Parameter-Component Dependency: Identifying the Biological Functions of Interchangeable Genetic Components

Seunghee Shelly Jang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Eric Klavins, Chair

Georg Seelig

James Carothers

Program Authorized to Offer Degree:
Electrical Engineering

University of Washington

Abstract

Parameter-Component Dependency:
Identifying the Biological Functions of Interchangeable Genetic Components

Seunghee Shelly Jang

Chair of the Supervisory Committee:
Associate Professor Eric Klavins
Department of Electrical Engineering

Synthetic biology can benefit from characterization and analysis of biological components that enable simulation and engineering of large scale networks with complex behavior. In this thesis, we introduce the Parameter Component Dependency (PCD) matrix, a characterization and analysis framework that enables users to quantify the biological functions of interchangeable genetic components, using datasets generated by combinatorial libraries composed of multiple components. We use two synthetic auxin signaling pathways to demonstrate that PCD matrices represent hypotheses about dependencies of model parameters to components. Using the PCD framework, we discriminated and verified multiple such hypotheses systematically and gained mechanistic insights into synthetic auxin signaling. We also present a case study of a synthetic biological system to demonstrate that the PCD framework can be used to analyze systems with little *a priori* information. By systematically searching through the PCD matrices, we showed that the dependency relationships used to simulate the dataset is recovered exactly.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

C_i	finite set of the i-th component variants
$(C_1^{(j_1)}, C_2^{(j_2)}, \cdots, C_n^{(j_n)})$	n-tuple representation of $S^{(j_1, j_2, \cdots, j_n)}$
$\mathbf{D}$	set of experimental data collected from $\mathbf{S}$
$D^{(j_1, j_2, \cdots, j_n)}$	data corresponding to the system variant $S^{(j_1, j_2, \cdots, j_n)}$
$f(\theta)$	model of the system S with parameter vector θ
h	hypothesis index
η_i	number of variants in the i-th component family
i	component index
j_i	component variant index
$k_\kappa^{(j_1, j_2, \cdots, j_n)}$	κ-th element of $\theta^{(j_1, j_2, \cdots, j_n)}$
m	length of the parameter vector θ
M	Parameter Component Dependency matrix
$\uparrow M$	Up-set of element M
$\downarrow M$	Down-set of element M
M_h	h-th M
$M_h(\kappa, i)$	entry of M_h at κ-th row and i-th column
n	number of components in a system

$nnz(M)$	number of non-zero entries in M
$S^{(j_1,j_2,\cdots,j_n)}$	a system composed of j_1-th variant of the first component, j_2-th variant of the second component, $\ldots$ j_n-th variant of the n-th component
$\mathbf{S}$	set of all $\prod_i^n \eta_i$ variants of the systems, S
$\mathscr{P}(X)$	Powerset of X
$\theta^{(j_1,j_2,\cdots,j_n)}$	parameter vector corresponding to the system variant $S^{(j_1,j_2,\cdots,j_n)}$
Θ	set of parameter vectors corresponding to $\mathbf{S}$
$X^{(m,n)}$	set of all positions in M of size m by n

ACKNOWLEDGMENTS

I thank my thesis advisor, Dr. Eric Klavins, who encouraged me to pursue the problems that deeply resonated with my interest. When that interest sometimes took me to unfamiliar territory, he was a willing and valuable collaborator. His ability to nurture his student's confidence as researchers and individual thinkers has been an irreplaceable positive influence in my graduate career. I also thank the members of the Klavins lab who blazed the path before me, Josh Bishop, Fayette Shaw, Rob Egbert and Kevin Oishi. Their commitment to their academic journey set fine examples for me to follow. Chris Takahashi, my cohort, deserves a special recognition for sharing his experiences with me and building priceless camaraderie to remind me that I am not alone. I owe much of my past five year's learning experience to Dr. David Thorsley, as it was his invitation to the lab that began my research in the Klavins lab. His words of encouragement and expertise inspired me throughout my research career.

I thank my committee members, Professors Georg Seelig, James Carothers and Maitreya Dunham, for sharing their invaluable views with me so that I may become a better thinker and researcher. Professor Radha Poovendran, for lack of a better word, adopted me and mentored me throughout my PhD path and for that I cannot thank him enough. His confidence in my ability gave me the courage to press on during challenging times.

The members of the auxin collaboration project, Professor Jennifer Nemhauser, Kyle Havens, Edith Pierre-Jerome, Jessica Guseman and Britney Moss, were integral in sparking the inspiration

in me to develop the analytical framework presented in this thesis. Their discussions widened my horizons to include scientific questions that go beyond those from engineering disciplines.

I'd like to thank my very dear friends, Jackie, Kathy and Roger, who never let me forget that I was loved and supported. I could not have done it without them. And finally, I thank my parents who taught me the value of hard work, my grandparents who support all that I do, my brother who loves me dearly, and my husband who quietly shows me everyday what true love is.

DEDICATION

To my family for their constant love.

To my husband, Tomoki.

You are my rock.

Chapter 1

INTRODUCTION

1.1 Motivation

Synthesis of complex behavior in engineered biological systems is, in part, facilitated by the characterization, standardization and modularity of constituent components. The provenance of this observation coincides with that of the consideration of living systems as being "programmable manufacturing systems" [1]. This inspired engineers to abstract away the details of molecular biology, such as DNA sequences and protein interactions, to pose biological networks as systems appropriate for the application of engineering solutions. One of the benefits of quantitative characterization and simulations of complex systems is that they enable high level conceptual design of system behaviors with a complexity that exceeds human intuition. Furthermore, simulation analysis allows engineers to test the limits of particular designs and analyze systems without physically having to build them. This results in overall reduction in the cost of the design-build-test cycle. Yet, the efforts in applying engineering principles to biological systems face many obstacles.

One such obstacle is the lack of clarity about what makes a component *engineerable*, a component being a recurring motif such as a transcription factor or promoter. Though knowing the compositional details of a biological component, such as its DNA sequence, is the first step in characterizing it, the information is not sufficient to conduct simulations analysis and apply engineering analysis methods. For a component to have practical utility and be engineerable in synthetic biology, it needs to be an encapsulation of one or more biological functions. Take electronics systems as an

example and imagine carbon and clay cylinders designed as resistors in electronic circuits. These components are useful for engineering circuit behaviors in that they serve the function of opposing electric current. In other words, the identified electrical function of resistors are their respective resistances. In addition, the identified function of a component not only needs to be descriptive ("opposing electric current"), but quantitative ("the resistance of this resistor is 1 ohm") to fully enable informative simulation. Once such information is obtained, it can be incorporated into mathematical representations of various system designs (structure or connectivity of the component in conjunction with other components) to simulate and tune the behaviors of the systems. Therefore, to fully elevate the utility of biological components to the level observed in other engineering fields, similar characterization approaches, such as quantitatively identifying the context relevant functions of components, are critical.

Often a biological component does not exist as a singular entity, but as a family of multiple mechanistically homogeneous members that are small variations of one another. Thus, multiple versions of a system, differing only by the identity of the specific component within, comprise a suite of system variants where a range of similar, but quantitatively varied behaviors are observed. There are many examples of naturally occurring components, such as the proteins found in the auxin signal pathways [2] and DNA sequences that serve as bacterial promoters [3]. Additionally, with advances in molecular biology, families of synthetic biological components are being developed as well [4, 5, 6].

There are two benefits to identifying the functions of a family of components in ensemble. First, identification reveals the range of this family's function that, when applied to the analysis of new systems built containing a member of the family, defines the limits of the system behavior.

Second, it reveals the rank order among the family members with respect to the identified functions, giving insights into which member ought to be substituted to tune the system behavior in the desired direction. Therefore, characterizing different biological components adds increased engineering and tuning capabilities to the synthetic biology toolkit.

Interestingly, a component existing as a multi-membered family may be critical to identifying the function of the component. One way to identify a component's function is to conduct a set of controlled experiments where all experimental variables are held constant while only the component of interest is varied. By analyzing the measured behavior of the suite of experiments for divergent characteristics, any significant changes in the behavior across the suite are hypothesized to originate from varying the component identity. Using this controlled approach, it is somewhat straightforward to determine whether varying a component has a significant effect on the overall behavior of the system. However, imagine the inverse of this process in simulation studies. To recreate the same varying behaviors in simulation, not only the structure of its corresponding model (i.e. equations) is needed, but information on which model parameters simulate the same change observed when the component is substituted. Said differently, we are interested in extrapolating the information of 'substituting this component causes the system to behave differently in this manner' to 'changing these parameters in simulation results in the same change in system behavior as substituting this component'. In this thesis, we consider the ability of a model parameter to simulate the same change in systems behavior caused by a component substitution as a 'dependency' relationship of the parameter to the component. Then we demonstrate that such dependent parameters of components are representative of the biological functions of the components.

The pursuit of identifying the dependency relationship of parameters to components is more

interesting when considering a suite of systems created by varying multiple families simultaneously. These suites are called combinatorial libraries, and each system in a library is a unique combination of multiple component variants joined in a cross-product manner. Imagine the following scenario where a system is designed with two distinct components: through a simulation analysis, a single model parameter was shown to significantly affect the system behavior. Engineers wishing to tune the system to match some specification would then benefit from knowing the dependent relationship of this parameter to either component, because there are fewer system variants to build – only the variants that differ by the component that the parameter is dependent on – instead of building and surveying the entire combinatorial library.

Through advances in DNA synthesis technology, the cost of the design-build-test cycles in synthetic biology continues to decrease. This trend hints that engineering approaches where a large number of system variants are generated by randomly introducing variations across the circuit is a feasible option, and possibly reducess the need for developing engineering approaches that enable rational design and deliberate tuning. But considering the continued progress of synthetic biology research and its growing areas of application that demand ever increasing engineered complexity of the field, it must be that the two approaches complement each other. Furthermore, perhaps driven by the decreasing cost of synthesis, there are a number of examples of combinatorial libraries of components in synthetic biology. Most of these systems are built to characterize, analyze and determine the engineerability of specific biological components. Nevertheless, we acknowledge the significance of this increasing trend in the number of combinatorial libraries as an indicator for the need for streamlined and cost-effective analytical framework applicable to combinatorial libraries. It is towards such a framework that this thesis is addressed.

1.2 Specific Contributions

In this thesis, we introduce a characterization and analysis framework that elucidates core functionalities of biological components, to elevate their engineerability and utility in synthetic biology. The framework is specifically applicable to combinatorial libraries built by combining multiple families of biological components in a cross-product manner. By revealing the quantitative rank orders within component families, it aids in the process of conceptual designs and tuning of systems composed of the characterized components. The analytical results have practical applications in both systems and synthetic biology. In systems biology, where practitioners probe existing systems with the focus on understanding the underlying mechanisms, model parameters represent specific system characteristics such as reaction rates. Thus, identifying a parameter's dependency on components is equivalent to identifying the source of variations in a specific characteristic to the same component. On the other hand, the same dependency relationship is beneficial in the construction-driven analytical approaches of synthetic biology in that it reduces the cost of engineering and tuning system behaviors by identifying optimal strategies for which subset of system variations to construct and test.

1.3 Overview

In Chapter 2, we review the motivation behind the need for developing biological components in synthetic biology, followed by a discussion on tuning strategies employed in synthetic biology. Additionally, a few relevant examples of synthetic combinatorial libraries and their major findings are discussed. Then, we discuss the black box modeling philosophy that complements the traditional mass-action kinetics oriented approach by adopting an agnostic view of the inner mechanisms of

systems of interest.

In Chapter 3, we introduce the Parameter Component Dependency (PCD) matrix for combinatorial libraries of biological systems and their models. We show that each PCD matrix is a mathematical representation of a hypothesis regarding the dependency of parameters to components, and for a given system and model pair, there exists a finite number of parameter-component dependency hypotheses, and thus a finite number of PCD matrices. We then present a metric for evaluating these hypotheses by computing the ability of each PCD matrix, corresponding to a hypotheses, to fit the model to the measured dataset. This is enabled by generating appropriate constraints of parameters in optimization problems that estimate the parameter values. We further discuss that systematic evaluation of PCD matrices is aided by representing the set of candidate of PCDs as a partially ordered set [7]. The final section of the chapter discusses the double-layer optimizations that make up the PCD framework.

In Chapter 4, we use a synthetic biological system to demonstrate the analytical workflow using the PCD framework. Using the relatively small system and simulated data, we conduct an exhaustive search of all possible hypotheses to show their varying performances, and show that the 'true' dependency relationship (used to simulate the dataset) can be fully recovered. We also present a greedy approach that reduces the cost of identifying the true relationship, along with the caveats of this approach and some heuristic solutions. To verify the framework against another method similar in its objective, we discuss a feature selection method called the Sparse Group Lasso (SGL), and present a comparative study of the PCD and the SGL. The chapter closes with a discussion on the utility of systematic PCD optimization as means to illuminate the unknown effects of changing a component to the system behavior that occur between components of combinatorial libraries.

In Chapters 5 and 6, we apply the PCD framework to two synthetic auxin signaling pathways engineered in *Saccharomyces cerevisiae* [8, 9]. Each chapter is prefaced with a brief introduction to the edited version of the corresponding article. In Chapter 5, a combinatorial library consisting of two of the three primary auxin pathway components is introduced. Using *a priori* knowledge of the pathway proteins and the deliberate design and construction of the synthetic pathway, a number of feasible hypothesis as well as some counterintuitive hypotheses are proposed and discriminated using the PCD framework. We show that the result allows us to identify a small number of model parameters that represent the synthetic pathway's core biological functions – the synthesis rate and auxin-mediated degradation rate of one of the components. Additionally, as the major outcome of the PCD framework is its ability to appropriate variations in systems behavior to one or more components, we show that a well-used statistical method, Analysis of Variance (ANOVA), illuminates a similar relationship, albeit under some specific assumptions regarding the parameter estimates and the linearity of the components contribution to behavior variations. In Chapter 6, a larger synthetic auxin signaling pathway, built by adding a third component to the one discussed in Chapter 5, is introduced. As in the previous chapter, the PCD framework is applied to achieve the same objective of verifying hypotheses regarding the dependency relationships between model parameters and system components. The resulting analysis verifies the findings from Chapter 5, that the auxin-mediated degradation rate is one of the core biological functions of the two-component pathway. Furthermore, as a larger system with a model with more parameters, this case-study highlights the existence of multiple competing hypotheses that cannot be distinguished nor discriminated without further experimentation. Thus, we discuss the possible utility of the PCD framework as the basis for developing an experimental design workflow.

Chapter 7 concludes the thesis with the summary of each chapter, discussions on outstanding questions, and suggestions for future projects that address these questions. In Appendix A, an overview of the Mathematica package written for the analysis discussed in Chapters 5 and 6, and a brief discussion of the workflow using the package are presented. In Appendix B, the R codes used for the Sparse Group Lasso and ANOVA analyses discussed in Chapters 4 and 5, respectively, are presented. In Appendix C, we delve into the recurrent observation made in each chapter, a set of indistinguishable PCD matrices, what we call an equivalency class of PCD matrices. We show that an equivalency class is a form of competing hypotheses and present a preliminary analysis on a real example of combinatorial library using a linear model [10].

Chapter 2

BACKGROUND

Early works of synthetic biology consisted of proof of concepts that demonstrated that naturally occurring small biological components can be repurposed – combined in new ways to give rise to behaviors that are different from their contexts. For instance, a ring oscillator was constructed by connecting three pairs of repressing transcription factors and promoters found in *E. coli* and λ phage [11]; this example demonstrated that it is possible to program cells with artificial behaviors. Similarly repurposing biological components from different origins, mutually repressing pairs of repressors and promoters gave rise to a toggle-switch behavior [12]; this example inspired a number of synthetic biological systems demonstrating that cellular states can be manipulated to mimic memory storage devices [13, 14, 15]. These examples signify the start of the synthesis based approach to understanding biology. In [16], Carlson states that "Only when we can build a system that quantitatively behaves as predicted should we say we understand (biology)", capturing the principles of synthetic biology. Thus, by building relatively simple circuits that mimic the complex behaviors in nature, engineers aim to uncover the core networks that give rise to these behaviors that are obstructed by various cumbersome evolutionary legacies as they exist currently in nature.

There are now even more examples of complex synthetic biology designs such as layered logic gates [17], synchronized behaviors in colony growth [18], pattern formations [19], and pulse generators [20]. The vast variety of synthetic designs indicates that the applications and resulting benefits of synthetic biology are only limited by our imagination. However, lack of fundamental

understanding of biological operations presents a challenge to this progress. In an article titled "Five Hard Truths for Synthetic Biology" [21], obstacles such as lack of clarity in what defines a part, unpredictable circuit behavior, unwieldy complexity, incompatible parts and lack of robustness to variability in system behavior are discussed. This discussion is not isolated, as others have brought forth similar observations in recent years [22, 23]. In the following sections, we discuss literature that focus on the standardization and characterization of biological parts, or components, and examples of analysis and experimental designs that utilize combinatorial libraries. Then, we present a strategy to address the main problem of insufficient characterization in engineering complex systems: tuning. Tuning is an integral process to engineering synthetic biological systems, as even the most carefully characterized parts tend to succumb to unpredictable emergent behaviors upon compositions with other parts. The chapter closes with a brief discussion on black-box modeling examples in synthetic biology, along with their advantages and disadvantages compared to models identified using first principles, such as mass action kinetics.

2.1 Components Characterization

Engineering sufficiently complex systems with tunable behaviors requires modular constituent components. Without them, the resulting system is a one-of-a-kind construction that requires multiple costly design-build-test cycles to satisfy its specifications. Currently, there is no explicit consensus of what is required to make a component modular, nor are there standardized approaches to characterizing components. However, strategies to improve or engineer component characteristics that increase modularity such as orthogonality (limited interference and interactions with hosts' native functionalities or other components of similar construction), reliability (equal levels of variations regardless of contexts in which they perform), tunability (malleable and allow a range

of system outputs) and composability (can be interconnected to build larger circuits with increased complexity) are being actively pursued [24]. One successful strategy for optimizing modularity in components is to borrow solutions from the field of controls engineering. For example, it is well known that a negative feedback structure is effective at disturbance rejection [25]. Since synthetic circuits built with open-loop structures were shown to be susceptible to stochastic fluctuations found in most cellular environments [26, 27], researchers hypothesized that they can increase the circuits' robustness by engineering negative feedback into the design [28, 29, 30]. Though the idea of posing biological systems as well-defined networks is attractive, given the high complexity found even in the smallest of biological system there are no standardized methods to how these systems can be structured to make the processes of engineering and tuning amenable to theoretical analysis as in other fields of engineering and science. Nevertheless, there are examples that demonstrate the successful outcome of such an approach. For instance, Del Vecchio *et al.* applied a mixture of control theory, mathematical biology and experimental techniques to characterize the unwanted effects of interconnection on the input-output relationship, called retroactivity [31, 32, 33, 34]. By casting transcriptional networks as input-output systems, the authors measure retroactivity as a function of measurable parameters. This allowed them to identify a set of strategies that minimize retroactivity (e.g. insulator devices), and develop concrete *in vivo* solutions. This is an exciting indication that the field is converging on developing streamlined engineering processes that take practitioners from identifying and characterizing components from nature, designing networks that match desired specifications, and constructing and tuning these networks in living systems [35, 36].

2.2 Combinatorial Libraries

Combinatorial libraries are constructed by combining multiple components from different families in a cross-product manner. It often leads to large numbers of combinations, or system variants, and is an effective optimization platform when coupled with high-throughput screening strategies [37, 38, 39]. There are some examples of using combinatorial libraries in synthetic biology to characterize and elucidate general rules to be used for engineering desired behavior. In [40], combinations of diverse promoter regions – divided into three segments depending on their respective positions to the transcription start sites – were analyzed to better understand the relationship between promoter function and architecture. The authors asked whether a set of general rules can be identified from the large scale library and proposed mathematical models and metrics to score individual performance. In [10], the authors addressed a similar objective of characterizing gene expression by constructing a combinatorial library of multiple genes and synthetically modified promoters. Using ANOVA to score each synthetic sequence's ability to result in consistent protein synthesis rate independent of the sequence, the authors identified general DNA sequences with high robustness to the changing downstream gene [10]. In [4], by combining both synthetic and naturally occurring transcription factors with many variations of bacterial promoters, the authors identified pairs of transcription factors and promoters that exhibit high specificity. Finally, in an attempt to decipher the complexity of chromatin regulation, Keung, *et al.* engineered combinatorial, spatial, and temporal patterns of individual chromatin regulators [41]. In all these examples, because the objectives of analyses were to survey the behavior of entire families of components, comprehensive combinatorial libraries that consider nearly all combinations of components were built for analysis and characterization, which tend to be labor intensive. However, if the objective of analysis was

more synthesis-oriented, as in seeking the optimal combination that matches some design specification, we may be abel to reduce the labor cost. For instance, an iterative approach can be implemented to reduce the synthesis, screening and analysis cost. Given some design specification, a smaller subset of combinations can be built and evaluated to reveal some variants that are closer to the desired specification. Those can then be used as benchmarks to build the next set of combinatorial libraries. This iterative strategy can be continued until the desired system is identified. Though this approach is conceivably a cheaper alternative to a comprehensive search, it requires that some rank order information regarding each family is known.

2.3 Tuning Strategies

Even with extensive characterizations of individual parts, the resulting systems behave differently than intended. Therefore, aside from parts characterization, an additional process of engineering circuit behavior using both coarse- and fine-grained modification to match the specification (*tuning*) is required. Tuning gene circuits has been identified as an integral process in engineering synthetic circuits [42], and there are many strategies, such as RBS tuning [43, 44, 45], directed evolution [46], high-throughput cloning [47], and promoter engineering [40]. However, with numerous tuning strategies and modifiable loci in the gene regulatory networks available, engineers are presented with the paradox of choice. For example, consider a simple gene expression cassette with a single promoter driving the expression of a gene. Even in this minimal system, there are multiple tuning knobs such as the promoter, RBS, or coding sequence. Considering that DNA recombination and synthesis still carry non-negligible price tags, and the relationship between tuning knobs and the downstream behavior may not be reliably predictive, it is prohibitively expensive to test all possible combinations of settings at each tuning knob. Therefore, the need for identifying tuning knobs

that result in the largest significant variations across the performance space is pressing. Such an approach allows engineers to effectively survey the possible behaviors using fewer combinations, identify the locus of variations that give rise to output performance that most closely resembles the desired performance, and construct another combinatorial library that surveys smaller portion of the performance landscape to pinpoint the optimal circuit [48, 49, 50, 51].

2.4 Model Identification

A well-used approach of model identification used for genetic regulatory networks is the identification of many known species and their interactions, then deriving ordinary differential equations models using first principles. Models identified using this approach often have large numbers of equations and parameters, which pose significant challenges during the parameter estimation process. This is because without an extensive measurements of the system behavior, users tend to be burdened with large estimation uncertainty [52, 53]. One way to address this problem is to reduce the complexity of a model through analytical and heuristic methods such as time-scale separation [54] or assumptions that some species have negligible contributions to the key dynamics of the network. Alternatively, using the black-box model approach, we can abstract away unverifiable details to build a model from the bottom up instead. Using this approach, the core dynamics of the system are modeled using an agnostic view towards the inner mechanisms of interacting molecules and biological species [55, 56]. For example, in [12], a simple two equation ODE model was used to capture the toggle-switch behavior of the genetic regulatory network. Though the model does not include well known interactions such as RNAp binding to a promoter site, transcript degradation and its interaction with ribosome, it is sufficient in describing the observed dynamics and delivering insightful predictions towards fine-tuning the behavior. In another example, an

iterative model identification approach is employed where details are added to a model, starting from a relatively simple model, increasing its ability to capture the observed dataset [57]. The approach, in an effort to avoid overfitting system behaviors and risk losing predictive power of the model, draws a threshold at the level of complexity allowed in the model. The approach sometimes results in multiple competing hypotheses, and the authors postulate that these can be used to design experiments that would yield hypothesis nullifying information. It is notable that the synthesis-based approach to understanding biology brings us to build systems from the bottom-up; and the corresponding approach to modeling requires us to discard as much satellite information as possible in order to capture the core dynamics from top-down.

Chapter 3

PARAMETER COMPONENT DEPENDENCY MATRIX

3.1 Notations

Each member of a combinatorial library is a system variant consisting of n interchangeable components and is denoted by the n-tuple,

$$(C_1^{(j_1)}, C_2^{(j_2)}, \cdots, C_n^{(j_n)}), \tag{3.1}$$

where the i-th component $C_i^{(j_i)}$ of Eq 3.1 belongs to a set C_i with $|C_i| = \eta_i$. The n-tuple is also alternatively denoted by $S^{(j_1, j_2, \cdots, j_n)}$, where S denotes an individual system variant. There exists a total of $\prod_i^n \eta_i$ unique combinations and the set containing the entire suite of system variants is denoted by **S**. Figure 3.1A shows an example system composed of *three* interchangeable components, C_1, C_2 and C_3 ($n = 3$). Figure 3.1B shows that each component family has 4, 5 and 3 variants – the library has total of $4 \times 5 \times 3 = 120$ system variants. Figure 3.1C shows two examples of system variants, $S^{(1,2,3)}$ and $S^{(4,1,2)}$.

The experimental data collected from each system variant is denoted by $D^{(j_1, j_2, \cdots, j_n)}$, and the set containing the entire suite of dataset collected from **S** is denoted by **D**. Because interchangeable components are mechanistically homogeneous, we assume that the architecture and the general behavior of all variants are shared. Therefore, a single mathematical model for the entire suite of the system variants is shared and denoted by $f(\theta)$, where $\theta = [k_1, k_2, \cdots, k_m]$ denotes the *elementary*

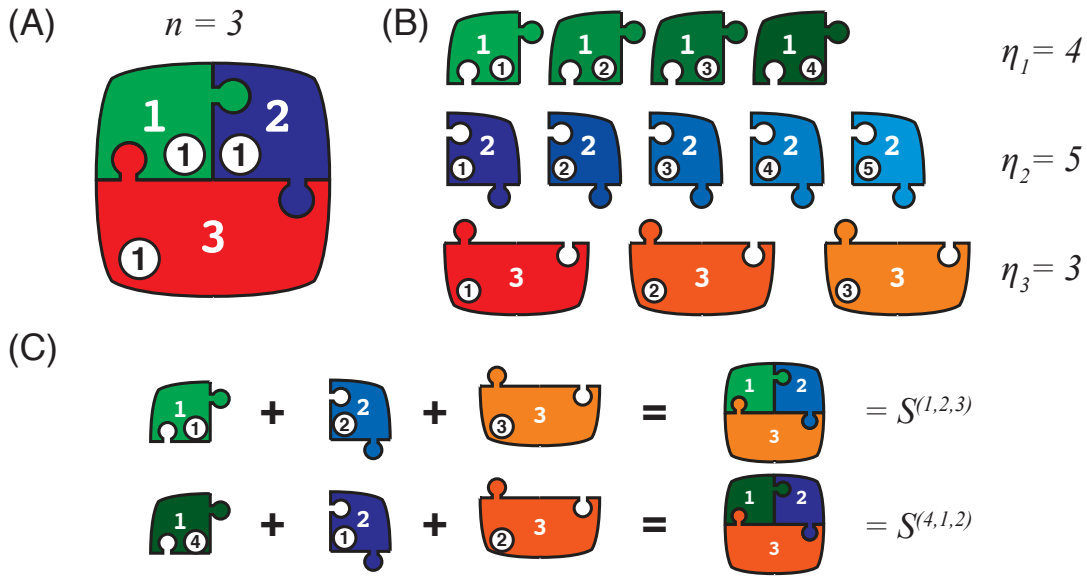


Figure 3.1: A schematic of the construction of combinatorial libraries with three constituent components. (A) Each member of the library, a system variant, is constituted of three interchangeable components. (B) Each component family has multiple members. (C) Each unique combination is a system variant, $S^{(j_1,j_2,\cdots,j_n)}$.

parameter vector of the model with length m. For each system variant, an elementary parameter vector is associated, $\theta^{(j_1,j_2,\cdots,j_n)} = [k_1^{(j_1,j_2,\cdots,j_n)},\cdots,k_m^{(j_1,j_2,\cdots,j_n)}]$. The set of all parameter vectors for all system variants in $\mathbf{S}$ is denoted by $\Theta = \{\theta^{(1,1,\cdots)},\theta^{(1,2,\cdots)},\cdots,\theta^{(\eta_1,\eta_2,\cdots)}\}$ and called the *expanded* parameter vector set. Finally, the model predicted data of a system variant is denoted by $\hat{D}^{(j_1,j_2,\cdots,j_n)}$ and is of the same type as the experimental data (e.g. time-series). Each $\hat{D}^{(j_1,j_2,\cdots,j_n)}$ is a function of $\theta^{(j_1,j_2,\cdots,j_n)}$, and the set of all model predicted behaviors is denoted by $\hat{\mathbf{D}}$. Note that $\hat{\mathbf{D}}$ is a function of the expanded parameter vector set, Θ.

To estimate Θ, we define the following cost function,

$$\begin{aligned} J(\Theta) &= \|D^{(j_1,j_2,\cdots,j_n)} - \hat{D}^{(j_1,j_2,\cdots,j_n)}\|_2 \\ &= \|D^{(j_1,j_2,\cdots,j_n)} - f(\theta^{(j_1,j_2,\cdots,j_n)})\|_2. \end{aligned} \tag{3.2}$$

The optimal Θ^* satisfies the following,

$$\Theta^* = \arg\min_{\Theta} J(\Theta). \tag{3.3}$$

For some systems, additional constraints are included to the optimization problem to adhere to limits on the parameters. For example, constraints can ensure that the estimated reaction rates are positive.

As discussed in Chapter 1, we seek dependency relationships between model parameters and system components. We informally state that a parameter is dependent on a component if substituting the component for another one will change the value of the parameter. The formal definition of dependency is as follows: consider that each dependency hypothesis – that some parameters are dependent on some components – is represented by a boolean matrix, called the *Parameter-Component Dependency* (PCD) matrix. For a given $\mathbf{S}$ and Θ, a PCD of size $m \times n$, M, is defined as follows.

$$M(\kappa, i) = \begin{cases} 1 & : \text{if the } \kappa\text{-th parameter is dependent on the } i\text{-th component} \\ 0 & : \text{otherwise,} \end{cases}$$

where $M(\kappa, i)$ denotes the entry in the κ-th row and i-th column in the matrix. There exist a total of $2^{m \times n}$ different PCDs for a given $\mathbf{S}$ and Θ pair. Each M may be indexed by the integer h, beginning

with 0, where M_0 corresponds to a matrix of zeros. Subsequent matrices are indexed using the following formula,

$$h = \sum_{\kappa=1}^{m}\sum_{i=1}^{n} M(\kappa,i)2^{mn-n(\kappa-1)-i}, \tag{3.4}$$

which is the flattened matrix interpreted as binary. For example, the following 2×2 matrices are indexed 13, 8, 3, and 0, respectively.

