ABSTRACT

WILLIAMS, CRANOS M. Engineering Methodologies and Design Concepts for Systems Biology. (Under the direction of Dr. Winser E. Alexander and Dr. William W. Edmonson).

The inability to adequately model and simulate many complex dynamic systems results from a lack of theoretical and computational tools that are capable of handling the increased complexity of these systems. This complexity contributes to our inability to identify appropriate mathematical representations that would advance the conceptual and theoretical understanding of these systems. This is particularly evident when attempting to understand the emergence of complex biological phenomenon from interactions that span many levels of biological organization. The complex properties of these systems limit the depth of understanding that can be acquired by more traditional trial and error, reductionists, and experimental analyses. Systems biology focuses on the development of specialized theoretical, analytical, and computational tools that can facilitate a paradigm shift in how we understand, model, manipulate and control biological processes.

The research outlined in this dissertation focuses on the development, refinement, and implementation of engineering methodologies and design concepts for systems biology applications. The work outlined here focuses on two approaches. The first was a procedure formulated around blind source separation techniques. This approach exploited assumptions about underlying biological characteristics in order to blindly extract hidden functional relationships embedded in large-scale gene expression data. The second approach presented an estimation algorithm that used an optimal control formulation to address biological constraints on the component and pathway dynamics. This type of formulation allowed for the direct integration of dynamic equality and inequality constraints into the solution, alleviating the need for constrained optimization routines.

The approaches presented here take advantage of fundamental analytical concepts in electrical engineering. Application specific modifications facilitate the integration of these analytical concepts with available experimental data and first-principles knowledge. This adds to the library of available tools that are capable of estimating and understanding the unknown or immeasurable characteristics of biological processes. Ultimately, these tools will help researchers attain the medicinal, nutritional, and social benefits that can be achieved through the systematic manipulation of these processes at various levels of biological organization.

Engineering Methodologies and Design Concepts for Systems Biology

by
Cranos M. Williams

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

Electrical Engineering

Raleigh, North Carolina

2007

APPROVED BY:

Dr. Wesley E. Snyder	Dr. Anne Stomp
Dr. Mladen A. Vouk	
Dr. Winser E. Alexander Chair of Advisory Committee	Dr. William W. Edmonson Co-chair of Advisory Committee

DEDICATION

I dedicate my dissertation to my family.

To my parents,
Irvin E. Williams, Sr. and Jasmin F. D. Williams,

my big brother and sister-in-law,
Irvin E. Williams, Jr. and Kimberly D. Williams,

and my big sister,
Nakia P. Williams

Without your love and support, none of this would be possible.

BIOGRAPHY

Cranos Monroe Williams was born on the beautiful island of St. Thomas in the US Virgin Islands to Irvin E. Williams, Sr. and Jasmin F. D. Williams. He grew up in the Virgin Islands, attending Joseph Gomez Elementary School, Bertha C. Boschulte Junior High School, and Ivanna Eudora Kean High School. After Hurricane Marilyn hit the island in 1995, Cranos transfered to Hugh M. Cummungs High School in Burlington, NC, where he received his high school diploma in 1996. He attended North Carolina Agricultural and Technical State University for his undergraduate studies where he received his Bachelor of Science in Electrical Engineering in 2001. Upon completion of his B.S., Cranos entered into the Master of Science program in Electrical Engineering at North Carolina State University. He successfully completed his M.S. in Electrical Engineering in 2002. In January 2003, he entered the Ph.D. program at North Carolina State University where his research interests focused on the development of engineering-based approaches for analyzing and understanding biological systems.

Cranos has participated in several internship and co-op opportunities. He worked between 1997 and 2001 for General Electric where he held several positions, which included Systems Analyst, Systems Test Engineer, and Applications Engineer. In 2001 and 2002, he participated in internship opportunities with Brookhaven National Laboratory, where he was a summer research student under the GEM Master's Fellowship. In 2003, he interned with the Eastman Kodak Company as a summer research student, this time under the GEM Ph.D. Fellowship.

Cranos has received several awards and has participated in several professional and scholarly activities. He received the National Science Foundation Graduate Research Fellowship, the NASA Harriett G. Jenkins Predoctoral Fellowship, and the National Consortium for Graduate Degrees for Minorities in Engineering and Science (GEM) doctoral and masters fellowships. He served as the president (2002/2003) and coordinator (2003/2004) for the Electrical and Computer Engineering Graduate Student Association. During his doctoral tenure, Cranos has served as a reviewer for two IEEE journals, worked on two research proposals, presented at several conferences, given several invited talks, and was invited to give a guest lecture for a course on systems biology.

ACKNOWLEDGMENTS

First and foremost, I would like to thank God for giving me the strength and endurance to get through this process. I would like to extend much thanks and appreciation to my advisor and mentor Dr. Winser E. Alexander. Your guidance, your wisdom, and your advice was instrumental in helping me to grow personally and professionally into the person that I am today. I would also like to thank my co-chair Dr. William W. Edmonson for his guidance and support. I also thank Dr. Wesley Snyder, Dr. Anne Stomp, Dr. Mladen Vouk, Dr. Heike Sederoff, Dr. Ron Sederoff, Dr. Vincent Chiang, and Dr. Joel Ducoste for their assistance throughout my dissertation research.

I would like to thank my girlfriend Teshanee Leonard for helping me to get over the final hump. The last few months were rough but you helped me get through it with your love, grace, and encouraging words; thank you. I have to give a special thanks to my "partners in Ph.D crime" Dr. Senanu Ocloo and Dr. Ramsey Hourani. We struggled together, we laughed together, we played together, we sank together, we swam together, and we passed together. I couldn't have asked for a better support system. I thank my close friends Gary Charles and Jamie Walls for being there to talk about life and things other than the Ph.D. I also thank my friends Karreem Hogan, Julian Taylor, Lalit Ponnala, Chuanhua Xing, Lisa McGee, Treshauna Wright, Youngsoo Kim, Kyle Luthy, Matt Craver, and all past and current members of the High Performance DSP group for all your support.

I would like to acknowledge the fellowship programs that gave me the opportunity to fund my graduate research. Sincere thanks go out to the National Science Foundation Graduate Research Fellowship Program, the Harriett G. Jenkins Predoctoral Fellowship Program, and the National Consortium for Graduate Degrees for Minorities in Engineering and Science Fellowship Program.

Last but certainly not least, I would like to acknowledge my family for all of the love and support they have given me over the years. I would like to thank my parents, Irvin 'Buckey' Williams Sr. and Jasmin Williams. I would like to thank my big brother Irvin for always being that solid role model. I would like to thank my sister-in-law Kim and my sister Nakia for being my big sisters. And finally I would like to thank my nieces and nephews, Kiki, Darion, Jeremy, and Molina for showing me the lighter side of life when things got rough.

TABLE OF CONTENTS

LIST OF TABLES ..... **vii**

LIST OF FIGURES ..... **ix**

1 Introduction **1**
1.1 Overview 1
1.2 Motivation 4
1.3 Contributions of This Work 5
1.3.1 Intellectual Merit 5
1.3.2 Broader Impacts 6
1.4 Organization 7

2 Systems Biology: Background and Related Work **9**
2.1 What is Systems Biology? 9
2.1.1 Notable Achievements in Systems Biology Research 11
2.1.2 Phases of Systems Biology Research 12
2.2 The Rise of "-omics" Data 13
2.3 Identification and Analysis of Functional Relationships 15
2.3.1 Topological Analysis Tools 15
2.3.2 Data-Driven Clustering 15
2.3.3 Blind Source Separation 16
2.4 Dynamic Modeling of Biochemical Pathways 17
2.4.1 Boolean Networks 17
2.4.2 Ordinary Differential Equations 18
2.4.3 Cybernetic Models 20
2.4.4 Petri Nets 20
2.4.5 Estimating Unknown Aspects of Dynamic Biochemical Models 21
2.5 Chapter Summary 22

3 Blind Analysis of High-Throughput Regulatory Data **24**
3.1 Motivation 24
3.2 Hidden Regulatory Characteristics in Microarray Data 26
3.3 Independent Component Analysis - Blind Estimation of Hidden Characteristics 31
3.3.1 Assumptions of ICA 35
3.3.2 Ambiguities of ICA 36
3.3.3 How Does ICA Work? 37
3.3.4 Computing Independent Components 39
3.3.5 Icasso MATLAB Software Package 40
3.4 Identifying Hidden Regulatory Characteristics using ICA 44
3.5 Simulations and Results 46

3.5.1 Data . . . 47
3.5.2 Analysis . . . 48
3.5.3 Results . . . 49
3.6 Chapter Summary . . . 68

4 State and Parameter Estimation in Biochemical Networks using Optimal Control Theory . . . 73
4.1 Motivation . . . 73
4.2 Mathematical Representation . . . 76
4.3 Estimation Methodology . . . 78
4.3.1 Optimal Control Formulation . . . 79
4.3.2 Conjugate Gradient Optimization Approach . . . 93
4.4 Simulations and Results . . . 96
4.4.1 Simulated Measurements Containing no Uncertainty . . . 97
4.4.2 Simulated Measurements Containing Uncertainty . . . 105
4.5 Chapter Summary . . . 130

5 Conclusions and Future Work . . . 134
5.1 Conclusions . . . 134
5.1.1 Blind Analysis of High-Throughput Regulatory Data . . . 134
5.1.2 State and Parameter Estimation in Biochemical Networks . . . 135
5.2 Future Work . . . 137

Bibliography . . . 139

LIST OF TABLES

Table 3.1 Cluster Quality Indices 52

Table 3.2 List of σ^B Genes 54

Table 3.3 Top Functional Groups for Differentially Expressed σ^B Genes 55

Table 3.4 σ^B Connectivity Matrix for Functional Groups 5, 6, and 7 57

Table 3.5 Alternate Expression σ^B Genes 57

Table 3.6 Top Functional Groups for Differentially Expressed Alternate Expression σ^B Genes 59

Table 3.7 Alternate Expression σ^B Connectivity Matrix for Functional Groups 6, 5, and 10 59

Table 3.8 List of Class III Heat Response Genes 61

Table 3.9 Top Functional Groups for Differentially Expressed Class III Heat Response Genes 61

Table 3.10 Class III Heat Response Connectivity Matrix for Functional Groups 6 and 5 61

Table 3.11 List of Arginine Biosynthesis and Transport Genes 62

Table 3.12 Top Functional Groups for Differentially Expressed Arginine Biosynthesis and Transport Genes 63

Table 3.13 Arginine Biosynthesis and Transport Connectivity Matrix for Functional Groups 11, 6, 20, 10, and 13 63

Table 3.14 List of 'Other' Heat Response Genes 66

Table 3.15 Top Functional Groups for Differentially Expressed 'Other' Heat Response Genes 67

Table 3.16 'Other' Heat Response Genes Connectivity Matrix for Functional Groups 11, 6, 10, 12, and 5 67

Table 3.17 List of Potential σ^B Related Genes 68

Table 3.18 Summary of Results from ICA Analysis 69

Table 4.1 Discrete-Time Measurement Profiles 98

Table 4.2 Sparse Uniform Observations: True Values & Average Estimates 101

Table 4.3 Uniform Observations: Relative Error Statistics of State Trajectories 103

Table 4.4 Non-uniform Observations: True Values & Average Estimates 103

Table 4.5 Non-uniform Observations: Relative Error Statistics of State Trajectories . . 104

Table 4.6 Uniform Observations w/ 4% Error: Uncertainty Statistics of Estimated State Trajectories 109

Table 4.7 Uniform Observations w/ 10% Error: Uncertainty Statistics of Estimated State Trajectories 114

Table 4.8 Uniform Observations w/ 15% Error: Uncertainty Statistics of Estimated State Trajectories 116

Table 4.9 Non-uniform Observations w/ 4% Error: Uncertainty Statistics of Estimated State Trajectories 121

Table 4.10 Non-uniform Observations w/ 10% Error: Uncertainty Statistics of Estimated State Trajectories 125

Table 4.11 Non-uniform Observations w/ 15% Error: Uncertainty Statistics of Estimated State Trajectories 129

LIST OF FIGURES

Figure 3.1 Underlying Regulatory Response to Experimental Stimuli 29

Figure 3.2 Recreation of the Blind Source Separation *Cocktail-Party Problem* 32

Figure 3.3 Two-Dimensional Projection of Cluster Similarities 51

Figure 3.4 Absolute Expression Activity for Functional Group 5 55

Figure 3.5 Absolute Expression Activity for Functional Group 6 60

Figure 3.6 Absolute Expression Activity for Functional Groups 6 and 11 64

Figure 3.7 Absolute Expression Activity for Experiment Profile $58\mu M$ of H_2O_2 at 3 min. 70

Figure 4.1 Template Signal Transduction Pathway 77

Figure 4.2 Measurements: Uniform Samples 99

Figure 4.3 Meaurements: Non-uniform samples 99

Figure 4.4 Uniform Observations: True vs. Estimated State Trajectories 102

Figure 4.5 Non-uniform Observations: True vs. Estimated State Trajectories 104

Figure 4.6 Uniform Observations w/ Uncertainty: 4% Error 107

Figure 4.7 Uniform Observations w/ 4% Error: Variation of Estimates 108

Figure 4.8 Uniform Observations w/ 4% Error: Range of Estimated Unmeasured Trajectories 109

Figure 4.9 Uniform Observations w/ 4% Error: Range of Estimated Measured Trajectories 110

Figure 4.10 Uniform Observations w/ Uncertainty: 10% Error 111

Figure 4.11 Uniform Observations w/ 10% Error: Variation of Estimates 112

Figure 4.12 Uniform Observations w/ 10% Error: Range of Estimated Unmeasured Trajectories 113

Figure 4.13 Uniform Observations w/ 10% Error: Range of Estimated Measured Trajectories . . . 113

Figure 4.14 Uniform Observations w/ Uncertainty: 15% Error . . . 115

Figure 4.15 Uniform Observations w/ 15% Error: Variation of Estimates . . . 116

Figure 4.16 Uniform Observations w/ 15% Error: Range of Estimated Unmeasured Trajectories . . . 117

Figure 4.17 Uniform Observations w/ 15% Error: Range of Estimated Measured Trajectories . . . 117

Figure 4.18 Non-uniform Observations w/ Uncertainty: 4% Error . . . 119

Figure 4.19 Non-uniform Observations w/ 4% Error: Variation of Estimates . . . 120

Figure 4.20 Non-uniform Observations w/ 4% Error: Range of Estimated Unmeasured Trajectories . . . 120

Figure 4.21 Non-uniform Observations w/ 4% Error: Range of Estimated Measured Trajectories . . . 121

Figure 4.22 Non-uniform Observations w/ Uncertainty: 10% Error . . . 123

Figure 4.23 Non-uniform Observations w/ 10% Error: Variation of Estimates . . . 124

Figure 4.24 Non-uniform Observations w/ 10% Error: Range of Estimated Unmeasured Trajectories . . . 124

Figure 4.25 Non-uniform Observations w/ 10% Error: Range of Estimated Measured Trajectories . . . 125

Figure 4.26 Non-uniform Observations w/ Uncertainty: 15% Error . . . 127

Figure 4.27 Non-uniform Observations w/ 15% Error: Variation of Estimates . . . 128

Figure 4.28 Non-uniform Observations w/ 15% Error: Range of Estimated Unmeasured Trajectories . . . 128

Figure 4.29 Non-uniform Observations w/ 15% Error: Range of Estimated Measured Trajectories . . . 129

Chapter 1

Introduction

1.1 Overview

The ability to fully understand the complex properties and characteristics of living systems has been an appealing yet elusive goal for decades. The primary difficulties involved in arriving at this goal stem from the enormous size and increasing complexity of these systems. The genomes of single cell organisms like bacteria can have thousands of genes. A simple eukaryote like yeast has protein interactions numbering in the tens of thousands. Humans are estimated to have on order one trillion cells. These numbers are overwhelming and are at times incomprehensible when trying to understand the number of combinatorial interactions and reactions that make living systems work. The complexity of such systems increases such confusions. The truth is that living systems are often not sustained by the activity that occurs at a single level of operation. In fact, living systems function at multiple scales which can extend across multiple levels of biological organization. For example, exposure to chemicals can trigger reactions at the gene level that change the activities of molecular receptors. This change in activity can then cause the deregulation of cell cycle checkpoints, which may in turn elevate the risk of tissue damage. A better understanding of the mechanistic behavior of these cascading events is needed to stop, predict, or even prevent such damage from occurring. This is only possible if the collection of intricate, robust, and complex phenomenon that makes living systems function are better

understood.

The approach to understanding living systems that has dominated the biological sciences for many years has been a hypothesis driven approach to research through conjecture and experimentation. This approach has sustained biological research for over a century. It has produced a tremendous amount of qualitative knowledge and has been instrumental in the progression of molecular biology. This qualitative knowledge is reflected in our overwhelming understanding of many subsystems of model organisms. These include subsytems like the *lac* operon [1], the λ phage genetic switch [2], the *trp* operon, and catabolite repression [3]. Examples of model organisms include *Escherichia coli* [4], *Saccharomyces cerevisiae* (yeast) [5], and *Arabidopsis thaliana* [6]. The understanding of the simplest aspects of these systems was accomplished by decades of experimentation and reasoning. Unfortunately, more complicated phenomenon, like the emergence of cancer or the onset of neurodegenerative diseases like Parkinson's disease are much more complex than the phenomenon seen in these model organisms. An understanding of these more complex phenomenon will require a paradigm shift in how we understand biological processes.

Tremendous advancements in experimental approaches have propelled the study of biological systems to the beginning of a new age of understanding. Starting with the sequencing of the genomes of many model organisms, including humans, experimental approaches are now able to capture both qualitative and quantitative information from several levels of biological organization. These include quantitative data that reflect gene expression, metabolite consumption, enzyme activation, and signaling across cells. These data characterize the interactions of very complex systems. Traditional statistical and heuristic approaches are not well suited to effectively make use of this information and rarely give adequate insight into the system-wide interactions that are contained in these large data sets. Instead, researchers are considering computational and analytical tools in the engineering, mathematical, and physical sciences that have been refined specifically to describe, analyze, and understand complex physical systems. These tools must be specially designed to handle the size and complexity of biological systems. Thus, an analytical, mathematical, and computational foundation must be developed that is able to integrate all of the qualitative and quantitative information that has been or can be acquired from biological systems and use it to decipher a system-level understanding of their behavior. Systems biology is an emergent research area that focuses on the development of methodologies for predictive systems modeling and analysis of biological systems. This is accomplished through multi-

faceted approaches that integrate analytical, computational, and experimental techniques as a means of understanding and eventually controlling the dynamics of these systems.

Electrical engineers have, for years, focused on the modeling and analysis of complex electrical, communication, and signaling systems. Specifically, areas like signal processing theory, systems theory, and control theory provide a strong base upon which to build a framework for analyzing biological systems. Many of the tools developed in this discipline were of course not specifically designed to analyze such systems. These tools were developed for applications like voice recognition, analog circuit control, signal detection, computer vision, and wireless communication just to name a few. This does not prevent these methodologies from being invaluable parts of the systems biology initiatives. As with any tool, one must always be careful how it is used because application of the wrong tools to the right problem can lead to misleading or even incorrect conclusions. Thus, the integration of existing or new tools from these disciplines require an understanding of the problem at hand. With this, one must understand the scope of the problem, the nature and limitations of the data, and the limitations of the tools that are being used. This intermediate, albeit, very important step in the progression of systems biology research is needed in order to achieve a meaningful integration of concepts and ideas from diverse disciplines for the purpose of solving complex problems.

The research described in this document focuses on the integration of existing tools and the development of new tools for the purpose of estimating and understanding the unknown or immeasurable characteristics of biological systems. We consider the development of tools for two different scenarios that are commonly encountered. The first looked at the situation where large-scale expression data is available but little is known about the functional relationships between the measured components. We explored the utility of the blind source separation tool, Independent Component Analysis, for establishing functional relationships between active groups of genes in response to changing environmental stimuli. We compared these results to functional characteristics established in the literature. The second looked at situations where complete dynamic representations of biochemical pathways are hindered by our inability to measure all time varying components and kinetic parameters of the systems. We formulated an algorithm based on optimal control that estimates both the component concentrations and the kinetic parameters of the pathway. Special modifications to the algorithm were made which enabled it to address biological constraints on the pathway dynamics without the need for constrained optimization techniques. Exploratory

analysis was used to assess how uncertainty within the observed measurements matriculated through the algorithm.

1.2 Motivation

Current experimental techniques are capable of generating large amounts of quantitative data that reflect the interactions of many components at the molecular level. This data may represent anything from large-scale measurements of metabolite concentrations or relative concentrations of gene transcripts. Often times, the features used to describe the data do not adequately capture the underlying functional relationships between biological processes because this information is often embedded within data. Feature extraction and data transformation methods are useful signal and data processing tools that have been used for many applications in the electrical engineering discipline. These tools have been successfully used to exploit useful similarities within data in order to group entities (data points) based on some common criteria. The most recognizable of such approaches are clustering methods. Although a variety of methods have been used to analyze large biological data sets [7] [8] [9] [10], the true relevance of the results have still not clearly been explained. Some approaches even violate characteristic about the biological entities that are known to be true. This is evident in traditional clustering algorithms which often restrict genes to a single group even though individual genes are known to participate in very diverse functions. The approach used in this document separated the data into statistically independent features. We conjectured that these features were representative of biologically relevant functional groups that have been activated as a result of internal or external perturbations. In this approach, genes could belong to multiple functional groups. This allowed us to infer connections between a single gene and multiple functional groups if necessary. Methods like these have the potential to yield more insight into the structural and functional relationships between groups of genes than some of the more traditional approaches.

Accurate dynamic models of biochemical pathways are the coveted end goal of most systems biology efforts. Although greatly desired, the generation of these models are often hindered by the inability of current experimental techniques to measure all of the necessary entities needed to simulate and analyze these systems. Parameter and state

estimation approaches have long been used in engineering applications. Issues specific to biochemical pathway modeling prevent the direct use of most of these approaches without some level of modification. These issues include but are not limited to incomplete first principles knowledge, unknown regulatory mechanisms, sparse data sets, nonlinear system characteristics, noisy or uncertain measurements, and biological constraints on components and pathway dynamics. Analytical frameworks that take a systematic approach to addressing these issues are needed. The approaches should be sound in their mathematical foundation. Optimal control presents a valuable foundation upon which to build such an approach. These methods are often used to find the best control signal that satisfies some global objective (e.g. minimum time, minimum energy, or minimum error). In our work, we formulated a trajectory tracking approach to an optimal control problem, which was used to estimate the unmeasured component concentrations and kinetic parameters of biochemical pathways. Further modifications were made to address the concerns of constraints on the dynamics of the system. These modifications fit directly into the strong analytical foundation of the optimal control problem and allowed us to find a solution to the constrained problem without the need for constrained optimization.

1.3 Contributions of This Work

1.3.1 Intellectual Merit

The intellectual merits associated with this work are separated into two parts:

Contribution 1

This research presents an exploratory approach to validating the use of independent component analysis as a viable tool for identifying functional relationships that are hidden within gene expressions data. Techniques like this that are shown to be consistent with current knowledge add to the wealth of available tools that are used to analyze biological systems. The consistency of our results represent the initial step in algorithm development, providing a foundation upon which to further refine such approaches. The consistency of these result also justifies the allocation of resources for further experimental

verification. This includes the verification of relationships identified by the algorithm and the verification of the assumptions inherent to the algorithm.

Contribution 2

This research also presents a component concentration and kinetic parameter estimation algorithm based on an optimal control solution to a least-squares estimation problem. The proposed estimation algorithm builds upon the solid foundation of optimal control theory. It addresses a major bottleneck in the creation of dynamic models by addressing several major issues inherent to modeling biochemical pathways, which include incomplete first principles knowledge, sparse data sets, nonlinear pathway dynamics, noisy or uncertain measurements, and biological constraints on pathway dynamics. The biological constraints are addressed using a scheme seen in the optimal control literature but, to this author's knowledge, has not been derived for solving these class of problems. The approach results in a solution to an optimization problem with dynamic constraints that does not require a constrained optimization algorithm. A detailed derivation of the approach for this class of problem is presented along with results demonstrating the applicability of the algorithm to biological template pathways. This methodology is again based on the solid foundation of optimal control theory. Variations to the current algorithm can be used to address more difficult problems like the estimation of time-varying control parameters. Tools that are able to estimate such parameters, which can reflect regulatory pressures on enzyme activity, would be a valuable addition to the suite of systems biology tools used to build dynamic models. These algorithms are also expandable to schemes for estimating the optimal specifications for biological experiments.

1.3.2 Broader Impacts

The design of mathematical approaches that are capable of capturing the immeasurable characteristics of biological systems is a very complicated problem. The complexity of this problem is superseded only by the potential medical and scientific benefits of such approaches. This research facilitates the further development of analytical techniques needed to develop mathematical models of biological systems. Successful development of these models will have a profound impact on the biological sciences. These models can be used

to design and build biological feedback controllers, which would allow the manipulation and exploitation of the processes at the cell's disposal in an effort to meet some unified design objective. These models can also have a direct impact on the development of disease treatments as well as drug design, where failures in these procedures often occur due to incomplete knowledge of the systems involved. From a scientific standpoint, this research integrates knowledge from multiple disciplines which include but is not limited to molecular biology, signal processing, systems and control theory, linear algebra, and optimization theory. Thus, this research further emphasizes the benefits of inter-disciplinary research and collaborative approaches for solving complex biological problems.

1.4 Organization

Chapter 2 presents a foundation for systems biology research. A concise explanation of systems biology research is given. Several notable achievements in systems biology research are also discussed. We elaborate on the primary phases of systems biology research and focus on the phases where active research is being conducted. Particularly, we elaborate on two areas: 1) the identification of functional relationships between biological components and 2) the dynamic modeling of biochemical pathways. For both areas, we elaborate on the current state of the art and discuss areas where contributions can still be made.

Chapter 3 presents an approach for identifying functional relationships between groups of genes by applying blind source separation techniques. The blind source separation approach use here, known as independent component analysis, separates the data into independent components which we propose represent groups of functionally diverse genes. We first set up the problem by presenting the motivation and the context in which hidden regulatory characteristics can produce the response seen in gene expression data. We then outline the theory behind independent component analysis and present useful tools for obtaining and analyzing independent components. We then describe an approach for extracting biological characteristics from the resulting independent components. We apply this approach to expression data obtained from bacteria under several experimental conditions. We then present a comparison between the biological characteristics obtained from our results and results seen in the literature.

Chapter 4 presents an approach for estimating the unknown or immeasurable char-

acteristics of dynamic models of biochemical pathways using an optimal control formulation. We first set up the problem by presenting the motivation behind this work. We then describe the mathematical model of the template biochemical pathway used to test the functionality of the algorithm. The next section describes the original constrained estimation problem and then outlines the optimal control approach used to solve it. A detailed derivation of the approach is given. We then describe the unconstrained optimization algorithm used. We present two scenarios under which the algorithm is tested. The first is the proof of concepts scenario where the estimation algorithm is tested based on measurements free of uncertainty. The next scenario tests the algorithm using measurements that contain varying amounts of bounded uncertainty. Exploratory analysis was used to assess how uncertainty within the observed measurements matriculated through the algorithm. These results are interpreted and summarized.

Chapter 5 summarizes the work presented in this document and presents ideas for future research interests. This includes further development of the algorithms presented in Chapters 3 and 4.

Chapter 2

Systems Biology: Background and Related Work

2.1 What is Systems Biology?

The analysis of systems has long been a focal point in engineering and various areas of the physical sciences. Researchers in the life sciences, however, have been reluctant to embrace these concepts to study the characteristics of living organisms. In a sense, living organism can be viewed as a complex system. This system is composed of networks of biological processes that combine cellular interactions with biochemical reactions to form robust processes that maintain the existence of the organism. These interactions often occur at multiple scales of biological organization, which can extend from the tissue and cellular level to the proteomic, metabolomic, and genomic level. This results in the potential interaction of thousands of proteins, genes, and other molecular components all working in concert to produce what are usually interpreted as qualitative characteristics seen through experimentation. The overwhelming thought of these exponentially increasing interactions did not stop some early biologists, physicists, and mathematicians, notably Lotka [11] and von Bertalanffy [12], from introducing mathematical concepts in the mid 1900s for describing biological systems and their dynamics. These concepts, by and large, were ignored

by the biology community [13]. Biological research at that time was primarily focused on hypothesis driven research whose conclusions were driven primarily by experiments that focused on small aspects of these larger systems. These works resulted in many successful advancements in molecular biology. These advancements were able to reduce these systems to their individual components, identifying many of the molecular mechanisms of these systems. In many cases, extensive knowledge about smaller subsystems like the λ phage genetic switch [2] and the *lac* operon [1] were obtained through decades of experimentation and analysis. More complicated phenomenon that involve tens to thousands of interacting components are still a mystery because the reductionist approaches of the past are not amenable to understanding such complex interactions. This is particular a problem when changes at the genetic, proteomic, or metabolic levels result in significant phenotypic changes (e.g. cancer).

Today, researchers are revisiting some of the "systems" concepts that were initially introduced by early scientists like von Bertalanffy as a means of addressing these problems. This paradigm shift in how biological research is conducted is not occurring primarily because systems analysis is seen as the ultimate answer to many of the difficult medical, social, and environmental problems scientists have encountered in the past several decades. Systems analysis is seen more as a useful tool that, when properly integrated with current experimental data and a-priori knowledge, can provide a more comprehensive approach to understanding biological systems. A series of advancements in biological experimentation, highlighted by the sequencing of the human genome, has resulted in the collection and storage of much of the available experimental data. These data capture characteristics of living systems at multiple levels of biological organization. Sequences of entire genomes provide the total potential inventory of resources available to the cell. Gene expression data provide snapshots into how the cell activates and deactivates these resources according to changing needs. Enzymatic and metabolic profiling provide better insight into the state of the cell, characterizing the chemical conversions and signaling mechanisms essential for cellular survival. A true systems approach facilitates the integration of these data, along with a-prior knowledge about these systems, to develop mathematical models that capture emergent properties and complex behavior that were not explainable by studies of the individual molecular components in isolation. This level of understanding will lead to more substantive explanations of observed functional modules [13]. Eventually, these models will be used to predict how alterations could affect the function and response of the

system, enabling us to avoid undesirable behaviors. This collective effort towards a multi-faceted approach of integrating analytical, computational, and experimental techniques for understanding biological processing is a part of an emerging research area known as *systems biology* [14] [15].

2.1.1 Notable Achievements in Systems Biology Research

An integrative approach to understanding the behavior of biological systems should not be seen as an academic exercise with no meaningful implications on the scientific community. Multidisciplinary research endeavors in systems biology range from small scale efforts which try to establish functional and control characteristics at the molecular level to large-scale projects which undertake the daunting task of providing *in-silico* representations of tissues, organs, and organisms. Hasty et al. investigated a mathematical framework based on regulatory logic that was capable of designing special purpose synthetic gene networks [16]. This framework attempted to provide a foundation for generating new forms of logical cellular control, which could have important applications in gene and cell therapy. Jeong et al. explored the difficult task of elucidating the structure of metabolic networks on a genome wide scale [17]. These networks were created based on graph theory and were used to reveal topological properties as well as recognize common aspects of organization across species. Anderson et al. successfully engineered a transgenic form of *Escherichia coli* that was capable of sensing and invading mammalian cell lines that were specific to cervical cancer. This has large implications on the development of alternative approaches for the sensing and destruction of tumors [18]. Biodiesel initiatives are underway that attempt to exploit the natural mechanisms of microalgae, which are capable of producing 30 times the amount of natural oil per unit area of land as compared to terrestrial oilseed crops [19]. The U.S. Environmental Protection Agency's National Center for Computational Toxicology has initiated a "virtual liver" project whose goal is to develop a multiscale, computational model of the liver that incorporates anatomical and biochemical information relevant to toxicological mechanisms and responses [20].

Other aspects of systems biology directly impact disease treatment and drug discovery. Cost effective ways of producing the highly effective malaria drug Artemisinin was accomplished by the manipulation of several biochemical pathways in yeast [21]. Musante et al. developed complex simulations of disease physiology using a framework of deterministic

differential equations based on empirical patient data [22]. Lewis et al. developed a computational model of chronic asthma that described complex interactions between interacting cells and their environment [23]. They were able to make predictions that were consistent with results from clinical trials [24]. Kansal presented a review of modeling approaches for type 2 diabetes, which attempted to analytically describe integrated responses of multiple tissues and their involvement in glucose metabolism [25].

2.1.2 Phases of Systems Biology Research

Systems biology is an iterative, integrative, and interdisciplinary research area. It requires an approach that integrates experimental and theoretical biology with diverse and often complex computational methods. This requires a unique partnership between biologist, chemists, physicists, mathematicians, and engineers. The precise roadmap to achieving this remains unclear but successful research in this area requires the integration of some or all of the following steps:

1. Top-down approach to identifying specific pathways components of interest and the molecular interactions of these biochemical constituents at the regulatory, metabolic, and signaling levels.

2. Integration of first-principles knowledge with genomic, proteomic, and metabolomic data at various levels of information quality to form an initial mathematical model that replicates the structural and dynamical characteristics of the true systems while reflecting the uncertainty associated with biological variation, uncertain measurements, and modeling errors.

3. An iterative approach to model refinement to verify and validate the analytical model. The model is used to design the most appropriate specifications for experiments. This involves establishing the most appropriate environmental, genetic, and signaling perturbations that best help to improve the model.

4. Culminate these modeling efforts by providing effective theoretical and realizable control schemes that are capable of manipulating the dynamics of the systems. Identification of the biological mechanisms that are capable of realizing these control schemes are essential.

Each step above represents huge challenges and genuinely stretches the limitations of current experimental, analytical, and computational approaches. The current state of systems biology research has primarily focused on variations of steps 1 and 2. In this document, we focus on the large scale identification of functional relationships between the components of the system and identification of the unknown or immeasurable characteristics that govern the dynamics of the system. A review of these areas, along with a review of the "-omics" data that is used to develop these approaches, is given in the following sections.

2.2 The Rise of "-omics" Data

In 1944, Avery and his colleagues discovered that deoxyribonucleic acid (DNA) was the genetic material that was responsible for exchanging hereditary traits to the offspring of organisms [26]. Since then, scientists have progressively searched for ways to better describe the mechanisms that contribute to the complex functional properties of living systems. The genome of an organism provides valuable insight into these mechanisms. The genome specifies the total potential inventory of resources available to the cell. It is composed of thousands of genes which, when expressed, translate into proteins that are responsible for the various functions operating within the cell. Thus, the ability to sequence the genome and elucidate these genes was an important milestone in the overall goal of understanding living systems. This milestone was achieved when genomes of many model organisms, including humans, were sequenced [4] [5] [6] [27] [28]. The sequencing of these organisms generated terabytes of information, which are now stored in databases scattered around the world. These large amounts of data gave rise to the field of Bioinformatics, which used data mining, statistics, and heuristics to analyze these enormous data sets. Tools like BLAST [29] and CLUSTALW [30] were used to perform sequence alignments with well studied organisms in order to identify the functional portions of genomes. This resulted in extensive knowledge of the building blocks of these complex machines. Unfortunately, systems that are as complex as biological processes are normally not easily described by there parts alone. The complexity normally arises from the complex interactions of these parts.

The sequencing of these genomes did have one interesting consequence. Knowledge of genes and gene sequences enabled the progression of very intricate experimental

approaches. These quantitative experimental procedures were able to acquire snap shots of behavioral responses that occurred due to perturbations to the overall system. Reverse transcription polymerase chain reaction (RT-PCR) and Northern blot [31] are useful tools for quantifying small concentrations of ribonucleic acid (RNA), which reflects the activation or deactivation of a given gene. These tools, although relatively accurate, are time consuming and expensive. This makes them not well suited for large-scale or genome-scale analysis. Large-scale analysis tools, like microarrays, are able to capture the relative abundance of RNA of thousands of genes at a time. These results are often not as accurate as RT-PCR but it does allow for the analysis of entire genomes at a much lower cost. Gene expression, however, is only an indirect indicator of the cellular response. Proteomics and metabolomics try to fill in the missing information that is not characterized by genomics. Proteomics focuses on the identity, abundance, processing, chemical modification, and interactions of proteins in order to assess their role in complex behaviors. This was successfully accomplished by experimental tools that quantified these characteristics. Two dimensional (2D) electrophoresis and mass spectrometry [31] are useful experimental approaches for identifying and quantifying large numbers of proteins. Enzyme-Linked ImmunoSorbent Assay or ELISA is another approach for quantifying proteins [31]. Metabolomics focus on assessing the conversion, consumption, and production of small molecules needed for the organism to survive. Mass spectrometry and nuclear magnetic resonance are experimental techniques that have been used to quantify metabolic profiles [31].

The quantitative data described above provides information that yields insight into the functional interconnections of components within the system. This data is often sparse and incomplete. Large-scale data tend to provide a course assessment of global characteristics. The functional relationships are often hidden within this data and are not easily interpreted without the use of additional tools. Time course measurements of smaller subsystems are sometime available, but only at limited scales. This data is often incomplete. This is a direct result of limitations on the experimental procedures which are unable to measure all time varying components of the system. Ideally, quantitative measurement of all components and parameters would be necessary for building dynamic models of the subsystems. These limitations on the data sets, however, are inescapable. This highlights the need for analytical and computational tools that are able to build functional relationships and model system dynamics based on these incomplete and sparse data sets. Below we describe some of the techniques that have been used in the literature.

