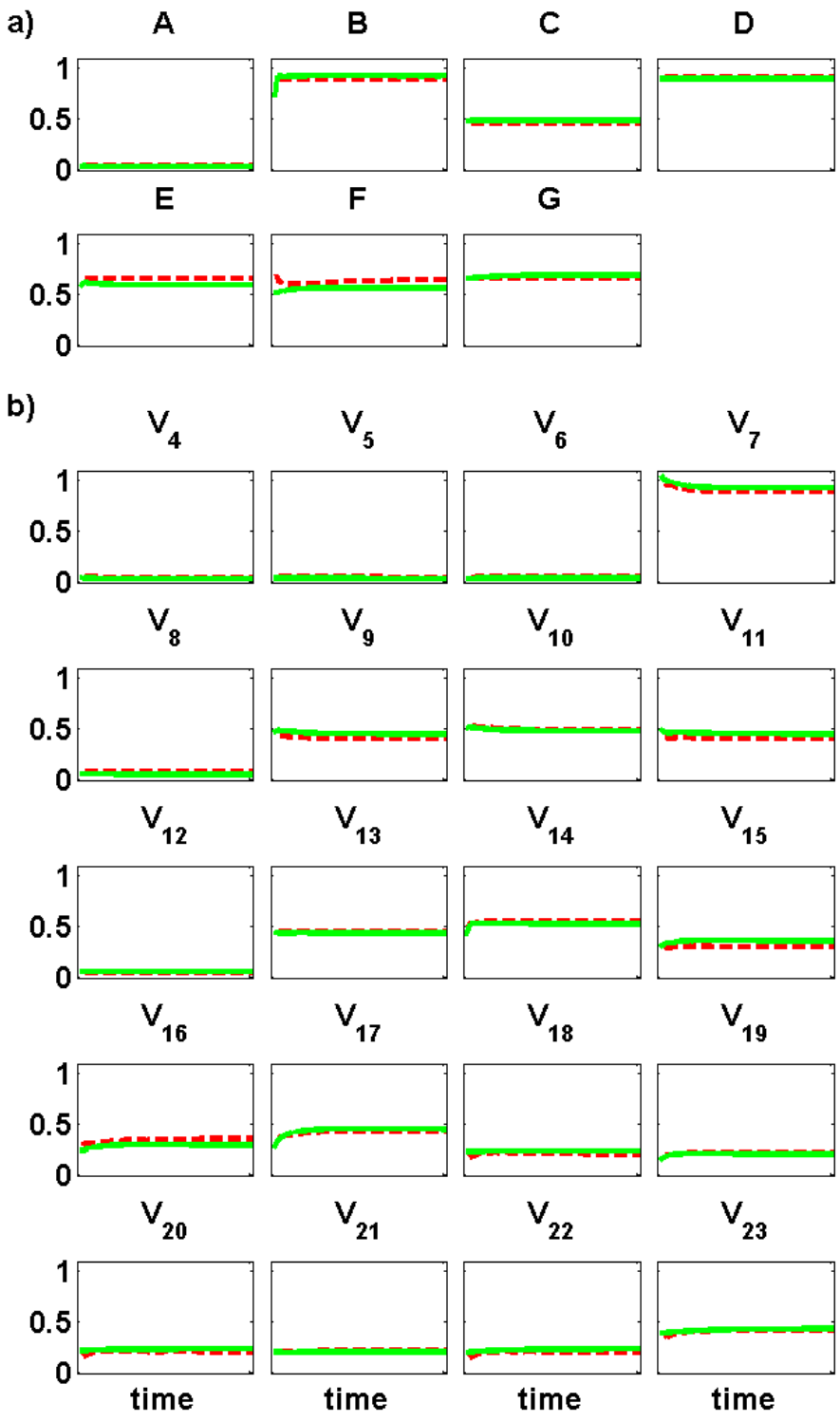

Figure 5.8 Inferred likely split ratios and flux distributions (dashed red) in comparison with the corresponding model features (green).

5.4 Discussion and Conclusions

In this chapter, I propose a method for inferring likely dynamic flux distributions in metabolic pathways from time series of metabolite concentrations. The method utilizes customized computational techniques that explore the space of solutions given by the stoichiometric matrix. Importantly, the method readily allows the inclusion of *a priori* knowledge regarding the pathway, including diverse biological constraints. To compare flux distributions collectively, we employed the Euclidean norm, which in some sense is a metric of the total metabolic energy consumption. High values of this norm indicate high flux rates, which seem wasteful, since all flux distributions yield exactly the same metabolic profile. Very low values are presumably disadvantageous as well, as they do not provide enough robustness and bandwidth to allow the system to respond to changes in metabolic demand.

Thus, we decided to focus on those combinations of split ratios that resulted in flux distributions in the center of the distribution of their norms, which we defined as $\mu \pm \sigma$, and which contains about two thirds of the mass of the distribution if it is more or less normal. This strategy led to some rather narrow and some wider bands of dynamic trends, which suggested which split ratios to fix in the next iteration. Ultimately, this process resulted in unique, time-dependent distributions.

For the illustration example of the pathway of lignin biosynthesis in switchgrass, the method converged within a few steps and yielded flux trends very similar to those in the model used to generate the data. This strong similarity between the inferred and actual

fluxes certainly does not validate the method, but lends it support. The method performed similarly well when we applied it to a simplified model of purine metabolism. These results can be found in the Appendix B. We furthermore tested a much simpler version of this method to a fermentation pathway with two branch points (results not shown; but see [119, 199, 200]).

Like any other modeling framework, the proposed method has its own advantages and drawbacks. As we demonstrated, it works well under favorable conditions. However, the method has two issues. First, it requires good, representative data, which at this point in time are rare. However, the methodologies of molecular biology have advanced very rapidly in recent years [201-204]. Whereas there were essentially no time series data two decades ago, they have become quite common, and it is to be expected that more, better and cheaper methods will emerge for measuring representative time series datasets. Secondly, the method emphasizes statistical likelihood, but it is of course possible that some pathway is naturally parameterized in a very specific, non-average fashion. The only guard against this situation is additional biological information that may be used to constrain some split ratios or fluxes.

A technical issue may ensue when none of the dynamic trends in split ratios or flux norms displays narrow bands. One possible reason for this situation is the existence of unrecognized relationships among fluxes. Such relationships may be identifiable with methods such as principal component analysis (PCA). For instance, it might be possible to identify a specific relationship between two fluxes that allows the numerical coupling between the two. This situation lowers the degrees of freedom by one, and allows the algorithm to continue.

The choice of an adequate model representation is paramount for uses of a model under new conditions. A considerable challenge in this model selection is the fact that compensation among its components may allow drastically different models that fit training data very well, but may fail spectacularly for other data. Dynamic Flux Estimation (DFE) addresses this issue by characterizing the shapes of individual fluxes within pathway systems from metabolic time series data. These flux shapes can subsequently be converted into explicit, parameterized functions [118] or into libraries for nonparametric modeling [119]. Unfortunately, DFE suffers from the existence of infinitely many equivalent solutions in underdetermined stoichiometric systems, and there is little guidance with respect to the best solutions within this set. The method introduced here reveals likely dynamic flux distributions, based on relatively general assumptions. Of course, the method is not failsafe, but with the quickly advancing development of techniques for generating high-quality experimental time series data, we expect it to become increasingly more powerful and provide a straightforward and relatively unbiased approach to the computational characterization of metabolic pathways.

CHAPTER VI

Nonparametric Dynamic Modeling[6]

6.1 Introduction

Ever since the digital revolution drove analog computing to the brink of extinction, the design of computational models for complex systems has become an effort in choosing optimal mathematical representations and their parameter values. For most physical and engineering systems, the choice of model functions is directly guided by our rather solid understanding of basic physical concepts, such as mechanical or electrical forces, dilution and dispersion phenomena, optical processes, and the features of electric circuits. Biological systems are, of course, objects of the physical world and must therefore obey the laws of physics, but most processes that govern even moderately sized biological systems are so convoluted that they cannot be dissected into elementary physical representations [205]. As an example, the transmission of a neuronal signal at a dopamine synapse requires electrical activation, the prior biochemical production of dopamine and its packaging into membrane vesicles, the move of these vesicles through the crowded cytoplasm toward the synapse, the merging of vesicle and cell membranes,

[6] The material in this chapter has been published as: 119. Faraji, M. and E.O. Voit, *Nonparametric dynamic modeling*. Math Biosci, 2016.

the opening of this membrane toward the synapse, the release of dopamine out of the vesicle and through the synaptic cleft to a receptor on the postsynaptic neuron, possible interactions with other neurotransmitters, and binding to the receptor. This binding in turn triggers a slew of additional mechanisms inside the signal receiving cell, including the complex process of signal interpretation which in the case of dopamine is often accomplished through multiple phosphorylation of the specific protein DARPP32, and the possible long-term adaptation to repeated stimuli [206-209]. Thus, a very coarse model could easily capture the fact that a signal moved from one neuron to another, but a detailed mechanistic model becomes quickly bogged down in the minutiae of the numerous intertwined biophysical processes that are involved in signal transduction.

Because elementary physical descriptions are often infeasible, the biological systems modeler is forced to resort to "higher-order" process representations, *ad hoc* models, suitable approximations, or combinations thereof. A good example is the Michaelis–Menten function of enzyme kinetics [108]. Its underlying concept is a process that postulates the reversible formation of a biochemical complex between an enzyme and its substrate and the subsequent release of the product of the reaction and of the enzyme, which is used over and over again. Under idealistic conditions *in vitro*, this concept is believed to be quite realistic. However, within living cells, the prerequisites for the involved mass-action functions are clearly not satisfied, and the so-called quasi-steady-state assumption, which is needed to formulate the process with a simple, explicit function, only holds under certain conditions. Thus, an idealized concept, formulated with the help of somewhat doubtful approximations, becomes a higher-order process representation for enzyme catalyzed reactions. Indeed, the Michaelis–Menten function performs well *in*

vitro and, in an approximate sense, presumably *in vivo*, although this is not really known. For simulations of large pathway systems, this function is often used as well, but its mathematical features become rather cumbersome, even for standard model assessments such as sensitivity analyses [210].

Notwithstanding these mathematical issues, it is common for the biological modeling community to base simulation studies in a variety of fields on a rather small set of functions, which are used time and again and prominently include mass-action, Michaelis–Menten and Hill functions, which often include regulatory terms [211]. The users of these functions rely on the argument that these functions suit their purposes—quasi as black boxes—and are sufficiently accurate if one considers the typical noise encountered in biological data. Furthermore, these particular functions at least have some foundation and rationale in biology, whereas the use of a function like a shifted arctangent has very little justification, except that its graph is s-shaped and therefore might resemble some saturation processes in biology.

True alternatives to these *ad hoc* approaches are generic approximations. Linearization, the simplest of these, has been enjoying enormous successes in engineering applications for many decades. For the representation of biological phenomena, by contrast, linear models tend to run into conflicts with the genuine nonlinearities that characterize living systems. For instance, common features like saturation, stable oscillations, threshold phenomena, synergisms, or chaos cannot directly be modeled with linear equations. A logical solution might seem to be the expansion of linear models to second-order Taylor approximations, but these become so awkward for larger systems [212] that very few modelers have resorted to this option. Instead, many biological

modeling groups have been using power-law approximations, which are nonlinear, but have linear characteristics in logarithmic space. Biochemical Systems Theory (BST; [122, 158, 213, 214]) and Metabolic Control Analysis (MCA; [82, 84, 215, 216]), which directly or indirectly utilize power-law representations [217, 218], respectively, have had success with analyses of a wide variety of complex biological systems (for a review, see [123]). Notwithstanding their successes, power-law representations are local approximations and therefore genuinely limited in their accuracy of capturing phenomena over large ranges of variation in the involved variables. As a case in point, univariate power-law functions in BST do not saturate for large substrate concentrations, and lin-log models, which are associated with MCA, become negative for small substrate concentrations and tend toward $-\infty$ for substrate concentrations approaching 0 [135-137]. As an alternative to these canonical power-law models, one could use sigmoidal basis functions, but for realistic models this option requires correspondingly larger numbers of parameters that need to be estimated [132, 219].

Even if reasonable guideposts could be found to justify the choice of appropriate model representations, the second step of model identification is still to be performed, namely the estimation of parameter values. For moderately sized or large models, this estimation is always challenging [220-222], due to noise in the data, non-convergence of the search algorithm and other problems, or because the wrong model was chosen after all. To make matters worse, even an excellent fit is not necessarily optimal, and the parameterized model may perform poorly in extrapolations, because the original fit was obscuring the compensation of errors among some terms within the model (*e.g.*, see [172,

185]). Furthermore, an excellent fit may be the result of overfitting with a model containing too many parameters.

These challenges and compromises lead to the obvious question of whether it might be possible to glean appropriate functions directly from experimental biological data, without presupposing potentially unjustified mathematical formats. The method of Dynamic Flux Estimation (DFE), which permits a relatively unbiased estimation of fluxes within a system and which will be reviewed later, took a first step toward answering this question affirmatively, at least for metabolic systems under ideal conditions [118]. Still, DFE requires some choices of model frameworks when the task is setting up a model from scratch.

Here, I describe a novel variant of DFE that makes such choices unnecessary, at least under favorable conditions. Given such conditions, the overall result of the proposed strategy is that it is possible to develop dynamic models in a *nonparametric* manner. Intriguingly, the resulting nonparametric models, which make no assumptions regarding parameter values or even mathematical formats, beyond the topology of the system, permit most of the typical diagnoses and analyses that are possible with a fully parametric model, which may be considered the gold standard in the field. As a consequence, simulations and other analyses can be performed without the complicated and often biased step of choosing models and parameterizing them, if suitable data are available. The data needed for this purpose consist of sets of time series that representatively capture the dynamics of a system under relevant inputs.

Both DFE and the nonparametric variant proposed here are particularly well suited for nonlinear, dynamic, regulated compartment models, because these possess the property of mass conservation, which imposes strong, unbiased constraints that greatly aid the formulation of appropriate models. As an illustration, and for ease of discussion, we will focus here on metabolic pathway systems, but it appears that other nonlinear compartment systems, such as SIR models of epidemiology and pharmacokinetic models, can be treated in the same manner.

6.2 Methods

6.2.1 Dynamic Flux Estimation (DFE)

The stoichiometric equation

$$\dot{\mathbf{X}} = \mathbf{S} \cdot \mathbf{V} \tag{6.1}$$

provides a generic description of the dynamics of a metabolic pathway system. This well-known equation collectively formulates dynamic changes in each metabolite of the system, $\frac{dX}{dt} = \dot{X}$, as a product between the stoichiometric matrix **S** and a vector of reactions or fluxes, **V**. This product formulation is remarkable, as it naturally separates the linear aspects of the system from its nonlinear features. Specifically, consider the situation where the slopes of all metabolites on the left-hand sides are known for some given time point. If so, Equation 6.1 is a system of linear algebraic equations, where each variable V_j represents the state of a flux at this time point, rather than a metabolite. The

nonlinear features enter the system secondarily, by virtue of the fact that each component of the flux vector is a possibly complicated function of metabolites and regulators, and therefore of time. Dynamic Flux Estimation (DFE) makes maximal use of this separation of the model into linear and nonlinear components.

In typical analyses, such as Flux Balance Analysis, the stoichiometric Equation 6.1 is studied at a steady state of the system [75-77], where the vector on the left-hand side contains zeros. DFE reaches beyond the steady state, by addressing the system at many time points of a system's trajectory, where the vector of derivatives is different from zero. In its first phase, DFE uses time series measurements of metabolite concentrations, X_1, …, X_n, along with estimates of the slopes of these time courses. Thus, DFE evaluates equations of the type

$$\frac{dX_i}{dt} \quad at \quad t_k = Slope \quad of \quad X_i \quad at \quad t_k = \sum_{j=1}^{k} s_{i,j} V_j(t_k) \tag{6.2}$$

where the slopes are numerical values, estimated from the data. Since data are typically noisy or incomplete, it is advisable to apply one of various available preprocessing, smoothing and data substitution techniques (*e.g.*, [180, 181, 183, 184]). We do not discuss these further here, because data smoothing techniques and the methods proposed here constitute clearly distinct steps within the model design procedure. Thus, we will assume in the following that the data had been successfully smoothed.

The slope substitution is performed for *m* time points, with the result that each differential equation in (6.1) is replaced with a set of *m* linear algebraic equations at these time points. Collectively, these equations may be formulated as a matrix equation, where

the variables are the fluxes V_j, rather than the metabolites [192-195]. If this matrix has full rank, the solution is unique, and if the equations are overdetermined, the best-fitting solution is computed via linear regression. In the most common case, the system is underdetermined. We will briefly skip this case here, but return to it later. The result of solving the linear equations is a set of numerical flux values at each time point. Collecting these sets for all time points it is straightforward to create a plot of each flux as a function of time. It is furthermore possible to plot each flux against the system variables upon which it depends. In the case of a single substrate and no regulation, the plot is a simple line graph. By the same token, if the flux depends on two or more variables, the result is a line on the manifold that is given by the unknown flux function in three or more dimensions. This first phase of DFE typically requires knowledge of the connectivity of the system. However, it is to some degree possible to infer formerly unknown reaction steps and regulatory signals [223, 224] (Figure 6.1).

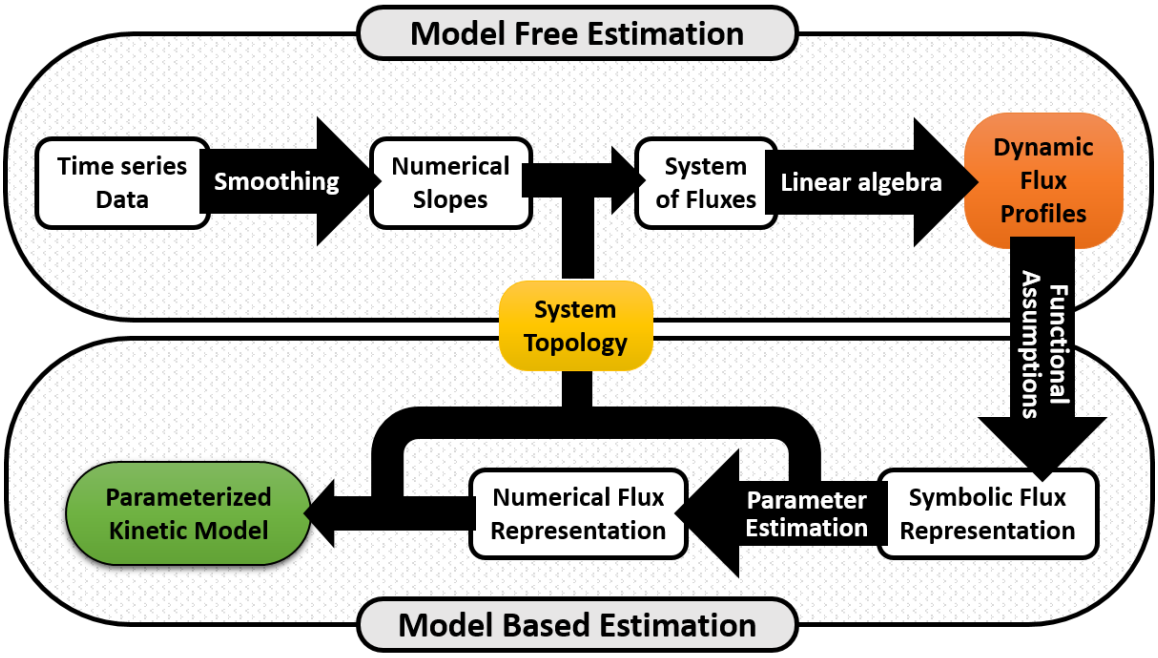

Figure 6.1 Dynamic Flux Estimation (DFE). The method consists of two phases. Phase 1 (top) is model free in a sense that only the stoichiometry is assumed to be known, whereas functional forms of the process representations are not. The procedure in this phase is based on raw experimental time series data in the form of metabolite concentrations. It is beneficial to smooth these data with some numerical algorithm, such as a spline. Next, the rate of change in each metabolite is obtained from the smoothed time courses. These numerical slopes correspond to the values on the left-hand side of Equation 6.2, so that numerical evaluation of the slopes at *m* time points converts each differential equation of the model into a set of *m* linear algebraic equations, in which the flux states are the driving variables. The system is solved, potentially with the aid of additional constraints, and the result is a time dependent, numerical profile of all fluxes. In Phase 2 (bottom), the dynamic flux profiles are fitted with appropriate functions or rate laws, and the result is a fully parameterized dynamic model.

In Phase 2 of DFE, the numerical flux representations are converted into mathematical functions. For this purpose, assumptions must be made regarding the functional format of each flux. In some cases, the shape of the flux profile or independent biological considerations may suggest a mathematical format, but this is not guaranteed (Figure 6.2). Once a format has been chosen, the parameters of each flux representation are to be

estimated from the available data, as it is typical for any other modeling effort. This estimation is much simpler than for the entire system, because it is performed for a single explicit function at a time, rather than simultaneously for all fluxes in the system of ODEs. The result of the two phases combined is a fully parameterized model of the pathway system.

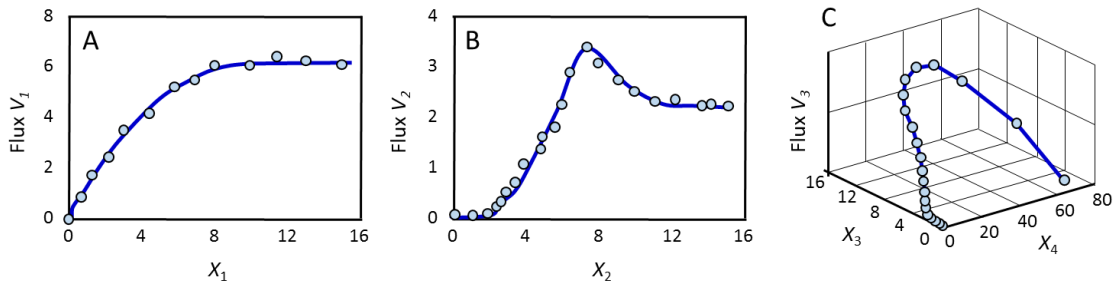

Figure 6.2 Typical results of Phase 1 of DFE. The flux in panel A may be representable with a Michaelis–Menten function. By contrast, it might be difficult to choose appropriate formats for the fluxes in panels B and C.

It is not directly possible to solve the linear system when the stoichiometric matrix is underdetermined. Unfortunately, this is actually the most common case for metabolic systems, which typically contain more reactions than metabolite pools. In the case of an underdetermined system, the stoichiometric equation admits infinitely many solutions, and these can differ tremendously, even if the system is small, with some having monotonic shapes, while others may overshoot or exhibit oscillations [139]. To study these sets of solutions, it is advisable to reduce the system mathematically to a system whose dimension

equals the degrees of freedom [139]; for instance, in a system with six metabolites and eight reactions, the dimension will typically be two.

A solution to the challenge of under-determination is the Moore–Penrose pseudoinverse [187-189]. While effective, the pseudoinverse usually contains negative values, which are not consistent with biological fluxes. Several other approaches have been proposed. First, characterizability analysis, based on the pseudoinverse, shows directly where additional information is needed about the system, or which metabolite pools could be merged, to make the equations uniquely solvable [186]. Second, it is sometimes possible to measure influxes or effluxes experimentally or to infer them from the data [223, 224]. Third, it might be feasible to obtain additional biological information regarding some of the internal fluxes. For instance, one might be able to deduce internal fluxes from biochemical features of the involved metabolites [141]. Fourth, one may use the dynamic concentration profile associated with one metabolite to estimate flux functions for its influx and efflux [140]. This procedure requires assumptions regarding the functional forms of these fluxes, but if the variable does not change much in magnitude, an approximate representation is likely to be sufficient. Fifth, sufficiently many datasets may allow the inference of a few flux profiles directly from the data [138]. Finally, it might be possible to utilize biological constraints, such as energy minimization, to reduce the degrees of freedom within the system [139].

If independent information regarding a flux can be found, the system of equations becomes simpler. For instance, suppose that flux V_3 is known in sufficient detail. Then, Equation 6.2 becomes

$$\frac{dX_i}{dt} - s_{i,3} \cdot V_3 = \sum_{j=1, j\neq 3}^{k} s_{i,j} V_j \tag{6.3}$$

where the left-hand side is numerically known and the number of free degrees is typically reduced by 1.

6.2.2 Concepts of Nonparametric Dynamic Modeling

6.2.2.1 Overview

The core idea of nonparametric modeling is to forgo Phase 2 of DFE and instead to replace the functional representations of the processes in the system with a library of numerical results directly obtained from Phase 1 of DFE. Of course, nothing comes for free: The proposed substitution requires good, comprehensive data. While such datasets may currently be scarce, the decreasing cost of generating rich datasets renders the proposed method increasingly more appealing.

Thus, let us suppose that rich data are available, where "richness" refers to more or less complete datasets and enough time series to represent the phenomenon under investigation appropriately. For instance, one could imagine sets of time series experiments with many combinations of different input and inhibitor concentrations. Collectively these datasets form the scaffold for the proposed modeling strategy. One should note that even a rich dataset rarely produces complete coverage of the imaginable metabolic profile space. One reason is that most metabolic systems have a single non-trivial steady state and that trajectories even from a variety of initial values tend to approach this steady state, so that many regions of the mathematically imaginable solution space are only scarcely covered or not at all. While this situation may seem to be undesirable, it actually reflects those

portions of the metabolic profile space that are most relevant and reliably represented, while ignoring situations that may be biologically less relevant. Figure 6.3 shows a generic example where the trajectories approach a common steady state.

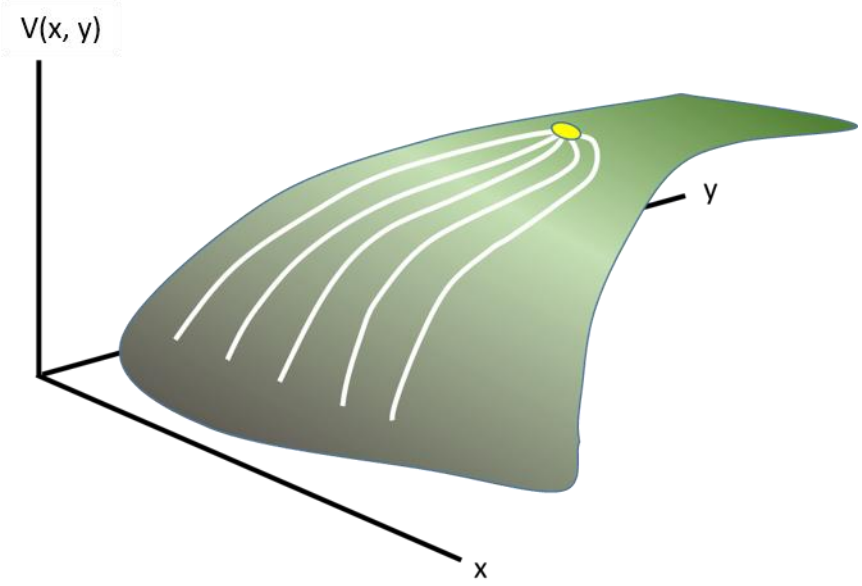

Figure 6.3 Illustrative example for scaffolding the library with different datasets. The green surface is the graph of the unknown flux function $V(x, y)$. Each white line represents the result of a time series experiment. All time series approach the same steady state (yellow) so that many regions of the surface, such as the top right, are not supported by measurements and may be biologically irrelevant. Thus, the set of time series results may be sufficient to interpolate between the data (white lines) but does not allow reliable extrapolations far beyond the data.