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}. \tag{3.5}$$

Using M, we can generate a set of constraints over Θ, which is denoted by $H(M)$. These constraints are added to the optimization problem, such that for a given M, we solve the following modified optimization problem.

$$\begin{aligned} \underset{\Theta}{\text{minimize}} \quad & J(\Theta) \\ \text{subject to} \quad & H(M). \end{aligned}$$

Using this constrained optimization formulation, we write the cost function as $J(\Theta,M)$ from now on, to indicate that the minimization of the cost is subject to M. In the next section, the procedure for generating $H(M)$ is introduced.

3.2 Contraints Generation

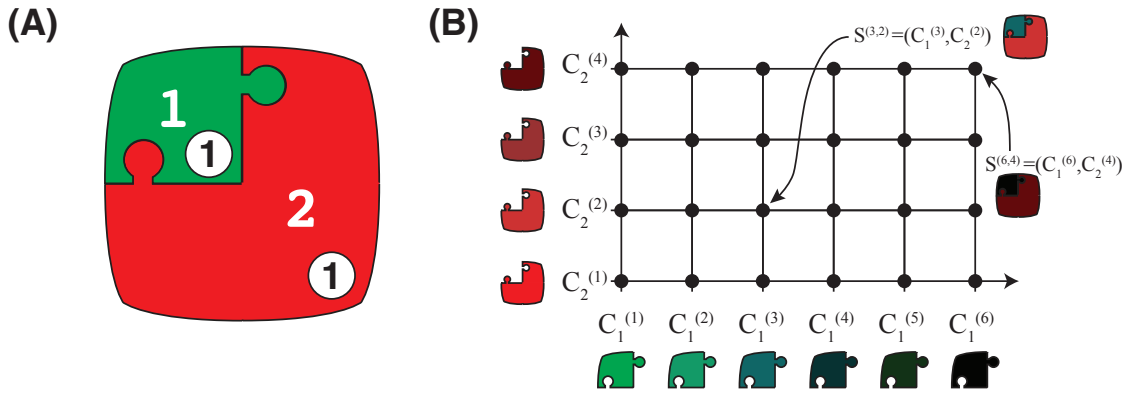
A set of equivalence constraints over the model parameters effectively reduces the number of parameters that are being optimized in the constrained optimization. It also reduces the number of unique quantities per system variant used to compare two or more system variants. From its

definition, if $M(\kappa,i)=1$, then the κ-th parameter is dependent on the i-th system component. This implies that the parameter can assume different values for different component variants. On the other hand, if $M(\kappa,i)=0$, then the κ-th parameter is *independent* of the i-th system component, implying that the parameter assumes a fixed value over the changing component variants. Therefore, when $M(\kappa,i)=0$, we generate a set of constraints over the set of all possible parameter vectors that ensure the κ-th parameter corresponding to all η_i component variants are equal to one another. In other words, when $M(\kappa,i)=0$,

$$k_{\kappa}^{(j_1,\cdots,j_{i-1},1,j_{i+1},\cdots,j_n)} = k_{\kappa}^{(j_1,\cdots,j_{i-1},2,j_{i+1},\cdots,j_n)} = \cdots = k_{\kappa}^{(j_1,\cdots,j_{i-1},j_{\eta_i}j_{i+1},\cdots,j_n)}, \tag{3.6}$$

where $j_i = 1,\cdots,\eta_i$ and $i = 1,\cdots,n$. To generate constraints, we begin by identifying the positions of all zeros in M. Then, for each position identify the set $K_{(\kappa,i)} = \{k_{\kappa}^{(j_1,j_2,\cdots,j_n)} | M(\kappa,i)=0\}$. The set $K_{(\kappa,i)}$ is then partitioned into $(\prod_i^n \eta_i)/\eta_i$ groups (each group has η_i elements) that have the same superscripts except for at the i-th entry. Finally, $H(M)$ is generated by asserting equivalence relationship within each group of parameters.

In Figure 3.2A, we show an example of composable system, S, with $n=2, \eta_1=6$ and $\eta_2=4$ to demonstrate the constraints generation process. Figure 3.2B shows a visualization of the two dimensional space on which the combinatorial library of system variants is placed. Total of twenty-four different system variants belong to the library, and two systems, $S^{(2,3)}$ and $S^{(6,4)}$, are indicated with arrows and corresponding pictographic representations. Assuming that the system has a model $f(\theta)$ with parameter vector of length 1, the elementary parameter vector is shown in Figure 3.2, along with the expanded parameter vector set with twenty-four elements, each element belonging to a system variant indicated in Figure 3.2B.



(C) elementary parameter vector

$$\theta = [k_1]$$

expanded parameter vector set

$$\Theta = \{[k_1^{(1,1)}], [k_1^{(1,2)}], [k_1^{(1,3)}], [k_1^{(1,4)}], [k_1^{(2,1)}], [k_1^{(2,2)}], [k_1^{(2,3)}], [k_1^{(2,4)}], [k_1^{(3,1)}], [k_1^{(3,2)}], [k_1^{(3,3)}], [k_1^{(3,4)}], [k_1^{(4,1)}], [k_1^{(4,2)}], [k_1^{(4,3)}], [k_1^{(4,4)}], [k_1^{(5,1)}], [k_1^{(5,2)}], [k_1^{(5,3)}], [k_1^{(5,4)}], [k_1^{(6,1)}], [k_1^{(6,2)}], [k_1^{(6,3)}], [k_1^{(6,4)}]\}$$

Figure 3.2: A system composed of two components with a model with a single parameter. (A) A system composed of two interchangeable components. (B) Visualization of the two dimensional space on which the combinatorial library of system variants is placed. Two of twenty-four variants, $S^{(2,3)}$ and $S^{(6,4)}$, are indicated with arrows on the two dimensional space. (C) The elementary vector of the system's model, $f(\theta)$, is of length 1. The expanded parameter vector set has twenty-four elements, each element being a vector of length 1.

In Figure 3.3, the four PCD candidates that correspond to the system shown in Figure 3.2 are shown, M_0, M_1, M_2 and M_3. In all four panels in Figure 3.3, the 3-dimensional space depicts the index of C_1 components along the x-axis, the index of C_2 components along the y-axis, and the estimated values of k_1 along the z-axis. The 2-dimensional space of (x,y) coordinates is the same as the one shown in Figure 3.2B. In panel (A), we see that using M_0 hypotheses, all values of k_1 belonging to the twenty-four system variants are constrained to the same value. In panel (B), M_1 implies that k_1 has a dependency relationship to the second component, C_2, but not to the first component, C_1. Therefore, the system variants that share the same second component variant and different first component are grouped together with equivalence constraints asserted within each group. In panel (C), M_2 implies that k_1 has a dependency relationship to the first component, but not to the second component. Therefore, the system variants that share the same first component variant and different second component are grouped together with equivalence constraints asserted within each group. Finally, in panel (D), M_3 implies that k_1 has dependency relationships to both components. Thus, there exits no specific equivalence constraints placed upon k_1 estimates and they are allowed to vary.

The objective of the PCD analysis framework is to find an M that results in low cost, $J(\Theta, M)$, with a sufficiently small number of nonzero, $nnz(M)$. We consider $nnz(M)$ in the optimization because this value is equal to the number of unique quantities identified for a system variant, the quantity we seek to minimize. Therefore, a parallel optimization,

$$\begin{aligned} \underset{\Theta}{\text{minimize}} \quad & J(\Theta, M) \\ \text{subject to} \quad & nnz(M) \leq \mu, \end{aligned}$$

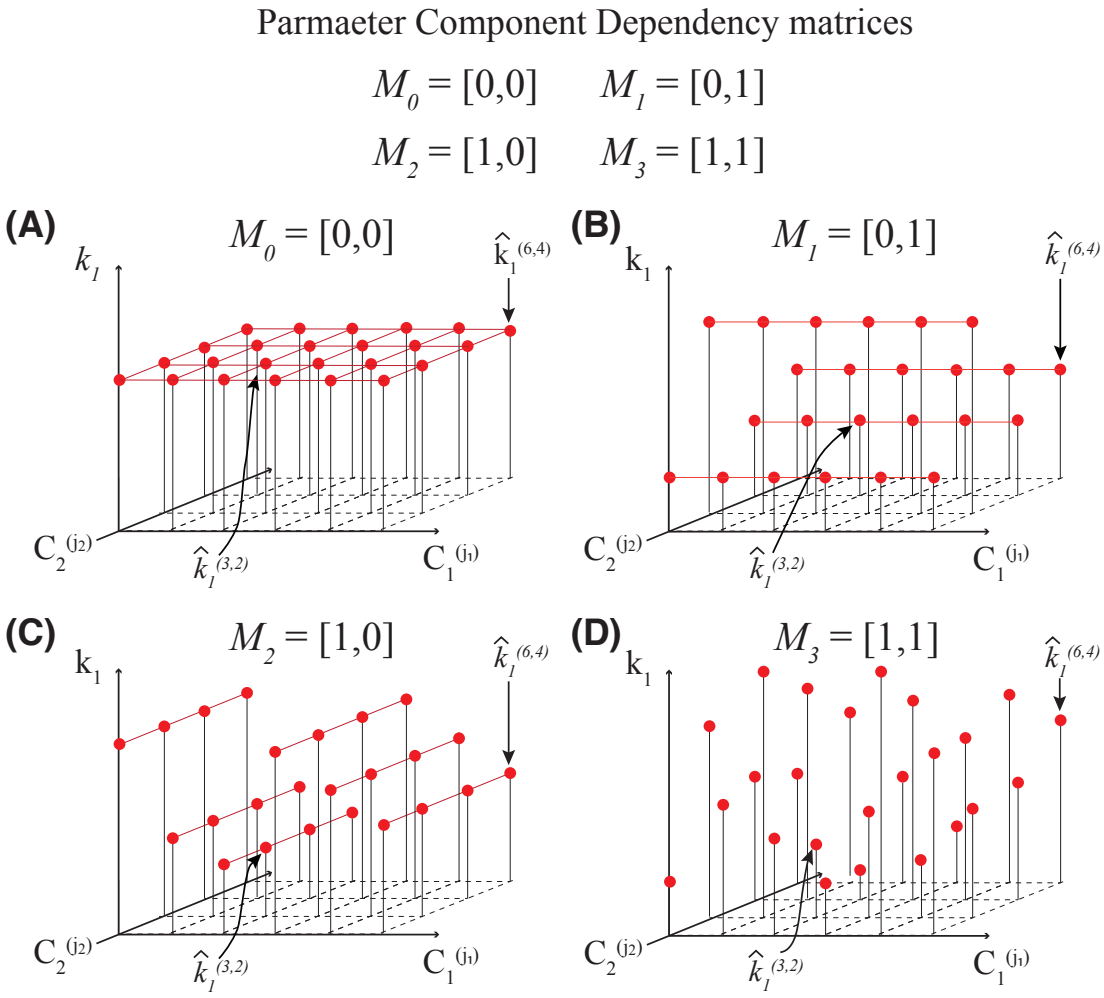


Figure 3.3: Parameter Component Dependency matrices and visualization of the constraints generated from the system shown in Figure 3.2. Given that there are two components in the system and the model has a parameter vector of length 1, there are four candidate PCDs, M_0, M_1, M_2 and M_3. (A) M_0 implies that the parameter k_1 does not have a dependency relationship to either of the two components, C_1 or C_2. Therefore, each parameter belonging to a system variant has an equal value as another parameter belonging to another system variant. The 3-dimensional space depicts the index of C_1 components across the x-axis, the index of C_2 components across the y-axis, and the estimated values of k_1 along the z-axis. The 2-dimensional space of (x,y) coordinates is the same as the one shown in Figure 3.2B. (B) M_1 implies that k_1 has a dependency relationship to the second component, C_2, but not to the first component, C_1. (C) M_2 implies that k_1 has a dependency relationship to the first component, but not to the second component. (D) M_3 implies that k_1 has dependency relationships to both the first and the second components.

is possible where μ constrains the number of nonzero entries in M. Therefore, the PCD analysis framework is composed of two layers of constrained optimization - the inner and the outer optimization levels. The inner optimization deals with estimating Θ and its performance largely depends on the accuracy of the optimization routines used in the process as well as by the quality of initial guesses given to the routine. However, it is difficult to guarantee that the set of solutions uncovered by the optimization routine is the global optimum, unless the model of the system results in a convex optimization cost function. On the other hand, the objective of the PCD method is identifying a set of quantities that capture the systematic functionalities of components. Therefore, when a set of (local) optimal parameter values are identified, we are primarily concerned with relationships (e.g. rankings) of these values with one another, and not the absolute values that may invite comparisons with parameter values estimated under different contexts (e.g. initial condition fed to an optimization routine). The outer optimization deals with estimating the optimal M and its primary objective is comparable to feature selection problems (Section 4.4). In this thesis, given that system analyses of combinatorial libraries and parts characterization problems are not generally defined, the optimality of $nnz(M)$ is determined heuristically.

3.3 PCD Candidates as a Powerset

In this section, we briefly introduce the theory of ordered sets, a branch of mathematics that formalizes our intuitive understanding of order using binary relations. Interestingly, the PCD method is well-suited for presentations using the definitions and structures presented in the theory of ordered sets. The following definitions can also be found in [7] along with more comprehensive discussion on the theory.

Order and Ordered Set. Let P be a set. An *order* on P is a binary relation $\leq$ on P such that, for

```
H(M) = { }
κ, i = 1, 1
while κ <= m
  while i <= n
    if M(κ,i) == 0,
      1. K is a set containing all k(κ,1, 2, ..., n)
      2. partition K by superscripts with the same entries in all
        positions except at the i-th position
      3. assert equivalence relations among the parameters in each
        partition
      4. H(M) = H(M) ∪ equivalence relations generated in step 3.
  i = i + 1
κ = κ + 1
```

Table 3.1: Pseudocode for generating $H(M)$.

all $x, y, z \in P$, the following three properties hold: reflexivity[1], anti-symmetry[2], and transitivity[3]. A set P along with an order relation $\leq$ is said to be an *ordered set* (or partially ordered set).

Order Preserving Map. Let P and Q be ordered sets. A map $\varphi : P \to Q$ is said to be *order preserving* if $x \leq y$ in P implies $\varphi(x) \leq \varphi(y)$ in Q. Order preserving maps also satisfy the following. Let $\varphi : P \to Q$ and $\psi : Q \to P$ be order preserving maps. Then the composite map $\varphi \circ \psi$, given by $\varphi(\psi(x))$ for $x \in Q$, is order-preserving.

Up-set and Down-set. Let P be an ordered set and $Q \subseteq P$. Q is a *down-set* if, whenever $x \in Q, y \in P$ and $y \leq x$, we have $y \in Q$. Dually, Q is an *up-set* if, whenever $x \in Q, y \in P$ and $y \geq x$, we have $y \in Q$. The down-set and up-set are denoted by $\downarrow P$ and $\uparrow P$, respectively.

The Covering Relation. Let P be an ordered set and let $x, y \in P$. We say x is covered by y, if $x < y$ and $x \leq z \leq y$ implies $z = x$, and denote it by $x \prec y$. The latter condition demands that there be

---

[1] $x \leq x$

[2] $x \leq y$ and $y \leq x$ imply $x = y$

[3] $x \leq y$ and $y \leq z$ imply $x \leq z$

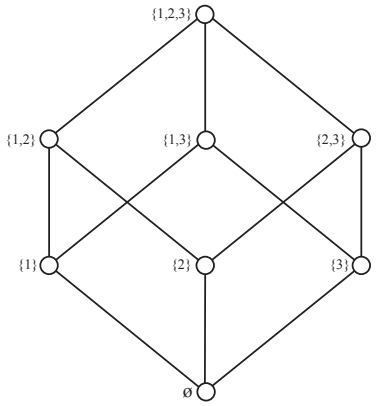


Figure 3.4: Hasse diagram of $\mathscr{P}(\{1,2,3\})$. Each vertex corresponds to an element $\mathscr{P}$ and each edge represents a covering pair.

no element $z \in P$ with $x < z < y$.

Hasse Diagrams. Let P be a finite ordered set. A Hasse diagram is a graph that represents P by a configuration of nodes and interconnecting lines. The construction goes as follows: 1) To each point $x \in P$, associate a point $P(x)$ of the euclidean plane R^2, depicted by a small circle with centre at $P(x)$. 2) For each covering pair x, y in P, take a line segment $\ell(x,y)$ joining the circle at $P(x)$ to the circle at $P(y)$. 3) Carry out 1) and 2) such that, if x is covered by y, then $P(x)$ is 'lower' than $P(y)$ and the circle at $P(z)$ does not intersect the line segment $\ell(x,y)$ if $z \neq x$ and $z \neq y$. Figure 3.4 shows the Hasse diagram of $\mathscr{P}(\{1,2,3\})$.

Among multiple PCD candidates, there exists a hierarchical relationship that can be exploited to find the optimal M. To highlight this hierarchical relationship, we use an alternative notation for M. First, we denote the set of all entry positions in M of size $m \times n$ by $X^{(m,n)}$,

$$X^{(m,n)} = \{(\kappa, i) \mid \{1,\cdots,m\} \times \{1,\cdots,n\}\}. \tag{3.7}$$

For example, a system-model pair with two interchangeable components and two parameters has a corresponding

$$X^{(2,2)} = \{(1,1),(1,2),(2,1),(2,2)\}. \tag{3.8}$$

Then, we let M be denoted by an ordered pair of 1) the set of positions of all non-zero entries and 2) the matrix size. For example,

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{3.9}$$

is denoted by $(\{(1,1),(2,2)\},(2,2))$. The size of M is included for completeness, but will be dropped from here on for simplicity. Using this notation, we see that the set of all candidate Ms of size $m \times n$ is equivalent to the powerset of $X^{(m,n)}$, $\mathscr{P}(X^{(m,n)})$. For example, if $(m,n) = (2,2)$,

$$\begin{aligned} \mathscr{P}(X^{(2,2)}) = \{&\phi, \\ &\{(1,1)\},\{(1,2)\},\{(2,1)\},\{(2,2)\} \\ &\{(1,1),(1,2)\},\{(1,1),(2,1)\},\{(1,1),(2,2)\}, \\ &\{(1,2),(2,1)\},\{(1,2),(2,2)\},\{(2,1),(2,2)\}, \\ &\{(1,1),(1,2),(2,1)\},\{(1,1),(1,2),(2,2)\}, \\ &\{(1,1),(2,1),(2,2)\},\{(1,2),(2,1),(2,2)\}, \\ &\{(1,1),(1,2),(2,1),(2,2)\}\}, \end{aligned} \tag{3.10}$$

where ϕ denotes the empty set. We then define an order on $\mathscr{P}(X^{(m,n)})$ as follows. For $M_{h_1}, M_{h_2} \in \mathscr{P}(X^{(m,n)})$,

$$M_{h_1} \subseteq M_{h_2} \Rightarrow \ \forall x: \ \{x \in M_{h_1} \Rightarrow x \in M_{h_2}\}. \tag{3.11}$$

For example, the following statement is true.

$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \subseteq \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{3.12}$$

Using this order relation, we say that $\mathscr{P}(X^{(m,n)})$ is an ordered set, or equivalently write $\left\langle \mathscr{P}(X^{(m,n)}), \subseteq \right\rangle$. Figure 3.5 shows a Hasse diagram for the ordered set $\left\langle \mathscr{P}(X^{(2,2)}), \subseteq \right\rangle$. Each vertex corresponds to an M and each edge corresponds to a pair $M_{h_1} \prec M_{h_2}$ with a covering relation. At the top of the diagram is M_0, a matrix of all zeros and at the bottom of the diagram is M_{15}, a matrix of all ones. Finally, a pair of (M_{h_1}, M) where $M \in \downarrow M_{h_1}$ is referred to as a *predecessor-successor* pair, and if a predecessor-successor pair also has a covering relationship, the successor is also known as an *immediate* successor[4]. Said plainly, an immediate successor is generated by replacing a single zero on its predecessor matrix with a one.

3.4 Cost as an Order-Preserving Map

Building upon the definitions introduced in the previous section, the optimization cost J is shown to be a map from the ordered set $\mathscr{P}(X^{(m,n)})$ to a real number, $J : M \to R$. Furthermore, it has the

[4] we often omit the term 'immediate'.

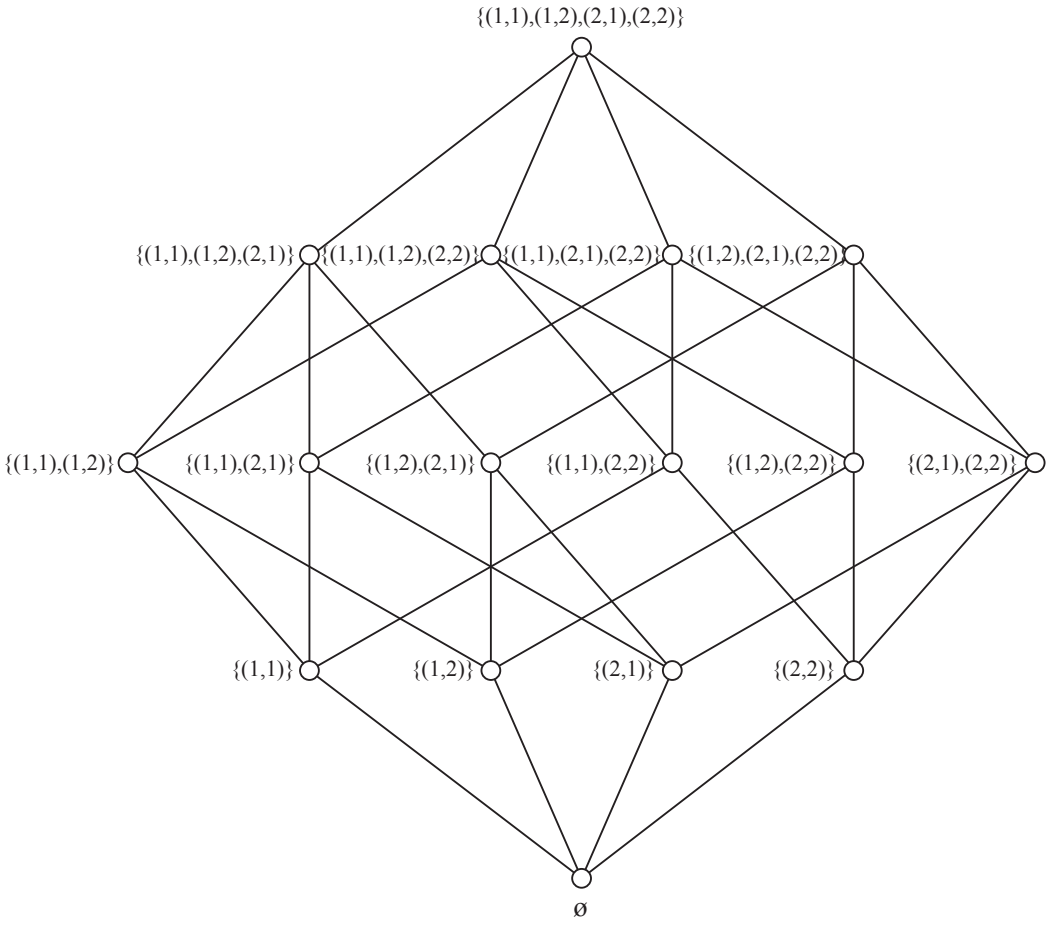


Figure 3.5: Hasse diagram of $\mathscr{P}(X^{(2,2)})$. Each vertex corresponds to an M and each edge represents a covering pair. The relation $\subseteq$ orders the set $\mathscr{P}(X^{(2,2)})$ from bottom to top. Inversely, the dual relation $\supseteq$ orders the set from top to bottom.

property of an order preserving map[5], such that,

$$J(\Theta, M_{h_1}) \geq J(\Theta, M), \quad \forall M \subseteq M_{h_1} \tag{3.13}$$

To prove Eq 3.13, we use the following toy-example: consider a system with *one* interchangeable component (with two variants) and a model of the system with one parameter. This example considers two PCD candidates, both of which are constants. Thus, $M_0 = [0]$ and $M_1 = [1]$, or equivalently $M_0 = (\phi)$ and $M_1 = (\{(1,1)\})$, and $M_0 \subseteq M_1$. The respective constraints over the parameters corresponding to the two matrices are,

$$\begin{aligned} H(M_0) &= \{k_1^{(1)} = k_1^{(2)}\}, \\ H(M_1) &= \phi. \end{aligned} \tag{3.14}$$

Figure 3.6 shows the two possible scenarios that exist for the cost landscape. On both panels, $H(M_0)$ and $H(M_1)$ are the diagonal line and the entire R^2 (the green surface), respectively. The optimal value, J^* for $H(M_0)$ and $H(M_1)$ are shown with green and black dots, respectively. In panel (A), $min\ J(k_1^{(1)}, k_1^{(2)})$ lies directly on $H(M_0)$ implying that for both constraints $H(M_0)$ and $H(M_1)$, we get $min\ J(\Theta, M_0) = min\ J(\Theta, M_1)$. Conversely, in panel (B), $min\ J(k_1^{(1)}, k_1^{(2)})$ lies outside of $H(M_0)$ line implying that $min\ J(\Theta, M_0) \geq min\ J(\Theta, M_1)$. Finally, because increasing the dimensions of this example setup does not distort the general conclusion, the proof holds with any arbitrary $m, n \geq 1$.

Generally speaking, Eq 3.13 implies a monotonicity of the function J over an ordered list of predecessor-successor matrices. Thus, starting from the bottom of the Figure 3.5 and traveling up,

[5]Though, strictly speaking, it reverses the order.

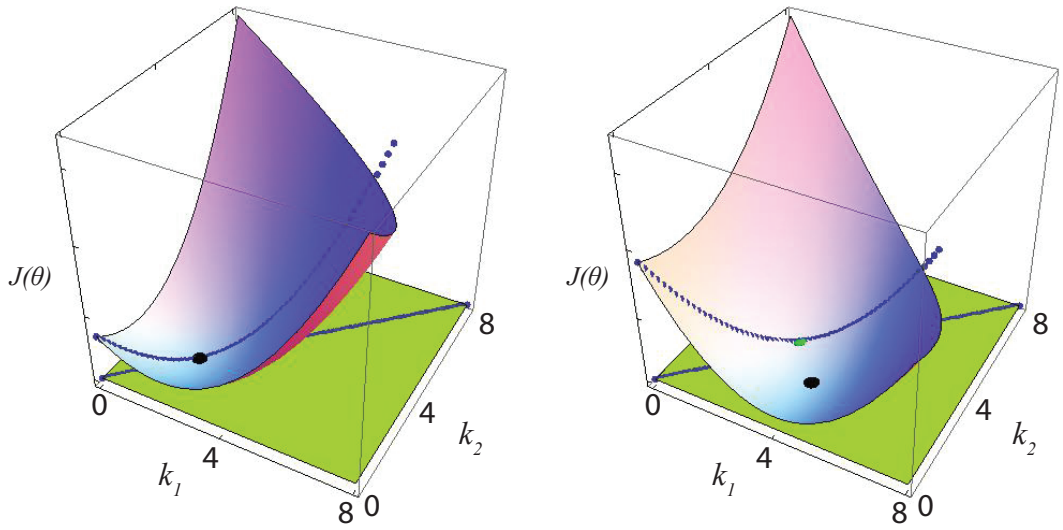


Figure 3.6: Two examples of constrained optimization in $m, n = 1$. The green surface represents the 2-dimensional space $(k_1^{(1)}, k_1^{(2)})$, the black diagonal line across the green surface corresponds to the $k_1^{(1)} = k_1^{(2)}$ line. The shaded purple curved space represents the cost landscape of J. The green and black dots represent the optimal J values under the constraints, $H(M_0)$ and $H(M_1)$, respectively. (A) Case (i) where the optimal J over the space $(k_1^{(1)}, k_1^{(2)})$ is directly on $H(M_0)$. (B) Case (ii) where the optimal J over the space $(k_1^{(1)}, k_1^{(2)})$ is not on $H(M_0)$.

the path will result in non-increasing J values. Finally, this observation ensures that M_0 and $M_{2^{mn}-1}$, the bottom and top nodes in Figure 3.5, correspond to the maximum and the minimum costs over all candidates of M, respectively. In the next chapter, we will demonstrate how this characteristic of J is used to increase the efficiency of searching for the optimal M.

Chapter 4

LEADER ELECTION SYSTEM

4.1 Introduction

In this chapter, we introduce a hypothetical synthetic biological system called the Leader Election (LE) system to demonstrate the PCD matrix framework. The LE system is an adaptation of a previously designed pattern-formation genetic circuit engineered in *Escherichia coli*, developed as a proof of concept to test 1) the engineerability of growing microcolonies to form patterns in 2D [58], and 2) the expressibility of genetic regulatory networks, namely in executing the leader election algorithm developed in distributed computing [59, 60, 61]. The algorithm uses randomness to break symmetry in a group of uniform agents to generate a heterogeneous state [62]. Because the developmental process of multicellular organisms exhibit the principle of symmetry breaking, studying the LE algorithm may lead to a deeper understanding of the process: by designing relatively simple and analytically tractable gene regulatory networks that can break symmetry, we gain an unobstructed view of the core design principles employed by nature. Furthermore, such information is an essential primitive for engineering novel synthetic behaviors in living systems involving multicellular behaviors, such as coordination, competition, communications and cheating.

There are three states in the LE system - *undecided, leader* and *followers*. The initial state of the LE system is a homogeneous population of *undecided* cells. At some point, some undecided cells transition to *leader* cells, triggered by a random event coupled. This event is communicated with the leader cells' neighbors and triggers them to transition into followers. For example, the random leader

selection event could be a gene expression initiated by a leaky promoter, and the communication channel can be realized through a signaling molecule that diffuses through a cellular membrane. The transition of undecided to followers can be realized through an expression of another gene (e.g. for a transcriptional repressor) triggered by the incoming signaling molecule. The system can be engineered so that the repressor inhibits the expression of GFP, so that the followers can be visually distinguished by their lower fluorescence intensity compared to leaders that continue to express GFP. Therefore, at the final state, the initially homogeneous population of cells is differentiated into two distinct populations, those that express GFP (the leaders), and those that do not (the followers).