2.3 Identification and Analysis of Functional Relationships

2.3.1 Topological Analysis Tools

Linear algebra provides a useful foundation upon which to base many topological analysis tools for systems biology. Metabolic Flux Analysis (MFA) and Flux Balance Analysis (FBA) provide useful tools for analyzing the theoretical capabilities and operative modes of metabolic pathways using linear algebra [32] [33]. This is done by performing structural analysis on the stoichiometric matrices of a given pathway. Generally, MFA requires that the number of measured fluxes in the pathway be greater than the number of unmeasured fluxes. This produces a solvable system of fluxes that is tied together by the functional relationships between the metabolites and the catalytic enzymes. FBA generally does not have this requirement. This technique uses a combination of linear algebra and optimization approaches to estimate conversion relationships between the metabolites and the enzymes. These relationships are dictated by some predetermined operational objective. Fluxes are estimated that satisfy both the structural relationship and maximizes or minimizes this objective. Examples of this operational objective include maximization of nutrient uptake, minimization of metabolite production, or minimization of adenosine triphophate (ATP) production.

2.3.2 Data-Driven Clustering

Tools like MFA and FBA require a significant amount of a-priori knowledge about the generally stoichiometry of the pathway in order to quantify functional production and consumption relationships between the enzymes and the metabolites. This insight is often difficult to obtain for larger networks. This is particularly true for gene regulatory pathways. Organisms often have genomes that contain anywhere from thousands to tens of thousands of genes. Establishing stoichiometric and/or functional relationships between all of these components is extremely difficult. Data-driven tools provide ways of deciphering functional relationships in the data based on empirical relationships between data feature variables. When using microarray data for example, these feature variables are the expression levels for a particular gene. Clustering tools attempt to find similarities between genes across a range of conditions or samples. Several types of clustering approaches have been

applied to expression data. They include hierarchical clustering [7], k-means clustering [8], graph-theoretic clustering [34], and self organizing maps [35]. Graphical inference models based on correlations have been used to establish relationships between genes [9]. Herrero et al. also formulated an alternative approach for clustering microarray data based on unsupervised neural networks [10]. Several interesting issues arise when using these types of data-driven tools to analyze biological data. Since the results are entirely based on the data, it is important to ensure that the metrics used will produce functional connections that are biologically relevant. Commonly used metrics like Euclidean distance and Pearson correlations that are applied to features like relative expression levels may lead to functional connections that are not necessarily true. For example, using a clustering method with Euclidean distance as the metric will tend to infer functional relationships between groups genes that have similar expression characteristics. We know from characteristics seen in the literature that this is not always true [36].

2.3.3 Blind Source Separation

Blind source separation techniques are special cases of system identification approaches that assume that both the structure and underlying relationships of the pathway are unknown and tries to estimate both based on some underlying relationship between the inputs. Techniques like Principal Component Analysis (PCA) and Factor Analysis (FA) assume that the inputs are statistically uncorrelated. Factor analysis imposes an additional statistical constraint on the inputs by restricting them to be Gaussian distributed. Independent Component Analysis is also a statistical separation tool that instead assumes that the inputs are statistically independent. Linear algebra techniques like Singular Value Decomposition (SVD) have been used to separate data but the assumptions are similar to those of PCA. Several of these tools have been used to analyze gene expression data with varying levels of success [37] [38] [39]. Tools like Network Component Analysis do not rely on statistics to perform the separation [40]. This tool does requires a-priori knowledge about a minimum number of relationships in the pathway in order to solve the problem uniquely. This may reduce the general applicability of this algorithm for some situations. These blind source separation techniques do provide a useful theoretical basis upon which to base a computational algorithm for identifying unknown functional relationships. The underlying mechanisms that are used to perform the separation is often mathematical in its

foundation and may not tie back to the biology. More work is needed to determine if these types of techniques do in fact produce results that are biologically significant. Establishing whether or not some of the criteria used in these source separation tools are valid will further promote the investigation and development of similar tools for building functional relationships in biological systems.

2.4 Dynamic Modeling of Biochemical Pathways

These models attempt to coordinate an accurate representation of the dynamics of biochemical processes by presenting an aggregate model of the nonlinear dependence of substrates, products, and effectors. These models are used to capture dynamic properties such as shifts in operating states, cooperative relationships between entities, system robustness, and system stability. Increased granularity in the detail of the representations requires larger and larger amounts of theoretical and empirical information in order to make these systems accurate. As a result, this need for information imposes restrictions on the size of the model given a particular level of granularity desired. Several choices for modeling and simulating the dynamics of biochemical pathways are available that vary in scale and granularity. They include Boolean Networks, Ordinary Differential Equations, Cybernetic Models, and Petri nets.

2.4.1 Boolean Networks

Boolean networks are logic-based representations that attempt to express the dynamic characteristics of large-scale biochemical networks when detailed kinetics are unavailable [41] [42]. The concept of Boolean networks was originally introduced by Kauffman [43]. Good reviews can be found in Huang [44], Kauffman [45], Somogyi and Sniegoski [46], and Aldana et al. [47]. An appealing property of Boolean networks is that they emphasize the general interactions of components instead of quantitative biochemical details. The networks, however, are not capable of simulating dynamic changes in component concentrations. This particular choice of model description is normally used to describe large-scale networks with hundreds and sometimes thousands of interacting components as is the case for genome-scale regulatory and metabolic pathways. These models represent each com-

ponent as a Boolean variable which can take on a value of on or off. The interactions between components are represented by Boolean functions, which determine the on/off status of a component based on the collective states of all other components in the system. This somewhat generic approximation gives valuable insight into the fundamental course-grained properties of large networks. These properties are used to simulate and predict how targeted perturbations affect complex biological phenomenon, such as cellular state dynamics, limit cycles, multiple steady states, and attractor characteristics. This provides a useful analytical tool for aiding in the potential for the discovery of novel targets for drugs [48]. The Boolean functional relationships between components are either inferred from prior knowledge or are estimated empirically based on high-throughput quantitative data (e.g. microarray analysis). Other variations of Boolean networks include probabilistic Boolean networks, which are able to cope with uncertainty in both the data and the model selection [41] [49].

2.4.2 Ordinary Differential Equations

Ordinary differential equations (ODEs) are usually the representation of choice when extensive information about the pathway kinetics has been acquired either experimentally or empirically. Models represented by ODEs incorporate information regarding structure and function of the pathway, applicable chemical and physical laws, material properties, components involved, and process rates. These models have been used to model a variety of biological processes, which include genetic regulatory pathways [50], metabolic pathways [51], signaling pathways [52], and viral infection of cells [53]. Unlike Boolean networks, these representations attempt provide the detailed descriptions of the time varying dynamics of all components with the system. This added level of detail comes at high information and data costs. These models often require detailed descriptions of molecule interactions and enzymatic kinetics. Several different representations have been used to describe these kinetics.

Mass-action kinetics are often used to build ODE models of biochemical networks using mathematical representations of the rate kinetics of elementary reactions [13] [54]. The rate kinetics for these models are usually determined by the participating reactants and their network relationships with other components in the system. It is given by a mathematical formula that expresses the kinetics as a product of kinetic rate parameters

and power functions of component concentrations. This provides an added advantage to models based on mass-action kinetics because it can be derived using network structure and knowledge of the participating components in a given reaction. It suffers from the fact that the number of rate parameters grow exponentially as more reactions are added to the system. These parameters can either be resolved through experimentation or empirically based on time-course measurements of component concentrations.

As an alternative, enzyme catalyzed reactions can be modeled using Michaelis-Menten kinetics, which represent the net flux through several elementary reactions that use the same enzyme [55]. These laws are represented as nonlinear relationships between the quantity of available substrate and the reaction rate and are define by two parameters, $V_{\max}$ and K_m. The variable $V_{\max}$ gives the saturation rate of the reaction and K_m gives the substrate quantity at which half of this maximum rate is achieved. These laws can be derived from mass-action kinetics under assumptions regarding the steady-state behavior of the complexes and the ratio of substrate and enzyme concentration [13]. In fact, the kinetic equations obtained through mass action kinetic are similar to the linear operating region of the Michaelis-Menten saturation curve when substrate concentrations are well below K_m. A reversible form of the Michaelis-Menten kinetics is also available [56].

Biochemical systems theory presents an assortment of other forms for representing the kinetics of biochemical pathways [57] [58] [59]. These representations were developed in order to generate a consistent set of analytical rules that can be applied to generating dynamic biochemical models. These representations are based on power-law approximations of all rate processes. General Mass Action Systems are the most general form. This representation presents the rates kinetics as a product function of component concentrations that each have kinetic orders that reflect molar participation in a given reaction. Other useful representations include S-Systems, Lotka-Voltera systems, generalize Lotka-Voltera systems, and half-systems or Ricatti systems [13]. These representations are capable of approximating other forms of kinetics, like Michaelis-Menten, while still providing a representation that is mathematically and computationally efficient. System analysis tools can be used to analyze the characteristics of any of these representations. Analysis tools include bifurcation analysis, parameter sensitivity analysis, steady state sensitivity analysis, and stability.

2.4.3 Cybernetic Models

In biochemical systems, many of the reactions of interest are catalyzed by enzymes, which may be inhibited or activated by regulatory effects on the genes that code for these enzymes. As a result, these changes at the regulatory level may impact the dynamics of a biochemical pathway and should be reflected in the respective model representation. The impact of regulatory changes at the gene level is difficult to capture with traditional ODE representations. Cybernetic models expand on traditional ODE representations by incorporating control laws that manipulate enzyme synthesis and activity in order to drive the pathway toward some nutritional or ecological goal [60]. As a result, this methodology is also capable of predicting how a metabolic network redirects enzymatic expression and activity in the face of environmental perturbations and genetic alterations. The framework is based on the postulate that biological systems function according to some optimal resource allocation objective [61]. These biological objectives are developed as a direct result of long-term evolutionary pressures imposed on the organism. Cybernetic modeling hypothesizes that the cellular objectives of the wild-type and the genetically altered systems are identical, supporting the use of the same objective function for predicting changes in dynamics due to regulatory perturbations. Generating a cybernetic model from an arbitrary metabolic model is no easy task. In fact, this process is very involved and requires a great deal of prior knowledge about the system to be effective. This knowledge may include a closed-form representation of all enzyme kinetics, intra- and extra-cellular measurements of component concentrations, kinetic rate parameters, and a quantitative representation of the true cellular objectives. This representation has proven useful in achieving the multiple steady states that occur in Hydridoma cultures in continuous flow bioreactors [62] and in capturing the regulatory effects for cellular organisms that exhibit diauxic growth under mixed substrate metabolism [60]. The demand for large amounts of information makes this methodology very difficult to employ for general metabolic networks.

2.4.4 Petri Nets

Discrete event simulation tools, like Petri nets, provide a means of modeling, analyzing, and simulating biological behavior using various levels of both qualitative and quantitative knowledge. This expands on the capabilities of purely qualitative dynamic

representations like Boolean networks and purely quantitative dynamic representations like ODEs and cybernetic models. Traditional and Colored Petri nets have been used to obtain qualitative dynamic characteristics of biochemical pathways using only structural information [63] [64] [65]. Hybrid Petri nets and Hybrid Functional Petri nets have been used as a quantitative counterpart for integrating various forms of the enzyme kinetics for predicting model dynamics [66], for model validation [67] and for determining the effects of gene regulation and enzyme inhibition on component concentrations [68]. Numerous tools and functional languages are useful for integrating information from biological databases to build, visualize, and simulate biochemical pathways using Petri nets [69] [70]. Detailed Petri net representations of biochemical networks that accurately model and simulate the dynamic characteristics of the pathway again require information about regulatory changes and enzyme kinetics with associated rate parameters. Missing information can severely hinder the accuracy of such representations.

2.4.5 Estimating Unknown Aspects of Dynamic Biochemical Models

A common theme that is seen in most of these dynamic representations is the need for a great deal of a-priori information about the modeled pathway. This includes the general structure and interactions of the biological components, the initial concentrations of each component, and the kinetic parameters that govern the transformation and conversions due to enzyme activity. There are many situations where some or all of these individual components are unknown and one needs to estimate characteristics based on the things that can be measured. There is evidence in the literature of methodologies that take many different approaches to estimating the unknown or unmeasurable characteristics of dynamic models of biochemical networks. These approaches involve numerical optimization, genetic algorithms, simulated annealing, and interval analysis. These procedures tend to focus on specific unknown aspects of the pathway. Gadkar et al. developed a method based on the quadratic state regulator problem for estimating the unmeasurable states in signal transduction networks when faced with missing information about the closed form analytical representation of the reaction rates [71]. Williams et al. expanded this work to explore the effectiveness of global stochastic search algorithms for performing the estimation [72]. Bastin and Dochain used an adaptive non-linear observer for estimating specific growth rates and biomass concentrations [73]. Both van Riel and Sontag along with Zheng

and Rundell studied approaches for estimating the kinetic parameters of biochemical pathways [74] [75]. Tucker et al. and Polisetty et al. both looked at deterministic global search approaches for the parameter estimation of biochemical models that are based on generalize mass action systems [76] [77]. The estimation of unobserved states can be a direct extension of approaches that estimate these parameters based on limited or incomplete concentration measurements of the network components. Voss et al. and Schittkowski approach this problem by using multiple shooting methods to solve what is known as the initial value problem where both the parameters and the states of the system are estimated concurrently [78] [79]. Timmer et al. applied similar methods to chaotic systems [80]. A-priori biological knowledge is often integrated into these estimation approaches, which allows for the implementation of equality or inequality constraints on the parameters in order to reduce the search space of the estimation problem [81]. A-priori knowledge about the system can lead to static constraints on the parameters of interest and dynamics constraints on the trajectories of the network. The implementation of an approach that would allow a systematic integration of dynamic constraints into the state and parameter estimation problem without significantly increasing the complexity of the estimation algorithm would be a useful addition to the library of tools used to better model the dynamic characteristics of these systems. We investigated one such approach in our work.

2.5 Chapter Summary

Systems biology is a an emerging research area that focuses on the development of experimental, analytical, and computational approaches for modeling, understanding, and eventually controlling biological system. The research is facilitated by the tremendous advancements that have been made in current experimental techniques, which are able to assess both qualitative and quantitative biological characteristics. Much of the research in systems biology has focused on deciphering the structural and dynamic characteristics of these systems. Current techniques are a long way from being as refined as those used to analyze physical systems. In this chapter, we presented the current state of the art in systems biology research. We presented a progressive assessment of techniques that have been used to identify structural and dynamic relationships in various forms of biochemical networks. Approaches for structural analysis included data-driven modeling techniques like

hierarchical clustering, k-means clustering, unsupervised neural networks, and graph theoretic approaches. Blind source separation techniques were also discussed as an alternative to these methods. Various forms of dynamic representations for biochemical pathways were given. This included large-scale course representations like Boolean networks and more detailed representations like ODEs and cybernetic models. We also discussed Petri nets, which are designed to incorporate both qualitative and quantitative information for providing models with varying levels of granularity. In the following chapters we investigate the viability of two computational approaches for identifying unknown characteristics. Chapter 3 focuses on a blind source separation techniques for uncovering regulatory relationships. Chapter 4 describes the formulation of an estimation approach that addresses issues inherent to the dynamic ODE modeling of biochemical pathways. Chapter 5 summarizes the results obtained from both approaches and presents ideas for future research topics.

Chapter 3

Blind Analysis of High-Throughput Regulatory Data

3.1 Motivation

The expression patterns of genes in the genome are influenced by changes at many levels. This is a result of a complex regulatory network of biochemical constituents that allows the organism to change its expression patterns in response to both internal and external changes. For example, an organism may change its expression patterns in order to accommodate for changes in available food sources. Changes in expression patterns associated with defense and cell preservation may be activated when the organism senses danger. This may include invading organisms or drastic changes in environmental conditions (e.g. drought, extreme heat, or extreme cold). Other changes in expression patterns may result from changes in the genome itself, which may occur from random or intentional mutations. These changes may affect the organism positively, allowing it to thrive in environments that it would otherwise not be able to. In other cases, these changes may be lethal or fatal, resulting in the eventual death of cells within the organism or the death of the organism itself. The former scenario may be very useful for manipulating food crops to survive under extreme climate conditions, thus enabling these crops to grow in otherwise desolate

areas. The latter scenario may be instrumental in fighting diseases like cancer and HIV. Effective, safe, and consistent manipulation of living organisms at the genetic level requires an in-depth understanding of the regulatory connections of the many functional groups in the genome; an understanding that has not been achieved for many organisms. The first step in acquiring this in-depth understanding is to develop approaches that are capable of deciphering these connections from global gene expression patterns of an organism in response to changing stimuli. Prevalent techniques found in the literature, like hierarchical clustering [7], identify functional groups by categorizing genes according to similarities in expression characteristics. In many cases, however, genes with similar expression patterns belong to very diverse functional groups and genes with different expression patterns belong to the same functional group. Thus, other criteria are often needed to effectively decipher these convoluted responses.

In this work, we used independent component analysis to extract hidden regulatory characteristics from microarray data obtained from the bacteria *Bacillus subtilis*. We first explored conditions under which the expression profiles in microarray data are formed by a mixture of underlying responses that are activated as the result of the experimental stimuli. We then investigated and reformulated the theory, assumptions, and ambiguities associated with independent component analysis. This understanding of the underlying assumptions of the algorithm was important in evaluating the biological relevance of the results it produces. We investigated two software packages that have been used to both extract and evaluate independent components derived from mixed data. The evaluation tool was used to assess the algorithmic integrity of the results. We then formulated an analysis scheme for extracting the biological characteristics from the resulting independent components. This analysis scheme was used to interpret the relative expression activity and potential functional relationships between groups of genes that were found in the literature to be highly responsive to the experimental stimulus. The functional relationships that were identified were better able to capture true biological characteristics due to the approach's unique ability to categorize genes as belonging to multiple functional groups. The characteristics from the evaluation tool and the extracted biological characteristics were used to evaluate the relevance of our results. This involved a comparison between our results and results that have been established in the literature. Consistent results confirmed the viability of this approach for uncovering functional relationships between groups of genes. This approach serves as another step towards attaining a global understanding of gene reg-

ulatory networks. It is imperative that we obtain this understanding in order to increase the success of gene manipulation endeavors.

This chapter is outlined as follows. Section 3.2 proposes a heuristic representation that describes microarray data as a convoluted response of active functional groups that respond to changing stimuli. This representation is used throughout the remainder of the analysis. Section 3.3 presents the theory, assumptions, and ambiguities associated with independent component analysis. The FastICA and Icasso software packages are described. These tools are used to extract and evaluate the independents components from the microarray data. Section 3.4 presents an analysis approach for extracting biological characteristics from the independent components that were obtained. Section 3.5 describes our analysis approach, which includes a description of the data used, the specifications used to perform the simulations, and the results from our analysis. Section 3.6 summarizes the results obtain from the approach described in this chapter.

3.2 Hidden Regulatory Characteristics in Microarray Data

Gene regulatory networks control how thousands of genes in the genome are turned on and off in response to changing stimuli. The patterns in which these genes are turned on and off are called regulatory patterns and are generated by complex interactions between many biochemical molecules. Dynamic changes in expression allows the organism to appropriately respond to changing environment, changes in external signaling, or changes in its internal structure such as random or induced mutations. Prior to the late 1990s, snapshots of the global response of all genes in the genome to a specific stimuli were generally unavailable. Regulatory analysis was often restricted to small groups of co-regulated genes [1] [2] , limiting our ability to analyze regulatory characteristics of the entire genome. Tremendous progress in high throughput experimental techniques has yielded large amounts of data. Using this data, we are able to quantify the behavior of the organism at the genomic, proteomic, and metabolomic levels [82] [83]. Microarray data have taken advantage of advancements in high throughput sequencing, revolutionizing the way we analyze gene regulatory patterns. Microarray data are able to globally quantify how changing stimuli correlates to changes in the regulation of gene transcription and expression [84]. Microarray technology has long been seen as having the potential to address many interesting questions

in genetics by revealing patterns of expression for genes and classifying samples and genes based on such patterns [85]. Microarrays quantify the relative increase and decrease in transcription levels of every gene in the genome by measuring the relative change of mRNA concentrations with respect to some control experiment. The output data from microarray experiments are typically the end products of low-dimensional regulatory signals driven through some interacting network [40]. Thus, it is thought that the change in transcription levels of each gene in the genome is orchestrated by regulatory mechanisms that are programmed to respond to a given stimulus in a very specific manor. The microarray array experiments capture the end result of the pressures of these regulatory mechanisms on gene regulation but the underlying characteristics of these regulatory mechanisms often remain hidden; convoluted in the overall transcriptional response of the genome. We envisioned these underlying mechanisms as sets of complex interactions between groups of genes that share common characteristics such as function, expression response, and regulatory mechanisms.

Denote $\mathcal{X}_i$ as a group or set of genes in the genome that all participate in a common function or response when presented with a particular stimuli. We called these sets of genes functional groups (FGs). These functional groups may be composed of genes that participate in an assortment of functions, which include phototropic response, starvation response, response to oxidative stress, or response to changes in food source. The set of stimuli used to perturb the natural response of the regulatory systems often determines which functional groups will have increased regulatory activity. Often, the genes contained in these groups are not mutually exclusive. In other words, there may be situations where for a given gene δ^i

$$\delta^i \in \mathcal{X}_i,\ \delta^i \in \mathcal{X}_j,\ i \neq j. \tag{3.1}$$

For example, the same gene δ^i may participate in a starvation response and a phototropic response. The relationship between genes in the same functional group does not necessarily imply that all of these genes have the same expression characteristics. For this reason, clustering approaches that attempt to separate genes into co-regulated groups of genes are often incapable of capturing the true underlying structure of these functional groups.

We explored a more systematic approach to understanding these functional groups by first proposing an analytical structure that described the participation of these groups

to produce a regulatory response for a given gene. Given the complexity of gene regulatory networks, it was practical to assume that the expression response of gene δ^i for a particular stimulus, stimulus j, was given as a nonlinear function of the functional groups that 1) contain gene δ^i as a member and 2) are activated as a result of stimulus j. The overall regulatory response of gene δ^i was written as a weighted sum of the nonlinear responses from all functional groups activated by stimulus j which contains gene δ^i as a member

$$x_j^{\delta^i} = \sum_{k \in \mathcal{S}} a_{jk} f_k^{\delta^i}([\mathcal{X}_k]). \tag{3.2}$$

Equation 3.2 represents a static model that describes the relationship between the overall expression level of a specific gene and the individual participation of all functional groups that impact the expression level of that gene. This representation provided a generalized nonlinear static structure that was akin to other types of regulatory models seen in the literature [10] [40] [86] [87]. The set $\mathcal{S}$ contains the indices of the relevant functional groups. The function $f_k^{\delta^i}(\cdot)$ is the nonlinear regulatory response from functional group k that contributes to gene δ^i. This is a function of $[\mathcal{X}_k]$, which represents the transcription levels of all genes from functional group k. The variable a_{jk} is a scalar weight that reflects the level of participation of the kth functional group on the overall expression activity of gene δ^i for stimulus j. Our ability to understand the gene makeup of each functional group, the nonlinear responses, and the weights associated with these responses was greatly inhibited by the shear complexity of such a system. This inhibition was exacerbated by the general lack of understanding of which genes and functional groups of genes responded to different stimuli. Ideally, quantitative data can be used to reverse engineer these connections to understand the genes that are particularly responsive to a given stimulus and to understand how these genes perform in concert with other genes that response to other stimuli. The lack of a-prior knowledge about the general structure of the underlying mechanisms limits the use of traditional system identification procedures like those seen in [88] [89]. Black box techniques like neural networks [10] [90] are capable of providing models that describe the data given little a-priori knowledge about the structure of the underlying mechanisms. The models are often phenomenological which provide analytical representations with little biological relevance. These issues emphasize the need for black box data driven analysis schemes that are able to 'blindly' identify and reconstruct biologically relevant connections from quantitative data obtained from the microarray experiments.

A complete blind estimation of the nonlinear interactions and the unknown connectivity of the regulatory pathway is an ill posed problem given the relatively small data sets that are usually obtained from microarray experiments. It was reasonable, however, to limit the estimation approach to the identification of the activated functional groups, the genes that are particularly active within these groups, and the overall impact or weight of relevant functional groups on the expression of a gene in the genome giving some stimulus. Figure 3.1 illustrates a pictorial representation of this relationship between functional groups and the transcriptional response seen from the microarray experiments.

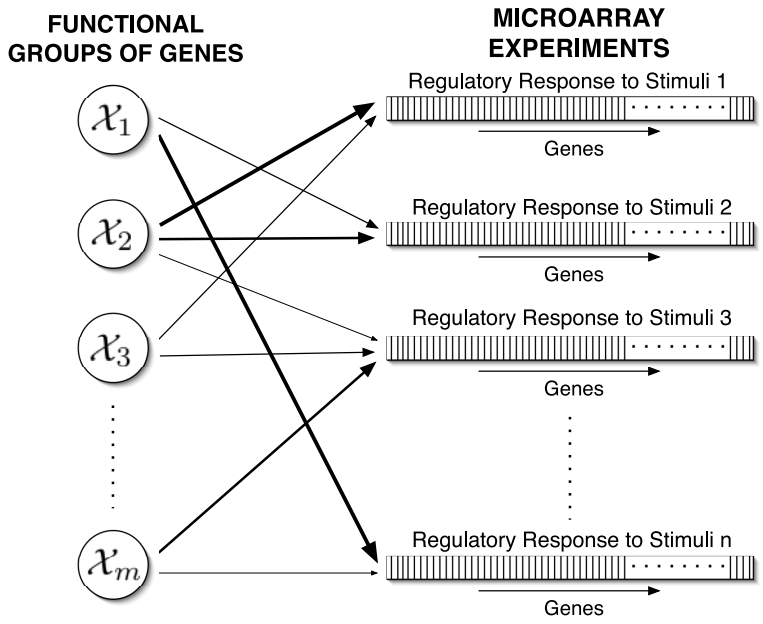

Figure 3.1: This figure illustrates the complex response of diverse functional groups of genes to a set of different experimental stimuli. The thickness of the lines illustrates the severity of the response of that group to a given stimuli. As shown, there may be situations where some groups respond heavily to certain stimuli while others may show little or no response.

This approximation resulted in an initial approach to reconstructing the regulatory characteristics that replaced the nonlinear model shown in (3.2) with a linear representation. We did this by replacing the nonlinear function with a representative value that indicated the level of participation of gene i in the kth functional group. This value was denoted as $s_k^{\delta^i}$. Given this approximation, we were able to rewrite the regulatory response of gene i to

stimulus j as

$$x_j^i = \sum_{k=1}^{m} a_{jk} s_k^i \tag{3.3}$$

where we replaced our reference to gene δ^i as just i. Here, we sum over all functional groups $k = 1, \ldots, m$, denoting the groups, k, that are not involved with stimulus j with a coefficient $a_{jk} = 0$. The expression of gene i for all stimuli was written in vector form as

$$\mathbf{x}^i = \mathbf{A}\mathbf{s}^i \tag{3.4}$$

where $\mathbf{A}$ is the matrix of weighting coefficients and $\mathbf{s}^i$ is the regulatory response of gene i for each functional group. The expression levels $\mathbf{x}^i$ for gene i in response to each stimuli are obtained from the microarray data. We regarded this as the known mixture data. From this, we wanted to blindly estimate 1) $\mathbf{s}^i$ which captures the level of participation of gene δ^i in all functional groups and 2) $\mathbf{A}$ which captures the participation of each functional groups given some stimuli. We saw that this problem closely resembled what is known as a blind source separation problem. As with most blind source separation problems, specific characteristics of the underlying functional groups must be identified in order to uniquely estimate both $\mathbf{A}$ and $\mathbf{s}^i$. We inferred some assumptions on the characteristics of the functional groups that comprise the elements of $\mathbf{s}^i$. These assumptions, which were based on biological heuristics, were used as the first step in the analysis. We first assumed a level of functional diversity across each functional group. This diversity allowed us to infer some level of independence across these groups. Secondly, we assumed that the regulatory activity of each functional group across all genes in the genome reflected functionally similar genes. This would in turn lead to regulatory distributions that are structured and that do not resemble white noise. Later, we will show how these assumptions were used to uniquely perform a type of blind source separation. The source separation approach was applied to data obtained from the bacteria *Bacillus subtilis*. The source separation approach, along with the underlying assumptions, were used as the first step in estimating the characteristics of the mechanisms that drive gene expression. We established the plausibility of the approach and assumptions by identifying whether the expression characteristics and functional relationships of the estimated functional groups were in fact biologically relevant. The validity of the results did not necessarily prove that the assumptions are true for all data but it did justify further exploration of blind source separation techniques for analyzing biological data.

3.3 Independent Component Analysis - Blind Estimation of Hidden Characteristics

This section outlines the theory, assumptions, and ambiguities associated with independent component analysis. Independent component analysis (ICA) is a statistical signal processing tool that belongs to a class of blind source separation approaches. These separation approaches are used to reveal hidden structural information from sets of random variable data, measurements, or signals [91] [92]. Assume these sets of random variable data, measurements, or signals were produced from a mixture of unknown components with specific characteristics. ICA is in a category of projection approaches that exploits these characteristics in order to blindly identify 1) the unknown components without a-priori knowledge of how these components are mixed and 2) the transformation that characterizes how these components are mixed to form the observed data. Other projection approaches that fall into this category include principal component analysis (PCA) [93], singular value decomposition (SVD) [94], and factor analysis (FA) [95]. ICA is seen as a more general approach and does not restrict its focus to uncorrelated or Gaussian characteristics as does PCA and FA, respectively. ICA assumes the data is a result of a mixture of statistically independent components and makes very few assumptions about the general form of the statistical distributions. This often makes ICA more appropriate for identifying hidden characteristics in data obtained from naturally occurring sources such as mixed speech data, neuro-imaging data from electroencephalograms (EEG) [96] and magnetoencephalograms (MEG) [97], and financial data [98] [99]. Other applications of ICA include dimension reduction [100], feature extraction [101], and face recognition [102].

One of the common applications of ICA is to estimate the solution to blind source separation problems. Blind source separation problems attempt to separate source signals that have been mixed through some unknown mixing process. A conceptual example of a simple blind source separation problem is shown in Figure 3.2, which is known as the *cocktail-party problem*. This example is often used in the literature to highlight the practicality of this approach [91] [92] and is recreated here for completeness. Consider the scenario where two people are speaking in a room. The combined speech is recorded by two microphones situated somewhere in the room, as shown in Figure 3.2. The mixed signals are considered known while the exact mixing process along with the original sources signals

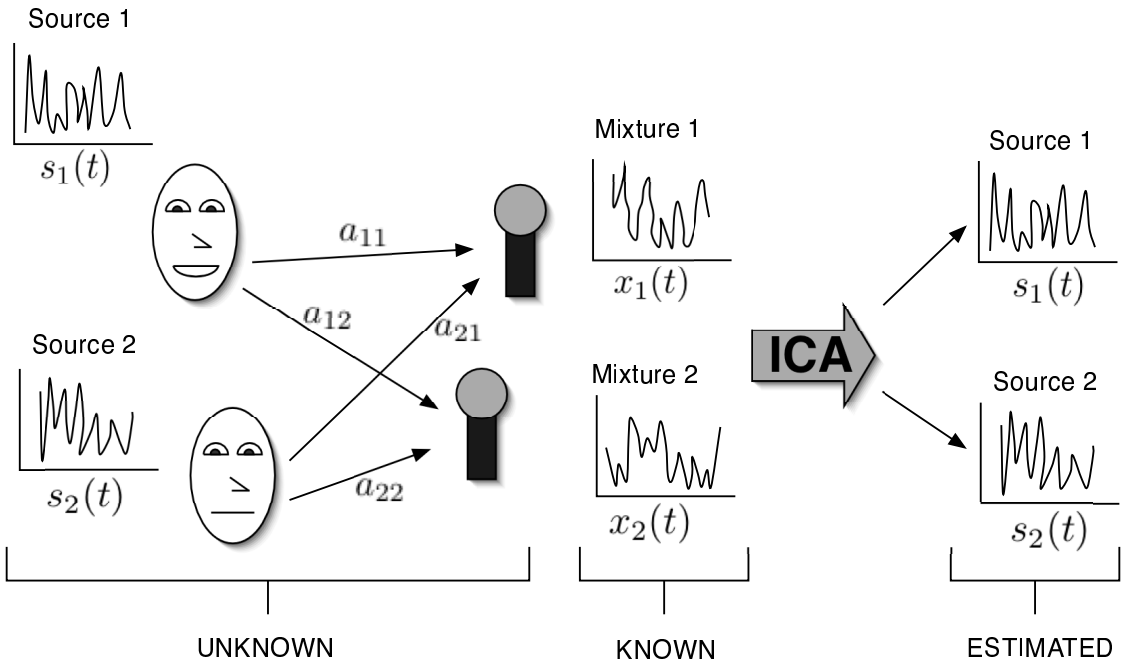

Figure 3.2: This figure illustrates a common blind source separation problem known as the *cocktail-party problem*. This concept was adapted from [91] and [92]

are considered unknown. For simplicity, this mixture is modeled as a linear process where the observed recorded signals are expressed as a linear combination of the original source signals.

$$\begin{aligned} x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) \\ x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t) \end{aligned} \tag{3.5}$$

The weights a_{ij}, $i, j = 1, 2$, represent the contributions of each source signal to the overall mixture. This linear model for the mixed signals may be somewhat of an oversimplification but is a reasonable first approximation for many scenarios [91]. It also provides a good base upon which to outline the mechanics of the approach. Both the coefficients, a_{ij} and the source signal $s_i(t)$, $i, j = 1, 2$, are unknown. The mixture signals $x_i(t)$, $i = 1, 2$ represent the only known information in this process. At first glance, the separation of $x_i(t)$ into a sum of weighted source signals appears ambiguous unless severe restrictions are imposed on both the sources and mixing coefficients. ICA attempts to perform this separation by imposing only a few assumptions about the sources, with the main assumption being that the signals s_i were produced by two completely different sources that are statistically independent of each other. ICA uses these characteristics to find an unambiguous system of weighted independent source signals for a given set of mixture signals, $x_i(t)$. Below, we further elaborate on the general process for accomplishing this separation. A more detailed derivation of ICA can be found in [91] and [92].

Consider a data set $\mathcal{D}$ which contains N n-dimensional data points with each point denoted as $\mathbf{x}^i$, $i = 1, \ldots, N$. We drop the reference to the independent variable t used in the cocktail party problem to generalize the phenomenon upon which the data points are observed. Without loss of generality, we can assume that the collection of points in this set have zero mean. If not, the data points in $\mathcal{D}$ are easily transformed into a zero mean data set using the following transformation

$$\tilde{\mathbf{x}}^i = \mathbf{x}^i - E\{\mathbf{x}\} \tag{3.6}$$

where $E\{\cdot\}$ is the expected value or mean of the data points. In practice, this is normally approximated by the sample mean

$$E\{\mathbf{x}\} \approx \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^i.$$

The assumption is that each data point $\mathbf{x}^i$ is a result of a linear mixture of statistically independent sources as stated previously in the cocktail-party problem. The idea behind ICA is to identify a linear transformation that is able to, once applied to this mixture data, separate each data point $\mathbf{x}^i$ into m statistically independent sources. Mathematically, this results in identifying the set of m basis vectors, $m \leq n$, that, once applied to $\mathbf{x}^i$ produces a vector $\mathbf{s}^i$ whose m elements are statistically independent. The kth component of $\mathbf{s}^i$ is written as

$$s_k^i = \mathbf{w}_k^T \mathbf{x}^i, \quad k = 1, \ldots, m \ , \tag{3.7}$$

where s_k^i is the kth independent source and $\mathbf{w}_k$, $k = 1, \ldots, m$, are the set of basis vectors that span $\mathcal{R}^{m \times 1}$ that are capable of separating the mixtures. This is rewritten in vector notation as

$$\mathbf{s}^i = \mathbf{W}\mathbf{x}^i. \tag{3.8}$$

Here, $\mathbf{w}_k^T$ represents the kth row of $\mathbf{W}$. For simplicity, we assume that $m = n$. In other words, we are interested in the same number of sources as there are dimensions in the data points. There are situations, however, where this assumption can be relaxed [91]. Now, given that $\mathbf{w}_k$, $k = 1, \ldots, m$ forms a basis in $\mathcal{R}^{m \times 1}$, we can rewrite (3.8) as

$$\mathbf{x}^i = \mathbf{A}\mathbf{s}^i \tag{3.9}$$

where for general m, the relationship between $\mathbf{A}$ and $\mathbf{W}$ is given as $\mathbf{A}^T = \left(\mathbf{W}^T\right)^\dagger$. The symbol † indicates the Moore-Penrose inverse or pseudo-inverse of $\mathbf{W}^T$. The pseudo-inverse of some general matrix $\mathbf{Y}$ is defined as

$$\mathbf{Y}^\dagger = \left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\mathbf{Y}^T.$$

Equation (3.9) is also written as

$$\mathbf{x}^i = \sum_{k=1}^{m} \mathbf{a}_k s_k^i \tag{3.10}$$

where $\mathbf{a}_k$ represents the kth column of $\mathbf{A}$. Equations (3.9) and (3.10) represent the model that describes how the observed data is generated from a mixture of independent sources.

The question is how do we obtain all variables in this model from just the data points $\mathbf{x}^i$. We further elaborate on the assumptions needed to uniquely identify these variables using ICA in the following section.