For metabolic pathway systems and many other compartment models, the topology of the underlying network of components is typically known, or assumed to be known. It is also often, although not always, known which metabolites affect each flux as substrates or modulators. Nonparametric modeling takes advantage of this situation. It directly

uses Equation 6.2, where the stoichiometric coefficients are assumed to be given and where the flux states are the dependent variables on the right-hand sides. In contrast to parametric modeling, the explicit functions on the right-hand sides of the differential equations are here replaced with numerical flux descriptions. Thus, no mathematical formalism is used to represent the fluxes and hence no parameterization is needed. Instead, the numerical flux-substrate profiles are generated directly from the smoothed experimental data processed in Phase 1 of DFE and recorded in a library that is subsequently used for looking up values required by the ODE solver. As a result, the library, combined with an interpolation algorithm, constitutes the primary simulation tool for analyses with the nonparametric model (Figure 6.4). Expressed differently, the processes of making functional assumptions and of parameterizing these functions are replaced by the use of a scaffolding library and geometric interpolations, once dynamic flux profiles have been obtained. The details of setting up the library and of the interpolation algorithm are discussed in the next sections.

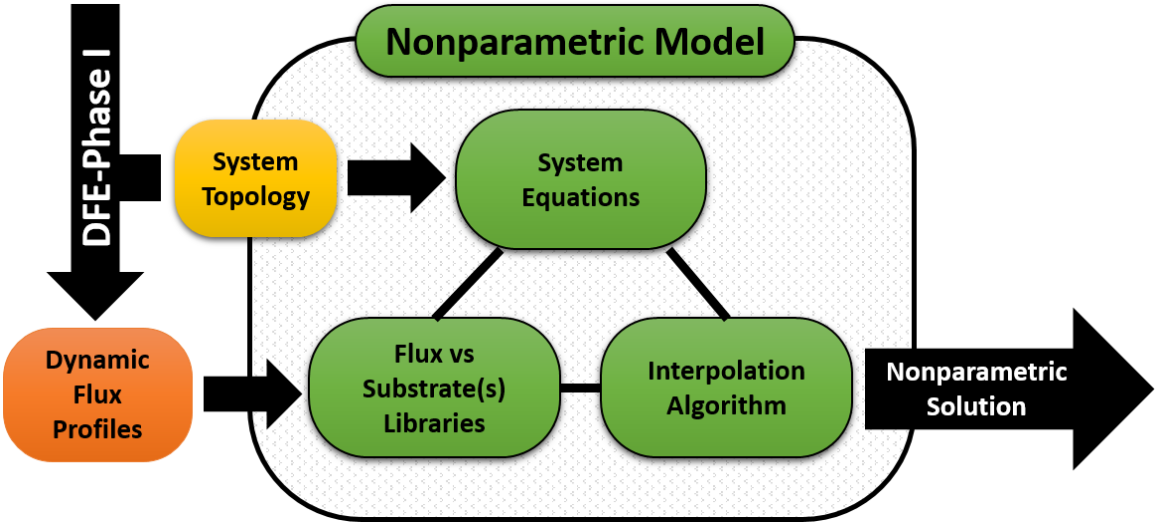

Figure 6.4 Nonparametric modeling framework. Functional assumptions and the parameterization of fluxes are circumvented in nonparametric modeling. Instead, the flux values needed at each step of a system simulation are provided by look-up tables that are generated from the measured dynamic metabolite profiles and the inferred flux profiles. They are stored in a library for call-up.

6.2.2.2 Library Construction

Phase 1 of DFE results in dynamic flux profiles that can be displayed against time or against the systems variables that affect the flux. Combining the smoothed metabolic time series and flux profiles, flux-substrate relationships are generated and recorded as arrays in which the columns contain values of substrates and regulatory agents that affect each flux, as well as the corresponding flux value itself. The rows represent snapshots of the state of the system at various time points and possibly from different experiments (Figure 6.5). To ensure sufficient scaffolding of the flux-substrate subspace, each array should ideally include the dynamic profile of a flux for several experimental settings, for instance, with different initial conditions.

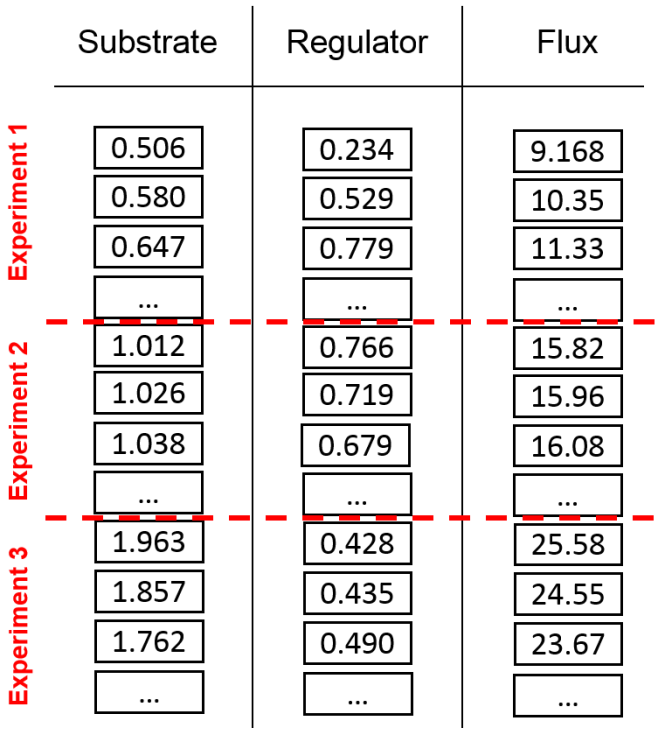

Figure 6.5 Generic format of the flux library. The (here artificial) data are arranged in arrays where the columns represent a specific flux along with all substrate(s) and regulator(s) that affect this flux directly. The rows represent snapshots of the state of the system for different time points and experiments. The specific numerical values shown here refer to the later example of a fermentation pathway but are not of import here.

Of course, the quality of further analyses and simulations depends on the quality of the library, which in turn depends on the quality of the available data. It is difficult to quantify how many data have to be available, as this quantity depends on the complexity of the flux functions. For example, oscillatory data will often require more data points than smooth monotonic data for a good representation of the true time trend. Thus, one can only pose a vague criterion that must be satisfied: the data must be representative of the trends they quantify.

In theory, every metabolite could be involved in every flux of the system, thereby causing an unmanageable combinatorial explosion for larger systems. In practice, this situation does not arise, and most metabolites are only affected by a few other metabolites.

6.2.2.3 Interpolation Algorithm

To be useful, the nonparametric model must permit simulations of the evolution of the modeled system over time. The model is formally specified as a set of ODEs whose right-hand sides are defined by fluxes whose values are stored in the libraries described before. By its nature, each ODE solver needs to have flux values available for essentially arbitrary substrate and modulator amounts within reasonable ranges, which are, for instance, bounded by the experimental datasets. In the case of nonparametric modeling, the library consists of a discrete set of values, which covers only a finite subset of all values that are possibly needed. To overcome this issue, a method is needed that allows the ODE solver quickly to estimate each required flux value from the data in the library. Obviously, the quality of this type of gap-filling is directly correlated with the density and quality of the data in the library.

A convenient tool for this task is the function *scatteredInterpolant* in MATLAB (version R2014a, The MathWorks, Natick, MA), which is applied to the smoothed data obtained from DFE. This function generates 2D or 3D interpolations from the dataset in the library in an unbiased manner. Specifically, the interpolant passes through the original data and uses adjustable methods such as linear or nearest-neighbor interpolation, which enable the algorithm to estimate flux values for arbitrary substrate and modulator values during the ODE simulation.

For our demonstration of the concepts of nonparametric modeling with fluxes of two or three arguments, we used the linear interpolation method. For fluxes with a single substrate and no regulators, the function *interp1* was preferred, as it is more efficient for one-dimensional interpolations. Figure 6.6 presents some output from this interpolation algorithm. Panel (a) exhibits data points for a flux V that depends on two substrates S_1 and S_2. Panel (b) visualizes the interpolated surface over a grid of inputs.

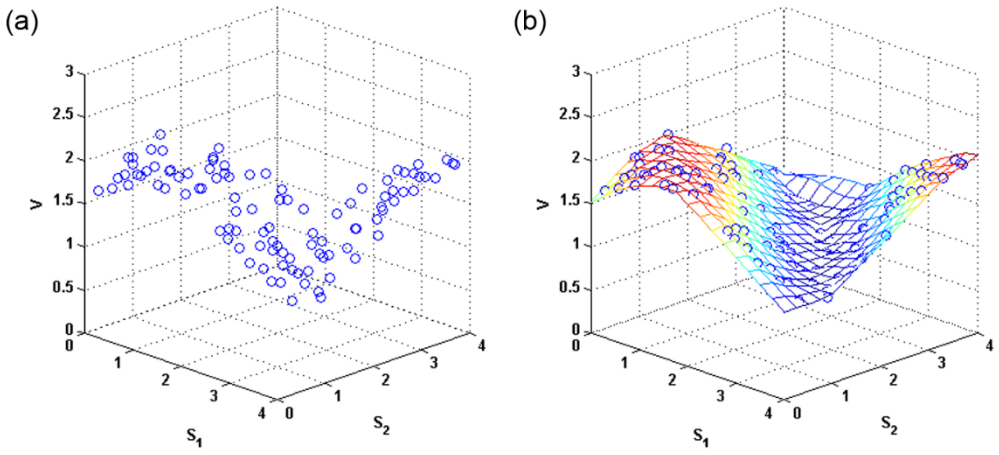

Figure 6.6 Interpolation of a flux-versus-substrate surface for two substrates. The algorithm uses an interpolant function that efficiently connects the data points. Panel (a) shows the raw data points, while panel (b) shows the interpolated surface.

Preliminary studies indicate that the interpolation, under favorable conditions, can generate sufficiently accurate libraries even for data that had not been smoothed. Indeed, skipping the smoothing step may lead to means of assessing the effects of intrinsic noise in the data. For instance, single data points or subsets could be removed from the library in Monte-Carlo simulations, thereby ultimately yielding distributions of slightly different

trajectories and steady states. This option will be addressed elsewhere. It also remains to be more formally investigated under what conditions the chosen interpolation algorithm is optimal for the tasks discussed here or whether different, specifically customized interpolation methods might perform better.

6.2.2.4 Expanded Data Coverage by Moderate Extrapolation

Parametric models may typically be extrapolated without bounds. However, one must ask to what degree such extrapolations beyond the domain of experimental data are really justified. In our situation of nonparametric modeling, the interpolating function in MATLAB is able to extrapolate datasets to some degree. However, the accuracy of extrapolation is only guaranteed for relatively small ranges in the vicinity of the scaffolding data subspace. While the lack of far-reaching extrapolations may be seen as a disadvantage over parametric models, it also provides a healthy warning against extrapolations that are not supported by data and might even be biologically infeasible. We used the nearest-neighbor variant of the interpolation technique in MATLAB for extrapolations. An example will be shown later.

6.2.2.5 Library-based Simulations

The flow of a typical simulation with a nonparametric model is illustrated in Figure 6.7. Starting from an arbitrary initial state, within or sufficiently near the recorded flux-substrate subspace, the library, together with the interpolation algorithm, provides for each set of metabolites the appropriate flux values V_i. The flux values are then used to compute

changes in the substrates, $\frac{d\mathbf{X}_\tau}{dt}$, using Equation 6.1. These changes are utilized to compute the state of the system at the next time point; the base concept of this procedure is shown in Equation 6.4.

$$\mathbf{X}_{\tau+\Delta t} \approx \mathbf{X}_\tau + \frac{d\mathbf{X}_\tau}{dt} \cdot \Delta t \tag{6.4}$$

where $\frac{d\mathbf{X}_\tau}{dt}$ represents the collection of appropriate flux values at time τ. While Equation 6.4 demonstrates the solution procedure with Euler's forward method, modern ODE solvers employ more sophisticated methods, some of which are expansions of Euler's method. Specifically, the standard ODE solver in MATLAB (version R2014a, The MathWorks, Natick, MA) uses the Runge–Kutta method with variable step size. In contrast to supplying the ODE solver with the parametric functions on the right-hand sides of the ODEs, as it is commonly done, we specify the appropriate interpolants from the library. This substitution does not affect the computation speed much.

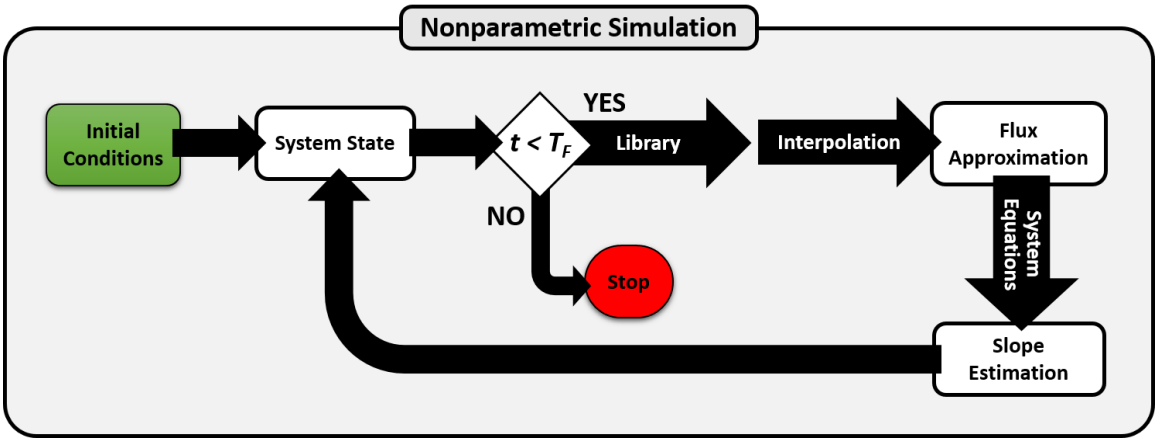

Figure 6.7 Flowchart for a typical nonparametric simulation. Once initial conditions are specified, the algorithm extracts from the library the flux values that correspond to the metabolite profile at the initial time point. From these values, the algorithm computes the slopes of all variables and moves the simulation to the next time step, according to Equation 6.4, until the desired final time point T_F is reached.

6.2.3 Typical Model Analyses

Analyses with the nonparametric model are primarily based on simulations. The simplest of these are changes in initial values of one or more of the dependent variables. Similarly easy to assess are changes in independent variables, which are often used to model constant quantities like an enzyme activity or a fixed input. As an example, suppose that some experimental technique raises the activity of an enzyme, which results in a concomitant 20% increase in the corresponding flux by. An analogous parametric situation would be a 20% raise in the V_{max} of a Michaelis–Menten or Hill rate law. In the nonparametric case, such an alteration is simply implemented by increasing all pertinent flux values in the library by 20%. The analogous strategy holds for an external inhibitor, as long as it decreases one or more fluxes in a multiplicative manner. A change in K_m corresponds to a different scale for the substrate concentration. For example, if K_m is to be doubled, an

appropriate solution is to double the substrate concentrations in the library without changing the corresponding flux values. To see the rationale for this strategy, consider the standard Michaelis–Menten rate law, $V = \frac{V_{max}S}{K_m+S}$. If K_m is doubled, the same functional values of V are achieved if each value of S is doubled.

To some degree, it is even possible to perform analyses that at first seem to require functional forms. For instance, once a steady state has been identified, one may use the plots of each flux against each contributing variable and determine the slope(s) of the flux at the steady-state metabolite concentration. Entering all slopes into an appropriately laid-out array yields an approximate Jacobian matrix, which may then be used for numerical analyses characterizing the model behavior close to the steady state and, in particular, the stability of the steady state. We will discuss this option, as well as sensitivity analysis, within the context of the case study in the *Results* section.

Of course, the nonparametric model cannot entirely replace its parametric analog. For instance, it seems difficult to perform formal investigations of bifurcations and other features of complex dynamics, at least in a direct manner.

6.3 Results

6.3.1 Case Study: Nonparametric Modeling of the Fermentation Pathway in Yeast

To illustrate the nonparametric modeling capabilities without being encumbered by the idiosyncrasies of experimental datasets, we created artificial "data" from a model in the literature. This model describes in a simplified manner the anaerobic fermentation pathway

in the baker's yeast *Saccharomyces cerevisiae*. The pathway is comparatively well understood, and a considerable body of *in vivo* measurements of metabolites and fluxes at various steady-states is available [225]. This model was originally proposed by Galazzo and Bailey [226] and subsequently converted into a power-law model by Curto *et al*. [199, 200, 227]*;* it has been used on numerous occasions to demonstrate new modeling and optimization techniques [159, 228-232].

For the illustration here, we use Curto's version of the model to generate datasets, which under typical conditions would have been obtained experimentally in the laboratory. We analyze the data without noise. Thus, we pretend that metabolic time courses had been measured and used in DFE to reveal plots of all fluxes *versus* time or *versus* the system variables that affect them. The system contains only a rather small number of metabolites and fluxes, which renders it a good candidate for illustration purposes.

The model captures the dynamics of the fermentation pathway from glucose uptake to ethanol yield (Figure 6.8). The pathway has essentially a linear structure, although two minor pathways branch off. It is regulated through negative feedback from glucose-6-phosphate (G6P) on glucose uptake and feedforward activation of the enzyme pyruvate kinase by fructose-1,6-bisphosphate (FBP).

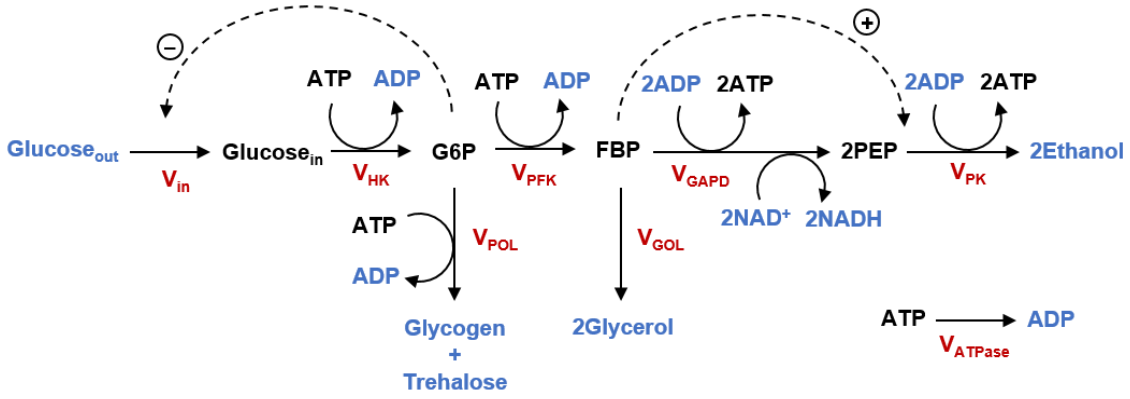

Figure 6.8 Anaerobic fermentation pathway in *Saccharomyces cerevisiae.* The model of the pathway contains five dependent variables and eight fluxes. Adapted from Curto *et al*. [199, 200, 227].

The independent variables and fluxes in the model (see Figure 6.8) are as follows:

X_1 : Glucose (Glc)
X_2 : Glucose – 6 – phosphate (G6P)
X_3 : Fructose – 1,6 – bisphosphate (FBP)
X_4 : Phosphoenolpyruvate (PEP)
X_5 : ATP

$V_1 : V_{in}$ $V_5 : V_{GAPD}$
$V_2 : V_{HK}$ $V_6 : V_{GOL}$
$V_3 : V_{PFK}$ $V_7 : V_{PK}$
$V_4 : V_{POL}$ $V_8 : V_{ATPase}$

Figure 6.9 shows the model scheme in a simplified fashion; although regulation is not explicitly shown, it is taken into account by the model. One notes that flux V_5 splits a 6-carbon molecule into two 3-carbon molecules, which causes a doubled rate of influx into the pools of X_4 and X_5, and mandates corresponding elements of 2 in the stoichiometric matrix (Equation 6.5).

$$\mathbf{S} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & -1 & 0 \\ 0 & -1 & -1 & -1 & 2 & 0 & 1 & -1 \end{bmatrix} \tag{6.5}$$

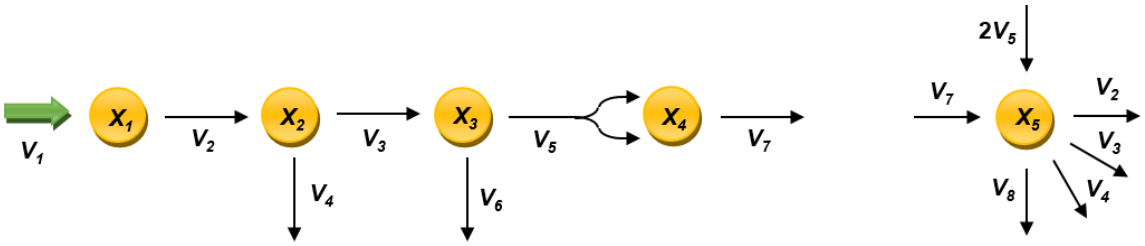

Figure 6.9 Simplified model scheme of the fermentation pathway in Figure 6.8.

The stoichiometric equation of the system, $\dot{\mathbf{X}} = \mathbf{SV}$, is equivalent to the set of differential equations in Equation 6.6.

$$\begin{aligned} \dot{X}_1 &= V_1 - V_2 \\ \dot{X}_2 &= V_2 - V_3 - V_4 \\ \dot{X}_3 &= V_3 - V_5 - V_6 \\ \dot{X}_4 &= 2V_5 - V_7 \\ \dot{X}_5 &= V_7 + 2V_5 - V_2 - V_3 - V_4 - V_8 \end{aligned} \tag{6.6}$$

The model has a stable steady state with concentration values, in [mM], Glc = 0.03456, G6P = 1.011, FBP = 9.188, PEP = 0.009532, and ATP = 1.128 [199].

6.3.1.1 Library for the Fermentation Model

In an actual systems analysis, experimental time series data would be used to populate the library. Instead, we use the Curto model and solve it multiple times, every time starting from different initial states of all metabolites. These states are located in a hypercube in $\mathbb{R}^5$ that corresponds to the five substrates represented in the pathway model. Of course, experimental data would be noisy. For clarity, we consider noise-free or well-smoothed data.

The collective result consists of time series data of the substrate concentrations and corresponding fluxes, which in reality would have been obtained from DFE. In addition to time plots, the data allow us to establish flux-*versus*-substrate trajectories, whose values are recorded in arrays, as it was illustrated in Figure 6.5. Figure 6.10 shows examples of trajectories for the fermentation model. In all figures, the metabolite concentrations are given in millimolar (mM), and all fluxes are given in millimolar per unit of time (mM/min).

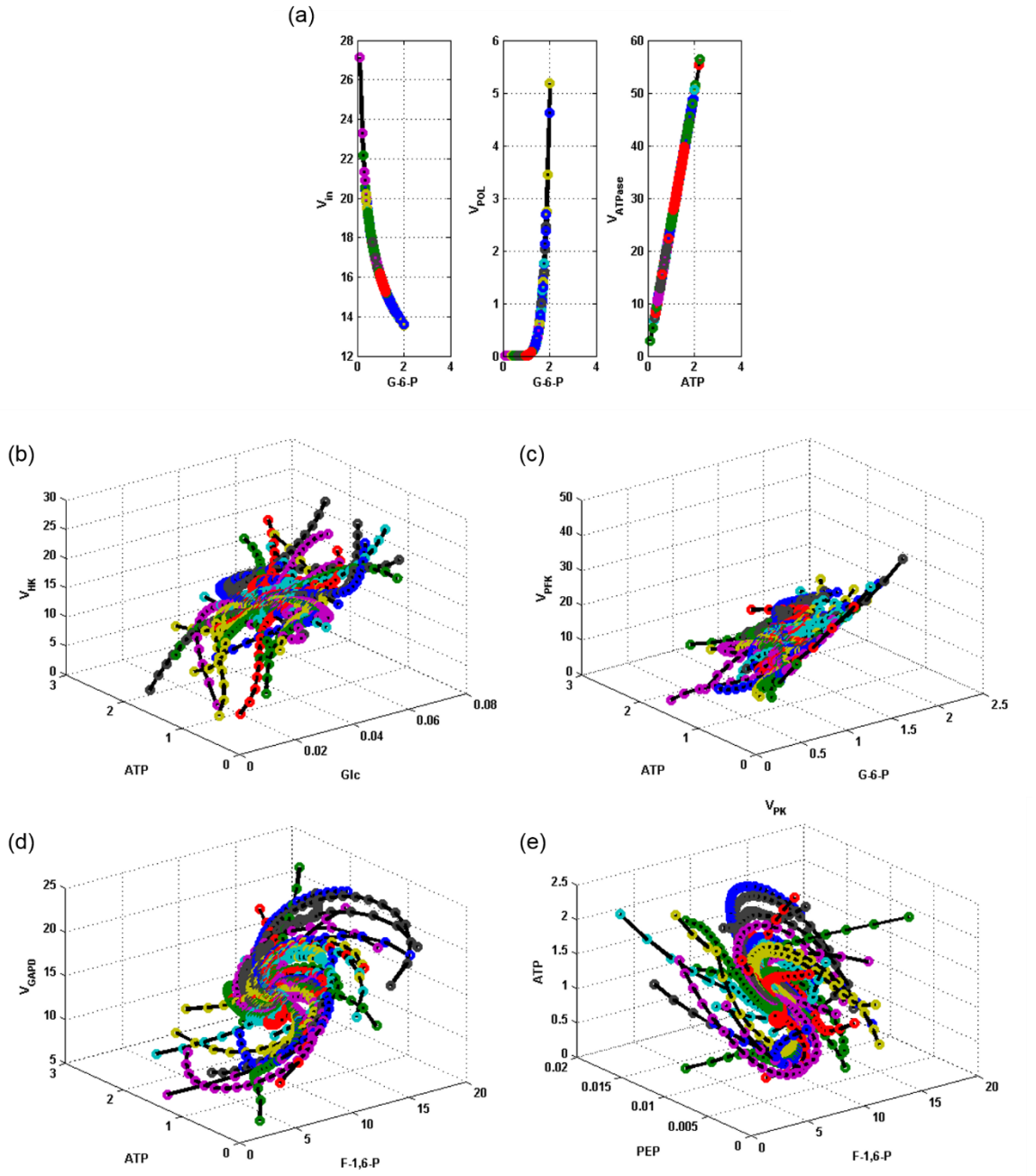

Figure 6.10 Flux-substrate profiles corresponding to different initial conditions. The dynamics of the system forms trajectories that are recorded and used in the form of look-up tables that substitute for explicit mathematical representations of the fluxes. The panels represent the different fluxes. For each panel, metabolite trajectories from different experiments are shown in different colors. Panel (a) shows V_{in}, V_{POL} and V_{ATPase}, which are single-substrate fluxes. Panels (b), (c) and (d) exhibit V_{HK}, V_{PFK} and V_{GAPD} respectively. These three fluxes have two substrates each. Lastly, panel (e) shows substrate trajectories for V_{PK}. This flux has three substrates so that the flux-substrate trajectory is four-

dimensional. For this reason, only the substrate trajectory is shown. According to Curto *et al.* [227], the flux V_{GOL} is proportional to V_{PK} with a proportionality constant of about 0.03 and therefore not shown.

6.3.1.2 Dynamic Simulations

Once the library is set up from the artificial data, the nonparametric model is ready to use. Figure 6.11 confirms with simulation results that the nonparametric model returns essentially the same results for three distinct initial conditions as a parametric simulation with Curto's model. For comparison, the nonparametric results (lines) are superimposed on the "data" (parametric results).