We assumed that the hypothetical LE system is engineered with three hypothetical interchangeable components, C_1, C_2 and C_3, with each component having five unique variants ($\eta_1 = \eta_2 = \eta_3 = 5$). For example, these components could be the leaky promoter, the diffusing signaling molecule and the transcriptional repressor from the example from the previous paragraph. By employing five different variants of each component, we can create $5^3 = 125$ unique combinations of LE system variants. To describe the general behavior of the LE system, we identified a population level ordinary differential equation model that captures the dynamics of the three states of the LE system as follows.

$$\begin{aligned} \dot{w} &= -k_1 w - k_2 w x \\ \dot{x} &= k_1 w \\ \dot{y} &= k_2 w x. \end{aligned} \tag{4.1}$$

The model has three states, w, x and y, representing the population density of undecided, leaders and

followers, respectively. We assume that y can be measured by monitoring the GFP intensity. The model has two parameters, k_1 and k_2, that denote the rate of transition from undecided to leader state and the rate of transition from undecided to followers, respectively. The elementary parameter vector of the model is $\theta = [k_1, k_2]$, where each parameter is potentially dependent on one, two, three or none of the three system components. In most real systems, the relationship between system components and model parameters is unknown. However, for the purpose of this case study, we invented arbitrary functions of the system components to generate the parameter values as follows

$$\begin{aligned} k_1^{(j_1,j_2,j_3)} &= \frac{1}{c_1^{(j_1)}} \\ k_2^{(j_1,j_2,j_3)} &= \frac{c_2^{(j_2)}}{0.1 + c_3^{(j_3)}}, \end{aligned} \tag{4.2}$$

where $c_1^{(j_1)}, c_2^{(j_2)}$ and $c_3^{(j_3)}$ are values dependent on the choices of the first, second and third component variants. For example, extending the example of the promoter, signaling molecule and repressor, we can argue that the transition rate, k_1, is dependent on the first component, the promoter, because the leakiness of regulated gene expression is a property of the promoter. By the same token, k_1 is arguably independent of the second or the third component, signaling molecule or repressor, because the leaky gene expression rate should not depend on the selection of these components. For the purpose of data simulation c_1, c_2 and c_3 were chosen from uniform distributions, $U(0.1, 10), U(1, 2)$, and $U(0.1, 1.5)$, respectively. Each simulated time-series was further augmented to mimic experimental reality by adding normally distributed random variable, $\eta \sim \mathrm{N}(0, \sigma)$. Given Eq 4.2, the PCD that correctly describes the dependency relationship of the model

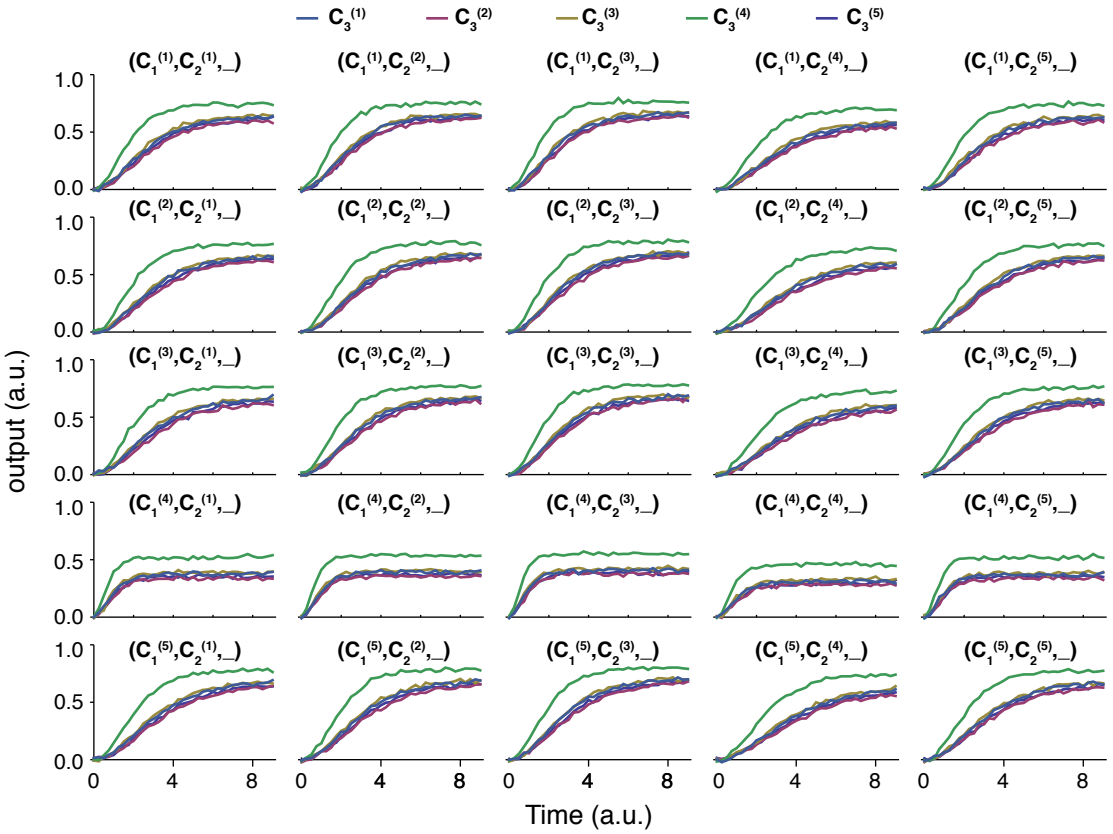


Figure 4.1: Simulated data set of the LE system variants. Each row and column corresponds to the five different first and second components, respectively. Within each panel, the time-series corresponding to the five different third components are shown with different plot markers.

parameters to the system components is

$$M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}. \tag{4.3}$$

Figure 4.1 shows the simulated LE data set, where the rows and columns correspond to the different C_1 and C_2, respectively. Within each panel are five time-series that correspond to a set of five LE variants that have the same C_1 and C_2 and different C_3. It is shown that within each group,

two different clusters of behaviors appear - one LE variant with $C_3^{(5)}$ that have relatively faster response and the other four LE variants that have slower response, and overlap with one another. This pattern is not characteristic of the LE systems, but is rather a result of the randomly chosen c values. Regardless, such patterns lend approaches that reduce the computational cost of parameter estimation. In the case of C_3 component variants, there exists a clear separation between $\{C_3^{(4)}\}$ and $\{C_3^{(1)}, C_3^{(2)}, C_3^{(3)}, C_3^{(5)}\}$, and more importantly, this separation is consistent across the different combinations of C_1 and C_2. This allows a smaller scale analysis that is targeted to the subset of the data corresponding to $\{C_1^{(1)}, C_1^{(2)}, C_1^{(3)}, C_1^{(4)}, C_1^{(5)}\} \times \{C_2^{(1)}, C_2^{(2)}, C_2^{(3)}, C_2^{(4)}, C_2^{(5)}\} \times \{C_3^{(1)}, C_3^{(4)}\}$. Upon further inspection, similar reductions can be made across the C_1 and C_2 variants and we limit our optimization of M to the subset of data corresponding to $\{C_1^{(1)}, C_1^{(4)}\} \times \{C_2^{(1)} C_2^{(4)}\} \times \{C_3^{(1)}, C_3^{(4)}\}$.

4.2 Exhaustive Search

There exist $2^6 = 64$ candidate PCD matrices M, and the relatively small number allows us to conduct a complete survey of the search space. Using the eight characteristically distinguishing system variants identified in the previous section, we conducted an exhaustive search of all 64 candidates, $M_0 - M_{63}$. Their individual performances are normalized by $J(\Theta, M_0)$ (Figure 4.2). It was observed that M_0, the matrix representation of a hypothesis with the most strict constraints on the parameter estimation, results in the highest cost, performing the worst in fitting the observed data. This is most likely due to the fact that by forcing the optimization to fit the average behavior, the cost function is equal to the overall variance of the simulated data, the theoretical upper limit on the cost function[1]. To investigate a general trend, if any, in the cost as a function of $nnz(M)$, the estimates are grouped and ordered by $nnz(M)$ (Figure 4.3). Though the mean cost decreases with

[1] within the boundary of earnest efforts of fitting the data using the given model

increasing $nnz(M)$, large one standard deviation bars indicate that there are large variances for some groups, especially the groups corresponding to $nnz = 2, 3$ or 4. This shows that not all M that share the same $nnz(M)$ perform equally, and indicates that $nnz(M)$ are not the primary determinant of M performance.

The matrix of all 1s, M_{63}, reaches the minimum $J(\Theta, M)$ over all candidate M because it has the largest degrees of freedom for parameter estimation. This trend was previously predicted by the order-preserving characteristic of J (Eq 3.13). Interestingly, it was shown that there are seven other matrices that perform equally well, $M_{35}, M_{39}, M_{43}, M_{47}, M_{51}, M_{55}$ and M_{59}. Since we search for a matrix with low J and small $nnz(M)$, we choose M_{35} with $nnz(M_{35}) = 3$ as the optimal M, which is equal to the matrix in Eq 4.3. Therefore, we demonstrated that through a complete survey of the PCD candidates, the original matrix used to generate the simulated dataset can be recovered. However, for moderately large m or n, the computational cost of a complete survey of the entire search space becomes prohibitively expensive.

4.3 Greedy Algorithm

To develop a more efficient approach to identifying the optimal M, we consider implementing a greedy approach, where the multivariate optimization is divided into multiple steps to choose a local optimum at each iteration. To implement it, we first set M_0, with the highest $J(\Theta, M_0)$, as the initial point, and call it the predecessor matrix, denoted by iteration index, $M_0[0]$. Using this predecessor matrix, we identify its successor matrices by replacing each 0 in the matrix with 1. We denote the generating function for identifying the set of successor matrices given M_h by $\mathscr{G}(M_h)$ and define it as

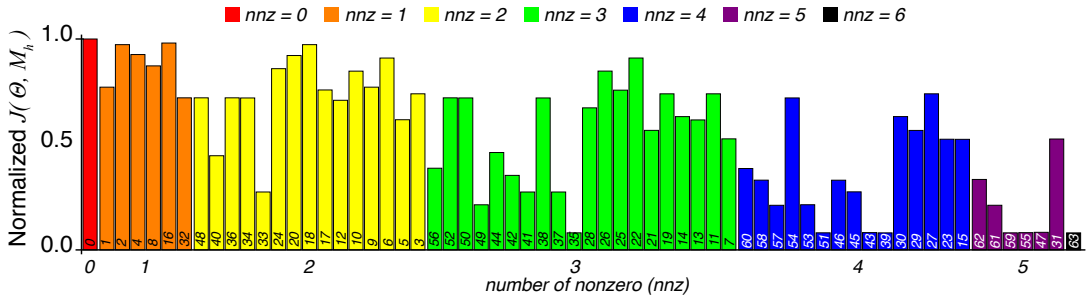


Figure 4.2: Normalized $J(\Theta, M_h)$ of 64 candidates matrices of LE. The performance is computed from the 2-norm difference between simulated data and model predicted output. Individual costs are normalized by the cost of $J(\Theta, M_0)$. The simulated data used to compute the cost functions looked at a subset of the system consisting of 8 unique system variants, which are combinations of 2 different first, second and third components. The numbers of non-zero (nnz) element are color coded (red = 0, orange = 1, yellow = 2, green = 3, blue = 4, purple = 5, black = 0). The numbers at the bottom of the bars indicate the candidate index (h).

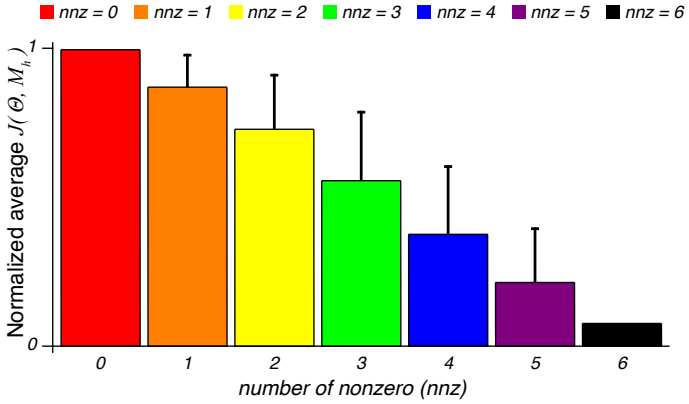


Figure 4.3: Average performances of 64 candidate M for LE, grouped by nnz. The error bars indicate 1 standard deviation of the performance within each group. There are 1 ($nnz = 0$), 6 ($nnz = 1$), 15 ($nnz = 2$), 20 ($nnz = 3$), 15 ($nnz = 4$), 6 ($nnz = 5$) and 1 ($nnz = 6$) matrices in each group, respectively.

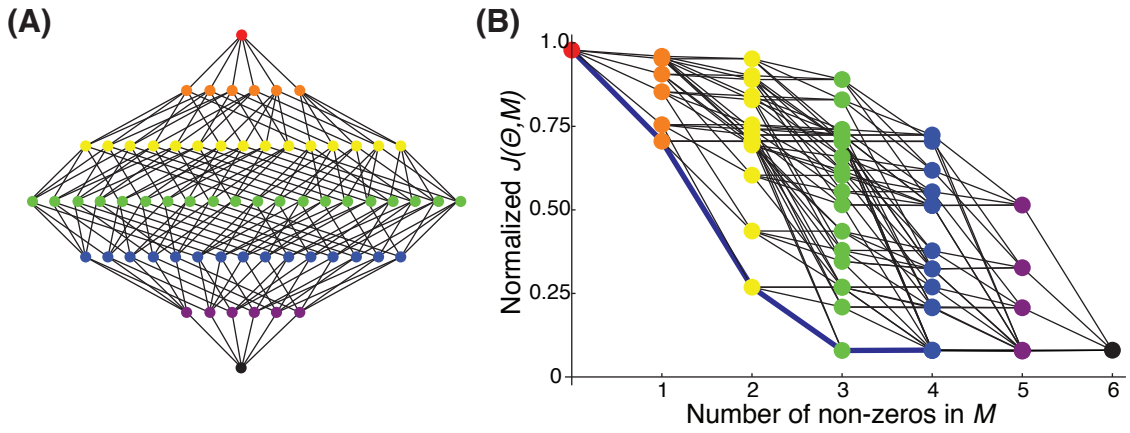


Figure 4.4: Hasse Diagram and the costs of the set of M candidates of LE. (A) The diagram shows pairs of M in covered relations and is a complete lattice, a graphical representation of the order relationship within a powerset. In this diagram, it represents the order of the set, $\mathscr{P}(X^{(2,3)})$, the candidate set of all M of LE. (B) The cost of each M are plotted and ordered by $nnz(M)$. Notice that the graph, by preserving the connectivities between predecessors and successors, is an isomorphism of the Hasse diagram on the left. The path of the greedy algorithm is depicted with a bold blue line.

follows

$$\mathscr{G}(M) \quad = \quad \{\overline{M} \mid \overline{M} \in\downarrow M \text{ and } nnz(\overline{M}) = nnz(M)+1\}. \tag{4.4}$$

For example, $\mathscr{G}(M_0) = \{M_h \mid h = 32,16,8,4,2,1\}$. Then, for each immediate successor, we minimize $J(\Theta \cdot M_h)$, and choose the matrix with the lowest cost as the next predecessor matrix,

$$\begin{aligned} M[t+1] \quad &= \quad \arg\min_{M_h} J(\Theta, M_h) \\ &\qquad \forall\, M_h \in \mathscr{G}(M[t]). \end{aligned}$$

This step is iterated until either 1) $M_{2^{mn}-1}$ is reached and no more successors can be generated, or 2) the decrease in cost value from $M[t]$ to $M[t+1]$ is less than some user-defined threshold τ.

1. t = 0
2. $M[t] = M\_0$
3. $M[t+1] = \underset{M\_h}{\arg\min} J(\Theta, M\_h)$
   $\forall\ M\_h \in \mathscr{G}(M[t])$
4. t = t + 1
5. if $\frac{J(\Theta, M[t+1]) - J(\Theta, M[t])}{J(\Theta, M[t+1]) + 1} < \tau$, terminate. $\hat{M*} = M[t]$
6. Otherwise, repeat 3-4

Table 4.1: Pseudocode for greedy search for optimal M.

Figure 4.4B plots each $(nnz(M), J(\Theta^*, M))$ pair on an (x, y) coordinate. Each point is connected with its immediate predecessors and successors, preserving the covering relationship shown in panel (A). The path of greedy algorithm is highlighted with a bold blue line. Because the cost reduction from the iteration $nnz = 3$ to $nnz = 4$ is negligible, the algorithm terminates and chooses the matrix of the previous iteration, M_{35}. Thus, we showed that the optimal matrix recovered by the greedy algorithm is the same as the one recovered by the complete survey.

4.4 Sparse Group Lasso and PCD

Linear models are frequently employed because their analytically tractable form yields fast results, though most real and interesting systems are nonlinear. Regardless, the analytical methods developed for linear models can apply to nonlinear systems with useful outcomes, if not in mechanical implementation but just for its logical approach to solving the problems at hand. Here, we apply a linear model fitting algorithm to the problem of identifying dependency relationships between model parameters and system components, and discuss whether the result is comparable to that of the PCD analysis. First, we begin with the following equation which formulates a linear

model in the matrix notation,

$$Y \quad = \quad X\beta + \varepsilon. \tag{4.5}$$

In it, Y is an $N \times 1$ vector of observations, X is an $N \times p$ feature matrix (or a design matrix), and β and ε are $N \times 1$ vectors of parameters (or coefficients) and zero-mean normally distributed noise, respectively. The general idea is that a system described by Eq 4.5 is tested or observed to yield the observations, Y; the observation is hypothesized to be a linear combination of some features, X; and the weights, or coefficients, of the combination is β, the vector we wish to estimate. To do so, the following optimization problem is solved,

$$\min_{\beta} \left(\|Y - X\beta\|_2 + g(\beta) \right). \tag{4.6}$$

The first term with ℓ_2-norm measures the distance between the observation and the model predicted outcome, and $g(\cdot)$ is a function of β that imposes extra constraints on β. The methods of constraining β are called regularization, it identifies and eliminates redundant and irrelevant features to avoid overfitting. Many different flavors $g(\cdot)$ exist such as ridge-regression [63, 64, 65], lasso [66], LARS [67] and subset selection [68, 69]. The ridge-regression, in particular, uses ℓ_2-norm for $g(\beta)$, which poses Eq 4.6 as a continuous variable optimization, allowing the application of many efficient solvers [70, 71]. However, because the resulting estimates, though small in values, do not get eliminated entirely, users are burdened with extraneous features. On the other hand, lasso and LARS uses ℓ_1-norm to induce sparsity of β, resulting in the identification of the features with the strongest effect on the system behavior.

It can be argued that the objective of the PCD optimization discussed in previous sections, specifically in regards to inducing parsimony by optimizing $nnz(M)$, is a form of feature subset selection problem, because both methods aim to reduce the number of features (or parameters) that are used to describe a set of variant observation (or output behavior). Furthermore, the entries of PCD are boolean, implying that the 'features' of parameter-component dependencies either exist or they do not (a feature is either selected or it is not). Among the various subset selection methods, group lasso considers the problem of selecting grouped variables for prediction in regression [72]. The goal is to select important main effects for accurate prediction, and this amounts to the selection of groups of derived input variables. The key observation here is that groups of variables are more relevant than individual variables, and the group lasso induces an entire group to be selected or drop-out. An extension of this method to induce sparsity in the solution was introduced in [73], called the sparse group lasso (SGL), and it solves the following convex optimization problem

$$\min_{\beta \in R^p} \left(\|Y - \sum_{\ell=1}^{L} X_\ell \beta_\ell\|_2^2 + \lambda_1 \sum_{\ell=1}^{L} \sqrt{p_\ell} \|\beta_\ell\|_2 + \lambda_2 \|\beta\|_1 \right). \tag{4.7}$$

Here, p predictors are divided into L groups, with p_ℓ denoting the number in group ℓ, and X_ℓ and β_ℓ denotes the predictors and the coefficient vector of the ℓ-th group, respectively. The method is an effective compromise between the lasso and the group lasso, yielding sparsity in both the group and individual levels [74].

In this section, we use the sparse group lasso method to demonstrate that its results are comparable with the result of the PCD. In Table 4.2, the estimated model parameter for the eight LE system variants are shown. These are parameter estimates obtained via M_{63}, hence, no constraints over the parameters. We use these values in the application of SGL, to avoid having any unverified

hypotheses affecting its outcome. Table 4.3 shows the resulting $\hat{\beta}$ for various hyper parameter value λ, where decreasing λ allows for solutions with less sparsity. It is shown that when λ is large, all of the $\hat{\beta}$ are zeros (the first column). As λ decreases, the penalty is lessened, and we can see that some of the $\hat{\beta}$ are allowed to be non-zeros. The key result here is the order in which different groups are allowed to take on nonzero parameter values with decreasing λ. For the SGL results for $\hat{k_1}$, the first group to take on non-zero estimates is the one corresponding to C_1, indicating that C_1 is the most significant source of variations in $\hat{k_1}$. This trend is maintained through the range of λ, implying that the variations in C_2 and C_3 are less significant in their effects to vary k_1. For the SGL results for $\hat{k_2}$, the first and second groups to take on non-zero estimates with decreasing λ are the ones corresponding to C_3 and C_2. This implies that these two components are the first and the second significant sources of variations for $\hat{k_2}$. This result is generally comparable to the optimal PCD found in previous sections, M_{35}. One of the interesting results of the SGL is that the parameter-component dependency relationship elucidated from it is quantitative. The relative values of β estimates indicate the rank order among the different pairs of parameter-component, and contain more information than 'True/False' as those found in a PCD matrix. However, we may be able to assign similar relevant rankings of the strengths of parameter-component pairs using the PCD matrix as well. For instance, in Figure 4.4B, the successive drops in cost at the different iterations are varied, specifically the largest drops in cost occur in the first few iterations. By mapping back the drop in cost, which is approximately equal to the increase in performance by the corresponding hypothesis, to the PCD that generated it, we may be able to derive a similar information as those of SGL.

Table 4.2: Parameter estimates corresponding to a subset of eight LE system variants using inherited initial guesses. Using the greedy algorithm, the parameter estimates were inherited starting from M_0 to M_{63}.

	$S^{(1,1,1)}$	$S^{(1,1,2)}$	$S^{(1,2,1)}$	$S^{(1,2,2)}$	$S^{(2,1,1)}$	$S^{(2,1,2)}$	$S^{(2,2,1)}$	$S^{(2,2,2)}$
$\hat{k_1}$	0.150	0.150	0.144	0.152	0.755	0.767	0.757	0.744
$\hat{k_2}$	1.278	3.453	0.902	2.340	1.277	3.454	0.903	2.337

4.5 Summary

In this chapter, we used a small simulated system called the Leader Election system, a synthetic biological circuit that implements the Leader Election distributed algorithm, to demonstrate the analytical framework of the PCD. We defined an arbitrary set of functions that predetermine the dependency relationships between three system components and two model parameters. The functions were used to generate arbitrary parameter values corresponding to 125 system variants, and these values were used to simulate a population-level ODE model. By taking advantage of the relative small number of PCD candidates, we were able to survey the performances of all candidates. The survey revealed seven other M that performed as well as M_{63}, which corresponds to the theoretical lower boundary of the cost function. We chose the M with the lowest $nnz(M)$ based on the criteria of the outer-level optimization problem, seeking to optimize the parsimony. Using the complete survey method, we demonstrate that the recovered M is equal to the one used in simulating the dataset originally. This implies that there exists in the dataset, some information regrading the dependency relationship between system components and the parameters of the system model. Furthermore, we demonstrate that such information can be formally represented with a boolean matrix, PCD, and by surveying all possible hypothesis regarding the dependency relationship, the "true" dependency

Table 4.3: Fitted parameters of $\hat{k_1}$ and $\hat{k_2}$ using the sparse group lasso.

$\hat{k_1}$	λ	0.107	0.075	0.052	0.036	0.025	0.052	0.036
$\hat{\beta}_{C_1^{(1)}}$		0.00	-0.13	-0.22	-0.28	-0.33	-0.36	-0.38
$\hat{\beta}_{C_1^{(4)}}$		0.00	0.13	0.22	0.28	0.33	0.36	0.38
$\hat{\beta}_{C_2^{(1)}}$		0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\beta}_{C_2^{(4)}}$		0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\beta}_{C_3^{(1)}}$		0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\beta}_{C_3^{(4)}}$		0.00	0.00	0.00	0.00	0.00	0.00	0.00

$\hat{k_2}$	λ	0.319	0.220	0.154	0.107	0.075	0.052	0.036
$\hat{\beta}_{C_1^{(1)}}$		0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\beta}_{C_1^{(4)}}$		0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\beta}_{C_2^{(1)}}$		0.00	0.00	0.00	0.10	0.23	0.32	0.38
$\hat{\beta}_{C_2^{(4)}}$		0.00	0.00	0.00	-0.10	-0.23	-0.32	-0.38
$\hat{\beta}_{C_3^{(1)}}$		0.00	-0.39	-0.66	-0.85	-0.98	-1.07	-1.13
$\hat{\beta}_{C_3^{(4)}}$		0.00	0.39	0.66	0.85	0.98	1.07	1.13

relationship can be recovered.

Next, in recognition of the prohibitively expensive cost of completely surveying all possible PCDs with growing numbers of system components and model parameters, we implemented the greedy algorithm. It was shown that the greedy search algorithm was successful in recovering the optimal M using the simulated data of the LE system. The stopping criterion of the greedy algorithm is determined heuristically; either by placing an upper-limit on the $nnz(M)$, or if the reduction in the cost from one iteration to the next is sufficiently small.

An alternative approach to optimize M is to take an advantage of the fact that if the algorithm

is allowed to continue until reaching its final iteration, $M_{2^{mn}-1}$, the $J(\Theta^*, M_{2^{mn}-1})$ approaches the *min* J over all M (Section 3.3). In the LE example, through both an exhaustive search and the greedy algorithm, we found that M_{35} is the optimal PCD, but only *after* both approaches were executed fully. In the future work, in lieu of exhaustive search, we can adopt the following forward and backward propagation approach: first, we let greedy search algorithm propagate forward until reaching the $M_{2^{mn}-1}$, then we back propagate along the path, and choose the last M that has cost $J(\Theta, M) \leq J(\Theta, M_{2^{mn}-1}) + \tau(J(\Theta, M_0) - J(\Theta, M_{2^{mn}-1}))$, where τ is a design parameter that allows users to arbitrarily set the desired threshold, and $(J(\Theta, M_0) - J(\Theta, M_{2^{mn}-1}))$ is the complete range of cost. For example, if $\tau = 0.1$, the algorithm chooses the first M that performs better than the 90% of the cost differential between the entire range of max and min of the $J(\Theta, M)$ over all M.

Because of the particular nature of hill-climbing optimization algorithms, it is possible for the greedy algorithm to *get stuck* in a local optimum and not reach the global optima. Such a case can be pictographically demonstrated using the Hasse diagram, such as the one shown in Figure 4.4(A). It should be noted that any instance of $J(\Theta^*, M)$ on an (x, y) coordinate plot is an isomorphism of the Hasse diagram, and that given the relationship in Eq 3.13, all line-segments in the graph have slopes less than or equal to zero. In these diagrams, the isomorphisms of a PCD Hasse diagram, the greedy algorithm is demonstrated to work if there is no lines intersecting the lowest line-segments of the graph (the path followed by the greedy algorithm). If such an intersection occurs, the greedy algorithm ends up following the path that diverges from the lowest line of the graph and it may not merge with it until the final iteration. In Figure 4.5, we show an isomorphism of the Hasse diagram that pictographically represents this case, while satisfying the non-positive slope constraints. Among the sixteen candidates, the optimal M corresponds to the lowest green dot, but the greedy algorithm

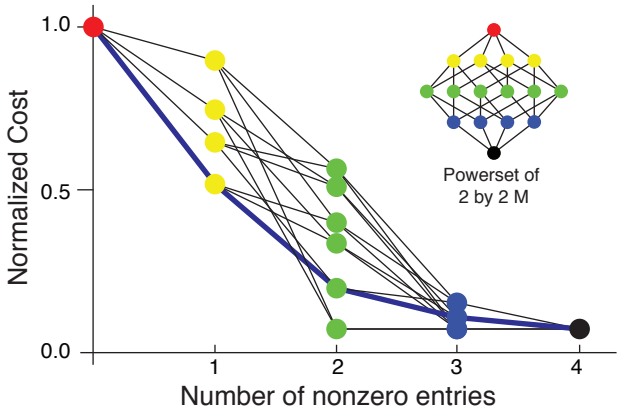


Figure 4.5: An isomorphism of a Hasse Diagram of $\mathscr{P}(\{1,2\})$ depicting an instance where the greedy algorithm fails. The path of the greedy algorithm is depicted with a bold blue line. Among the sixteen candidates, the optimal M corresponds to the lowest green dot, but the greedy algorithm is shown to not recover this M. At the second iteration, the algorithm chooses from one of the successors the best performer and updates. By getting stuck in a local optimal path, the algorithm continues the iteration until the algorithm terminates with the M_{15}.

is shown to not recover this M. The divergence occurs at the second iteration, and because it is stuck in a local optimal path, the algorithm continues until the algorithm terminates with M_{15}. As of now, it is not known whether such a behavior can arise from some fundamental characteristics of a composed system or of the model of the system. An in-depth investigation of the existence of such characteristic would be beneficial in conducting the PCD analysis, as it would indicate the possible greedy algorithm failure and directs users to implement alternative optimization algorithms.

Lastly, we demonstrated that the PCD analysis yields results consistent with that of a feature selection algorithm called the sparse group lasso (SGL). The results of the SGL R package can be interpreted to yield insights regarding dependency relationships between model parameters and system components. Though it was the case in the LE system that both PCD and SGL yielded the same results, in certain cases, the SGL method may yield inconsistent results as those of the PCD.

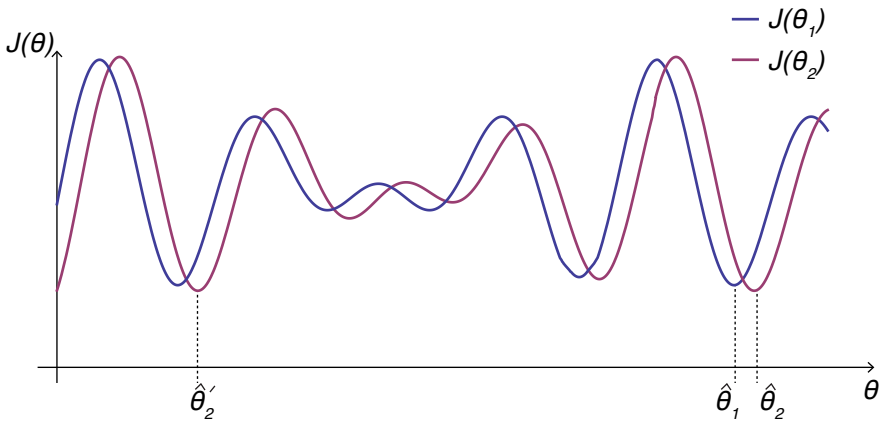


Figure 4.6: An example of multiple local optima of two different system variants. When the free-for-all estimation individually finds the θ_1 and θ_2 on the far right side of the cost landscape, it is fair to attribute the parameter values to the systematic difference between the constituent components. However, if the free-for-all estimation for θ_2 settles on the value on the left side of the cost landscape, it is difficult to attribute the variations in the parameter values to the systematic difference, but rather estimation error most likely caused by different initial conditions and the optimization solver's algorithm for searching the parameter space.