3.3.1 Assumptions of ICA

The main assumption of the ICA algorithm pertains to the pairwise statistical relationship between the underlying sources. The assumption is made that each component is in fact statistically independent of all other components. This assumption implies that each source represents independent entities and provides no statistical information about the other sources. Some other projection approaches assume weaker statistical relationship between the components. The projection approach known as PCA [93] separates the mixed data into components that are statistically uncorrelated. Statistically uncorrelated is a limited statistical measure, only relating the second order statistics of the components. Statistical independence, however, is a stronger statistical measure that relates the higher order statistics between two components. In fact, if two sources are statistically independent then they are also statistically uncorrelated. The reverse however is not true. Hence, statistically independent implies statistically uncorrelated but statistically uncorrelated does not imply statistically independent in the general case. Statistical independence has been shown to be a more biologically relevant criteria than statistically uncorrelated when attempting to identify relevant regulatory components from real microarray data [103].

ICA makes no explicit assumptions about the distributions of each component as is the case for Factor Analysis. It does, however, require that the components have non-Gaussian distributions. As stated previously, statistical independence incorporates higher order statistics in order to effectively identify the sources of interest. These higher order statistics are zero for Gaussian distributions. This effectively makes it impossible for ICA to unambiguously identify components with Gaussian distributions. The assumptions made here are general and do not significantly reduce the range of practical problems upon which ICA can be applied.

3.3.2 Ambiguities of ICA

Specific ambiguities arise from the ICA model shown in (3.9) and (3.10). These ambiguities are described below:

1. ICA cannot determine the energies of the independent sources

ICA assumes that both the mixing matrix, $\mathbf{A}$, and the source vector, $\mathbf{s}^i$, are unknown. Thus, any scalar multiplier in one of the sources could always be canceled by multiplying the corresponding column of $\mathbf{A}$ by the reciprocal of that scalar. This is illustrated through a simple modification of (3.10).

$$\mathbf{x}^i = \sum_{k=1}^{m} \left(\frac{1}{\alpha_k} \mathbf{a}_k \right) \left(\alpha_k s_k^i \right) \tag{3.11}$$

This ambiguity in the analysis is eliminated by ensuring that each component have unit variance. In other words

$$\mathrm{Var}\{s_k\} = E\{(s_k)^2\} \approx \frac{1}{N-1} \sum_{i=1}^{N} \left(s_k^i\right)^2 = 1, \qquad k = 1, \ldots, m.$$

The final equality results from an approximation commonly used for the variance. This still leaves an ambiguity of the sign which makes it impossible for ICA to disctinguish between s_j^i and $-s_j^i$. This will come into play when attempting to distinguish between repressed and induced expression activities of genes in the genome.

2. ICA cannot uniquely determine the order of the independent sources

This is again a result of that fact that ICA assumes that both the mixing matrix and the source vector are unknown. Thus, any change in the order of the sources in $\mathbf{s}^i$ can again be accommodated for by a similar change in the order of the columns of $\mathbf{A}$. We can formally state this by introducing a permutation matrix $\mathbf{P}$. A simple modification of (3.9) results in

$$\mathbf{x}^i = \mathbf{A}\mathbf{P}^{-1}\mathbf{P}\mathbf{s}^i$$

where $\mathbf{P}\mathbf{s}^i$ are the original independent sources reordered and $\mathbf{A}\mathbf{P}^{-1}$ is just another mixing matrix for ICA to find. Methods can be implemented that are able to order the sources in terms of importance where the measure of importance is determined by some cost measure.

3.3.3 How Does ICA Work?

As stated previously, ICA attempts to find the collection of basis vectors that, once applied to the observed data points, produce a set of statistically independent components. Each component is obtained by projecting a data point, $\mathbf{x}^i$, onto one of the basis vectors as shown in (3.7). An estimate of the kth component is written as

$$\hat{s}_k^i = \hat{\mathbf{w}}_k^T \mathbf{x}^i \tag{3.12}$$

where $\hat{\mathbf{w}}_k$ is an estimate of the true basis vector used to produce the independent sources. This can be rewritten as

$$\hat{s}_k^i = \hat{\mathbf{w}}_k^T \mathbf{A}\mathbf{s}^i = \mathbf{q}_k^T \mathbf{s}^i \tag{3.13}$$

with the first equality resulting from a substitution of (3.9) into the equation. Upon further inspection, it is clear that the estimate $\hat{s}_k^i$ is one of the original sources of $\mathbf{s}^i$ if and only if $\mathbf{q}_j$ has only one nonzero component. If $\hat{s}_k^i$ is not a true estimate, then it is clear from (3.13) that it will be a linear combination of the original sources. The Central Limit Theorem states that the distribution of a sum of independent random variables tends toward a Gaussian distribution under certain conditions [104]. Hence, it is thought that the sum of two or more independent sources will tend to have a distribution that is more Gaussian than any of the original sources [91]. ICA capitalizes on these concepts and establishes a reasonable estimate of s_k^i by searching for the $\hat{\mathbf{w}}_k$ that produces an estimate $\hat{s}_k^i$ that is as far away from Gaussian as possible. This means minimizing the Gaussianity (or maximizing the non-Gaussianity) of the distribution of $\hat{\mathbf{w}}_k^T \mathbf{x}^i$ over i. Such a vector would correspond to $\mathbf{q}_j^T = \hat{\mathbf{w}}_j^T \mathbf{A}$ having only one nonzero component. The result is an estimate $\hat{s}_j^i$ that corresponds to one of the independent sources and not a sum of the independent sources. Although somewhat heuristic, this concept appears to be quite robust and works for many practical applications [91].

Estimation of an appropriate basis vector $\hat{\mathbf{w}}_k$ that maximizes the non-Gaussianity of the estimate $\hat{s}_k^i$ requires a suitable cost function that serves as an adequate measure of the Gaussianity of the distribution of a random variable. Ideally, the optimization landscape for non-Gaussianity in the m-dimensional space of vectors $\mathbf{w}_k$ has $2m$ local maxima, two for each independent component corresponding to s_k^i and $-s_k^i$ [91]. Several reasonable measures of non-Gaussianity exist for performing this estimation. Some of these measures, however, are

not suitable for practical implementation. The kurtosis, or the fourth-order cummulant of a random variable, is a classical theoretical measure for distinguishing Gaussian distributions from other distributions. The kurtosis of $\hat{s}_k^i$ is defined as

$$\text{kurt}(\hat{s}_k^i) = E\{(\hat{s}_k^i)^4\} - 3(E\{(\hat{s}_k^i)^2\})^2. \tag{3.14}$$

In general, the kurtosis of most random variables is nonzero while the kurtosis of a Gaussian random variable is zero. Thus maximizing the kurtosis, or absolute value of the kurtosis, in a sense maximizes the non-Gaussianity of the estimate [91]. Kurtosis, however, is not very robust when estimated from sampled data due to its high sensitivity to outliers. Another measure of non-Gaussianity used in the literature is negentropy. It incorporates the concept from information theory known as entropy, which is a measure of the uncertainty associated with a random variable. The entropy of a random variable estimate $\hat{s}_k^i$ with associated probability density function $p_s(\gamma)$ is defined as

$$H(s_k^i) = -\int p_s(\gamma) \log p_s(\gamma) d\gamma. \tag{3.15}$$

A fundamental result from information theory is that a Gaussian random variable has the largest entropy among all other random variables of equal variance [91]. A measure of non-Gaussianity that is zero for Gaussian random variables and nonzero otherwise is defined as

$$J(\hat{s}_k^i) = H(\nu) - H(\hat{s}_k^i) \tag{3.16}$$

where ν is a Gaussian random variable with the same variance as $\hat{s}_k^i$. This measure is known as the negentropy. The negentropy is considered to be the optimal estimator of non-Gaussianity in the statistical sense [91]. Calculation of the negentropy, however, requires estimates of potentially non-parametric probability density functions, which makes the definition in (3.16) computationally difficult to realize. The next section presents simpler approximations of negentropy and incorporates it into an algorithm that is used for estimating the independent sources. Other robust measures of Gaussianity include maximum likelihood and mutual information [91].

3.3.4 Computing Independent Components

The ability to implement computational algorithms that effectively and efficiently compute the independent sources are as important as the theoretical foundation described in the previous section. An approach developed by Hyvärinen [91] [105] for estimating the individual basis vectors used to obtain the independent sources is summarized here. This Newton based batch mode algorithm uses an accurate and easy to calculate approximation of the negentropy to perform the estimation. This algorithm has been packaged into the FastICA MATLAB toolbox which is freely available on the world-wide-web [106].

The definition of negentropy shown in (3.16) requires estimates of potentially non-parametric probability density functions, thus making it computationally difficult to realize. A computationally efficient approximation of the negentropy is given as

$$J(\hat{s}_k^i) \propto \left[E\{G(\hat{s}_k^i)\} - E\{G(\nu)\}\right]^2. \tag{3.17}$$

Here, $G(\cdot)$ is any non-quadratic function and ν is a zero mean Gaussian random variable with unit variance. Further inspection of (3.17) reveals that choosing $G(s_k^i) = (s_k^i)^4$ yields a cummulant-based calculation similar to the kurtosis. The point here, however, is to generate a measure or cost function that is more robust than the kurtosis and yields an accurate approximation of the negentropy. This is accomplished by choosing a function $G(\cdot)$ that does not grow too fast. This style of function aids in the algorithms ability to not be affected by noisy outliers in the data. Hyvarinen listed the following as useful choices of $G(\cdot)$

$$G_1(u) = \frac{1}{a_1} \log\cosh a_1 u \tag{3.18}$$

$$G_2(u) = -e^{-u^2/2} \tag{3.19}$$

where a_1 is some suitable constant in the range $1 \leq a_1 \leq 2$. The nonlinearity in (3.18) (or (3.19)) in combination with the cost function defined in (3.17) are used to generate the following optimization problem

$$\max_{\mathbf{w}_k} \sum_{i=1}^{N} J(\mathbf{w}_k^T \mathbf{x}^i) \tag{3.20}$$

subject to a normalization constraint

$$\frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{w}_k^T\mathbf{x}^i\right)^2 = 1. \tag{3.21}$$

This optimization is used to estimate one of the m independent sources. Recall that statistically independence implies statistically uncorrelated. Hence, the procedure for solving several independent components involves a scheme that ensures that all subsequent vectors $\mathbf{w}_j$ produce sources that are not correlated with any of the sources that were previously estimated. Several methods for accomplishing this are described in [105].

The FastICA package [105] [106] implements a fast batch-mode algorithm based on a Newton-like derivation for solving this optimization problem. It is called a batch algorithm because a large amount of data (in some cases all of the data) are used in the calculations for a single iteration. For situations where the data perfectly matches the ICA model and there are an adequate number of data points that sufficiently represent the distribution of the data, the optimization landscape for non-Gaussianity in the m-dimensional space of vector $\mathbf{w}_k$ has $2m$ local maxima. In practice however, real scenarios result in data of finite sample size that may not exactly follow the ICA model. This may result in situations where more than $2m$ local maxima exists. This affects the reproducibility of the estimated components resulting in a algorithm that estimates different components across multiple runs. In such situations the reliability of the estimated sources can be evaluated by assessing the variability of the results across the multiple runs. Himberg and Hyvärinen [107] and Himberg et al. [108] present the software package, Icasso, for testing the reliability of the estimates resulting from the FastICA algorithm.

3.3.5 Icasso MATLAB Software Package

Uncertainty and variability in the estimated independent components may arise from two specific sources. The first may occur from the statistical variability of the mixed data used to identify the independent sources. This may occur as a result of a small sample size, non-stationary characteristics in the process used to create the data, or uncertainty associated with the observation tools used to capture the mixed data. Another source of uncertainty is the algorithmic variability associated with the convergence of search algorithms used to estimate the components. Cost function irregularities caused by an inexact

fit of the data to the ICA model may cause local search algorithms, such as those used in the FastICA algorithm described above, to sometimes converge to inaccurate local minima in the search space if the independent components are not well defined. Both of these sources of uncertainty in the estimates can be addressed by running the algorithm multiple times and assessing the variability of the resulting estimates. The statistical uncertainty associated with the data is often difficult to ascertain, particularly when the actual distribution of the data is not parameterizable. Bootstrapping is a well-known computational method for computing the statistical reliability of estimates when no mathematical formula of the distribution is known [109]. It is a method that randomly resamples the set of data points in $\mathcal{D}$ to produce another data set of N data points $\hat{\mathbf{x}}^i$, $i = 1, \ldots, N$ with the same or similar statistical properties. The estimates of the independent components are then obtained using this new set of data points. This allows one to assess how the resampling of the observation process affects the estimates. The uncertainty due to irregularities in the search space can be addressed by using randomized initial conditions for the optimization. It is then possible to assess the variability of the converged estimates of all the runs based on randomizing the initial conditions.

Icasso is a free software package [110] that combines the two randomization methods of data bootstrapping and randomized initial conditions to assess the reliability of the estimated independent components. It is a toolbox written for the MATLAB platform and integrates directly into the FastICA toolbox. The Icasso toolbox also provides an assortment of tools for clustering and visualizing the independent components obtained from the multiple runs. The general steps for the algorithm are listed in [108] and are restated below:

1. Parameters for the FastICA estimation algorithm are selected. This includes the nonlinearity, nonlinearity parameters, number of independent components, etc.

2. The FastICA algorithm is ran M times with each run using bootstrapped data and/or randomized initial conditions for the optimization

3. The resulting estimated independent components from each run are clustered according to their mutual similarities. The clustering method is freely selected with the default clustering approach being agglomerative clustering with average-linkage criterion.

4. The clusters of independent components are visualized on 2-D plots using nonlinear

projection approaches. Each cluster is bounded by a convex hull. Compact, tight, and isolated clusters correspond to reliable independent component estimates that were repeatedly identified despite the randomization.

5. The best estimates from each cluster, particularly the reliable clusters, are identified and used for further analysis.

These steps of the Icasso algorithm are described in more detail below.

Generating Estimates

Steps 1 and 2 focus on generating sets of independent component estimates from M runs of the FastICA algorithm. Each run uses bootstrapped data and/or random initial conditions for the optimization in order to try to assess how statistical and algorithmic variability affects the estimates. The result is a set of M estimates of the demixing matrix denoted as $\mathbf{W}_i$, $i = 1, \ldots, M$. This set of M demixing matrices are collected and placed in a block matrix $\mathbf{W} = \left[\mathbf{W}_1^T \ \mathbf{W}_2^T \cdots \mathbf{W}_M^T\right]^T$. The matrix $\mathbf{W}$ is used to generate an estimate of the correlation matrix $\mathbf{R} = \mathbf{W}\mathbf{C}\mathbf{W}^T$, which contains pairwise correlations of the estimated independent components. Here, $\mathbf{C}$ is the correlation matrix of the observed data. The correlation coefficients r_{ij} for the ith and j^{th} components provide an adequate measure for comparing the resulting components and can be used to compute a distance metric for clustering groups of similar estimated independent components. This metric is given as

$$d_{ij} = 1 - |r_{ij}| \,. \tag{3.22}$$

Note that this distance metric is maximum (equal to 1) when the ith component and the jth component are uncorrelated and the metric is minimum (equal to 0) when the ith component and the jth component are fully correlated.

Clustering Estimates

Steps 3 focuses on partitioning the resulting estimates of the independent components into m clusters using some basic clustering algorithms [108]. Any clustering algorithm can be integrated into the process but the toolbox has readily available implementations of agglomerative hierarchical clustering. The clustering is performed using the distance metric

in (3.22). The toolbox employs three basic agglomeration strategies, which include single-linkage or minimum distance strategy, complete-linkage or maximum distance strategy, and average linkage strategy. A heuristic measure call the cluster quality index is also used to quantify the compactness of a given cluster of independent components [108]. The cluster quality index for the mth cluster is defined as

$$I_q(C_m) = \frac{1}{|C_m|^2} \sum_{i,j \in C_m} |r_{ij}| - \frac{1}{|C_m||C_{-m}|} \sum_{i \in C_m} \sum_{j \in C_{-m}} |r_{ij}|, \tag{3.23}$$

where C_{-m} is the set of indices that do not belong to the mth cluster and $|C_m|$ is the size of the mth cluster. Ideally, $I_q(C_m)$ is equal to one if it contains components that are identical which are fully uncorrelated with all other estimated components, corresponding to an very tight cluster. The value of $I_q(C_m)$ decreases as the cluster becomes less compact and less isolated.

2D Visualization of Estimates

The goal of this 2D visualization of the estimated components is to map the $M \times m$ estimates of the independent components onto a two dimensional plane. This mapping attempts to preserve the pairwise distances in (3.22) between all estimated components. Recall the relationship between the distance metric and the pairwise correlation between the individual components. Thus, there should be m groups of M independent components if the source separation algorithm work ideally on each of the M runs, m being the total number of estimated independent component for each run. According to the metric and the assumptions made about the pairwise independence of each of the components, the M estimates in each group would be completely correlated with each other and uncorrelated with the estimates in the other groups. Thus, if the projection method adequately preserved the distances for all points, the ideal results would be m adequately distanced dots on the 2D plane with each dot corresponding to M fully correlated estimates. This would not be the case for practical problems of course. Inconsistency due to statistical and algorithmic variability may cause the m clusters to expand on the 2D plane. The spread of these clusters determine the its reliability and reproducibility. Reliable clusters translate to tight groups of estimates that are highly correlated with each other and minimally correlated with the estimates in the other clusters. Unreliable clusters result in large spacious clusters that may even appear to

cross cluster boundaries at a result of the dimension reduction. Projection techniques available in this toolbox include linear metric multidimensional scaling, Sammon's projection, and curvilinear component analysis.

Calculating the Best Estimates From Each Cluster

The centrotype is identified as the representative independent component for each cluster. The centrotype is defined as the point in the cluster that has the maximum sum of similarities (as measured by the correlation coefficients) to the other points in the cluster. This provides a more meaningful independent component than would be obtained by any signal run. These representative components are further analyzed according to the scope of the application.

3.4 Identifying Hidden Regulatory Characteristics using ICA

This section describes the primary contribution of this work: the application of ICA to microarray data for the purpose of blindly identifying hidden regulatory characteristics and functional relationships across groups of genes. In our work, each data point from the mixed data, $\mathbf{x}^i$, represented the regulatory response of the ith gene for a given set of experiments. Upon separation, the corresponding point $\mathbf{s}^i$ represented the contribution of the ith gene to all of the estimated independent components. The initial conjecture that was made assumed that the estimated independent components represented functionally independent groups which respond to the sets of stimuli used to obtain the microarray data. The validity of this conjecture was tested by examining whether or not the results from the ICA analysis produce components that were in fact biologically relevant. We first obtained the estimated components labeled as the centrotype by the Icasso software as the meaningful representative for each of the m clusters. With that, we also obtained the corresponding mixing matrix for that set of m components. Given our initial conjecture, one would expect that each functional group contains subsets of important genes that are particularly active for that specific function. The level of activity for each gene was indicated by s_k^i, which shows the participation of gene δ^i in the kth functional group. Largely positive or negative values of s_k^i indicated a gene that was particularly responsive to that functional group, corresponding to a highly induced or highly inhibited response. The ambiguity of the sign

of s_k^i in ICA estimation, however, made it virtually impossible to differentiate between induced and repressed members of the functional group. We therefore looked at the absolute value of s_k^i to identify the important sets of genes that were highly responsive in any given functional group. These sets of important genes were identified using a simple thresholding scheme

$$\mathcal{X}_k = \{\text{genes } \delta^i \mid |s_k^i| > \beta^k\}, \qquad i = 0, \ldots, N \tag{3.24}$$

where β^k is a hard threshold for the kth component. This threshold was set statistically, identifying all gene that were greater than n standard deviations away from the mean of that component. These genes represented what we termed as the *differentially expressed* genes of that functional group. The overlap of differentially expressed genes in multiple functional groups allowed us to infer on possible functional relationships across genes with dissimilar expression activity.

The set of differentially expressed genes for each component provided the set of genes that were the most responsive to that particular functional group. This, however, only gave a partial view of the activation of the regulatory mechanisms within the organism. The mixing matrix $\mathbf{A}$ associated with these components also provided insight in how each functional group of genes participated in the overall response. We restate again that the source separation for ICA yields

$$\mathbf{X} = \mathbf{AS}. \tag{3.25}$$

When microarray data is used, the $(i, j)^{th}$ component of the data matrix $\mathbf{X}$ represents the regulatory response of the jth gene on the ith experiment. Given the matrix separation in (3.25), we know from matrix multiplication that the $(i, j)^{th}$ value is obtained from the inner product of the ith row of $\mathbf{A}$ with the jth column of $\mathbf{S}$. Thus, each element in a given row of $\mathbf{A}$ represented the contribution of a single component to that experiment. Hence, each of the columns of $\mathbf{A}$ represented the participation or expression activity of each functional group across all experiments where the profile for the jth functional group corresponded to the jth column of the mixing matrix. Conversely, we also see the rows of $\mathbf{A}$ capture the participation or expression activity of each functional group for the experiment corresponding to that row. In other words, the ith row of the mixing matrix contained the activity of each functional group for the ith experiment. Again, our ability to only estimate the mixing

matrix within a multiplicative factor of ± 1 limits our ability to distinguish negative participation from positive participation. We therefore used the absolute value of the activities in **A** and denoted this as the *Absolute Expression Activity*. These absolute expression activity values, along with the set of differentially expressed genes for each functional group, gave a reasonable assessment of how the different groups of genes are regulated under varying stimuli. The absolute regulatory activity and the relationships across functional groups seen in the results were compared with analyses that have been performed by other researchers on the same data. Components that were consistent with the characteristics seen in the literature and that were characterized as reliable in the presence of statistical and algorithmic variation are seen as biological relevant results. The presence of biologically relevant components signify the relevance of using these and other types of source separation techniques for inferring functional connections across groups of genes.

3.5 Simulations and Results

The analysis presented in this section identifies whether independent components obtained from applying ICA to microarray data do in fact represent functional groups of genes whose regulatory activities and functional relationships are consistent with characteristics seen in the literature. This analysis used data obtained from exposing the gram positive bacteria *B. subtilis* to adverse environmental conditions. These conditions included heat shock, hydrogen peroxide exposure, tert-butyl peroxide exposure, and tert-butyl alcohol exposure. The FastICA package was used to implement the ICA algorithm, which attempts to separate the convoluted response of major functional groups of genes that were activated by such harsh conditions. We simplified the approach by selecting the algorithm to estimate m independent components, where m is equal to the number of experiments used to acquire the data. The Icasso software package was used as a tool to identify how the components vary due to statistical and algorithmic variability. This was done by running the FastICA algorithm 50 times. The resulting estimates obtained from these 50 independent runs were classified into m clusters. The centrotypes for each cluster were extracted and used as the meaningful component or functional group for that cluster. The reliability of each component was examined visually using a projection utility in the Icasso software package that projects the clusters of estimated components onto two dimensions. This

visual assessment was used in combination with the cluster quality index, I_q, to identify reliable, unreliable, and questionable clusters. This cluster assessment was used to aid in the identification of biologically relevant clusters. The centrotype or functional group from each cluster was then used to perform the following analysis: 1) identify the most important genes of each functional group by identifying all differentially expressed genes given a specific threshold, 2) infer on the relationships across functional groups by looking at the overlaps in differentially expressed genes, and 3) examine the absolute regulatory activity of important functional groups by looking at the mixing matrix associated with the selected components. We tested the significance of the results obtained in this section by comparing our results with established regulatory characteristics and functional relationships of groups of gene known to respond to harsh environmental stimuli. Particularly, we looked at the characteristics of the following groups of genes: genes belonging to the general stress response regulon known as σ^B, a group of genes known as the Class III heat response genes, three operons responsible for arginine biosynthesis and transport, uncharacterized genes known to respond specifically to heat shock, and genes belonging to the PerR regulon. Our results attempt to validate the appropriateness of using the ICA model and ICA assumptions to estimate hidden and unknown regulatory characteristics from microarray data.

3.5.1 Data

The microarray data used for this analysis capture the regulatory response of *B. subtilis* to adverse changes in environmental conditions. Originally generated by Helman et al. [36] [111], the data are now available online through the Stanford Microarray Database [112]. The data reflects the regulatory response of *B. subtilis* to 1) heat shock, 2) hydrogen peroxide (H_2O_2), 3) tert-butyl peroxide (t-buOOH), and 4) tert-butyl alcohol (t-buOH). The heat shock experiment involved transferring bacteria grown at 37° C to a 48° C environment. Time samples for this experiment were acquired at 0, 3, 10, and 20 minutes. The peroxide stress experiment resulted from adding hydrogen peroxide at concentrations of $8\mu M$ and $58\mu M$ to bacteria grown at 37° C. The other stress experiments resulted from adding tert-butyl peroxide and tert-butyl alcohol at concentrations of $100\mu M$ for each. Time samples for RNA isolation for these experiments were taken at 0, 3, 10 and 20 minutes, although some were also recorded at 40 minutes. This yielded a total of 30 experiments where each experiment contained RNA probes for 4116 open reading frames or genes in *B. subtilis*.

The microarray data used here were previously analyzed by Helmann et al. [36] [111] who used a variety of analysis techniques which included quantitative reverse transcriptase (RT)-PCR, statistical analysis, and hierarchical clustering. Refer to [36] and [111] for details on experiment materials, RNA extraction, RT-PCR analysis, and microarray data production and normalization.

3.5.2 Analysis

The microarray data acquired from the Stanford Microarray Database resulted in a 30×4116 data matrix where the rows corresponded to the experiments and the columns corresponded to the genes in the genome. This matrix was denoted by the matrix $\mathbf{X}$ where the ith column of $\mathbf{X}$ corresponds to the ith data point $\mathbf{x}^i$. The collection of data points $\mathbf{x}^i$, $i = 1, \ldots, 4116$ were used as the mixture data to be separated. The FastICA algorithm was used to estimate the unknown matrices in the linear regulatory model shown in (3.25). Equation (3.18) was used as the nonlinearity for the FastICA algorithm with a nonlinearity parameter of $a_1 = 1$. This value for the nonlinearity parameter was used because it produced a nonlinear function that would not grow too fast and that performed well for similar applications [38]. A symmetric decorrelation approach was used to identify all components simultaneously. The number of actual components hidden within the data was unknown and strictly depended on the functional groups that were activated as a result of the experimental stimuli. For this reason, the algorithm was set to estimate the maximum amount of components given the mixture data used. This resulted in the estimation of 30 independent components, which corresponded to the number of experiments used to obtain the microarray data. Further analysis was used to identify if all of the resulting 30 components were both algorithmically and biologically relevant.

Icasso was used to run the FastICA algorithm 50 times using both bootstrapping and randomized initial conditions to assess the variability of the estimated components. Of the 50 total runs, 40 runs were executed using only randomized initial conditions and 10 runs were executed using both bootstrapped data and randomized initial conditions. This scheme was used so as to not have the inaccuracies associated with the resampling scheme overly affect the assessment of the variation of the estimates. The resulting estimates of the independent components were grouped into 30 clusters using agglomerative clustering with an average linkage strategy. Curvilinear component analysis was used to visualize the

compactness of each cluster by projecting the components onto two dimensions while preserving the pairwise correlation-based distance metric. A visual assessment of the reliability of each cluster was made based on how compact and isolated each cluster was on the two dimensional plane. A quantitative assessment of the reliability of each cluster was made based on the cluster quality index, I_q.

The centrotype was selected as the meaningful representative for each cluster and was used as the operable functional group or FG for further analysis. The differentially expressed genes for each functional group were obtained statistically using a threshold of $\beta^k = 2\sigma_k,\ k = 1, \ldots, 30$, where σ_k was the standard deviation of the collection of gene values for FGk. These differentially expressed genes were used in combination with the profiles acquired from the mixing matrix $\mathbf{A}$ in order to identify functional relationships and expression patterns of groups of genes that were responsive to these stress experiments. We first identified individual genes that belonged to regulons or groups of genes that were known to respond specifically to environmental stress. We identified the top FGs where the majority of these genes from a given gene group were differentially expressed. We assessed the algorithmic reliability of these top FGs qualitatively using the 2-D projection and quantitatively using the cluster quality index in (3.23). We then looked at the differentially expressed genes that each top FG had in common in order to assess any possible functional relationship. This information was contained in a *connectivity matrix* which showed the pairwise percent similarities between top FGs. This allowed us to identify possible diverse functional and expression characteristics in subsets of genes within the same regulon. This also allowed us to infer on possible functional relationships between multiple regulons. The absolute expression activity for a given group of genes was determined by the expression activity of the top FG(s). The compactness of the corresponding clusters along with the comparison of the analysis to characteristics established in the literature allowed us to assess the biological relevance of the major FGs.

3.5.3 Results

These results focused on capturing the underlying regulatory characteristics and functional connections of several groups of genes known in the literature to be highly responsive to stressful conditions. These groups included genes belonging to the general response regulon known as σ^B, a group of genes known as Class III heat response genes, three operons

involved in the biosynthesis and transport of the amino acid arginine, uncharacterized genes known to respond specifically to heat shock, and peroxide inducible members of the PerR regulon. The analysis was able to highlight potential relationships between groups of genes, identify alternative activity within the same regulon, and separate group of genes that were functionally diverse even though they had similar regulatory responses to the experimental stimuli. Overall, our results were used to 1) identify a level of consistency between the analysis and the a-priori knowledge, 2) highlight the biological signification of the ICA model and assumptions, and 3) collectively assess the practicality of this approach in exploratory analysis schemes for identifying unknown regulatory relationships from microarray data.

Consistency of Estimated Components

Figure 3.3 shows the 2-D projection of the clustered independent components obtained from the 50 independent runs of the FastICA algorithm. The estimates from each run are indicated by black points in the figure. The estimates belonging to the same cluster are bound by a convex hull which fits like an elastic band around the set of points belonging to that cluster. The shaded regions inside the bounded area indicate the similarities between the elements of that cluster. Dark shades of red signifies strong similarities and lighter shades signifies weak similarities. The pairwise similarities are replaced with a uniform dark shade of red if the average intra-cluster similarity is above 0.9. Table 3.1 shows the cluster quality index for all 30 clusters. In comparing the visual characteristics of the clusters in Figure 3.3 and the cluster quality indices in Table 3.1, it should be apparent that cluster index values closer to 1.0 corresponds to tight clusters that are spatially isolated from the other components. Examples of this include clusters 1, 4, and 6. Lower cluster quality indices correspond to clusters that are not tight and, in some cases, overlap because they contain estimates that are overly correlated with members of other clusters. We characterize these unreliable or inconsistent clusters. Examples of this include clusters 26, 29, and 30. All unreliable clusters are indicated in Table 3.1 by † and all questionable clusters are indicated by ‡. The results here show that a majority of the estimated components are considered to be reliable. Next, we focused on the analysis of the centrotype of reliable clusters in order to qualify their biological relevance. Note that the clustering approach also addresses the concern pertaining to the ambiguity of the order of estimated components. Since all related components are grouped into the same cluster, the order in which the algorithm estimates

these components is irrelevant.

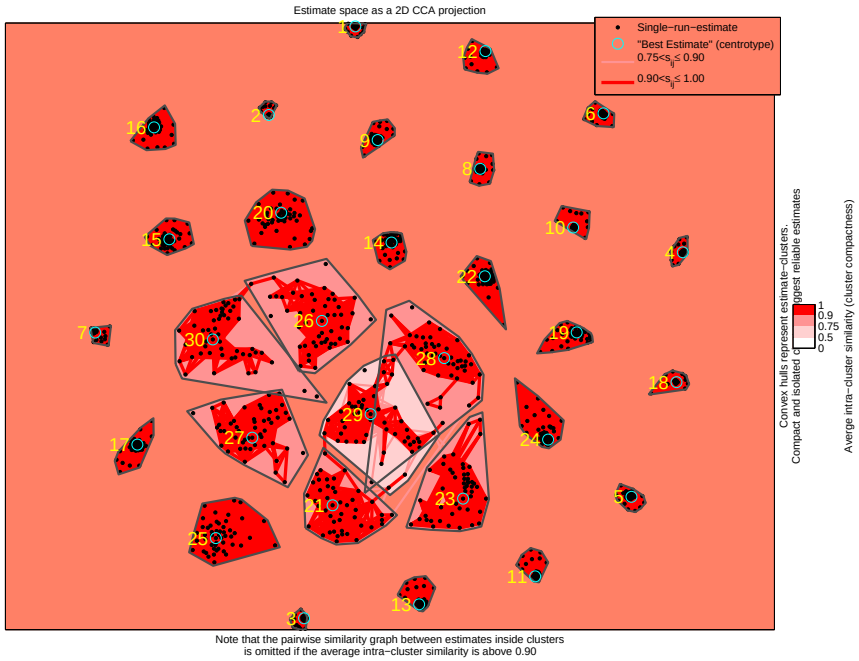

Figure 3.3: This figure shows a 2-D projection of the similarities between the estimated components. The size and tightness of each cluster serves as an indicator of the overall reliability and stability of the cluster.

σ^B Regulon

The σ^B regulon is a class of co-regulated genes made up of well over 100 members [36] [113] [114] [115]. The genes in this regulon comprise a functional group known as general stress response genes whose expression patterns are activated by a variety of stress environments which include heat, ethanol, and salt stress [116]. The size and high sensitivity to stress environments makes this group an ideal candidate for our analysis. We collected a set of 128 genes that were confirmed or suspected members of the σ^B regulon. These genes were identified as being highly responsive to heat and peroxide stress [36] [111]. A list of these genes are shown in Table 3.2. The purpose of our analysis was to use the approach outlined in Section 3.2 to 1) identify the functional group or groups that

Table 3.1: Cluster Quality Index of Clusters of Estimated Components

Cluster Number	**1**	**2**	**3**	**4**	**5**
Cluster Quality Index (I_q)	0.9876	0.978	0.9835	0.9798	0.9806
Cluster Number	**6**	**7**	**8**	**9**	**10**
Cluster Quality Index (I_q)	0.9825	0.9844	0.9738	0.9739	0.9799
Cluster Number	**11**	**12**	**13**	**14**	**15**
Cluster Quality Index (I_q)	0.9759	0.9772	0.9738	0.9682	0.9694
Cluster Number	**16**	**17**	**18**	**19**	**20**
Cluster Quality Index (I_q)	0.9694	0.9648	0.9811	0.9656	0.9336‡
Cluster Number	**21**	**22**	**23**	**24**	**25**
Cluster Quality Index (I_q)	0.82916†	0.9559	0.8285†	0.9534‡	0.932‡
Cluster Number	**26**	**27**	**28**	**29**	**30**
Cluster Quality Index (I_q)	0.7755†	0.8247†	0.7934†	0.675†	0.801†

† Components labeled as inconsistent or unreliable

‡ Components labeled questionable

contained a large number of differentially expressed genes belonging to the σ^B regulon, 2) identify consistent expression characteristics across the entire regulon, and 3) identify any alternative expression characteristics consistent with smaller subsets of genes belonging to this regulon. These results were then compared to the characteristics established in the literature in order to verify some level of consistency between our results and results that have been previously established.

A total of 111 of the 128 genes listed in Table 3.2 were identified as differentially expressed in at least one of the 30 functional groups obtained from the ICA algorithm. Approximately 74% of these genes were differentially expressed in the functional group labeled 5 which we term as FG5. This was almost three times more than any amount of differentially expressed genes found in any other functional group (see Table 3.3). Note that the sum of percentages in Table 3.3 will not equal 100% because there is overlap in the sets of differentially expressed genes for each functional group. The importance of having these genes capable of belonging to multiple functional groups will be highlighted later in the analysis. The results in Table 3.3 show that FG5 is the reasonable candidate to represent the majority of general stress response genes belonging to the σ^B regulon. We inspected the gene expression levels of FG5 by looking at the 5th column of the mixture matrix $\mathbf{A}$. As stated previously, we looked at the absolute value of these entries to assess the absolute expression activity of the functional group. Figure 3.4 shows the absolute expression activity of FG5 for four of the stress experiments at 0, 3, 10, and 20 minutes. This figure shows a very distinctive absolute expression pattern that is consistent across all four of the stress experiments shown. The figure shows a transient-like response, with the absolute expression activity peaking at the 3 minute experiment and decaying rapidly at later experiments. Helmann et al. [36] [111] found this to be a distinct expression characteristic of most, but not all, of the genes belonging to the σ^B regulon in response to heat and peroxide stress. The consistency of these characteristics confirmed that FG5 may in fact be representative of the collection of genes that responded to general stress. The relevance of these results were reinforced by the statistical and algorithmic consistency of FG5 (see Figure 3.3 and Table 3.1). Figure 3.3 shows FG5 to be a tight cluster isolated from the remaining groups of independent components. Table 3.1 also shows that this consistency is reflected by an I_q value of 0.9806 for FG5.