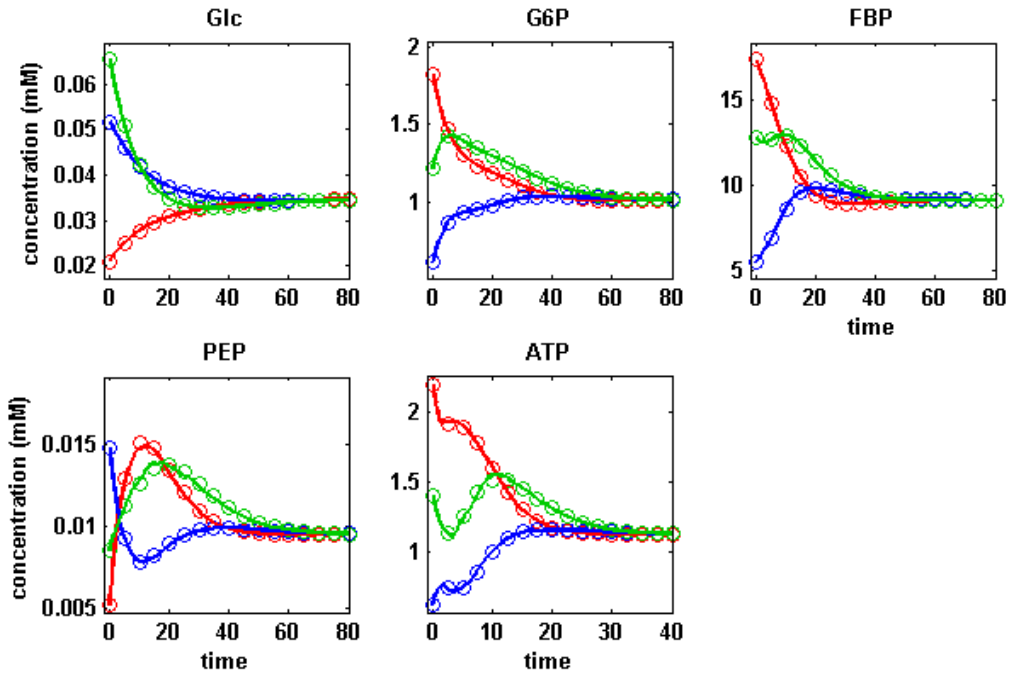

Figure 6.11 Nonparametric and parametric simulations of the fermentation pathway. Three different initial conditions were used; the results are shown in different colors. All simulations converge to the same steady state: Glc: 0.03456, G6P: 1.011, FBP:

9.188, PEP: 0.009532, ATP: 1.128. The initial conditions are: green: [0.0657 1.213 12.86 0.0086 1.410]; blue: [0.0518 0.6066 5.513 0.0148 0.6203]; red: [0.0207 1.820 17.46 0.0052 2.199]. The nonparametric simulation results (solid line) match the artificial data (circles) very closely.

6.3.1.3 Jacobian at the Steady-state

The library permits the estimation of the Jacobian of the system. Namely, once a steady-state metabolite profile has been determined, one estimates the slopes of the fluxes at this profile with respect to the various variables. The slopes of each flux, $\frac{dV}{dX_i}$, are computed numerically, using the flux values at the steady-state, $\mathbf{V}(X_i)$, and the flux values at points in the vicinity of the steady-state, $\mathbf{V}(X_i + \Delta X_i)$. Using Equation 6.1, the flux slopes together yield $\frac{d\dot{X}}{dX_i}$. The quantities $\frac{d\dot{X}}{dX_i}$ of all system variables constitute the elements of the Jacobian matrix. The most prominent use of the Jacobian is the determination of eigenvalues for local stability analysis. This analysis is directly comparable with the corresponding analysis of a linearized parametric model.

For the fermentation pathway, the eigenvalues of the nonparametric model at the steady-state are [−1734.0, −333.5, − 14.04 ± 7.581i, −1.635]. For comparison, the corresponding analysis for the parametric model yields quite similar eigenvalues: [−1689.0, −341.1, −13.66 ± 7.853i, −1.843].

6.3.1.4 Sensitivity Analysis

An important component of modeling is sensitivity and gain analysis. In the former case, a parameter is slightly altered and corresponding changes in critical system features are recorded. In the latter case, an independent variable is altered. Most prominent is the effect of changes in parameters or independent variables on the steady-state values of the system. While the nonparametric model obviously has no parameters, it is still possible to study sensitivities if one of the fluxes is either raised or decreased by a small percentage, which mimics a situation where the activity of an enzyme is altered. The results of such an analysis are shown in Figure 6.12. As one would expect, the system reaches a new steady-state, which depends on the specific flux alteration. Comparisons with Curto's model demonstrate that the nonparametric model matches the corresponding analytical results from the parametric model very well.

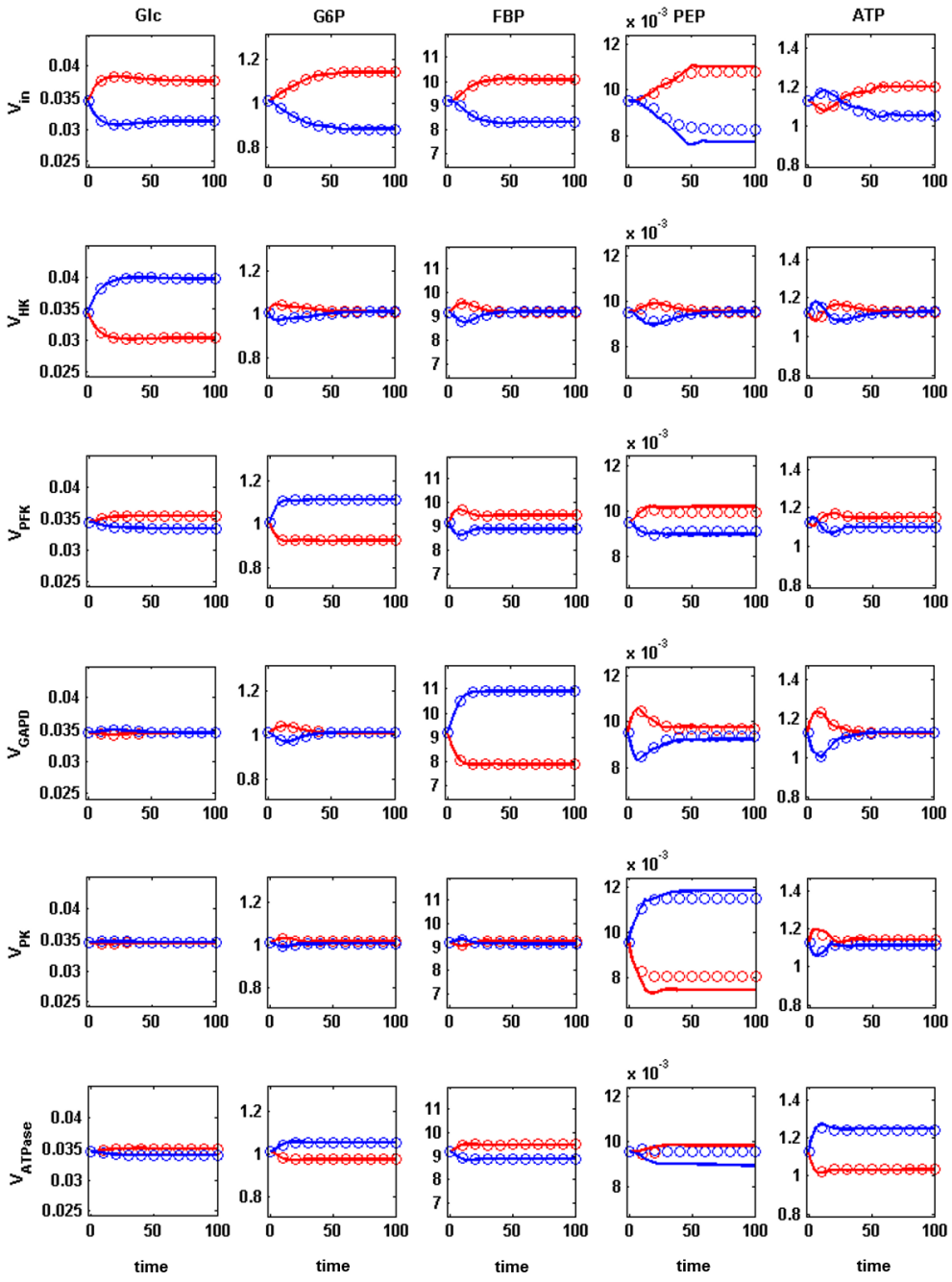

Figure 6.12 Sensitivity analysis. Each of the fluxes was separately perturbed by ±10% for the entire duration of each simulation. With each change in a flux, the simulation starts from the original steady state of the system and moves toward the new steady state. For all the flux perturbations other than V_{POL} and V_{GOL}(not shown here because they are insensitive), the system converges to new steady-states. The nonparametric model results (solid lines) closely match the parametric results (circles). One should note different scales for the five variables.

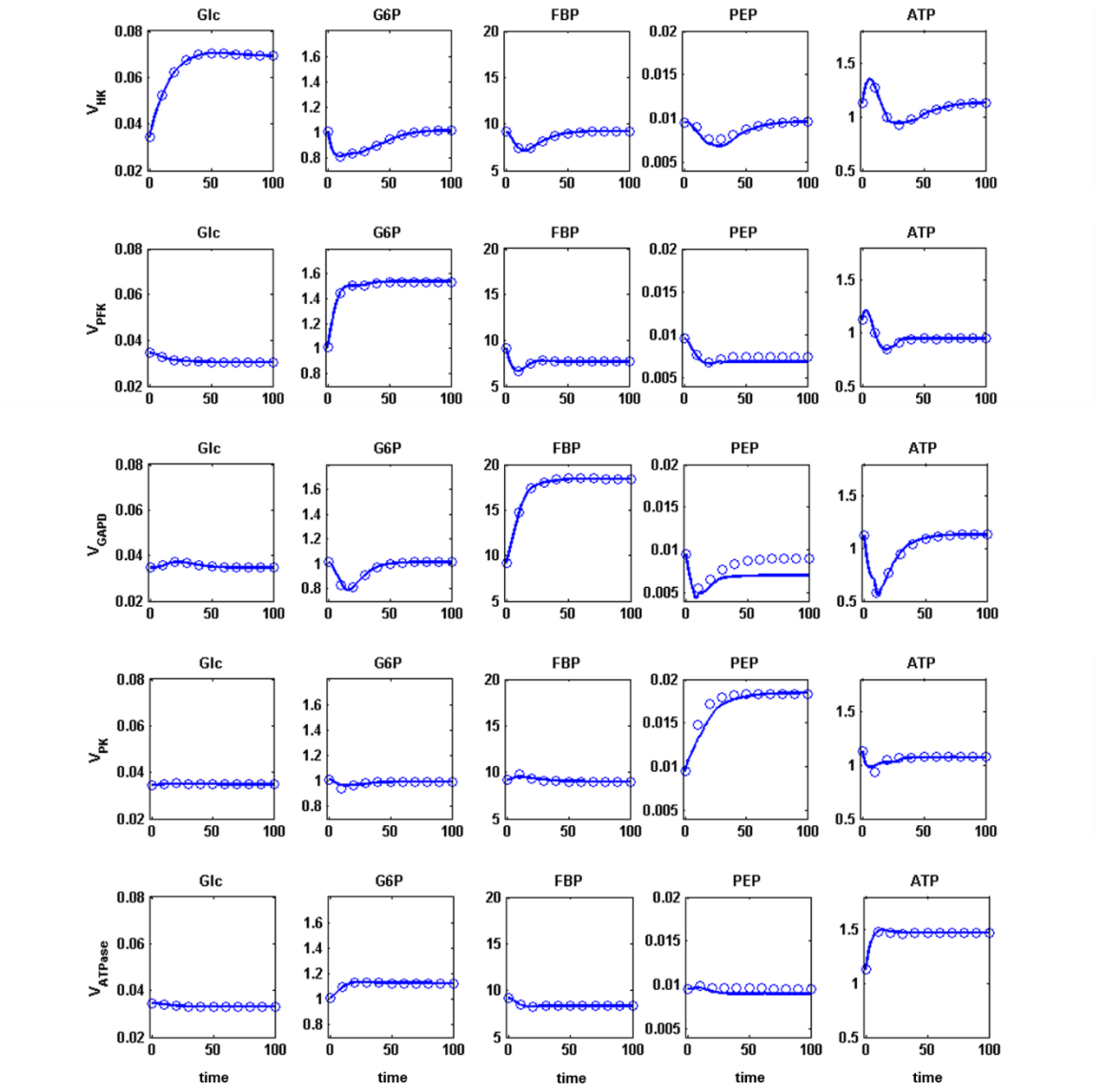

Figure 6.13 Trajectories and steady-states of the system for enzymes with altered K_m. The system dynamics was replaced such that fluxes depict enzymes with increased K_m values. Each row represents simulation results for a flux catalyzed by an enzyme with an altered K_m, where the corresponding substrates respectively from top are: [Glc, G6P, FBP, PEP, ATP]. Each flux was modified to mimic a two-fold increase in K_m. Only V_{ATPase} was simulated for a 30% increase, because larger perturbations lead to instabilities in the parametric and nonparametric models. V_{POL} and V_{GOL} did not show any noticeable changes in steady-states and hence are not shown in here. Solid lines show the nonparametric results, while circles represent parametric results from Curto's model.

Figure 6.13 exhibits simulation results representing the case where the K_m of an enzyme is altered. To mimic an increase in a K_m, more substrate is needed to compensate for the same flux magnitude. As was explained earlier, for example, doubling the K_m of a reaction corresponds to replacing the substrate concentration S with $S/2$ in the library. As can be seen, the steady-states take new values. Specifically, the substrate of the enzyme with altered K_m shows an increased steady-state value in each row. Again, the nonparametric and parametric model results agree quite closely.

6.3.2 Data Collection from a Bolus Experiment

6.3.2.1 Model Set-up

To test the nonparametric approach under more realistic circumstances, we used the Curto model to simulate an *in vivo* nuclear magnetic resonance experiment, which typically uses a bolus input rather than a constant substrate influx (for a pertinent example, see [198]).

In the original Curto model, the input flux of the system, V_1, is constant and set to simulate a saturated flux that corresponds to an exterior glucose concentration in excess. To model the transient dynamics of a bolus experiment, we converted external glucose into a dependent variable X_0, thereby expanding the system by one ODE:

$$\dot{X}_0 = -0.01V_1 \tag{6.7}$$

The coefficient 0.01 reflects the compensation between the concentration within the large exterior volume of the medium and the much smaller interior volume of the cells. This coefficient quantifies how much a change in the external glucose concentration (*e.g.*, from

4 mM to 3.99 mM) affects the internal glucose concentration (from 0 to 1 mM). For the uptake function, we use a Michaelis–Menten function with noncompetitive inhibition, where the inhibitor, G6P, reduces the overall activity of the enzyme. This choice was based on the original article which considered the inhibition as not competitive, because G6P binds to the glucose transporter inside the cell while glucose binds on the outside [225]. Thus, we have

$$V_1 = \frac{V_{\max} X_0}{K_M + X_0}\left(1 + \frac{X_2}{K_I}\right)^{-1}, \tag{6.8}$$

with K_m = 4 mM [233]. $V_{\max}$ and K_I were retrieved from Curto's model as about 16 and 394, respectively. With the exception of these two adaptations, the modified model is the same as in Equation 6.6 and the pathway diagram is the same as in Figure 6.9, where the input arrow now originates at X_0.

Commensurate with the K_m value for V_1, we used the modified Curto model to perform experiments with three different bolus amounts, where the initial values corresponded to a concentration of half the K_m, twice the K_m, and close to saturation, respectively. The results are shown in Figure 6.14. Panel (a) corresponds to a bolus of exterior glucose resulting in an external initial concentration of 2 mM. In panel (b) the initial concentration is 8 mM, and in panel (c) it is 60 mM. In each panel, the left plot shows the dynamic metabolite profiles, while the right plot shows the dynamic flux profiles, which in a real experiment would have been inferred per DFE. In all cases, the input V_{in} initially causes the interior glucose concentration to rise, and this rise migrates throughout the pathway. The plots on the right side show how V_{in} and other fluxes start to

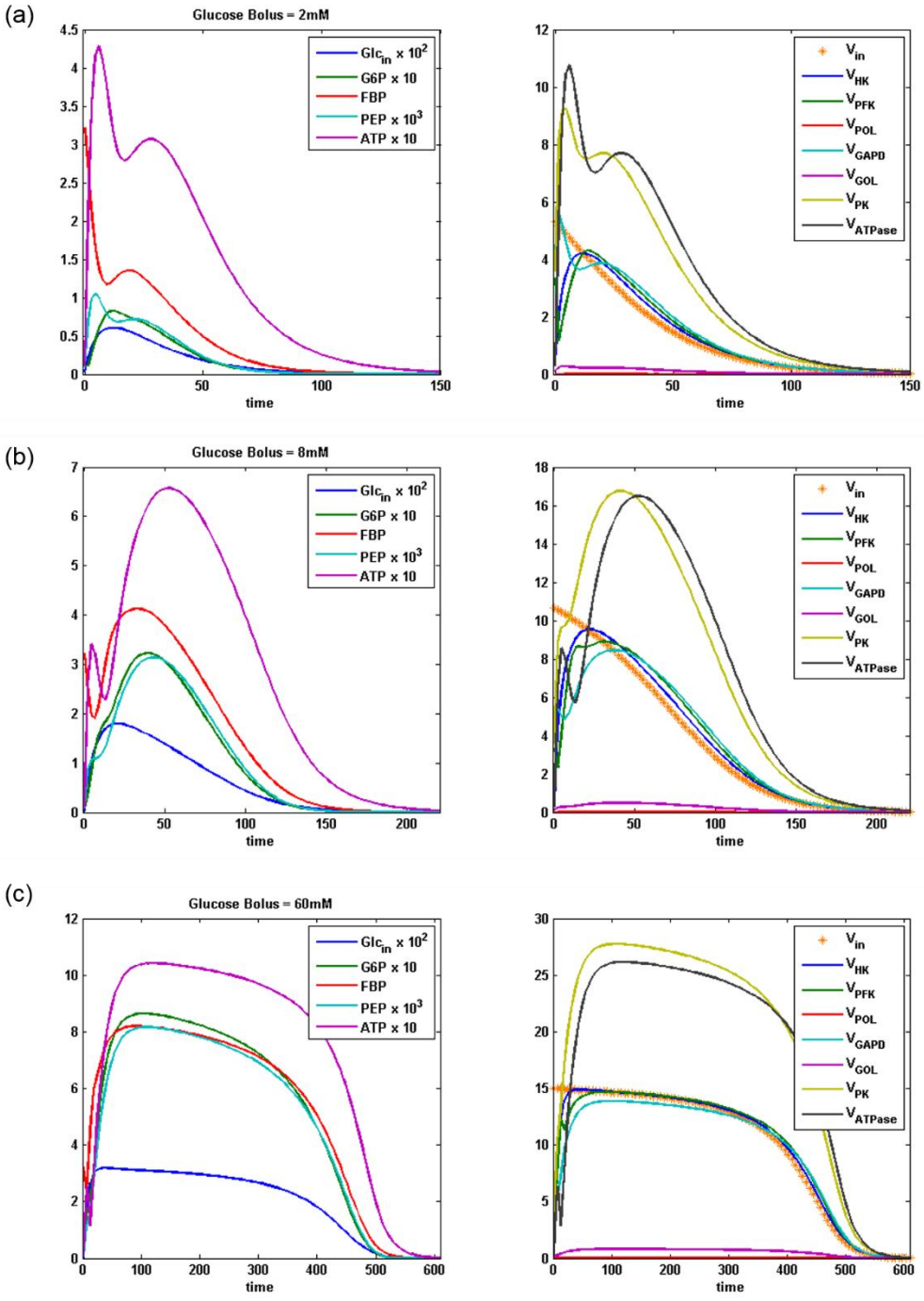

Figure 6.14 Simulation results for bolus experiments with different external glucose concentrations. The time courses capture the transient behavior of the system very clearly. The left column exhibits concentration profiles, while the right column shows flux profiles. Amount of glucose bolus: Panel a: 2 mM; panel b: 8 mM; panel c: 60 mM.

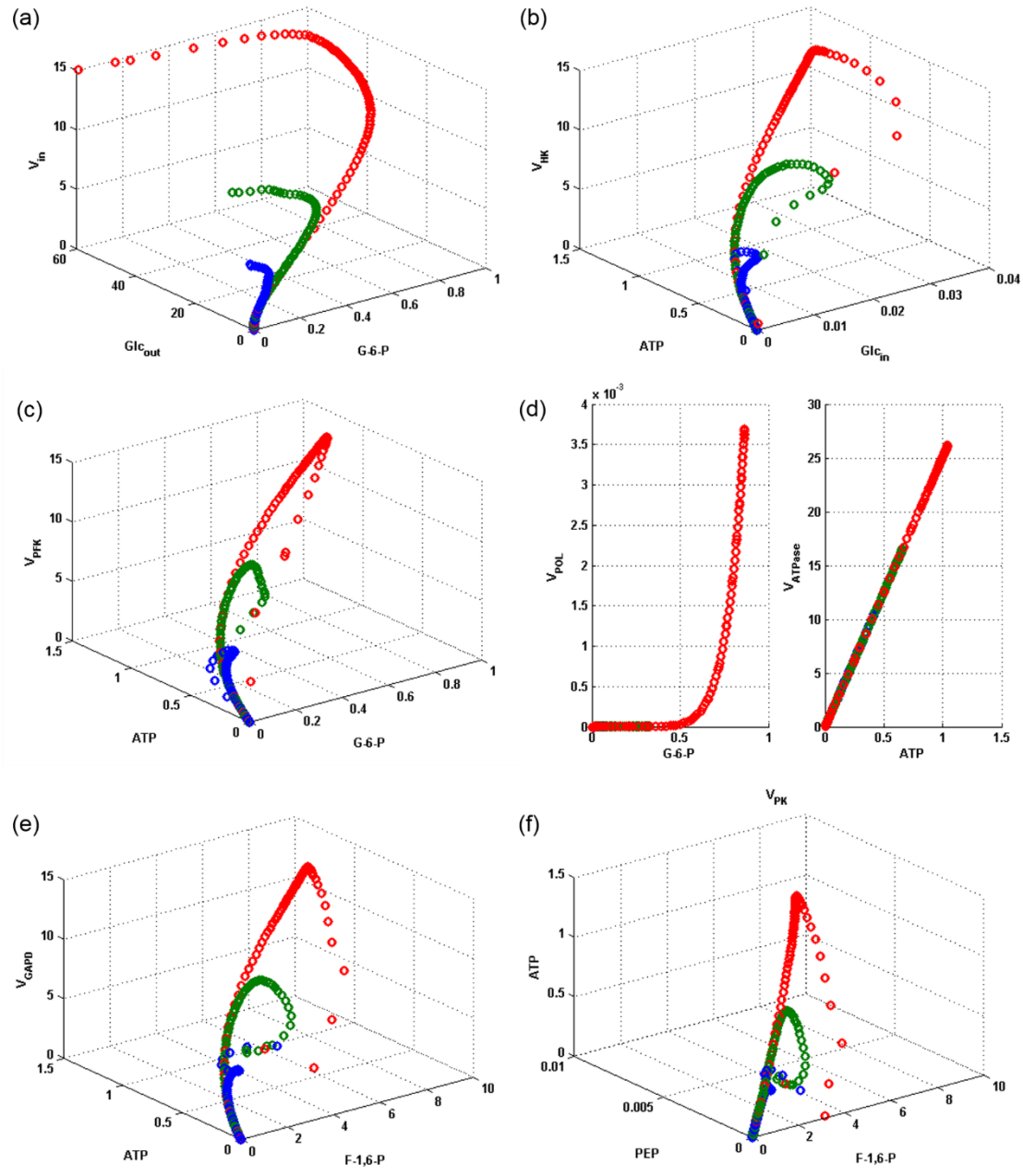

Figure 6.15 Flux-*versus*-substrate trajectories retrieved from the bolus experiment results. The three experiments lead to different trajectories, which are shown in different colors.

decline once the exterior glucose becomes depleted. Correspondingly, each metabolite accumulates until the efflux from its pool becomes larger than the influx to the pool, due to the decline in exterior glucose. Eventually all metabolites become consumed and approach zero concentration.

The three experiments allow us to plot flux-*versus*-substrate trajectories (Figure 6.15). Panels (a), (b), (c) and (e) correspond to V_{in}, V_{HK}, V_{PFK} and V_{GAPD} respectively. Panel (d) shows the trajectories for V_{POL} and V_{ATPase}, both of which are single-substrate fluxes with two-dimensional trajectories. Naturally, the trajectories converge since each flux value is a function of a substrate and the substrates approach steady state. In a bolus experiment, however, the system only converges to a trivial steady-state. Panel (f) only shows the substrate trajectories since the flux-*versus*-substrate trajectories for V_{PK} are entities in a four-dimensional space.

6.3.2.2 Library Construction

We took the simulation results of the three bolus experiments as artificial metabolite and flux data and used them to construct the library. The MATLAB function *scatteredInterpolant* (Figure 6.6) provided a fast and effective interpolation for the data. Two examples of resulting surfaces are shown for V_{in} and V_{HK} in Figure 6.16(a) and (b), respectively. Panels (c) and (d) show the flux-*versus*-substrate surfaces with moderate extrapolations beyond the boundaries of the dataset.

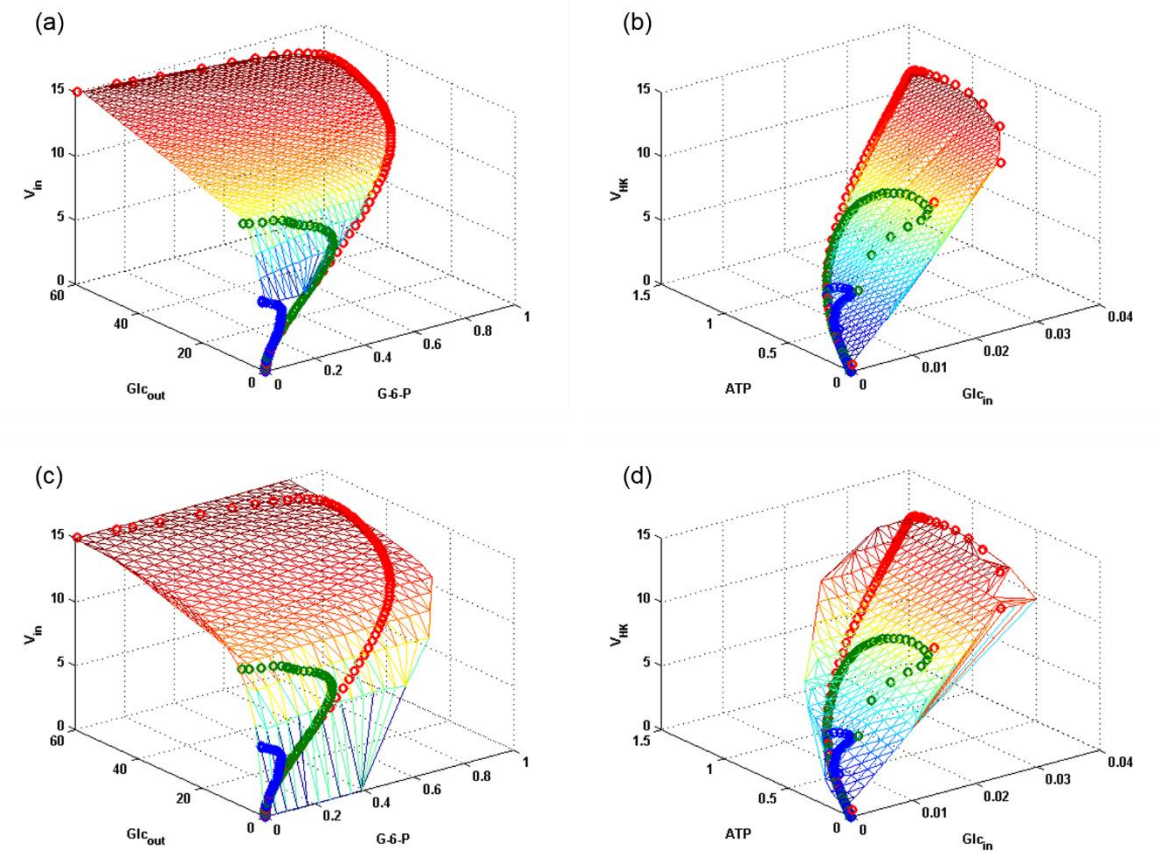

Figure 6.16 Interpolation and extrapolation of flux-versus-substrate surfaces. Panels (a) and (b) exhibit how the interpolations provide approximations for desired values not included in the library. The algorithm can also provide approximate values slightly outside the dataset boundaries. The quality of this extrapolation often declines with the distance from the boundary (panels (c) and (d)).