The general idea in using the SGL to elucidate the parameter-component dependency is to use the estimated model parameters as the response vector, Y, and assign the group indices to correspond to the component variant indices, the collection of which becomes the observation matrix, X, in the regression formula. In doing so, it is critical that the model parameters are estimated with no constraints regarding the existence of any parameter-component dependency. Said differently, the parameters are estimated with *no equivalence constraints*, or with the $M_{2^{mn}-1}$ matrix (matrix of all ones). This is to ensure that the parameters fairly capture the variations among component variants without any *a priori* assumptions about their relationship to the components. In fact, this particular relationship is precisely what the spare group lasso uncovers. Furthermore, the estimated parameters of each system variant need be consistent to those of other system variants, in ways that they belong

to relatively close proximity to one another. The reason for such a requirement is because of the existence of multiple local optima. In Figure 4.6, an illustration of this is shown using an example of 1-dimensional vector θ and a combinatorial library with two system variants, $S^{(1)}$ and $S^{(2)}$, where the curves indicate hypothetical cost landscapes of the two θ. Assume unconstrained optimization yielded the estimated values of $\hat{\theta^{(1)}}$ and $\hat{\theta^{(2)}}$ as the ones towards the right hand side of the figure. The SGL applied to this set of estimates is likely to elucidate the variations in parameter estimates arising from systematic differences between the components. However, if for a variety of reasons (different starting guess given to the optimization solver, ruggedness of the cost landscape, particular approaches in traversing the cost landscape used by optimization routines, etc.), estimated $\hat{\theta^{(2)}}$ value corresponds to the value shown towards the left hand side of the figure, it is difficult to compare the $\hat{\theta^{(1)}}$ and $\hat{\theta^{(2)}}$ fairly, such that the difference between them is caused only by the difference between the component variants.

The systematic searches for the optimal PCD is useful in analyzing combinatorial libraries of systems where little is known about the internal mechanism. Assuming some high level dynamics of the system is captured by its model, the interpretation of model parameters and their dependency relationships to system components revealed through the PCD framework may give practitioners better insights about the system, such as which components contribute more strongly to variations of system behavior.

Chapter 5

AUXIN SIGNAL PATHWAY I: IAA DEGRADATION SYSTEM

The material presented in this chapter has been published in [8]. Our collaboration with a plant biology research laboratory[1] provided an opportunity to develop a custom analysis framework that suits the unique construction scheme of the synthetic auxin signal pathway presented here. The pathway satisfies the requirements of our problem statement, being composed of multiple interchangeable components that each belongs to a family of mechanistically homogeneous proteins. This collaboration aimed to address the origin of diversity observed in the auxin signal pathway proteins as well as their potential use in building multicellular consortia-like behavior in *Saccharomyces cerevisiae*. The article [8] is the first of the two publication series that comprised the foundation for the PCD analytical framework presented in this thesis. The material has been edited with updated notations where possible and marked with footnotes for commentaries.

5.1 Introduction

Auxin directs almost every aspect of plant biology, yet how specificity is generated from auxin signaling components remains largely unresolved. A range of auxin-associated phenotypes, including profound disruptions in development and severely compromised responses to environmental signals, are caused by single amino acid substitutions that stabilize transcriptional co-repressors called the Aux/IAAs or IAAs [75]. The diversity of these phenotypes and the size of the IAA family

[1] Dr. Jennifer Nemhauser, Department of Biology, University of Washington, Seattle, WA

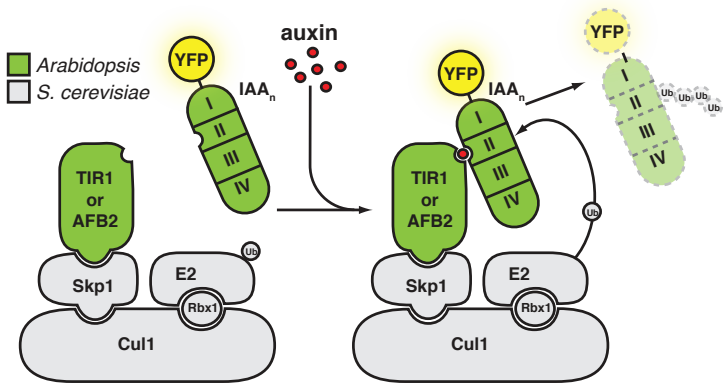


Figure 5.1: Plant auxin AFBs and IAAs integrated into the yeast ubiquitin pathway. Plant auxin AFBs (TIR1 or AFB2) and YFP-tagged IAA repressors were integrated into the yeast ubiquitin pathway, shown in grey.

suggest IAAs may provide specificity in auxin responses [76]. Functional studies support this idea, as stabilized IAAs provoke different phenotypes even when expressed from the same promoter [77, 78].

Auxin activates gene expression by enhancing IAA turnover through interaction with auxin AFBs, a family of F-box proteins called TRANSPORT INHIBITOR RESISTANT 1 (TIR1) / AUXIN SIGNALING F-BOX PROTEINS (AFBs) [2, 79], referred to here collectively as AFBs. Variation in the affinities of IAA|AFB pairs has recently been observed [80]. How such differences relate to degradation kinetics is still unclear. Labor-intensive seedling studies on a small number of IAA proteins, in combination with analysis of stabilized IAA mutants, uncovered the importance of a conserved region, termed domain II, in determining protein stability. The degron-containing IAA domain II is both necessary and sufficient for interaction with TIR1 and the resulting auxin-induced degradation [2, 79, 81, 82]. In addition, IAA-reporter fusions with diverse domain II sequences show a range of degradation rates when over-expressed in Arabidopsis seedlings [83]. However, the

ubiquity of the auxin pathway in plants and the difficulty in reconstituting the complete degradation machinery *in vitro* have hindered further characterization of the molecular determinants of IAA degradation rates.

As a complement to existing systems and to systematically characterize the potential tunability of different IAA|AFB pairs, we engineered the auxin-induced degradation of IAA proteins in *Saccharomyces cerevisiae* (yeast) (Figure 5.1). Our synthetic system has several advantages: precise control of auxin input levels, the ability to study IAA|AFB pairs in isolation, and the absence of the many other plant pathways known to impact auxin signaling [84]. Our system allowed a comprehensive survey of IAA turnover while recapitulating behaviors observed in plants. We discovered that the particular AFB receptor used greatly impacted the rate of degradation and that sequences outside of the degron-containing domain II accelerated or decelerated IAA degradation in an IAA-specific manner. Moreover, we identified a mathematical model that provides a single parameter to quantitatively describe degradation behavior.

5.2 Model Development and Discrimination

The primary objective of our quantitative analysis is to identify a small model, with as few parameters as possible, that describes the differential degradation of IAA|AFB pairs observed in experiments. Having few parameters avoids over-fitting, facilitates comparison among available pairs, and provides a guide for the selection of parts during the rational design of new networks. A computationally intractable number of candidate models can be generated by, for example, various assumptions about the mechanism of degradation and the many molecular species involved. Computational limits, therefore, require that we limit the number of candidate models by employing current knowledge of the auxin signal pathway, basic assumptions about protein synthesis, and

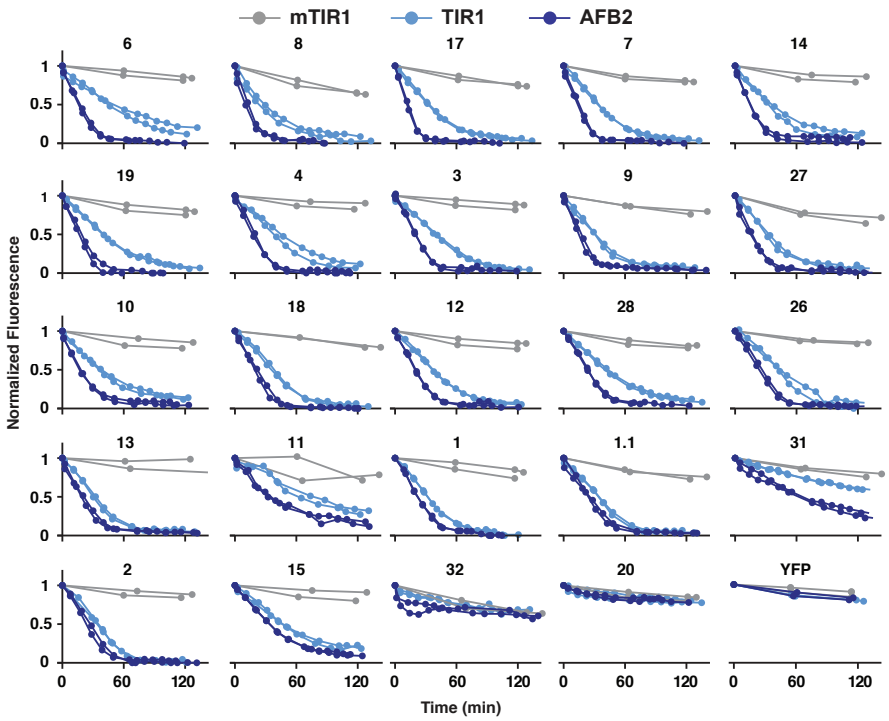


Figure 5.2: IAA degradation is highly variable. A range of IAA|AFB degradation rates were obtained using time-lapse flow cytometry. Degradation curves were normalized to starting fluorescence. IAAs are listed in order of the relative difference in degradation in the presence of TIR1 (light blue) versus AFB2 (dark blue). Strains expressing the F-box deficient mTIR1 (grey) show no auxin-dependent degradation.

simple input-output concepts. We assess candidate models by fitting their parameters to the entire data set, analyzing their qualitative fit, and then computing their residual fit. We fit our models to experimental data using the nonlinear optimization function FindFit in *Mathematica*™ where the cost function is defined as

$$J_1(\theta_s^{(j_1,j_2)}, f_s) = \left[\sum_{(j_1,j_2)} \sum_{t_m \in \mathrm{T}} \left(y^{(j_1,j_2)}(t_m) - \hat{y}^{(j_1,j_2)}(t_m, \theta_{(s)}^{(j_1,j_2)}, f_s) \right)^2 \right]^{1/2}, \qquad (5.1)$$

where j_1 denotes the index of the first component (IAA), j_2 denotes the index of the second component (AFB), s denotes the model index, $y^{(j_1,j_2)}(t_m)$ denotes the measured output at $t = t_m$, T denotes the set of measurement times, $\theta_s^{(j_1,j_2)}$ denotes the parameter vector for a model f_s corresponding to a system variant $(C_1^{(j_1)}, C_2^{(j_2)})$, or alternatively $S^{(j_1,j_2)}$, $\hat{y}^{(j_1,j_2)}(t_m, \theta^{(j_1,j_2)}, f_s)$ denotes the predicted output using the model f_s and its parameter vector $\theta^{(j_1,j_2)}$ at $t = t_m$, and Θ_s denotes the set of parameter vectors of all system variants (**S**) for model f_s[2]. The optimal Θ satisfies the following,

$$\Theta_s^{(j_1,j_2)*} = \underset{\Theta_s^{(j_1,j_2)}}{\arg\min}\, J_1(\theta_s,^{(j_1,j_2)}). \tag{5.2}$$

Following the estimation, the residual of a model is defined as

$$J_1(f_s) = \left[\sum_{j_1,j_2} \sum_{t_m \in \mathrm{T}} \left(\frac{y^{(j_1,j_2)}(t_m) - \hat{y}_{(j_1,j_2)}\left(t_m, f_s, \theta_s^{(j_1,j_2)*}\right)}{y_{(j_1,j_2)}(t_m)} \right)^2 \right]^{1/2}. \tag{5.3}$$

Our approach is to develop candidate models with good qualitative fits, low residuals, and a small number of parameters by incrementally increasing complexity via refinements that are based on known mechanisms. In the model discrimination step, both quantitative and qualitative metrics of each model are evaluated. It is true that quantitatively low residual is desired in general, however, the analytical metric takes precedence. If a model is shown to be analytically incapable of fitting critical features of experimental observation, the model is deemed structurally flawed, and is eliminated, even if the quantitative metric is relatively low. Following are the two general qualitative features observed in the experimental data.

1. Each degradation time-course includes an inflection point, such that the curve switches from

[2] $\eta_1 = 24, \eta_2 = 2$. The subscript on J_1 is to differentiate it with Eq 3.2, introduced in Chapter 3.

concave to convex (Figure 5.2).

2. Each dose-response curve (steady-state fluorescence measurement vs auxin concentration) is nonlinear (Figure 5.3B).

These two features were used as qualitative criteria to which all models were subjected to. The number of parameters for the model is computed over all experimental data sets, such that the total number of parameters equals the number of parameters in the model multiplied by the number of system variants (i.e. the number of distinct IAA|AFB pairs). After a model has been selected, we reduced the total number of parameters by assuming that some of the parameters can be consolidated for a specific group of system variants.

5.2.1 f_0

We first consider the synthetic yeast system as a grey-box with a single input (auxin) and a single output (YFP intensity). Therefore, we propose the following model, f_0, which is a combination of simple synthesis/degradation dynamics and exponential decay:

$$\dot{y}(t) = k_1 - k_2\, y - k_3\, u\, y(t), \tag{5.4}$$

where t denotes the time, y denotes the output, and u denotes the input. This model encodes the hypothesis that the YFP-IAA fusion protein is subject to zeroth-order synthesis and first-order degradation, and that the output degrades at a rate proportional to the input. The first two terms, k_1 and $k_2\, y$, represent the input-independent synthesis and degradation of the output, respectively. The last term represents the degradation of output with the overall rate proportional to itself and the

input. Thus, k_1 is the synthesis rate, k_2 is the basal degradation rate, and k_3 is the input mediated degradation rate. This model has residual $J_1(f_0) = 49.31$, and requires 144 distinct parameters (3 per IAA|AFB pair) to fit the entire data set. The model has a closed form analytical solution,

$$y(t) = \frac{k_1}{k_2 + k_3 u}\left(1 + \frac{e^{-t(k_2+k_3u)}k_3 u}{k_2}\right), \tag{5.5}$$

which demonstrates a nonlinear relationship between u and the y, consistent with the qualitative observation of the experimental data (Figure 5.3B, purple curve). However, the model cannot capture the inflection of the curve (Figure 5.3A, purple curve). For a function to shift its convexity, the second derivative has to equal zero at some $t > 0$ (and that t is the inflection point). However, f_0 has second-order derivative,

$$\ddot{y}(t) = \frac{k_1 k_3 u(k_2 + k_3 u)}{k_2} e^{-t(k_2+k_3u)} \tag{5.6}$$

which shows that for all parameters $k_1, k_2, k_3 > 0$ and $u \geq 0, \ddot{y}(t) > 0$. This represents f_0 is fundamentally unable to capture one of the crucial features of the system.

5.2.2 f_1

The inflection point and the related initial delay in the degradation curves suggest that the IAA degradation mechanism comprises additional intermediate processes (e.g. formation of an intermediate species). To model this feature, we add an internal state x. At this point, we do not assume that x is any specific species or combination of species, just that it is formed and degraded. In particular, we assume that the rate at which x is synthesized is proportional to the input, and that

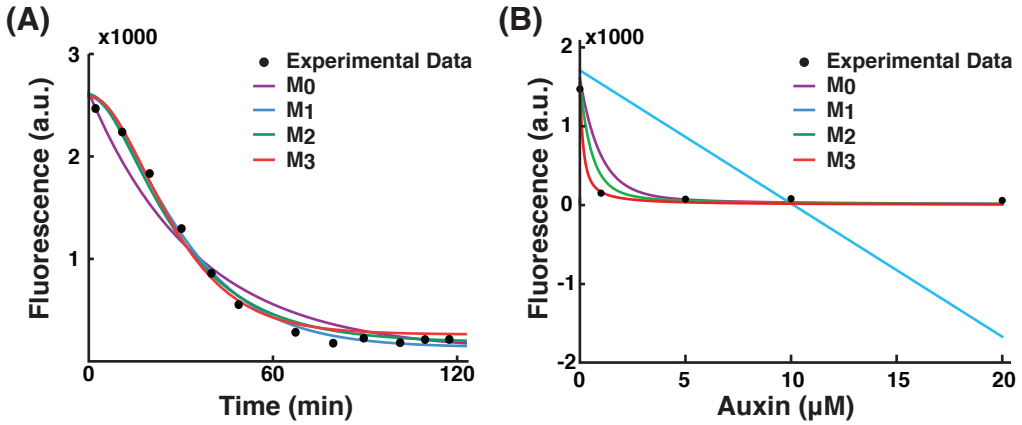


Figure 5.3: Sample time-course IAA degradation data, dose-response data and model fits of IAA14|TIR1. Experimental data are shown in black circles, and the fits are shown in purple, blue, green and red for f_0, f_1, f_2 and f_3 models, respectively.

x affects the non-basal degradation of y. This second-order linear model, f_1, is defined by

$$\begin{aligned} \dot{x}(t) &= k_1 u - k_2 x(t) \\ \dot{y}(t) &= k_3 - k_4 y(t) - k_5 x(t), \end{aligned} \tag{5.7}$$

The model has residual $J_1(f_1) = 81.68$, and requires 240 distinct parameters (5 per IAA|AFB pair) to fit the entire data set. The model captures the inflection point in time-course degradation curve (Figure 5.3A, blue curve). However, the dose-response predicted by f_1 is qualitatively different from our observations. Analytically, the steady-state solution of f_1 has a closed form,

$$\begin{aligned} x^* &= \frac{k_1 u}{k_2} \\ y^* &= \frac{k_3 - k_5 k_1 u / k_2}{k_4}. \end{aligned} \tag{5.8}$$

which demonstrates that y^* has a linear relationship to u. Furthermore, because of this linear relationship, the model predicts that the steady-state fluorescence decreases indiscriminately with increasing amount of input, to the point where negative fluorescence is predicted (Figure 5.3B, blue curve). The fact that f_1 makes predictions that starkly contradict physical constraints demonstrates that the model is qualitatively unfit for the given experimental data.

5.2.3 f_2

Since the dose-response is nonlinear, we modify f_1 by introducing a nonlinear term, making the degradation term dependent on $y(t)$. This second-order nonlinear model, f_2, is,

$$\begin{aligned} \dot{x}(t) &= k_1\,u - k_2\,x(t) \\ \dot{y}(t) &= k_3 - k_4\,y(t) - k_5\,x(t)\,y(t), \end{aligned} \tag{5.9}$$

which has a residual of $J_1(f_2) = 24.54$ and requires 240 distinct parameters (5 per IAA|AFB pair) to fit the entire data set. The steady-state of output predicted by f_2 are,

$$\begin{aligned} x^* &= \frac{k_1\,u}{k_2} \\ y^* &= \frac{k_2 k_3}{k_2 k_4 + k_1 k_5 u}. \end{aligned} \tag{5.10}$$

which demonstrates a nonlinear relationship between y^* and u, and satisfies one of the qualitative features. There is no closed-form solution for the model, but numerical solution demonstrates that the model captures the inflection point in the predicted time-course curves (Figure 5.3, green curve).

5.2.4 f_3

As mentioned before, the internal state x represents an unknown intermediate species (or combination of species) that interacts with auxin input and the YFP-IAA output. It is feasible that more than a single intermediate species is required to encompass the underlying dynamics. To investigate whether this might be the case, we added a second internal state, z, as an intermediate state between x and y. This third-order nonlinear model, f_3, is

$$\begin{aligned} \dot{x}(t) &= k_1\,u(t) - k_2\,x(t) \\ \dot{z}(t) &= k_3\,x(t) - k_4\,z(t) \\ \dot{y}(t) &= k_5 - k_6\,y(t) - k_7\,z(t)\,y(t), \end{aligned} \tag{5.11}$$

which has a residual of $J_1(f_3) = 33.30$ and requires 336 distinct parameters (7 parameters per IAA|AFB pair) to fit the entire data set. As in the case with f_2, f_3 captures the two qualitative features of the system (Figure 5.3, red curve). It is notable that the computed residual is larger than $J_1(f_2)$, however, this is most likely the result of standard estimation error (e.g. non-optimal initial condition or insufficient search space). Alternatively, f_3 may represent a point at which the saturation of model benefit gained by increasing model complexity is saturated.

5.2.5 *Summary*

We generated and evaluated four candidate models, f_0, f_1, f_2, and f_3, with respect to how well each model captures the experimental observation, both quantitatively and qualitatively. f_0 did not capture the inflection observed in time-course degradation curves; f_1 did capture the inflection, but

did not capture the nonlinear dose-response behavior; f_2 captures both the time-course and dose-response data qualitatively; and f_3 matches these behaviors with comparable quantitative metric to f_2. However, given the limitations of the experimental dataset (time-course to auxin step-input and auxin dose response), f_3, nor any more complex models, may not provide verifiable insights to the system. Said differently, more complex models may result in lower quantitative fit metric, however, to balance the estimation uncertainty associated with higher complexity models, we require richer perturbation of the system that reveals the appropriately higher complexity of the internal mechanism. Therefore, f_2 is arguably the simplest explanation for the observed phenomenon.

5.3 Parameter Reduction

Aside from f_0, the model candidates considered in the previous section include one or more internal states, x (and z), that are not readily associated with biological complexes. One way to approach this is to consider the internal state simply as a mathematical necessity for fitting the experimental observation. Another approach is to consider hypotheses as to what biological complex – that we know to exist in the auxin-mediated IAA degradation pathway – can be associated with x, and examine whether these hypotheses lend a useful quality to our model. One of the hypothetical interpretations of x is that it is a complex formed between auxin and AFB. A simple description of the mechanism in which IAA degrades in the presence of auxin and AFB is shown in Figure 5.1, where auxin serves as a molecular glue, binding with an AFB and allowing the protein to bind with IAA, further triggering the ubiquitylation of IAA. Therefore, we postulate that x is a proxy of such a species.

Building on the hypothesis of model interpretation that x is a proxy of the auxin-bound AFB complex, we associate each of the model parameters to the identity of the two different proteins,

IAA and AFB. Firstly, parameters k_1 and k_2 are rates associated with synthesis and degradation of auxin-bound receptor protein, where the synthesis is directly proportional to the input. In which case, we further hypothesize that these two parameters are dependent on the identity of AFB but are independent of the identity of the co-expressed IAA. Secondly, analogous interpretations are given to k_3 and k_4; these are rates of synthesis and basal (input-independent) degradation of IAA-YFP, and they are only dependent on the identity of the IAA and not on the identity of the AFB. Finally, k_5 is unique among the model parameters such that it is dependent on the identities of both proteins. Mathematically, k_5 is the rate constant for how fast the output degrades, and such a degradation process is proportional to the amount of x and y in the system. One can then assign a biochemically feasible interpretation to the term, "when active AFB and IAA interact, the IAA degrades at a rate proportional to k_5". As shown in Figure 5.2, the unifying feature of all IAA|AFB pairs is that they degrade when auxin is added to the system in varying degrees. Such variation in degradation among the IAA|AFB pairs can serve as a unique identifier for each pair, and we investigate whether a subset of parameters can serve as the quantitative identifier. Note that these association of parameters to reaction rates are based on a possible interpretation of the model, and should not be taken as a direct representation of a true system. The parameters are crude amalgamations of multiple biochemical processes but the resulting model is an elegant abstraction rid of unwieldy details.

In the previous section, when we fit candidate models to experimental data, all five of the f_2 parameters,

$$\theta^{(j_1,j_2)} = [k_1^{(j_1,j_2)}, k_2^{(j_1,j_2)}, k_3^{(j_1,j_2)}, k_4^{(j_1,j_2)}, k_5^{(j_1,j_2)}] \tag{5.12}$$

were allowed to vary, resulting in five different parameter values for each IAA|AFB pair. This

method, while allowing us to easily compare candidate models, makes it difficult to compare the parameters of one IAA|AFB pair to another. For example, suppose we want to directly compare the degradation rates of IAA17|TIR1 and IAA17|AFB2, which have mean fitted k_5 values of 0.034 and 0.21, respectively. The large discrepancy in the two values suggests a large difference in the respective degradation dynamics, but because each pair has other parameters that vary as well, the differences in k_5 are not the only source of the variation.

Therefore, we asked if some parameters may be common to a subset of the IAA|AFB pairs, and if consolidating these common parameters would reduce estimation uncertainty. This approach is sometimes called global curve fitting. A systematic generation of all possible global fitting hypotheses for parameter reduction (identifying all possible combinations of common parameters across the 48 IAA|AFB pairs, and five parameter model) would have been computationally expensive. Fortunately, the model interpretation discussed previously provides a feasible guide as to which hypotheses are more likely than others. With this estimation approach, we can reduce the variability caused by other parameters and fairly compare parameter values. As k_5 is the parameter associated with both species in the model, it is unique to the identity of each IAA|AFB pair. Therefore, the parameter is the primary quantity of interest in comparing the differential degradation among the pairs.

First, we address an extreme case of parameter reduction, where we only let k_5 vary for all IAA|AFB pairs, and fix the rest of the parameters to be the same across all pairs. We denote this

hypothesis ζ_1, which provides an additional constraint on the optimization problem, defined by

$$\begin{aligned} \min\ J_1(\Theta, f_s) \quad &= \quad \sum_{j_1,j_2} \sum_{t_m \in \mathrm{T}} |y^{(j_1,j_2)}(t_m) - \hat{y}^{(j_1,j_2)}(t_m, \theta^{(j_1,j_2)}, f_s)|_2, \\ \text{subject to} \quad & k_\kappa^{(j_1,j_2)} = k_\kappa^{(j_1',j_2')}, \\ & \forall j_1, j_1' \in [1, \eta_1],\ \forall j_2, j_2' \in [1, \eta_2] \\ & \text{and } \kappa = 1,2,3,4. \end{aligned} \tag{5.13}$$

where, j_1 and j_2 denote the sets of IAA and AFB indices, respectively, and κ denotes the parameter index. The resulting parameter vector, Θ_{ζ_1}, has 52 distinct parameters (4 + 48 per pair) and results in $J_1(f_2, \Theta_{\zeta_1}) = 48.9$. Thus, ζ_1 reduces the cardinality of the parameter vector to the point where an IAA|AFB pair has only one parameter that differs from another pair. However, because ζ_1 makes both k_3 and k_4 be the same for all IAA|AFB pairs, it implies that the initial output value at $t = 0$ for all IAA|AFB pairs are equal ($y^{(j_1,j_2)}(0) = k_3^{(j_1,j_2)}/k_4^{(j_1,j_2)}$). This is inconsistent with the experimental data where initial levels of expression vary (Figure 5.4A). Therefore, ζ_1 is invalidated based on its conflict between the interpretation of the model and the experimental data [3].

Next, we propose a hypothesis alternative to ζ_1 in which k_1, k_2, k_3 and k_4 are constrained, but across a smaller subset of IAA|AFB pairs. For example, it was shown that AFBs do not have differential effect on basal dynamics of IAA. Therefore, two IAA|AFB pairs with the same IAA have a similar initial output - suggesting equal k_1, k_2, k_3 and k_4 values for these two pairs. We denote

[3] Hypothesis ζ_1 is equal to $M_3 = (\{(5,1),(5,2)\},(5,2))$, following the COO convention. Note that $J_1(f_2, \Theta_{\zeta_1})$ is then equal to $J(\Theta, M_3)$ accepting that f_2 is the default model in the analysis.

this hypothesis by ζ_{11}, and define the following optimization constraints.

$$k_{\kappa}^{(j_1,j_2)} = k_{\kappa}^{(j_1,j_2')} \qquad \forall j_2, j_2' \in [1,\eta_2], \text{ and } \kappa = 1,2,3,4. \tag{5.14}$$

The resulting parameter vector set, denoted $\Theta_{\zeta_{11}}$, has 144 distinct parameters, and results in $J_1(f_2,\Theta_{\zeta_{11}}) = 24.33$. This decreased residual is a validation in which we let the model interpretation and the fundamental details of the system guide our hypotheses generation. To verify this approach, we generated an alternative hypothesis, ζ_{12}, where IAA|AFB pairs with the same AFB, versus those with the same IAA (ζ_{11}), have equal k_1, k_2, k_3 and k_4. This hypothesis has the same cost function as Eq 5.14 but with the following optimization constraint[4],

$$k_{\kappa}^{(j_1,j_2)} = k_{\kappa}^{(j_1',j_2)} \qquad \forall j_1, j_1' \in [1,\eta_1], \text{ and } \kappa = 1,2,3,4. \tag{5.15}$$

Note that ζ_{12} was not generated based on the model interpretation as ζ_{11}, but was proposed as a counter example to our approach. The resulting residual is $J_1(f_2,\Theta_{\zeta_{12}}) = 36.62$, where $\Theta_{\zeta_{12}}$ has 56 distinct parameters, and the increased residual validates our approach. Comparing ζ_1, ζ_{11} and ζ_{12} suggests that

1. larger numbers of distinct parameters tend to decrease the residual, and

2. when constraining the parameters across a smaller subset of IAA|AFB pairs, the model interpretation serves as a helpful guide in generating reasonable hypotheses.

The higher residuals of hypotheses ζ_1, ζ_{11} and ζ_{12} also suggest that for any two IAA|AFB

[4]Hypotheses ζ_{11} and ζ_{12} are equal to $M_{683} = (\{(1,1),(2,1),(3,1),(4,1),(5,1),(5,2)\},(5,2))$ and $M_{343} = (\{(1,2),(2,2),(3,2),(4,2),(5,1),(5,2)\},(5,2))$, respectively.

pairs, more than one parameter ought to be allowed to vary (increasing degrees-of-freedom in the parameter estimation) to fit both data sets. Therefore, we further investigate whether a specific grouping of experimental data and parameter reduction hypotheses is possible. Each case of varying parameter (in addition to k_5), k_1, k_2, k_3 or k_4, is denoted with $\zeta_{21}, \zeta_{22}, \zeta_{23}$ or ζ_{24}, respectively (the four hypotheses are elements of the set $\zeta_{2\kappa}$, where $\kappa = 1, 2, 3$ or 4). Each hypothesis is associated with data grouping that are supported by the model interpretation, depending on which parameter is allowed to vary.

1. For ζ_{21} and ζ_{22}, the additional parameter allowed to vary is dependent on the identity of AFB and independent of the identity of the IAA. Therefore, IAA|AFB pairs are grouped by their IAAs (each group containing two pairs, IAA_{j_1}|TIR1 and IAA_{j_1}|AFB2), resulting in 24 distinct groups. For each group, the parameters are estimated under the constraint,

$$\begin{aligned} \text{subject to} \qquad & k_\kappa^{(j_1,j_2)} = k_\kappa^{(j_1,j_2')} \\ & \forall j_2, j_2' \in [1,2], \quad \kappa \in \{2,3,4\} \text{ or } \{1,3,4\}. \end{aligned} \tag{5.16}$$

   These hypotheses imply that a group of IAA|AFB pairs that have the same IAA have the same basal synthesis and degradation rates (k_3 and k_4) of the output. Furthermore, they imply that the variations among each group can be captured by varying k_5 and k_1 for ζ_{21}, or k_5 and k_2 for ζ_{22}. [5]

2. For ζ_{23} and ζ_{24}, the additional parameter allowed to vary is dependent on the identity of IAA and independent of the identity of the AFB. Therefore, IAA|AFB pairs are grouped by their

[5] Hypotheses ζ_{21} and ζ_{22} are equal to M_{939} and M_{747}, respectively.