As stated previously, the differentially expressed genes in each of the functional groups were not mutually exclusive. This overlap allowed us to investigate potential rela-

Table 3.2: List of σ^B Genes

bmrU	*bmr*	*bmrR*†	*bofC*	*csbA*	*csbB*	*yfhO*	*yxcC*
ywmG	*csbX*	*ctc*	*dps*	*gsiB*	*gspA*	*katB*	*trxA*
yacH	*yacI*	*yacL*	*ycdF*	*yycE*	*ycdG*	*ydaP*	*yfkM*
yflT	*yhdF*	*yhdG*	*yhdN*	*yjbC*	*yjbD*†	*ykzA*	*yocK*
ysdB†	*ytkL*†	*ytxG*	*ytxH*	*ytxJ*	*yvyD*†	*yxkO*	*ydaD*
katX	*opuE*	*ydaE*	*ydaF*	*ydaG*	*aldY*	*lctE*	*ycnH*
ydaT	*ydaS*	*ydhK*	*yfhK*	*yfhL*	*yfhM*	*yflA*	*yjgB*
yjgA	*ykgA*	*ykgB*†	*yoxC*	*yoxB*	*yoaA*	*yqhA*	*yqxL*
yvrE	*ydbP*	*yoxA*	*ypuB*	*ypuC*	*yqhQ*†	*yqhP*†	*yqiS*†
yrvD†	*yaaI*	*yaaH*	*ybyB*	*ycbP*	*ydaB*	*ydaC*	*ydaJ*
ydaK	*ydaL*	*ydaM*	*ydaN*	*ydbD*	*ydgC*†	*yerD*	*yfhF*
yfhD	*yfkE*	*yfkD*	*yfkJ*	*yfkH*	*yflS*	*ygxB*	*yhaS*
yhaT	*yhaU*	*yhcM*	*yheK*	*yhxD*	*yjcE*	*yjgC*	*yjgD*
yocB	*ypuD*	*yqgZ*	*yqhB*	*ytaB*	*yunG*†	*yuzA*	*yvaA*
yvaG	*yvaK*†	*yvaJ*†	*yvbG*†	*yvgN*	*yvgO*	*yvgW*	*yvgZ*†
ywdD†	*ywiE*	*ywsC*	*ywtG*	*yxbG*	*yxlJ*	*yxzF*	*yxnA*

† Gene Not Differentially Expressed in any Functional Group

Table 3.3: Percentage of Differentially Expressed Genes From σ^B Regulon

Differentially Expressed σ^B Genes (110 Total)	
Functional Group (FG)	% of Genes Differentially Expressed
5	73.87% (82/111)
6	26.13% (29/111)
23	25.23% (28/111)
7	12.61% (14/111)
Remaining 25 components containing $< 11\%$	

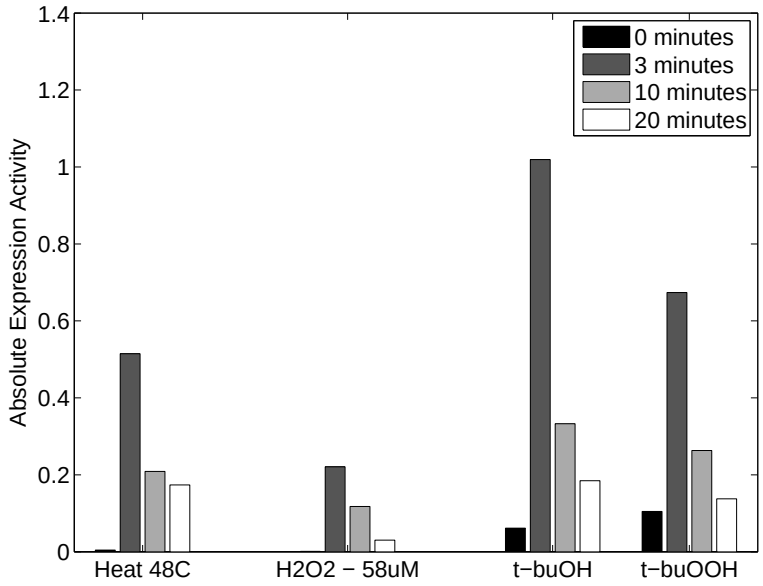

Figure 3.4: Absolute expression activity for functional group 5 (FG5). The figure shows a transient response of the genes differentially expressed in FG5 to the stress experiments for heat and peroxide with maximum activity at 3 minutes.

tionships between groups of genes that were differentially expressed in multiple functional groups. We used what we called a connectivity matrix to assess possible ties between groups of differentially expressed genes across multiple functional groups. An example of this connectivity matrix for the top functional groups that contained differentially expressed σ^B genes is shown in Table 3.4. The $(i,i)^{th}$ element assesses the exclusivity of the regulon by quantifying the number of genes that were differentially expressed in FGi only. This allowed us to gauge how self contained the regulon was in any one group. The ith row of the last column quantifies the number of genes that were differentially expressed in both FGi and any of the remaining functional groups, excluding those that were also differentially expressed in any of the other top functional groups. This allowed us to gauge the similarity between FGi and functional groups other than those considered to be in the list of top FGs. The remaining $(i,j)^{th}$ elements quantifies the number of genes that were differentially expressed in both FGi and FGj. High percentage values correspond to large overlaps that may indicate a functional connection between the genes from the regulon that are differentially expressed in FGi and FGj. The connectivity matrix in Table 3.4 shows very interesting pairwise relationships between genes from the σ^B regulon that were differentially expressed in FGs 5, 6, and 7. Note that FG23 is not included in the table because it was labeled as an inconsistent or unreliable estimate (see Table 3.1). Table 3.4 shows that 32 (39%) of the differentially expressed genes from the σ^B regulon were only differentially expressed in FG5. This was by far that largest number of genes isolated to any single functional group. This reinforces the notion that the expression activity for FG5 was a good representative of the regulatory activity of general σ^B genes. The other top functional groups for the differentially expressed σ^B genes were FG6 and FG7. Contrary to FG5, differentially expressed genes from the σ^B regulon were not isolated to these groups. The fact that a moderate amount of genes from σ^B were differentially expressed in these other functional groups indicated that these genes may take on expression characteristics that are not similar to the larger groups of σ^B genes. One interesting point to note is that a large percentage of genes differentially expressed in FG6 and FG7 are also differentially expressed in FG5. This overlap allow us to infer a functional relationship between the genes differentially expressed in FG6/7 and FG5 even though they may have dissimilar expression characteristics. To test this possibility, we applied our analysis to a group of genes that were known to be functionally related to the σ^B genes but exhibited alternative expression characteristics.

Table 3.4: Analysis of Differentially Expressed σ^B Regulon Across FGs 5, 6, and 7 (111/128)

	FG5	FG6	FG7	Other FGs Only
FG5 (82/111)	**32** (**39**%)	20 (24%)	11 (13%)	22 (27%)
FG6 (29/111)	20 (69%)	**3** (**10**%)	4 (14%)	5 (17%)
FG7 (14/111)	11 (79%)	4 (29%)	**1** (**7**%)	1 (7%)

Helmann established that most, but not all, of the genes in the σ^B regulon exhibited expression characteristics similar to those seen for FG5. He stated that a small group of genes did exist that belonged to the σ^B regulon but exhibited *alternative expression characteristics*. These genes, instead, had expression activity that peaked at 10 minutes instead of 3 minutes for the heat shock experiments. A set of 21 genes confirmed by Helmann as having these characteristics are listed in Table 3.5. We used our analysis to explore 1) if we can identify expression activities for these groups of genes that are consistent with prior knowledge, 2) if we can establish a connection between this group and the main FG associated with the majority of the σ^B genes, and 3) if the connectivity matrix for this group of genes can help explain the overlap characteristics seen in Table 3.4.

Table 3.5: List of the σ^B regulon known to have alternative expression characteristics

gsiB	*trxA*	*yacH*	*yacI*	*yflT*	*yhdG*	*yjbD*†
yxkO	*ydaD*	*ydaE*	*lctE*	*ycnH*	*ydbD*	*yhaS*
yhaT	*yhaU*	*yhxD*	*ypuD*	*yvgN*	*yvgW*	*ywdD*†

† Gene Not Differentially Expressed in any Functional Group

Table 3.6 shows the top functional groups that contained differentially expressed genes belonging to the alternate expression σ^B genes. Of the 21 genes in this list, 19 were found to be differentially expressed in at least one functional group. We see that FG6 and FG5 contained the most differentially expressed genes from this group with 68% and 47%, respectively. A connectivity matrix for FG6, FG5, and FG10 was created and is

shown in Table 3.7. We again eliminated FG23 from the analysis because this functional group was labeled as inconsistent with a cluster quality index of 0.8285. Upon initial inspection, it was clear from Table 3.7 that very few of these alternate expression σ^B genes were isolated to a single component and very few were differentially expressed across functional groups other than those seen in Table 3.6 (indicated by the last column of Table 3.7). We do see a noticeable overlap between the genes differentially expressed in FG6 and FG5. This relationship is similar to the characteristics seen in Table 3.4. We looked at the absolute expression activity of FG6, the functional group with the most differentially expressed genes from the alternate expression σ^B genes. This is shown in Figure 3.6. We saw that the absolute expression activity for FG6 also exhibited a transient-like characteristic. This time, however, the maximum absolute activity occurred at 10 minutes instead of 3 minutes. The expression activity for the remaining stress experiments were negligible. This negligible response indicated that these alternate expression genes may take on the expression characteristics of the majority of the σ^B genes for stress experiments other than heat. These results confirmed characteristics already established by Helmann, which stated that these genes exhibited expression characteristics that were different from the majority of the σ^B genes even though they were still functionally related to them. This allowed us to infer that genes belonging to the group of alternate expression σ^B genes tend to be differentially expressed in FG6 but have many genes that are also differentially expressed in FG5. This allowed us to speculate, based on the characteristics seen in Table 3.4, that the additional genes that were differentially expressed in both FG6 and FG5 may also be possible members of the alternate expression σ^B genes. Further experimental analysis is needed to verify if the remaining genes identified by this analysis but not tested by Helmann do in fact belong to the group of alternate expression σ^B genes.

We analyzed another set of genes known as the class III heat shock genes, which are known to be responsive to heat stress. Helmann identified the nine genes listed in Table 3.8 as members of this gene group. We used our analysis to identify expression patterns for these genes along with possible connections to other functional groups. Our analysis identified these genes as being primarily differentially expressed in FG6 and FG5 (see Table 3.9). We used the connectivity matrix shown in Table 3.10 to examine the overlap of these differentially expressed genes that were contained in both FG6 and FG5. We saw once again that the majority of genes from this group that were differentially expressed were shared between FG6 and FG5. This again allows us to infer a functional relationship between

Table 3.6: Percentage of Differentially Expressed Genes From Alternative Expression σ^B Genes

Differentially Expressed Alternative Expression σ^B Genes (19 Total)	
Functional Group (FG)	% of Genes Differentially Expressed
6	68.42% (13/19)
5	47.37% (9/19)
23	31.58% (6/19)
10	21.05% (4/19)
16 components containing $< 11\%$	

Table 3.7: Analysis of Differentially Expressed Alternate Expression σ^B Genes Across FGs 6, 5, and 10 (19/21)

	FG6	FG5	FG10	Other FGs Only
FG6 (13/19)	**2** (**15**%)	7 (54%)	3 (23%)	1 (8%)
FG5 (9/19)	7 (78%)	**1** (**11**%)	0 (0%)	1 (11%)
FG10 (4/19)	3 (75%)	0 (0%)	**0** (**0**%)	1 (25%)

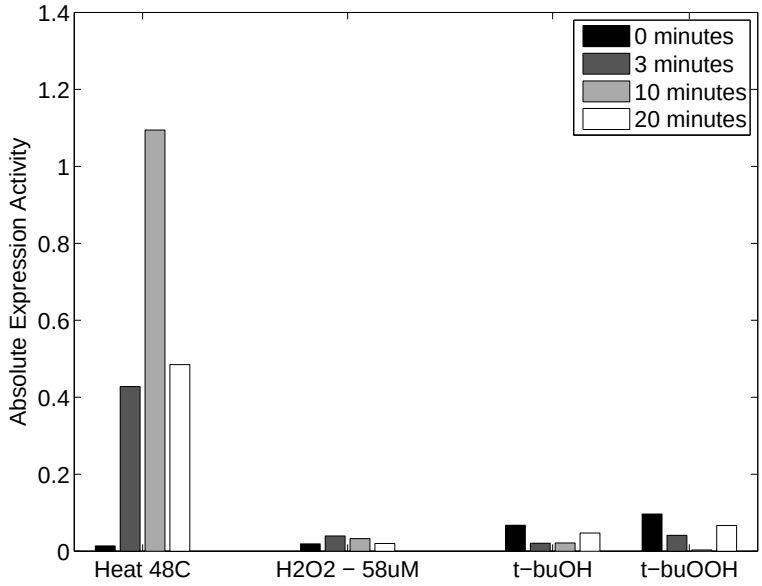

Figure 3.5: Absolute expression activity for functional group 6 (FG6). The figure shows a transient response to the heat shock experiments with maximum activity at 10 minutes. Other stress experiments do not show much activity.

these genes and the majority of genes belonging to the σ^B regulon. We suspect these genes take on the absolute expression activity characteristics seen for FG6 with the maximum transient activity occurring at 10 minutes (see Figure 3.6). Thus, based on our analysis, we concluded that these genes were functionally related to the σ^B genes but do not exhibit the same expression characteristics seen for the majority of the genes belonging to the σ^B regulon. Instead, we concluded that these genes possess characteristics that are more similar to the alternate expression σ^B genes. These results are in line with the characteristics of these genes established in the literature. The class III heat shock genes are identified in the literature as being regulated by σ^B and other alternative heat shock pathways controlled by σ^A [36]. These genes were also identified by Helmann et al. as having the same transient expression behavior with the activity peaking at 10 minutes [36].

Arginine Biosynthesis and Transport

Next, we analyzed the regulatory characteristics of a group of three operons involved in the biosynthesis and transport of the amino acid arginine. The total of 13 genes

Table 3.8: List of the Class III Heat Response Genes

ctsR	*yacH*	*yacI*	*clpC*	*sms*
yacK†	*clpP*	*clpE*	*clpC*	

† Gene Not Differentially Expressed in any Functional Group

Table 3.9: Percentage of Differentially Expressed Class III Heat Shock Genes

Differentially Expressed Class III Genes (8 Total)	
Functional Group (FG)	% of Genes Differentially Expressed
6	87.5% (7/8)
5	62.5% (5/8)
2 other components = 12.5% (1/9)	

Table 3.10: Analysis of Differentially Expressed Class III Heat Responsive Genes Across FGs 6, and 5 (8/9)

	FG6	FG5	Other FGs Only
FG6 (7/8)	**2** (**29**%)	5 (71%)	0 (0%)
FG5 (5/8)	5 (100%)	**0** (**0**%)	0 (0%)

belonged to 1) the *argC* biosynthetic operon, 2) the *argG* biosynthetic operon, and 3) the *yqiX* arginine transport system. A list of these genes are shown in Table 3.11. Our analysis identified at least 12 of these genes as differentially expressed in three functional groups, which included FGs 11, 6, and 29 (see Table 3.12). There were another three functional groups, which included FGs 20, 10, and 13, that contained at least 45% of the differentially expressed genes from this gene group. The connectivity matrix shown in Table 3.13 was used to establish possible relationships across these functional groups. Note that FG29 was excluded from the analysis because it represented a component that was labeled as inconsistent with a cluster quality index value of 0.675. FG20 was included in the analysis because it was only labeled as questionable. We see from Table 3.13 that none of the genes differentially expressed from this group were isolated to any one functional group. This truly emphasizes the inter- relationship across the top functional groups. It is also apparent from the first two columns of the connectivity matrix that all of the functional groups showed a maximum overlap of 100% with the top two functional groups, FG11, and FG6. This strong overlap with FG11 and FG6 along with no sign of isolated activation motivated us to infer that the response associated with both FGs 11 and 6 contribute to the regulatory behavior of these genes. Figure 3.6 shows the absolute expression activity for the heat shock experiments for FG6, FG11, and the combined activity from both.

Table 3.11: List of Genes Involved in Arginine Biosynthesis and Transport

argC	*argJ*	*argB*	*argD*	*carA*	*carB*	*argF*
argG	*argH*	*ytzD*	*yqiX*	*yqiY*	*yqiZ*	

We first notice that FG11 shows expression characteristics that are dominated by the expression characteristics of FG6. The combined expression activity shows characteristics that are more similar to the characteristics of FG6 where there is lower activity at 3 minutes with the peak in expression activity occurring at 10 minutes. Hence, we concluded that these genes should exhibit expression transient expression characteristics similar to those of the alternative expression σ^B genes where maximum activity occurred at the 10 minute point for the heat shock experiments. Unlike the alternative expression σ^B genes and the class III heat shock genes, a majority of the arginine biosynthesis and transport genes found to be differentially expressed in FG6 and FG11 were not differentially expressed

Table 3.12: Percentage of Differentially Expressed Arginine Biosynthesis and Transport Genes

Differentially Expressed Arginine Biosynthesis and Transport Genes (13 Total)	
Functional Group (FG)	% of Genes Differentially Expressed
11	100.0% (13/13)
6	92.30% (12/13)
29	92.30% (12/13)
20	69.23% (9/13)
10	53.84% (7/13)
13	46.15% (6/13)
12	30.77% (4/13)
7 other components = 7.69% (1/13)	

Table 3.13: Analysis of Differentially Expressed Arginine Biosynthesis and Transport Across FGs 11, 6, 20, 10, and 13 (13/13)

	FG11	FG6	FG20	FG10	FG13	Other FGs Only
FG11 (13/13)	**0** (**0**%)	12 (92%)	9 (69%)	7 (54%)	6 (46%)	1 (8%)
FG6 (12/13)	12 (100%)	**0** (**0**%)	9 (75%)	7 (58%)	6 (50%)	0 (0%)
FG20 (9/13)	9 (100%)	9 (100%)	**0** (**0**%)	6 (67%)	3 (33%)	0 (0%)
FG10 (7/13)	7 (100%)	7 (100%)	6 (86%)	**0** (**0**%)	4 (57%)	0 (0%)
FG13 (6/13)	6 (100%)	6 (100%)	3 (50%)	4 (67%)	**0** (**0**%)	0 (0%)

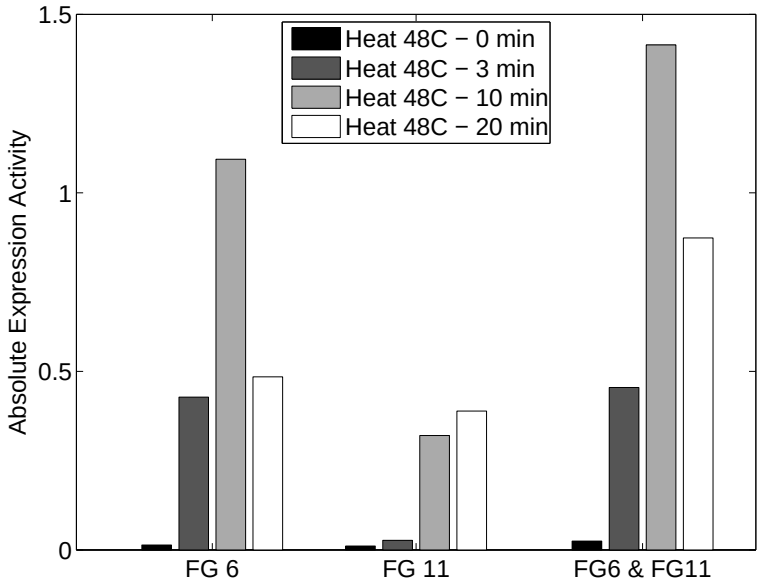

Figure 3.6: This figure shows the absolute expression activity of FG6, FG11, and the combined activity of both FG6 and FG11. Due to the dominant characteristics of FG6, the characteristics of the combined activity resembles that of FG6.

in FG5. In fact, only one of these genes, *argB*, was found to be differentially expressed in FG5. Hence, our results found no functional link between these genes and the genes belonging to the general stress response regulon σ^B. These characteristics again matched the conclusions determined by Helmann who identified these genes as having relatively low expression activity at the 3 minute time point, a large increase in activity at the 10 minute point, and diminishing activity at the 20 minute point. A regulatory or functional relationship between these genes and the general stress response genes in the σ^B regulon has not been established in the literature.

Other Heat Shock Response Genes

We also examined the regulatory characteristics of another set of genes known again to be responsive to the heat shock experiments. These genes, unlike most of the gene groups we have been working with, do not collectively have a dominant functional characteristic. We used this ICA based approach to perform exploratory analysis in order to identify possible ties to known functional groups. The set consisted of 66 genes which are

listed in Table 3.14. A total of 59 of these genes were found to be differentially expressed in at least one of the 30 functional groups. The functional diversity of these gene scattered them across several different functional groups, where the top five groups were FG11, FG6, FG10, FG12, and FG5 (see Table 3.15). Note that the percentage values for these functional groups are small compared to our previous results. The connectivity matrix in Table 3.16 was used to establish possible relationships between these differentially expressed genes across the top functional groups. Unlike the results seen for the groups of differentially expressed arginine biosynthesis and transport genes, no dominant relationships across these functional groups stood out. In fact, FG6 seemed to be more related to the other functional groups than to any of the remaining top functional groups in the list. One characteristic that particularly stands out is the isolation of the differentially expressed genes found in FG5 (12/59). As much as 5 of the 12 genes (42%) were only differentially expressed in FG5. This was the largest percentage of isolated genes of any other functional groups in the list. This, however, by no means signified a dominant relationship between the list of 66 genes and the σ^B genes. In fact, this, along with the small overlap of differentially expressed genes with FGs 11, 6, and 10, showed a weak functional relationship between this class of genes and the σ^B regulon. This supported the statements made by Helmann who concluded that there was no significant evidence that established that these group of genes were controlled by σ^B even though they had expression response patterns to heat shock that were similar to both the majority σ^B genes and the alternate expression σ^B genes. The notable isolated characteristics of the differentially expressed genes in FG5, however, justifies the use of further in-depth analysis through knock-out experimentation and RT-PCR on this small subset in order to identify if these 12 genes may in fact be directly or indirectly affected by σ^B. This list of 12 genes are shown in Table 3.17.

PerR Regulon

Initially, we looked almost entirely at genes identified as being responsive to the heat shock experiments. We also used this approach to identify and differentiate between groups of genes that were highly responsive to the peroxide experiments. We looked at the experiment profile generated for the $58\mu M$ H_2O_2 experiment at the 3 minute time point. This was done by obtaining the values from the corresponding row of **A**. The absolute expression activity for this experiment over all functional groups is shown in Figure

Table 3.14: List of Heat Responsive Genes with no Obvious Regulator Ties to σ^B

appD	*appF*	*appA*	*appB*	*appC*	*amhX*	*cydA*†
gerCA	*gntR*	*gntK*	*gntZ*†	*ilvB*	*ilvN*	*ilvC*
leuA	*leuB*	*leuC*	*leuD*	*murC*†	*nrgB*	*opuBA*
opuBB	*opuBC*	*opuBD*	*phoA*	*sacV*	*spoIIR*	*ywlB*
ssuD†	*ybfB*	*yceB*	*yceK*	*ycgJ*	*yclA*	*ydcN*
ydcM	*ydcL*	*yfjR*	*yfkF*	*yjcI*	*yjcJ*	*yjlB*
yoaD	*yoaC*	*yoaB*	*yoeB*	*yomZ*†	*yonA*	*yxeI*†
yqaB	*yqiK*†	*yqkE*	*yqjM*	*yqjL*	*yrhH*	*yrhI*
yrhJ	*ytiA*	*ytzE*	*yugS*	*yugU*	*yumD*	*yutK*
yuxJ	*ywfB*	*yxaB*				

† Gene Not Differentially Expressed in any Functional Group

Table 3.15: Percentage of Differentially Expressed Heat Response Genes from the 'Other' Category

Differentially Expressed Heat Response Genes (59 Total)	
Functional Group (FG)	% of Genes Differentially Expressed
11	32.20.% (19/59)
6	28.81% (17/59)
10	27.12% (16/59)
12	23.73% (14/59)
5	20.34% (12/59)
17 other components $< 14\%$	

Table 3.16: Analysis of 'Other' Differentially Expressed Heat Responsive Genes Across FGs 11, 6, 10, 12, and 5 (59/65)

	FG11	FG6	FG10	FG12	FG5	Other FGs Only
FG11 (19/59)	**4** (**21**%)	5 (26%)	9 (47%)	4 (21%)	1 (5%)	4 (21%)
FG6 (17/59)	5 (29%)	**3** (**18**%)	3 (18%)	5 (29%)	1 (6%)	6 (35%)
FG10 (16/59)	9 (56%)	3 (19%)	**2** (**13**%)	3 (19%)	1 (6%)	3 (19%)
FG12 (14/59)	4 (29%)	5 (36%)	3 (21%)	**3** (**21**%)	4 (29%)	0 (0%)
FG5 (12/59)	1 (8%)	1 (8%)	1 (8%)	4 (33%)	**5** (**42**%)	2 (17%)

Table 3.17: List of Potential σ^B Related Genes

opuBB	*opuBC*	*opuBD*	*spoIIR*	*ywlB*	*yjlB*
ytiA	*yugS*	*yugU*	*yutK*	*ywfB*	*yxaB*

3.7. Figure 3.7 clearly shows that the most active functional groups for this experiment were FG5 and FG8. The large amount of activity in FG5 was not surprising considering we have previously established that a large majority of the general stress response genes belonging to the σ^B regulon were differentially expressed in FG5. Further investigation of the differentially expressed genes in FG8 revealed a small set of genes belonging to the PerR regulon, particularly *mrgA*, *katA*, and *ykvW* (also known as *zosA*). None of these genes were found to be differentially expressed in FG5. Thus we are able to conclude that these genes, along with the majority of genes belonging to the σ^B regulon, were highly responsive to peroxide stress at 3 minutes. The algorithm was able to separate these genes into two groups allowing us to infer that these three genes are not functionally related to the σ^B regulon. These results again correlate well with those established in the literature. Helmann identified these two groups, the σ^B regulon and the PerR regulon, as two of the most responsive groups of genes in this experiment [111]. No functional relationship between these two groups of genes were highlighted. Our results highlight ICAs ability to identify and distinguish between two functionally diverse groups of genes even though they both had similar responses to hydrogen peroxide. Table 3.18 summarizes our results, highlighting the function and expression characteristics of the functional groups.

3.6 Chapter Summary

Identifying the complex regulatory relationships of functional groups of genes in response to changing stimuli is the first step to understanding how the genome can be adjusted and modified to accommodate a particular response. The complexity and size of these regulatory networks makes the reverse engineering of these connections very difficult, particularly when little a-priori knowledge about the regulatory mechanisms of interest is known. This motivates the generation and implementation blind analysis tools that are

Table 3.18: Summary of Regulons Identified in the Major Functional Groups

Summary of Results		
Regulon	Characteristics	Dominant FGs
σ^B Regulon	Transient expression characteristics with expression activity peaking at 3 minutes for all major stress experiments	FG5
Alternative Expression σ^B Regulon Genes	Transient expression characteristics with expression activity peaking at 10 minutes	FG6, FG5
Class III Heat Shock Genes (σ^B & σ^A)	Transient expression characteristics with expression activity peaking at 10 minutes	FG6, FG5
Arginine Biosynthesis and Transport	Transient expression characteristics with expression activity peaking at 10 minutes; no functional relationship with σ^B	FG11, FG6
'Other' Heat Response Genes	Varied expression characteristics; most have no functional relationship with σ^B	FG11, FG6, FG10, FG12, FG5
mrgA, *katA*, and *ykvW* (PerR Regulon)	Highly responsive to hydrogen peroxide at 3 minutes	FG8 ($58\mu M$ H_2O_2)

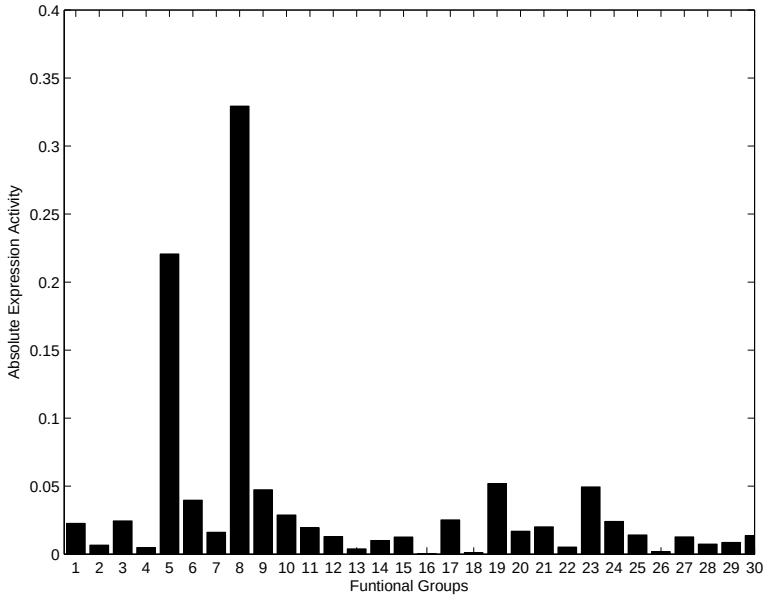

Figure 3.7: Graphical representation of the experiment profile for the $58\mu M$ of H_2O_2 experiment at the 3 minute time point. This shows the absolute expression activity of all functional groups for this experiment.

capable of gaining insight into these regulatory mechanism with only little or no a-prior information. We outlined a blind analysis procedure based on independent component analysis that builds connections between commonly regulated and expressed genes using only global gene expression data. The information from this analysis is used to identify strong relationships and isolate important groups of genes that are most responsive to a series of stimuli.

Our analysis applied independent component analysis to microarray data obtained from exposing *B. subtilis* to a variety of stress environments which included heat shock stress, peroxide stress, and alcohol stress. These stress experiments change the regulatory responses of the genome. The resulting overall expression activity is comprised of a combination of several functional responses that are activated due to changing stimuli. The analysis was used to separate the microarray into overlapping functional groups of genes that conserved the true characteristics of these functional responses. Our analysis qualified the relative expression activity of each functional group across all experimental stimulus and identified possible relationships between genes differentially expressed in multiple functional groups. We analyzed the regulatory characteristics of several groups of stress responsive

genes, which included genes belonging to the general stress response regulon known as σ^B, a groups of genes known as the Class III heat response genes, three operons responsible for arginine biosynthesis and transport, and uncharacterized genes known to respond specifically to heat shock. We also looked at the characteristics of a subset of genes in the PerR regulon that were responsive to large amounts of hydrogen peroxide. The regulatory characteristics interpreted from our analysis were compared to known stress related regulatory characteristics of *B. subtilis* that were previously observed in the literature.

Our results identified a large majority of genes in the σ^B regulon as differentially expression in functional group 5 (FG5). The absolute expression activity associated with FG5 was transient-like, with the peak activity occuring at 3 minutes for all major stress experiments. This activity was in-line with the expression characteristics identified by Helmann et al. in [36] for most, but not all, of the genes belonging to the σ^B regulon. Connectivity analysis also identified a subset of genes in the σ^B regulon as differentially expressed in FG5 and FG6. We identified these genes as having expression characteristics that were similar to the majority of the σ^B gene for the hydrogen peroxide, tert-butyl peroxide, and tert-butyl alcohol experiments. The characteristics for the heat shock experiments, however, showed what are called alternative expression characteristics where the maximum activity occurred at 10 minutes instead of 3 minutes. This analysis was able to detect subtle differences and differentiate these genes from the main group of σ^B genes even though they had similar responses for most of the other experiments. We found similar expression characteristics for a small subset of class III heat response genes that were also differentially expressed in FG5 and FG6. A connectivity analysis between the functional groups that contained differentially expressed genes from this gene group showed possible functional links between these genes and the majority of the genes belonging to the σ^B regulon. These absolute expression activities, however, more resembled those of the alternative expression σ^B genes. Our results were consistent with that of Helmann who identified these genes as related to the σ^B regulon but exhibiting these alternative expression characteristics whose transient behavior had maximum activity at 10 minutes [36].

This algorithm was also able to identify the expression activity of other groups of genes that were known as highly responsive to heat shock. We looked at a set of three regulons responsible for the biosynthesis and transport of the amino acid arginine. We identified these genes as primarily differentially expressed in FG11 and FG6. Connectivity analysis between the top 5 reliable functional groups showed functional dependencies on the top 2

functional groups, FG11 and FG6. This strong overlap along with no clear sign of isolated activity from either FG11 or FG6 motivated us to infer that both of these functional groups contribute to the expression characteristics of these genes. The overall combined absolute response of FG11 and FG6 showed expression characteristics for the heat shock experiments that peaked in activity at 10 minutes but showed a decrease in activity at 3 minutes and 20 minutes. Overall, these expression characteristics were akin to the transient behavior of the σ^B genes with alternative expression characteristics. These arginine biosynthesis and transport genes, however, were not identified as functionally related to the σ^B genes because they were not identified in FG5. These results were again consistent with the results established by Helmann where the genes in these regulons were not identified as having an obvious relationship to other general stress response genes other than the fact that they both exhibited similar responses to the heat shock experiments [36]. Thus, our analysis successfully distinguished the function of these regulons from the σ^B regulon even though the expression activity was somewhat similar.

We examined another set of heat responsive genes established as having a broad range of possible functions. As a result, these gene were differentially expressed over a broad range of functional groups with no one group possessing a significant majority. Connectivity analysis showed very little, if any, functional relationship between the top functional groups and the functional group associate with the majority of the σ^B genes. This allowed us to conclude that most of these genes were not functionally related to the σ^B regulon. These results again supports the statements made by Helmann who concluded that there was no dominant relationship between these genes and the σ^B regulon [36]. We also mention here that the analysis identified 12 genes from this gene group as differentially expressed in FG5. The isolated characteristics of these genes seen in our connectivity analysis highlight a possible relationship between this small subset of genes and the σ^B regulon. This justifies more in-depth experimental analysis through knockout mutations and RT-PCR in order to identify if these genes may in fact be directly or indirectly affected by σ^B. Other results that supports the utility of this algorithm involved the organism's the response to hydrogen peroxide. Our results were able to differentiate between genes from the σ^B regulon and genes from the PerR regulon, which Helmann identified as two of the most responsive groups of genes for the 3 minute $58\mu M$ hydrogen peroxide experiment. This further highlights the algorithm's ability to distinguish between functionally diverse groups of genes with similar expression activities.

Chapter 4

State and Parameter Estimation in Biochemical Networks using Optimal Control Theory

4.1 Motivation

Biochemical networks are an ensemble of complex chemical reactions that form the heart of most biological processes at the molecular level. These networks are responsible for manufacturing and manipulating the compounds necessary for maintaining proper cellular function. Signal transduction and metabolic networks are at the forefront of these processes and play an important role in normal cellular operations. Signal transduction networks form the basis of communications that allow the cell to interpret and respond to external signaling mechanisms such as changes in nutrients, temperature, mechanical stress, etc. Metabolic networks incorporate responses at gene regulatory levels to properly guide the conversions of small molecules. This is used to manage the biochemical and physiological composition of dietary minerals, vitamins, and other molecules used by the organism to survive. These networks combine to form complex operational structures that contribute to the robustness

of living organisms. The overall dynamic characteristics of these networks significantly govern the complex characteristics of cellular behavior. For example, malfunctions in the signal transduction networks associated with programmed cell death can play a significant role in neurodegenerative diseases like Parkinson's, Alzheimer's, and Huntington's disease [117]. Disruptions in the widely used Extracellular signal Regulated Kinase (ERK) signaling pathways are common in many forms of human cancers [118]. The metabolic production of compounds like polyamines are important to cellular processes like cell proliferation, transcription, and translation. The deregulation of these compounds are involved in other relevant processes like cancer and neurodegenerative disorders [119]. The manipulation of metabolites involved in carotenoid biosynthesis can attribute to increased nutrition in plants like tomato fruit [120]. Alternate configurations of lignin can lead to more efficient ways to produce biofuels from cellulose in plants [121]. Thus, it is an important mission to better understand how these biochemical networks behave and understand how we can better manipulate the dynamics of these networks to meet some global objective.

Both signal transduction and metabolic networks are powered by functional proteins or enzymes that facilitate the activation of other proteins or the conversion of compounds within the network. Complete characterization of the nonlinear mechanisms associated with these activities along with the kinetic rate parameters associated with the reactions is a particularly elusive goal that has been hindered by several factors. This includes 1) the lack of first principles knowledge about the system and 2) our inability to measure all time-varying components of the system. Some realizations of the nonlinear activities in the network can be obtained through the use of mass action kinetics. Without detailed knowledge of all of the dynamics of the components and complete first principles knowledge of all of the kinetic rate parameters, these nonlinear descriptions are incomplete at best. This prevents researchers from fully characterizing such networks. The lack of these mathematical descriptions prevent the progression and development of analytical approaches for manipulating and controlling these networks. It also limits approaches for analytically testing hypotheses in order to verify the first principles knowledge used to develop these models. Thus, the development of approaches for adequately modeling these networks is imperative to the pursuit of the medicinal, nutritional, and social benefits that can be achieved through the systematic manipulation of these networks.

The main issues that are faced in developing analytical procedures for modeling the full nonlinear characteristics of biochemical networks are based in the subtle issues

that are listed above. Newly created and reformulated modeling approaches must therefore address these road blocks, which include nonlinear systems characteristics, sparse uncertain measurements, and biological constraints on pathway dynamics. These methods for creating and refining predictive models of living systems will require approaches that go beyond the traditional heuristic and statistical data-mining techniques used in computational biology and bioinformatics. A strong theoretical base is needed that will allow the development of novel approaches and still provide a means of expanding and scaling such methods according to the growing size and complexity that has been seen in living systems. Signal processing theory, systems theory, and control theory provide a strong base upon which to build a framework for the analytical modeling of biological systems. For years, these approaches have been used successfully to model many aspects of mechanical and electrical systems. The added complexity and subtle nuances that are encountered with biological systems inhibit a strictly transparent integration of these techniques into the biological genre. The careful integration of existing analytical tools along with the creation of new tools is essential for modeling these networks. These tools are an important first step toward better understanding, manipulating, and controlling living systems without having the systems behave in an adverse or unexpected manor.