6.3.2.3 Steady-state Simulation

We simulated the system for a constant concentration of exterior glucose within the ranges recorded in the library. This experiment is fundamentally different from the bolus experiments that were used to stock the library. The results are shown in Figure 6.17 for one of the fixed glucose inputs. Three different sets of initial conditions for the dependent variables were simulated, and the system converged to the same steady-state for all three.

This steady state is essentially identical to that of Curto's model if the same glucose input is simulated.

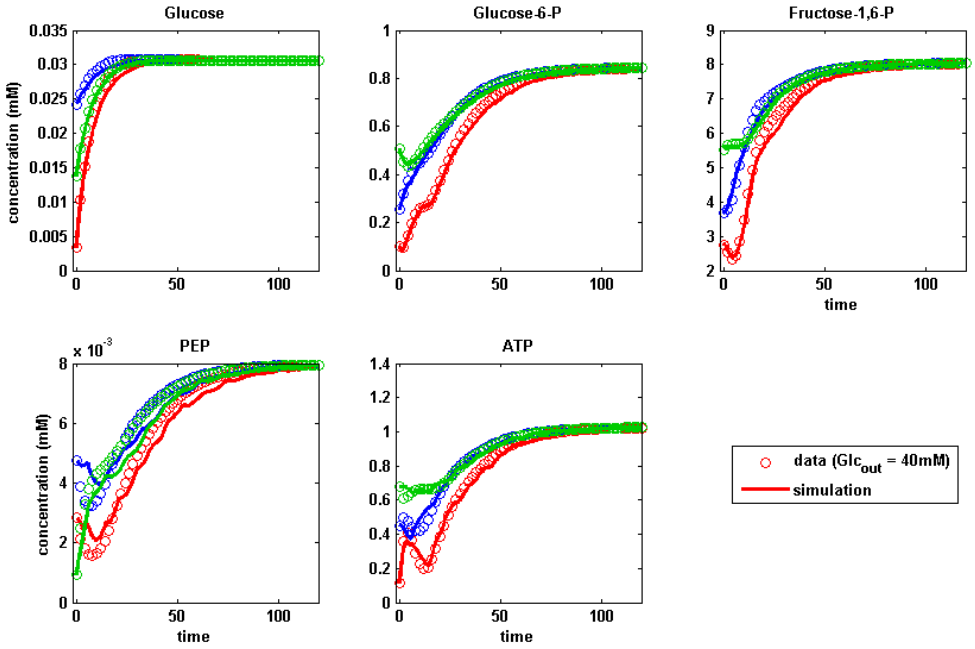

Figure 6.17 Steady-state simulation results based on data from bolus experiments. Although the data in the library were obtained from bolus experiments, the nonparametric simulation results for constant glucose inputs are very similar to their parametric analogues. Three different initial conditions were chosen, and the system was simulated for a constant exterior glucose concentration of 40 mM.

6.3.2.4 Computation of a Steady-state

The computation of steady states is of great import, as most biological systems operate within the vicinity of a normal homeostatic state. For the case of parametric models, some structures, including linear systems, S-systems [121] and Lotka–Volterra

systems (*e.g.*, [234]), permit the algebraic calculation of steady-state solutions, while most other nonlinear systems require search algorithms.

Interestingly, the nonparametric model permits not only simulations toward a stable steady state, but even the option of direct computational assessments of steady states. In fact, the flux-*versus*-substrate library appears to be as effective as an explicit functional form. Here, we applied the MATLAB function *fsolve* to the task at hand. Similar to the traditional case involving explicit ODEs, *fsolve* starts from a given initial condition and at each iteration estimates the system derivatives using system equations and the library. These are used to determine the direction and size of the next step toward optimizing the objective function. It turned out that *fsolve* is quite sensitive to initial guesses. We therefore randomly sampled 10,000 initial conditions from the space of metabolite concentrations in the library, ran *fsolve*, and recorded the solutions that achieved zero derivatives within a reasonable tolerance. Figure 6.18 compares the computed steady-states (blue circles) with the true steady-state from the parametric model (yellow circles). The left panel exhibits the metabolites, while the right panel shows the corresponding steady-state flux values.

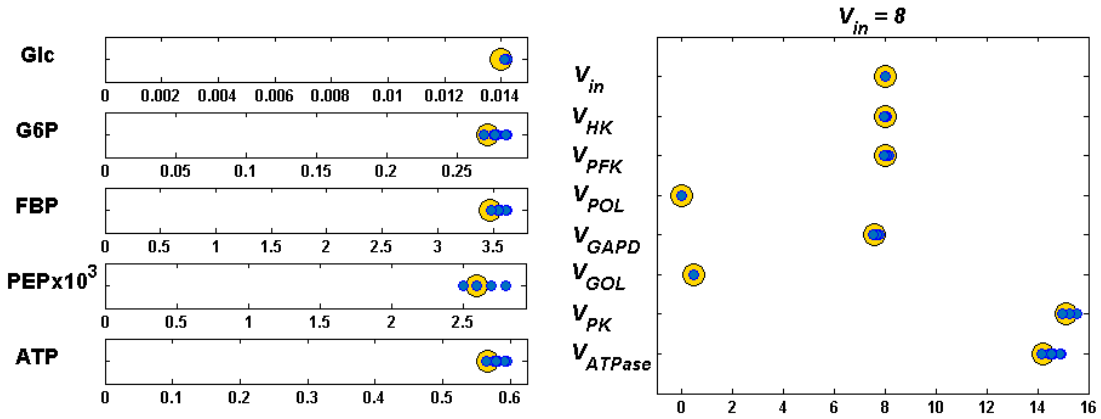

Figure 6.18 Steady-state metabolite and flux values. The yellow circles show the true solutions for the constant input V_{in} = 8, according to Curto's model, while the blue dots exhibit the ensemble of computed nonparametric solutions, which resulted from different initial values in *fsolve*.

6.4 Discussion and Conclusions

The term "nonparametric" is currently used almost exclusively within the context of statistics, where nonparametric tests are sometimes preferred to their parametric analogues because they require fewer assumptions and are often simpler to execute, although they are not always as discerning as parametric tests [235]. It is impossible to date the first use of nonparametric statistical methods, as standard features like histograms and sample means do not require *a priori* choices of parameters and have been used for a very long time. The terminology itself was apparently introduced by Wolfowitz in 1942 to characterize those methods that did not require specialized assumptions regarding the functional forms of the distributions characterizing the populations from which samples were analyzed [236, 237]. As an alternative to "nonparametric," some authors proposed the term "distribution-free,"

while Ury suggested “assumption-freer” or ISD (incompletely specified distribution) statistics, openly admitting, however, that neither term was “especially felicitous” [238].

Notwithstanding the terminology, the statistics community at the time was of course not unanimously welcoming and initially considered nonparametric methods as “short cuts for well-established parametric methods” and later as “rough and ready (quick and dirty), inefficient methods, … that were wasteful of information” [237, 239]. In today's view, the advantages and drawbacks appear to have found a healthy equilibrium, and nonparametric methods are considered true alternatives to more traditional parametric methods, because they make less stringent demands on the data and sometimes reveal quick, although possibly coarser answers with a lesser amount of calculation. It is even possible that nonparametric methods utilize more information in large datasets, especially if the data do not stem from processes that have well-parameterized representations. Then again, by their very nature, nonparametric methods do not involve parameters that permit succinct descriptions, and they sometimes throw away information to a point where, for instance, differences between two populations can no longer be quantified [240]. It is also openly acknowledged that the interpretation of nonparametric statistical methods is sometimes difficult [241]. Overall, nonparametric methods are recognized as sometimes useful and maybe necessary, and on occasion superior, but certainly not perfect.

Similar to nonparametric models in statistics, the nonparametric dynamic models proposed here do not sweepingly solve all problems in nonlinear compartment modeling. In fact, one could say that they have similar advantages and drawbacks. They use fewer assumptions regarding mathematical formats, which immediately implies that these models cannot be described or compared succinctly with a set of parameter values. Similarly, they

may be more difficult to interpret. They do not throw away information, however. For instance, if an appropriate format of a flux with correct kinetic parameter values is known from some outside source, this information can easily be used to reduce the degrees of freedom or to replace a unique analytical solution with a regression solution, which one might expect to be more robust. At any rate, the method and its solutions depend on the quantity and quality of available data, thus yielding a situation that is not genuinely different from parametric models.

Similar to parametric models that are obtained with DFE, the models here require solvability of the linear system of fluxes in the ODE model. This solvability issue can be assessed using the topology of the pathway through characterizability analysis, even in the absence of specific data [186]. Non-characterizable pathways incur estimation issues no matter what methods are applied, because such pathways admit entire spaces of solutions, some of which differ drastically from others [139]. In this situation, the DFE approach offers the advantage of identifying whether any, and if so which, fluxes can be determined uniquely, or what additional information would remedy the situation. For instance, Iwata and colleagues [140] demonstrated that it is possible to estimate a few fluxes with traditional methods and thereby to make a DFE solution unique. Similarly, it was shown that reasonable biological constraints may be able to ameliorate or even solve the under-determinedness of a pathway system [139]. The same arguments apply to our nonparametric approach, although the infusion of additional information may result in a hybrid model where some fluxes are parameterized and others are not.

Not long ago, the experimental characterization of metabolic time courses was considered a very difficult task, and metabolic time series data—and corresponding

models— were extremely rare [64]. This situation has changed dramatically during the past decade, and various methods of molecular biology have rendered it feasible to generate time series data with reasonable effort (*e.g.*, [201-204]). This type of data generation is crucial here, because the construction of nonparametric models depends on a data scaffold, from which a library of nonparametric flux representations can be established.

In contrast to metabolite concentrations, fluxes of a pathway system are presently difficult to measure, except possibly for influxes and effluxes that communicate with the exterior milieu. However, in analogy to the progress in time series measurements of concentrations, one might hope that flux measurements will follow in due time; *e.g.*, see [99]. Indeed, if it becomes more widely known that a limited set of flux measurements may lead to a complete pathway characterization that does not make mathematical assumptions beyond the stoichiometry of the system, the community of molecular biologists may devote concerted effort toward such measurements in the future.

In addition to the difficulties of obtaining the right data, one could argue that experimental data often cover only a relatively small subspace of what an explicit mathematical function would be able to represent. This argument is certainly true, but one must ask to what degree extrapolations with these functions far outside the data domain are justifiable and biologically relevant.

Once experimental concentration and flux measurements become more readily available, the nonparametric method proposed here may actually become preferable over the tried and true approaches of parametric pathway modeling, at least for initial, unbiased assessments and various types of simulation studies. For deeper analyses, such as formal

bifurcation studies, it appears that parametric representations will be difficult to replace. In this case, DFE may be executed to the end, where parametric functions are selected based on the inferred flux profiles.

CHAPTER VII

Conclusions

7.1 Conclusions

This dissertation contributes to the field of plant metabolic modeling at three levels. **First**, the literature review in this dissertation presents an extensive review of state-of-the-art in the field of plant metabolic modeling and models of lignin biosynthesis, and provides a concise platform for future modelers and investigators. **Second**, novel computational models of lignin biosynthesis in switchgrass and *Brachypodium* give insights into the control mechanisms, topological and spatial characteristics, and most importantly the necessity of species specific modeling. **And lastly**, at the level of methodology, two novel methods ("Stepwise inference of likely dynamic flux distributions from metabolic time series data" and "Nonparametric dynamic modeling") provide powerful techniques to tackle a number of arguably the most challenging steps in metabolic modeling, *i.e.*, the choice of an objective function, of a mathematical formalism, and the parameterization of the latter.

Investigations in this dissertation reveal that the different pathway models for lignin synthesis in a number of plant species [2, 3, 18, 57-59] show commonality, but also differences in regulatory and spatial features. One could thus come to the conclusion that

every species manages its lignin production differently. However, the fact that some distinctive features of one model are not part of the other models should not be over-interpreted, at least not quite yet. It is well possible that the pathway in alfalfa and switchgrass is as compartmentalized as the one in *Brachypodium*, but the dictum of simplicity in modeling, and the data that these models were based upon, suggested that specific compartments were not needed to match the data in these species. The same is true for other apparently distinguishing features, such as product inhibition, which we found necessary in switchgrass, and which may well be in effect in other species. These features were needed to make the models consistent with specific data typesets, and if one re-analyzed the models with other types of data types, the same features could well be suggested for other species. As it stands, the collective experimental database is sparse, and the published models are minimalistic special cases of the same "master model," which can even account for the fact that some species seem to be missing certain pathway metabolites or enzymatic reactions. It remains to be the subject of further data generation and analysis to determine whether these differences disappear toward one common, complex model, whether they are immaterial byproducts of evolution that did not exert strong selective pressure, or whether they evolved for reasons that are germane to these species and their environments. The cooperation between experimentalists and modelers has led to early successes. Some of these are narrowly focused by explaining observations that had been puzzling before. We described some of these in the context of lignin synthesis and recalcitrance. To study some of the *in vivo* complexity in an *in vitro* system, the lignifying cell suspension cultures reported in several species (*Arabidopsis* [242], poplar [243], switchgrass [166]) could be useful for modeling purposes of the lignin pathway.

These systems can be studied along a time course when lignin deposition and cell differentiation occur, allowing the evaluation of different parameters such as pH or temperature and the use of dynamic models that could be proposed as potential *in vivo* validation systems.

7.2 Future Work

Some potential directions and needs for future research in computational metabolic modeling, especially in plants, were already presented in Chapter II. This section proposes some additional extensions of models and describes further information that will be required to achieve a fuller understanding of the lignin pathway. This section also describes improvements of the novel modeling methods presented in Chapters V and VI that appear to be beneficial for future applications.

Lignin Pathway

- As it was emphasized throughout this thesis, different plant species exhibit different pathway structures, regulatory mechanisms, and spatial characterizations in their lignin biosynthesis. From the taxonomical point of view, and considering the economy of experimental research, it seems that the only viable approach to recognizing distinct lignin pathway features in monocots and dicots is to collect additional data from the species investigated here, and also from other species, and

to subject them to similar computational analyses. It appears that labeling experiments would very well complement gene knock-down studies.

- The lignin pathway is usually initiated from a ubiquitous amino acid like phenylalanine, and in the case of *Brachypodium* from phenylalanine and tyrosine. Hence, the lignin pathway is closely tied to the metabolism of these two amino acids, as well as other amino acids such as alanine, glutamine and asparagine. Therefore, one might expect that a comprehensive understanding of the lignin pathway and its systemic responses to gene alterations would be tremendously improved by extending the model scope into amino acid metabolism of at least phenylalanine and tyrosine, if not beyond.
- To predict lignin content and composition in transgenic *Brachypodium* strains, a dynamic model is needed, as a static model like the one presented in Chapter V is not able to simulate dynamic adjustments in metabolite and flux steady-states in a perturbed system.

Method Development

- The methods of "Nonparametric dynamic modeling" presented in Chapter V and "Stepwise inference of likely dynamic flux distributions from metabolic time series data" presented in Chapter VI, are both developed based on the assumption that experimental techniques and devices provide us with high-quality metabolic time series data. Even if the availability of time series data increases, however, and even if the data are of high quality, some data regarding less-studied pathway metabolites

will always be missing. Therefore, effective inference methods for estimating missing metabolite profiles from the available data would substantially extend the current nonparametric and stepwise inference methods.

APPENDIX A

SUPPLEMENTARY MATERIAL FOR CHAPTER III

1.1 Supporting Information

A.1.1 Analysis of the Lignin Pathway with Inclusion of Caffeyl Aldehyde.

The lignin pathway in switchgrass is still not entirely understood. In particular, some species synthesize caffeyl aldehyde, but this does not seem to be the case in switchgrass. The account for this intermediate metabolite complicates the structural analysis, which is discussed in the following, because many more topological pathway configurations are possible.

Recent enzyme kinetic experiments suggest that a reaction catalyzed by CCR1 with caffeoyl CoA as substrate is not likely to occur in switchgrass. The K_{cat}/K_M values are up to 10 times higher for feruloyl CoA compared to caffeoyl CoA [176]. However, because the *in vivo* concentration of caffeoyl CoA is unknown, the reaction could be plausible. The reason is that a higher steady-state concentration of caffeoyl CoA would compensate for the lower K_{cat}/K_M ratio, and a new metabolite, caffeyl aldehyde, would have to become a component of the pathway.

The most significant change in comparison with the results of the main text is that the COMT/F5H complex could constitute a second functional channel. As a consequence, the number of candidates for topological configurations jumps to 19, and large-scale simulations confirm that 12 among these are compatible with experimental data. Between the 12, four configuration are significantly more abundant in parameter space. Interestingly, all four include at least one channel. Results for the pathway containing caffeyl aldehyde are shown in Figures A1- A5. They correspond to the main result figures in the text. In addition, all theoretically admissible pathway configurations that include caffeyl aldehyde can be mapped according to their total numbers of reaction steps (Fig. S6). It is interesting to note that in this type of representation the compatible configurations are closely clustered, considering that the order of configurations in the second and third row is more or less arbitrary.

A.1.2 Principal Component Analysis

In order to characterize the parameter space of the system, we performed principal component analysis (PCA) on the parameter sets that had been filtered by the model criteria (Figure S1). Principal Components 1 through 4 collectively account for 88% of the variance. The blue circles show the parameters from the original set of simulations, while the red circles represent the set of parameters that was secondarily generated from the blue points using PCA. Finally, the green area is a subset of the red area; it contains the parameter values that successfully passed all model criteria. The straight edges of the red and green areas are due to the imposed biological constraints on the parameter values. Note that the generated area is not a convex hull, that is, the smallest space containing all possible

linear combinations of the blue points, but a multi-dimensional "hyperdomain" expanded along the principal directions. Each edge of this hyperdomain is elongated by 20% of the original length in order to cover the principal direction up to 10% off the original blue point.

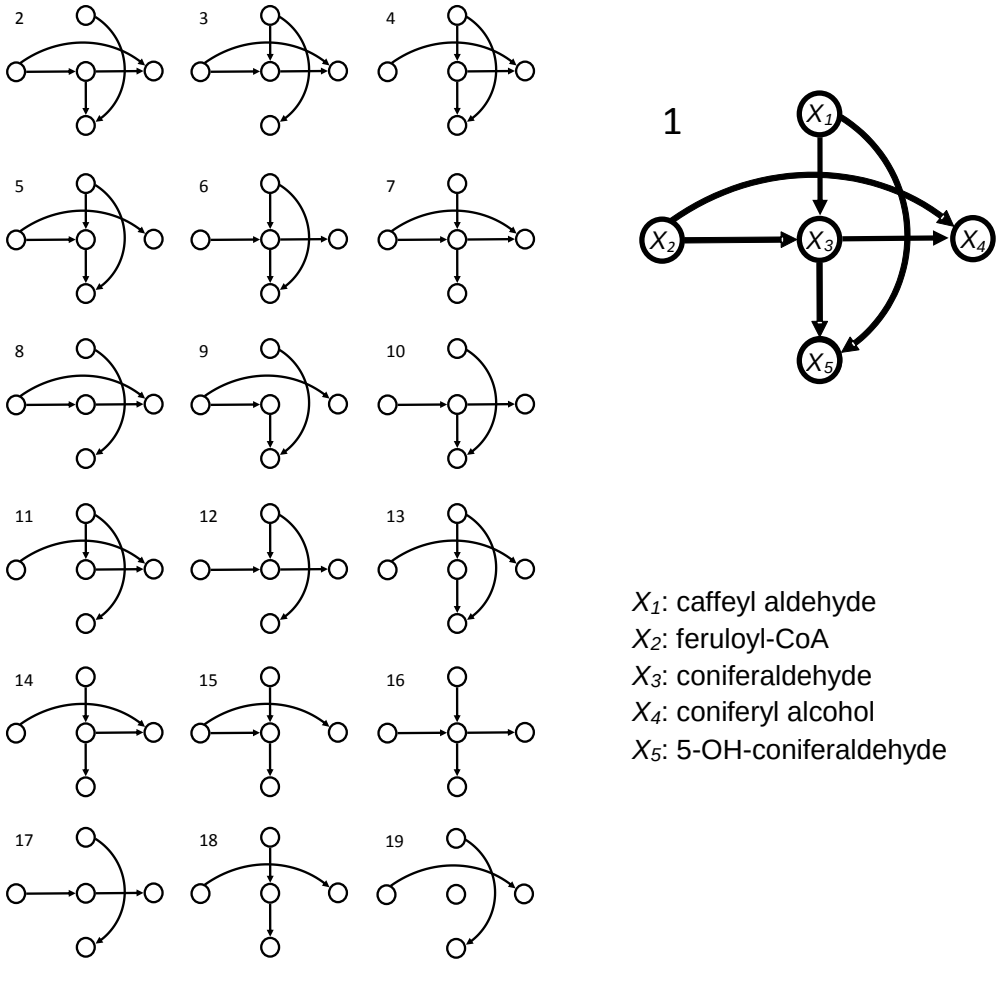

Figure A.1 Topological Configurations. A set of 19 structures is plausible when CCR1/CAD and COMT/F5H channels are considered. Only Configuration 16 lacks both channels. Other configurations represent different combinations of the absence and presence of channels.

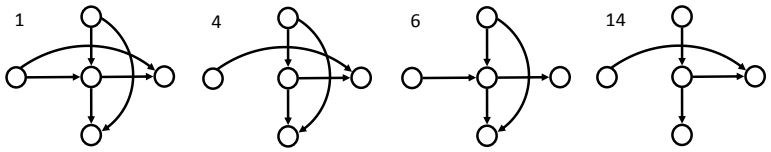

Figure A.2 Topological configurations that are best compatible with all available experimental data. Models with any of these topologies, which in addition account for product inhibition and competitive inhibition for CCR1, are able to reproduce all available experimental results. Note that at least one channel is present in all configurations.

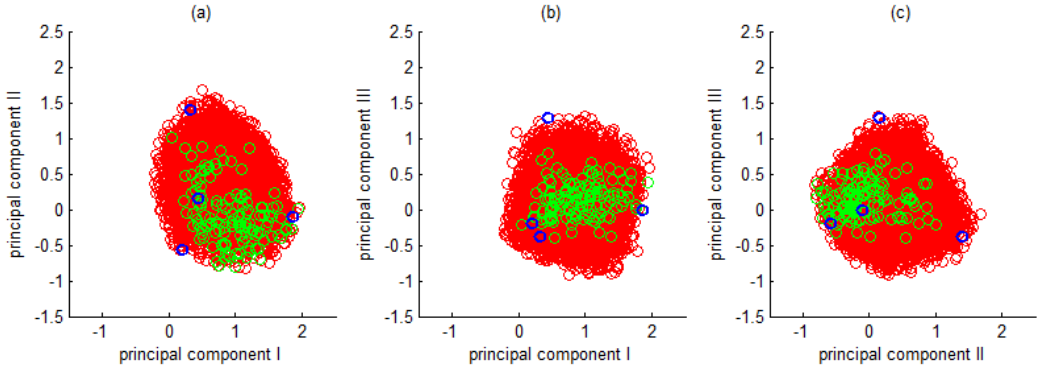

Figure A.3 Parameter distribution along the principal components of the parameter space. From the initial random sampling in the original parameter space, simulations led to only a limited number of points that successfully satisfied the model criteria. Using PCA, principal directions of the admissible points were identified and used to resample the space (red points). A second round of simulations filtered the randomly generated points again to insure that the model criteria were satisfied. The result is a set of fully admissible parameter values (green point).

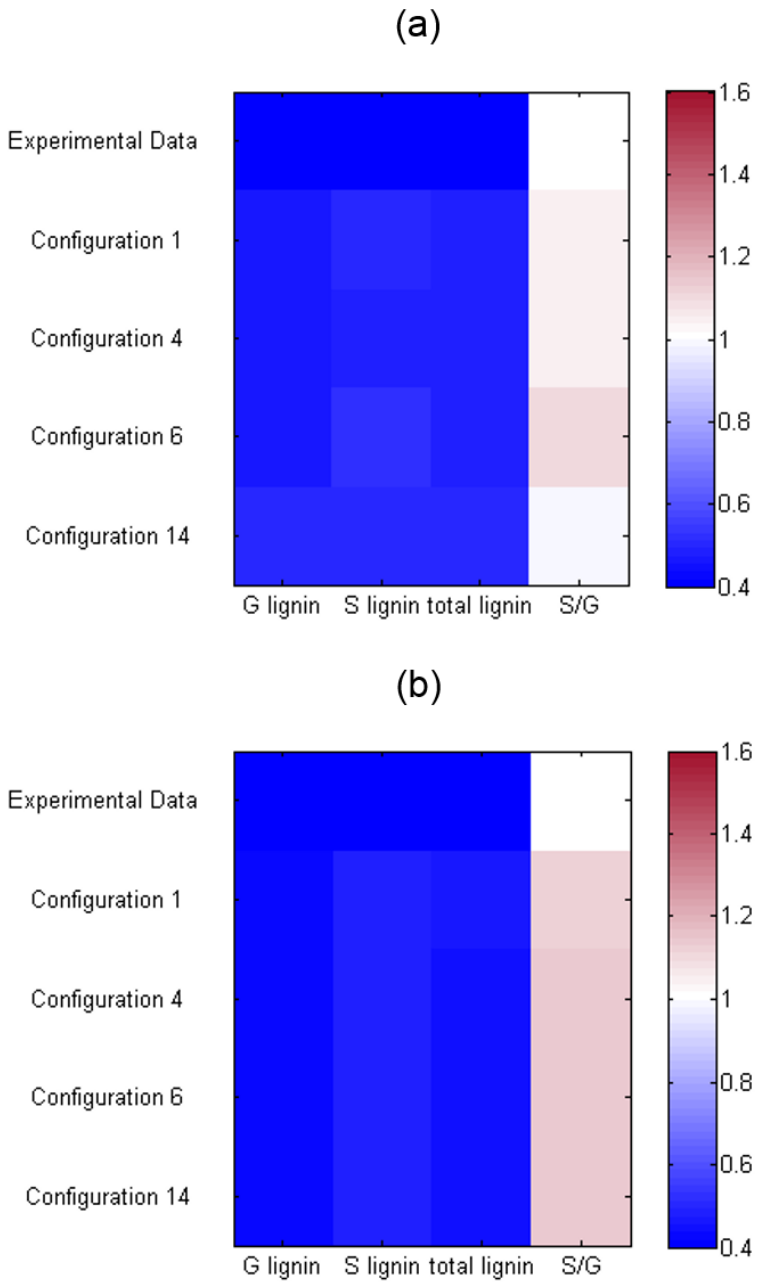

Figure A.4 Fold changes in lignin monomer concentrations in PvMYB4 transgenic plants. The top row represents the average of experimental data normalized with respect to the average of the control plants. Other rows represent the perturbed for PvMYB4 model results normalized with respect to wild type model results in compatible topological configurations. Wild type is set to white in the color bar. H lignin only counts for 3% of total lignin and is not shown in here. (a) The same common parameters are used for all configurations. (b) Best results, where each configuration is simulated using a separate, optimized parameter set.