AFBs (each group containing 24 pairs, $\text{IAA}_{j_1}|\text{AFB}$, where $j_1 = [1,\cdots,24]$), resulting in two groups. For each group, the parameters are estimated under the constraint,

$$\text{subject to} \qquad k_{\kappa}^{(j_1,j_2)} = k_{\kappa}^{(j_1',j_2)}$$
$$\forall j_1, j_1' \in [1,24], \;\; \kappa \in \{1,2,4\} \text{ or } \{1,2,3\}. \tag{5.17}$$

The hypotheses imply that a group of IAA|AFB pairs that have the same AFB has the same synthesis and degradation rates of the internal state (k_1 and k_2). Furthermore, they imply that the variations among each group can be captured by varying k_5 and k_3 for ζ_{23}, or k_5 and k_4 for ζ_{24}.[6]

Table 5.1 shows the residuals and the number of distinct parameters for various hypotheses investigated. Compared to ζ_1 hypotheses, ζ_2 hypotheses tend to have lower residuals, owing to the larger degrees-of-freedom given in the estimation constraints. Additionally, ζ_{24} has the lowest residual even with the lowest number of distinct parameters. This suggests that for any two IAA|AFB pairs, the differential dynamics between the two can be captured through varying the basal rate parameters of IAA and the input-mediated degradation by the AFBs (k_5). These two dynamics, however, are independent of one another as discussed previously (Figure 5.4). Also, single parameter variation studies suggests that a set of time-courses predicted by varying k_3 or k_4 have identical degradation profiles when normalized to their initial conditions (Figure 5.5). The simulation study is also supported by experimental data of IAA1 and IAA1.1, where IAA1.1 is a codon-optimized version of IAA1. This results in higher expression level for IAA1.1 relative

[6]Hypotheses ζ_{23} and ζ_{24} are equal to M_{375} and M_{351}, respectively.

Table 5.1: The residuals and the number of distinct parameters for f_0, f_1, f_2 (including its hypotheses), and f_3.

Model & Hypothesis	$J_1(f)$	Number of distinct parameters for 48 IAA|AFB pairs
f_0	49.31	144
f_1	81.68	240
f_2	24.54	240
f_2, ζ_{21}	32.29	168
f_2, ζ_{22}	46.32	168
f_2, ζ_{23}	48.11	102
f_2, ζ_{24}	27.44	102
f_3	30.63	336

to IAA1, which under the model interpretation, is equivalent to increasing k_3 and leaving other parameters the same. When the two curves are normalized to their initial conditions (Figure 5.3A), the curves overlap closely. This is reflected by the close k_5 values of IAA1 and IAA1.1 (Tables 5.2 and 5.3).

Table 5.2: Estimated parameters for IAA|TIR1 pairs using ζ_{23}.

AFB	IAA	Replicate	k_1	k_2	k_3	k_4	k_5
TIR1	1	1	1.08E-01	9.22E-02	4.29E+00	3.14E-03	3.44E-02
TIR1	1.1	1	1.08E-01	9.22E-02	8.69E+00	3.14E-03	3.22E-02
TIR1	2	1	1.08E-01	9.22E-02	6.86E+00	3.14E-03	3.54E-02
TIR1	3	1	1.08E-01	9.22E-02	9.03E+00	3.14E-03	2.62E-02
TIR1	4	1	1.08E-01	9.22E-02	9.74E+00	3.14E-03	2.58E-02
TIR1	6	1	1.08E-01	9.22E-02	8.44E+00	3.14E-03	2.02E-02
TIR1	7	1	1.08E-01	9.22E-02	7.62E+00	3.14E-03	3.05E-02
TIR1	8	1	1.08E-01	9.22E-02	3.29E+00	3.14E-03	3.17E-02
TIR1	9	1	1.08E-01	9.22E-02	7.39E+00	3.14E-03	3.34E-02
TIR1	10	1	1.08E-01	9.22E-02	5.02E+00	3.14E-03	2.55E-02
TIR1	11	1	1.08E-01	9.22E-02	4.61E+00	3.14E-03	9.46E-03
TIR1	12	1	1.08E-01	9.22E-02	8.91E+00	3.14E-03	2.67E-02
TIR1	13	1	1.08E-01	9.22E-02	8.24E+00	3.14E-03	3.54E-02
TIR1	14	1	1.08E-01	9.22E-02	5.16E+00	3.14E-03	3.17E-02
TIR1	15	1	1.08E-01	9.22E-02	6.67E+00	3.14E-03	1.91E-02
TIR1	17	1	1.08E-01	9.22E-02	6.74E+00	3.14E-03	3.03E-02
TIR1	18	1	1.08E-01	9.22E-02	7.44E+00	3.14E-03	2.89E-02
TIR1	19	1	1.08E-01	9.22E-02	5.71E+00	3.14E-03	2.41E-02
TIR1	20	1	1.08E-01	9.22E-02	9.99E+00	3.14E-03	2.21E-03
TIR1	26	1	1.08E-01	9.22E-02	8.03E+00	3.14E-03	2.66E-02
TIR1	27	1	1.08E-01	9.22E-02	4.06E+00	3.14E-03	3.18E-02
TIR1	28	1	1.08E-01	9.22E-02	8.20E+00	3.14E-03	2.09E-02
TIR1	31	1	1.08E-01	9.22E-02	6.72E+00	3.14E-03	4.88E-03
TIR1	32	1	1.08E-01	9.22E-02	2.47E+00	3.14E-03	3.81E-03
TIR1	1	2	1.42E-01	1.72E-01	3.89E+00	3.19E-03	4.43E-02
TIR1	1.1	2	1.42E-01	1.72E-01	9.00E+00	3.19E-03	3.77E-02
TIR1	2	2	1.42E-01	1.72E-01	5.77E+00	3.19E-03	4.14E-02
TIR1	3	2	1.42E-01	1.72E-01	8.67E+00	3.19E-03	3.19E-02
TIR1	4	2	1.42E-01	1.72E-01	1.05E+01	3.19E-03	2.59E-02
TIR1	6	2	1.42E-01	1.72E-01	8.79E+00	3.19E-03	2.00E-02
TIR1	7	2	1.42E-01	1.72E-01	7.07E+00	3.19E-03	4.00E-02
TIR1	8	2	1.42E-01	1.72E-01	2.92E+00	3.19E-03	5.02E-02
TIR1	9	2	1.42E-01	1.72E-01	7.56E+00	3.19E-03	4.07E-02
TIR1	10	2	1.42E-01	1.72E-01	4.76E+00	3.19E-03	2.74E-02
TIR1	11	2	1.42E-01	1.72E-01	4.71E+00	3.19E-03	1.54E-02
TIR1	12	2	1.42E-01	1.72E-01	8.83E+00	3.19E-03	3.45E-02
TIR1	13	2	1.42E-01	1.72E-01	8.46E+00	3.19E-03	4.07E-02
TIR1	14	2	1.42E-01	1.72E-01	5.32E+00	3.19E-03	3.05E-02
TIR1	15	2	1.42E-01	1.72E-01	5.71E+00	3.19E-03	2.14E-02
TIR1	17	2	1.42E-01	1.72E-01	6.30E+00	3.19E-03	4.23E-02
TIR1	18	2	1.42E-01	1.72E-01	8.19E+00	3.19E-03	3.51E-02
TIR1	19	2	1.42E-01	1.72E-01	5.39E+00	3.19E-03	3.22E-02
TIR1	20	2	1.42E-01	1.72E-01	9.59E+00	3.19E-03	3.54E-03
TIR1	26	2	1.42E-01	1.72E-01	9.27E+00	3.19E-03	2.85E-02
TIR1	27	2	1.42E-01	1.72E-01	3.92E+00	3.19E-03	3.78E-02
TIR1	28	2	1.42E-01	1.72E-01	7.87E+00	3.19E-03	3.02E-02
TIR1	31	2	1.42E-01	1.72E-01	6.81E+00	3.19E-03	6.23E-03
TIR1	32	2	1.42E-01	1.72E-01	2.55E+00	3.19E-03	5.39E-03

Table 5.3: Estimated parameters for IAA|AFB2 pairs using ζ_{23}.

AFB	IAA	Replicate	k_1	k_2	k_3	k_4	k_5
AFB2	1	1	1.49E-01	1.18E-01	4.48E+00	3.63E-03	4.59E-02
AFB2	1.1	1	1.49E-01	1.18E-01	9.77E+00	3.63E-03	4.47E-02
AFB2	2	1	1.49E-01	1.18E-01	8.09E+00	3.63E-03	4.65E-02
AFB2	3	1	1.49E-01	1.18E-01	1.13E+01	3.63E-03	6.14E-02
AFB2	4	1	1.49E-01	1.18E-01	1.16E+01	3.63E-03	6.05E-02
AFB2	6	1	1.49E-01	1.18E-01	8.55E+00	3.63E-03	7.06E-02
AFB2	7	1	1.49E-01	1.18E-01	7.19E+00	3.63E-03	8.35E-02
AFB2	8	1	1.49E-01	1.18E-01	3.16E+00	3.63E-03	1.05E-01
AFB2	9	1	1.49E-01	1.18E-01	8.01E+00	3.63E-03	7.29E-02
AFB2	10	1	1.49E-01	1.18E-01	5.80E+00	3.63E-03	5.16E-02
AFB2	11	1	1.49E-01	1.18E-01	5.87E+00	3.63E-03	2.08E-02
AFB2	12	1	1.49E-01	1.18E-01	9.16E+00	3.63E-03	5.04E-02
AFB2	13	1	1.49E-01	1.18E-01	9.57E+00	3.63E-03	5.52E-02
AFB2	14	1	1.49E-01	1.18E-01	6.25E+00	3.63E-03	6.45E-02
AFB2	15	1	1.49E-01	1.18E-01	7.94E+00	3.63E-03	2.47E-02
AFB2	17	1	1.49E-01	1.18E-01	6.04E+00	3.63E-03	1.04E-01
AFB2	18	1	1.49E-01	1.18E-01	8.57E+00	3.63E-03	5.80E-02
AFB2	19	1	1.49E-01	1.18E-01	5.96E+00	3.63E-03	5.65E-02
AFB2	20	1	1.49E-01	1.18E-01	1.07E+01	3.63E-03	2.67E-03
AFB2	26	1	1.49E-01	1.18E-01	1.06E+01	3.63E-03	4.42E-02
AFB2	27	1	1.49E-01	1.18E-01	4.49E+00	3.63E-03	6.19E-02
AFB2	28	1	1.49E-01	1.18E-01	8.85E+00	3.63E-03	4.14E-02
AFB2	31	1	1.49E-01	1.18E-01	7.50E+00	3.63E-03	1.07E-02
AFB2	32	1	1.49E-01	1.18E-01	2.85E+00	3.63E-03	2.51E-03
AFB2	1	2	1.27E-01	8.77E-02	3.72E+00	3.25E-03	4.46E-02
AFB2	1.1	2	1.27E-01	8.77E-02	9.23E+00	3.25E-03	3.59E-02
AFB2	2	2	1.27E-01	8.77E-02	7.84E+00	3.25E-03	3.82E-02
AFB2	3	2	1.27E-01	8.77E-02	8.45E+00	3.25E-03	5.28E-02
AFB2	4	2	1.27E-01	8.77E-02	1.02E+01	3.25E-03	5.87E-02
AFB2	6	2	1.27E-01	8.77E-02	8.10E+00	3.25E-03	6.08E-02
AFB2	7	2	1.27E-01	8.77E-02	6.33E+00	3.25E-03	7.79E-02
AFB2	8	2	1.27E-01	8.77E-02	2.66E+00	3.25E-03	9.07E-02
AFB2	9	2	1.27E-01	8.77E-02	7.16E+00	3.25E-03	5.58E-02
AFB2	10	2	1.27E-01	8.77E-02	5.49E+00	3.25E-03	5.23E-02
AFB2	11	2	1.27E-01	8.77E-02	5.63E+00	3.25E-03	2.04E-02
AFB2	12	2	1.27E-01	8.77E-02	8.07E+00	3.25E-03	4.96E-02
AFB2	13	2	1.27E-01	8.77E-02	8.90E+00	3.25E-03	4.47E-02
AFB2	14	2	1.27E-01	8.77E-02	5.43E+00	3.25E-03	7.30E-02
AFB2	15	2	1.27E-01	8.77E-02	7.37E+00	3.25E-03	2.29E-02
AFB2	17	2	1.27E-01	8.77E-02	5.05E+00	3.25E-03	1.12E-01
AFB2	18	2	1.27E-01	8.77E-02	8.16E+00	3.25E-03	4.71E-02
AFB2	19	2	1.27E-01	8.77E-02	4.67E+00	3.25E-03	7.35E-02
AFB2	20	2	1.27E-01	8.77E-02	9.29E+00	3.25E-03	3.03E-03
AFB2	26	2	1.27E-01	8.77E-02	9.99E+00	3.25E-03	3.77E-02
AFB2	27	2	1.27E-01	8.77E-02	3.93E+00	3.25E-03	6.17E-02
AFB2	28	2	1.27E-01	8.77E-02	7.56E+00	3.25E-03	4.12E-02
AFB2	31	2	1.27E-01	8.77E-02	6.90E+00	3.25E-03	1.06E-02
AFB2	32	2	1.27E-01	8.77E-02	2.68E+00	3.25E-03	4.25E-03

The normalized degradation curves in Figure 5.5 further provide two notable features of the model. First, they suggest that the two parameter k_3 and k_4 are, in fact, dependent parameters that can be consolidated into a single parameter by normalizing the output. The normalized version of f_2 is

$$\begin{aligned} \frac{dx}{dt} &= k_1\,u(t) - k_2\,x(t) \\ \frac{dz}{dt} &= k_4 - k_4\,z(t) - k_5\,x(t)\,z(t), \end{aligned} \tag{5.18}$$

where $z = y/y_0 = yk_4/k_3$, the output normalized to its initial value. This procedure further reduces the number of parameters to be estimated and decreases computational cost in estimation. The differences in the estimated values of k_5 are negligible between the two versions of model. A second notable feature of the model is that the range of degradation profiles in experimental data can be fitted just as well by varying k_1 instead of k_5. This feature is not surprising as varying k_1 increase the rate at which x increases, ultimately having the same effect on the output. However, because of the way we interpret the model parameters, k_1 is independent of the identity of IAA and conflicts with our objective of identifying a degradation rate parameter that are unique to each IAA|AFB pair. Therefore, we choose k_5 as the single parameter that captures the differential degradation range we observe among the IAA|AFB pairs.

Through iterative searches in model discrimination and parameter reduction, we identified a single parameter that captures the differential degradation exhibited by the family of IAA|AFB pairs. The reasoning behind finding a single parameter is largely inspired by the engineering principles of modularity and composable parts. For example, an electric circuit is composed of individual modular parts, such as resistors and transistors. The function of the circuit is tunable by swapping

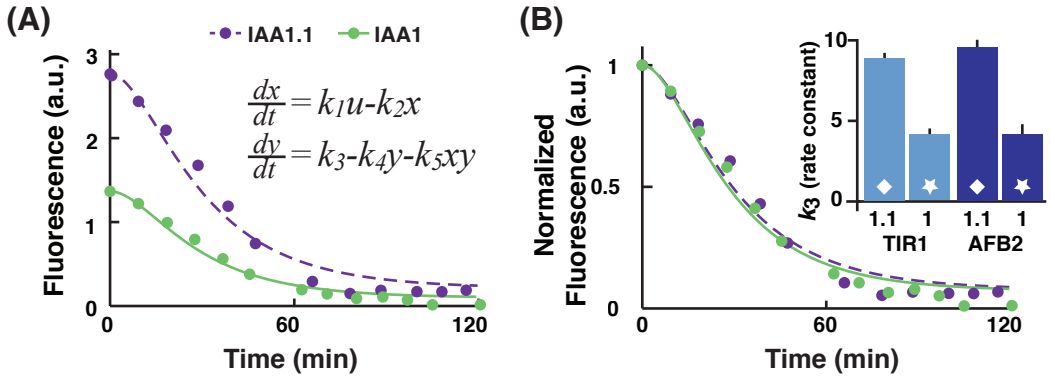


Figure 5.4: Degradation dynamics can be described using few parameters. (A) Our model is described by two ODEs. Degradation curves for AFB2 strains expressing IAA1 (blue) or yeast codon-optimized IAA1.1 (purple) are shown. (B) k_5 is largely independent of expression levels. IAA1 and IAA1.1 degradation curves overlap after normalization, although there is an approximately 2-fold difference in k_3 values.

out these modular parts, where the electrical functionality of each type of modular part is specified by a number (i.e. a resistor is specified by its resistance). Casting this principle on the auxin signaling pathway, we ask whether the IAA|AFB pair degradation module (in the larger scheme of the auxin signal pathway) is a modular part, and if so, whether the biological functionality of the module can be specified by a number. The functional feature of the IAA|AFB module, degradation, demonstrates a large range of responses. Therefore, we hypothesize that the number that specifies each IAA|AFB pair's unique biological functionality is its degradation rate (k_5).

Now that we have a data-sheet of k_5 for a large group of IAA|AFB pair (Figure 5.6), the hypothesis regarding the modularity of these pairs must be verified. It will require an auxin synthetic network containing an IAA|AFB pair, where the part can be easily replaced with another pair. Using a similar approach, a succinct mathematical representation of such a network can be devised, and preferably, the k_5 identified in this work will be a part of such a model. If a coherent function

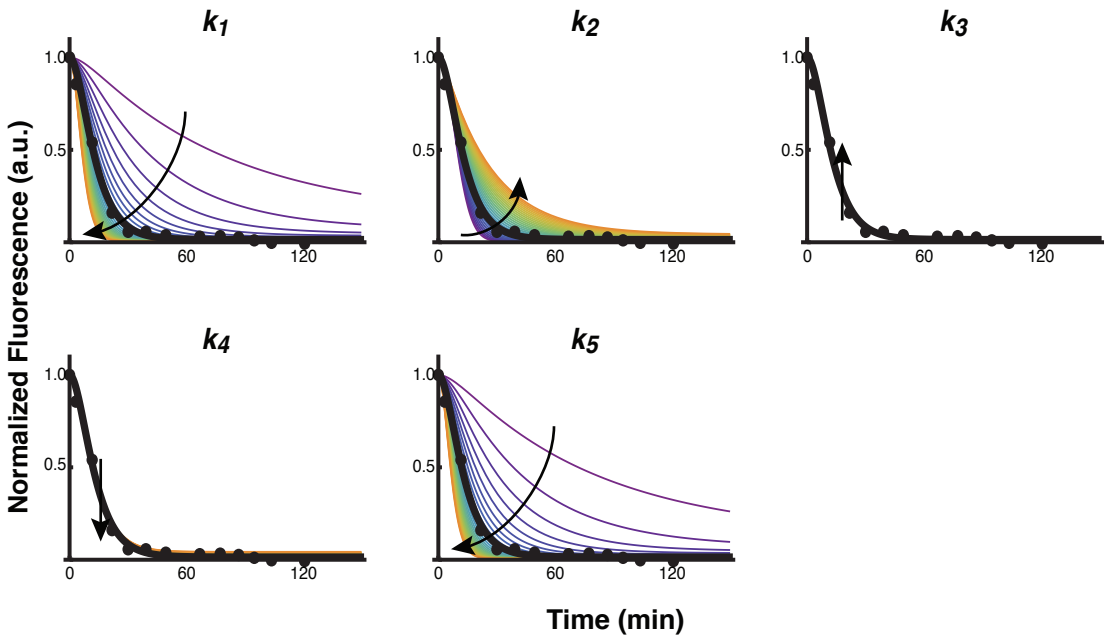


Figure 5.5: Parameter variations study of f_2. Parameters of f_2 are varied from 10% to 300% of the nominal (estimated) values for IAA17|AFB2. The range was chosen based on the demonstrated range of estimated parameters shown in Supplementary Table 7. The trend demonstrated here is general for all IAA |AFB pairs. The resulting degradation curves are normalized to its initial condition. The experimental data are shown in black dots and the black arrows indicate the direction of parameter value increase.

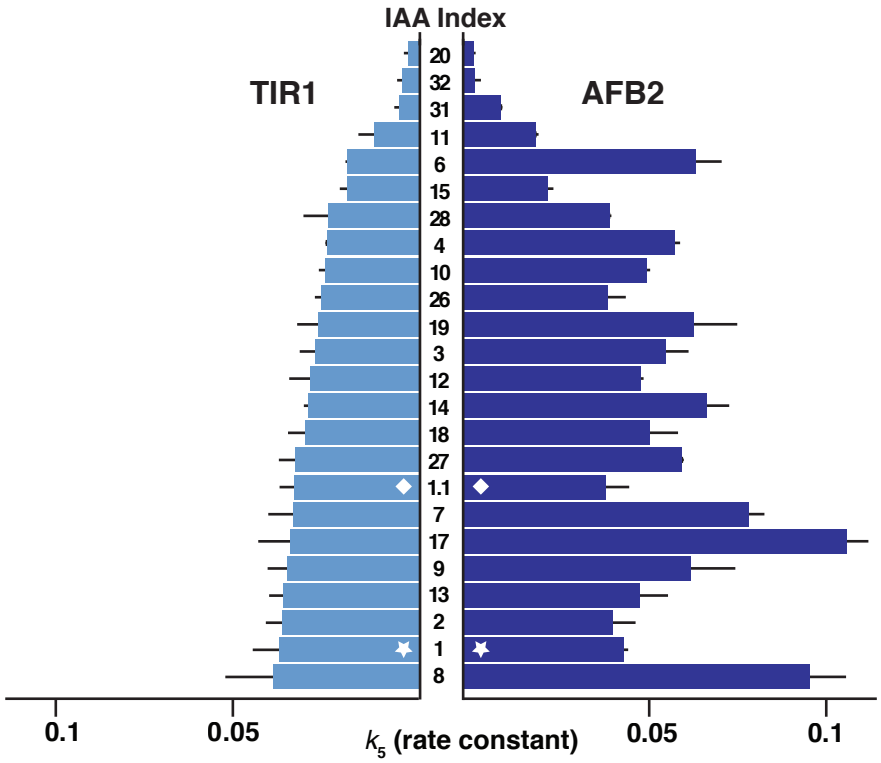


Figure 5.6: The biological function of the IAA|AFB module quantified. IAA|AFB2 pairs have increased degradation rates (k_5), a different rank order when compared to IAA|TIR1 pairs, and an increased dynamic range between the slowest and fastest pairs. Parameters were estimated for two independent replicates. All error bars represent one standard deviation.

that maps the differential degradation rates (caused by using different IAA|AFB pairs) and the differential outputs of the larger network can be identified, it will allow us to test to what extent and under what context these pairs are modular. And if this is not the case, the question remains whether there is another set of identifiable quantities for the IAA|AFB pairs that is a better specification of their biological functionality and predictors of the composite network output. Furthermore, the relationship between the k_5 and the output of the network will illuminate the core interactions within the network, aiding in increasingly accurate mathematical representations of the synthetic network. These approaches will not only answer questions regarding the engineerability of auxin signaling pathway, but also provide the basis for novel ways of system identification in biology.

5.4 ANOVA and PCD

The PCD approach of identifying the dependency relationship between model parameters and system components benefits engineering new synthetic biology systems by guiding the strategies to tune new systems. This is done by 1) simulation studies with identified models that illuminate a small subset of parameters that can significantly alter the outcome in a desired way, 2) by adjusting the components that these parameters are dependent to, achieve the same alteration of the output. Therefore, the PCD approach can be viewed as a way of establishing a relationship between system behavior and system components, aided by establishing a relationship between the system components and model parameters. Our objective is then similar to the objective of Analysis of Variance (ANOVA), in which the source of variation in experimental outcome is appropriated to experimental variables. In [85], it was shown that using ANOVA, variations in experimental data can be attributed to multiple elements in a genetic expression cassette by inspecting a cross-product family of these genetic expression elements. Our approach aims to uncover the relationship between

Table 5.4: ANOVA table for components and their interactions for $\hat{k}_{1-5}$.

	Component	Df	Sum.Sq	% Sum. Sq	Mean.Sq	% Mean Sq	F value	P value
$\hat{k}_1$	AFB	1	6.92E-04	0.1	6.92E-04	2.5	5.52E-02	8.15E-01
	IAA	23	1.59E-01	17.1	6.90E-03	25.2	5.51E-01	9.39E-01
	AFB:IAA	23	1.66E-01	17.9	7.22E-03	26.4	5.77E-01	9.23E-01
	Residuals	48	6.01E-01	64.9	1.25E-02	45.8		
$\hat{k}_2$	AFB	1	4.37E-07	0.0	4.37E-07	0.0	7.38E-06	9.98E-01
	IAA	23	1.87E+00	33.7	8.12E-02	46.0	1.37E+00	1.76E-01
	AFB:IAA	23	8.29E-01	15.0	3.60E-02	20.4	6.09E-01	9.01E-01
	Residuals	48	2.84E+00	51.3	5.92E-02	33.6		
$\hat{k}_3$	AFB	1	1.60E+01	2.1	1.60E+01	35.3	5.05E+00	2.93E-02
	IAA	23	4.95E+02	64.2	2.15E+01	47.4	6.78E+00	1.33E-08
	AFB:IAA	23	1.07E+02	13.9	4.67E+00	10.3	1.47E+00	1.28E-01
	Residuals	48	1.52E+02	19.8	3.17E+00	7.0		
$\hat{k}_4$	AFB	1	7.81E-06	11.0	7.81E-06	79.3	1.17E+01	1.31E-03
	IAA	23	1.55E-05	21.7	6.73E-07	6.8	1.00E+00	4.78E-01
	AFB:IAA	23	1.59E-05	22.3	6.91E-07	7.0	1.03E+00	4.50E-01
	Residuals	48	3.22E-05	45.1	6.70E-07	6.8		
$\hat{k}_5$	AFB	1	5.66E-02	24.0	5.66E-02	89.2	6.26E+01	3.00E-10
	IAA	23	8.64E-02	36.6	3.76E-03	5.9	4.15E+00	1.59E-05
	AFB:IAA	23	4.97E-02	21.1	2.16E-03	3.4	2.39E+00	5.47E-03
	Residuals	48	4.34E-02	18.4	9.05E-04	1.4		

system components (experimental variables) and model parameters, rather than experimental outcome. Therefore, we explore whether using the ANOVA methods on this particular problem set up is suitable for achieving our objective.

We used a short R code (Appendix B) to generate the ANOVA summary coefficients of the IAA|AFB experimental data (shown in Table 5.4). For each parameter estimate (estimated assuming no constraints over the parameters to ensure that the analysis is uncorrupted by *a priori* assumption), we assume it is a linear combination of effects caused by AFB, IAA and the interaction between AFB and IAA. The relative fractions of Sum of Squares and the Mean of Squares are indicative of the significance of each component attributing to the variations in the estimated parameters. For $\hat{k}_1$ and $\hat{k}_2$, it is shown that residuals (replicate errors) are a significant source of variations. However, the large variations between replicate estimates can be attributed to various factors such as measurement

noise, estimation noise and parameter uncertainty. For $\hat{k}_3$, 35.3% and 47.4% of variations are attributed to the difference arising from AFB and IAA, respectively. Given that there is a very small difference in the initial expression levels between each pair of system variants having the same IAA and different AFB, and because $\hat{k}_3$ is effectively a proxy for the initial expression level, this result is surprising. For $\hat{k}_4$ and $\hat{k}_5$, majority of variations in the estimated values are attributed to AFB, and very little is attributed to IAA, or the interaction between AFB and IAA. However, rough inspection of the experimental data suggests that this may not be the case. Since this is a bastardization of the ANOVA method to solve our problem of associating variations in parameter to variations in system components, we conclude that due to the large uncertainty that arises in unconstrained parameter estimation, it is inconclusive to determine whether the ANOVA analysis is appropriate for uncovering dependent relationships between model parameters and system components[7].

5.5 Summary

The quantitative analysis presented in this chapter was divided into two phases: in the first phase, the structure of the model was optimized and in the second phase, parameter reduction options were explored with the model from the first phase. Both phases of analysis were driven by the goal of having a succinct, quantitative representation that captured the source of variation observed in the degradation rates of IAA due to AFBs. The results discussed in Section 5.3 precede, and therefore motivate, the formalization of the PCD method introduced in Chapter 3.

Section 5.2 asks the question "is the model good enough?", as in whether the model is adequate by comparing the error variances estimated by the model and those obtained experimentally. If the model predicted error variance is deemed smaller than the experimental error variance, the model is

[7] In Appendix C, we further investigate the approaches presented in [85], by applying the PCD analysis to their system.

determined to be adequate [86]. This approach is quantitative in the sense that the performance of the model is assumed to be purely numerical. However, we observed cases where the prediction of a model is categorically different from an experimental response. For example, a linear model was used to fit an experimental dose response adequately, even when the system behavior is fundamentally nonlinear. Because we had a fairly strong conviction about the synthetic auxin signal pathway and its underlying mechanism, we rejected models that are fundamentally faulty.

Section 5.3 asks the question "among several rival models, which is the best one?", and demonstrates a form of model discrimination that requires modelers to weigh the different aspects of models - performance, complexity and predictive power. As the alternative hypotheses discussed in Section 5.3 explore the dependent relationships between the model parameters (k_1 through k_5) and the interchangeable system components (IAA and AFB), it is noted that the search for a smaller subset of parameters to represent each unique system variant leans on heuristics and our interpretation of the model. The synthetic auxin signal pathway system and its 2 ODEs model, compared to the one in Chapter 4, result in a search space 2^4 times larger and the computational cost of applying the same systematic search methods is prohibitively expensive. Therefore, taking advantage of the *a priori* knowledge of the underlying mechanisms is preferred, which allowed us to verify and discriminate multiple hypotheses comparing their performances to the benchmark performance of the unconstrained hypothesis. Considering that we observed multiple M that perform equally well, it was acceptable to choose the hypothesis that is the closest to its description with the known mechanism.

Moving forward, it would be beneficial to systematically design experiments for gaining better insights of the system that discriminate competing hypothesis (i.e. multiple M with equal cost).

For example, identifying a metric over the the possible input signals that measures their ability to distinguish previously equally performing M would help us determine whether some experiments are worth doing [87, 88]. Additionally, it would be interesting to investigate whether a set of competing M reveals fundamental characteristics of the identified model of the system. It is shown that owing to the structure of a model itself and various experimental limitations, some model parameters are indistinguishable [89]. This observation implies that for a pair of Ms that are only different in the position of the same pairs of indistinguishable parameters in its non-zero entries will yield an approximately same cost[8] because variations caused by these indistinguishable parameters would result in the same model-predicted outcome. By using the analytical methods discussed in [89] to determine whether the auxin model (Eq 5.9) contains indistinguishable parameters[9], and examining groups of Ms that contain these parameters, we would be able to develop an alternative methods by which to identify parameter indistinguishability framed by PCD matrices[10].