In this paper, we present an algorithm for estimating the kinetic parameters and unmeasured components or states of a biochemical network in the presence of equality and/or inequality constraints on the dynamics of the network. Similar to the initial value approaches described in [78] and [79], this method used a least-squares estimation approach to numerically calculate the unmeasured dynamics of the network. This approach, however, solved the problem using an optimal control formulation. The strength of the optimal control approach to this problem is the formal integration of dynamic equality and inequality constraints into the solution without the need for constrained optimization. This method was applied to a model of a single-step signal transduction pathway. This pathway serves as a template and can be used as the base element for generating larger more complex biochemical networks. For benchmarking purposes, we used partial synthetic measurement profiles acquired from two different sampling schemes to estimate unknown characteristics of this pathway. We also explored how bounded uncertainty in the measurements impacted the algorithms ability to correctly estimate the unknown dynamics of the pathway. The optimal control approach to this estimation problem provides a powerful foundation that can be extended to estimating time varying biological systems in the presence of dynamic

constraints and for generating actions needed to drive the dynamics of the network to achieve some overall goal.

This chapter is outlined as follows. Section 4.2 describes the single-step signal transduction pathway used in the analysis. This section describes the characteristics of this system and presents a mathematical model that describes the dynamics of the pathway. The estimation algorithm is formulated in Section 4.3. As a proof of concept, this algorithm is applied to an estimation problem involving the sample pathway. The results of the estimation algorithm are presented in Section 4.4. Section 4.5 summarizes of the performance of the algorithm.

4.2 Mathematical Representation

Most biochemical pathways are comprised of a complex network of chemical reactions that are controlled by catalyst proteins called enzymes. These enzymes are used to catalyze the conversion of one substance to another at some kinetic rate. Understanding the full dynamics of these networks means understanding all component dynamics as well as all kinetic rate parameters associated with these systems. Conditions often arise, however, where all component concentrations and all kinetic rate parameters cannot be measured. This limits our ability to completely assess the time varying characteristics of biochemical pathways. This profoundly impacts long-term goals of engineering task specific pathways that are capable of achieving predetermined biological objectives. Analytical algorithms that are specifically formulated to estimate these unmeasured characteristics are essential for creating mathematical models that represent the complete dynamics of biochemical pathways. Well understood pathways are needed to aid in the development and formulation of such algorithms. In this work, we looked at a single-step signal transduction pathway consisting of a small series of chemical reactions that take place in many larger biochemical networks. Kutalik et al. presented a description of this simple pathway in [122]. This model is of practical interest because it is simple and well understood. Also, subtle variants and combinations of this simple model can be used to build more complex pathways. A block diagram of this pathway along with the associated chemical reactions is shown in Figure 4.1.

This pathway is comprised of four components, the substrate (S), the enzyme (E),

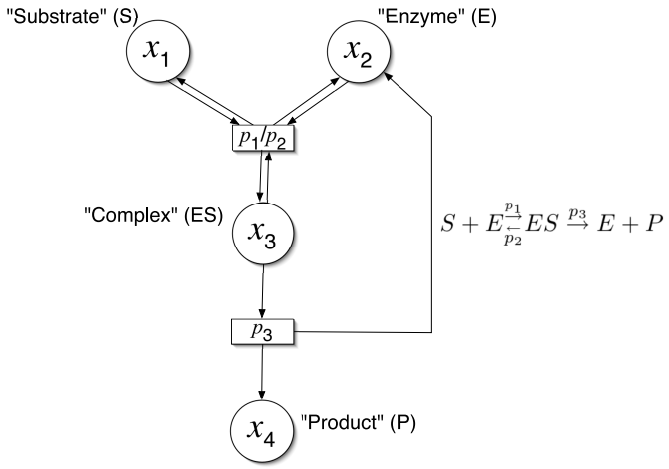

Figure 4.1: Template model of a single-step signal transduction pathway [122].

the enzyme-substrate complex (ES), and the product (P). The diagram shows that the presence of both the substrate and the enzyme can lead to the formation of an enzyme-substrate complex with kinetic rate constant p_1. The complex can either disassociate with a kinetic rate constant p_2 or proceed to form the product with kinetic rate constant p_3. Kutalik et al. used mass action kinetics to obtain the following set of ordinary differential equations (ODEs) to represent the pathway

$$\dot{x}_1(t) = -p_1 x_1(t) x_2(t) + p_2 x_3(t) \tag{4.1}$$

$$\dot{x}_2(t) = -p_1 x_1(t) x_2(t) + (p_2 + p_3) x_3(t) \tag{4.2}$$

$$\dot{x}_3(t) = p_1 x_1(t) x_2(t) - (p_2 + p_3) x_3(t) \tag{4.3}$$

$$\dot{x}_4(t) = p_3 x_3(t). \tag{4.4}$$

Here, the states x_1, x_2, x_3, and x_4 refer to the concentrations of the substrate, the enzyme, the enzyme-substrate complex, and the product, respectively. We generated an equivalent system of discrete-time nonlinear difference equations using an Euler approximation of the continuous-time system

$$
\begin{aligned}
x_1(k+1) &= x_1(k) + \Delta T(-p_1 x_1(k) x_2(k) \\
&\quad + p_2 x_3(k)) \qquad (4.5) \\
x_2(k+1) &= x_2(k) + \Delta T(-p_1 x_1(k) x_2(k) \\
&\quad + (p_2 + p_3) x_3(k)) \qquad (4.6) \\
x_3(k+1) &= x_3(k) + \Delta T(p_1 x_1(k) x_2(k) \\
&\quad - (p_2 + p_3) x_3(k)) \qquad (4.7) \\
x_4(k+1) &= x_4(k) + \Delta T p_3 x_3(k) \qquad (4.8)
\end{aligned}
$$

where the index k is an integer representing the discrete time t_k and ΔT represents the sample period. It is important to note that the dynamics of this system are completely characterized by the initial concentrations of the system components and the values of the kinetic rate parameters because the system was modeled with no external inputs.

4.3 Estimation Methodology

The inability to measure all important system-level characteristics of biochemical networks has hindered our ability to model, understand, and control the dynamics of such systems. Generally speaking, it is time consuming, expensive, and often impossible to measure all the components and parameters that affect these dynamics. This is a direct result of the increased complexity of biological processes and limitations in current experimental techniques. Analytical procedures developed to estimate both the component trajectories and the kinetic parameters can be used to help capture the complete dynamics of these systems. Various levels of difficulty are encountered when using traditional approaches for estimating the unmeasured characteristics of biological processes. Obstacles include high-dimensional sparse data sets, nonlinear system characteristics, and biological constraints on components and pathway dynamics. Estimation techniques that are formulated specifically to address these obstacles can be used to help build mathematical models of biochemical networks. These models will aid in the development of more extensive analytical procedures that will allow us to gain insight into the mechanisms needed to manipulate and control the dynamics of these systems. In this section, we present an estimation procedure that

was formulated based on a constrained optimal control solution to a least squares estimation problem that addressed many of these concerns. Particularly, this formulation took into consideration possible linear and/or nonlinear constraints on the system dynamics and formulated a solution without the need for complex constrained optimization approaches. The optimal control framework used to formulate the solution to this problem provides a powerful foundation that can be extended to estimate time varying parameters in biological systems in the presence of dynamic constraints and to identify the actions needed to drive the dynamics of the network to some predetermined goal.

4.3.1 Optimal Control Formulation

Many complex systems are described by a system of nonlinear ordinary differential equations (ODEs). ODEs are often the representation of choice when developing mechanistic models for biochemical pathways [71] [123] [124]. These representations are often derived using mass action kinetics, which identifies the rate of elementary reactions in the pathway as being proportional to the product of the concentration of the participating molecules [54]. These ODEs are conveniently represented in a nonlinear state-space form as described in [125].

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{p}) \tag{4.9}$$

Here $\mathbf{x} \in \mathcal{R}^{n\times 1}$ is the state vector containing concentrations for the pathway components, $\mathbf{p} \in \mathcal{R}^{m\times 1}$ is the vector of kinetic rate parameters, $\mathbf{f}(\cdot) : \mathcal{R}^{n\times 1} \rightarrow \mathcal{R}^{n\times 1}$ is a nonlinear vector function, and $t \in [t_0,\ t_f]$ represents the time of interest. The representation in (4.9) has no external inputs. This implies that the dynamics of the pathway are completely characterized by the initial condition of the state vector, $\mathbf{x}(t_0)$, and the kinetic parameters, $\mathbf{p}$. Limitations on first principles knowledge and current measurement technologies impede our ability to acquire these vectors in their entirety for many biochemical pathways, limiting our ability to model these processes.

Measurements acquired from the system are modeled as

$$\mathbf{y}(t) = \mathbf{h}(\mathbf{x}(t)) \tag{4.10}$$

where $\mathbf{y} \in \mathcal{R}^{p\times 1}$ represents the measurements and $\mathbf{h}(\cdot) : \mathcal{R}^{n\times 1} \rightarrow \mathcal{R}^{p\times 1}$ describes the general

transformation associated with the measuring apparatus. The continuous-time state-space system described in (4.9) and (4.10) is often approximated by its discrete-time counterpart,

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{x}_k + \Delta T \mathbf{f}(\mathbf{x}_k, \mathbf{p}), \quad k = 0, \ldots, N-1 \\ \mathbf{y}_i &= \mathbf{h}(\mathbf{x}_i), \qquad\qquad i \in \mathcal{I}. \end{aligned} \tag{4.11}$$

The subscript k indicates the discrete time t_k and ΔT represents the sampling period. The set $\mathcal{I}$ is a subset of integers between 0 and $N-1$ corresponding to the set of discrete time points t_k, $k \in \mathcal{I}$, where measurements from the systems are observed. Given an appropriate sampling period, many naturally occurring systems can be approximated using the discrete representation in (4.11). This eliminates the need for complex integration techniques for solving and simulating continuous-time nonlinear ordinary differential equations. We approached the problem of modeling and understanding the dynamics of a biochemical network by first correlating the system equations in (4.11) with known information about the pathway. The nonlinear function $\mathbf{f}(\cdot)$ described the nonlinear mechanisms that governed the dynamics of the pathway, which resulted from the system of chemical reactions associated with the pathway. This is often obtained using some kinetics modeling approach like mass action kinetics [54], Michaelis-Menten kinetics [55], or general mass action system representations [13]. Direct measurements of specific pathway components would result in a trivial analytical representation of $\mathbf{h}(\cdot)$. Indirect measurements of these components are more difficult to characterize and would require empirical approximations in order to get an analytical form of the observation transformation. The actual discrete-time measurements, $\mathbf{y}_i$, $i \in \mathcal{I}$, used in this work were obtained using simulated measurements that modeled the characteristics of existing measurement techniques. The goal of this work was to provide an approach that was capable of estimating the complete dynamics of the pathway using only available information such as the discrete-time measurements of accessible pathway components and first principles knowledge of the nonlinear mechanisms that governed the dynamics of the pathway. Estimating the complete dynamics of the system equation in (4.11) is equivalent to estimating the correct initial component concentrations, $\mathbf{x}_0^*$, and the correct kinetic parameter vector, $\mathbf{p}^*$, where the superscript * represents the true value. This requires a cost function that, when minimized, provides the "best" estimates of $\mathbf{x}_0^*$ and $\mathbf{p}^*$. Least-squares is a well accepted criteria for determining "best" estimates from error cal-

culations based on measured data points [126] and has often been used in approaches for estimating unknown or unmeasurable characteristics of biochemical networks [78] [79]. The procedures used to obtain concentration measurements of the pathway components often involves extensive laboratory work and post-processing. Thus, this problem does not fit well into a real-time estimation scheme. This allowed us to formulate a batch-mode cost function where error calculations were obtained using all measured component concentrations over the time interval of interest. We defined the following cost function

$$\min_{\mathbf{x}_0,\ \mathbf{p}} J_0(\mathbf{x}_0, \mathbf{p}) = \frac{1}{2} \sum_{i \in \mathcal{I}} (\mathbf{h}(\mathbf{x}_i) - \mathbf{y}_i)^T \mathbf{Q} (\mathbf{h}(\mathbf{x}_i) - \mathbf{y}_i) \tag{4.12}$$

which was subjected to the following dynamics constraints

$$\begin{array}{ll} \mathbf{x}_{k+1} = \mathbf{x}_k + \Delta T \mathbf{f}(\mathbf{x}_k, \mathbf{p}), & k = 0, \ldots, N-1 \\ \mathbf{x}_k \geq 0, & k = 0, \ldots, N-1 \end{array}, \tag{4.13}$$

where $\mathbf{Q} \in \mathcal{R}^{p \times p}$ represented a weighting matrix that could be used to adjust the significance of any measurement. Minimizing J_0 with respect to (wrt) $\mathbf{x}_0$ and $\mathbf{p}$ is equivalent to searching for the $\hat{\mathbf{x}}_0$ and the $\hat{\mathbf{p}}$ that produces the component trajectory that, once measured, best matches the discrete-time measurements acquired from the system. The equality constraint ensures that the estimated state trajectory follows the nonlinear rules that govern the dynamics of the system. The inequality constraint ensures that the dynamics for the component concentrations are nonnegative at all discrete times $t_k,\ k = 0, \ldots, N-1$. This constraint is intuitive since concentrations are never negative. Equations (4.12) and (4.13) do not form your traditional constrained optimization problem that just limits the variables of interest, $\mathbf{x}_0$ and $\mathbf{p}$. They form an optimization problem that is bounded by nonlinear system dynamics and by constraints on the pathway dynamics due to practical biological limitations. An approach was needed that was well suited to handle this type of problem. We approached this using an optimal control formulation of this least squares estimation problem with the added capability of handling both equality [127] and inequality [128] constraints on pathway dynamics. Optimal control is a powerful tool for estimating the appropriate input that produces system dynamics that are optimal with respect to some cost function. The approach presented here reformulated the optimal control approach to work specifically with the initial value models used to describe biochemical pathways.

Other applications of optimal control for state estimation include state and flux estimation in signal transduction networks [71] and data assimilation approaches for state estimation in oceanic and weather models [129].

A technique for handling the inequality constraints on the pathway dynamics was first integrated into the procedure. This technique has previously been used for traditional optimal control problems [128]. First, the n inequality constraints seen in (4.13) were transformed into a single equality constraint represented as an additional state difference equation. This difference equation was then appended to the original system of difference equations in (4.11). This new state variable was defined as

$$\begin{aligned} x_{n+1}(k+1) &= x_{n+1}(k) + \left[x_1^2(k)u(-x_1(k)) \right. \\ &\quad + x_2^2(k)u(-x_2(k)) \\ &\quad \left. + \ldots + x_n^2(k)u(-x_n(k))\right]. \end{aligned} \tag{4.14}$$

The scalar state space variable $x_i(k)$ in (4.14) has a slightly different notation where i represents the ith state and k is an integer representing the discrete time t_k. The function $u(\cdot)$ is a unit-step function where

$$u(-x_i(k)) = \begin{cases} 0 & x_i(k) \geq 0 \quad i = 1, 2, \ldots, n \\ 1 & x_i(k) < 0 \quad i = 1, 2, \ldots, n \end{cases}.$$

Based on the definition of the unit-step function, it is clear that the term inside the brackets of (4.14) is nonnegative with the result of the sum strictly greater than zero if any of the inequality constraints in (4.13) are violated and equal to zero if and only if all inequality constraints on the system dynamics are met. This state variable was defined as having a fixed initial condition, $x_{n+1}(0) = 0$, and a fixed final condition, $x_{n+1}(N) = 0$. Given that all terms in (4.14) are greater than or equal to zero for all times t_k, an initial and final condition equal to zero guaranteed that all values of $x_{n+1}(k)$, $k = 1, \ldots, N-1$ are also zero, thus satisfying all inequality constraints. These initial and final constraints on the appended state were used as boundary conditions in the resulting optimal control formulation. This additional state was augmented onto the original state vector forming a new state vector $\tilde{\mathbf{x}}_k = \left[\mathbf{x}_k^T \; x_{n+1}(k)\right]^T$. This was used in combination with the equality constraint in (4.14)

to redefine the minimization problem as

$$\min_{\tilde{\mathbf{x}}_0,\ \mathbf{p}} J_0(\tilde{\mathbf{x}}_0, \mathbf{p}) = \frac{1}{2}\sum_{i\in\mathcal{I}}(\mathbf{h}(\tilde{\mathbf{x}}_i) - \mathbf{y}_i)^T\mathbf{Q}(\mathbf{h}(\tilde{\mathbf{x}}_i) - \mathbf{y}_i) \tag{4.15}$$

subject to

$$\tilde{\mathbf{x}}_{k+1} = \tilde{\mathbf{f}}(\tilde{\mathbf{x}}_k, \mathbf{p}), \quad k = 0, \ldots, N-1, \tag{4.16}$$

where $\tilde{\mathbf{x}}_k \in \mathcal{R}^{(n+1)\times 1}$ and $\tilde{\mathbf{f}}(\cdot) : \mathcal{R}^{(n+1)\times 1} \rightarrow \mathcal{R}^{(n+1)\times 1}$. The function $\tilde{\mathbf{f}}(\cdot)$ was defined as

$$\tilde{\mathbf{f}}(\tilde{\mathbf{x}}, \mathbf{p}) = \begin{bmatrix} \mathbf{x}_k + \Delta T\mathbf{f}(\mathbf{x}_k, \mathbf{p}) \\ x_{n+1}(k) + \left[x_1^2(k)u(-x_1(k)) + \ldots + x_n^2(k)u(-x_n(k))\right] \end{bmatrix}.$$

Optimal control theory provided a systematic approach to solving this problem. In traditional optimal control problems, the initial conditions and parameters are known and fixed and the external inputs of the systems are varied to determine the best control that minimizes the cost function. This problem differs slightly in that some or all of the initial conditions and parameters were unknown. With no external inputs, the initial concentration vector $\mathbf{x}_0$ and the kinetic rate parameter vector $\mathbf{p}$ were varied in order to minimize the cost function while still satisfying the equality constraints in (4.16) and the initial and final conditions on x_{n+1}. With the inequality constraints absorbed into (4.16), Lagrange multipliers were used to transform the constrained optimization in (4.15) and (4.16) into as unconstrained optimization problem. This was done by adjoining the constraints to the cost function and using the Lagrange multipliers to enforce the constraints

$$\begin{aligned} \min_{\tilde{\mathbf{x}}_0,\ \mathbf{p}} \hat{J}_0 \quad = \quad & \frac{1}{2}\sum_{i\in\mathcal{I}}(\mathbf{h}(\tilde{\mathbf{x}}_i) - \mathbf{y}_i)^T\mathbf{Q}(\mathbf{h}(\tilde{\mathbf{x}}_i) - \mathbf{y}_i) \\ & + \sum_{k=0}^{N-1} \boldsymbol{\lambda}_{k+1}^T(\tilde{\mathbf{f}}(\tilde{\mathbf{x}}_k, \mathbf{p}) - \tilde{\mathbf{x}}_{k+1}). \end{aligned} \tag{4.17}$$

The vector $\boldsymbol{\lambda}_{k+1} \in \mathcal{R}^{(n+1)\times 1}$ are the set of undetermined Lagrange multipliers, $k = 0, \ldots, N-1$. These undetermined vectors were solved using the following approach. First, the Lagrange multipliers and the error argument from the cost function were used to generate the Hamiltonian function [127]

$$H_k(\mathbf{x}_k, \mathbf{p}, \boldsymbol{\lambda}_{k+1}) = \begin{cases} \frac{1}{2}(\mathbf{h}(\tilde{\mathbf{x}}_k) - \mathbf{y}_k)^T \mathbf{Q}(\mathbf{h}(\tilde{\mathbf{x}}_k) - \mathbf{y}_k) + \boldsymbol{\lambda}_{k+1}^T \tilde{\mathbf{f}}(\tilde{\mathbf{x}}_k, \mathbf{p}), & k \in \mathcal{I} \\ \boldsymbol{\lambda}_{k+1}^T \tilde{\mathbf{f}}(\tilde{\mathbf{x}}_k, \mathbf{p}), & \text{otherwise} \end{cases} . \quad (4.18)$$

This was back substituted into the modified cost function in (4.17), yielding the following equation

$$\min_{\tilde{\mathbf{x}}_0,\ \mathbf{p}} \hat{J}_0 = H_0(\tilde{\mathbf{x}}_0, \mathbf{p}, \boldsymbol{\lambda}_1) - \boldsymbol{\lambda}_N^T \tilde{\mathbf{x}}_N + \sum_{k=1}^{N-1} \left[H_k(\tilde{\mathbf{x}}_k, \mathbf{p}, \boldsymbol{\lambda}_{k+1}) - \boldsymbol{\lambda}_k^T \tilde{\mathbf{x}}_k \right]. \quad (4.19)$$

A stationary point for $\hat{J}_0(\cdot)$ occurs at the point where the differential of the cost function $d\hat{J}_0 = 0$. This yields a constrained minimum solution of (4.19). It is clear from (4.19) that $\hat{J}_0$ is a function of $\tilde{\mathbf{x}}_k$, $k = 0, \ldots, N$, $\boldsymbol{\lambda}_{k+1}$, $k = 0, \ldots, N-1$, and $\mathbf{p}$. That being the case, the differential of the cost function $d\hat{J}_0$ was written as

$$\begin{aligned} d\hat{J}_0 &= (-\boldsymbol{\lambda}_N)^T d\tilde{\mathbf{x}}_N + \sum_{k=1}^{N-1} \left(\frac{\partial H_k}{\partial \tilde{\mathbf{x}}_k} - \boldsymbol{\lambda}_k \right)^T d\tilde{\mathbf{x}}_k \\ &\quad + \sum_{k=1}^{N} \left(\frac{\partial H_{k-1}}{\partial \boldsymbol{\lambda}_k} - \tilde{\mathbf{x}}_k \right)^T d\boldsymbol{\lambda}_N + \sum_{k=0}^{N-1} \left(\frac{\partial H_k}{\partial \mathbf{p}} \right)^T d\mathbf{p}. \end{aligned} \quad (4.20)$$

A constrained minimum solution required that $d\hat{J}_0 = 0$ for any $d\tilde{\mathbf{x}}_k$, $d\boldsymbol{\lambda}$, and $d\mathbf{p}$. This was accomplished by setting the necessary parts of (4.20) to zero. Setting the necessary terms in (4.20) equal to zero and simplifying the terms algebraically yielded a set of coupled difference equations

$$\begin{aligned} \tilde{\mathbf{x}}_{k+1} &= \tilde{\mathbf{f}}(\tilde{\mathbf{x}}_k, \mathbf{p}), \\ &\quad k = 0, \ldots, N-1 \quad (4.21) \\ \boldsymbol{\lambda}_k &= \begin{cases} \left(\tilde{\mathbf{f}}_{\tilde{\mathbf{x}}_k} \right)^T \boldsymbol{\lambda}_{k+1} \\ \quad + (\mathbf{h}_{\tilde{\mathbf{x}}_k})^T \mathbf{Q}(\mathbf{h}(\tilde{\mathbf{x}}_k) - \mathbf{y}_k), & k \in \mathcal{I} \\ \left(\tilde{\mathbf{f}}_{\tilde{\mathbf{x}}_k} \right)^T \boldsymbol{\lambda}_{k+1}, & \text{otherwise} \end{cases} \\ &\quad k = N-1, \ldots, 1 \quad (4.22) \end{aligned}$$

with the following boundary conditions

$$\text{BC1:} \qquad (\boldsymbol{\lambda}_N)^T \, d\tilde{\mathbf{x}}_N = 0 \tag{4.23}$$

$$\text{BC2:} \qquad \left(\left(\tilde{\mathbf{f}}_{\tilde{\mathbf{x}}_0} \right)^T \Big|_{\tilde{\mathbf{x}}_0} \boldsymbol{\lambda}_1 + (\mathbf{h}_{\tilde{\mathbf{x}}_0})^T \Big|_{\tilde{\mathbf{x}}_0} \mathbf{Q}(\mathbf{h}(\tilde{\mathbf{x}}_0) - \mathbf{y}_0) \right)^T d\tilde{\mathbf{x}}_0 = 0 \tag{4.24}$$

$$\text{BC3:} \qquad \left(\sum_{k=0}^{N-1} \left(\tilde{\mathbf{f}}_{\mathbf{p}} \right)^T \Big|_{\tilde{\mathbf{x}}_k} \boldsymbol{\lambda}_{k+1} \right)^T d\mathbf{p} = 0. \tag{4.25}$$

Equation (4.21) represents the original systems of equations that govern the dynamics of the network. This is called the *state* equation and runs forward in time from $k = 0$ to $k = N - 1$. Equation (4.22) is a secondary set of difference equations and is known as the *costate* equation. This system runs backward in time from $k = N - 1$ to $k = 1$. Both equations are autonomous. The *state* equation is completely characterized by the initial condition $\mathbf{x}_0$ and the kinetic parameter vector $\mathbf{p}$. The *costate* equation is completely characterized by the final condition $\boldsymbol{\lambda}_N$, the initial condition $\mathbf{x}_0$, and the parameter vector $\mathbf{p}$. These coupled equations with boundary conditions represented a two point boundary value problem whose solution was also the solution to the least squares estimation problem with dynamic constraints. The next obstacle we faced was identifying the initial condition $\mathbf{x}_0$, the final condition $\boldsymbol{\lambda}_N$, and the kinetic parameter vector $\mathbf{p}$ that solved this two point boundary value problem. We first examined the boundary conditions of the two point boundary value problem and used the knowledge of the fixed initial and final values of x_{n+1} to find the most appropriate values of $\mathbf{x}_0$, $\boldsymbol{\lambda}_N$, and $\mathbf{p}$ that provided a solution to this problem.

We first examined BC1, which limited the range of possible values for $\boldsymbol{\lambda}_N$. It was clear from the definition of $\tilde{\mathbf{x}}_k$ that both vectors could be partitioned as follows

$$\begin{bmatrix} \hat{\boldsymbol{\lambda}}_N \\ \lambda_{n+1}(N) \end{bmatrix}^T \begin{bmatrix} d\mathbf{x}_N \\ dx_{n+1}(N) \end{bmatrix} = \left(\hat{\boldsymbol{\lambda}}_N \right)^T d\mathbf{x}_N + \lambda_{n+1}(N) dx_{n+1}(N) = 0. \tag{4.26}$$

Here, $\hat{\boldsymbol{\lambda}}_N = [\lambda_1(N), \ \lambda_2(N), \ldots, \lambda_n(N)]^T$ and $\mathbf{x}_N$ is the original state vector without the augmented state variable. As stated previously, the final condition $x_{n+1}(N)$ was considered known and fixed, thus $dx_{n+1}(N) = 0$. This meant that $\lambda_{n+1}(N)$ could be any value without

violating this condition. Equation (4.26) was thus rewritten as

$$\left(\hat{\boldsymbol{\lambda}}_N\right)^T d\mathbf{x}_N = 0. \tag{4.27}$$

In general, the final condition of the remaining states were not fixed so $d\mathbf{x}_N \neq 0$. This meant that (4.27) was true only if $\hat{\boldsymbol{\lambda}}_N = 0$. Thus, BC1 was true only if $\lambda_i(N) = 0$, $i = 1, \ldots, n$. The value for λ_{n+1}, however, still remained unknown.

We then examined BC2. This equation in combination with the costate equation in (4.22) was used to identify an appropriate value for the unknown constant $\lambda_{n+1}(N)$. The analytical form of $\boldsymbol{\lambda}_1$ was first needed in order to proceed with the analysis. The following variables were defined in order to simplify later derivations

$$\begin{aligned} \mathbf{A}_k &= \left(\tilde{\mathbf{f}}_{\tilde{\mathbf{x}}_k}\right)^T\Big|_{\tilde{\mathbf{x}}_k} \\ \mathbf{B}_k &= \left(\mathbf{h}_{\tilde{\mathbf{x}}_k}\right)^T\Big|_{\tilde{\mathbf{x}}_k} \\ \mathbf{e}_k &= \begin{cases} \left(\mathbf{h}(\tilde{\mathbf{x}}_k) - \mathbf{y}_k\right) & k \in \mathcal{I} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{4.28}$$

where $\mathbf{A}_k$ represented the Jacobian of $\tilde{\mathbf{f}}(\cdot)$ wrt $\tilde{\mathbf{x}}_k$, $\mathbf{B}$ represented the Jacobian of $\mathbf{h}(\cdot)$ wrt $\tilde{\mathbf{x}}_k$, and $\mathbf{e}_k$ represented the error vector. We also set $\mathbf{Q} = \mathbf{I}$ to simplify the derivation but this could easily be changed and had no significant impact on the results of the derivation. Using the *costate* equation in (4.22) and the definitions in (4.28), the analytical form of $\boldsymbol{\lambda}_1$ was defined as

$$\boldsymbol{\lambda}_1 = \left\{ \left(\prod_{j=1}^{N-1} \mathbf{A}_j \right) \boldsymbol{\lambda}_N \right\} + \left\{ \mathbf{B}_1\mathbf{e}_1 + \sum_{i=2}^{N-1} \left(\prod_{m=1}^{i-1} \mathbf{A}_m \right) \mathbf{B}_i\mathbf{e}_i \right\}. \tag{4.29}$$

This equation was partitioned into two main parts as shown by the groups of terms in curly brackets. The first term is the *natural response* of the system and depends on the final condition of the costate. The second term is the *forced response* and is driven by the error between the estimated trajectory and the observed measurements. Based on the conclusions of the analysis of BC1, it was understood that

$$\boldsymbol{\lambda}_N = \begin{bmatrix} \mathbf{0} \\ \lambda_{n+1}(N) \end{bmatrix}, \tag{4.30}$$

where $\mathbf{0} \in \mathcal{R}^{n\times 1}$ is a vector of zeros and $\lambda_{n+1}(N)$ is still an unknown constant. The analytical form of the product of matrices $\mathbf{A}_j$, $j = 1, \ldots, N-1$ was then investigated. A single matrix $\mathbf{A}_k$ takes the form

$$\mathbf{A}_k = \left(\tilde{\mathbf{f}}_{\tilde{\mathbf{x}}_k}\right)^T\Big|_{\tilde{\mathbf{x}}_k} = \left[\begin{array}{cccc} \frac{\partial f_1}{\partial x_1(k)} & \frac{\partial f_2}{\partial x_1(k)} & \cdots & \frac{\partial f_{n+1}}{\partial x_1(k)} \\ \frac{\partial f_1}{\partial x_2(k)} & \ddots & & \vdots \\ \vdots & & \ddots & \\ \frac{\partial f_1}{\partial x_{n+1}(k)} & \frac{\partial f_2}{\partial x_{n+1}(k)} & & \frac{\partial f_{n+1}}{\partial x_{n+1}(k)} \end{array}\right]\Bigg|_{\tilde{\mathbf{x}}_k}. \tag{4.31}$$

It is clear that the elements in the last row are all zero except for the $(n+1, n+1)^{th}$ position which is equal to 1 because $x_{n+1}(k)$ is present only in $f_{n+1}(\cdot)$. The last column contains derivatives of $f_{n+1}(\cdot)$ wrt $x_i(k)$. These terms were defined as

$$\frac{\partial f_{n+1}}{\partial x_i(k)} = 2x_i(k)u(-x_i(k)) \tag{4.32}$$

which is again nonzero only if the constraints have been violated. This general structure of $\mathbf{A}_k$ caused the resulting product matrix $\left(\prod_{j=1}^{N-1}\mathbf{A}_j\right)$ to also have very specific characteristics. This matrix has a 1 in the $(n+1, n+1)$ position and each element in the last column, except for position $(n+1, n+1)$, will be a sum of terms where each term is multiplied by some combination of $u(-x_i(k))$, $i = 1, \ldots, n$; $k = 1, \ldots, N-1$. Thus these values in the last column are controlled by the constraints and are nonzero only if the constraints have been violated. Our focus on only the last column of this matrix becomes apparent when looking at the resulting product of this matrix and the vector $\boldsymbol{\lambda}_N$. Since the vector $\boldsymbol{\lambda}_N$ was defined as having all zeros except for the last element $\lambda_{n+1}(N)$, the product of $\left(\prod_{j=1}^{N-1}\mathbf{A}_j\right)$ and $\boldsymbol{\lambda}_N$ was equal to the last column of $\left(\prod_{j=1}^{N-1}\mathbf{A}_j\right)$ times $\lambda_{n+1}(N)$. Thus we were able to derive the following relationship

$$\left(\prod_{j=1}^{N-1}\mathbf{A}_j\right)\boldsymbol{\lambda}_N = \left(\prod_{j=1}^{N-1}\mathbf{A}_j\right)\left[\begin{array}{c}\mathbf{0}\\ \lambda_{n+1}(N)\end{array}\right] = \left[\begin{array}{c} g_1(\mathbf{u}) \\ g_2(\mathbf{u}) \\ \vdots \\ g_n(\mathbf{u}) \\ 1 \end{array}\right]\cdot\lambda_{n+1}(N) \tag{4.33}$$

where $[g_1(\mathbf{u}), \ldots, g_n(\mathbf{u}),\ 1]^T$ represented the last column of $\left(\prod_{j=1}^{N-1}\mathbf{A}_j\right)$. The vector $\mathbf{u}$ was defined as $\mathbf{u} = [u(-x_1(1)),\ u(-x_2(1)),\ \ldots,\ u(-x_i(k)), \ldots,\ u(-x_n(N-1))]$ and the

function $g_i(\mathbf{u})$ was a sum of terms where each term was the product of some combination of $u(-x_i(k))$, $i = 1, \ldots, n$; $k = 1, \ldots, N-1$. As previously stated, the term $g_i(\mathbf{u})$ was nonzero if any of the constraints had been violated. This means that the natural response was dependent on whether or not the constraints for the problem had been satisfied. We were able to rewrite $\boldsymbol{\lambda}_1$ as

$$\boldsymbol{\lambda}_1 = \left\{ \begin{bmatrix} g_1(\mathbf{u}) \\ g_2(\mathbf{u}) \\ \vdots \\ g_n(\mathbf{u}) \\ 1 \end{bmatrix} \cdot \lambda_{n+1}(N) \right\} + \left\{ \mathbf{B}_1\mathbf{e}_1 + \sum_{i=2}^{N-1} \left(\prod_{m=1}^{i-1} \mathbf{A}_m \right) \mathbf{B}_i\mathbf{e}_i \right\}. \tag{4.34}$$

We then looked at the individual elements of $\boldsymbol{\lambda}_1$ and focused particularly on how the natural response and the error response contributed to the values of these elements. Looking first at the last element of $\boldsymbol{\lambda}_1$, $\lambda_{n+1}(1)$, it was apparent that the natural response was constant. Specifically, the natural response for this element was equal to the unknown constant $\lambda_{n+1}(N)$. The error response of $\lambda_{n+1}(1)$ also had straight forward characteristics. It is important to first note that $x_{n+1}(k)$ does not appear in $\mathbf{h}(\tilde{\mathbf{x}}_k)$ because this is an appended variable that is not observed by the measuring apparatus. Thus, the derivative of $\mathbf{h}(\cdot)$ wrt $x_{n+1}(k)$ was zero. This caused all elements in the $(n+1)^{th}$ row of $\mathbf{B}_k$ to be zero. This in combination with the previously established characteristics of $\left(\prod_{j=1}^{N-1} \mathbf{A}_j\right)$ caused the $(n+1)^{th}$ elements of the error response to equal zero for any value of $\mathbf{e}_k$, $k = 1, \ldots, N-1$, thus causing the error response to have no affect on $\lambda_{n+1}(1)$. It was then clear that the value of $\lambda_{n+1}(1)$ was equal to the unknown constant, $\lambda_{n+1}(N)$. The natural response of the remaining elements, $\lambda_i(1)$, $i = 1, \ldots, n$, depended on whether or not the constraints on the dynamics of the network had been satisfied, with the response equaling to zero if the constraints were satisfied or if $\lambda_{n+1}(N) = 0$. The error response for $\lambda_i(1)$, $i = 1, \ldots, n$ depended on the error between the estimated trajectory and the true trajectory. It also depended on the Jacobian of the system with respect to $\mathbf{x}_k$, $k = 1, \ldots, N-2$. This reflected the sensitivity of the system with respect to small changes in the component concentrations. Using this information, we explored what characteristics were needed from both the natural and error responses in order to satisfy BC2.