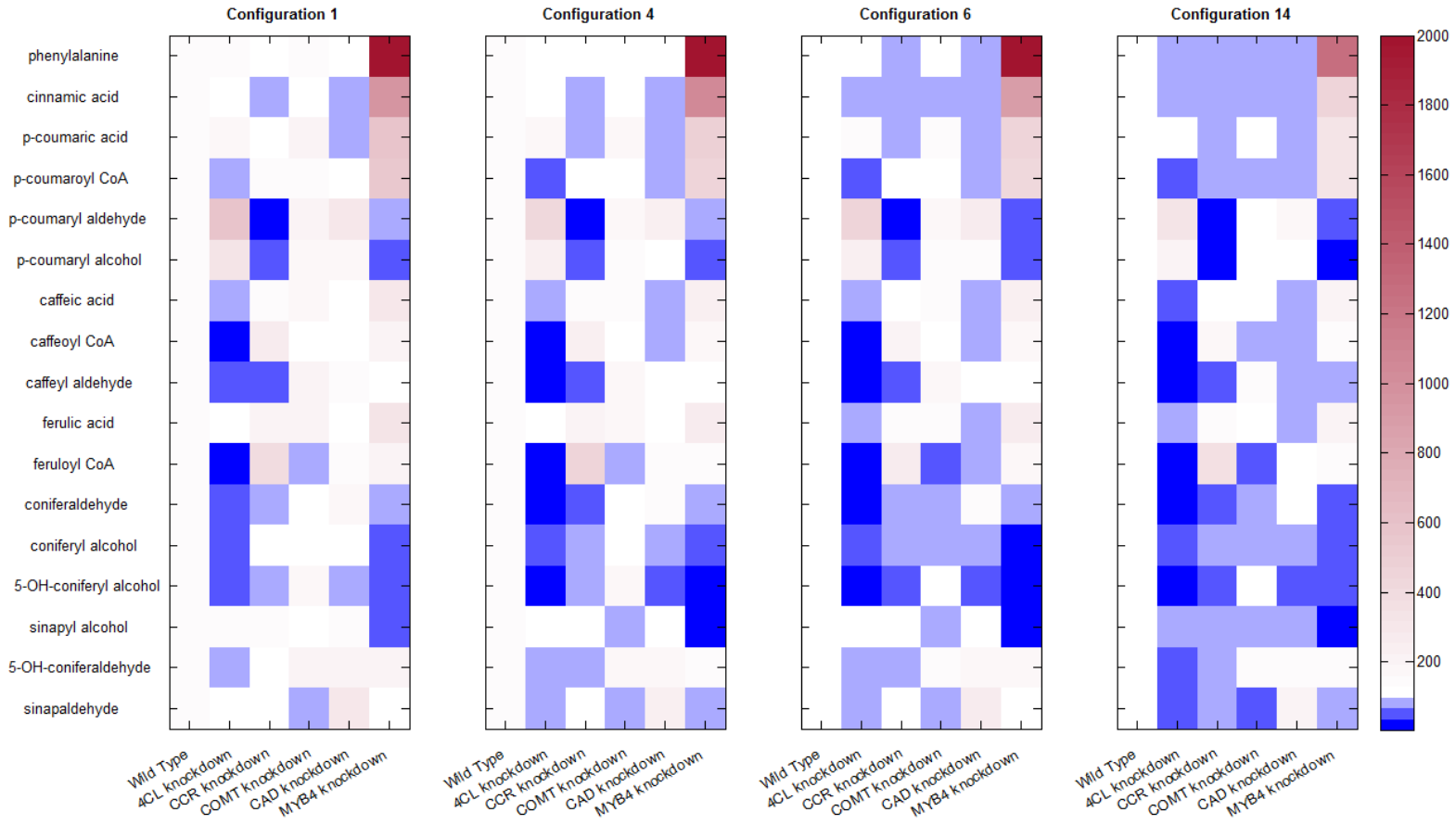

Figure A.5 Predicted steady-state profiles of key pathway metabolites in the wild type and in single knockdowns and PvMYB4 overexpression as predicted by the model. Concentrations are normalized and the base value is set to 100, which corresponds to white in the color bar. Any increases with respect to the wild type steady state are reflected in the red spectrum and any decreases in the blue spectrum. It is quite evident that the specific model configuration has no significant effect on the predictions.

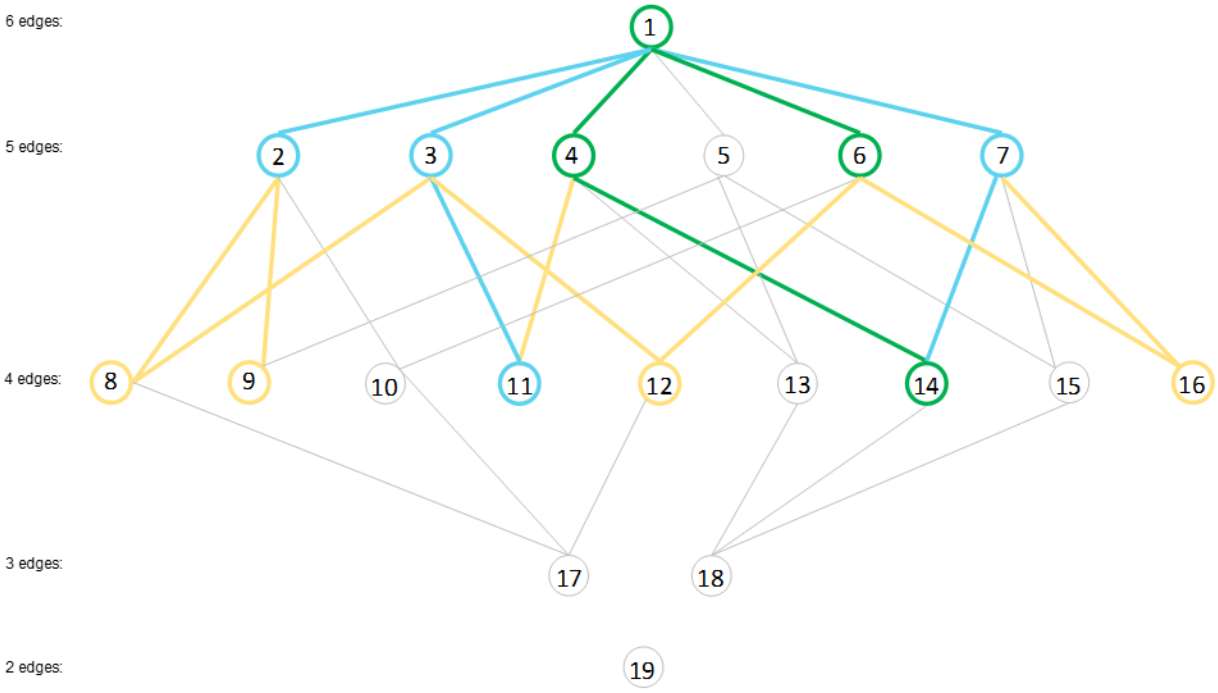

Figure A.6 Map of connectedness of admissible topological configurations. Two configurations are connected if they differ in only one edge. The best compatible configurations (1, 4, 6, 14; highlighted in green) are connected. The second best compatible configurations (2, 3, 7, 11; highlighted in blue) are also connected. The rest of the compatible configurations (8, 9, 12, 16; highlighted in orange) represent the lease abundant configurations in the solution space. The compatible configurations form a rather tight cluster.

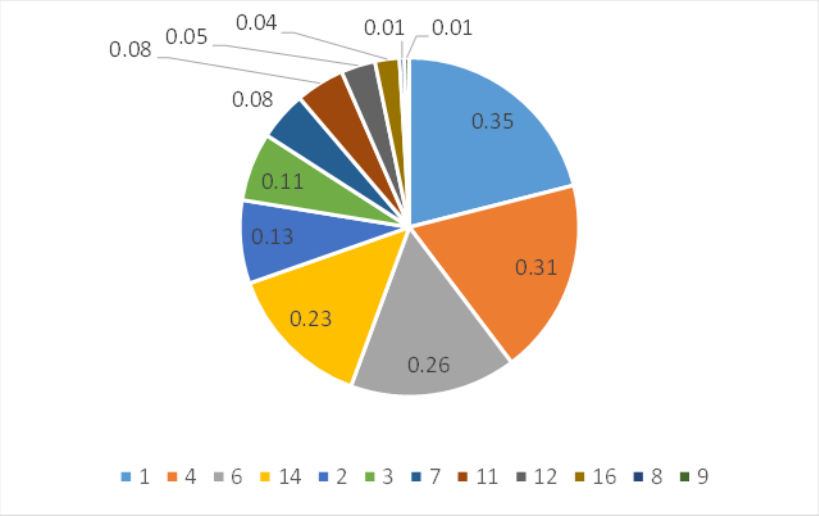

Figure A.7 Compatibility ratio in different topological configurations. Compatibility ratio for each configuration is defined as the number of parameter sets working with this configuration divided by the number of the parameter sets working for all the configurations collectively.

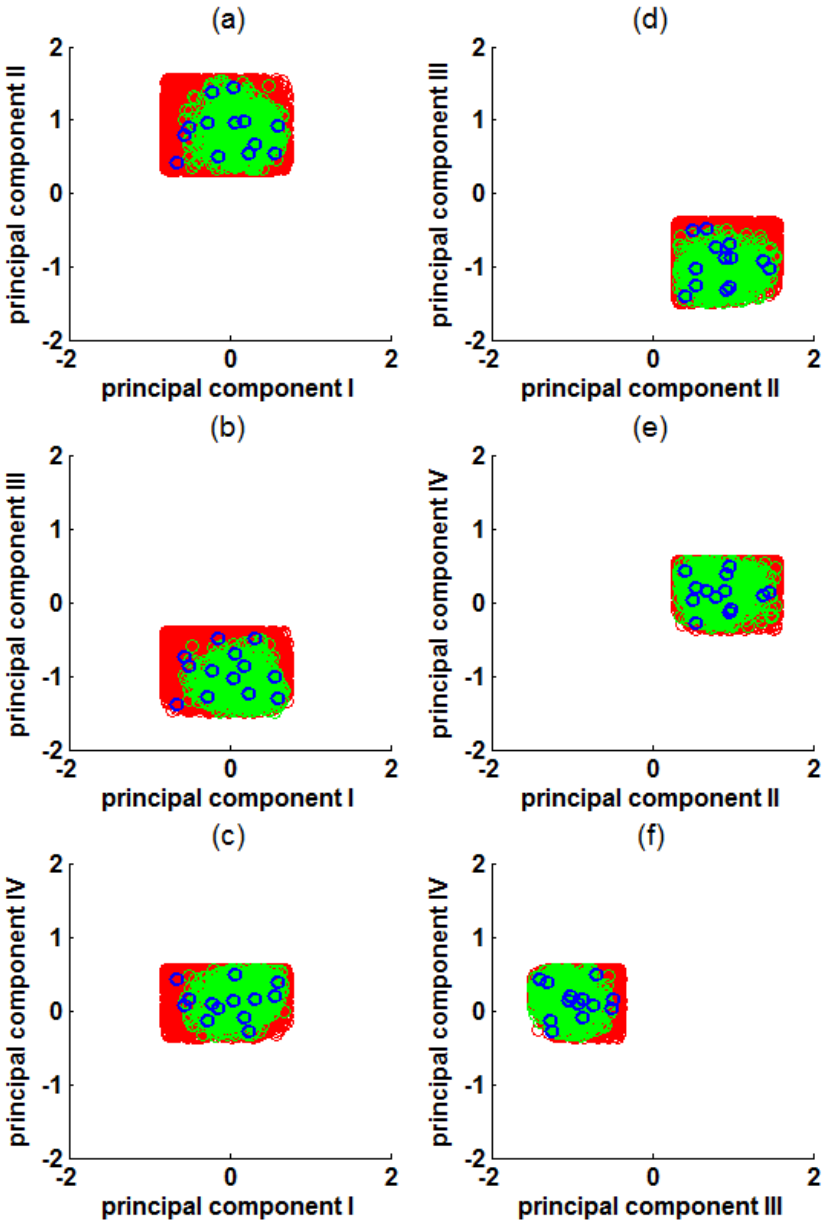

Figure A.8 Parameter distribution along the three principal components of the parameter space. From the initial sampling in the original parameter space, simulations led to only 13 points that successfully satisfied all model criteria (blue circles). Using PCA, principal directions of the admissible points were identified and used to resample the space with higher density (red points). A second round of simulations filtered the generated points again to insure that all model criteria were satisfied. The result is a set of fully admissible parameter values (green point). Note that the admissible parameter ranges are quite small.

APPENDIX B

SUPPLEMENTARY MATERIAL FOR CHAPTER V

B.1 Additional Figures Accompanying the Illustration Example of the Lignin Biosynthesis Pathway in Switchgrass

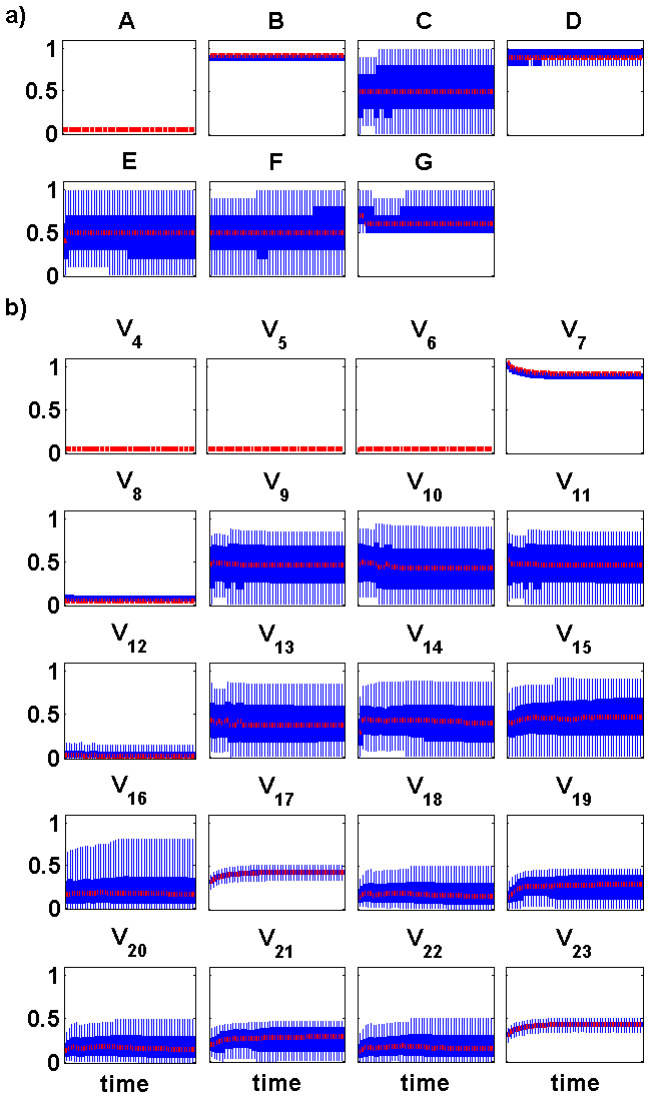

Figure B.1 Split ratios and flux distributions within the range $\mu \pm \sigma$ of admissible solutions in iteration 2.

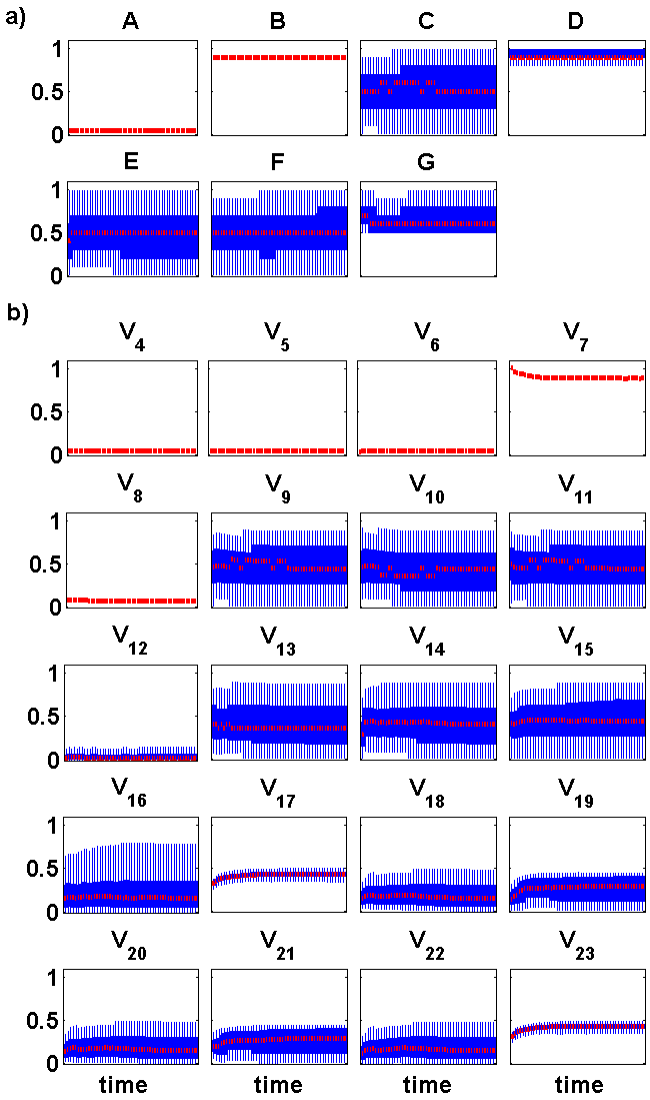

Figure B.2 Split ratios and flux distributions within the range $\mu \pm \sigma$ of admissible solutions in iteration 3.

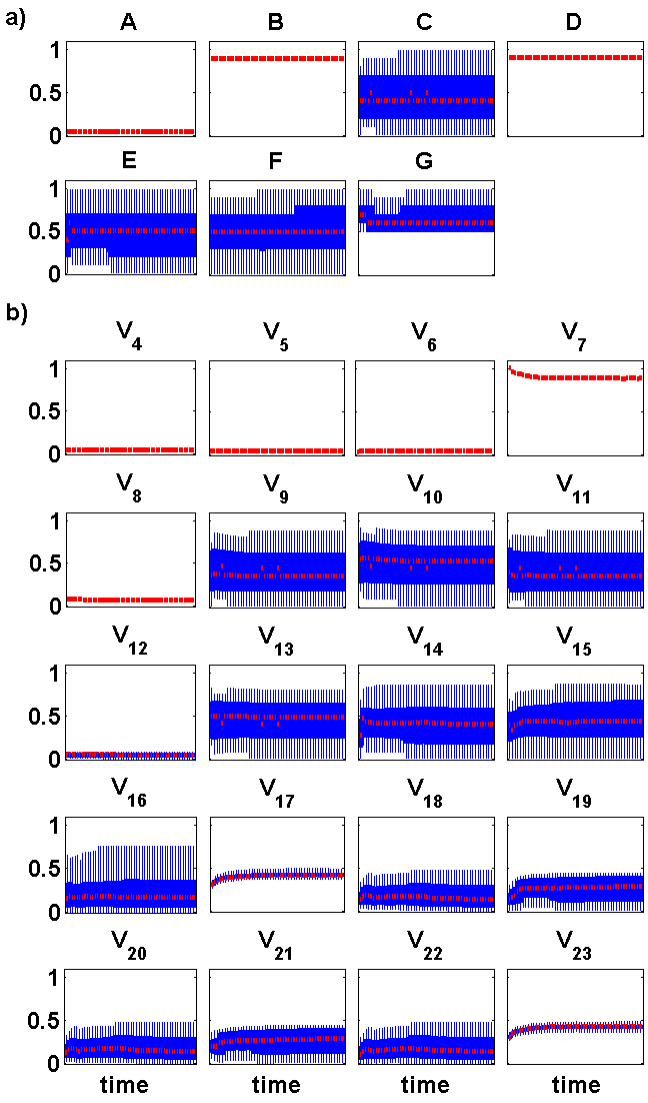

Figure B.3 Split ratios and flux distributions within the range $\mu \pm \sigma$ of admissible solutions in iteration 4.

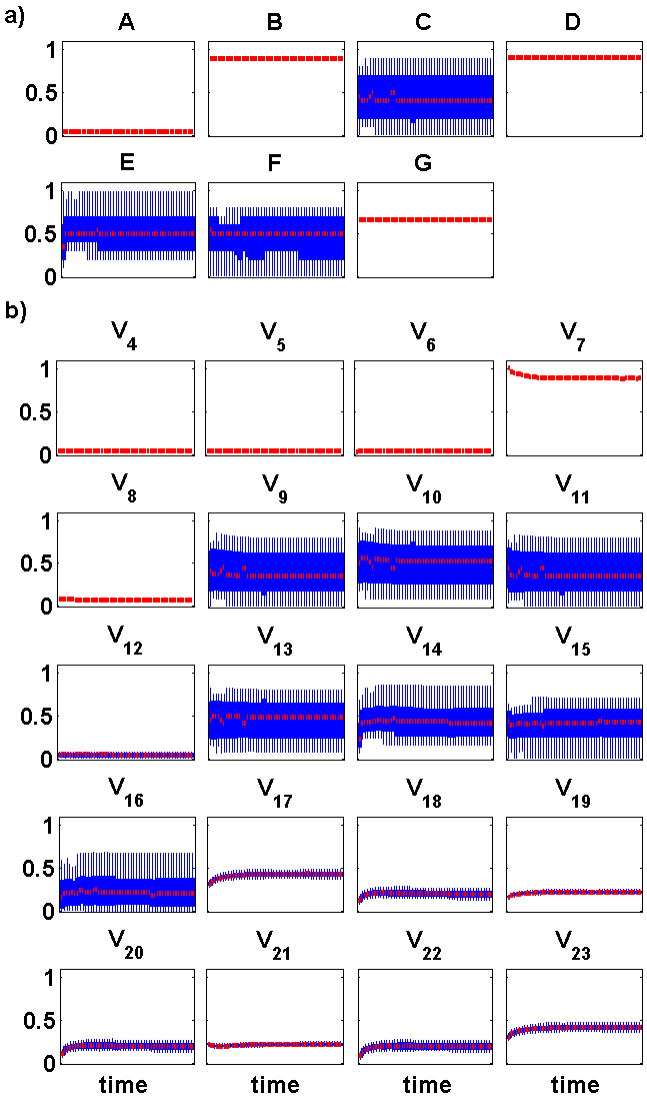

Figure B.4 Split ratios and flux distributions within the range $\mu \pm \sigma$ of admissible solutions in iteration 5.

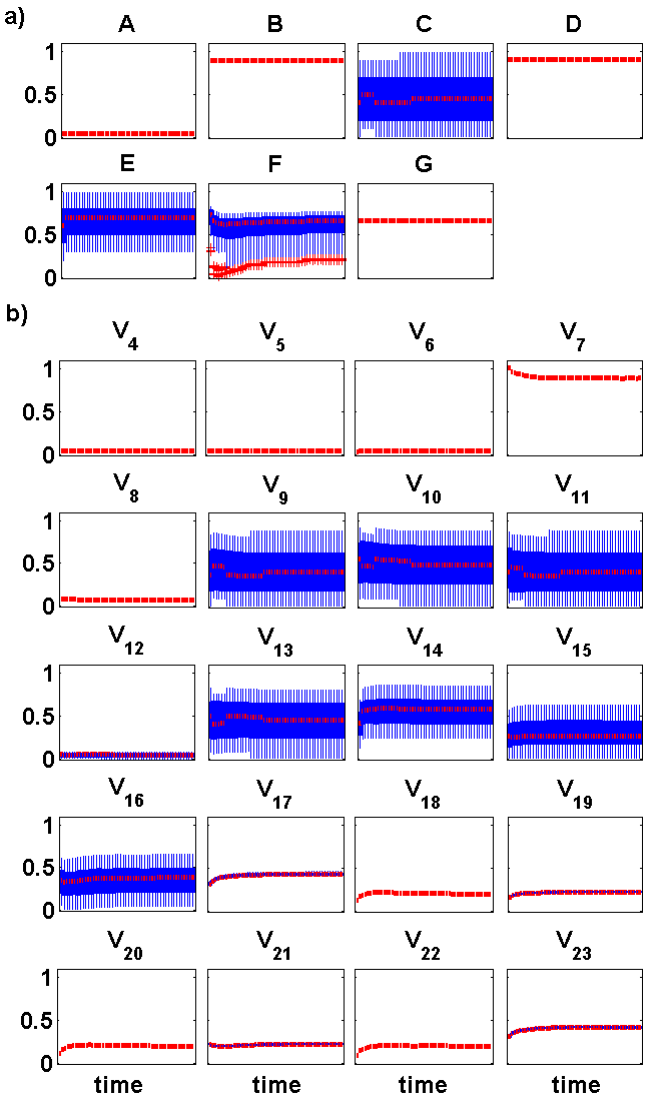

Figure B.5 Split ratios and flux distributions within the range $\mu \pm \sigma$ of admissible solutions in iteration 6.

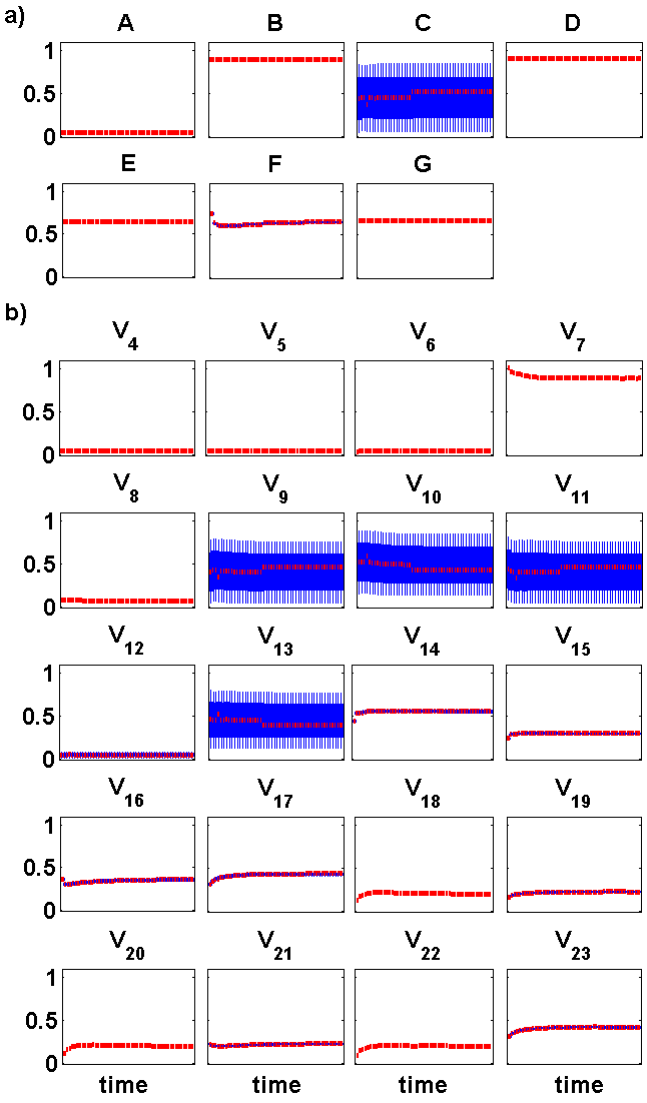

Figure B.6 Split ratios and flux distributions within the range $\mu \pm \sigma$ of admissible solutions in iteration 7.