[8] within some error due to estimation error

[9] From Figure 5.5, and experimental observations, one would suspect that it does

[10] Some analysis addressing this problem is discussed in Appendix C.

Chapter 6

AUXIN SIGNAL PATHWAY II: ARF ACTIVATION SYSTEM

The material presented in this chapter has been published in [9]. The pathway presented in this chapter is an extension of the pathway discussed in Chapter 5 synthesized by adding another family of mechanistically homogeneous proteins called the Auxin Response Factors (ARFs) found in *Arabidopsis thaliana*. We aimed to identify a model describing the new system with minimal modification to the model of the previous system. Additionally, we verified the result from Chapter 5 by testing the various possible hypotheses regarding the parameter-component dependency arising within the new system. The material has been edited with updated notations where possible and marked with footnotes for commentaries.

6.1 Introduction

Evolution depends on the plasticity of existing signaling pathways. The small molecule auxin is linked to signaling modules that allowed plants to move to land, develop new organs and respond to the environment [90, 91]. Despite the wide range of auxin responses, the core auxin signal transduction pathway is quite simple, involving a small number of components from perception through transcription. Auxin triggers the rapid turnover of Aux/IAA (IAA) proteins, which repress the activity of Auxin Response Factor (ARF) transcription factors through recruitment of TOPLESS (TPL) co-repressors [92, 93]. Auxin receptors (AFBs) are F-box proteins, which act as part of an E3 ubiquitin ligase to catalyze the ubiquitination and subsequent degradation of IAAs when auxin is

present [79, 94]. ARFs bound to auxin-responsive cis-regulatory elements (AuxREs) are then free to regulate the expression of auxin target genes, which include the IAAs themselves [95, 96, 97].

How such a simple pathway can orchestrate the large number of context-specific responses regulated by auxin is a long-standing question. Notably, each component in the auxin signaling pathway belongs to a large gene family ([2, 91, 98]. In Arabidopsis, there are 6 AFBs, 23 ARFs and 29 IAAs. Functional divergence between component family members could provide variation in response to a generic auxin signal [76]. Members of the ARF family can be classified as transcriptional activators or repressors ([95, 96], and the role of repressor ARFs in regulating auxin signaling is still a matter of debate [99]. In addition, expression studies of auxin signaling gene families support the idea of an "auxin pre-pattern" where certain combinations of co-expressed components generate auxin circuits with distinct auxin response capabilities [100, 101]. However, due to the high level of redundancy and co-expression of auxin signaling component family members, feedback, and interference from other signaling pathways [102], the contributions of individual components to auxin circuit behavior have remained elusive.

Here, we ported several variations of the auxin transcriptional response pathway from *Arabidopsis thaliana* to *Saccharomyces cerevisiae*. In so doing, we defined a minimal Auxin Response Circuit (ARC) sufficient to recapitulate auxin-induced transcription in a heterologous context. As no feedback was engineered into these ARCs, they captured the dynamic potential of the simplest forward response architecture. The ARC consists of five plant components: an AFB, an IAA, TPL, an activating ARF transcription factor, and an auxin-responsive plant promoter (Figure 6.1). The implementation of ARCs in yeast (ARC^{Sc}) relied on successful interfacing of the ARC with other essential elements of the yeast degradation and transcriptional machinery (Figure 6.1B). This

modular, plug-in nature of the ARC, similar to what was observed with Auxin-Induced Degradation [103], highlights the likelihood that the ARC could be implemented in a variety of other eukaryotic contexts. The suite of ARC variants represented in the ARCSc collection presented an opportunity to quantitatively investigate the dynamic capabilities of the auxin response in isolation.

6.2 Model Identification

The model mirrors the construction of the ARC^{sc} (Figure 6.2): Just as an ARF and a promoter were added to the previously engineered AFB|IAA system, an additional equation for GFP (represented by g) was added to the previously identified mathematical model of the AFB|IAA system [8] to obtain the following.

$$\begin{aligned} \dot{x} &= k_1 u - k_2 x \\ \dot{y} &= k_3 - k_4 y - k_5 x y \\ \dot{g} &= -k_6 g + \frac{k_7}{1 + k_8 y}. \end{aligned} \tag{6.1}$$

The time-dependent variables u, x, y and g represent the concentrations of auxin input, a lumped internal state (combining the intermediate reactions involving the binding of auxin to the AFB receptor and related molecular machinery), the IAA levels, and the GFP reporter, respectively. The parameters describe the synthesis of the internal state (k_1), the degradation and dilution of the internal state (k_2), expression of IAA (k_3), degradation and dilution of IAA (k_4), auxin-induced degradation of IAA (k_5), degradation and dilution of GFP (k_6), non-repressed expression of the reporter (k_7), and the repression on the reporter (k_8). The first two equations in the model are identical to the model of the AFB|IAA system reported previously in Chapter 5. The third equation

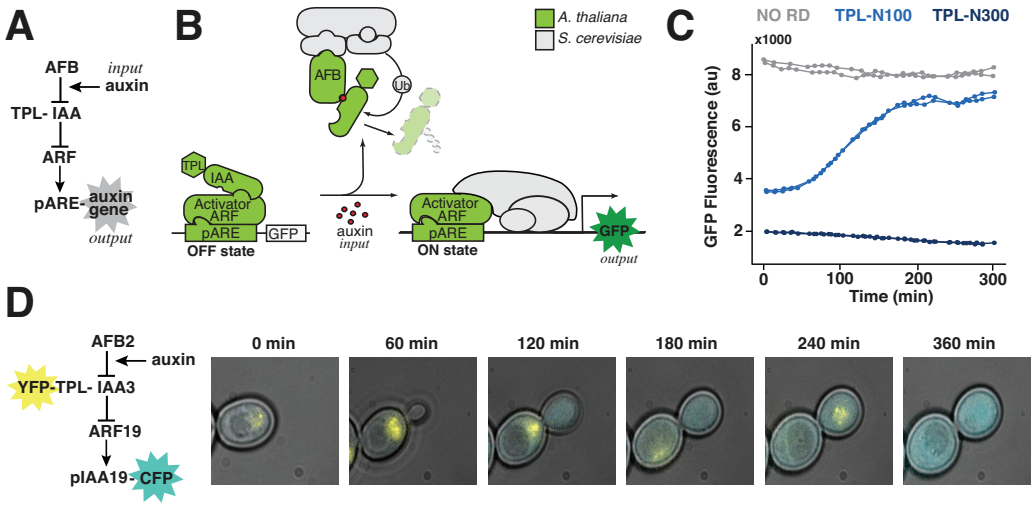


Figure 6.1: Auxin-induced transcription in yeast. (A) Network diagram of the forward auxin response pathway in yeast. An auxin input increases association of a member of the TIR1/AFB family of F-box proteins (AFB) and an IAA protein fused to the first 100 amino acids of the TPL co-repressor (TPL-IAA). Auxin-induced association of an AFB and a TPL-IAA leads to degradation of the TPL-IAA, thereby freeing a transcriptional activator in the ARF family to induce expression of an output gene driven by a promoter containing an Auxin Response Element (pARE). (B) The five *A. thaliana* components needed to recapitulate auxin response in *S. cerevisiae* are shown in light green. They were: an AFB F-box receptor, an IAA, a TPL co-repressor, an ARF transcription factor and an auxin responsive promoter. The remaining cellular machinery (grey) was supplied by yeast. Fluorescence from a GFP reporter was used as a quantitative output. (C) Synthetic auxin-reversible repression required fusion of a specific TPL truncation to the IAA protein. Flow cytometry was used to monitor the induction of a GFP reporter following auxin treatment in circuits containing either IAA14 with no repression domain (NO RD), shown in grey, or two different C-terminal TPL truncations fused to IAA14. Auxin was added at time zero. TPL-N300 (dark blue) includes the first 300 amino acids of TPL and excludes the multiple C-terminal WD repeats. Fusion of TPL-N300 to IAA proteins results in reporter repression that is largely auxin insensitive. TPL-N100 (light blue) includes only the first 100 amino acids of TPL. When fused to an IAA, TPL-N100 provides auxin-reversible repression. Two replicate induction curves are shown for each circuit. (D) Auxin-induced IAA degradation and subsequent transcriptional activation could be simultaneously monitored in dual-labeled yeast strains. YFP-TPL-IAA3 and a Cerulean reporter driven by the auxin responsive IAA19 promoter (pIAA19-CFP) were monitored following auxin treatment using time-lapse microscopy and a microfluidic chamber.

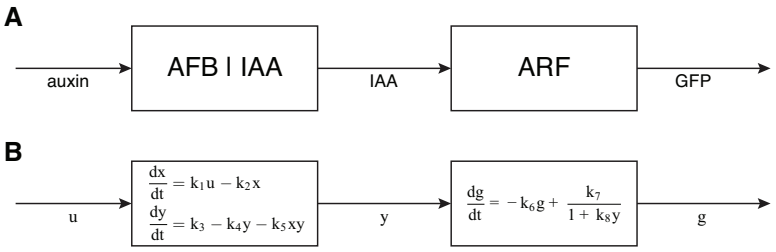


Figure 6.2: The block diagram of the ARC^{sc} and its model. (A) The subcomponent, AFB|IAA, was characterized in Chapter 5. The ARF and GFP components are added to synthesize the full circuit ARC^{sc} presented in this chapter. The output of the first component, IAA, interacts with ARF and affects the system output, GFP. (B) The model for the ARC^{sc} is extended by adding a third equation for the GFP dynamics. The block diagram illustrates the analogous composition of both the physical system and the mathematical model identification presented.

captures the dynamics of GFP output – the basal degradation and dilution rate, and the IAA dependent expression rate. Because IAAs bind with ARF and inhibit the activation of GFP expression, the GFP expression rate is inversely proportional to the amount of IAA. The model is not intended to be mechanistic. Nevertheless, even without additional terms modeling ARF dynamics (e.g. by directly modeling the inhibitory effect of IAA on GFP expression), the model fits the wide range of GFP induction responses (Figure 6.3). In fact, adding a species representing the concentration of ARF would introduce uncertainty to the parameter estimation that cannot be resolved because of the low number of observable outputs.

6.3 PCD Analysis

The main question we answer in our analysis of the model is: which of the parameters k_1 through k_8 are tuned by the choice of ARF or IAA. We call the proteins that make up the circuit, and in particular the ARF and IAA, *components*. As can be seen from the data, the auxin response clearly depends on the choice of components. For those aspects of the dynamics that are independent of the changing components, we expect that the model parameters associated with the components

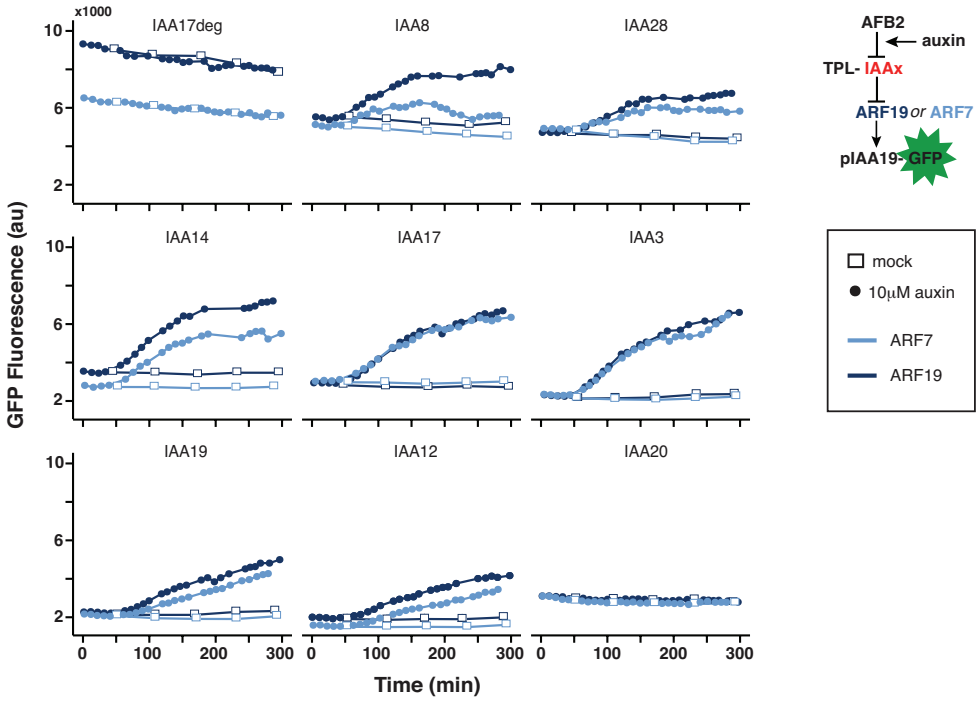


Figure 6.3: IAAs drive auxin response dynamics. Representative auxin-induced reporter fluorescence curves are shown for ARCs with ARF7 (light blue) or ARF19 (dark blue) in the context of different IAAs. Auxin was added at time zero in all graphs. ARF7 and ARF19 showed qualitatively similar patterns of auxin response, while the identity of the IAA had a dramatic effect on ARC dynamics. ARC^{sc} recapitulated regulatory features of plant ARC function. Transcriptional repression required the known ARF-IAA interaction domain, as an IAA lacking this domain (IAA17deg) had no effect on ARF activity. In addition, auxin response was mediated by IAA degradation, as a naturally occurring IAA lacking a degron (IAA20) rendered the circuit insensitive to auxin treatment.

remain constant across ARC^{sc} variants. For example, k_1 is a parameter associated with the choice of AFB [8] and when estimating the parameters of ARC^{sc} circuits, all with the same AFB variant, the value for k_1 is constant for all members of the set. However, k_3 is associated with the choice of IAA; therefore, when estimating the parameters of ARC^{sc} circuits with different IAAs, we expect the parameter values to be different for each variant. This dependency implies that every dependent parameter becomes a variable unique to an ARC^{sc} variant (made up of unique constituent components), and the set of these unique variables add to the number of characteristic features we use to compare ARC^{sc} variants with one another. Since we seek a small set of quantitative features for simpler comparison across ARC^{sc} variants for parsimonious representation, we aim to keep the number of dependencies small.

We use a matrix to represent the relationship between parameters (k_1 through k_8) and components (the IAAs and ARFs). In particular, we define the $m \times n$ Boolean matrix M for m-parameters and n-components. Note that n refers to the number of component types that compose the system, and not the number of interchangeable variants for a given component type. Here $m = 8$ and $n = 2$. The entry $M(\kappa, i)$ is 1 if the κ-th parameter is dependent on the choice of the i-th component, and is 0, otherwise.

To validate our interpretation of the parameters k_1 through k_8, we estimate M using pooled experimental data. Note that there exists $2^{m \times n}$ candidates for M and each candidate is a hypothesis regarding the parameter-component dependency. Some hypothesis perform better than others with respect to how well it fits the range of behavior across a suite of system variants. These candidates tend to have more non-zero entries because each non-zero entry effectively allows another degree-of-freedom for fitting the experimental data. At the same time, the candidates with more non-zero

entries ultimately result in larger number of unique variables per system variant. Therefore, we search for an M that results in low model-fit residual with small number of non-zero entries.

We implemented the model fitting and component matrix search in Mathematica (Appendix A). Given that there are 2^{16} possible candidate matrices for the 2-component ARC^{sc} circuits and the 8-parameter model, directly examining all of the candidates is prohibitively expensive in computational cost (approximately 2 hours per matrix using the NMinimize function in Mathematica). Instead, we selectively choose and evaluate candidate matrices using a direct evaluation and greedy search algorithm (Figure 6.4). Though efficient, the greedy algorithm sometimes fails to consider other candidate matrices that are nearly equivalent in model-fit residual to the lowest value. Repeated searches from different initial conditions, however, gave the same results in this case. Generally, the model-fit residual decreases with increasing number of non-zero entries. However, because of various uncertainties in measurement/experimental noise and estimation errors, the residual eventually saturates at some low value and further addition of non-zero entries do not decrease it further. We defined an arbitrary threshold at which point the iterative search for optimal M halts.

The optimal matrix M^* (Figure 6.4) suggests the following interpretations regarding the ARC^{sc} and the model parameter.

- The expression strength of IAA (k_3) is one of the strongest determinants of the ARC^{sc} response dynamics as it was consistently chosen as the first non-zero entry in multiple iterations of greedy search algorithm. Additionally, the parameter is dependent on the choice of IAA alone, and is independent of the ARF.

- The basal expression of GFP (k_7) is dependent on the choice of ARF, but independent of the choice of IAA.

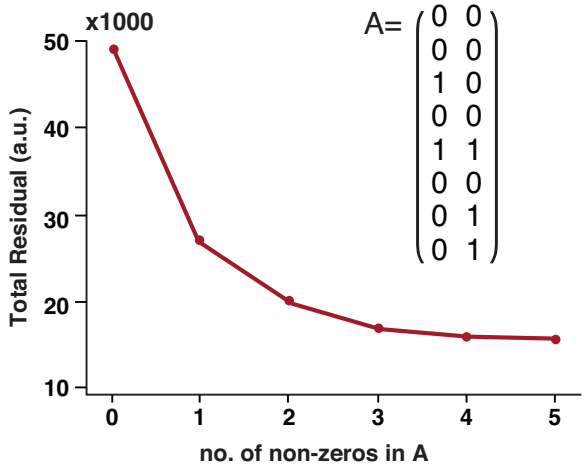


Figure 6.4: The model-fit residual as a function of the number of non-zero entries in the PCD. The figure shows an instance of greedy search algorithm. The model-fit residual decreases as the number of non-zero entries in the PCD increases.

– Somewhat unexpectedly, the auxin-mediated IAA degradation (k_5) is dependent on the choice of both IAA and ARF. One hypothesis is that ARF and AFB compete for binding with IAA and when IAA is bound to ARF, AFB cannot bind – inhibiting auxin-mediated IAA degradation.

– The affinity of IAA to ARF (k_8) is dependent on the choices of IAA and ARF, which is consistent with the model interpretation.

The findings of parameter-component dependency identified here could have been obtained, to some degree, through qualitative observation of ARC^{sc} responses (e.g. The minimal dependency of k_3 on the choice of ARF could have been elicited from noticing that the initial state of ARC^{sc} remains constant when a different ARF is used). However, by using the approach discussed here, we systematically investigated different hypotheses and determined that not all of the model parameters need be varied to account for the range of dynamic responses in ARC^{sc}. Additionally, we have

identified a minimal set of quantities that identify the biological functions of ARC^{sc}.

6.4 Sensitivity Analysis

We defined two performance metrics to capture some of the characteristics of the system: The initial state and the activation time. These metrics enable quantitative comparisons among the homologous yeast synthetic systems (and later on guide the rational engineering of new synthetic gene networks using the yeast synthetic system analyzed here).

The pre-auxin steady state. The following analytical expression for the initial state with $u = 0$ is

$$g_0 = \frac{k_7 k_4}{k_6 (k_4 + k_3 k_8)}. \tag{6.2}$$

Experimentally, it was shown that the maximal GFP intensity is dependent on the choice of ARF. Therefore, to fairly compare initial state of GFP among the yeast synthetic systems with different ARFs, the values were normalized by k_7/k_6, which corresponds to the maximal GFP expression rate when y is zero. Thus the normalized value represents the relative fraction of the initial states of GFP with respect to the maximal GFP intensity possible for the given ARF. Note that the relative rankings of IAAs in ARF 19 and ARF7 are conserved for this metric (Figure 6.5).

Sensitivity analysis reveals how the initial state changes with a change in the model parameters (Figure 6.6D) – specifically k_3, k_5, and k_8. Each parameter was varied one at a time from the minimum and the maximum estimated values. The pre-auxin steady state depends on the synthesis rate of IAA and the affinity of IAA to ARF. At the same time, because the pre-auxin steady state is measured before auxin is added to the system, the quantity is independent of the auxin-induced

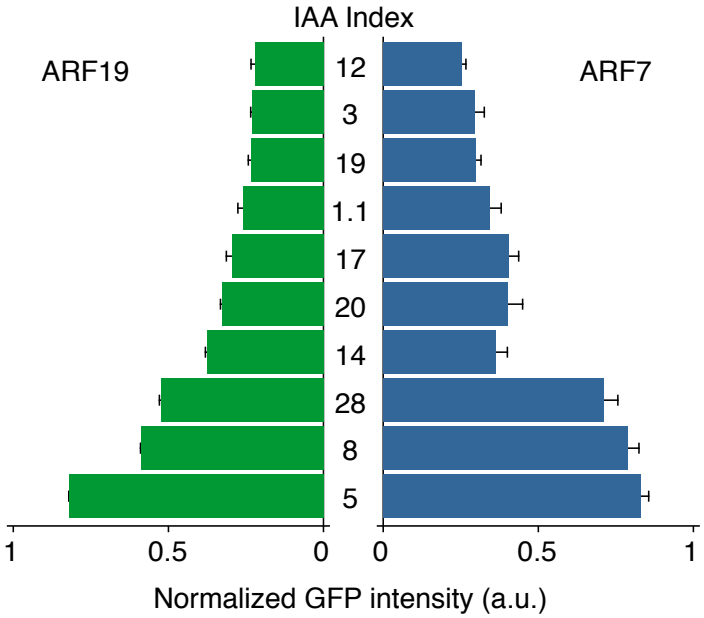


Figure 6.5: Initial state of the yeast synthetic system homologs. Total of 20 different system variants – the combinations of 2 different ARFs and 10 different IAAs – were evaluated. They are presented in the order of increasing GFP intensity from top to bottom in the ARF19. For each yeast synthetic system variants, two to four replicates exist and the error bar shows one standard deviation among the replicates.

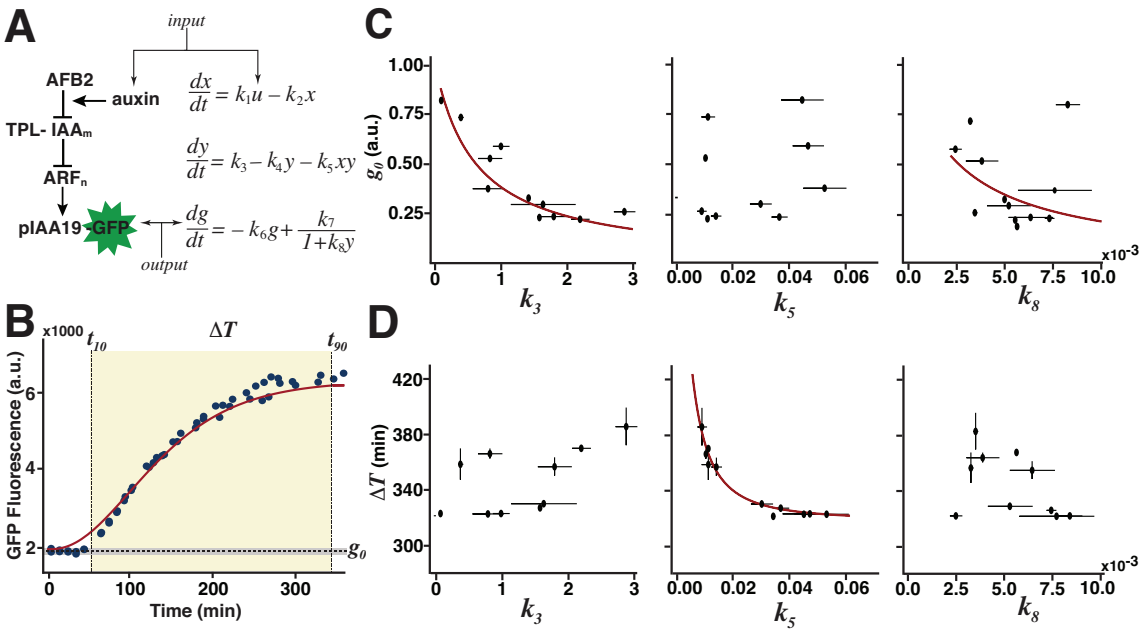


Figure 6.6: Model selection and sensitivity analysis of the auxin response pathway. (A) Grey-box model of the auxin response pathway. The auxin input is represented by the variable u and the GFP output of the reporter is represented by the variable g. The variable x represents a lumped internal state, which combines multiple reactions including the binding of auxin to the AFB receptor, and the variable y represents IAA protein levels. The model has eight parameters (k_1 to k_8) that intuitively correspond to the biological processes listed in Table 6.1. We focused on three parameters with strong effect on ARC dynamics: IAA expression level (k_3), rate of auxin-induced IAA degradation (k_5), and IAA-ARF affinity (k_8). (B) A graphical representation of the two performance metrics used for sensitivity analysis: the pre-auxin steady state g_0 and the activation time $\Delta T(t_{90} - t_{10})$. Sample data from one ARC is shown as blue dots with a sample model fit in red. (C, D) Sensitivity analysis of pre-auxin steady state g_0 (C) and activation time ΔT (D) to model parameter values of k_3, k_5 and k_8 for each IAA. Each parameter was varied across its entire range of estimated values derived from our experimental dataset, while all other parameters were held constant. Each IAA is plotted as a single point, and the red line indicates the sensitivity curve computed from the model. The pre-auxin steady state g_0 was accurately predicted by k_3 (expression level) and to a lesser extent by k_8 (ARF affinity) with k_5 having little effect. In contrast, activation time ΔT was predicted with high accuracy by k_5 (auxin-induced degradation) alone. Error bars represent one standard deviation ($n = 2$).

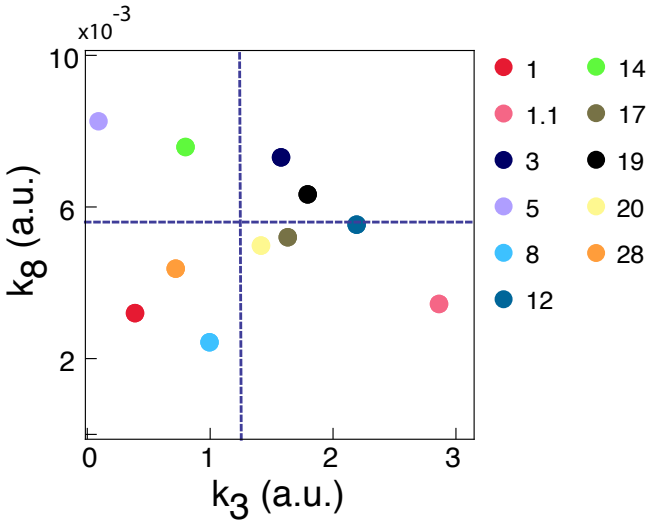


Figure 6.7: k_3 and k_8 may be controlled independently within the IAA sequence. The scatter plot shows k_3 and k_8 values for the 11 IAAs assayed with ARF19. Each strain has two to four replicates, and each dot on the scatter plot represents the mean of the estimated parameters over the replicates. The dashed line indicates the mean of k_3 and k_8 across the IAAs.

IAA degradation (k_5). Therefore, the initial state does not vary predictably as k_5 is changed, and the sensitivity analysis is consistent with this observation. Additionally, the effects of varying k_3 and k_8 are qualitatively equivalent – when either of the value is decreased, the initial state increases, and vice versa. This result is consistent with the interpretation of the model parameters. As k_3 represents the basal expression rate of IAA, increasing the value has an effect of increasing the concentration of IAA (y) present in the system (as long as k_4 and k_5 are fixed). Thus, it has the the effect of increasing the $k_8 y$ term in Eq 6.1. Considering the same effect can be simulated via increasing k_8, it is not surprising that these values would exhibit roughly equivalent sensitivity trend on the output, g.

The activation time, ΔT. We define the *activation time* as the difference in time between when 10% GFP expression intensity (t_{10}) and 90% GFP expression (t_{90}) are reached. We set the initial

Table 6.1: Parameter Interpretations

Parameter	Biological interpretation
k_1	assembly of auxin primed AFB receptor
k_2	basal degradation/dissociation of auxin primed AFB receptor
k_3	IAA expression
k_4	basal degradation & dilution of IAA
k_5	auxin-induced degradation of IAA
k_6	basal degradation & dilution of GFP
k_7	GFP expression
k_8	ARF affinity of IAA

state as the 0% and the final state (refers to the post-auxin steady-state GFP intensity) as 100%. The ODEs are numerically solved to approximate the activation time, because there is no analytical solution to the model. As in the sensitivity analysis of the pre-auxin steady state, each parameter was varied between the range of the estimated parameter values from the data, while the other parameters are held constant (Figure 6.6E). The most striking trend is that the activation time is predicted with high accuracy by k_5 alone and that k_3 or k_8 are poor predictors of the activation time. This suggests that when engineering a synthetic gene network using the ARC^{sc} circuit which requires faster/slower activation rate, it is far more advantageous to select an IAA with a higher/lower k_5 than one with a different k_3 or k_8.

The sensitivity analysis also suggests that varying k_3 and k_8 have similar effect on the ΔT. This observation enforces our suspicion that these parameters are indistinguishable when the output measurement is limited to the variable g. However, the ambiguity can be resolved if the output measurements are extended to the variable y, because it allows us to estimate k_3 – the expression rate of IAA – independently of the k_8. To verify this prediction, we pooled two data sets - GFP induction

data and YFP-IAA degradation data to simulate the ability to observe the IAA concentration as well as the GFP intensity. Though the YFP-IAA degradation data is collected from variations of ARC^{sc} circuit that lack the plant promoter, it is a close approximation of the standard ARC^{sc} circuit, and allowed us to estimate the approximate range of k_3 for ARC^{sc} circuits. Though mathematically indistinguishable in sensitivity, IAAs with different k_3 and k_8 values have differential effects on the circuit performance. In particular, the scatter plot of k_3 and k_8 shows little correlation, suggesting that the two parameters are controlled independently within the IAA sequence and have partially independent roles in output dynamics (Figure 6.7). We hypothesize that k_3 is determined by the transcriptional and translational efficiency of the IAA in question, whereas k_8 is determined by the affinity of the IAA for its target ARF.

6.4.1 Competition of multiple IAAs

We next engineered ARC^{Sc} competition variants that co-expressed pairs of IAAs to test whether the inherent properties of competing IAAs could create a hierarchical sequence of auxin responses. For the model identification, instead of modifying the model structure and introducing new variables and parameters to represent the secondary IAA, we reinterpret the standard model[1]. Specifically, the variable y was interpreted as the *effective* IAA. This is because without engineering the multiple-IAA ARC^{sc} to measure either or both of the IAA dynamics, it is difficult to decouple the effects of individual IAAs. Therefore, we assume that there exists a lumped species whose behavior is a combination of behaviors of the two IAAs.

As discussed before, the characteristic features of an IAA in ARC^{sc} are quantified by k_3, k_5 and k_8. Similarly, we assume the characteristic features of the effective IAA species of a mixed-

[1]The estimated parameters for the IAA-competition yeast synthetic systems are shown in the Table 6.3.