Boundary condition 2 was rewritten using the definitions in (4.28) as

$$(\mathbf{A}_0\boldsymbol{\lambda}_1 + \mathbf{B}_0\mathbf{e}_0)^T d\tilde{\mathbf{x}}_0 = (\mathbf{A}_0\boldsymbol{\lambda}_1)^T d\tilde{\mathbf{x}}_0 + (\mathbf{B}_0\mathbf{e}_0)^T d\tilde{\mathbf{x}}_0 = 0 \tag{4.35}$$

The objective now was to establish the conditions under which BC2 would be satisfied given a general system, a general observation function, and general changes in $d\tilde{\mathbf{x}}_0$. For general Jacobian matrices, this could only be accomplished if both summed terms in (4.35) went to zero. We first looked at $d\mathbf{x}_0$ and identified all fixed initial conditions. Any initial condition with a known and fixed value resulted in $dx_i(0) = 0$. This was the case for $x_{n+1}(0)$ which was known, fixed, and equal to zero. This meant that $dx_{n+1}(0) = 0$. The matrices and vectors in the equation were rewritten to reflect the separation of states with fixed initial conditions from state with initial conditions that were not fixed. Looking first at the second term in (4.35), we partitioned $\mathbf{B}_0$ according to the rows corresponding to ICs that were free and ICs that were fixed.

$$\mathbf{B}_0 = \begin{bmatrix} \hat{\mathbf{B}}_0 \\ \mathbf{b}_{n+1}^T \end{bmatrix} \tag{4.36}$$

Here, $\mathbf{b}_{n+1}$ represents the $(n+1)^{th}$ row of $\mathbf{B}_0$ in column vector format and $\hat{\mathbf{B}}_0$ are the remaining rows of $\mathbf{B}_0$ which represents the Jacobian of $\mathbf{h}(\cdot)$ wrt the original state vector $\mathbf{x}_0$. This was used to rewrite the second term in (4.35) as

$$\begin{aligned} (\mathbf{B}_0\mathbf{e}_0)^T d\tilde{\mathbf{x}}_0 &= \begin{bmatrix} \hat{\mathbf{B}}_0\mathbf{e}_0 \\ \mathbf{b}_{n+1}^T\mathbf{e}_0 \end{bmatrix}^T \begin{bmatrix} d\mathbf{x}_0 \\ dx_{n+1}(0) \end{bmatrix} \\ &= \left(\hat{\mathbf{B}}_0\mathbf{e}_0\right)^T d\mathbf{x}_0 + \left(\mathbf{b}_{n+1}^T\mathbf{e}_0\right)^T dx_{n+1}(0) \\ &= \left(\hat{\mathbf{B}}_0\mathbf{e}_0\right)^T d\mathbf{x}_0 = 0 \end{aligned} \tag{4.37}$$

with the last equality resulting from the fact that $dx_{n+1}(0) = 0$. In general, $d\mathbf{x}_0 \neq 0$. Thus (4.37) is true if and only if $\hat{\mathbf{B}}_0\mathbf{e}_0 = 0$. This is true for a general measurement scheme only if the initial error is driven to zero or in other words, only if $\mathbf{e}_0 = 0$. A similar partitioning of $\mathbf{B}_0$ can be made if additional states have fixed and known initial conditions. This would result in either $\mathbf{e}_0$ equaling zero or the resulting partition of $\mathbf{B}_0$ equaling zero in order for this condition to hold true.

We then focused on the first term in (4.35). Again, because the initial value, $x_{n+1}(0)$, is fixed, we again partitioned $\mathbf{A}_0$ similar to how $\mathbf{B}_0$ was partitioned.

$$\mathbf{A}_0 = \begin{bmatrix} \hat{\mathbf{A}}_0 \\ \mathbf{a}_{n+1}^T \end{bmatrix} \tag{4.38}$$

Again, $\mathbf{a}_{n+1}$ represents the $(n+1)^{th}$ row of $\mathbf{A}_0$ in column vector format and $\hat{\mathbf{A}}_0$ represents the first n rows of $\mathbf{A}_0$. The first term in (4.35) was rewritten as

$$\begin{aligned} \left(\mathbf{A}_0\boldsymbol{\lambda}_1\right)^T d\tilde{\mathbf{x}}_0 &= \begin{bmatrix} \hat{\mathbf{A}}_0\boldsymbol{\lambda}_1 \\ \mathbf{a}_{n+1}^T\boldsymbol{\lambda}_1 \end{bmatrix}^T \begin{bmatrix} d\mathbf{x}_0 \\ dx_{n+1}(0) \end{bmatrix} \\ &= \left(\hat{\mathbf{A}}_0\boldsymbol{\lambda}_1\right)^T d\mathbf{x}_0 + \left(\mathbf{a}_{n+1}^T\boldsymbol{\lambda}_1\right)^T dx_{n+1}(0) \\ &= \left(\hat{\mathbf{A}}_0\boldsymbol{\lambda}_1\right)^T d\mathbf{x}_0 = 0 \end{aligned} \tag{4.39}$$

with the last equality again resulting from the fact that $dx_{n+1}(0) = 0$. Again in general, $d\mathbf{x}_0 \neq 0$ because all remaining initial conditions are not fixed. Thus $\hat{\mathbf{A}}_0\boldsymbol{\lambda}_1$ must equal zero in order for (4.39) to be true. The matrix $\hat{\mathbf{A}}_0$ was further partitioned into two parts: the Jacobian matrix of the original system $\mathbf{f}(\cdot)$ and the gradient vector of the appended function $f_{n+1}(\cdot)$ wrt $\mathbf{x}_0$

$$\hat{\mathbf{A}}_0 = \begin{bmatrix} \left(\mathbf{f}_{\mathbf{x}_0}\right)^T\Big|_{\mathbf{x}_0} & \frac{\partial f_{n+1}}{\partial \mathbf{x}_0}\Big|_{\mathbf{x}_0} \end{bmatrix}. \tag{4.40}$$

We rewrote (4.37) as

$$\begin{bmatrix} \left(\mathbf{f}_{\mathbf{x}_0}\right)^T\Big|_{\mathbf{x}_0} & \frac{\partial f_{n+1}}{\partial \mathbf{x}_0}\Big|_{\mathbf{x}_0} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\lambda}}_1 \\ \lambda_{n+1}(N) \end{bmatrix} = \left(\mathbf{f}_{\mathbf{x}_0}\right)^T\Big|_{\mathbf{x}_0} \hat{\boldsymbol{\lambda}}_1 + \frac{\partial f_{n+1}}{\partial \mathbf{x}_0}\Big|_{\mathbf{x}_0} \lambda_{n+1}(N) = 0 \tag{4.41}$$

where $\hat{\boldsymbol{\lambda}}_1 = [\lambda_1(1), \ldots, \lambda_n(1)]$. For this equation to equal zero, it was necessary for

$$\left(\mathbf{f}_{\mathbf{x}_0}\right)^T\Big|_{\mathbf{x}_0} \hat{\boldsymbol{\lambda}}_1 = 0 \tag{4.42}$$

and

$$\frac{\partial f_{n+1}}{\partial \mathbf{x}_0}\Big|_{\mathbf{x}_0} \lambda_{n+1}(N) = 0. \tag{4.43}$$

Looking first at (4.42), we saw that this was true only if $(\mathbf{f}_{\mathbf{x}_0})^T\Big|_{\mathbf{x}_0} = 0$ or if $\hat{\boldsymbol{\lambda}}_1 = 0$. The matrix $(\mathbf{f}_{\mathbf{x}_0})^T\Big|_{\mathbf{x}_0}$ is a matrix whose structure is dependent on the system and is, in general, not equal to zero. Thus this equation would be satisfied only if $\hat{\boldsymbol{\lambda}}_1 = 0$. Recall that $\hat{\boldsymbol{\lambda}}_1$ has two responses that contribute to its overall value; the natural response and the error response. Hence, the only way to make $\hat{\boldsymbol{\lambda}}_1$ equal to zero would be to drive both the natural response and the error response to zero. The error response of $\hat{\boldsymbol{\lambda}}_1$ is equal to zero if the sum of all errors over time, weighted by the overall sensitivity of the the system wrt the measured components, goes to zero. This would result in an estimated trajectory that matches the observed data at the time points of interest. Model error or uncertainty in the measurements may make it impossible to match all of the observed points exactly. In that case driving the sum of errors to zero would result in a trajectory that was the most appropriate fit to the observed data in the least-squares sense. The natural response of $\hat{\boldsymbol{\lambda}}_1$ is driven to zero by ensuring that either the constraints are met or by setting $\lambda_{n+1}(N) = 0$. Setting $\lambda_{n+1} = 0$, however, would result in a trivial solution that ignores the inequality constraints. Setting $\lambda_{n+1}(N) = C \neq 0$, where C is a large, would ensure that the constraints have been satisfied if the natural response is zero. Satisfying the constraints would also ensure that $\frac{\partial f_{n+1}}{\partial \mathbf{x}_0}\Big|_{\mathbf{x}_0} = 0$, which satisfies (4.43) even if $\lambda_{n+1}(N) \neq 0$ was some arbitrary large constant. Hence, these result show that if $\lambda_{n+1}(N) \neq 0$, the results that satisfy BC2 would also provide an estimated trajectory that satisfied all constraints while still providing the most appropriate fit to the observed data points in the least squares sense.

A similar analysis of boundary condition 3 was also made. We restate the equation for BC3 as

$$\sum_{k=0}^{N-1} \left(\mathbf{C}_k \boldsymbol{\lambda}_{k+1}\right)^T d\mathbf{p} = 0 \tag{4.44}$$

where

$$\mathbf{C}_k = \left(\tilde{\mathbf{f}}_{\mathbf{p}}\right)^T\Big|_{\tilde{\mathbf{x}}_k} = \left[\begin{array}{cccc} \frac{\partial f_1}{\partial p_1(k)} & \frac{\partial f_2}{\partial p_1(k)} & \cdots & \frac{\partial f_{n+1}}{\partial p_1(k)} \\ \frac{\partial f_1}{\partial p_2(k)} & \ddots & & \vdots \\ \vdots & & \ddots & \\ \frac{\partial f_1}{\partial p_m(k)} & \frac{\partial f_2}{\partial p_m(k)} & & \frac{\partial f_{n+1}}{\partial p_m(k)} \end{array}\right]\Bigg|_{\tilde{\mathbf{x}}_k}. \tag{4.45}$$

The matrix was partitioned into two parts which were defined as

$$\mathbf{C}_k = \begin{bmatrix} \hat{\mathbf{C}}_k & \frac{\partial f_{n+1}}{\partial \mathbf{p}} \end{bmatrix} \tag{4.46}$$

where $\hat{\mathbf{C}}_k = (\mathbf{f_p})^T\Big|_{\mathbf{x}_k}$, represents the Jacobian of the original system wrt $\mathbf{p}$. Substituting this back into (4.44), we got

$$\left\{ \sum_{k=0}^{N-1} \left(\begin{bmatrix} \hat{\mathbf{C}}_k & \frac{\partial f_{n+1}}{\partial \mathbf{p}} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\lambda}}_{k+1} \\ \lambda_{n+1}(k+1) \end{bmatrix} \right)^T \right\} d\mathbf{p} = 0 \tag{4.47}$$

It was clear that $\frac{\partial f_{n+1}}{\partial \mathbf{p}} = 0$ since $\mathbf{p}$ does not appear in $f_{n+1}(\cdot)$. Thus, (4.47) was rewritten as

$$\left(\sum_{k=0}^{N-1} \left(\hat{\mathbf{C}}_k \hat{\boldsymbol{\lambda}}_{k+1} \right)^T \right) d\mathbf{p} = 0. \tag{4.48}$$

In general, $d\mathbf{p} \neq 0$. This allowed us to conclude that (4.48) was true if and only if

$$\sum_{k=0}^{N-1} \hat{\mathbf{C}}_k \hat{\boldsymbol{\lambda}}_{k+1} = \mathbf{0}. \tag{4.49}$$

This was rewritten as

$$\hat{\mathbf{C}}_0 \hat{\boldsymbol{\lambda}}_1 + \hat{\mathbf{C}}_1 \hat{\boldsymbol{\lambda}}_2 + \cdots + \hat{\mathbf{C}}_{N-1} \hat{\boldsymbol{\lambda}}_N = \begin{bmatrix} \hat{\mathbf{C}}_0 & \hat{\mathbf{C}}_1 & \cdots & \hat{\mathbf{C}}_{N-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\lambda}}_1 \\ \hat{\boldsymbol{\lambda}}_2 \\ \vdots \\ \hat{\boldsymbol{\lambda}}_N \end{bmatrix} = \mathbf{0}. \tag{4.50}$$

This is true for a general set of matrices $\hat{\mathbf{C}}_i$, $i = 0, \ldots, N-1$ only if $\hat{\boldsymbol{\lambda}}_i = 0$, $i = 1, \ldots, N$. We previously concluded that the latter was true only if the error response of $\hat{\boldsymbol{\lambda}}_i$, $i = 1, \ldots, N$ was equal to zero and if the natural response of $\hat{\boldsymbol{\lambda}}_i$, $i = 1, \ldots, N$ was driven to zero. All of this is wrt small changes in the parameter vector $\mathbf{p}$.

We concluded that a solution to the two-point boundary value problem that satisfied these boundary conditions produced component trajectories that provided an appropriate fit to the data that minimized the error between the estimated trajectory and the observed trajectory while satisfying all of the constraints. Hence, this solution provided

the solution to the constrained least-squared estimation problem. A procedure was then needed that effectively found the best $\mathbf{x}_0$ and the best $\mathbf{p}$ that satisfies these boundary conditions. Looking closer at BC2 and BC3, we saw that these were equivalent to the gradient vectors of $\hat{J}_0$ wrt $\mathbf{x}_0$ and wrt $\mathbf{p}$, respectively. We concluded that these could be used in a gradient based search algorithm that iteratively finds the best $\mathbf{x}_0$ and $\mathbf{p}$ that satisfies these conditions. Thus, this approach effectively transformed the constrained least-squares optimization problem with dynamic constraints in (4.12) and (4.13) into an approach that only requires the forward and backward simulation of two dynamic equations and a gradient-based unconstrained optimization approach. An adaptation of an unconstrained conjugate gradient algorithm was used to perform the optimization and is described in the following section.

4.3.2 Conjugate Gradient Optimization Approach

This problem presented a search grid that was represented by the space of unknown initial concentrations and unknown kinetic parameters. A vector was therefor defined that properly represented this list of unknown variables. This vector was used to instantiate an iterative gradient-based search of the most appropriate values in this space. An augmented vector composed of the initial condition vector and the parameter vector was defined as

$$\boldsymbol{\xi} \triangleq \begin{bmatrix} \tilde{\mathbf{x}}_0 \\ \mathbf{p} \end{bmatrix},$$

with the ith guess of this combined vector denoted as $\boldsymbol{\xi}^{(i)}$. Starting at some random initial starting point within an appropriate biological range, the iterative update for a gradient-based search method was defined as

$$\boldsymbol{\xi}^{(i+1)} = \boldsymbol{\xi}^{(i)} + \alpha^{(i)}\mathbf{d}^{(i)}, \tag{4.51}$$

where $\mathbf{d}^{(i)}$ was an appropriate descent direction for the guess at the $(i+1)^{th}$ iteration, and $\alpha^{(i)}$ was a positive step size parameter that adjusts the step along direction $\mathbf{d}^{(i)}$. Descent methods are characterized by the definition of the descent direction $\mathbf{d}^{(i)}$ [130]. The conjugate gradient approach is a useful approach in the vast library of gradient search methods that has distinct advantages over some of the other search methods. It avoids the computation of costly Hessian matrices and avoids redundant search directions, which are issues encountered

when implementing Newton-based methods and traditional steepest descent approaches, respectively. The conjugate gradient approach defines the descent direction vector as [130]

$$\mathbf{d}^{(i)} = -\frac{\partial \hat{J}}{\partial \boldsymbol{\xi}}\bigg|_{\boldsymbol{\xi}^{(i)}} + \beta^{(i)}\mathbf{d}^{(i-1)} \tag{4.52}$$

where $\beta^{(i)}$ is a scalar that ensures the descent vectors $\mathbf{d}^{(i)}$ are mutually orthogonal or conjugate. The conjugate nature of the descent vectors allows the algorithm to avoid redundant search directions. This aids in the convergence of the optimization, ideally allowing the algorithm to converge in l iterations where l is the dimension of $\boldsymbol{\xi}$. Two of the most widely used definitions of $\beta^{(i)}$ for nonlinear conjugate gradient search are the Fletcher-Reeves formula given as

$$\beta^{(i)} = \begin{cases} 0, & i = 0 \\ \dfrac{\left\| \frac{\partial \hat{J}}{\partial \boldsymbol{\xi}}\big|_{\boldsymbol{\xi}^{(i)}} \right\|_2^2}{\left\| \frac{\partial \hat{J}}{\partial \boldsymbol{\xi}}\big|_{\boldsymbol{\xi}^{(i-1)}} \right\|_2^2} & i > 0. \end{cases} \tag{4.53}$$

and the Polak-Ribiére formula given as

$$\beta^{(i)} = \begin{cases} 0, & i = 0 \\ \dfrac{\left(\frac{\partial \hat{J}}{\partial \boldsymbol{\xi}}\big|_{\boldsymbol{\xi}^{(i)}} \right)^T \left(\frac{\partial \hat{J}}{\partial \boldsymbol{\xi}}\big|_{\boldsymbol{\xi}^{(i)}} - \frac{\partial \hat{J}}{\partial \boldsymbol{\xi}}\big|_{\boldsymbol{\xi}^{(i-1)}} \right)}{\left\| \frac{\partial \hat{J}}{\partial \boldsymbol{\xi}}\big|_{\boldsymbol{\xi}^{(i-1)}} \right\|^2}, & i > 0 \end{cases} . \tag{4.54}$$

Both formulas are equivalent for quadratic functions but their use for nonlinear functions depends on the nature of the nonlinear optimization problem. An appropriate value for $\alpha^{(i)}$ was still needed to complete the definition of the search algorithm. The value of the step-size is critical because a value of $\alpha^{(i)}$ that is too large may cause the algorithm to diverge and a value that is too small may cause the algorithm to converge excessively slow. The optimum value of $\alpha^{(i)}$ requires the solution of a 1-dimensional optimization problem

$$\min_{\alpha^{(i)}} \hat{J}\left(\boldsymbol{\xi}^{(i)} + \alpha^{(i)}\mathbf{d}^{(i)}\right). \tag{4.55}$$

When the function $\hat{J}(\cdot)$ is not quadratic, $\alpha^{(i)}$ cannot be determined *a-priori* and requires the implementation of another minimization routine. Running another optimization routine to find the best $\alpha^{(i)}$ at every iteration of the conjugate gradient algorithm can

add greatly to the computational complexity of the overall estimation routine. A more useful approach would be to find a reasonable approximation for $\alpha^{(i)}$ that yields a value that will still exhibit favorable convergence characteristics while saving on computations. For this algorithm, we used what is known as a backtracking algorithm to approximate $\alpha^{(i)}$ [130]. This algorithm approximates $\alpha^{(i)}$ by reducing $\alpha^{(i)}$ from some starting value by a factor of ρ until the Armijo criterion is satisfied

$$\hat{J}\left(\boldsymbol{\xi}^{(i)}+\alpha^{(i)}\mathbf{d}^{(i)}\right)\leq\hat{J}\left(\boldsymbol{\xi}^{(i)}\right)+\sigma\alpha^{(i)}\left(\mathbf{d}^{(i)}\right)^{T}\left.\frac{\partial\hat{J}}{\partial\boldsymbol{\xi}}\right|_{\boldsymbol{\xi}^{(i)}}. \tag{4.56}$$

Here, σ is an constant on the interval $[0,1]$. This approximation often distorts the conjugate nature of the descent vectors, resulting in the algorithm converging in greater than l iterations. This approximation, however, often results in an overall savings in computations.

With everything defined, the optimization was performed as follows. First, an initial guess for $\boldsymbol{\xi}$ was obtained and denoted as

$$\boldsymbol{\xi}^{(0)}=\begin{bmatrix}\tilde{\mathbf{x}}_0^{(0)}\\ \mathbf{p}^{(0)}\end{bmatrix}.$$

The initial guess for the initial condition vector, $\tilde{\mathbf{x}}_0^{(0)}$, and the initial guess for the parameter vector, $\mathbf{p}^{(0)}$, were used to simulate the system in (4.21) to find $\tilde{\mathbf{x}}_k^{(0)}$, $k=0,\ldots,N-1$. As previously stated, $\lambda_{n+1}(N)$ was set to an arbitrary large constant. This value was set large enough to ensure that the inequality constraints were enforced but small enough to make sure the optimization problem was not ill-conditioned. We used this arbitrary constant in combination with the remaining boundary conditions $\lambda_i(N)=0,\ i=1,2,\ldots,n$, the state vectors $\tilde{\mathbf{x}}_k^{(0)}$, and $\mathbf{p}^{(0)}$ to simulate the costate system in (4.22) to find $\boldsymbol{\lambda}_k^{(0)}$, $k=N-1,\ldots,1$. The resulting states, $\tilde{\mathbf{x}}_k^{(0)}$, the resulting costates, $\boldsymbol{\lambda}_k^{(0)}$, and $\mathbf{p}^{(0)}$ were used to calculate

$$\frac{\partial\hat{J}_0}{\partial\boldsymbol{\xi}^{(0)}}=\left.\begin{bmatrix}\frac{\partial\hat{J}_0}{\partial\tilde{\mathbf{x}}_0}\\ \frac{\partial\hat{J}_0}{\partial\mathbf{p}}\end{bmatrix}\right|_{[\tilde{\mathbf{x}}_0^{(0)},\mathbf{p}^{(0)}]}.$$

This partial derivative was then used in (4.52) and (4.53) (or (4.54)) to calculate the descent direction used in the update equation in (4.51). The algorithm was iterated until one of the following stopping criteria were met:

$$\begin{aligned} &1: \quad \left\|\mathbf{d}^{(i)}\right\| < TOL1 \\ &2: \quad \max\left|\hat{J}(\boldsymbol{\xi}^{(i)}) - \hat{J}(\boldsymbol{\xi}^{(i-j)})\right| < TOL2, \\ &\qquad j = 1, \ldots, M. \end{aligned} \tag{4.57}$$

Stopping criterion 1) checked the norm of the descent direction vector which indicated convergence when the norm was equal to zero. Stopping criterion 2) checked if there had been any significant change in the objective function over M consecutive iterations.

4.4 Simulations and Results

The analysis described in this section posed a problem that is in-line with an issue experimental biologists often encounter, where complete dynamics of a pathway are needed and only limited information about the pathway is known. The goals of this analysis were to: 1) verify the functionality of the estimation algorithm using ideal measurements free of uncertainty and 2) assess how bounded uncertainty added to the measurements affected the resulting estimates provided by the algorithm. We used the single-step signal transduction pathway described in Section 4.2 as a model system for testing the estimation algorithm. The limited information associated with this problem included 1) nonlinear equations derived from mass action kinetics [122], 2) simulated discrete-time measurements of the substrate and product concentrations, and 3) knowledge of the initial substrate and product concentrations (obtained from the simulated measurements). Time, costs, and available resources are the primary factors that come into play when collecting measurements from biological systems. Although these factors are usually fixed, a large amount of flexibility is involved in selecting the times at which these samples are taken. For benchmarking purposes, we generated synthetic measurement profiles that reflected common sampling schemes used for collecting data from biological systems where a fixed number of samples were obtained at uniformly and non-uniformly spaced intervals.

First, the system described in (4.5)-(4.8) was simulated using an initial component concentration vector $\mathbf{x}_0^* = [12,\ 12,\ 0,\ 0]^T$ and a kinetic parameter vector $\mathbf{p}^* = [0.18,\ 0.02,\ 0.23]^T$ as demonstrated in Kutalik et al. [122]. A sample period of $\Delta T = 0.01$ min was used and the system was simulated over $N = 1000$ time steps (10 min). The simulated discrete-time measurements of the substrate and the product were obtained at 11

uniformly and 11 non-uniformly space intervals over the simulation period. With the initial concentration of the substrate and product assumed to be known, the algorithm was used to provide estimates of the enzyme and the enzyme-substrate complex initial concentrations along with estimates of all kinetic parameters. The functionality of the algorithm was tested under ideal conditions where uniformly and non-uniformly spaced simulated measurements containing no uncertainty were used. The consistency of the algorithm was tested by executing 100 independent runs at different initial guesses where all initial guesses were within valid biological ranges. Our secondary analysis executed another 100 independent runs, adding varying amounts of bounded uncertainty to the measurements for each run. The variation in the estimates obtained from the algorithm were analyzed in order to gauge the role uncertainty plays in estimating the pathway dynamics. For these analyses, we set the weighting matrix $\mathbf{Q}$ to an identity matrix, $\lambda_{n+1}(N) = 10^4$, and $TOL1 = TOL2 = 1e-8$. The estimation algorithm was implemented using the C programming language.

4.4.1 Simulated Measurements Containing no Uncertainty

The estimation algorithm outlined in Section 4.3 was tested on the single-step signal transduction pathway shown in Figure 4.1 under ideal observation conditions. We posed a practical problem where the enzyme was not easily measured using mass spectroscopy or any other common technique used for quantifying protein concentrations. Consequently, this scenario would hinder our ability to measure the enzyme-substrate complex. Discrete-time measurements of the substrate and the product, however, would be obtainable at either uniform or non-uniform sample times. A table describing the respective sampling profiles used is shown in Table 4.1. Our inability to measure all proteins associated with the kinetic parameters p_1, p_2 and p_3 would make it difficult to acquire knowledge of these parameters *a-priori*. Thus, our objective was to estimate all unmeasured protein dynamics and all kinetics parameters using only the sparse discrete-time measurements from the substrate and the product.

We defined the measurement scheme described above as

$$\mathbf{y}_k = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \mathbf{x}_k, \quad k \in \mathcal{I} \tag{4.58}$$

Table 4.1: Discrete-time measurement profiles for uniformly and non-uniformly spaced measurements

Uniform Samples	Non-uniform Samples
~ 0 min (k = 0)	~ 0 min (k = 0)
~ 1 min (k = 99)	~ 15 sec (k = 24)
~ 2 min (k = 199)	~ 30 sec (k = 49)
~ 3 min (k = 299)	~ 1 min (k = 99)
~ 4 min (k = 399)	$\sim 1\frac{1}{2}$ min (k = 149)
~ 5 min (k = 499)	~ 2 min (k = 199)
~ 6 min (k = 599)	~ 3 min (k = 299)
~ 7 min (k = 699)	~ 4 min (k = 399)
~ 8 min (k = 799)	~ 6 min (k = 599)
~ 9 min (k = 899)	~ 8 min (k = 799)
~ 10 min (k = 999)	~ 10 min (k = 999)

where $\mathbf{y}_k$ corresponded to the simulated measurement vector containing measurements of the substrate and product concentrations at time index k. The set $\mathcal{I}$ contained the points shown in Table 4.1 associated with the measurement profile used. The simulated measurements for the uniform and non-uniform sampling profiles are shown in Figures 4.2 and 4.3, respectively. The reliability of the initial measurement (no uncertainty in observations) allowed us to consider the initial concentration for the substrate and the product as known and fixed. The initial value of the appended state was also fixed ($x_{n+1}(0) = 0$). Given that these values were fixed and constant, the associated columns of the Jacobian matrices, $\tilde{\mathbf{f}}_{\tilde{\mathbf{x}}_0}$ and $\mathbf{h}_{\tilde{\mathbf{x}}_k}$, were removed when calculating the boundary condition in (4.24). The initial guesses for the remaining initial concentrations, $x_2^{(0)}(0)$ and $x_3^{(0)}(0)$, and the initial guess for $\mathbf{p}^{(0)}$ were set to values that were biologically significant. This includes ensuring the initial

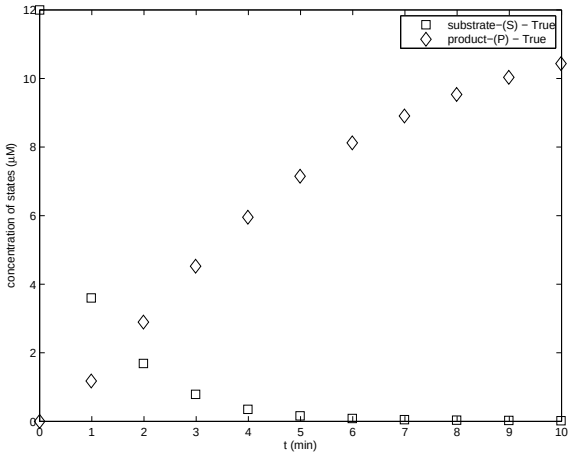

Figure 4.2: Meaurements obtained from uniformly spaced samples.

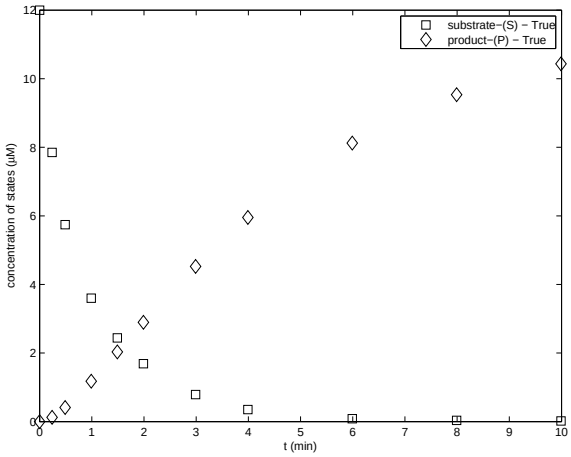

Figure 4.3: Measurements obtained from non-uniformly spaced samples.

concentrations were all positive and ensuring that the initial parameters were within valid ranges. A-priori biological knowledge and knowledge based on network characteristics and topology are often used to help select an initial guess that is within reasonable biological ranges [131]. We used

$$\begin{aligned} x_2^{(0)}(0) &= x_2^*(0) + \delta_1 \\ x_3^{(0)}(0) &= x_3^*(0) + \delta_2 \end{aligned} \tag{4.59}$$

where $\delta_i,\ i = 1, 2$, could be any value on the interval $[0,\ 4]$. The initial parameter vector was initiated to a value that was of the same order of magnitude as the true parameter. Generally, some knowledge about the order of magnitude of the kinetic parameters is known. For example, if the kinetic parameter p_1 is thought to of order of magnitude 10^{-2} to 10^{-1} because of a-priori knowledge of the reaction speeds, the range for the initial guess would be $0.01 \leq p_1^{(0)} \leq 1$. The range of the initial guess can be adjusted to reflect the amount of a-priori knowledge known about the parameter. The initial guesses were then updated according to the estimation algorithm outlined in Section 4.3 until one of the stopping criteria in (4.57) was met. We executed 100 independent runs using different initial guesses, recording the converged results from each run. We used the mean and variance of the 100 converged results to evaluate the algorithms ability to consistently estimate $\mathbf{x}_0^*$ and $\mathbf{p}^*$. We then assessed the algorithms ability to estimate the dynamics of the unmeasured components.

Results: Uniform Sparse Sampling

Table 4.2 summarizes the quality of the estimates obtained using sparse uniform measurements. It shows a comparison between the true elements of $\mathbf{x}_0^*$ and $\mathbf{p}^*$ and the elements of the average converged estimates, $\bar{\mathbf{x}}_0$ and $\bar{\mathbf{p}}$, over 100 independent runs. The average converged estimates, $\bar{\mathbf{x}}_0$ and $\bar{\mathbf{p}}$ were defined as

$$\begin{aligned} \bar{\mathbf{x}}_0 &= \tfrac{1}{100}\textstyle\sum_{i=1}^{100} \hat{\mathbf{x}}_0^*(i) \\ \bar{\mathbf{p}} &= \tfrac{1}{100}\textstyle\sum_{i=1}^{100} \hat{\mathbf{p}}^*(i) \end{aligned}, \tag{4.60}$$

Table 4.2: Sparse Uniform Observations: True Values & Average Estimates

	$\mathbf{x}_0^*,\ \mathbf{p}^*$	$\bar{\mathbf{x}}_0,\ \bar{\mathbf{p}}$	Var $(\hat{\mathbf{x}}_0,\ \hat{\mathbf{p}})$
$x_2(0)$	12.0000000	12.0000078	$7.6898330e-09$
$x_3(0)$	$0.0000000e+00$	0.0000004	$2.5645230e-12$
p_1	0.1800000	0.1799998	$6.5456570e-12$
p_2	0.0200000	0.0200000	$1.0708880e-13$
p_3	0.2300000	0.2300000	$4.4775580e-15$

where $\hat{\mathbf{x}}_0^*(i)$ and $\hat{\mathbf{p}}^*(i)$ represented the converged results for the ith run. The initial concentrations for x_1 and x_4 were observed and assumed known and constant. Thus, these initial concentrations were not included in the comparison. The differences between the true values and the average estimates were very small, with the largest absolute difference of order $O(10^{-6})$. The results also showed very small variation between the estimates over the 100 independent runs, with the largest variation on order $O(10^{-9})$. These results confirmed the algorithm's ability to consistently estimate the true unknown initial concentrations and kinetic parameters of the biochemical pathway when no uncertainty in the measurements were present. This verified the functionality of the algorithm under ideal conditions.

The average estimates of the initial concentrations of the enzyme ($x_2(0)$) and the enzymes-substrate complex ($x_3(0)$) along with the average kinetic parameter estimates provided in Table 4.2 were used to simulate the pathway and produce the dynamics of the component concentrations that were previously unavailable. This includes the unmeasured dynamics of the enzyme and the enzyme-substrate complex over time and the unknown dynamics of the substrate and product at times where the components were not measured. We compared these estimated dynamics against the true dynamics of the systems when simulated using $\mathbf{x}_0^*$ and $\mathbf{p}^*$. The error statistics between the true and estimated trajectories were calculated based on the relative error over time for each state

$$E_{x_i}(k) = \frac{|\bar{x}_i(k) - x_i^*(k)|}{\max_j x_i^*(j)} \times 100 \tag{4.61}$$

where $\bar{x}_i(k)$ represented the trajectory produce by $\bar{\mathbf{x}}_0,\ \bar{\mathbf{p}}$ for the ith state at the kth time step and $x_i^*(k)$ represents the trajectory produced by $\mathbf{x}_0^*,\ \mathbf{p}^*$. The definition of the relative error shown in (4.61) gives a reasonable assessment of the percentage error without generating error results that are overly biased by the relative magnitude of the state. Figure 4.4 gives a visual assessment of the differences between the estimated dynamics and the true dynamics. The two trajectories are almost identical, with almost perfect overlay at the observation sampling points. This error is quantified in Table 4.3 which shows a small average relative error for each state with the largest average error of order $O(10^{-5})$. The small average error highlighted the algorithms ability to accurately estimate the molar concentrations of unoccupied enzymes ($x_2(k)$) and occupied enzymes ($x_3(k)$) over time. It also reinforces the algorithms ability to estimate the molar concentrations of substrate ($x_1(k)$) and product ($x_4(k)$) at times where these components were not measured.

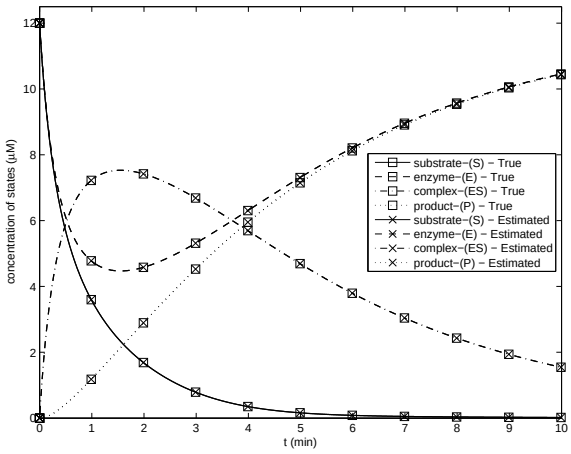

Figure 4.4: Uniform Observations: True vs. Estimated State Trajectories

Results: Non-uniform Sparse Sampling

Table 4.4 summarizes the quality of the estimates obtained using sparse non-uniform measurements. The true initial unknown concentrations and kinetic parameters were again compared to the average estimates over 100 independent runs. The results show that the estimation algorithm was also able to accurately estimate these values using

Table 4.3: Uniform Observations: Relative Error Statistics of State Trajectories

	Max $E_{x_i(k)}$	Avg $E_{x_i(k)}$	Std Dev $E_{x_i(k)}$
x_1	0.0000167	0.0000016	0.0000034
x_2	0.0000750	0.0000653	0.0000041
x_3	0.0000133	0.0000058	0.0000066
x_4	0.0000096	0.0000004	0.0000019

Table 4.4: Non-uniform Observations: True Values & Average Estimates

	$\mathbf{x}_0^*$, $\mathbf{p}^*$	$\bar{\mathbf{x}}_0$, $\bar{\mathbf{p}}$	Var ($\hat{\mathbf{x}}_0$, $\hat{\mathbf{p}}$)
$x_2(0)$	12.0000000	12.0000001	$4.5549440e-09$
$x_3(0)$	$0.0000000e+00$	0.0000002	$1.3322310e-12$
p_1	0.1800000	0.1800000	$2.6207090e-14$
p_2	0.0200000	0.0200000	$1.3460240e-14$
p_3	0.2300000	0.2300000	$1.2679610e-15$

non-uniform observations with the largest absolute difference of order $O(10^{-7})$. The small variation across the 100 runs again shows that the algorithm was consistent in estimating the true values. We see that a comparison between Table 4.2 and 4.4 shows that the non-uniform observations produce slightly better estimates even though the errors for both analyses were small in magnitude.

The average estimates in Table 4.4 were used to produce estimated trajectories. The error statistics betwen the true and estimated trajectories were again based on the relative error formula in (4.61). Figure 4.5 shows plots of the true and estimated trajectories. The very small error between these trajectories resulted in an almost perfect overlay of the true and estimated dynamics. This error is quantified in Table 4.5. The small average errors

Table 4.5: Non-uniform Observations: Relative Error Statistics of State Trajectories

	Max $E_{x_i(k)}$	Avg $E_{x_i(k)}$	Std Dev $E_{x_i(k)}$
x_1	0.0000083	0.0000002	0.0000012
x_2	0.0000083	0.0000006	0.0000021
x_3	0.0000133	0.0000024	0.0000051
x_4	0.0000096	0.0000002	0.0000014

(largest average error of order $O(10^{-6})$) show that the algorithm was able to estimate the unknown dynamics of the pathway using non-uniform observations as well. Comparing these results to those in Table 4.3 does show that the estimates obtained from the uniform measurements produced average errors in the trajectories that were an order of magnitude larger than those produced by the non-uniform measurements. This may be attributed to the non-uniform measurements having an increased number of samples at the transient response of the system.