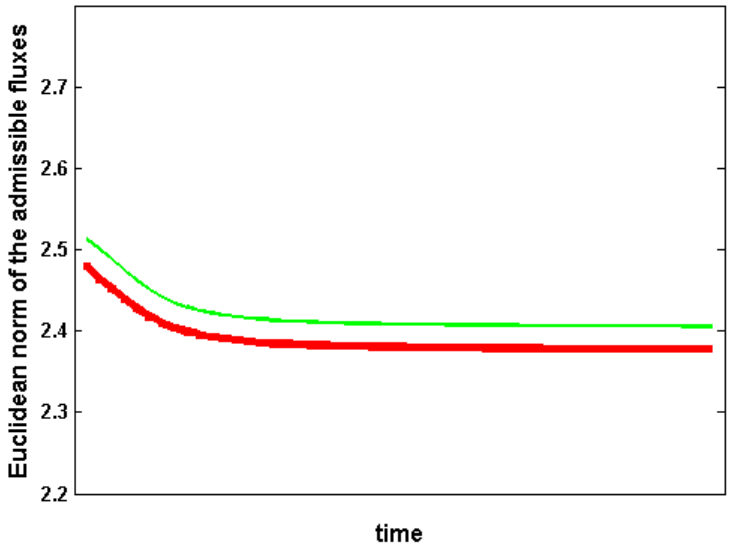

Figure B.7 Trends in norms of the inferred (red) flux distribution in comparison to the norms computed from the model (green). The error is less than 1.2%.

B.2 Analysis of a Simplified Model of Purine Metabolism

As a further illustration, a simplified model of purine metabolism (Figure B.8) was used to demonstrate the proposed method of stepwise inference of likely dynamic flux distribution. The model [172] is a simplified version of a computational model developed by Curto et al. [244], which represents purine metabolism in man. Purines form a family of aromatic organic compounds that have important roles for the synthesis of DNA, RNA, ATP and other vital macromolecules. The main metabolic breakdown product of purine metabolism is uric acid. The model was designed to analyze the dynamics of hyperuricemia that leads to gout and other diseases.

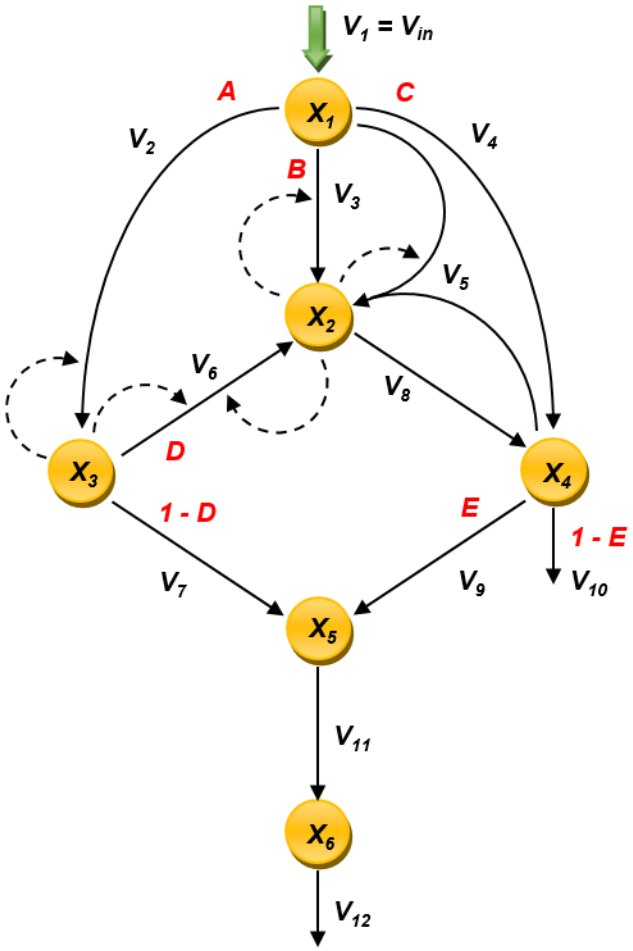

Figure B.8 Simplified representation of purine metabolism.

Figure B.8 shows the pathway diagram in mathematical representation. Similar to the illustration example of the lignin pathway in switchgrass, each branch point here introduces a degree of freedom to the pathway. The pathway thus has five degrees of freedom, which are represented in the diagram by split ratios, shown in red. The system variables are

$$
\begin{aligned}
&X_1 : \text{phosphoribosylpyrophosphate}\\
&X_2 : \text{inosine monophosphate}\\
&X_3 : \text{guanylates}\\
&X_4 : \text{hypoxanthine and inosine}\\
&X_5 : \text{xanthine}\\
&X_6 : \text{uric acid.}
\end{aligned}
\tag{B.1}
$$

The system equations are rewritten for the split ratios as follows:

$$
\begin{aligned}
&V_1 = V_{in}, && V_7 = (1-D)\cdot\left(V_2 - \dot{X}_3\right),\\
&V_2 = A\cdot\left(V_1 - \dot{X}_1\right), && V_8 = V_3 + V_5 + V_6 - \dot{X}_2,\\
&V_3 = B\cdot\left(V_1 - \dot{X}_1\right), && V_9 = E\cdot\left(V_4 + V_8 - V_5 - \dot{X}_4\right),\\
&V_4 = C\cdot\left(V_1 - \dot{X}_1\right), && V_{10} = (1-E)\cdot\left(V_4 + V_8 - V_5 - \dot{X}_4\right),\\
&V_5 = (1-A-B-C)\cdot\left(V_1 - \dot{X}_1\right), && V_{11} = V_7 + V_9 - \dot{X}_5,\\
&V_6 = D\cdot\left(V_2 - \dot{X}_3\right), && V_{12} = V_{11} - \dot{X}_6.
\end{aligned}
\tag{B.2}
$$

Here $V_{in} = 5$ is the input, and $[A, B, C, D, E]$ is the vector of split ratios. *A priori* knowledge about the pathway is that V_{10}/V_9 is less than 0.05, and V_7 accounts for about 12% of the mass leaving the pool of X_3. Therefore we set the sampling space as follows:

$$
\begin{aligned}
&[A, B, C, D, E] \in [0.01, 0.99],\\
&V_{10}/V_9 < 0.1,\\
&V_7/V_6 < 0.25.
\end{aligned}
\tag{B.3}
$$

Using the settings in (B.3) and (B.2), we executed the same procedure as illustrated in the *Text* to compute the likely flux distribution of the pathway. The results are shown in Figures B.9 to B.16.

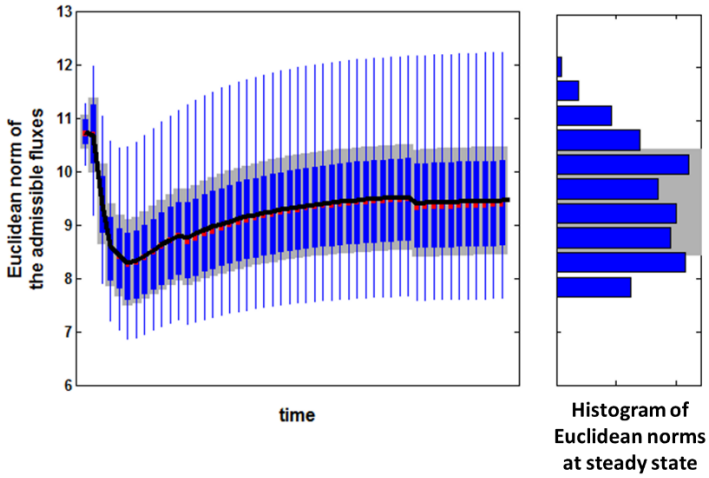

Figure B.9 Distribution of flux norms in iteration 1. Left panel: The gray band depicts the range $\mu \pm \sigma$, which contains about two thirds of the solutions, at each time point; the mean is shown in black. The thick blue boxes represent the second and third quartiles, and the red line is the median, which is similar to the mean. The thin blue lines are the first and fourth quartiles. Right panel: Histogram of norms in the left panel at the steady state, and range $\mu \pm \sigma$ (grey).

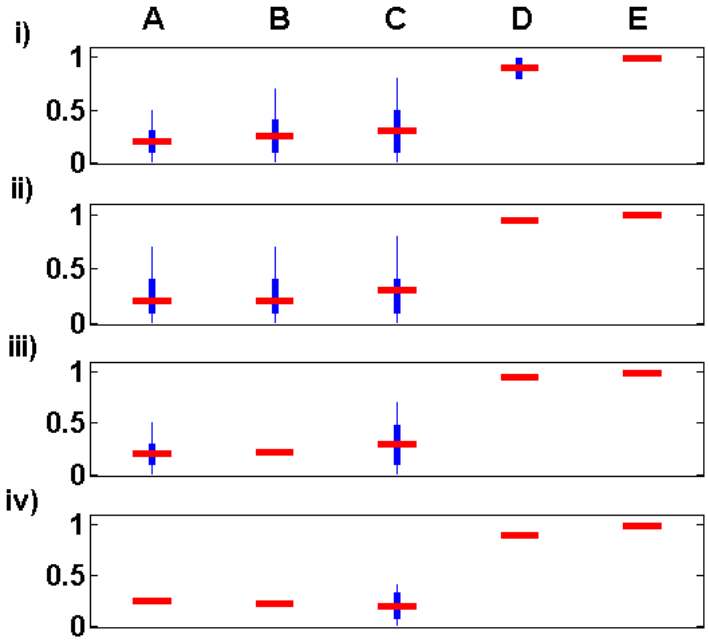

Figure B.10 Steady-state split ratios within the range $\mu \pm \sigma$ of admissible solutions. The subpanels correspond to the last time point (steady state) of each iteration. Each red line shows the median at the given iteration, and the boxplots represent the quartiles of split ratios.

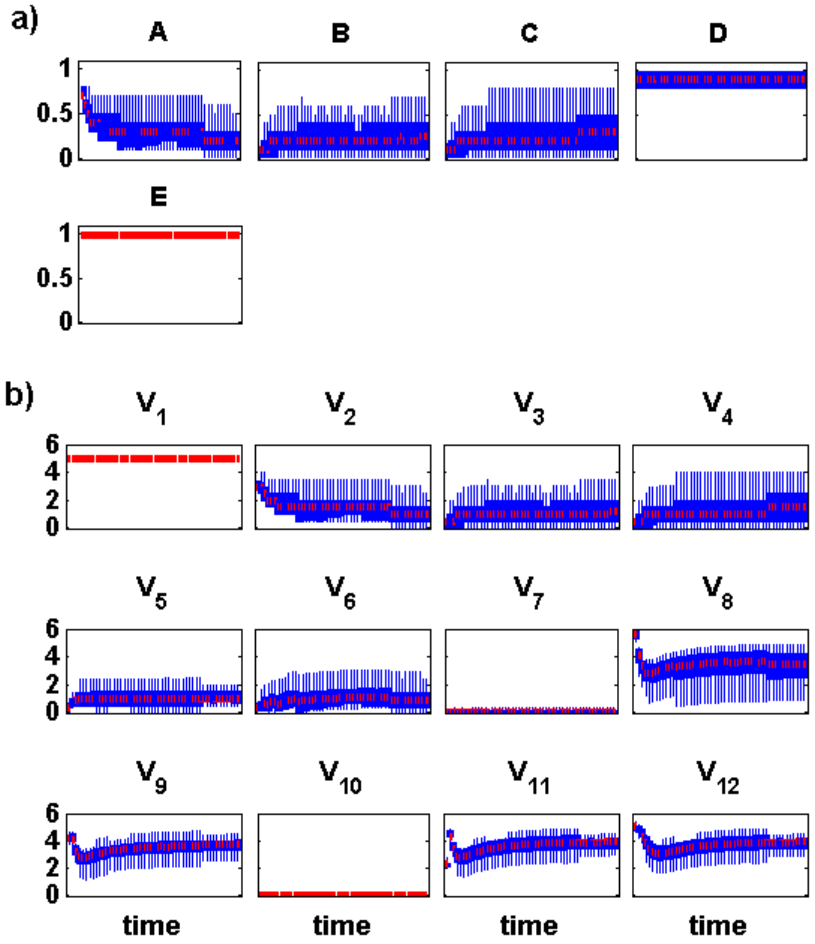

Figure B.11 Split ratios and flux distributions within the range $\mu \pm \sigma$ of admissible solutions in iteration 1. Split ratios D and E have very small variations, and thus are fixed on their mean values, $D = 0.95$ and $E = 0.99$, for the next iteration.

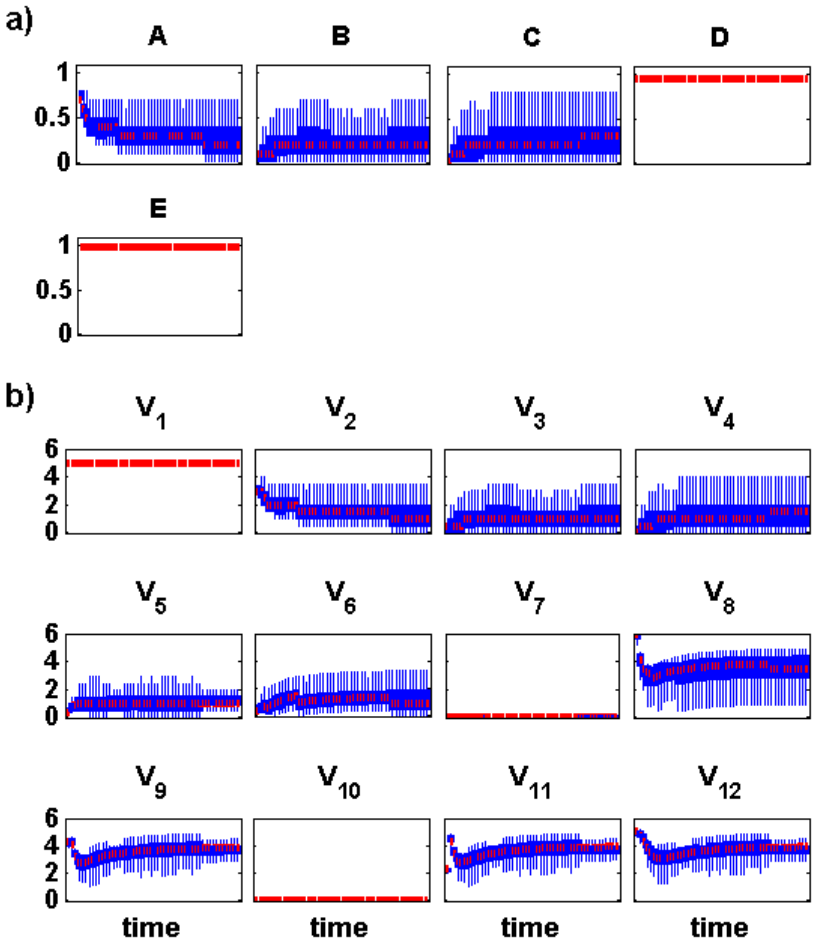

Figure B.12 Split ratios and flux distributions within the range $\mu \pm \sigma$ of admissible solutions in iteration 2. Split ratio B shows a relatively small variation considering the distribution of the quartiles, and thus is fixed on its smoothed mean profile for the next iteration.

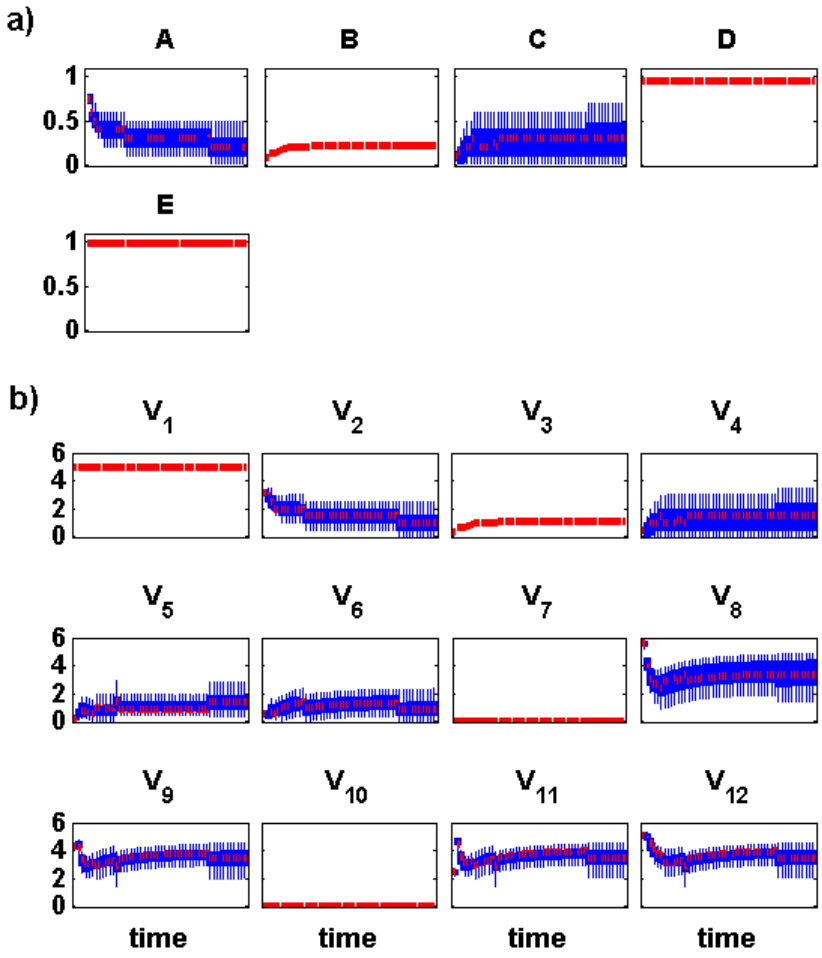

Figure B.13 Split ratios and flux distributions within the range $\mu \pm \sigma$ of admissible solutions in iteration 3. Split ratio A shows a relatively small variation considering the distribution of the quartiles, and thus is fixed on its smoothed mean profile for the next iteration.

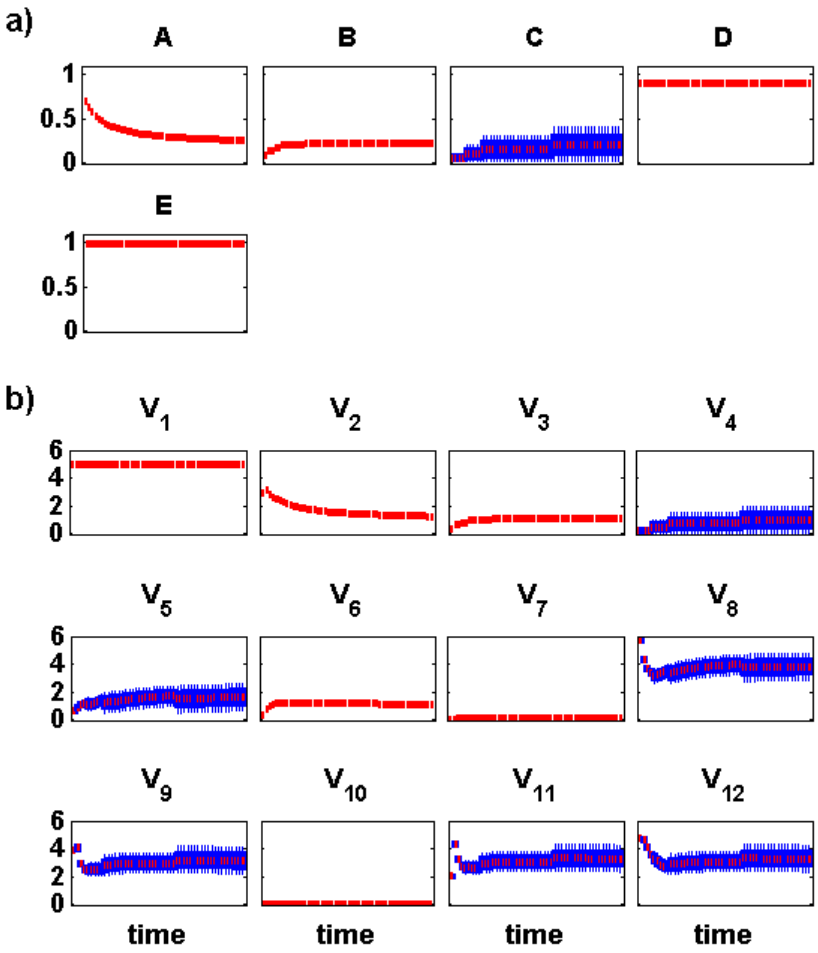

Figure B.14 Split ratios and flux distributions within the range $\mu \pm \sigma$ of admissible solutions in iteration 4. By the results of this iteration, split ratio C as the last degree of freedom is fixed on its smoothed mean profile.

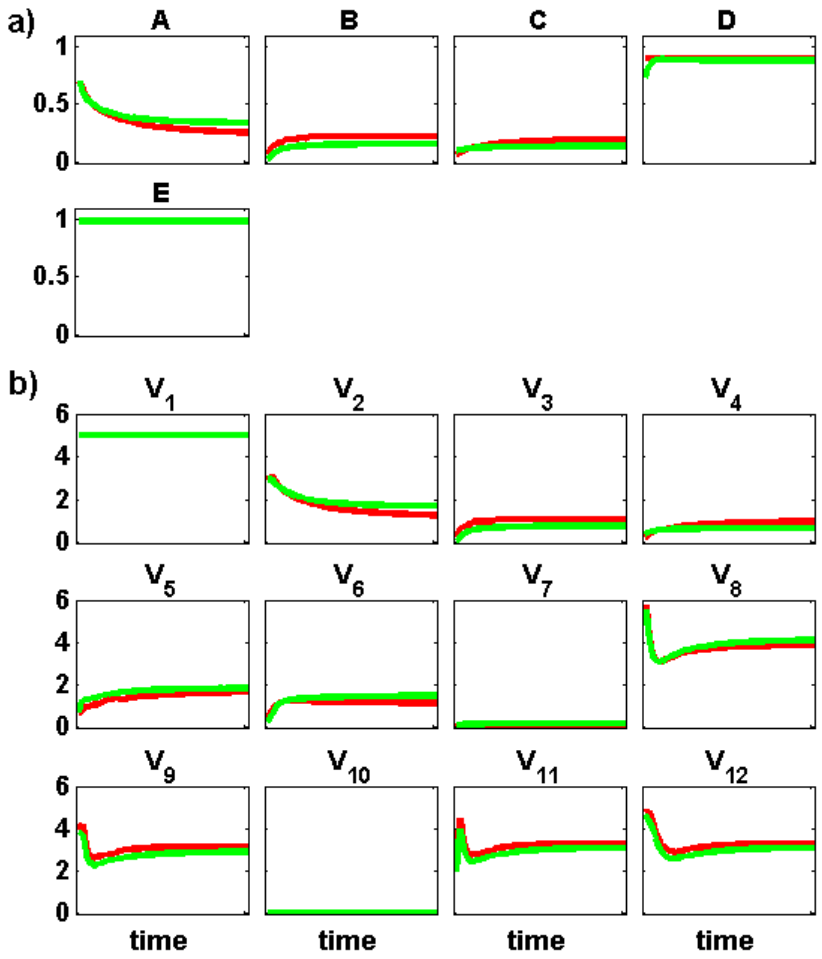

Figure B.15 Inferred likely split ratios and flux distributions (red) in comparison with the corresponding model features (green).

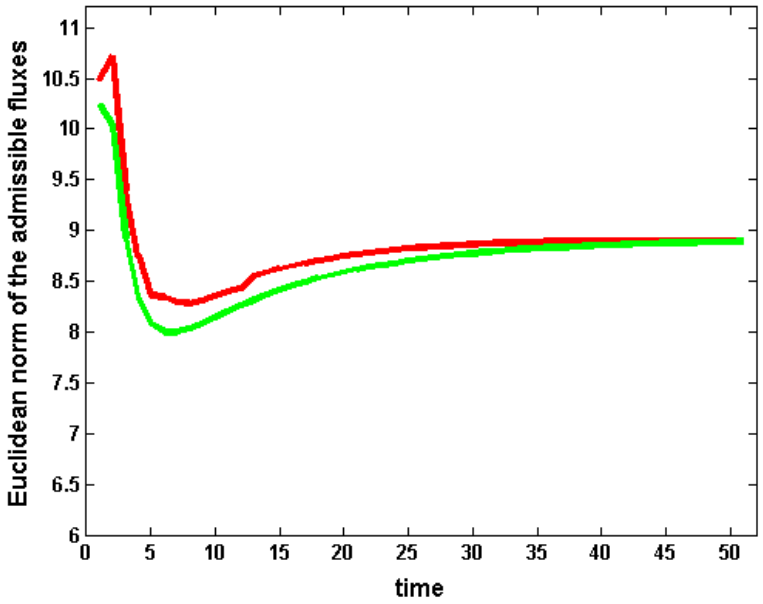

Figure B.16 Trends in norms of the inferred (red) flux distribution in comparison to the norms computed from the model (green).

REFERENCES

1. Faraji, M., L.L. Fonseca, L. Escamilla-Treviño, J. Barros-Rios, N. Engle, Z.K. Yang, T.J. Tschaplinski, R.A. Dixon, and E.O. Voit, *Mathematical models of lignin biosynthesis.* Biotechnology for Biofuels, 2018. **11**(1): p. 34.

2. Faraji, M. and E.O. Voit, *Improving Bioenergy Crops through Dynamic Metabolic Modeling.* Processes, 2017. **5**(4): p. 61.

3. Faraji, M., L.L. Fonseca, L. Escamilla-Treviño, R.A. Dixon, and E.O. Voit, *Computational inference of the structure and regulation of the lignin pathway in Panicum virgatum.* Biotechnology for Biofuels, 2015.

4. Lang, W.H. and I.C. Cookson, *On a flora, including vascular land plants, associated with Monograptus, in rocks of Silurian age, from Victoria, Australia.* Philosophical Transactions of the Royal Society of London B 1935. **224 (517)**: p. 421–449.

5. Kotyk, M.E., J.F. Basinger, P.G. Gensel, and T.A. de Freitas, *Morphologically complex plant macrofossils from the Late Silurian of Arctic Canada.* American Journal of Botany 2002. **89**: p. 1004–1013.

6. Xu, B., L.L. Escamilla-Trevino, N. Sathitsuksanoh, Z. Shen, H. Shen, Y.H. Zhang, R.A. Dixon, and B. Zhao, *Silencing of 4-coumarate:coenzyme A ligase in switchgrass leads to reduced lignin content and improved fermentable sugar yields for biofuel production.* New Phytol, 2011. **192**(3): p. 611-25.

7. Tschaplinski, T.J., R.F. Standaert, N.L. Engle, M.Z. Martin, A.K. Sangha, J.M. Parks, J.C. Smith, R. Samuel, N. Jiang, Y. Pu, A.J. Ragauskas, C.Y. Hamilton, C. Fu, Z.Y. Wang, B.H. Davison, R.A. Dixon, and J.R. Mielenz, *Down-regulation of the caffeic acid O-methyltransferase gene in switchgrass reveals a novel monolignol analog.* Biotechnol Biofuels, 2012. **5**(1): p. 71.

8. Fu, C., J.R. Mielenz, X. Xiao, Y. Ge, C.Y. Hamilton, M. Rodriguez, Jr., F. Chen, M. Foston, A. Ragauskas, J. Bouton, R.A. Dixon, and Z.Y. Wang, *Genetic manipulation of lignin reduces recalcitrance and improves ethanol production from switchgrass.* Proc Natl Acad Sci U S A, 2011. **108**(9): p. 3803-8.

9. Zhang, Z.Y., D.W. Rackemann, W.O.S. Doherty, and I.M. O'Hara, *Glycerol carbonate as green solvent for pretreatment of sugarcane bagasse.* Biotechnology for Biofuels, 2013. **6**.

10. Naik, S.N., V.V. Goud, P.K. Rout, and A.K. Dalai, *Production of first and second generation biofuels: A comprehensive review.* Renewable & Sustainable Energy Reviews, 2010. **14**(2): p. 578-597.