IAA ARC^{sc} are quantified by the same parameters. We further hypothesize that k_5 and k_8 of the mixed IAA circuit are linear combinations of the parametric values of each individual IAA with the weighting coefficient parametrized by α and β, respectively. These parameters approximate the relative competitiveness in AFB- and ARF-binding of one of the two IAA in a mixed ARC^{sc} system. This relationship is written as follows.

$$k_{5,\text{mixed-IAA}} = \alpha k_{5,\text{IAA}_{j_1}} + (1-\alpha) k_{5,\text{IAA}_{j_2}} \tag{6.3}$$

$$k_{8,\text{mixed-IAA}} = \beta k_{8,\text{IAA}_{j_1}} + (1-\beta) k_{8,\text{IAA}_{j_2}}. \tag{6.4}$$

It was shown that mixed-IAA response dynamics often mimic the response of one of the individual IAA ARC^{sc} circuit more closer than the other. To determine whether this trend is bolstered by either a strong relative competitiveness of AFB- or ARF-binding of a given IAA, we simulated the effect of varying α and β simultaneously. To examine the preferential response dynamics of a mixed ARC^{sc} system, we define a metric ω, that quantifies how closely the mixed ARC^{sc} mimics the IAA_{j_1} behavior, as follows.

$$\omega = \frac{\int |g(\theta_{(j_1,j_2)},t) - g(\theta_{j_2},t)|dt}{\int |g(\theta_{j_1},t) - g(\theta_{j_2},t)|dt}, \tag{6.5}$$

where $g(\theta_{(j_1,j_2)})$ is the response dynamics of a mixed ARC^{sc} that co-expresses IAA_{j_1} and IAA_{j_2}. IAA_{j_1} has near 0 dominance, when α and β are both zeros, and complete dominance when they are both ones. It was shown that with a low β (e.g. $\beta = 0$, IAA_{j_1} entirely loses the competition over ARF-binding to IAA_{j_2}), IAA_{j_1} can still dominate the response by having a high relative competitiveness in AFB-binding, α. On the other hand, with a low α, IAA_{j_1} cannot dominate

even with maximum β value. It is not surprising that competitiveness is better captured by α, since it corresponds to k_5 which was shown to affect the auxin response much more than k_8. Note that one cannot conclude from this analysis that a faster or slower k_5 predicts the preferential behavior in the two competing IAA circuits. Rather, we simply use α to describe which of two IAAs dominate by showing which effective k_5 results from the competition.

Table 6.2: Estimated parameter values of the 21 synthetic yeast system homologs and the replicates.

IAA	ARF	num. of rep	k_1	k_2	k_3	k_4
1.1	7	4	1.382E-01	1.027E-01	1.539E+00	3.440E-03
3	7	4	1.382E-01	1.027E-01	1.492E+00	3.440E-03
5	7	4	1.382E-01	1.027E-01	1.063E-01	3.440E-03
8	7	4	1.382E-01	1.027E-01	1.280E-01	3.440E-03
12	7	3	1.382E-01	1.027E-01	1.562E+00	3.440E-03
14	7	4	1.382E-01	1.027E-01	9.678E-01	3.440E-03
17	7	3	1.382E-01	1.027E-01	9.741E-01	3.440E-03
19	7	4	1.382E-01	1.027E-01	1.955E+00	3.440E-03
20	7	4	1.382E-01	1.027E-01	9.367E-01	3.440E-03
28	7	3	1.382E-01	1.027E-01	3.809E-01	3.440E-03
1	19	2	1.382E-01	1.027E-01	3.885E-01	3.440E-03
1.1	19	2	1.382E-01	1.027E-01	2.862E+00	3.440E-03
3	19	2	1.382E-01	1.027E-01	1.577E+00	3.440E-03
5	19	2	1.382E-01	1.027E-01	9.136E-02	3.440E-03
8	19	2	1.382E-01	1.027E-01	9.948E-01	3.440E-03
12	19	2	1.382E-01	1.027E-01	2.192E+00	3.440E-03
14	19	5	1.382E-01	1.027E-01	7.994E-01	3.440E-03
17	19	2	1.382E-01	1.027E-01	1.631E+00	3.440E-03
19	19	2	1.382E-01	1.027E-01	1.794E+00	3.440E-03
20	19	2	1.382E-01	1.027E-01	1.413E+00	3.440E-03
28	19	2	1.382E-01	1.027E-01	8.286E-01	3.440E-03

IAA	ARF	num. of rep	k_5	k_6	k_7	k_8
1.1	7	4	1.242E-02	6.890E-03	4.329E+01	4.391E-03
3	7	4	7.248E-02	6.890E-03	4.329E+01	6.288E-03
5	7	4	2.397E-02	6.890E-03	4.329E+01	6.724E-03
8	7	4	4.736E-02	6.890E-03	4.329E+01	7.869E-03
12	7	3	1.346E-02	6.890E-03	4.329E+01	6.542E-03
14	7	4	9.224E-02	6.890E-03	4.329E+01	6.509E-03
17	7	3	3.683E-02	6.890E-03	4.329E+01	4.406E-03
19	7	4	1.116E-02	6.890E-03	4.329E+01	4.307E-03
20	7	4	-8.677E-05	6.890E-03	4.329E+01	5.796E-03
28	7	3	3.248E-02	6.890E-03	4.329E+01	3.376E-03
1	19	2	1.130E-02	6.890E-03	6.228E+01	3.200E-03
1.1	19	2	9.072E-03	6.890E-03	6.228E+01	3.441E-03
3	19	2	3.650E-02	6.890E-03	6.228E+01	7.309E-03
5	19	2	4.457E-02	6.890E-03	6.228E+01	8.261E-03
8	19	2	4.664E-02	6.890E-03	6.228E+01	2.431E-03
12	19	2	1.129E-02	6.890E-03	6.228E+01	5.533E-03
14	19	5	5.248E-02	6.890E-03	6.228E+01	7.582E-03
17	19	2	2.988E-02	6.890E-03	6.228E+01	5.201E-03
19	19	2	1.410E-02	6.890E-03	6.228E+01	6.332E-03
20	19	2	0.000E+00	6.890E-03	6.228E+01	4.981E-03
28	19	2	1.045E-02	6.890E-03	6.228E+01	3.801E-03

Table 6.3: Estimated parameter values of the competition synthetic yeast system homologs and the replicates.

Primary IAA	Secondary IAA	num. of rep	k_1	k_2	k_3	k_4
3	3	4	1.38E-01	1.03E-01	5.07E+00	3.44E-03
3	12	4	1.38E-01	1.03E-01	7.51E+00	3.44E-03
3	14	4	1.38E-01	1.03E-01	5.28E+00	3.44E-03
3	28	4	1.38E-01	1.03E-01	6.32E+00	3.44E-03
12	12	2	1.38E-01	1.03E-01	6.39E+00	3.44E-03
12	14	4	1.38E-01	1.03E-01	6.33E+00	3.44E-03
12	28	5	1.38E-01	1.03E-01	5.64E+00	3.44E-03
14	14	2	1.38E-01	1.03E-01	4.72E+00	3.44E-03
14	28	4	1.38E-01	1.03E-01	3.74E+00	3.44E-03
28	28	2	1.38E-01	1.03E-01	4.23E+00	3.44E-03

Primary IAA	Secondary IAA	num. of rep	k_5	k_6	k_7	k_8
3	3	4	6.72E-03	6.89E-03	9.91E+01	4.31E-03
3	12	4	4.76E-03	6.89E-03	9.91E+01	3.45E-03
3	14	4	9.48E-03	6.89E-03	9.91E+01	3.60E-03
3	28	4	7.54E-03	6.89E-03	9.91E+01	3.39E-03
12	12	2	2.46E-03	6.89E-03	9.91E+01	3.28E-03
12	14	4	5.73E-03	6.89E-03	9.91E+01	3.17E-03
12	28	5	4.26E-03	6.89E-03	9.91E+01	2.80E-03
14	14	2	1.50E-02	6.89E-03	9.91E+01	3.10E-03
14	28	4	1.17E-02	6.89E-03	9.91E+01	3.45E-03
28	28	2	3.36E-03	6.89E-03	9.91E+01	3.98E-03

6.5 Summary

In this chapter, we discussed the analysis of a synthetic ARC system recapitulated in yeast for characterization that is often challenging *in planta*. Building upon the results from the previous chapter, the quantitative analysis of ARC^{sc} and its model follows the same approach – directed by the need to have a succinct, quantitative representation that captures the source of variation observed in the combinatorial library of auxin signal pathway components. The analysis resulted in a small set of parameters that uniquely identified the biological functions of IAAs and ARFs.

Similar to the results of the previous chapter, we observed sets of Ms that exhibit approximately equal costs. The overlapping performances of multiple M with increasing $nnz(M)$ possibly provides us with an alternative stopping criteria for our greedy search algorithm. It would be interesting to investigate the efficacy and efficiency of stopping the greedy search if all (or majority of) successor matrices have costs with standard deviation smaller than some threshold (a metric for determining how close these values are). Though exploring the different options for stoping criteria is an interesting exercise that may lead to discovering generally applicable solutions for efficient search of the PCD space, it is prudent to keep in mind that most such stopping criteria are heuristic strategies to determine at which point we refuse to increase the degrees of freedom in the parameter estimation for the sake of easier model interpretability and to avoid having too many 'unique' identifiers for each system variant. The resulting M must be scrutinized for its ability to aptly represent the underlying mechanism of system in question through further analysis and experimental design.

Chapter 7

CONCLUSION & DISCUSSION

Until recently, approaches to understanding complex biological systems have been limited to top-down methodologies, where existing organisms are studied by way of decomposition. Synthetic biology, on the other hand, complements these approaches through methodologies of bottom-up construction. By designing relatively simple and analytically tractable gene regulatory networks and studying their characteristics, the synthesis-focused principles of synthetic biology embodies the words of Richard Feynman - "What I cannot create, I do not understand". Thus, understanding the design principles of nature is an essential primitive for engineering and modifying complex behaviors in living systems to realize practical applications in renewable energy resources, clinical and pharmaceutical research, and bioremediation, to name a few. However, the progress of the field is challenged by the lack of systematic understanding of biology, including clear identification of constituent components and their corresponding biological functions. In this thesis, we aimed to address this problem and develop an analytical framework to quantitatively identify the relevant biological functions of system components to elevate their utility as engineerable parts.

In Chapter 3, we presented the PCD matrix that provides the framework for systematically identifying the dependent relationships of model parameters to system components, and showed that the interpretation of these relationships are used to identify and quantify the functions of biological components. Given a combinatorial library of system variants composed of one or more components and a general mathematical model of the system, we introduced how to formulate a

complete set of PCD candidates. There are multiple hypotheses regarding dependency relationships of model parameters to system components, and using the PCD framework to generate a set of parameter constraints for each hypothesis, we presented a metric for measuring the ability of each hypothesis to describe the systematic variations observed in the experimental dataset. Finally, we presented the representation of the complete set of PCD candidates as a partially ordered set. This representation allowed us to highlight the order-preserving characteristic of the cost function J over PCD candidates leading to heuristic approaches that improve the systematic search for the optimal PCD.

In Chapter 4, we used a hypothetical synthetic biology system called the Leader Election system to demonstrate the application of the PCD framework, and presented two systematic approaches to identify the optimal PCD – exhaustive search and a greedy search method. We demonstrated that for a sufficiently small system (made up of few components) and its model (with few parameters), the computational cost of screening performances of all candidate M is cheap and that a single or a set of optimal PCDs is identified by using an arbitrary threshold on the number of key system functionalities to be identified. However, as a PCD search space tends to increase exponentially with the numbers of constituent components and model parameters, we proposed an alternative approach using a greedy algorithm. The efficiency of searching for the optimal PCD can be improved using the greedy algorithm by introducing a threshold on the number of desired key functionalities to be identified. Because the greedy approach is a series of local optimizations, it is possible for the algorithm to get stuck in local optima. The existence of such a scenario was supported by a diagrammatic representation of the PCD search space as a complete lattice. It showed that though the feature of the cost function as an order-preserving map, it does not guarantee a global optimum

in a greedy search. Investigating whether the behavior of the greedy algorithm getting stuck in local optima can be predicted by some specific and identifiable features of the system and/or model of interest may help practitioners apply an alternative method for identifying the optimal PCD. For such a scenario, we proposed two alternative methods to address this problem, a forward-and-backward propagation approach and a greedy search coupled with random searches. A compare-and-contrast discussion of analytical results of the PCD and the SGL methods was presented using the LE system, as well as the limitations of applying the SGL to help identify the biological functions of genetic components as discussed in the scope of the thesis. Finally, we discussed that systematic searches for the optimal PCD is useful in analyzing combinatorial libraries of systems where little is known about the internal mechanism. Assuming some high level dynamics of the system is captured by its model, the interpretation of model parameters and dependency relationships revealed through the PCD framework may give practitioners better insights about the system, such as which components contribute more strongly to variations of system behavior.

In Chapter 5, we introduced a synthetic signal pathway built inside of yeast using the plant hormone, auxin. Taking advantage of the fact that each component of the auxin pathway naturally belongs to a family of interchangeable proteins, a combinatorial library consisting of 48 unique system variants was synthesized. In this chapter, the application of the PCD framework demonstrated its ability to systematically discriminate and verify hypotheses about the inner mechanism of systems, which is subtly different from its application in Ch. 4. The analytical results revealed a small number of features of the IAA family that define their biological functions: the expression level and the auxin-mediated degradation rate. This demonstrated the utility of the PCD framework specifically for systems biology – by identifying quantities that allow practitioners to reason and

hypothesize about the naturally occurring system, thereby studying the 'what is' aspect of biology. Furthermore, because the PCD framework is a platform for formally representing each hypothesis regarding the relationships between a combinatorial library system and its corresponding model, it brought formality to hypotheses invalidation process as well as enabling clear dialogues among practitioners of differing backgrounds.

Additionally, a series of simulation studies of IAA raised the question of whether independently tuning the key characteristics of IAA is possible. Inspired by this question, a synthetic variation of an IAA with only its expression level increased, while leaving the degradation rates consistent (both basal and auxin-mediated) was built and tested. The experimental data verified that independent tuning was indeed possible. This raised further questions regarding which region of the coding sequence of IAA is responsible for the different characteristics (e.g. expression level, degradation rates, AFB-binding rates) and whether they can be modified to create synthetic variations that exactly match the specifications of new circuits. Moving forward, it would be interesting to build and analyze a combinatorial library of IAAs that are cross-product combinations of the four internal IAA domains. This would require identification of lower-level models of the IAA domains with modified parameters that differentiate its key functions. Applying the PCD framework to this experimental setup will probe the relationship between the compositional details of IAA and its functions, answering the plant-biology inspired questions such as the evolutionary paths taken by the IAA proteins, and provide synthetic biologists a set of fine-grained tuning strategies.

In Chapter 6, we applied the PCD framework to an extension of the synthetic auxin pathway from Chapter 5. The iterative construction approach proved to be an interesting exercise in how the analytical PCD result from one system translates to its extension system. A similar objective

of verifying hypotheses about the relationship between model parameters and system components was addressed. The analysis quantitatively identified a subset of parameters that are representative of biological functions of ARF and IAA families. The results suggested that some of the previously identified characteristics of IAAs suffer from retroactivity, or context-dependent behavior, common in synthetic biology. Though the low dimensionality of the experimental data limited the scope of analysis, it raised questions regarding whether new experiments can be designed from the results of PCD analysis to definitively distinguish some of the competing hypotheses.

We believe that the PCD framework has practical applications in both systems and synthetic biology. In systems biology, where practitioners probe existing systems with the focus on understanding the underlying mechanisms, distributing the functions of biological components into multiple parameters help appropriate the sources of variations in system behaviors. On the other hand, the dependency relationships reduce the cost of engineering and tuning systems by identifying the optimal strategies for which system variations to construct and test.

Appendix A

MATHEMATICA PACKAGE DETAILS

A.1 Summary

A custom Mathematica™ package for the PCD analysis is available on https://github.com/shellyjang. The package requires three objects - model f, data $\mathbf{D}$ and cost function J - formatted in the following way.

Model. The model function, `f`, accepts two arguments - an expanded parameter set (`p`) and an index of the system variant (`tup`). `p` is a multi-dimensional array with an indexing ordered by $\kappa, j_1, \cdots, j_n$, where κ denotes the index of the elementary parameter vector, js denote the variant index of each component, and `tup` is a single list of form (`[`$j_1, \cdots, j_n$`]`). For example, to fetch the value $k_2^{(2,3,1)}$, we query `p[[2, {2,3,1}]]`. To simulate the behavior of the system variant $C^{(2,3,1)}$ given `p`, we write `f[p, {2, 3, 1}]`. Within the function definition, each κ-th parameter should be specified as `Fold[Part,p,Flatten[`κ`, tup]]`. The model can handle any arbitrary function as long as the type of the output from the model matches the type of the data.

Data. The data array, `data`, is a multi-dimensional array with a flattened indexing order of 1 through $\prod_i \eta_i$. The order of entry follows the standard combinations rule and follows the order generated via `Tuples[Table[Range[`η`[[ii]]],ii,n]]`. The number of measurements per entry (corresponding to a single system variant) must be equal to the number of output given by the model `f`.

Cost Function. Users can define any arbitrary metric for the model error (e.g. 2-norm, ∞-norm).

It takes the experimental data and the model predicted data, each specified by the same `tup`.

A.2 Functions

- `parameter`: The function outputs the elementary parameter vector of size m.

- `totalParameter`: The function outputs the expanded parameter vector of size m for a system with n components, each component with η_i variants. The vector of η_is is the last input μ. The output of the function is a multi-dimensional array with indices in the order of $\kappa, j_1, \cdots, j_n$.

- `constraints`: The function outputs the parameter constraints ($H(M)$) over the expanded parameter vector generated from the input μ.

- `initGuess`: The function outputs the initial guess within a specified range formatted for the `NMinimize` function used in `fitInit`. The range by default is set to 50% and 150% of the nominal value. However, for inherited estimation, we suggest that these ranges are specified with much smaller margins to ensure that the `NMinimize` does not choose an initial point far enough from the inherited estimate parameters.

- `res`: The function outputs the user-defined cost between the data and the model prediction.

- `fitInit`: The function outputs the $J(\Theta^*, M)$ and Θ^* with input arguments M, the sub-selection of the dataset (user may choose to use all or some of the dataset to speed up the estimation) and the initial guess for the parameters. Optionally, the user can specify the maxIteration value, along with other optional optimization arguments such as accuracy and precision goals for the `NMinimize`.

– `TolSearch`: The function executes the greedy search algorithm and terminates when the relative decrease in the cost from the predecessor to the successor is less than some user-specified threshold.

A.3 Sample Code

The following code shows how the greedy search algorithm is implemented using the LE system dataset from Chapter 4. Lines 1-10 specify the model; line 12 specifies the dimension of the PCD; line 13 species the components indices for selecting subset of the dataset; lines 14-15 specify the expanded parameter set Θ; lines 16 and 17 specify the tolerance of the greedy algorithm and the maximum number of iterations allowed for the `NMinimize`; and line 19 runs the TolSearch algorithm and saves the intermediate results as a Mathematica readable file (`/private/tmp/11.mx`). The experimental data is assumed to be preloaded.

```
1   f[p_, tup_] := (model[p] =
2     First[y /.
3       NDSolve[{w'[t] == -Fold[Part, p, Flatten[{1, tup}]] w[t] -
4           Fold[Part, p, Flatten[{2, tup}]] w[t] x[t],
5         x'[t] == Fold[Part, p, Flatten[{1, tup}]] w[t],
6         y'[t] == Fold[Part, p, Flatten[{2, tup}]] w[t] x[t],
7         w[0] == 1,
8         x[0] == 0,
9         y[0] == 0}, {w, x, y}, {t, 10}]]) /;
10  VectorQ[Flatten[p], NumberQ];
11
12  {m, n} = {2, 3};
```

```
13  sel = {{1, 4}, {1, 4}, {1, 4}};
14  pInit = {0.342, 1.459};
15  pInit = Table[#, {m}, {n}]&/@pInit;
16  tol = 0.1;
17  mI = 250;
18
19  TolSearch[parameter[m], m, n, sel, pInit , "/private/tmp/11.mx", tol , mI]
```

Appendix B

R CODE

B.1 SGL analysis for LE

The following code shows the SGL implementation on the LE experimental data and model (Section 4.4). Lines 4-6 specify the matrix indicating group indices of each system variant that a parameter estimate corresponds to. Lines 8-9 specify the parameters obtained under unconstrained optimization from the experimental data (i.e. using $M_{2^{mm}-1}$). Line 12 loads the SGL library [104] and line 13 and 15 structure the matrix specified in lines 4-6 and a parameter estimate vectors, k_1 and k_2 respectively, into a format appropriate for SGL library. Lines 14 and 16 fit a regularized generalized linear model via penalized maximum likelihood.

./code/LE_SGL_v1.R

```
1 ## LE system
2 ## Parameter level group lasso
3 ## Using the free-for-all fitted parameters, determine whether there are
      significant groupings
4 X <- cbind( rep(c(1,0), each=4), rep(c(0,1), each=4),
5             rep(rep(c(1,0),each=2),2), rep(rep(c(0,1),each=2),2),
6             rep(c(1,0), 4), rep(c(0,1), 4) )
7 ## free-for-all estimates
8 k1 <- c(0.1500, 0.1497, 0.1440, 0.1523, 0.755067, 0.7671, 0.756938, 0.744231)
9 k2 <- c(1.27848, 3.453, 0.901603, 2.340, 1.27688, 3.454, 0.9030, 2.337)
10
```

```
11 ## SPARSE GROUP LASSO
12 library (SGL)
13 datak1 <- list(x=X, y=k1)
14 fitSGLk1 <- SGL(data=datak1, index=rep(1:3, each=2), type="linear")
15 datak2 <- list(x=X, y=k2)
16 fitSGLk2 <- SGL(data=datak2, index=rep(1:3, each=2), type="linear")
17 fitSGLk1$beta[,seq(from=1,to=20,by=3)]
18 fitSGLk2$beta[,seq(from=1,to=20,by=3)]
```

B.2 ANOVA analysis for Auxin Signal Pathway I

The following code shows the ANOVA implementation on the Auxin signal pathway experimental data and model (Section 5.4). Lines 2-6 import and structure the parameters obtained under unconstrained optimization from the experimental data (i.e. using $M_{2^{mm}-1}$). Lines 9-10 and 12 specify the model parameter to be analyzed, and line 11 fits an ANOVA model.

./code/Auxin_ANOVA.R

```
1 ## parametric ANOVA for AFB-IAA free-for-all fitting
2 library (xlsx)
3 thetaHat = read.xlsx('~/.dropbox-alt/Dropbox/auxsynbio/Quantitative Analysis/FinalValues/Compiled List of
      Parameter Estimates.xlsx', 3)
4 thetaHat = thetaHat[1:96,]
5 thetaHat$IAA = rep(c(letters [1:24]) , 4)
6 thetaHat$Replicate = rep(c(rep(toupper(letters [1]) , 24),rep(toupper(letters [2]) , 24)),2)
7
8 ## ANOVA (kk = 1, 2, 3, 4, 5)
9 kk = 1
```

```
colnames(thetaHat)[3+kk] = 'k'
aov.k = aov(k ~ Fbox * IAA, data=thetaHat)
colnames(thetaHat)[3+kk] = paste('k',toString(kk),sep='')
```

Appendix C

BICISTRONIC DESIGN

C.1 Introduction

The bicistronic design (BCD) [10] is an engineered translational coupling architecture that addresses the lack of reliability in conventional components used in synthetic biology. By encoding short leader peptides followed by a downstream Gene of Interest (GOI), the engineered sequence demonstrates a *context independent* behavior. Furthermore, the mathematical analysis defines the 'quality' of a genetic element, quantifying 'the extent to which the activity of a part varies across changes in context'. This is done by interpreting the results of standard ANOVA using R specific to the objective of the research. The authors discuss the design and analysis of BCD along with monocistronic designs (MCD), which serves as a control architecture. Pairwise combinations of different BCDs (or MCDs) and GOIs are constructed and the resulting steady-state GFP fluorescence levels are measured. This experiment design readily lends itself as a suitable dataset to test the PCM analysis method, because the general system architecture its conserved while swapping out mechanistically homogeneous system components. There exist two interchangeable components in the system, the BCD (or MCD) and the GOI.

To draw a comparison between the PCD analysis method and the ANOVA framework discussed in [10], we conducted two comparative analyses. First, to demonstrate the PCD analysis using the BCD/MCD system, we formulate the system, the data set and the model following the conventions introduced in Chapter 3. Second, to demonstrate the ANOVA framework and compare the

interpretation with the PCD analysis in another system, we apply it to the Auxin Response Circuit (ARC^{Sc}) introduced in Chapter 6.

C.2 PCD formulation

Components. There are $n = 2$ interchangeable components that make the BCD/MCD system - the translation elements (i.e., MCD and BCD) and the GOI. These are denoted by C_1 and C_2, respectively, with $\eta_1 = 22$ and $\eta_2 = 8$.

System. The system is denoted by a pair, $(C_1^{(i)}, C_2^{(j)}$, where $i = 1, \cdots, \eta_1$ and $j = 1, \cdots, \eta_2$, and there are 176 unique combinations of C_1 and C_2.

Data. For each $S^{(j_1, j_2)}$, there is a corresponding measurement of the output signal - GFP fluorescence, $D^{(j_1, j_2)}$.

Model. In the article, the relationship between the GFP fluorescence and the varying components (the BCD/MCD elements and the GOI) is defined as follows.

$$\begin{aligned} \log(\text{Fluorescence}_{j_1 j_2}) &= f(\theta) \\ &= \alpha + U_{j_1} + GOI_{j_2} + U : GOI_{j_1 j_2} \end{aligned}$$

This approach is commonly employed in statistical analyses and is a simple linear black-box model. Though easier to manipulate analytically, the ability to infer the underlying mechanism of the system from the model is limited because it assumes linear relationships among the individual components.

Parameters. The model has $m = 4$ parameters that each has an explicit relationship with BCD/MCD and GOI:

α	independent of the choices of the BCD/MCD and the GOI
U_{j_1}	dependent on the choice of the BCD/MCD
GOI_{j_2}	dependent on the choice of the GOI
$U : GOI_{j_1 j_2}$	dependent on the choices of BCD/MCD and GOI

The elementary model parameter is denoted by $\theta = [k_1, k_2, k_3, k_4]$[1], where we opted to replace the authors' nomenclature to be consistent with that of the previous chapters. The expanded parameter set for all of the system variants is denoted by

$$\Theta = \{\theta^{(1,1)}, \theta^{(1,2)}, \cdots, \theta^{(\eta_1,\eta_2)}\}$$

M. The $\mathbf{S}$ and Θ result in a set of M of dimension 4×2, with $2^8 = 512$ possible candidates. From the implicit relationship discussed in the previous section, the following M is considered.

$$M_{39} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}$$

However, before we proceed with the analysis of M_{39}, we build our case by inspecting several other M, such as M_0 and M_{255}.

$M_{\mathbf{0}}$. All entries in M_0 are 0, and the matrix is one of the two extreme cases. The matrix represents the hypothesis that the parameters of the proposed model are insensitive to the changing components of the system. Thus, the model is hypothetically unable to capture the variations of the system and

[1] $\alpha = k_1, U_i = k_2, GOI_j = k_3, U : GOI_{ij} = k_4$

that the best the model can do is to represent the overall average behavior. The set of optimization constraints, $H(M_0)$, is

$$\bigcup_{\kappa=1,2,3,4} \{k_\kappa^{(1,1)} = k_\kappa^{(1,2)} = \cdots = k_\kappa^{(\eta_1,\eta_2)}\}$$

Following the constraints, we can assume that there is only one representative parameter vector, $\theta^{(1,1)}$, and the rest of the parameter vectors in the set Θ are equal to $\theta^{(1,1)}$. Thus, the optimization problem is formulated as

$$\begin{aligned}
\Theta^* &= \arg\min_{\Theta} \quad J(\Theta) \\
&= \arg\min_{\Theta} \left(\sum_{j_1}^{\eta_1} \sum_{j_2}^{\eta_2} \left(f(\theta^{(1,1)}) - D^{(j_1,j_2)} \right)^2 \right)^{1/2}
\end{aligned}$$

Solving the least-squares equation results in the optimal $\theta^{(1,1)}$ that satisfies the following relationship,

$$\begin{aligned}
f(\theta^{(1,1)}) &= k_1^{(1,1)} + k_2^{(1,1)} + k_3^{(1,1)} + k_4^{(1,1)} \\
&= \mu_D \\
&= \frac{1}{\eta_1 \eta_2} \sum_{i}^{\eta_1} \sum_{j}^{\eta_2} D^{(i,j)},
\end{aligned}$$

where μ_D is the average experimental data over $\mathbf{D}$. From the linear model definition, it is shown that there exists an infinite number of solutions for $\theta^{(1,1)}$ and the cost of the optimal parameter vector is,

$$J(\Theta, M_{\mathbf{0}}) = \sqrt{\eta_1 \eta_2 \sigma_{\mathbf{D}}^2},$$

where $\sigma^2_{\mathbf{D}}$ is the variance of $\mathbf{D}$.

M_{255}. All entries in M_{255} are 1, and is the other extreme member in the set of candidate M. This matrix represents the hypothesis that all parameter vectors are unique to their corresponding system variants. Said differently, it hypothesizes that all four members of the parameter vector capture some quantitative features of each system variant. The set of optimization constraints is an empty set.

$$H(M_{255}) \quad = \quad \phi.$$

Following the convention, the optimal parameter set is obtained by solving the following,

$$\begin{aligned} \Theta^* &= \arg\min_{\Theta} \quad J(\Theta) \\ &= \arg\min_{\Theta} \sum_{j_1}^{\eta_1} \sum_{j_2}^{\eta_2} \left(f(\theta^{(j_1,j_2)}) - D^{(j_1,j_2)} \right)^2 \end{aligned}$$

Similar to the optimization problem using M_0, which has the strictest constraints on the variables and were underdetermined with an infinite number of solutions, M_{255} yields a result where there is an infinite number of solutions. But, since the four parameters in individual $\theta^{(i,j)}$ can be chosen arbitrarily to satisfy

$$\begin{aligned} f(\theta^{(j_1,j_2)}) &= k_1^{(j_1,j_2)} + k_2^{(j_1,j_2)} + k_3^{(j_1,j_2)} + k_4^{(j_1,j_2)} \\ &= D^{(j_1,j_2)}, \end{aligned}$$

the cost is 0.