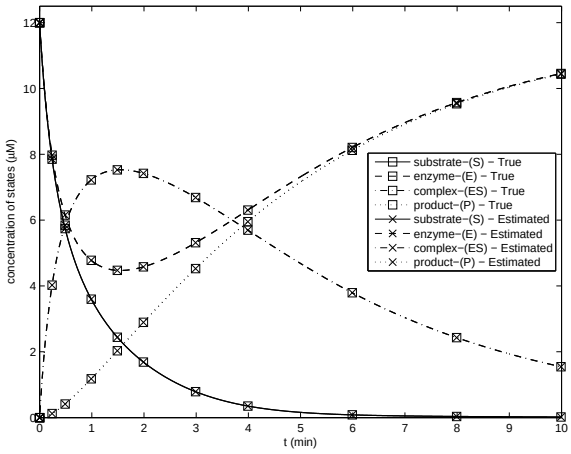

Figure 4.5: Non-uniform Observations: True vs. Estimated State Trajectories

4.4.2 Simulated Measurements Containing Uncertainty

The results shown in Chapter 4.4.1 demonstrated the ability of this algorithm to correctly and consistenly estimate the unmeasured characteristics of the biochemical network without the need for constrained optimization routines. These results were obtained under ideal experimental conditions where the observations were free of uncertainty. Uncertainty, however, is a practical problem that is encountered in almost every real-world estimation application. It often has a negative impact on such algorithms, forcing it to provide an alternate trajectory that is the most adequate fit in the least squares sense to the uncertain observations. Given the practical nature of this issue, it was important to understand how uncertainty in the observations affected the overall behavior and characteristics of the results of the estimation algorithm. This was done by assessing the range of possible estimates and by quantifying the characteristics of the resulting trajectories of these estimates given some range of uncertainty in the observations. The uncertainty in experimental approaches for quantifying the molecular concentrations in biochemical networks often cannot be characterized using parameterizable probability density functions. Uncertainty in protein quantification approaches like the Enzyme-Linked ImmunoSorbent Assay or ELISA are often quantified in terms of some form of bounded error [132]. A useful form of this error is defined as percentage error where the error is given as a percentage of the true value. We defined the percentage error for a given observation as

$$e_{pct} = \frac{|\text{true value} - \text{measure value}|}{\text{true value}} \times 100. \tag{4.62}$$

The uncertain observations used here were defined as having uncertainty up to $\pm e_{pct}\%$. This concept was incorporated into the previous measurement scheme using the following equation

$$\mathbf{y}_k = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \mathbf{x}_k \left(\mathbf{I}_{p\times p} + \frac{1}{100}\boldsymbol{\gamma} \right), \quad k \in \mathcal{I}, k \neq 0, \tag{4.63}$$

where $\mathbf{I}_{p\times p}$ was an identity matrix of size $p \times p$ and $\boldsymbol{\gamma} \in \mathcal{R}^{p\times 1}$ was a vector whose elements were selected within the interval of uncertainty $[-e_{pct}, e_{pct}]$. Equation (4.63) produces observations of the substrate and the product that contained up to $\pm e_{pct}\%$ of uncertainty wrt the true value at that time point. Note that uncertainty was added to all observed concentrations except for the initial concentrations. This was because the initial concen-

trations of the substrate and products were assumed known with little to no uncertainty. This is a reasonable assumption when working with in-vitro experimentation for estimating the conversion dynamics for elementary reactions, particularly when there is direct access to the sources and sinks of the reactions. The confidence in the accuracy of the initial concentrations of the observed internal components for larger networks in-vivo can be adjusted depending on the specific experimental scenario. A total of 100 realizations of the observations containing up to $\pm e_{pct}\%$ of uncertainty was generated. We executed the estimation algorithm for each of the 100 realizations and recorded the range of estimated initial conditions and kinetic parameters. The concentration trajectories generated by this range of estimates were obtained and the characteristics of these trajectories were assessed. The range of uncertainty in these estimated trajectories were quantified based on the following total error signal

$$E_{x_i}^{TOT}(k) = \max_j \hat{x}_i^{*(j)}(k) - \min_j \hat{x}_i^{*(j)}(k) \tag{4.64}$$

where $\hat{x}_i^{*(j)}(k)$, $j = 1, \ldots, 100$ was the concentration of the ith component at time k that resulted from estimates obtained based on the jth realization of the observations with $\pm e_{pct}\%$ measurement error. Thus, $E_{x_i}^{TOT}(k)$ is an uncertainty signal that assesses the range of possible values for the ith component at time k. The maximum of this uncertainty signal, the average of this uncertainty signal, and the spread or standard deviation of this uncertainty signal were used to assess the performance of the estimation algorithm under several error percentage cases. Here, we test scenarios for 3 cases; up to $\pm 2\%$, $\pm 5\%$, and $\pm 7.5\%$ of measurement error. This results in a total error of up to 4%, 10%, and 15%. These three cases were each tested for the two sampling schemes shown in Table 4.1 yielding six cases in all. In the end, all of the results were used to get a quantitative assessment of how increasing uncertainty and changes in the sampling scheme affected our ability to estimate the unmeasured characteristics of the biochemical pathway.

Results: Uniform Sparse Sampling

Case 1: *Uniformly Sampled Observations with up to* 4% *Measurement Error*

This case presents the results we obtained from the estimation algorithm after executing it using 100 realizations of uniformly sampled observations that contained up to

4% of measurement error. The range of possible observations that fell within the interval of ±2% of measurement error for the substrate and product concentrations is illustrated in Figure 4.6.

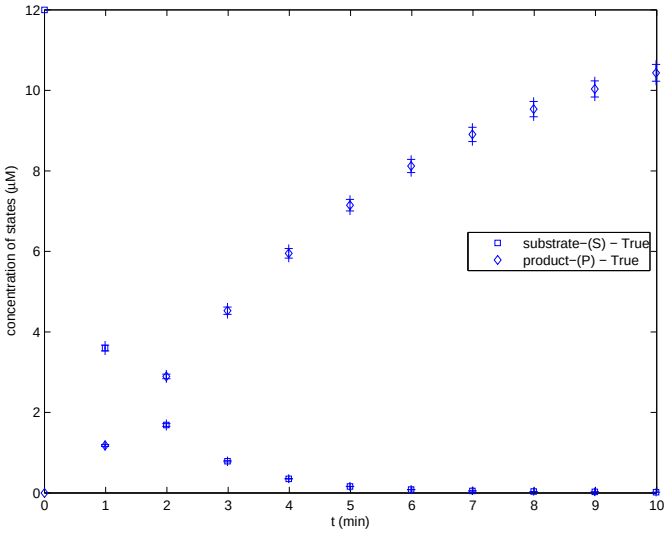

Figure 4.6: Uniform Observations w/ Uncertainty: 4% Error

Executing the estimation algorithm for any valid observation of the substrate and product concentrations within these intervals resulted in estimates of the unknown initial concentrations and the unknown parameters that produced trajectories that best fit the uncertain data points in the least squares sense. Executing the algorithm for 100 different data points within these observation intervals produced a range of possible estimates that could result given the range of uncertain observations. The range of possible estimates for the initial conditions and kinetic parameters is shown in Figure 4.7. This figure indicates the upper and lower bounds of the range of estimates by a plus marker (+) and indicates the true value by an asterisk (∗). Results are shown for the initial concentrations of the enzyme and enzyme-substrate (ES) complex and for the kinetic parameters p_1, p_2, and p_3.

The estimates of the initial enzyme concentration ranged over $\sim 1.66\mu M$ and the estimates of the initial ES complex concentration ranged over $\sim 0.5\mu M$. The variation of the estimates for the initial enzyme concentration was noticeably larger than the variation of the estimates for the initial ES complex concentration with the former being greater than

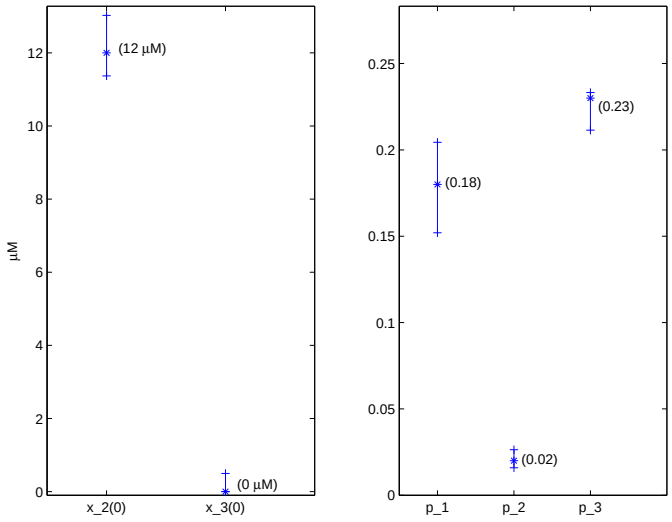

Figure 4.7: Uniform Observations w/ 4% Error: Variation of Estimates

3 times that of the latter. The kinetic rate parameters had variations of ~ 0.05, ~ 0.01, and ~ 0.02 for p_1, p_2, and p_3, respectively. The kinetic parameter p_1 had the largest amount of absolute variation and p_2 had the largest amount of relative variation with respect to its true value. The kinetic parameter p_3 yielded the smallest amount of variation compared to both p_1 and p_2. Note that the true values of both the initial concentrations and the kinetic parameters are within the range of possible estimates shown.

The range of estimated initial conditions and kinetic parameters shown in Figure 4.7 were used to produce sets of concentration trajectories for both the unobservable components (enzyme and ES complex) and the components that were observed at uniformly spaced sparse time points (substrate and product). These sets of concentration trajectories were generated by simulating the network using the 100 estimates obtained from the algorithm. These sets of concentration trajectories for the unobserved and observed components are shown in Figures 4.8a and 4.9a, respectively. These figures give a graphical assessment of the uncertainty associated with estimating the complete dynamics of the biochemical network in the presence of erroneous observations. The upper and lower bounds on the concentrations for a given component are indicated by $(\cdot -)$ lines and the true trajectory is indicated by a solid line with $\square$ or $\diamond$ markers at the appropriate sampling points. Figures 4.8b and 4.9b show the uncertainty signal defined in (4.64). Again, the uncertainty

associated with a specific component is indicated by □ or ◇ markers and the mean of the uncertainty signal over the time of interest is indicated by the line (· · ·). The maximum uncertainty, the average uncertainty, and the spread or standard deviation of the uncertainty for each component in the network is shown in Table 4.6.

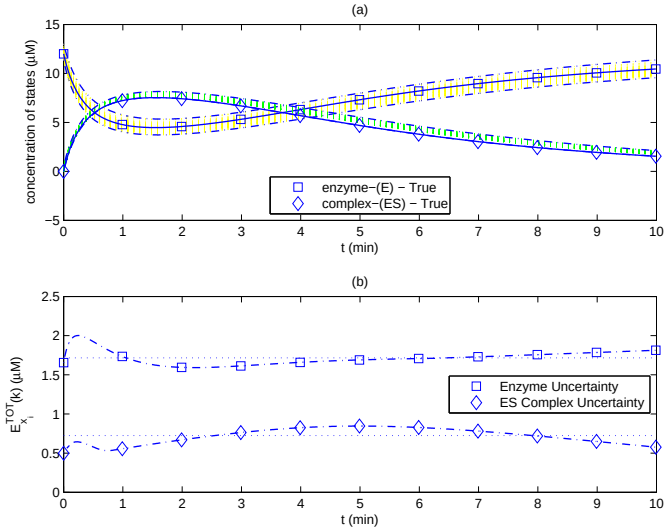

Figure 4.8: Uniform Observations w/ 4% Error: Range of Estimated Unmeasured Trajectories

Table 4.6: Uniform Observations w/ 4% Error: Uncertainty Statistics of Estimated State Trajectories

	Max $E^{TOT}_{x_i(k)}$ (μM)	Avg $E^{TOT}_{x_i(k)}$ (μM)	Std Dev $E^{TOT}_{x_i(k)}$ (μM)
x_1	0.3521139	0.0678185	0.0683826
x_2	2.0016418	1.7165321	0.0870790
x_3	0.8454356	0.7236619	0.0961129
x_4	0.7516428	0.3391079	0.2145150

The widths of the uncertainty interval illustrated in Figure 4.8a show that the uncertainty in the enzyme concentrations over time exhibited slightly greater variation than

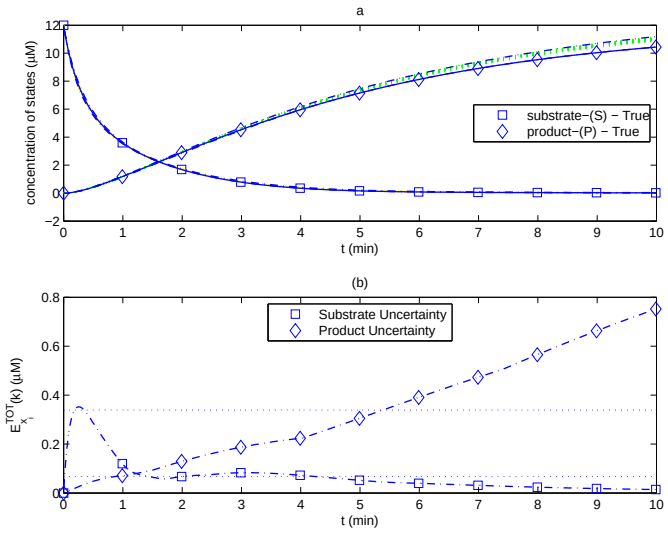

Figure 4.9: Uniform Observations w/ 4% Error: Range of Estimated Measured Trajectories

the concentrations of the ES complex. This was further emphasized by the uncertainty signals for the unobserved components shown in Figure 4.8b which shows that the uncertainty of the enzyme concentration ranged between $1.5\mu M$ and $2\mu M$ and the uncertainty of the ES complex concentration ranged between $0.5\mu M$ and $1\mu M$. The difference in uncertainty between the two components was a direct consequence of the larger variation in the initial enzyme concentration estimate.

As expected, Figure 4.9 shows that the observed components have a smaller amount of uncertainty associated with their resulting estimated trajectories than was the case for the components that were unobserved. Increasing widths in the range of estimated concentration trajectories for the product shows that the algorithm had difficulty correctly estimating the product concentrations as the simulation got further away from the initial starting point. This was further emphasized by the almost linear increase in uncertainty shown in Figure 4.9b for the product concentrations. The results in Figure 4.9b also illustrates that the algorithm had increased difficulty in providing accurate estimates of the substrate trajectories for early times, particularly between 0 and 1 minute. Table 4.6 shows that the average uncertainty for the enzyme and the ES complex was in fact close to the variation in the initial concentration estimates seen in Figure 4.7. The small spread for these components also indicates that the uncertainty was relatively constant over

the simulation time. The uncertainty for the observed components had slightly different characteristics. Although the mean uncertainty for these components were small, the large maximum uncertainty and the relatively large standard deviation highlighted the fact that the uncertainty for these components changed significantly over time. Particularly, the large standard deviation of $\sim .21\mu M$ for the product uncertainty quantified the monotonically increasing uncertainty seen in Figure 4.9b.

Case 2: *Uniformly Sampled Observations with up to* 10% *Measurement Error*

This case presents the results obtained from executing the estimation algorithm on 100 realizations of uniformly sampled observations that contained up to 10% of measurement error. The range of possible observations that fell within the interval of $\pm 5\%$ of measurement error for the substrate and product concentrations is illustrated in Figure 4.10.

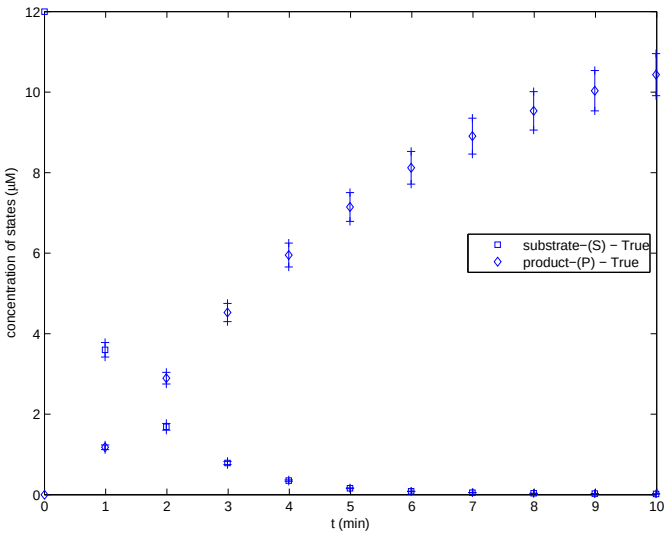

Figure 4.10: Uniform Observations w/ Uncertainty: 10% Error

A range of estimates of the initial concentrations of the enzyme and ES complex as well of kinetic parameters p_1, p_2, and p_3 were obtained based on 100 different realizations of observed data points that fell within the intervals of the measurement uncertainty is shown in Figure 4.11. As expected, the range in these estimates were noticeably larger than those observed in the results for Case 1 due to the increase in measurement error. The estimates

of the initial enzyme concentration ranged over $\sim 5.5\mu M$ and the estimates of the initial ES complex concentration ranged over $\sim 2\mu M$. The increase in the range of the initial concentration estimates was $3-4$ times greater than the ranges seen in Case 1. The range in these estimate indicated once again that the variation associated with the initial ES complex concentration was smaller than the variation for the initial enzyme concentration. The estimates of the kinetic rate parameters had a range of ~ 0.15, ~ 0.03, and ~ 0.08 for p_1, p_2, and p_3, respectively. These estimates also had a similar increase in variation with p_1 and p_2 increasing by ~ 3 times and p_3 increasing by ~ 4 times the variation seen in Case 1.

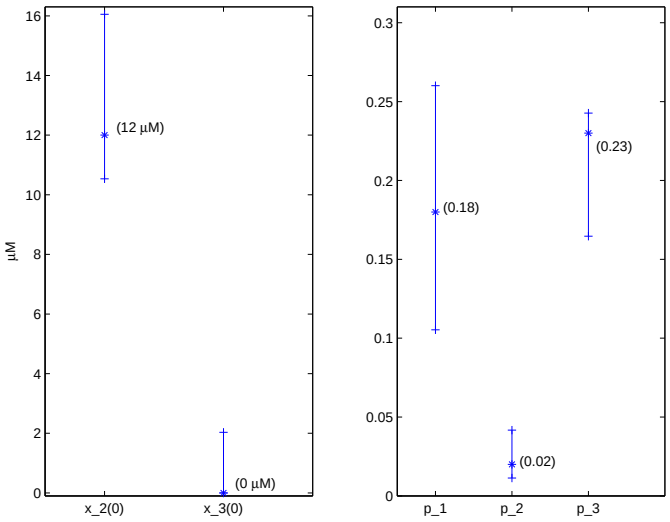

Figure 4.11: Uniform Observations w/ 10% Error: Variation of Estimates

The range of estimates in Figure 4.11 was used to produce the sets of concentration trajectories seen in Figures 4.12a and 4.13a for both the unobserved and observed components, respectively. Figures 4.12b and 4.13b show the uncertainty associated with the sets of concentration trajectories for each component over the simulation time. The maximum uncertainty, average uncertainty, and standard deviation of the uncertainty signal for each component in the network is shown in Table 4.7.

Figure 4.12a shows an increase in the uncertainty associated with the unobserved components. The overall uncertainty associated with the estimated enzyme concentration trajectories was again larger than the uncertainty associate with the ES complex, with the uncertainty for the enzyme ranging between $5.5\mu M$ and $6.5\mu M$ and the uncertainty for the

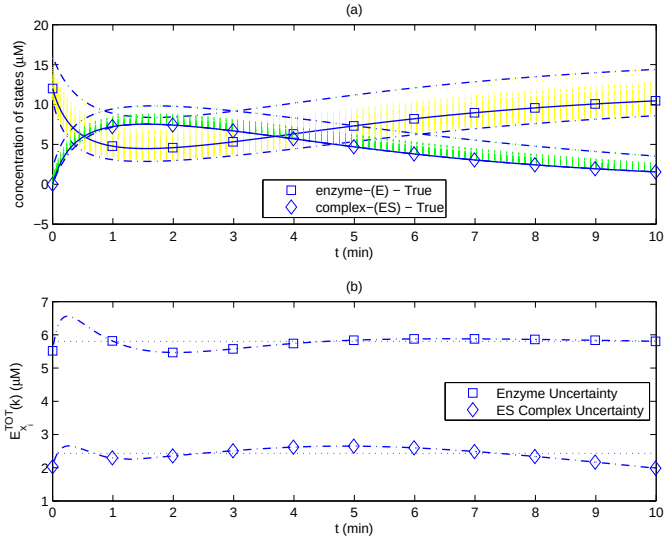

Figure 4.12: Uniform Observations w/ 10% Error: Range of Estimated Unmeasured Trajectories

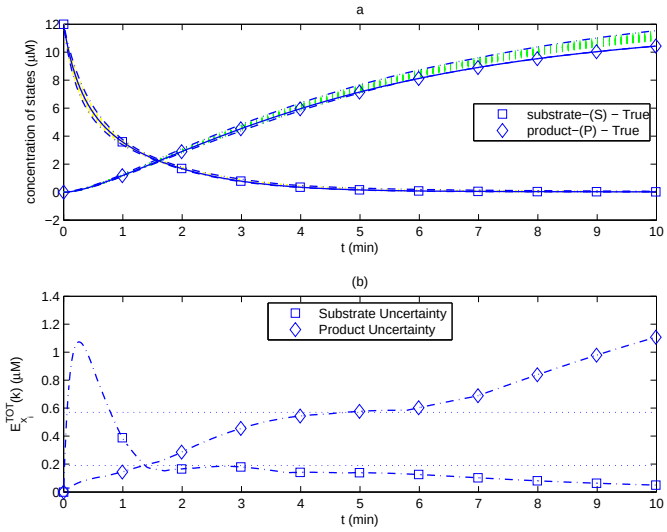

Figure 4.13: Uniform Observations w/ 10% Error: Range of Estimated Measured Trajectories

Table 4.7: Uniform Observations w/ 10% Error: Uncertainty Statistics of Estimated State Trajectories

	Max $E^{TOT}_{x_i(k)}$ (μM)	Avg $E^{TOT}_{x_i(k)}$ (μM)	Std Dev $E^{TOT}_{x_i(k)}$ (μM)
x_1	1.0732661	0.1898327	0.2097330
x_2	6.5609535	5.8049833	0.2074041
x_3	2.6568115	2.4344054	0.1813486
x_4	1.1067879	0.5692848	0.2829216

ES complex ranging between $2\mu M$ and $2.7\mu M$. The increase in the average uncertainty for these components shown in Table 4.7 was about 3.5 times the average uncertainty seen in Case 1. This was in line with the relative increase in the range of initial concentration estimates that were seen. The small standard deviation in the uncertainty signals for both the enzyme and the ES complex along with the characteristics of the uncertainty signal shown in Figure 4.12b indicated that the uncertainty for these components are relatively constant across all time points of interest.

The set of concentration trajectories for the observed components continued to have an overall uncertainty that was smaller than the uncertainty for the unobserved components, as is shown in Figures 4.13a and 4.13b. The uncertainty associated with the product concentration trajectories still had a relatively large spread with the largest standard deviation of all uncertainty signals of $0.28\mu M$. Figure 4.13b shows that the uncertainty signal for the product concentrations did continue to grow monotonically as time increases. Figure 4.13b also illustrates that the substrate had a more dramatic increase in the concentration trajectory uncertainty at early times, again between the 0 and 1 minute time points. In fact, the substrate had a maximum uncertainty that was almost equal to the maximum uncertainty associated with the product. This was slightly different than the characteristics seen for Case 1 where the maximum uncertainty associated with the product was more than two times that of the maximum uncertainty associated with the substrate.

Case 3: *Uniformly Sampled Observations with up to* 15% *Measurement Error*

This case presents the results obtained from executing the estimation algorithm on

100 realizations of uniformly sampled observations that contained up to 15% of measurement error. The range of possible observed data points that fell within this error interval of ±7.5% for the substrate and product concentration is illustrated in Figure 4.14

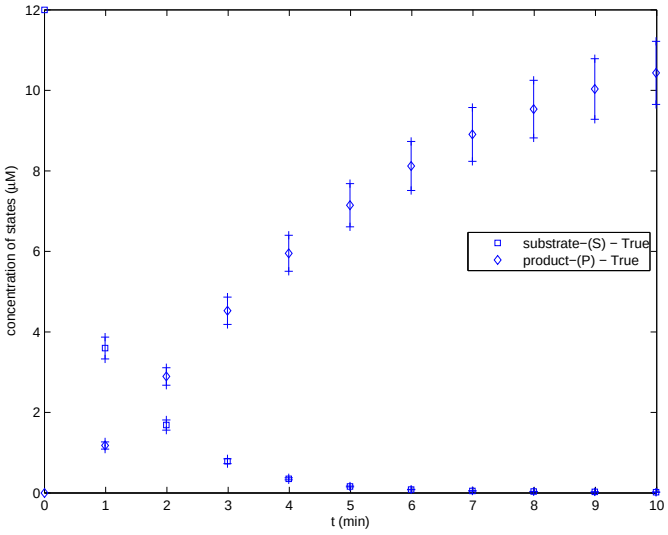

Figure 4.14: Uniform Observations w/ Uncertainty: 15% Error

A range of estimates of the initial concentrations of the enzyme and ES complex as well as the kinetic parameters p_1, p_2, and p_3 were obtained based on the 100 different realizations of the observed data points that fell within the intervals of measurement uncertainty is shown in Figure 4.15. The estimates of the initial enzyme concentration ranged over $\sim 9.12\mu M$ and the estimates of the initial ES complex concentration ranged over $\sim 2.85\mu M$. The ranges for both estimates were about 5.5 times greater than the ranges observed in Case 1. The estimates of the kinetic parameters had ranges of ~ 0.27, ~ 0.037, ~ 0.096, for p_1, p_2, and p_3, respectively. This translated to an increase of $4-5$ times the ranges seen in Case 1 for the kinetic rate parameters.

The range of estimates in Figure 4.15 were used to produce the sets of concentration trajectories seen in Figures 4.16a and 4.17a for both the unobserved and observed components, respectively. Figures 4.16b and 4.17b show the uncertainty associated with the sets of concentration trajectories for each component over the simulation time. The maximum uncertainty, average uncertainty, and standard deviation of the uncertainty signal for

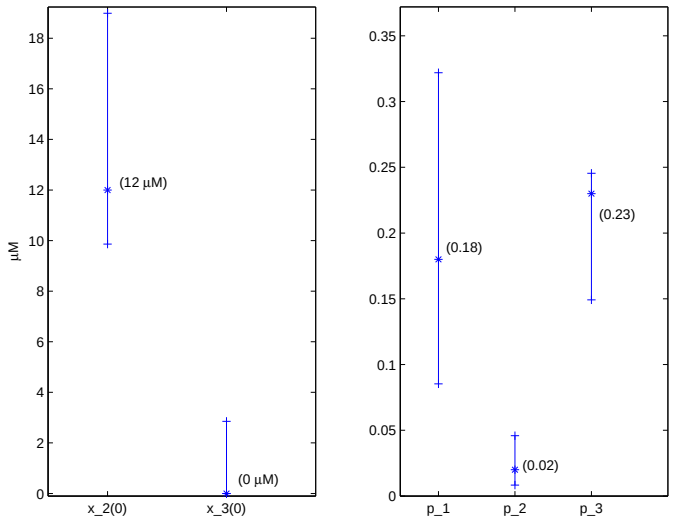

Figure 4.15: Uniform Observations w/ 15% Error: Variation of Estimates

each component in the network is shown in Table 4.8.

Table 4.8: Uniform Observations w/ 15% Error: Uncertainty Statistics of Estimated State Trajectories

	Max $E^{TOT}_{x_i(k)}$ (μM)	Avg $E^{TOT}_{x_i(k)}$ (μM)	Std Dev $E^{TOT}_{x_i(k)}$ (μM)
x_1	1.4708617	0.2596060	0.2939241
x_2	10.5364197	9.3580747	0.2975133
x_3	3.3690266	3.1122293	0.2169677
x_4	1.2542764	0.7553656	0.3109945

Figure 4.16a shows a dramatic increase in the overall uncertainty associated with the unobserved components. The sets of concentration trajectories for the enzymes continued to have a significantly larger uncertainty than the ES complex with the uncertainty for the enzyme ranging between $9\mu M$ and $10.5\mu M$ and the uncertainty for the ES complex ranging between $2.25\mu M$ and $3.5\mu M$. The average uncertainty for the enzyme concentration trajectories in Table 4.8 had an increase of 5.5 times the average uncertainty seen in Case 1.

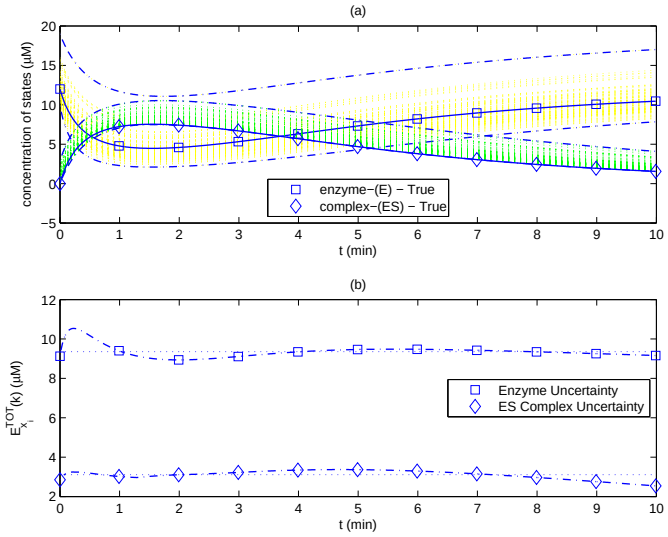

Figure 4.16: Uniform Observations w/ 15% Error: Range of Estimated Unmeasured Trajectories

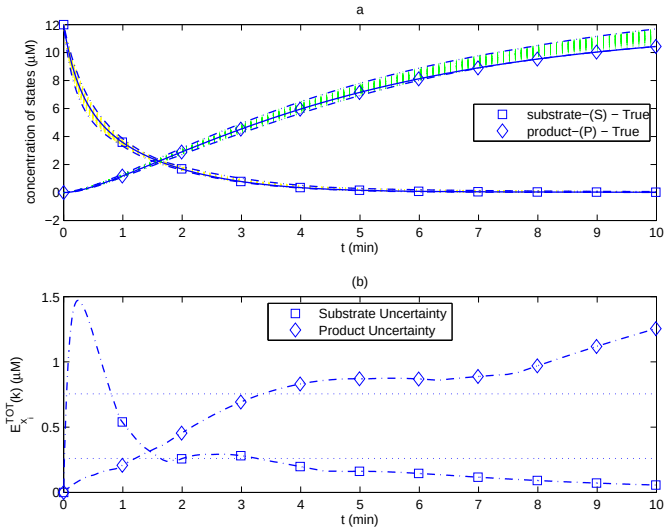

Figure 4.17: Uniform Observations w/ 15% Error: Range of Estimated Measured Trajectories

The increase in the average uncertainty for the ES complex concentration trajectories was slightly smaller with an increase of about 4.5 times. The relatively small standard deviation in the uncertainty signal for both the enzyme and the ES complex along with the characteristics of the uncertainty signal shown in Figure 4.16b indicated that the uncertainty for those components was relatively constant across all time points of interest.

The sets of concentration trajectories for the observed components continued to have an overall uncertainty that was smaller than the uncertainty seen for the unobserved components, as shown in Figure 4.17a and 4.17b. The characteristics of the uncertainty signal of these components appeared to change significantly with the increase in measurement error. The linear increase in the uncertainty associated with the product concentrations appeared to be somewhat piecewise with a constant uncertainty between 4 and 7 minutes. This characteristic was further emphasized by a standard deviation that was not as large with respect to the mean uncertainty. The increase in the uncertainty of the substrate estimate was significantly larger, particularly between the 0 minute and the 1 minute time points. We also saw that the standard deviation for the substrate uncertainty was greater than the average uncertainty, which further emphasizes the large variation in the uncertainty for this component. We also note that Table 4.8 also showed that the maximum uncertainty for the substrate estimates was greater than the maximum uncertainty for the product estimates which was a characteristic that had not been seen in the previous cases.

Results: Non-uniform Sparse Sampling

Case 4: *Non-uniformly Sampled Observations with up to* 4% *Measurement Error*

This case presents the results we obtained from executing the estimation algorithm on 100 realizations of non-uniformly sampled observations that contained up to 4% of measurement error. These samples were obtained based on the non-uniform sampling scheme in Table 4.1. The range of possible observed data points that fell within the error interval of $\pm 2\%$ for the substrate and the product concentrations is illustrated in Figure 4.18.

A range of estimates of the initial concentrations of the enzyme and ES complex as well as the kinetic parameters p_1, p_2, and p_3 that were obtained based on the 100 different realizations of the observed data points that fell within the intervals of measurement uncertainty is shown in Figure 4.19. The estimates of the initial enzyme concentration ranged over $\sim 1.68 \mu M$ and the estimates of the initial ES complex concentration ranged

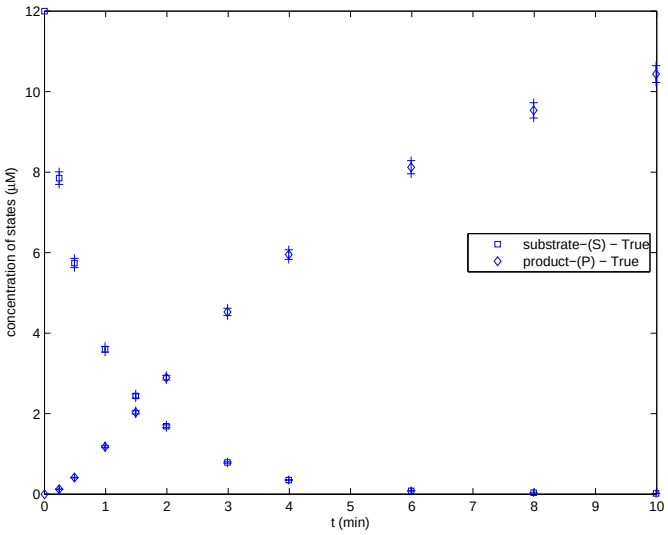

Figure 4.18: Non-uniform Observations w/ Uncertainty: 4% Error

over $\sim 0.64\mu M$. The range for the initial enzyme concentration was similar to the range seen for Case 1 where 4% of measurement error was added to uniformly sampled observations. The range for the estimate of the initial ES complex concentration was $\sim 0.14\mu M$ greater than the range seen in Case 1, which is small compared to the true value. The estimates of the kinetic parameters had a range of ~ 0.04, ~ 0.02, and ~ 0.03 for p_1, p_2, and p_3, respectively. Comparing these ranges to the ranges seen in Case 1, we see that the range for p_1 was reduce but the ranges for p_2 and p_3 both increased by 0.01. This was considered a significant increase for p_2 given the small order of magnitude for the true value.

The range of valid estimates shown in Figure 4.19 were used to produce the sets of concentration trajectories seen in Figures 4.20a and 4.21a for both the unobserved and observed components, respectively. Figures 4.20b and 4.21b show the uncertainty associated with the sets of concentrations trajectories for each components over the entire simulation time. Table 4.9 lists the maximum uncertainty, average uncertainty, and standard deviation of the uncertainty signals in Figures 4.20b and 4.21b.