11. Jung, H.G. and K.P. Vogel, *Influence of lignin on digestibility of forage cell wall material.* J Anim Sci, 1986. **62**(6): p. 1703-12.

12. Boerjan, W., Ralph, J., Baucher, M., *Lignin biosynthesis.* Annu Rev Plant Biol, 2003. **54**: p. 519-46.

13. Weng, J.K., L. X, S. J, and C. C. *Independent origins of syringyl lignin in vascular plants*. in *Natl Acad Sci*. 2008. USA.

14. Weng, J.K., T. Akiyama, N.D. Bonawitz, X. Li, J. Ralph, and C. Chapple, *Convergent evolution of syringyl lignin biosynthesis via distinct pathways in the lycophyte Selaginella and flowering plants.* Plant Cell, 2010. **22**(4): p. 1033-45.

15. Weng, J.K., A. T, R. J, and C. C, *Independent recruitment of an O-methyltransferase for syringyl lignin biosynthesis in Selaginella moellendorffii.* Plant Cell 2011. **23**: p. 2708–2724.

16. Bulik, S., S. Grimbs, C. Huthmacher, J. Selbig, and H.G. Holzhutter, *Kinetic hybrid models composed of mechanistic and simplified enzymatic rate laws--a promising method for speeding up the kinetic modelling of complex metabolic networks.* FEBS J, 2009. **276**(2): p. 410-24.

17. van Eunen, K., B. J, D.-L. P, P. J, C. AB, M. FI, O. R, T. I, v.d.B. J, S. GJ, v.G. WM, B. S, H. JJ, d.W. JH, d.M. MJ, K. C, N. J, W. HV, and B. BM, *Measuring enzyme activities under standardized in vivo-like conditions for systems biology.* FEBS J., 2010. **277(3)**: p. 749-60.

18. Lee, Y., L. Escamilla-Trevino, R.A. Dixon, and E.O. Voit, *Functional analysis of metabolic channeling and regulation in lignin biosynthesis: a computational approach.* PLoS Comput Biol, 2012. **8**(11): p. e1002769.

19. Doebley, J. *Teosinte as a Grain Crop*. [cited 2017 August]; Available from: http://teosinte.wisc.edu/grain_Crop.html.

20. *List of sequenced plant genomes*. [cited 2017 August]; Available from: http://en.wikipedia.org/wiki/List_of_sequenced_plant_genomes#Gymnosperm.

21. Williams, T.C.R., L. Miguet, S.K. Masakapalli, N.J. Kruger, L.J. Sweetlove, and R.G. Ratcliffe, *Metabolic network fluxes in heterotrophic Arabidopsis cells: Stability of the flux distribution under different oxygenation conditions.* Plant Physiology, 2008. **148**(2): p. 704-718.

22. Yuan, J.S., D.W. Galbraith, S.Y. Dai, P. Griffin, and C.N. Stewart, *Plant systems biology comes of age.* Trends in Plant Science, 2008. **13**(4): p. 165-171.

23. *Human metabolome Database*. [cited 2017 August]; Available from: http://www.hmdb.ca/statistics.

24. Dixon, R.A. and D. Strack, *Phytochemistry meets genome analysis, and beyond.* Phytochemistry, 2003. **62**(6): p. 815-816.

25. Saito, K. and F. Matsuda, *Metabolomics for Functional Genomics, Systems Biology, and Biotechnology.* Annual Review of Plant Biology, 2010. **61**(1): p. 463-489.

26. Cao, H.X., W. Wang, H.T.T. Le, and G.T.H. Vu, *The Power of CRISPR-Cas9-Induced Genome Editing to Speed Up Plant Breeding.* International Journal of Genomics, 2016. **2016**: p. 10.

27. Shan, Q., Y. Wang, J. Li, Y. Zhang, K. Chen, Z. Liang, K. Zhang, J. Liu, J.J. Xi, J.-L. Qiu, and C. Gao, *Targeted genome modification of crop plants using a CRISPR-Cas system.* Nat Biotech, 2013. **31**(8): p. 686-688.

28. Nekrasov, V., B. Staskawicz, D. Weigel, J.D.G. Jones, and S. Kamoun, *Targeted mutagenesis in the model plant Nicotiana benthamiana using Cas9 RNA-guided endonuclease.* Nat Biotech, 2013. **31**(8): p. 691-693.

29. Li, J.F., J.E. Norville, J. Aach, M. McCormack, D. Zhang, J. Bush, G.M. Church, and J. Sheen, *Multiplex and homologous recombination-mediated genome editing in Arabidopsis and Nicotiana benthamiana using guide RNA and Cas9.* Nat Biotechnol, 2013. **31**(8): p. 688-91.

30. Cai, Y., L. Chen, X. Liu, C. Guo, S. Sun, C. Wu, B. Jiang, T. Han, and W. Hou, *CRISPR/Cas9-mediated targeted mutagenesis of GmFT2a delays flowering time in soya bean.* Plant Biotechnol J, 2017.

31. Tian, S., L. Jiang, Q. Gao, J. Zhang, M. Zong, H. Zhang, Y. Ren, S. Guo, G. Gong, F. Liu, and Y. Xu, *Efficient CRISPR/Cas9-based gene knockout in watermelon.* Plant Cell Rep, 2017. **36**(3): p. 399-406.

32. Soyk, S., N.A. Muller, S.J. Park, I. Schmalenbach, K. Jiang, R. Hayama, L. Zhang, J. Van Eck, and J.M. Jimenez-Gomez, *Variation in the flowering gene SELF PRUNING 5G promotes day-neutrality and early yield in tomato.* 2017. **49**(1): p. 162-168.

33. Aharoni, A. and G. Galili, *Metabolic engineering of the plant primary-secondary metabolism interface.* Curr Opin Biotechnol, 2011. **22**(2): p. 239-44.

34. Ratcliffe, R.G. and Y. Shachar-Hill, *Measuring multiple fluxes through plant metabolic networks.* Plant J, 2006. **45**(4): p. 490-511.

35. Morgan, J.A. and D. Rhodes, *Mathematical Modeling of plant metabolic pathways.* Metabolic Engineering, 2002. **4**(1): p. 80-89.

36. Sweetlove, L.J., T.C. Williams, C.Y. Cheung, and R.G. Ratcliffe, *Modelling metabolic CO(2) evolution--a fresh perspective on respiration.* Plant Cell Environ, 2013. **36**(9): p. 1631-40.

37. Nepali, M.R. *Polyploidy breeding*. 2013 [cited 2017 August]; Available from: http://mukeshramjalipb.blogspot.com/2013/03/polyploidy-breeding.html.

38. Meru, G. *Polyploidy*. 2013 [cited 2017 August]; Available from: http://plantbreeding.coe.uga.edu/index.php?title=5._Polyploidy.

39. Lukhtanov, V.A., *The blue butterfly Polyommatus (Plebicula) atlanticus (Lepidoptera, Lycaenidae) holds the record of the highest number of chromosomes in the non-polyploid eukaryotic organisms.* Comparative Cytogenetics, 2015. **9**(4): p. 683-690.

40. Janick, J. and American Society for Horticultural Science., *Plant breeding reviews. Volume 31*, in *Plant breeding reviews v 31*. 2009, Wiley Blackwell,: Hoboken, NJ.

41. Yu, J., J. Wang, W. Lin, S. Li, H. Li, J. Zhou, P. Ni, W. Dong, S. Hu, C. Zeng, J. Zhang, Y. Zhang, R. Li, Z. Xu, S. Li, X. Li, H. Zheng, L. Cong, L. Lin, J. Yin, J. Geng, G. Li, J. Shi, J. Liu, H. Lv, J. Li, J. Wang, Y. Deng, L. Ran, X. Shi, X. Wang, Q. Wu, C. Li, X. Ren, J. Wang, X. Wang, D. Li, D. Liu, X. Zhang, Z. Ji, W. Zhao, Y. Sun, Z. Zhang, J. Bao, Y. Han, L. Dong, J. Ji, P. Chen, S. Wu, J. Liu, Y. Xiao, D. Bu, J. Tan, L. Yang, C. Ye, J. Zhang, J. Xu, Y. Zhou, Y. Yu, B. Zhang, S. Zhuang, H. Wei, B. Liu, M. Lei, H. Yu, Y. Li, H. Xu, S. Wei, X. He, L. Fang, Z. Zhang, Y. Zhang, X. Huang, Z. Su, W. Tong, J. Li, Z. Tong, S. Li, J. Ye, L. Wang, L. Fang, T. Lei, C. Chen, H. Chen, Z. Xu, H. Li, H. Huang, F. Zhang, H. Xu, N. Li, C. Zhao, S. Li, L. Dong, Y. Huang, L. Li, Y. Xi, Q. Qi, W. Li, B. Zhang, W. Hu, Y. Zhang, X. Tian, Y. Jiao, X. Liang, J. Jin, L. Gao, W. Zheng, B. Hao, S. Liu, W. Wang, L. Yuan, M. Cao, J. McDermott, R. Samudrala, J. Wang, G.K.-S. Wong and H. Yang, *The Genomes of Oryza sativa: A History of Duplications.* PLOS Biology, 2005. **3**(2): p. e38.

42. Arnold, A. and Z. Nikoloski, *In search for an accurate model of the photosynthetic carbon metabolism.* Mathematics and Computers in Simulation, 2014. **96**: p. 171-194.

43. Szecowka, M., R. Heise, T. Tohge, A. Nunes-Nesi, D. Vosloh, J. Huege, R. Feil, J. Lunn, Z. Nikoloski, M. Stitt, A.R. Fernie, and S. Arrivault, *Metabolic Fluxes in an Illuminated Arabidopsis Rosette.* Plant Cell, 2013. **25**(2): p. 694-714.

44. Zhu, X.G., Y. Wang, D.R. Ort, and S.P. Long, *e-photosynthesis: a comprehensive dynamic mechanistic model of C3 photosynthesis: from light capture to sucrose synthesis.* Plant Cell and Environment, 2013. **36**(9): p. 1711-1727.

45. Arnold, A. and Z. Nikoloski, *A quantitative comparison of Calvin-Benson cycle models.* Trends in Plant Science, 2011. **16**(12): p. 676-683.

46. Cheung, C.Y.M., M.G. Poolman, D.A. Fell, R.G. Ratcliffe, and L.J. Sweetlove, *A Diel Flux Balance Model Captures Interactions between Light and Dark Metabolism during Day-Night Cycles in C-3 and Crassulacean Acid Metabolism Leaves.* Plant Physiology, 2014. **165**(2): p. 917-929.

47. Boyle, N.R. and J.A. Morgan, *Computation of metabolic fluxes and efficiencies for biological carbon dioxide fixation.* Metabolic Engineering, 2011. **13**(2): p. 150-158.

48. Guo, Y. and J.L. Tan, *A kinetic model structure for delayed fluorescence from plants.* Biosystems, 2009. **95**(2): p. 98-103.

49. Pearcy, R.W., L.J. Gross, and D. He, *An improved dynamic model of photosynthesis for estimation of carbon gain in sunfleck light regimes.* Plant Cell and Environment, 1997. **20**(4): p. 411-424.

50. Poolman, M.G., L. Miguet, L.J. Sweetlove, and D.A. Fell, *A Genome-Scale Metabolic Model of Arabidopsis and Some of Its Properties.* Plant Physiology, 2009. **151**(3): p. 1570-1581.

51. Lakshmanan, M., Z.Y. Zhang, B. Mohanty, J.Y. Kwon, H.Y. Choi, H.J. Nam, D.I. Kim, and D.Y. Lee, *Elucidating Rice Cell Metabolism under Flooding and Drought Stresses Using Flux-Based Modeling and Analysis.* Plant Physiology, 2013. **162**(4): p. 2140-2150.

52. Sweetlove, L.J., K.F.M. Beard, A. Nunes-Nesi, A.R. Fernie, and R.G. Ratcliffe, *Not just a circle: flux modes in the plant TCA cycle.* Trends in Plant Science, 2010. **15**(8): p. 462-470.

53. Baghalian, K., M.R. Hajirezaei, and F. Schreiber, *Plant Metabolic Modeling: Achieving New Insight into Metabolism and Metabolic Engineering.* Plant Cell, 2014. **26**(10): p. 3847-3866.

54. Rohwer, J.M., *Kinetic modelling of plant metabolic pathways.* Journal of Experimental Botany, 2012. **63**(6): p. 2275-2292.

55. Shen, H., C.R. Poovaiah, A. Ziebell, T.J. Tschaplinski, S. Pattathil, E. Gjersing, N.L. Engle, R. Katahira, Y. Pu, R. Sykes, F. Chen, A.J. Ragauskas, J.R. Mielenz, M.G. Hahn, M. Davis, C.N. Stewart, Jr., and R.A. Dixon, *Enhanced characteristics*

of genetically modified switchgrass (Panicum virgatum L.) for high biofuel production. Biotechnol Biofuels, 2013. **6**(1): p. 71.

56. Fu, C.X., X.R. Xiao, Y.J. Xi, Y.X. Ge, F. Chen, J. Bouton, R.A. Dixon, and Z.Y. Wang, *Downregulation of Cinnamyl Alcohol Dehydrogenase (CAD) Leads to Improved Saccharification Efficiency in Switchgrass.* Bioenergy Research, 2011. **4**(3): p. 153-164.

57. Lee, Y. and E.O. Voit, *Mathematical modeling of monolignol biosynthesis in Populus xylem.* Math Biosci, 2010. **228**(1): p. 78-89.

58. Lee, Y., F. Chen, L. Gallego-Giraldo, R.A. Dixon, and E.O. Voit, *Integrative analysis of transgenic alfalfa (Medicago sativa L.) suggests new metabolic control mechanisms for monolignol biosynthesis.* PLoS Comput Biol, 2011. **7**(5): p. e1002047.

59. Wang, J.P., P.P. Naik, H.C. Chen, R. Shi, C.Y. Lin, J. Liu, C.M. Shuford, Q. Li, Y.H. Sun, S. Tunlaya-Anukit, C.M. Williams, D.C. Muddiman, J.J. Ducoste, R.R. Sederoff, and V.L. Chiang, *Complete proteomic-based enzyme reaction and inhibition kinetics reveal how monolignol biosynthetic enzyme families affect metabolic flux and lignin in Populus trichocarpa.* Plant Cell, 2014. **26**(3): p. 894-914.

60. Amthor, J.S., *Efficiency of lignin biosynthesis: a quantitative analysis.* Ann Bot, 2003. **91**(6): p. 673-95.

61. Saha, R., P.F. Suthers, and C.D. Maranas, *Zea mays iRS1563: a comprehensive genome-scale metabolic reconstruction of maize metabolism.* PLoS One, 2011. **6**(7): p. e21784.

62. Marshall-Colon, A., S.P. Long, D.K. Allen, G. Allen, D.A. Beard, B. Benes, S. von Caemmerer, A.J. Christensen, D.J. Cox, J.C. Hart, P.M. Hirst, K. Kannan, D.S. Katz, J.P. Lynch, A.J. Millar, B. Panneerselvam, N.D. Price, P. Prusinkiewicz, D. Raila, R.G. Shekar, S. Shrivastava, D. Shukla, V. Srinivasan, M. Stitt, M.J. Turk, E.O. Voit, Y. Wang, X. Yin, and X.-G. Zhu, *Crops In Silico: Generating Virtual Crops Using an Integrative and Multi-scale Modeling Platform.* Frontiers in Plant Science, 2017. **8**(786).

63. Bogart, E. and C.R. Myers, *Multiscale Metabolic Modeling of C4 Plants: Connecting Nonlinear Genome-Scale Models to Leaf-Scale Metabolism in Developing Maize Leaves.* PLOS ONE, 2016. **11**(3): p. e0151722.

64. Voit, E.O., *Models-of-data and models-of-processes in the post-genomic era.* Math Biosci, 2002. **180**: p. 263-74.

65. Wiechert, W., *C-13 metabolic flux analysis.* Metabolic Engineering, 2001. **3**(3): p. 195-206.

66. Wiechert, W., M. Mollney, S. Petersen, and A.A. de Graaf, *A universal framework for C-13 metabolic flux analysis.* Metabolic Engineering, 2001. **3**(3): p. 265-283.

67. Maarleveld, T.R., R.A. Khandelwal, B.G. Olivier, B. Teusink, and F.J. Bruggeman, *Basic concepts and principles of stoichiometric modeling of metabolic networks.* Biotechnol J, 2013. **8**(9): p. 997-1008.

68. Libourel, I.G. and Y. Shachar-Hill, *Metabolic flux analysis in plants: from intelligent design to rational engineering.* Annu Rev Plant Biol, 2008. **59**: p. 625-50.

69. Kruger, N.J. and R.G. Ratcliffe, *Insights into plant metabolic networks from steady-state metabolic flux analysis.* Biochimie, 2009. **91**(6): p. 697-702.

70. Allen, D.K., I.G. Libourel, and Y. Shachar-Hill, *Metabolic flux analysis in plants: coping with complexity.* Plant Cell Environ, 2009. **32**(9): p. 1241-57.

71. Schwender, J., F. Goffman, J.B. Ohlrogge, and Y. Shachar-Hill, *Rubisco without the Calvin cycle improves the carbon efficiency of developing green seeds.* Nature, 2004. **432**(7018): p. 779-82.

72. Sweetlove, L.J. and R.G. Ratcliffe, *Flux-balance modeling of plant metabolism.* Frontiers in Plant Science, 2011. **2**.

73. Varma, A. and B.O. Palsson, *Metabolic Flux Balancing - Basic Concepts, Scientific and Practical Use.* Bio-Technology, 1994. **12**(10): p. 994-998.

74. Edwards, J.S. and B.O. Palsson, *Systems properties of the Haemophilus influenzae Rd metabolic genotype.* Journal of Biological Chemistry, 1999. **274**(25): p. 17410-17416.

75. Heinrich, R. and S. Schuster, *The regulation of cellular systems*. 1996, New York: Chapman & Hall. xix, 372 p.

76. Gavalas, G.R., *Nonlinear differential equations of chemically reacting systems*. Springer tracts in natural philosophy, v 17. 1968, New York,. 106 p.

77. Palsson, B., *Systems biology : properties of reconstructed networks*. 2006, New York: Cambridge University Press. xii, 322 p.

78. Mahadevan, R. and C.H. Schilling, *The effects of alternate optimal solutions in constraint-based genome-scale metabolic models.* Metabolic Engineering, 2003. **5**(4): p. 264-276.

79. Schuster S, H.S., *On Elementary Flux Modes in Biochemical Reaction Systems At Steady State.* Journal of Biological Systems, 1994. **2**: p. 165-182.

80. Trinh, C.T., A. Wlaschin, and F. Srienc, *Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism.* Appl Microbiol Biotechnol, 2009. **81**(5): p. 813-26.

81. Kruger, N.J., S.K. Masakapalli, and R.G. Ratcliffe, *Strategies for investigating the plant metabolic network with steady-state metabolic flux analysis: lessons from an Arabidopsis cell culture and other systems.* J Exp Bot, 2012. **63**(6): p. 2309-23.

82. Kacser, H. and J.A. Burns, *The control of flux.* Symp Soc Exp Biol, 1973. **27**: p. 65-104.

83. Heinrich, R. and T.A. Rapoport, *A linear steady-state treatment of enzymatic chains. Critique of the crossover theorem and a general procedure to identify interaction sites with an effector.* Eur J Biochem, 1974. **42**(1): p. 97-105.

84. Heinrich, R. and T.A. Rapoport, *A linear steady-state treatment of enzymatic chains. General properties, control and effector strength.* Eur J Biochem, 1974. **42**(1): p. 89-95.

85. Fell, D.A., *Metabolic control analysis: a survey of its theoretical and experimental development.* Biochem J, 1992. **286 (Pt 2)**: p. 313-30.

86. Orth, J.D., I. Thiele, and B.O. Palsson, *What is flux balance analysis?* Nat Biotechnol, 2010. **28**(3): p. 245-8.

87. David Orlando Páez Melo, R.J.-P.M., Flavia Vischi Winck and Andrés Fernando González Barrios, *In Silico Analysis for Biomass Synthesis under Different CO2 Levels for Chlamydomonas reinhardtii Utilizing a Flux Balance Analysis Approach*, in *Advances in Intelligent Systems and Computing*, E. Pietka, Editor. 2014. p. 279-285.

88. Chang, R.L., L. Ghamsari, A. Manichaikul, E.F. Hom, S. Balaji, W. Fu, Y. Shen, T. Hao, B.O. Palsson, K. Salehi-Ashtiani, and J.A. Papin, *Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism.* Mol Syst Biol, 2011. **7**: p. 518.

89. Flassig, R.J., M. Fachet, K. Hoffner, P.I. Barton, and K. Sundmacher, *Dynamic flux balance modeling to increase the production of high-value compounds in green microalgae.* Biotechnol Biofuels, 2016. **9**: p. 165.

90. Sengupta, T., M. Bhushan, and P.P. Wangikar, *Metabolic modeling for multi-objective optimization of ethanol production in a Synechocystis mutant.* Photosynth Res, 2013. **118**(1-2): p. 155-65.

91. Villaverde, A.F., S. Bongard, K. Mauch, E. Balsa-Canto, and J.R. Banga, *Metabolic engineering with multi-objective optimization of kinetic models.* J Biotechnol, 2016. **222**: p. 1-8.

92. Barros, J., J.C. Serrani-Yarce, F. Chen, D. Baxter, B.J. Venables, and R.A. Dixon, *Role of bifunctional ammonia-lyase in grass cell wall biosynthesis.* Nat Plants, 2016. **2**(6): p. 16050.

93. Segre, D., D. Vitkup, and G.M. Church, *Analysis of optimality in natural and perturbed metabolic networks.* Proc Natl Acad Sci U S A, 2002. **99**(23): p. 15112-7.

94. Hay, J. and J. Schwender, *Metabolic network reconstruction and flux variability analysis of storage synthesis in developing oilseed rape (Brassica napus L.) embryos.* Plant J, 2011. **67**(3): p. 526-41.

95. Hay, J. and J. Schwender, *Computational analysis of storage synthesis in developing Brassica napus L. (oilseed rape) embryos: flux variability analysis in relation to (1)(3)C metabolic flux analysis.* Plant J, 2011. **67**(3): p. 513-25.

96. Steuer, R., A.N. Nesi, A.R. Fernie, T. Gross, B. Blasius, and J. Selbig, *From structure to dynamics of metabolic pathways: application to the plant mitochondrial TCA cycle.* Bioinformatics, 2007. **23**(11): p. 1378-1385.

97. Schuster, S., T. Dandekar, and D.A. Fell, *Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering.* Trends Biotechnol, 1999. **17**(2): p. 53-60.

98. Llaneras, F. and J. Pico, *Which metabolic pathways generate and characterize the flux space? A comparison among elementary modes, extreme pathways and minimal generators.* J Biomed Biotechnol, 2010. **2010**: p. 753904.

99. Sherry, A.D. and C.R. Malloy, *Integration of 13C Isotopomer Methods and Hyperpolarization Provides a Comprehensive Picture of Metabolism*, in *eMagRes*. 2007, John Wiley & Sons, Ltd.

100. Roscher, A., N.J. Kruger, and R.G. Ratcliffe, *Strategies for metabolic flux analysis in plants using isotope labelling.* J Biotechnol, 2000. **77**(1): p. 81-102.

101. Dieuaide-Noubhani, M. and A.P. Alonso, *Application of Metabolic Flux Analysis to Plants.* Plant Metabolic Flux Analysis: Methods and Protocols, 2014. **1090**: p. 1-17.

102. Moreno-Sanchez, R., E. Saavedra, S. Rodriguez-Enriquez, and V. Olin-Sandoval, *Metabolic control analysis: a tool for designing strategies to manipulate metabolic pathways.* J Biomed Biotechnol, 2008. **2008**: p. 597913.

103. Ap Rees, T. and S.A. Hill, *Metabolic control analysis of plant metabolism.* Plant, Cell & Environment, 1994. **17**(5): p. 587-599.

104. Giersch, C., D. Lämmel, and G. Farquhar, *Control analysis of photosynthetic CO2 fixation.* Photosynthesis Research, 1990. **24**(2): p. 151-165.

105. Waage, P. and C.M. Gulberg, *Studies concerning affinity*. Journal of Chemical Education, 1986. **63**(12): p. 1044.

106. Farré, G., S. Maiam Rivera, R. Alves, E. Vilaprinyo, A. Sorribas, R. Canela, S. Naqvi, G. Sandmann, T. Capell, C. Zhu, and P. Christou, *Targeted transcriptomic and metabolic profiling reveals temporal bottlenecks in the maize carotenoid pathway that may be addressed by multigene engineering.* The Plant Journal, 2013. **75**(3): p. 441-455.

107. Bai, C., S.M. Rivera, V. Medina, R. Alves, E. Vilaprinyo, A. Sorribas, R. Canela, T. Capell, G. Sandmann, P. Christou, and C. Zhu, *An in vitro system for the rapid functional characterization of genes involved in carotenoid biosynthesis and accumulation.* The Plant Journal, 2014. **77**(3): p. 464-475.

108. Michaelis, L. and M.L. Menten, *Die Kinetik der Invertinwirkung.* Biochemische Zeitschrift, 1913. **49**: p. 333-369.

109. Henri, V., *Lois générales de l'action des diastases*. 1903, Paris,: Librairie Scientifique A. Hermann. viii, 129 pages.

110. Cornish-Bowden, A., *One hundred years of Michaelis–Menten kinetics.* Perspectives in Science, 2015. **4**: p. 3-9.

111. Teusink, B., J. Passarge, C.A. Reijenga, E. Esgalhado, C.C. van der Weijden, M. Schepper, M.C. Walsh, B.M. Bakker, K. van Dam, H.V. Westerhoff, and J.L. Snoep, *Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry.* Eur J Biochem, 2000. **267**(17): p. 5313-29.

112. Schulz, A.R., *Enzyme kinetics : from diastase to multi-enzyme systems*. 1994, Cambridge ; New York: Cambridge University Press. x, 246 p.

113. Hill, A.V., *The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves.* J Physiol (Lond), 1910. **40**: p. 4-7.

114. Nag, A., M. Lunacek, P.A. Graf, and C.H. Chang, *Kinetic modeling and exploratory numerical simulation of chloroplastic starch degradation.* BMC Syst Biol, 2011. **5**: p. 94.

115. Cornish-Bowden, A., *Fundamentals of enzyme kinetics*. 3rd ed. 2004, London: Portland Press. xvi, 422 p.

116. Voit, E.O., *The best models of metabolism.* Wiley Interdiscip Rev Syst Biol Med, 2017.

117. Steuer, R., T. Gross, J. Selbig, and B. Blasius, *Structural kinetic modeling of metabolic networks.* Proc Natl Acad Sci U S A, 2006. **103**(32): p. 11868-73.

118. Goel, G., I.C. Chou, and E.O. Voit, *System estimation from metabolic time-series data.* Bioinformatics, 2008. **24**(21): p. 2505-11.

119. Faraji, M. and E.O. Voit, *Nonparametric dynamic modeling.* Math Biosci, 2016.

120. Faraji, M. and E.O. Voit, *Stepwise Inference of Likely Dynamic Flux Distributions from Metabolic Time Series Data.* Bioinformatics, 2017.

121. Savageau, M.A., *Biochemical systems analysis. II. The steady-state solutions for an n-pool system using a power-law approximation.* J Theor Biol, 1969. **25**(3): p. 370-9.

122. Savageau, M.A., *Biochemical systems analysis : a study of function and design in molecular biology*. 1976, Reading, Mass.: Addison-Wesley Pub. Co. Advanced Book Program. xvii, 379 pages.