Decomposition of M_{39}. We would like to decompose M_{39} from earlier and discuss how each

M can be interpreted and be used for invalidating the model. The decomposition results in the following three M.

$$M_{32} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \; M_4 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \; M_3 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 1 \end{pmatrix}$$

- M_{32}: The matrix represents the hypothesis that k_2 is dependent on the choice of BCD/MCD, and that the rest of the parameters are independent from the changing components. This results the following $H(M_{32})$,

$$\left(\bigcup_{j_1}^{\eta_1} \{ k_2^{(j_1,1)} = k_2^{(j_1,2)} = \cdots = k_2^{(j_1,\eta_2)} \} \right) \bigcup \left(\bigcup_{\kappa=1,3,4} \{ k_\kappa^{(1,1)} = k_\kappa^{(1,2)} = \cdots = k_\kappa^{(\eta_1,\eta_2)} \} \right)$$

Following the constraints, we define a new set of η_1 parameter variables as follows to reduce the number of k_2 parameters to be estimated from $\eta_1\eta_2$ to η_1,

$$k_2^{(j_1,i_2)} \quad = \quad k_2^{(j_1,-)}, \;\; \forall i_2 = 1, \cdots, \eta_2.$$

Furthermore, there each of the parameters, $k_1^{(j_1,j_2)}$,$k_3^{(j_1,j_2)}$ and $k_4^{(j_1,j_2)}$, is reduced to a single value

k_1, k_3 and k_4, respectively. The cost function for M_{32} then becomes,

$$\begin{aligned} J(\Theta, M_{32}) &= \sum_{j_1}^{\eta_1}\sum_{j_2}^{\eta_2}\left(f(\theta^{(j_1,j_2)}) - D^{(j_1,j_2)}\right)^2 \\ &= \sum_{j_1}^{\eta_1}\sum_{j_2}^{\eta_2}\left(\sum_{\kappa}^{4} k_{\kappa}^{(j_1,j_2)} - D^{(j_1,j_2)}\right)^2 \\ &= \sum_{j_1}^{\eta_1}\sum_{j_2}^{\eta_2}\left(k_1^{(j_1,j_2)} + k_2^{(j_1,j_2)} + k_3^{(j_1,j_2)} + k_4^{(j_1,j_2)} - D^{(j_1,j_2)}\right)^2 \\ &= \sum_{j_1}^{\eta_1}\sum_{j_2}^{\eta_2}\left(k_1 + k_2^{(j_1,-)} + k_3 + k_4 - D^{(j_1,j_2)}\right)^2 \end{aligned}$$

To simplify the problem further, we assume that $k_1 + k_3 + k_4 = \varepsilon$, where $\varepsilon > 0$ is a small positive number. This results in the following solution for the $k_2^{(j_1,-)}$.

$$\begin{aligned} k_2^{(j_1,-)} &= \frac{1}{\eta_2}\sum_{j_2}^{\eta_2} D^{(j_1,j_2)} - \varepsilon \\ &= \mu_{D^{(j_1,-)}} - \varepsilon, \end{aligned}$$

where $\mu_{D^{(j_1,-)}}$ is the average value of the experimental data of the set $\{D^{(j_1,j_2)} \mid \forall\ j_2 = 1, \cdots, \eta_2\}$. The cost of the optimal parameter vector is approximately,

$$J(\Theta, M_{32}) \approx \sum_{j_1}^{\eta_1} \eta_2 \sigma^2_{D^{(j_1,-)}},$$

where $\sigma^2_{D^{(j_1,-)}}$ is the variance of experimental data set $\{D^{(j_1,j_2)} \mid \forall\ j_2 = 1, \cdots, \eta_2\}$.

- M_4: The matrix represents the hypothesis that k_3 is dependent on the choice of GOI, and that the rest of the parameters are independent from the changing components. This results the following

$H(M_4)$,

$$\left(\bigcup_{j_2}^{\eta_2}\{k_3^{(1,j_2)} = k_3^{(2,j_2)} = \cdots = k_2^{(\eta_1,2)}\}\right)\bigcup\left(\bigcup_{\kappa=1,2,4}\{k_\kappa^{(1,1)} = k_\kappa^{(1,2)} = \cdots = k_\kappa^{(\eta_1,\eta_2)}\}\right)$$

Following the constraints, we define a new set of η_2 parameter variables as follows to reduce the number of k_3 parameters to be estimated from $\eta_1\eta_2$ to η_2,

$$k_3^{(i_1,j_2)} \quad = \quad k_3^{(-,j_2)},\ \forall i_1 = 1,\cdots,\eta_1.$$

Furthermore, there each of the parameters, $k_1^{(j_1,j_2)}, k_2^{(j_1,j_2)}$ and $k_4^{(j_1,j_2)}$, is reduced to a single value k_1, k_2 and k_4, respectively. Using the same procedure as for M_{32}, the optimization problem results in the following solution for the $k_2^{(j_1,-)}$.

$$\begin{aligned} k_3^{(-,j_2)} &= \frac{1}{\eta_2}\sum_{j_1}^{\eta_1} D^{(j_1,j_2)} - \varepsilon \\ &= \mu_{D^{(-,j_2)}} - \varepsilon, \end{aligned}$$

where $\mu_{D^{(-,j_2)}}$ is the average value of the experimental data of the set $\{D^{(j_1,j_2)} \mid \forall\ j_1 = 1,\cdots,\eta_1\}$. The cost of the optimal parameter vector is approximately,

$$J(\Theta, M_4) \quad \approx \quad \sum_{j_2}^{\eta_2} \eta_1 \sigma^2_{D^{(-,j_2)}},$$

where $\sigma^2_{D^{(-,j_2)}}$ is the variance of experimental data set $\{D^{(j_1,j_2)} \mid \forall\ j_1 = 1,\cdots,\eta_1\}$.

Comparing the performances of M_{32} and M_4, we make the following observation. Because $\sigma^2_{D^{(j_1,-)}}$ and $\sigma^2_{D^{(-,j_2)}}$ correspond to the sum of variances in the experimental data set with respect

to BCD/MCD and GOI, respectively, if the variance in the experimental data across BCD/MCD is larger than the variance in the experimental data across GOI, M_{32} will perform worse than M_4, and the reverse is also true. Applying this observation to the result of [10], we can make the following claim that M_{32} of the BCD data set will perform better than M_{32} of the MCD data set, because BCD is less sensitive to the chaining GOI than MCD. This is indeed true - when the $J(\Theta, M_{32})$ is normalized by $J(\Theta, M_0)$, BCD shows increased reduction in the performance than MCD.

- M_3: The matrix represents the hypothesis that the k_4 is dependent on the choices of both BCD/MCD and GOI, and the rest of the parameters are independent from the composition of the system. The resulting $H(M_3)$ is,

$$\bigcup_{\kappa=1,2,3} \{k_\kappa^{(1,1)} = k_\kappa^{(1,2)} = \cdots = k_\kappa^{(\eta_1,\eta_2)}\}$$

As before, we reduce the parameters $k_1^{(j_1,j_2)}, k_2^{(j_1,j_2)}$ and $k_3^{(j_1,j_2)}$ to a single value k_1, k_2 and k_3, respectively. The cost function for M_3 then becomes,

$$J(\Theta, M_3) = \left(\sum_{j_1}^{\eta_1} \sum_{j_2}^{\eta_2} \left(k_1 + k_2 + k_3 + k_4^{(j_1,j_2)} - D^{(j_1,j_2)} \right)^2 \right)^{1/2}$$

As this is an underdetermined problem, we arbitrarily choose k_1, k_2, k_3 to satisfy $k_1 + k_2 + k_3 = \varepsilon$, and the optimal $k_4^{(j_1,j_2)}$ are chosen to equal $D^{(j_1,j_2)} - \varepsilon$, respectively. And the performance is,

$$\begin{aligned} J(\Theta, M_3) &= \eta_1 \eta_2 \varepsilon^2 \\ &\approx 0 \end{aligned}$$

Because we can choose ε to be arbitrarily small, $J(\Theta, M_3)$ can also be arbitrarily small.

C.3 Equivalence Classes

Because BCD/MCD has a linear model, the above analysis for M_{32}, M_4 and M_3 applies to a subset of M. For instance, the analysis of M_{32} is equivalent for M_{128}, M_8 and M_2, all of which has the common characteristic that the $nnz = 1$ and the single 1-entry is found in the first column. Thus, these M belong to an equivalency class, where it is defined as,

$$\begin{aligned} E_1 &= \{M \mid J(\Theta, M) = \sum_{j_1}^{\eta_1} \eta_2 \sigma^2_{D^{(j_1,-)}}\} \\ &= \{M_{128}, M_{32}, M_8, M_2\} \end{aligned}$$

Interestingly, there are other members of E_1 and they satisfy the following:

- all entries in the second column are 0
- there is at least one non-zero entry in the first column

There exits other equivalency classes in the search space of M. For instance,

$$E_2 = \{M \mid J(\Theta, M) = \sum_{j_2}^{\eta_2} \eta_1 \sigma^2_{D^{(-,j_2)}}\}$$

E_2 is consisted of M that are opposite of members of E_1 with respect to which column of the matrix are all 0. Therefore, each matrix in E_2, all entries in the first column are 0 and there is at least one non-zero entry in the second column.

Identifying the mapping between indistinguishable parameters and set building rules for equivalence classes of M may serve as an interesting foundation for experimental design project

discussed in Chapter 7. Here, a brief and incomplete theorem regarding equivalence classes in M is introduced.

Theorem C.3.1 *Two PCDs M' and M'' are said to belong to an equivalence class E, if all elements in $(M' \cup M'') \setminus (M' \cap M'')$ can be paired up to $((\kappa', i'), (\kappa'' i''))$ where*

1. $i' = i''$,

2. κ' *and* κ'' *are a pair of indistinguishable parameters.*

It would be of interest to expand and verify this theorem by applying it to the case studies appearing in [89].

BIBLIOGRAPHY

[1] G. Church, *Regenesis*. Basic Books, 2012.

[2] N. Dharmasiri, S. Dharmasiri, D. Weijers, E. Lechner, M. Yamada, L. Hobbie, J. S. Ehrismann, G. Jürgens, and M. Estelle, "Plant development is regulated by a family of auxin receptor f box proteins," *Developmental Cell*, vol. 9, no. 1, pp. 109 – 119, 2005.

[3] Y. Yokobayashi, R. Weiss, and F. H. Arnold, "Directed evolution of a genetic circuit," *Proceedings of the National Academy of Sciences*, vol. 99, no. 26, pp. 16587–16591, 2002.

[4] B. C. Stanton, A. A. K. Nielsen, A. Tamsir, K. Clancy, T. Peterson, and C. A. Voigt, "Genomic mining of prokaryotic repressors for orthogonal logic gates," *Nat Chem Biol*, vol. 10, pp. 99–105, 02 2014.

[5] K. E. Thompson, C. J. Bashor, W. A. Lim, and A. E. Keating, "Synzip protein interaction toolbox: in vitro and in vivo specifications of heterospecific coiled-coil interaction domains," *ACS synthetic biology*, vol. 1, no. 4, pp. 118–129, 2012.

[6] J. A. Hurt, S. A. Thibodeau, A. S. Hirsh, C. O. Pabo, and J. K. Joung, "Highly specific zinc finger proteins obtained by directed domain shuffling and cell-based selection," *Proceedings of the National Academy of Sciences*, vol. 100, no. 21, pp. 12271–12276, 2003.

[7] B. Davey and H. Priestley, *Introduction to Lattices and Order*. Cambridge Mathematical Textbooks, 1990.

[8] K. Havens, J. Guseman, S. Jang, E. Pierre-Jerome, N. Bolten, E. Klavins, and J. Nemhauser, "A synthetic approach reveals extensive tunability of auxin signaling," *Plant Physiology*, 2012.

[9] E. Pierre-Jerome, S. S. Jang, K. Havens, J. L. Nemhauser, and E. Klavins, "Recapitulation of the forward nuclear auxin response pathway in yeast," *PNAS*, 2014.

[10] V. Mutalik, J. Guimaraes, G. Cambray, C. Lam, M. Christoffersen, Q. Mai, A. Tran, M. Paull, J. Keasling, A. Arkin, and D. Endy, "Precise and reliable gene expression via standard transcription and translation initiation elements," *Nat Meth*, vol. 10, 2013.

[11] M. Elowitz and S. Leibler, "A synthetic oscillatory network of transcriptional regulators," *Nature*, vol. 403, 2000.

[12] T. Gardner, C. Cantor, and J. Collins, "Construction of a genetic toggle swithch in Escherichia coli," *Nature*, vol. 403, 339-342 2000.

[13] J. Bonnet, P. Subsoontorn, and D. Endy, "Rewritable digital data storage in live cells via engineered control of recombination directionality," *Proceedings of the National Academy of Sciences*, vol. 109, no. 23, pp. 8884–8889, 2012.

[14] A. Padirac, T. Fujii, and Y. Rondelez, "Bottom-up construction of in vitro switchable memories," *Proceedings of the National Academy of Sciences*, vol. 109, no. 47, pp. E3212–E3220, 2012.

[15] J. W. Kotula, S. J. Kerns, L. A. Shaket, L. Siraj, J. J. Collins, J. C. Way, and P. A. Silver, "Programmable bacteria detect and record an environmental signal in the mammalian gut," *Proceedings of the National Academy of Sciences*, vol. 111, no. 13, pp. 4838–4843, 2014.

[16] R. H. Carlson, *Biology is technology: the promise, peril, and new business of engineering life*, vol. 2010. Harvard University Press Cambridge, 2010.

[17] T. S. Moon, C. Lou, A. Tamsir, B. C. Stanton, and C. A. Voigt, "Genetic programs constructed from layered logic gates in single cells," *Nature*, vol. 491, no. 7423, pp. 249–253, 2012.

[18] A. Prindle, P. Samayoa, I. Razinkov, T. Danino, L. S. Tsimring, and J. Hasty, "A sensing array of radically coupled genetic/biopixels/'," *Nature*, vol. 481, no. 7379, pp. 39–44, 2012.

[19] S. Basu, R. Mehreja, S. Thiberge, M. Chen, and R. Weiss, "Spatiotemporal control of gene expression with pulse-generating networks.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 17, pp. 6355–60, 2004.

[20] S. Basu, Y. Gerchman, J. Collins, F. Arnold, and R. Weiss, "A synthetic multicellular system for programmed pattern formation," *Nature*, vol. 434, April 2005.

[21] R. Kwok, "Five hard truths for synthetic biology," *Nature*, 2010.

[22] E. Purnick and S. Weiss, "The second wave of synthetic biology: from modules to systems," *Nature Reviews Molecular Cell Biology*, vol. 10, pp. 410–422, 2009.

[23] A. A. Cheng and T. K. Lu, "Synthetic biology: an emerging engineering discipline," *Annual review of biomedical engineering*, vol. 14, pp. 155–178, 2012.

[24] J. B. Lucks, L. Qi, W. R. Whitaker, and A. P. Arkin, "Toward scalable parts families for predictable design of biological circuits," *Current Opinion in Microbiology*, vol. 11, no. 6, pp. 567 – 573, 2008. Growth and Development: Eukaryotes/Prokaryotes.

[25] N. S. Nise, *CONTROL SYSTEMS ENGINEERING, (With CD)*. John Wiley & Sons, 2007.

[26] H. H. McAdams and A. Arkin, "Stochastic mechanisms in gene expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, pp. 814–819, 1997.

[27] M. Elowitz, A. Levine, E. Siggia, and P. Swain, "Stochastic gene expression in a single cell," *Science*, vol. 297, pp. 1183–1186, 2002.

[28] D. Del Vecchio and R. Murray, *Biomolecular Feedback Systems*. MIT Press, 2010.

[29] P. A. Iglesias and B. P. Ingalls, *Control theory and systems biology*. MIT Press, 2010.

[30] E. D. Sontag, "Some new directions in control theory inspired by systems biology," *Systems biology*, vol. 1, no. 1, pp. 9–18, 2004.

[31] A. C. Ventura, J. Peng, L. Van Wassenhove, D. Del Vecchio, S. D. Merajver, and A. J. Ninfa, "Signaling properties of a covalent modification cycle are altered by a downstream target," *Proceedings of the National Academy of Sciences*, 2010.

[32] D. Del Vecchio, A. J. Ninfa, and E. D. Sontag, "Modular cell biology: retroactivity and insulation," *Mol Syst Biol*, vol. 4, 02 2008.

[33] D. Del Vecchio and E. D. Sontag, "Engineering Principles in Bio-Molecular Systems: From Retroactivity to Modularity," *European Journal of Control*, 2009.

[34] A. Levin-Karp, U. Barenholz, T. Bareia, M. Dayagi, L. Zelcbuch, N. Antonovsky, E. Noor, and R. Milo, "Quantifying translational coupling in e. coli synthetic operons using rbs modulation and fluorescent reporters," *ACS Synthetic Biology*, vol. 2, no. 6, pp. 327–336, 2013.

[35] J. Beal, T. Lu, and R. Weiss, "Automatic compilation from high-level biologically-oriented programming language to genetic regulatory networks," *PLoS ONE*, vol. 6, p. e22490, 08 2011.

[36] J. Beal, R. Weiss, D. Densmore, A. Adler, E. Appleton, J. Babb, S. Bhatia, N. Davidsohn, T. Haddock, J. Loyall, R. Schantz, V. Vasilev, and F. Yaman, "An end-to-end workflow for engineering of biological networks from high-level specifications," *ACS Synthetic Biology*, vol. 1, no. 8, pp. 317–331, 2012.

[37] R. A. Houghten, C. Pinilla, S. E. Blondelle, J. R. Appel, C. T. Dooley, and J. H. Cuervo, "Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery," *Nature*, vol. 354, no. 6348, pp. 84–86, 1991.

[38] M. A. Gallop, R. W. Barrett, W. J. Dower, S. P. Fodor, and E. M. Gordon, "Applications of combinatorial technologies to drug discovery. 1. background and peptide combinatorial libraries," *Journal of medicinal chemistry*, vol. 37, no. 9, pp. 1233–1251, 1994.

[39] T. E. Gorochowski, E. van den Berg, R. Kerkman, J. A. Roubos, and R. A. L. Bovenberg, "Using synthetic biological parts and microbioreactors to explore the protein expression characteristics of escherichia coli," *ACS Synthetic Biology*, vol. 3, no. 3, pp. 129–139, 2014.

[40] R. Cox, M. Surette, and M. Elowitz, "Programming gene expression with combinatorial promoters," *Molecular Systems Biology*, 2007.

[41] A. J. Keung, C. J. Bashor, S. Kiriakov, J. J. Collins, and A. S. Khalil, "Using targeted chromatin regulators to engineer combinatorial and spatial transcriptional regulation," *Cell*, vol. 158, pp. 110–120, 2014/07/14 2014.

[42] A. Prindle and J. Hasty, "Making gene circuits sing," *Proceedings of the National Academy of Sciences*, 2012.

[43] H. Salis, E. Mirsky, and C. Voigt, "Automated design of synthetic ribosome binding sites to control protein expression," *Nat Biotech*, vol. 27, pp. 946–950, 10 2009.

[44] F. J. Isaacs, D. J. Dwyer, C. Ding, D. D. Pervouchine, C. R. Cantor, and J. J. Collins, "Engineered riboregulators enable post-transcriptional control of gene expression," *Nat Biotech*, vol. 22, pp. 841–847, 07 2004.

[45] R. Egbert and E. Klavins, "Fine-tuning gene networks using simple sequence repeats," *PNAS*, 2012.

[46] E. Haseltine and F. Arnold, "Synthetic gene circuits: Design with directed evolution," *Annual Review of Biophysics and Biomolecular Structure*, vol. 36, no. 1, pp. 1–19, 2007. PMID: 17243895.

[47] H. Wang, F. Isaacs, P. Carr, Z. Sun, G. Xu, C. Forest, and G. Church, "Programming cells by multiplex genome engineering and accelerated evolution," *Nature*, vol. 460, pp. 894–898, 08 2009.

[48] X. Feng, S. Hooshangi, D. Chen, G. Li, R. Weiss, and H. Rabitz, "Optimizing genetic circuits by global sensitivity analysis," *Biophysical Journal*, vol. 87, no. 4, pp. 2195 – 2202, 2004.

[49] S. Hooshangi, S. Thiberge, and R. Weiss, "Ultrasensitivity and noise propagation in a synthetic transcriptional cascade," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 10, pp. 3581–3586, 2005.

[50] M. Rodriguez-Fernandez, J. Banga, and F. Doyle, "Novel global sensitivity analysis methodology accounting for the crucial role of the distribution of input parameters: application to systems biology models," *International Journal of Robust and Nonlinear Control*, vol. 22, no. 10, pp. 1082–1102, 2012.

[51] P. Wang, J. LÃ, and M. Ogorzalek, "Global relative parameter sensitivities of the feed-forward loops in genetic networks," *Neurocomputing*, vol. 78, no. 1, pp. 155 – 165, 2012. ¡ce:title¿Selected papers from the 8th International Symposium on Neural Networks (ISNN 2011)¡/ce:title¿.

[52] J. Beck and K. Arnold, *Parameter Estimation in Engineering and Science*. John Wiley, 1977.

[53] H. W. Sorenson, *Parameter estimation: principles and problems*. Marcel Dekker New York, 1980.

[54] H. K. Khalil and J. Grizzle, *Nonlinear systems*, vol. 3. Prentice hall Upper Saddle River, 2002.

[55] B. Canton, A. Labno, and D. Endy, "Refinement and standardization of synthetic biological parts and devices," *Nat Biotech*, vol. 26, pp. 787–793, 07 2008.

[56] R. Weiss and S. Basu, "The device physics of cellular logic gates," in *NSC-1: The First Workshop of Non-Silicon Computing. Boston, Massachusetts*, 2002.

[57] G. Neuert, B. Munsky, R. Tan, L. Teytelman, M. Khammash, and A. van Oudenaarden, "Systematic Identification of Signal-Activated Stochastic Gene Regulation," *Science*, vol. 330, 2013.

[58] S. Jang, K. Oishi, R. Egbert, and E. Klavins, "Specification and simulation of multicelled behaviors in gro," *Journal of American Chemical Society*, 2012.

[59] N. Lynch, *Distributed Algorithms*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1996.

[60] G. Le Lann, "Distributed systems - towards a formal approach," *Information Processing*, 1977.

[61] E. Chang and R. Roberts, "An improved algorithm for decentralized extrema-finding in cirular configurations of processes," *Comunications of the ACM*, vol. 22, no. 5, pp. 281–283, 1979.

[62] R. Li and B. Bowerman, "Symmetry breaking in biology," *Cold Spring Harbor Perspectives in Biology*, vol. 2, no. 3, 2010.

[63] A. N. Tikhonov, ed., *Ill-posed Problems in Natural Sciences: Proceedings of the International Conference Held in Moscow*. VSP, 1992.

[64] A. N. Tikhonov and V. I. Arsenin, *Solutions of ill-posed problems*. WInston, 1977.

[65] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *J. R. Statist. Soc.B*, 2011.

[66] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc B.*, vol. 58, no. 1, pp. 267–288, 1996.

[67] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[68] A. Miller, *Subset selection in regression*. Chapman and Hall, 2 ed., 2002.

[69] Z. Zhao, S. Sharma, A. Anand, F. Morstatter, S. Aleyani, and H. Liu, "Advancing feature selection research," tech. rep., Arizona State University, 2014.

[70] A. Hoerl, "Application of ridge analysis to regression problems," *Chemical Engineering Progress*, pp. 54–59, 1958.

[71] G. Wahba, G. Golub, and M. Heath, "Generalized cross-validation as a method for choosing a good ridge parameter.," *Technometrics*, 1979.

[72] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal. Statist. Soc B.*, vol. 68, pp. 49–67, 2006.

[73] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a spares group lasso," *arXiv*, 2010.

[74] A. Puig, A. Wiesel, and A. Hero, "A multidimensional shrinkage thresholding operator," in *IEEE/SP 15th Workshop on Statistical Signal Processing*, 2009.

[75] E. J. Chapman and M. Estelle, "Mechanism of auxin-regulated gene expression in plants," *Annual Review of Genetics*, vol. 43, no. 1, pp. 265–285, 2009. PMID: 19686081.

[76] A. Lokerse and D. Weijers, "Auxin enters the matrix - assembly of response machineries for specific outputs," *Current Opinion in Plant Biology*, vol. 12, no. 5, pp. 520 – 526, 2009.

[77] D. Weijers, E. Benkova, K. E. Jager, A. Schlereth, T. Hamann, M. Kientz, J. C. Wilmoth, J. W. Reed, and G. Jurgens, "Developmental specificity of auxin response by pairs of arf and aux/iaa transcriptional regulators," *EMBO J*, vol. 24, pp. 1874–1885, 05 2005.

[78] H. Muto, M. K. Watahiki, D. Nakamoto, M. Kinjo, and K. T. Yamamoto, "Specificity and similarity of functions of the aux/iaa genes in auxin signaling of arabidopsis revealed by promoter-exchange experiments among msg2/iaa19, axr2/iaa7, and slr/iaa14," *Plant Physiology*, vol. 144, no. 1, pp. 187–196, May 2007.

[79] S. Kepinski and O. Leyser, "The Arabidopsis F-box protein TIR1 is an auxin receptor," *Nature*, vol. 435, 2005.

[80] L. Calderon-Villalobos, X. Tan, N. Zheng, and M. Estelle, "Auxin Perception - Structural Insights," *Cold Spring Harbor Perspect Biology*, 2010.

[81] J. A. Ramos, N. Zenser, O. Leyser, and J. Callis, "Rapid degradation of auxin/indoleacetic acid proteins requires conserved amino acids of domain ii and is proteasome dependent," *The Plant Cell Online*, vol. 13, no. 10, pp. 2349–2360, 2001.

[82] X. Tan and N. Zheng, "Mechanism of auxin perception by the TIR1 ubiquitin ligase," *Nature*, vol. 446, pp. 640–645, 2007.

[83] K. A. Dreher, J. Brown, R. E. Saw, and J. Callis, "The arabidopsis aux/iaa protein family has diversified in degradation and auxin responsiveness," *The Plant Cell Online*, vol. 18, no. 3, pp. 699–714, March 2006.

[84] J. L. Stewart and J. L. Nemhauser, "Do trees grow on money? auxin as the currency of the cellular economy," *Cold Spring Harbor Perspectives in Biology*, vol. 2, no. 2, 2010.

[85] V. K. Mutalik, J. C. Guimaraes, G. Cambray, Q.-A. Mai, M. J. Christoffersen, L. Martin, A. Yu, C. Lam, C. Rodriguez, G. Bennett, J. D. Keasling, D. Endy, and A. P. Arkin, "Quantitative estimation of activity and quality for collections of functional genetic elements," *Nat Meth*, vol. 10, pp. 347–353, 04 2013.

[86] M. Srivastava and E. Carter, *Introduction to Applied Multivariate Statistics*. Elsevier Science Publishers,, 1983.

[87] D. Georgiev and E. Klavins, "Model discrimination of polynomial systems via stochastic inputs," in *Decision and Control, 2008. CDC 2008. 47th IEEE Conference on*, pp. 3323–3329, IEEE, 2008.

[88] D. Georgiev, M. Fazel, and E. Klavins, "Model discrimination of chemical reaction networks by linearization," in *American Control Conference (ACC), 2010*, pp. 5916–5922, IEEE, 2010.

[89] E. Walter and L. Pronzato, "On the identifiability and distinguishability of nonlinear parametric models," *Mathematics and Computers in Simulation*, vol. 42, pp. 125–134, 1996.

[90] C. Finet, A. Berne-Dedieu, C. P. Scutt, and F. Marlétaz, "Evolution of the arf gene family in land plants: old domains, new tricks.," *Molecular Biology and Evolution*, 2012.

[91] I. Paponov, W. Teale, D. Lang, M. Paponov, R. Reski, S. Rensing, and K. Palme, "The evolution of nuclear auxin signalling," *BMC Evolutionary Biology*, 2009.

[92] H. Szemenyei, M. Hannon, and J. A. Long, "Topless mediates auxin-dependent transcriptional repression during arabidopsis embryogenesis," *Science*, vol. 319, no. 5868, pp. 1384–1386, 2008.

[93] T. Ulmasov, J. Murfett, G. Hagen, and T. J. Guilfoyle, "Aux/iaa proteins repress expression of reporter genes containing natural and highly active synthetic auxin response elements.," *The Plant Cell Online*, vol. 9, no. 11, pp. 1963–71, 1997.

[94] N. Dharmasiri, S. Dharmasiri, and M. Estelle, "The F-box protein TIR1 is an auxin receptor," *Nature*, vol. 435, 2005.

[95] T. Ulmasov, G. Hagen, and T. J. Guilfoyle, "Activation and repression of transcription by auxin-response factors," *Proceedings of the National Academy of Sciences*, vol. 96, no. 10, pp. 5844–5849, 1999.

[96] S. B. Tiwari, G. Hagen, and T. Guilfoyle, "The roles of auxin response factor domains in auxin-responsive transcription," *The Plant Cell Online*, vol. 15, no. 2, pp. 533–543, 2003.

[97] S. Abel, M. D. Nguyen, and A. Theologis, "Theps-iaa4/5-like family of early auxin-inducible mrnas inarabidopsis thaliana," *Journal of Molecular Biology*, vol. 251, no. 4, pp. 533 – 549, 1995.

[98] D. L. Remington, T. J. Vision, T. J. Guilfoyle, and J. W. Reed, "Contrasting modes of diversification in the aux/iaa and arf gene families," *Plant Physiology*, vol. 135, no. 3, pp. 1738–1752, 2004.

[99] E. Pierre-Jerome, B. L. Moss, and J. L. Nemhauser, "Tuning the auxin transcriptional response," *Journal of Experimental Botany*, vol. 64, no. 9, pp. 2557–2563, 2013.

[100] E. H. Rademacher, A. S. Lokerse, A. Schlereth, C. I. Llavata-Peris, M. Bayer, M. Kientz, A. Freire Rios, J. W. Borst, W. Lukowitz, G. Jürgens, and D. Weijers, "Different auxin response machineries control distinct cell fates in the early plant embryo," *Developmental Cell*, vol. 22, pp. 211–222, 2014/05/14 2012.

[101] T. Vernoux, G. Brunoud, E. Farcot, V. Morin, H. Van den Daele, J. Legrand, M. Oliva, P. Das, A. Larrieu, D. Wells, Y. Guedon, L. Armitage, F. Picard, S. Guyomarc'h, C. Cellier, G. Parry, R. Koumproglou, J. H. Doonan, M. Estelle, C. Godin, S. Kepinski, M. Bennett, L. De Veylder, and J. Traas, "The auxin signalling network translates dynamic input into robust patterning at the shoot apex," *Mol Syst Biol*, vol. 7, 07 2011.

[102] M. Del Bianco and S. Kepinski, "Context, specificity, and self-organization in auxin response," *Cold Spring Harbor Perspectives in Biology*, vol. 3, no. 1, 2011.

[103] K. Nishimura, T. Fukagawa, H. Takisawa, T. Kakimoto, and M. Kanemaki, "An auxin-based degron system for the rapid depletion of proteins in nonplant cells," *Nat Meth*, vol. 6, pp. 917–922, 12 2009.

[104] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, *Package 'SGL'*. Stanford, 2010.