123. Voit, E.O., *Biochemical Systems Theory: A Review*. ISRN Biomathematics, 2013. **2013**: p. 53.

124. Voit, E.O., *Dynamics of Self-Thinning Plant Stands.* Annals of Botany, 1988. **62**(1): p. 67-78.

125. Torsella, J. and A. Bin Razali, *An analysis of forestry data.* Canonical Nonlinear Modeling: S-System Approach to Understanding Complexity, 1991: p. 181-199.

126. Torres, N.V., *S-system modelling approach to ecosystem: application to a study of magnesium flow in a tropical forest.* Ecological modelling, 1996. **89**(1-3): p. 109-120.

127. Sands, P. and E. Voit, *Flux-based estimation of parameters in S-systems.* Ecological Modelling, 1996. **93**(1-3): p. 75-88.

128. Voit, E.O. and P.J. Sands, *Modeling forest growth II. Biomass partitioning in Scots pine.* Ecological modelling, 1996. **86**(1): p. 73-89.

129. Martin, P.-G., *The use of canonical S-system modelling for condensation of complex dynamic models.* Ecological modelling, 1997. **103**(1): p. 43-70.

130. Kaitaniemi, P., *A canonical model of tree resource allocation after defoliation and bud consumption.* Ecological modelling, 2000. **129**(2): p. 259-272.

131. Renton, M., P. Kaitaniemi, and J. Hanan, *Functional–structural plant modelling using a combination of architectural analysis, L-systems and a canonical model of function.* Ecological Modelling, 2005. **184**(2): p. 277-298.

132. Sorribas, A., B. Hernandez-Bermejo, E. Vilaprinyo, and R. Alves, *Cooperativity and saturation in biochemical networks: a saturable formalism using Taylor series approximations.* Biotechnol Bioeng, 2007. **97**(5): p. 1259-77.

133. Wu, L., W. Wang, W.A. van Winden, W.M. van Gulik, and J.J. Heijnen, *A new framework for the estimation of control parameters in metabolic pathways using lin-log kinetics.* European Journal of Biochemistry, 2004. **271**(16): p. 3348-3359.

134. Visser, D. and J.J. Heijnen, *Dynamic simulation and metabolic re-design of a branched pathway using linlog kinetics.* Metabolic Engineering, 2003. **5**(3): p. 164-176.

135. Heijnen, J.J., *Approximative kinetic formats used in metabolic network modeling.* Biotechnology and Bioengineering, 2005. **91**(5): p. 534-545.

136. del Rosario, R.C.H., E. Mendoza, and E.O. Voit, *Challenges in lin-log modelling of glycolysis in Lactococcus lactis.* Iet Systems Biology, 2008. **2**(3): p. 136-U30.

137. Wang, F.S., C.L. Ko, and E.O. Voit, *Kinetic modeling using S-systems and lin-log approaches.* Biochemical Engineering Journal, 2007. **33**(3): p. 238-247.

138. Chou, I.C. and E.O. Voit, *Estimation of dynamic flux profiles from metabolic time series data.* BMC Syst Biol, 2012. **6**: p. 84.

139. Dolatshahi, S. and E.O. Voit, *Identification of Metabolic Pathway Systems.* Front Genet, 2016. **7**: p. 6.

140. Iwata, M., F. Shiraishi, and E.O. Voit, *Coarse but efficient identification of metabolic pathway systems.* International Journal of Systems Biology, 2013. **4**(1): p. 57.

141. Voit, E.O., G. Goel, I.C. Chou, and L.L. Fonseca, *Estimation of metabolic pathway systems from different data sources.* IET Syst Biol, 2009. **3**(6): p. 513-22.

142. Hartmann, A. and F. Schreiber, *Integrative Analysis of Metabolic Models – from Structure to Dynamics.* Frontiers in Bioengineering and Biotechnology, 2015. **2**(91).

143. Wu, Q. and T. Tian, *Stochastic modeling of biochemical systems with multistep reactions using state-dependent time delay.* Sci Rep, 2016. **6**: p. 31909.

144. Yu, Y., W. Dong, C. Altimus, X. Tang, J. Griffith, M. Morello, L. Dudek, J. Arnold, and H.-B. Schüttler, *A genetic network for the clock of Neurospora crassa.* Proceedings of the National Academy of Sciences, 2007. **104**(8): p. 2809-2814.

145. Deng, Z., S. Arsenault, C. Caranica, J. Griffith, T. Zhu, A. Al-Omari, H.B. Schuttler, J. Arnold, and L. Mao, *Synchronizing stochastic circadian oscillators in single cells of Neurospora crassa.* Sci Rep, 2016. **6**: p. 35828.

146. Guerriero, M.L., O.E. Akman, and G. van Ooijen, *Stochastic models of cellular circadian rhythms in plants help to understand the impact of noise on robustness and clock structure.* Frontiers in Plant Science, 2014. **5**: p. 564.

147. Guerriero, M.L., A. Pokhilko, A.P. Fernández, K.J. Halliday, A.J. Millar, and J. Hillston, *Stochastic properties of the plant circadian clock.* Journal of the Royal Society Interface, 2012. **9**(69): p. 744-756.

148. Akman, O.E., F. Ciocchetta, A. Degasperi, and M.L. Guerriero, *Modelling Biological Clocks with Bio-PEPA: Stochasticity and Robustness for the Neurospora crassa Circadian Network*, in *Computational Methods in Systems Biology: 7th International Conference, CMSB 2009, Bologna, Italy, August 31-September 1, 2009. Proceedings*, P. Degano and R. Gorrieri, Editors. 2009, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 52-67.

149. Gonze, D., J. Halloy, and A. Goldbeter, *Deterministic Versus Stochastic Models for Circadian Rhythms.* Journal of Biological Physics, 2002. **28**(4): p. 637-653.

150. Sweetlove, L.J. and A.R. Fernie, *The Spatial Organization of Metabolism Within the Plant Cell.* Annual Review of Plant Biology, 2013. **64**(1): p. 723-746.

151. Grafahrend-Belau, E., A. Junker, A. Eschenroder, J. Muller, F. Schreiber, and B.H. Junker, *Multiscale metabolic modeling: dynamic flux balance analysis on a whole-plant scale.* Plant Physiol, 2013. **163**(2): p. 637-47.

152. Vanholme, R., I. Cesarino, K. Rataj, Y. Xiao, L. Sundin, G. Goeminne, H. Kim, J. Cross, K. Morreel, P. Araujo, L. Welsh, J. Haustraete, C. McClellan, B. Vanholme, J. Ralph, G.G. Simpson, C. Halpin, and W. Boerjan, *Caffeoyl shikimate esterase (CSE) is an enzyme in the lignin biosynthetic pathway in Arabidopsis.* Science, 2013. **341**(6150): p. 1103-6.

153. Savageau, M.A., *Chapter 5 Enzyme kinetics in vitro and in vivo: Michaelis-Menten revisited*, in *Principles of Medical Biology*, E.E. Bittar and N. Bittar, Editors. 1995, Elsevier. p. 93-146.

154. van Eunen, K. and B.M. Bakker, *The importance and challenges of in vivo-like enzyme kinetics.* Perspectives in Science, 2014. **1**(1): p. 126-130.

155. Albe, K.R., M.H. Butler, and B.E. Wright, *Cellular concentrations of enzymes and their substrates.* Journal of Theoretical Biology, 1990. **143**(2): p. 163-195.

156. van Eunen, K., J.A.L. Kiewiet, H.V. Westerhoff, and B.M. Bakker, *Testing Biochemistry Revisited: How In Vivo Metabolism Can Be Understood from In Vitro Enzyme Kinetics.* PLOS Computational Biology, 2012. **8**(4): p. e1002483.

157. Torres, N.V. and E.O. Voit, *Pathway analysis and optimization in metabolic engineering*. 2002, Cambridge, UK ; New York: Cambridge University Press. xiv, 305 p.

158. Voit, E.O., *Computational Analysis of Biochemical Systems : A Practical Guide for Biochemists and Molecular Biologists*. 2000, New York: Cambridge University Press. xii, 531 p.

159. Torres, N.V., E.O. Voit, C. Glez-Alcon, and F. Rodriguez, *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, 1997. **55**(5): p. 758-72.

160. Chapple, C., *Molecular-genetic analysis of plant cytochrome P450-dependent monooxygenases.* Annual Review of Plant Physiology and Plant Molecular Biology, 1998. **49**(1): p. 311-343.

161. Guo, D., F. Chen, and R.A. Dixon, *Monolignol biosynthesis in microsomal preparations from lignifying stems of alfalfa (Medicago sativa L.).* Phytochemistry, 2002. **61**(6): p. 657-667.

162. Tummler, K., T. Lubitz, M. Schelker, and E. Klipp, *New types of experimental data shape the use of enzyme kinetics for dynamic network modeling.* FEBS Journal, 2014. **281**(2): p. 549-571.

163. Hoffmann, L., S. Besseau, P. Geoffroy, C. Ritzenthaler, D. Meyer, C. Lapierre, B. Pollet, and M. Legrand, *Silencing of Hydroxycinnamoyl-Coenzyme A Shikimate/Quinate Hydroxycinnamoyltransferase Affects Phenylpropanoid Biosynthesis.* The Plant Cell, 2004. **16**(6): p. 1446.

164. Escamilla-Treviño, L.L., H. Shen, T. Hernandez, Y. Yin, Y. Xu, and R.A. Dixon, *Early lignin pathway enzymes and routes to chlorogenic acid in switchgrass (Panicum virgatum L.).* Plant Molecular Biology, 2014. **84**(4): p. 565-576.

165. Ralph, J., J.H. Grabber, and R.D. Hatfield, *Lignin-ferulate cross-links in grasses: active incorporation of ferulate polysaccharide esters into ryegrass lignins.* Carbohydrate Research, 1995. **275**(1): p. 167-178.

166. Shen, H., M. Mazarei, H. Hisano, L. Escamilla-Trevino, C. Fu, Y. Pu, M.R. Rudis, Y. Tang, X. Xiao, L. Jackson, G. Li, T. Hernandez, F. Chen, A.J. Ragauskas, C.N. Stewart, Z.-Y. Wang, and R.A. Dixon, *A Genomics Approach to Deciphering Lignin Biosynthesis in Switchgrass.* The Plant Cell, 2013. **25**(11): p. 4342.

167. Lin, C.-Y., Jack P. Wang, Q. Li, H.-C. Chen, J. Liu, P. Loziuk, J. Song, C. Williams, David C. Muddiman, Ronald R. Sederoff, and Vincent L. Chiang, *4-Coumaroyl and Caffeoyl Shikimic Acids Inhibit 4-Coumaric Acid:Coenzyme A Ligases and Modulate Metabolic Flux for 3-Hydroxylation in Monolignol Biosynthesis of Populus trichocarpa.* Molecular Plant. **8**(1): p. 176-187.

168. Shen, H., X. He, C.R. Poovaiah, W.A. Wuddineh, J. Ma, D.G. Mann, H. Wang, L. Jackson, Y. Tang, C.N. Stewart, Jr., F. Chen, and R.A. Dixon, *Functional characterization of the switchgrass (Panicum virgatum) R2R3-MYB transcription factor PvMYB4 for improvement of lignocellulosic feedstocks.* New Phytol, 2012. **193**(1): p. 121-36.

169. Davison, B.H., S.R. Drescher, G.A. Tuskan, M.F. Davis, and N.P. Nghiem, *Variation of S/G ratio and lignin content in a Populus family influences the release of xylose by dilute acid hydrolysis.* Applied Biochemistry and Biotechnology, 2006. **130**(1): p. 427-435.

170. Van Acker, R., R. Vanholme, V. Storme, J.C. Mortimer, P. Dupree, and W. Boerjan, *Lignin biosynthesis perturbations affect secondary cell wall composition and saccharification yield in Arabidopsis thaliana.* Biotechnology for Biofuels, 2013. **6**(1): p. 46.

171. Escamilla-Treviño, L.L., H. Shen, S.R. Uppalapati, T. Ray, Y. Tang, T. Hernandez, Y. Yin, Y. Xu, and R.A. Dixon, *Switchgrass (Panicum virgatum) possesses a divergent family of cinnamoyl CoA reductases with distinct biochemical properties.* New Phytologist, 2010. **185**(1): p. 143-155.

172. Voit, E.O., *A First course in systems biology*. 2012, New York: Garland Science; Taylor & Francis distributor.

173. Winkel, B.S.J., *Metabolic channeling in plants.* Annual Review of Plant Biology, 2004. **55**(1): p. 85-107.

174. Achnine, L., E.B. Blancaflor, S. Rasmussen, and R.A. Dixon, *Colocalization of L-Phenylalanine Ammonia-Lyase and Cinnamate 4-Hydroxylase for Metabolic*

Channeling in Phenylpropanoid Biosynthesis. The Plant Cell, 2004. **16**(11): p. 3098.

175. Rasmussen, S. and R.A. Dixon, *Transgene-Mediated and Elicitor-Induced Perturbation of Metabolic Channeling at the Entry Point into the Phenylpropanoid Pathway.* The Plant Cell, 1999. **11**(8): p. 1537.

176. Zhou, R., L. Jackson, G. Shadle, J. Nakashima, S. Temple, F. Chen, and R.A. Dixon, *Distinct cinnamoyl CoA reductases involved in parallel routes to lignin in Medicago truncatula.* Proc Natl Acad Sci U S A, 2010. **107**(41): p. 17803-8.

177. Bassard, J.-E., L. Richert, J. Geerinck, H. Renault, F. Duval, P. Ullmann, M. Schmitt, E. Meyer, J. Mutterer, W. Boerjan, G. De Jaeger, Y. Mely, A. Goossens, and D. Werck-Reichhart, *Protein–Protein and Protein–Membrane Associations in the Lignin Pathway.* The Plant Cell, 2012. **24**(11): p. 4465.

178. Chen, H.-C., J. Song, J.P. Wang, Y.-C. Lin, J. Ducoste, C.M. Shuford, J. Liu, Q. Li, R. Shi, A. Nepomuceno, F. Isik, D.C. Muddiman, C. Williams, R.R. Sederoff, and V.L. Chiang, *Systems Biology of Lignin Biosynthesis in Populus trichocarpa: Heteromeric 4-Coumaric Acid:Coenzyme A Ligase Protein Complex Formation, Regulation, and Numerical Modeling.* The Plant Cell, 2014. **26**(3): p. 876.

179. Savageau, M.A., *Critique of the enzymologist's test tube*, in *Fundamentals of Medical Cell Biology*, E.E. Bittar, Editor. 1992, JAI Press Inc.: Greenwich, CT. p. 45-108.

180. Dolatshahi, S., B. Vidakovic, and E.O. Voit, *A constrained wavelet smoother for pathway identification tasks in systems biology.* Computers & Chemical Engineering, 2014. **71**: p. 728-733.

181. Eilers, P.H.C., *A Perfect Smoother.* Analytical Chemistry, 2003. **75**(14): p. 3631-3636.

182. Seatzu, C., *A fitting based method for parameter estimation in S-systems.* 2000. **9**: p. 77-98.

183. Vilela, M., C.C.H. Borges, S. Vinga, A.T.R. Vasconcelos, H. Santos, E.O. Voit, and J.S. Almeida, *Automated smoother for the numerical decoupling of dynamics models.* BMC Bioinformatics, 2007. **8**(1): p. 305.

184. Whittaker, E.T., *On a New Method of Graduation.* Proceedings of the Edinburgh Mathematical Society, 1922. **41**: p. 63-75.

185. Voit, E.O., *What if the Fit is Unfit? Criteria for Biological Systems Estimation beyond Residual Errors*, in *Applied Statistics for Network Biology*. 2011, Wiley-VCH Verlag GmbH & Co. KGaA. p. 181-200.

186. Voit, E.O., *Characterizability of metabolic pathway systems from time series data.* Mathematical Biosciences, 2013. **246**(2): p. 315-325.

187. Albert, A.E., *Regression and the Moore-Penrose pseudoinverse*. Mathematics in Science and Engineering Vol. 94. 1972, New York: Academic Press.

188. Moore, E.H., *On the reciprocal of the general algebraic matrix.* Bulletin of the American Mathematical Society, 1920. **26**: p. 394–395.

189. Penrose, R., *A generalized inverse for matrices.* Mathematical Proceedings of the Cambridge Philosophical Society, 2008. **51**(3): p. 406-413.

190. Campbell, D. and R.J. Steele, *Smooth functional tempering for nonlinear differential equation models.* Statistics and Computing, 2012. **22**(2): p. 429-443.

191. Ramsay, J.O., G. Hooker, D. Campbell, and J. Cao, *Parameter estimation for differential equations: a generalized smoothing approach.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2007. **69**(5): p. 741-796.

192. Varah, J.M., *A Spline Least Squares Method for Numerical Parameter Estimation in Differential Equations.* SIAM Journal on Scientific and Statistical Computing, 1982. **3**(1): p. 28-46.

193. Voit, E.O. and J. Almeida, *Decoupling dynamical systems for pathway identification from metabolic profiles.* Bioinformatics, 2004. **20**(11): p. 1670-1681.

194. Voit, E.O. and M.A. Savageau, *Power-law approach to modeling biological systems; III. Methods of analysis.* J. Ferment. Technol . 1982(60): p. 223–241.

195. Voit, E.O. and M.A. Savageau, *Power-law approach to modeling biological systems; II. Application to ethanol production.* J. Ferment. Technol . 1982(60): p. 229–232.

196. Gumaa, K.A. and P. McLean, *The pentose phosphate pathway of glucose metabolism. Enzyme profiles and transient and steady-state content of intermediates of alternative pathways of glucose metabolism in Krebs ascites cells.* Biochemical Journal, 1969. **115**(5): p. 1009.

197. Loreck, D.J., J. Galarraga, J. Van der Feen, J.M. Phang, B.H. Smith, and C.J. Cummins, *Regulation of the pentose phosphate pathway in human astrocytes and gliomas.* Metabolic Brain Disease, 1987. **2**(1): p. 31-46.

198. Fonseca, L.L., C. Sanchez, H. Santos, and E.O. Voit, *Complex coordination of multi-scale cellular responses to environmental stress.* Molecular BioSystems, 2011. **7**(3): p. 731-741.

199. Cascante, M., R. Curto, and A. Sorribas, *Comparative characterization of the fermentation pathway of Saccharomyces cerevisiae using biochemical systems theory and metabolic control analysis: Steady-state analysis.* Mathematical Biosciences, 1995. **130**(1): p. 51-69.

200. Sorribas, A., R. Curto, and M. Cascante, *Comparative characterization of the fermentation pathway of Saccharomyces cerevisiae using biochemical systems theory and metabolic control analysis: Model validation and dynamic behavior.* Mathematical Biosciences, 1995. **130**(1): p. 71-84.

201. Bruggner, R.V., B. Bodenmiller, D.L. Dill, R.J. Tibshirani, and G.P. Nolan, *Automated identification of stratifying signatures in cellular subpopulations.* Proceedings of the National Academy of Sciences, 2014. **111**(26): p. E2770.

202. El-Aneed, A., A. Cohen, and J. Banoub, *Mass Spectrometry, Review of the Basics: Electrospray, MALDI, and Commonly Used Mass Analyzers.* Applied Spectroscopy Reviews, 2009. **44**(3): p. 210-230.

203. Li, S., Y. Park, S. Duraisingham, F.H. Strobel, N. Khan, Q.A. Soltow, D.P. Jones, and B. Pulendran, *Predicting Network Activity from High Throughput Metabolomics.* PLoS Computational Biology, 2013. **9**(7): p. e1003123.

204. Neves, A.R., W.A. Pool, J. Kok, O.P. Kuipers, and H. Santos, *Overview on sugar metabolism and its control in Lactococcus lactis - the input from in vivo NMR.* FEMS Microbiol Rev, 2005. **29**(3): p. 531-54.

205. Voit, E.O., *Modelling metabolic networks using power-laws and S-systems.* Essays In Biochemistry, 2008. **45**: p. 29.

206. Beaulieu, J.-M. and R.R. Gainetdinov, *The Physiology, Signaling, and Pharmacology of Dopamine Receptors.* Pharmacological Reviews, 2011. **63**(1): p. 182.

207. Qi, Z., G.W. Miller, and E.O. Voit, *Computational Systems Analysis of Dopamine Metabolism.* PLOS ONE, 2008. **3**(6): p. e2444.

208. Qi, Z., G.W. Miller, and E.O. Voit, *The internal state of medium spiny neurons varies in response to different input signals.* BMC Systems Biology, 2010. **4**(1): p. 26.

209. Surmeier, D.J., S.M. Graves, and W. Shen, *Dopaminergic modulation of striatal networks in health and Parkinson's disease.* Current Opinion in Neurobiology, 2014. **29**: p. 109-117.

210. Shiraishi, F. and M.A. Savageau, *The tricarboxylic acid cycle in Dictyostelium discoideum. I. Formulation of alternative kinetic representations.* Journal of Biological Chemistry, 1992. **267**(32): p. 22912-22918.

211. Voit, E.O., H.A. Martens, and S.W. Omholt, *150 Years of the Mass Action Law.* PLOS Computational Biology, 2015. **11**(1): p. e1004012.

212. Cascante, M., A. Sorribas, R. Franco, and E.I. Canela, *Biochemical systems theory: Increasing predictive power by using second-order derivatives measurements.* Journal of Theoretical Biology, 1991. **149**(4): p. 521-535.

213. Savageau, M.A., *Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions.* J Theor Biol, 1969. **25**(3): p. 365-9.

214. Voit, E.O.e., *Canonical Nonlinear Modeling. S-System Approach to Understanding Complexity*. 1991, NY: Van Nostrand Reinhold.

215. Fell, D., *Understanding the control of metabolism.* Trends in Biochemical Sciences, 1997. **22**(6): p. 231-232.

216. Visser, D. and J.J. Heijnen, *The Mathematics of Metabolic Control Analysis Revisited.* Metabolic Engineering, 2002. **4**(2): p. 114-123.

217. Savageau, M.A., E.O. Voit, and D.H. Irvine, *Biochemical systems theory and metabolic control theory: 1. fundamental similarities and differences.* Mathematical Biosciences, 1987. **86**(2): p. 127-145.

218. Savageau, M.A., E.O. Voit, and D.H. Irvine, *Biochemical systems theory and metabolic control theory: 2. the role of summation and connectivity relationships.* Mathematical Biosciences, 1987. **86**(2): p. 147-169.

219. Sorribas, A., E. Vilaprinyo, and R. Alves, *Approximate kinetic formalisms for modeling metabolic networks: does anything work?* 2008.

220. Chou, I.C. and E.O. Voit, *Recent developments in parameter estimation and structure identification of biochemical and genomic systems.* Mathematical Biosciences, 2009. **219**(2): p. 57-83.

221. Gennemark, P. and D. Wedelin, *Benchmarks for identification of ordinary differential equations from time series data.* Bioinformatics, 2009. **25**(6): p. 780-786.

222. Gennemark, P. and D. Wedelin, *Efficient algorithms for ordinary differential equation model identification of biological systems.* IET Systems Biology, 2007. **1**(2): p. 120-129.

223. Dolatshahi, S., L.L. Fonseca, and E.O. Voit, *New insights into the complex regulation of the glycolytic pathway in Lactococcus lactis. II. Inference of the precisely timed control system regulating glycolysis.* Molecular BioSystems, 2015. **12**(1): p. 37-47.

224. Dolatshahi, S., L.L. Fonseca, and E.O. Voit, *New insights into the complex regulation of the glycolytic pathway in Lactococcus lactis. I. Construction and diagnosis of a comprehensive dynamic model.* Molecular BioSystems, 2016. **12**(1): p. 23-36.

225. Galazzo, J.L. and J.E. Bailey, *Fermentation pathway kinetics and metabolic flux control in suspended and immobilized Saccharomyces cerevisiae.* Enzyme and Microbial Technology, 1990. **12**(3): p. 162-172.

226. Bailey, J.E., S. Birnbaum, J.L. Galazzo, C. Khosla, and J.V. Shanks, *Strategies and Challenges in Metabolic Engineeringa*. Annals of the New York Academy of Sciences, 1990. **589**(1): p. 1-15.

227. Curto, R., A. Sorribas, and M. Cascante, *Comparative characterization of the fermentation pathway of Saccharomyces cerevisiae using biochemical systems theory and metabolic control analysis: Model definition and nomenclature.* Mathematical Biosciences, 1995. **130**(1): p. 25-50.

228. Sorribas, A., C. Pozo, E. Vilaprinyo, G. Guillén-Gosálbez, L. Jiménez, and R. Alves, *Optimization and evolution in metabolic pathways: Global optimization techniques in Generalized Mass Action models.* Journal of Biotechnology, 2010. **149**(3): p. 141-153.

229. Polisetty, P.K., E.P. Gatzke, and E.O. Voit, *Yield optimization of regulated metabolic systems using deterministic branch-and-reduce methods.* Biotechnology and Bioengineering, 2008. **99**(5): p. 1154-1169.

230. Polisetty, P.K., E.O. Voit, and E.P. Gatzke, *Identification of metabolic system parameters using global optimization methods.* Theoretical Biology and Medical Modelling, 2006. **3**(1): p. 4.

231. Vera, J., P. De Atauri, M. Cascante, and N.V. Torres, *Multicriteria optimization of biochemical systems by linear programming: Application to production of ethanol by Saccharomyces cerevisiae.* Biotechnology and Bioengineering, 2003. **83**(3): p. 335-343.

232. Voit, E.O. and T. Radivoyevitch, *Biochemical systems analysis of genome-wide expression data.* Bioinformatics, 2000. **16**(11): p. 1023-1037.

233. Goncalves, T. and M.C. Loureiro-Dias, *Aspects of glucose uptake in Saccharomyces cerevisiae.* Journal of Bacteriology, 1994. **176**(5): p. 1511-1513.

234. Edelstein-Keshet, L., *Mathematical Models in Biology*. Classics in Applied Mathematics. 2005: Society for Industrial and Applied Mathematics. 615.

235. Kvam, P.H. and B. Vidakovic, *Nonparametric Statistics with Applications to Science and Engineering (Wiley Series in Probability and Statistics)*. 2007: Wiley-Interscience.

236. Wolfowitz, J., *Additive Partition Functions and a Class of Statistical Hypotheses.* Ann. Math. Statist., 1942. **13**(3): p. 247-279.

237. Dudewicz, E.J., *Nonparametric Methods: The History, the Reality, and the Future (with Special Reference to Statistical Selection Problems)*, in *Contributions to Stochastics: In Honour of the 75th Birthday of Walther Eberl, Sr.*, W. Sendler, Editor. 1987, Physica-Verlag HD: Heidelberg. p. 63-83.

238. Ury, H., *Letter to the editor.* Am. Stat., 1967. **21**: p. 53.

239. Noether, G., *Needed-A New Name (Letter to the Editor).* The American Statistician, 1967. **21**: p. 41.

240. Dallal, G.E. *Nonparametric Statistics*. 2000; Available from: http://www.jerrydallal.com/lhsp/npar.htm.

241. Hoskin, T. *Parametric and Nonparametric: Demystifying the Terms*. 2016; Available from: www.mayo.edu/mayo-edu-docs/center-for-translational-science-activities-documents/berd-5-6.pdf

242. Oda, Y., T. Mimura, and S. Hasezawa, *Regulation of Secondary Cell Wall Development by Cortical Microtubules during Tracheary Element Differentiation in Arabidopsis Cell Suspensions.* Plant Physiology, 2005. **137**(3): p. 1027.

243. Christiernin, M., A.B. Ohlsson, T. Berglund, and G. Henriksson, *Lignin isolated from primary walls of hybrid aspen cell cultures indicates significant differences in lignin structure between primary and secondary cell wall.* Plant Physiology and Biochemistry, 2005. **43**(8): p. 777-785.

244. Curto, R., E.O. Voit, A. Sorribas, and M. Cascante, *Mathematical models of purine metabolism in man.* Math Biosci, 1998. **151**(1): p. 1-49.