The widths of the uncertainty intervals illustrated in Figure 4.20a once again show that the uncertainty associated with the enzyme concentration trajectories is greater than the uncertainty associated with the ES complex. This point was clearly illustrated in Figure

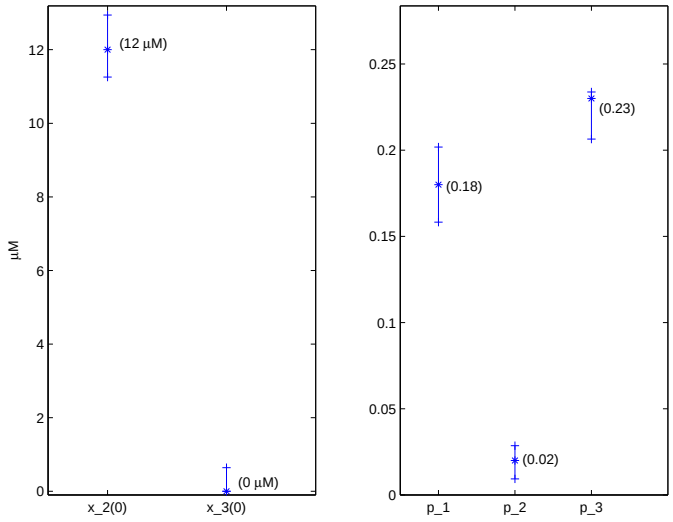

Figure 4.19: Non-uniform Observations w/ 4% Error: Variation of Estimates

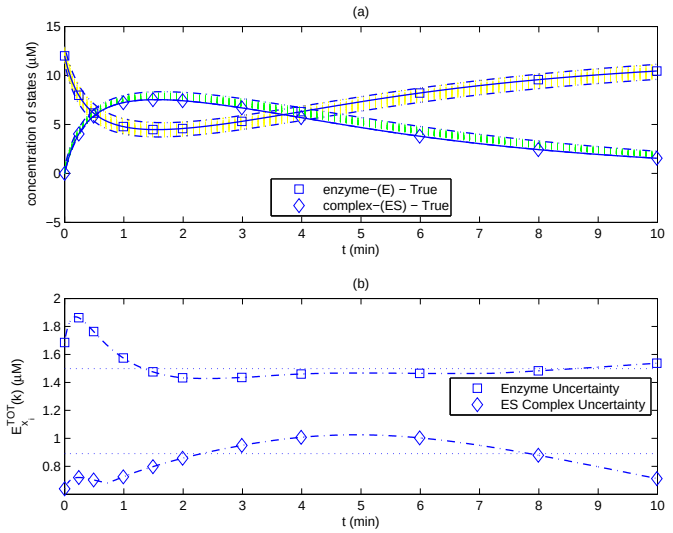

Figure 4.20: Non-uniform Observations w/ 4% Error: Range of Estimated Unmeasured Trajectories

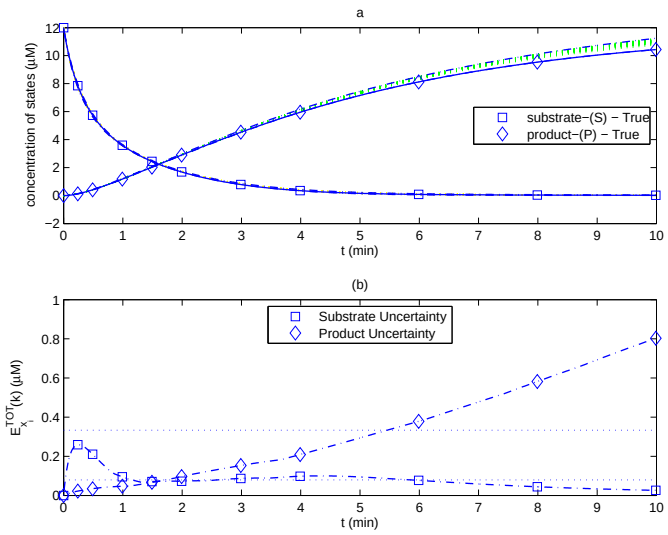

Figure 4.21: Non-uniform Observations w/ 4% Error: Range of Estimated Measured Trajectories

Table 4.9: Non-uniform Observations w/ 4% Error: Uncertainty Statistics of Estimated State Trajectories

	Max $E^{TOT}_{x_i(k)}$ (μM)	Avg $E^{TOT}_{x_i(k)}$ (μM)	Std Dev $E^{TOT}_{x_i(k)}$ (μM)
x_1	0.2591635	0.0793267	0.0437712
x_2	1.8686553	1.4988599	0.0902957
x_3	1.0253696	0.8910777	0.1097281
x_4	0.8028519	0.3333216	0.2352149

4.20b where the uncertainty in the enzyme concentration trajectories ranged from $1.4\mu M$ to $1.9\mu M$ and the uncertainty in the ES complex concentration trajectories ranged from $0.6\mu M$ to $1\mu M$. These results were comparable to the ranges seen for Case 1. The maximum uncertainty for the enzyme and the average uncertainty for both the enzyme and the ES complex were slightly smaller for this case than for Case 1 but the reduction was not overwhelmingly significant. The uncertainty associated with the observed components, which is shown in Figures 4.21a and 4.21b, was still relatively small compared to the uncertainty seen for the other components. The characteristics for the uncertainty associated with these components were slightly different than those seen in Case 1. Specifically, the uncertainty associated with the product concentration trajectories seemed to increase at a faster rate as time increased. This signified a decrease in our ability to estimate the product at later times in the simulation. There was however a significant reduction in the uncertainty associated with the substrate between the 0 minute and 1 minute time points. Specifically, there was a decrease in the maximum uncertainty of $0.1\mu M$ during this time. This may be attributed to the increase in the number of samples in this area.

Case 5: *Non-uniformly Sampled Observations with up to* 10% *Measurement Error*

This case presents the results obtained from executing the estimation algorithm on 100 realizations of non-uniformly sampled observations that contained up to 10% of measurement error. The range of possible observed data points that fell within this error interval of $\pm 5\%$ for the substrate and product concentrations is illustrated in Figure 4.22.

The range of estimates for the initial concentrations of the enzymes and ES complex as well as the kinetic rate parameters p_1, p_2, and p_3 that were obtained based on 100 different realizations of observed data points that fell within the interval of measurement uncertainty is shown in Figure 4.23. The estimates of the initial enzyme concentration ranged over $\sim 3.8\mu M$ and the estimates of the initial ES complex ranged over $\sim 2\mu M$. The range of the estimate of the initial enzyme concentration was significantly smaller than the increase seen in Case 2, where 10% of measurement error was added to uniformly sampled observations. The range of the estimates for the initial ES complex concentration was equal to the range seen in Case 2. The estimates of the kinetic rate parameters show a range of ~ 0.094, ~ 0.046, and ~ 0.07 for p_1, p_2, and p_3, respectively. Comparing these ranges with the results seen in Case 2, we saw that the range in the estimates of p_1 and p_3 were smaller for this case while the range in the estimates of p_2 increased. This significant decrease in the

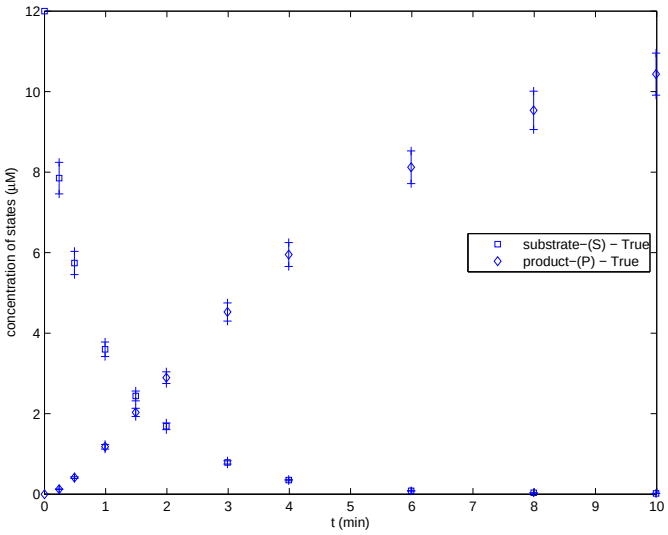

Figure 4.22: Non-uniform Observations w/ Uncertainty: 10% Error

range of the forward conversion rate parameter, p_1, and the estimate of the initial enzyme concentration helped to later improve our ability to estimate both the enzyme and the ES complex concentration trajectory.

The range of the resulting estimates shown in Figure 4.23 were used to produce the sets of concentration trajectories seen in Figure 4.24a and 4.25a for both the unobserved and observed components, respectively. Figures 4.24b and 4.25b show the uncertainty associated with the set of concentration trajectories for each component over the entire simulation time. Table 4.10 lists the maximum uncertainty, average uncertainty, and standard deviation of the uncertainty signals in Figure 4.24b and 4.25b.

Figure 4.24a shows that the overall uncertainty associated with both unobserved components increased proportionally due to the increase in measurement error. Figure 4.24b shows that the uncertainty associated with the enzyme concentration trajectories ranged between $3.25\mu M$ and $4.25\mu M$ and the uncertainty associated with the ES complex concentration trajectories ranged between $1.75\mu M$ and $2.5\mu M$. We see that the upper and lower bounds on the uncertainty for the enzyme concentrations were noticeably smaller than the upper and lower bounds on the uncertainty for the enzyme shown in Case 2 resulting in more accurate estimates of the enzyme concentration trajectories. The maximum uncer-

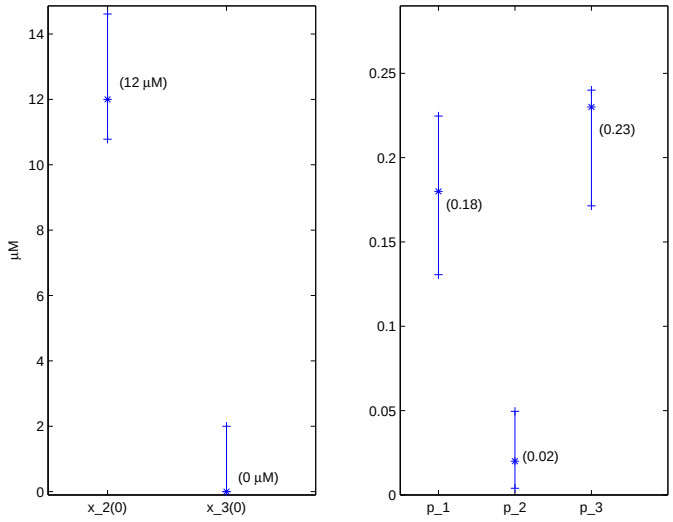

Figure 4.23: Non-uniform Observations w/ 10% Error: Variation of Estimates

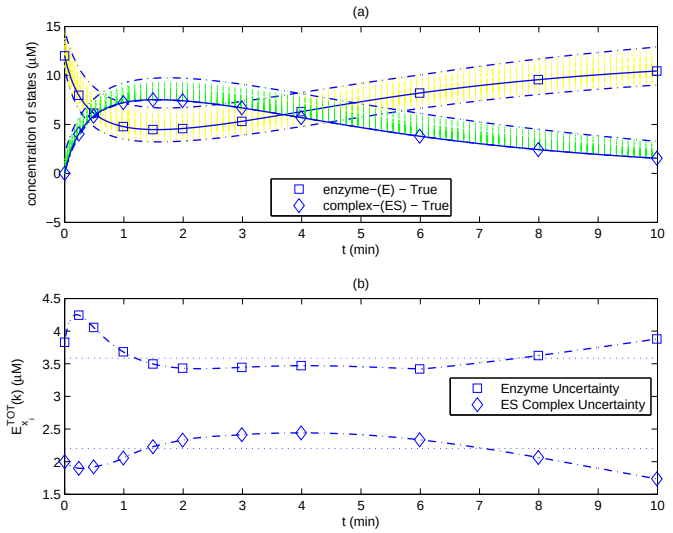

Figure 4.24: Non-uniform Observations w/ 10% Error: Range of Estimated Unmeasured Trajectories

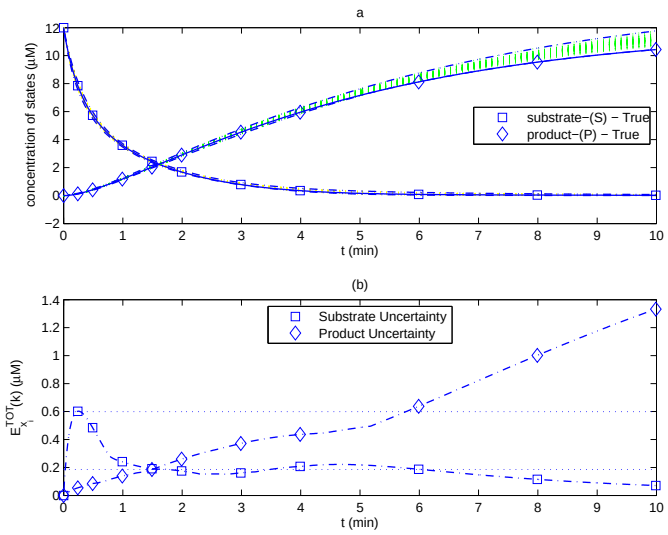

Figure 4.25: Non-uniform Observations w/ 10% Error: Range of Estimated Measured Trajectories

Table 4.10: Non-uniform Observations w/ 10% Error: Uncertainty Statistics of Estimated State Trajectories

	Max $E^{TOT}_{x_i(k)}$ (μM)	Avg $E^{TOT}_{x_i(k)}$ (μM)	Std Dev $E^{TOT}_{x_i(k)}$ (μM)
x_1	0.6027067	0.1854317	0.0976506
x_2	4.2510576	3.5847814	0.1938339
x_3	2.4408713	2.2005079	0.2135438
x_4	1.3326252	0.5998849	0.3707996

tainty and the average uncertainty associated with the enzyme shown in Table 4.10 was more than $2\mu M$ smaller than the maximum and average uncertainty seen in Table 4.7 for Case 2. These results show an increase in our ability to better estimate the trajectories of the enzyme concentrations. The upper and lower bounds on the uncertainty of the ES complex trajectories as well as the maximum and average uncertainty were also reduced, but not by as significant an amount.

The uncertainty associated with the observed components is shown in Figures 4.25a and 4.25b. The characteristics of the uncertainty associated with these variables were again slightly different than the characteristics seen in Case 2. The uncertainty associated with the product concentrations still monotonically increased as the simulation time increased but it appeared to increase at a faster rate. This caused the maximum uncertainty for the product seen in Table 4.10 to be larger for this case than in Case 2. The uncertainty of the substrate, however, was significantly less. This was attributed to smaller uncertainty in the transient region between 0 and 1 minute. Table 4.10 showed the maximum uncertainty in this region to be $\sim 0.6\mu M$ as opposed to the $1.07\mu M$ seen in Table 4.7. This significant drop in the maximum uncertainty also reduced the spread or standard deviation of this uncertainty resulting in a more accurate estimate of the substrate concentrations.

Case 6: *Non-uniformly Sampled Observations with up to* 15% *Measurement Error*

This case presents the results obtained from executing the estimation algorithm on 100 realizations of non-uniformly sampled observations that contained up to 15% of measurement error. The range of possible observed data points that fell within this error interval of $\pm 7.5\%$ for the substrate and product concentrations is illustrated in Figure 4.26.

The ranges of estimates for the initial concentrations of the enzyme and ES complex as well as the kinetic rate parameter p_1, p_2, and p_3 that were obtained based on the 100 realizations of observed data points that fell within the intervals of measurement uncertainty is shown in Figure 4.27. The estimates of the initial enzyme concentration ranged over $\sim 7.24\mu M$ and the estimates of the initial ES complex concentration ranged over $3\mu M$. The range of the initial enzyme concentration continued to be significantly less that the range seen in Case 3 where 15% of measurement error was added to uniformly samples observations. This case did, however, produce a range for the estimate of the initial ES complex concentration that was larger than the range seen in Case 3. The estimates of the kinetic rate parameters show a range of 0.14, 0.055, and 0.092 for p_1, p_2, and p_3, respectively.

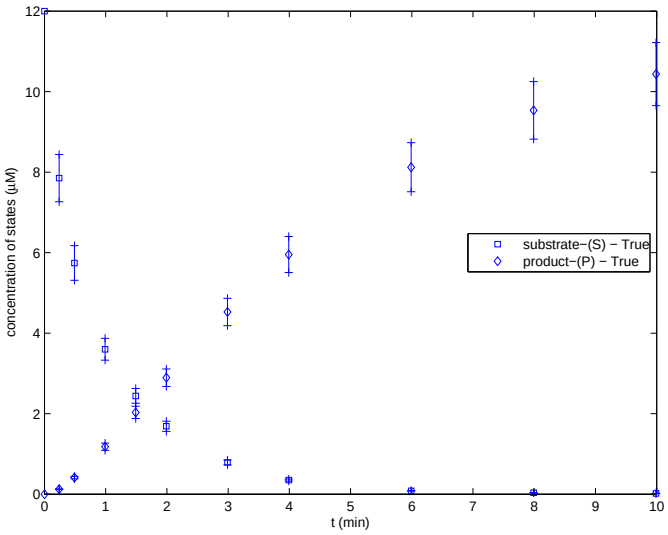

Figure 4.26: Non-uniform Observations w/ Uncertainty: 15% Error

Comparing the results to the ranges seen in Case 3, we found that the range for p_1 was again significantly smaller and the range for p_3 was slightly smaller. The range for p_2 did show a significant increase.

The range of the resulting estimates shown in Figure 4.27 were used to produce the sets of concentration trajectories seen in Figure 4.28a and 4.29a for both the unobserved and observed components, respectively. Figures 4.28b and 4.29b show the uncertainty associated with the set of concentration trajectories for each component over the entire simulation time. Table 4.11 lists the maximum uncertainnty, average uncertainty, and standard deviation of the uncertainty signal in Figures 4.28b and 4.29b.

Figure 4.28a does show that the overall uncertainty for the unobserved components did increase due to the increase in measurement error. Figure 4.28b shows that the uncertainty associated with the enzyme concentration trajectories ranged between $6.5\mu M$ and $8\mu M$ and the uncertainty associated with the ES complex concentration trajectories ranged between $2.75\mu M$ and $3.75\mu M$. The uncertainty associated with the enzyme concentrations was clearly smaller for this case when compared to the results in Case 3. In fact, the average uncertainty associated with the enzyme (Table 4.11) is about $2.5\mu M$ smaller. The uncertainty associated with the ES complex concentrations did see a small increase.

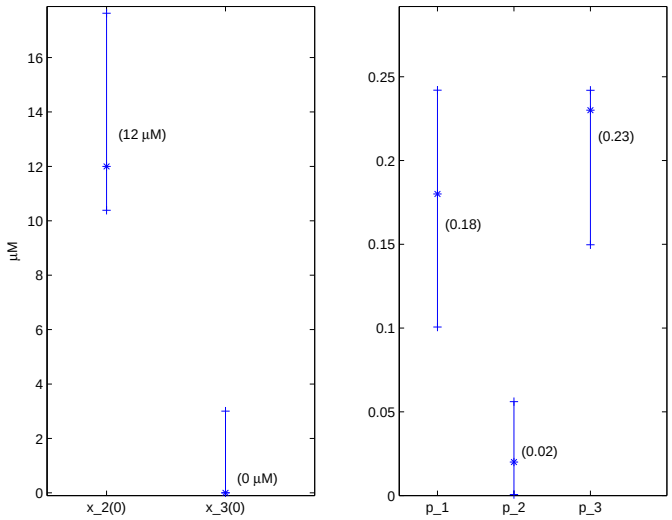

Figure 4.27: Non-uniform Observations w/ 15% Error: Variation of Estimates

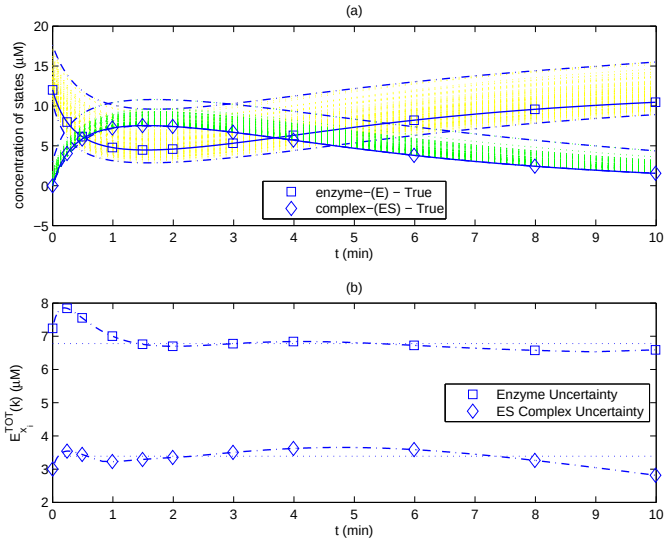

Figure 4.28: Non-uniform Observations w/ 15% Error: Range of Estimated Unmeasured Trajectories

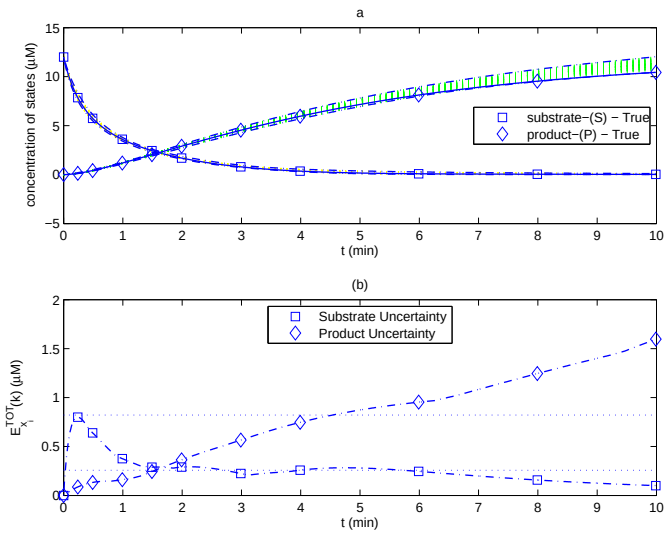

Figure 4.29: Non-uniform Observations w/ 15% Error: Range of Estimated Measured Trajectories

Table 4.11: Non-uniform Observations w/ 15% Error: Uncertainty Statistics of Estimated State Trajectories

	Max $E^{TOT}_{x_i(k)}$ (μM)	Avg $E^{TOT}_{x_i(k)}$ (μM)	Std Dev $E^{TOT}_{x_i(k)}$ (μM)
x_1	0.7998046	0.2571985	0.1318383
x_2	7.8541706	6.7820289	0.2680208
x_3	3.6539114	3.3893959	0.2202297
x_4	1.5964353	0.8209956	0.4315600

The average uncertainty associated with the ES complex in Table 4.11 was about $0.2\mu M$ larger. Both the maximum uncertainty and the standard deviation of the uncertainty had small increases as well.

The uncertainty associated with the observed components are shown in Figure 4.29a and 4.29b. The uncertainty associated with the product concentrations continued to have a monotonically increasing characteristic. This was different than the characteristics seen in Case 3 which started to exhibit a piecewise characteristic with a constant uncertainty at periods of times throughout the simulation. We also saw in Figure 4.29 that the uncertainty continued to grow at a fast rate as time increases, with the uncertainty having a maximum value of $1.6\mu M$. This was $0.35\mu M$ greater than the maximum uncertainty seen in Case 3 for the product concentrations. The uncertainty associated with the substrate did appear to improve for this case. The uncertainty between 0 and 1 minute still had the maximum amount of uncertainty but the maximum uncertainty was reduced to $0.8\mu M$. This was $0.7\mu M$ smaller than the maximum uncertainty seen in Case 3. This reduction in the uncertainty was also evident in the reduced standard deviation in the uncertainty associated with the substrate shown in Table 4.11, which was $0.16\mu M$ less than the standard deviation for the substrate uncertainty seen in Case 3.

4.5 Chapter Summary

The algorithm presented in Section 4.3 was used to estimate the unobserved dynamic characteristics of a template biochemical network under various conditions. We first verified the functionality of the algorithm by testing it under conditions where no measurement error was added to the observations used to estimate the unknown characteristics of the network. This was done for two sampling schemes; observations that were taken at uniformly spaced time points and observations that were taken at non-uniformly space time points. The algorithm was executed for 100 independent runs using different initial conditions. The resulting 100 estimates were analyzed to verify the accuracy and consistency of the algorithm. This was observed by looking at the average and variance of these 100 resulting estimates. The mean of these estimate were used to verify the accuracy of the algorithm for both cases where uniformly and non-uniformly spaced observations were used. Both situations resulted in mean estimates that had an absolute error that was less

than $O(10^{-6})$ when compared to the true initial condition and kinetic parameter values. The largest variance in the estimates for the initial concentrations was of $O(10^{-9})$ and the largest variance in the estimates of the kinetic parameters was of $O(10^{-12})$. This low upper limit on the variance and small absolute error between the true values and the mean estimates confirmed the algorithms ability to consistently produce correct estimates of the unobserved and unknown characteristics of the network. The mean estimates were then used to produce concentration trajectories for both the observed and unobserved components in the biochemical network. The comparison between these estimated trajectories and the true trajectories revealed that the algorithm was able to produce consistent estimates of the trajectories of the unobserved components and estimate the trajectories of the observed components between sample points. The results obtained from the observations that were non-uniformly spaced did produce results that had slightly less error but the difference was not significant.

Next, we implemented the algorithm on uniformly and non-uniformly spaced observations that had varying amounts of measurement error added to it. A range of possible estimates were obtained by executing the algorithm on 100 different realizations of observations that fell within a specific range of measurement uncertainty. This range of possible estimates was used to produce sets of concentration trajectories where the width of these trajectories over time defined the uncertainty associated with a specific observation scheme and measurement error. We were able to identify very distinct characteristics for both observation schemes for increasing amounts of measurement error. For example, increasing the measurement error for the uniformly sampled observations produced very large variations in the estimates of the initial concentrations of the enzyme and in the forward conversion kinetic rate parameter p_1. This translated to very large uncertainty in our overall ability to estimate the concentrations of this component over time. The uniformly sampled observations also made it difficult to accurately estimate the substrate concentrations between 0 and 1 minute. This was a direct result of having no samples between these time points. Our analysis showed a significant increase in the uncertainty associated with the substrate concentrations around this time as the total measurement error was increased from 4% to 15%. The uncertainty associated with the product concentrations did show some favorable characteristics as the measurement error was increased. At low measurement error, the uncertainty associated with this component appeared to be monotonically increasing. As the measurement error was increased, these characteristics tended to level off with the uncer-

tainty. At times, these characteristics even displayed a somewhat piecewise characteristic where the uncertainty remained constants at certain periods of time. This allowed us to infer that there may be some upper limit on the uncertainty for this components as the measurement error increases.

The estimates that were obtained from the observations that were non-uniformly spaced achieved slightly different characteristics and in some cases provided very distinct trade-offs when compared to the results obtained from the uniformly spaced observations. The most notable change was the increased ability to reduce the uncertainty in the concentration trajectories for the enzyme. This reduction was not initially evident when small amounts of measurement error was added to the observations. In fact, the results from the uniformly sampled observations and the results from the non-uniformly sampled observations exhibited very similar characteristics when 4% of measurement error was added to the observations. The reduction in enzyme uncertainty became more evident when the measurement error was increased to 10% and 15%. This produced a reduced range in the estimate of the initial enzyme concentration and the estimate of the forward converting rate parameter p_1. Thus, the resulting trajectories from the estimates obtained from the non-uniformly spaced observations showed a reduction in the overall uncertainty and a reduction in the average uncertainty associated with the enzyme. The concentration of the observed samples at earlier times also helped to reduce the uncertainty associated with the substrate because the increased number of samples helped to better resolve the transient characteristics of this component. This resulted in a significant reduction in the maximum uncertainty and in the standard deviation of the uncertainty associated with the substrate which signifies an improved ability to estimate this component. These improvements did come at the cost of a decreased ability to estimate the product concentrations. We saw the uncertainty associated with the product concentrations get significantly worse as the measurement error was increased. In fact, the results from the non-uniformly sample observations resulted in increased average uncertainty, maximum uncertainty, and standard deviation of uncertainty for this component. This may be a result of a decreased number of samples at later times in the simulation. Thus, the improved ability to estimate the enzyme and the substrate came at the cost of a decreased ability to estimate the product. This highlights the usefulness of strategically identifying the best time points to observe the system in order to maximize the reliability of the estimates while minimizing the uncertainty associated with the concentration trajectories. In both case we saw that small changes in

the substrate and product concentrations could result from a wide range of enzyme and ES complex concentrations. Thus, we conclude that the production of the product is not extremely sensitive to changes in the initial enzyme concentrations. For larger networks of such template pathways, this may have direct implication on the development of control algorithms for increase, decreasing, or maintaining constant yield of a pathway product over a specified range.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

5.1.1 Blind Analysis of High-Throughput Regulatory Data

ICA was applied to microarray data in the blind source separation analysis outline in Chapter 3. The data was obtained after exposing the bacteria to several stress conditions, which included heat shock, peroxide stress, and alcohol stress. The algorithm was able to separate the data into statistically independent components. We conjectured that these components represented functional groups of genes that were activated as a result of the stress experiments. We used our analysis approach to extract biological characteristics from the resulting components and compared our results with characteristics established in the literature. We examined the expression characteristics and functional relationships between several groups of genes that were known to response to these stressful experimental stimuli. We identified the results that we obtained as consistent with the results established in the literature. We also identified small groups of genes that may be related to larger groups of stress response genes. This functional link will need to be verified experimentally.

The results obtained from our analysis were consistent with regulatory characteristics identified in the literature, which reemphasizes the usefulness of this approach. Many aspects of this analysis, however, still need to be investigated. Further verification of

the consistency of our results with actual functional characteristics of groups of genes are needed. This requires more rigorous cross-validation approaches through further comparisons with other results in the literature and the generation of new experimental techniques for hypothesis testing. Also, ICA is based on the premise that the underlying components that comprise the data are statistically independent. This has not yet been verified to be a biologically relevant assumption even though our results did show much consistency with known characteristics. Further investigation is needed to either identify the relevance of this assumption or identify further biological characteristics upon which to base the source separation. Also, this analysis assumes that the individual source components are linearly mixed to form the resulting quantitative response that is seen in the microarray data. The use of linear approximations may be appropriate for initial stages of analysis but extensions to the separation of nonlinear mixtures may be needed in order to identify more detailed characteristics of the responsive regulons. Overall, the use of ICA to identify hidden regulatory characteristics from global expression data produced results that were consistent with the literature. These positive results promote the need to further explore such techniques as a means of initially analyzing regulatory changes in the genome that result from changes in environment, external signaling, and/or mutations.

5.1.2 State and Parameter Estimation in Biochemical Networks

The advancements in quantitative experimental approaches and the increase in first-principles knowledge of biological systems has prompted increased interests in analytical techniques for modeling and understanding biochemical pathways. The traditional form of these techniques, however, must be slightly modified in order to accommodate the nonlinear and uncertain aspects of pathway modeling. In this work, we implemented a least-squares approach to estimating the states and parameters of a single-step signal transduction pathway. The algorithm was formulated based on an optimal control approach and addressed several issues inherent to biochemical pathway modeling, which included limited sparse measurements, nonlinear characteristics, uncertainty due to experimental errors, and biological constraints on the components. The most novel of these attributes is the ability to address biological constraints on the pathway dynamics. The approach used a technique from optimal control theory that estimates input signals under equality and inequality constraints on system dynamics. This approach was modified and reformulated to fit the zero

input initial value models normally used to characterize biological pathways. We performed a proof of concept of the approach by applying it to sparse simulated measurements free of uncertainty. Our results showed that the algorithm was able to consistently identify the unknown or unmeasured characteristics of the pathway and was able to successfully reproduce the true pathway dynamics.

Our secondary analysis tested how bounded uncertainty added to the simulated measurements affected the resulting estimates provided by the algorithm. We saw that the bounded uncertainty did not greatly hinder our ability to generally characterize the substrate, enzyme-substrate complex, and product concentrations for both uniformly and non-uniformly space measurements. We saw that intuitive placement of sampled measurements over non-uniform intervals increased our ability to estimate the enzyme concentrations. We also saw that the same sampling scheme decreased our ability to accurately estimate the product concentrations. The positive and negative results associated with the arbitrary sampling of the observed data encourages the development of more systematic approaches for strategically identifying the most appropriate sampling schemes. These sampling schemes are identified according to some predefined criterion, e.g. minimizing the overall uncertainty associated with the estimated dynamics. These results also allowed us to qualify the impact of a given component on the remaining pathway dynamics. In this case, we saw that large changes in the enzyme concentrations did not greatly impact the dynamics of the remaining three components. This type of analysis has implications on the development of control algorithms for increasing, decreasing, or maintaining constant yield of the pathway product over a specified range.

The problem of generating predictive models for biochemical pathways is a very difficult albeit important problem that has medicinal, nutritional, and social benefits. Our analysis serves as another step in the long progression of developing tools for generating these models. This suggests that efforts spent towards further development and improvement of these approaches may lead to more successful applications of pathway modeling to biological and environmental problems. These problems include the development of approaches for impeding cancer growth, the development of alternative techniques for biofuel synthesis, and the development of advanced bioprocessing techniques for carbon emissions reduction.

5.2 Future Work

The work outlined in this document fits into a larger research plan that focuses on building a library of sophisticated engineering tools that are designed specifically to model, analyze, and control complex biological systems. These tools serve as the base of information integration, facilitating the integration of large amounts of data and knowledge from several research areas for the purpose of 1) understanding internal interactions, 2) producing and simulating predictive mathematical models, and 3) systematically manipulating and controlling biological systems. Future work will include the following topics.

Component and Parameter Estimation:

Tremendous progress in experimental approaches, particularly the extensive quantification of characteristics at the genomic, metabolomic, and proteomic levels, has increased our ability to effectively analyze biochemical pathways. Limitations of these state of the art approaches, however, inhibit our ability to measure all time-varying components and all kinetic parameters in the pathway. I plan to continue to develop and improve on techniques that systematically identify unknown or unmeasured kinetic parameters and component concentrations by calibrating the dynamics of the pathway to the steady state and transient data observed from wet-lab experiments. The integration of global search optimization algorithms aid in the development of estimation techniques that are capable of finding the best solution from sets of adequate solutions. Also, expansion of the current optimal control algorithm for estimating time varying regulatory control parameters would be a valuable tool for assessing regulatory effects on metabolic and signaling pathways.

Bounded Uncertainty Analysis:

Sources that attribute to uncertainty within biological systems are difficult to parameterize. This uncertainty is often best characterized as bounded error. I plan to extend on the uncertainty analysis performed in this work by developing algorithms that are capable of analytically tracking and bounding the uncertainty. These algorithms will be used to model and simulate the pathway in the presence of uncertain measurements, biological uncertainty, and variation due to unknown or under-modeled aspects of the pathway (modeling error). This analysis will be integrated into estimation procedures. This will result in a useful tool for providing bounds on estimated kinetic parameters and component concentrations. This

will also allow for the assessment of the predictive capabilities of the model through quantification of the width of possible valid pathway trajectories that result from the uncertain model.

Model Refinement:

An intricate exchange of information between wet-lab experimentation and model refinement approaches are needed in order to assess and eventually reduce the uncertainty associated with the model. I plan to develop algorithms that use uncertain models to design specifications for perturbation experiments and identify the most appropriate times to sample the system. These algorithms will be incorporated into an iterative modeling-and-refinement scheme whose objective is to reduce modeling error while increasing the predictive power of the *in-silico* model. The exchange of information will be used to identify unknown pathway mechanisms and reveal emergent pathway properties.

Sensitivity Analysis:

Analytical approaches for sensitivity analysis will highlight the kinetic parameters and components that have the greatest impact on the overall dynamics of the pathway. This analysis will have a direct impact on genetic engineering and control of these systems. It also has the potential of eliminating the pursuit of meaningless perturbations. Added functionality based on the bounded analysis procedures described above will allow these approaches to focus on the components and parameters capable of increasing, decreasing, or maintaining constant component concentrations over specified ranges of uncertainty.

Stable Nonlinear Control:

Uncertain mathematical models will be used to facilitate the creation of *in-silico* nonlinear control schemes that manipulate the necessary components in order to drive the pathway to specific steady states. These control schemes must be robust, stable, and able to maintain the desired operating state through various internal and external signal perturbations. The appropriate biological mechanisms can be identified in order to implement the control *in-vivo*. Optimal control schemes can also be developed which drive the system to a stable operating point that represents some optimal objective. Biological objectives include maximum yield and minimum flux.

Bibliography

[1] B. Müller-Hill, The Lac Operon: A Short History of a Genetic Paradigm. Walter de Gruyter, 1996.

[2] M. Ptashne, A Gene Switch: Phage λ and Higher Organisms. Blackwell Publishing, 1992.

[3] R. L. P. Adams, J. T. Knowler, and D. P. Leader, The Biochemistry of the Nucleic Acids. Springer, 1992.

[4] "Escherichia coli Genome Project." [Online]. Available: http://www.genome.wisc.edu/

[5] "Saccharomyces Genome Database." [Online]. Available: http://www.yeastgenome.org/

[6] "Arabidopsis thaliana Genome Project and Database." [Online]. Available: http://mips.gsf.de/proj/plant/jsf/athal/index.jsp

[7] M. B. Eisen, P. T. Spellman, P. O. Brown, and K. M. Devine, "Cluster analysis and display of genome-wide expression patterns," Proc. Nat. Acad. Sci., vol. 95, pp. 14 863–14 868, 1998.

[8] M. Bittner, P. Meltzer, Y. chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, G. Marincola, C. Godden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Lija, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, and V. Sondak, "Molecular classification of cutaneous malignant melanoma by gene expression profiling," Nature, vol. 406, pp. 536–540, 2000.

[9] P. J. Waddell and H. Kishino, "Cluster Inference Methods and Graphical Models Evaluated on NCI60 Microarray Gene Expression Data," Genome Informatics, vol. 11, pp. 129–140, 2000.

[10] J. Herrero, A. Valencia, and J. Dopazo, "A hierarchical unsupervised growing neural network for clustering gene expression patterns," Bioinformatics, vol. 17, no. 2, pp. 126–136, 2001.

[11] A. J. Lotka, Elements of Mathematical Biology. Dover Publications, Inc., 1956.

[12] L. Bertalanffy, General Systems Theory; Foundations, Development, Applications. New York: Braziller, 1968.

[13] E. O. Voit and J. H. Schwacke, "Understanding through Modeling: A Historical Perspective and Review of Biochemical Systems Theory as a Powerful Tool for Systems Biology," in System Biology: Principles, Methods, and Concepts, A. K. Konopka, Ed. CRC Press, Inc., 2007, pp. 27–82.

[14] H. Kitano, "Systems Biology: A Brief Overview," Science, vol. 295, pp. 1662–1664, 2002.

[15] ——, "Computational systems biology," Nature, vol. 420, pp. 206–210, 2002.

[16] J. Hasty, D. McMillen, and J. J. Collins, "Engineeered gene circuits," Nature, vol. 420, no. 6912, pp. 224–230, 2002.

[17] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási, "The large-scale organization of metabolic networks," Nature, vol. 407, pp. 651–654, 2000.

[18] J. C. Anderson, E. J. Clarke, A. P. Arkin, and C. A. Voigt, "Environmentally Controlled Invasion of Cancer Cells by Engineered Bacteria," J. Mol. Biol., vol. 355, pp. 619–627, 2006.

[19] Y. Chisti, "Biodiesel from microalgae," Biotechnology Advances, vol. 25, no. 3, pp. 294–306, 2007.

[20] R. J. Kavlock, G. Ankley, J. Blancato, M. Breen, R. Conolly, D. Dix, K. Houck, E. Hubal, R. Judson, J. Rabinowitz, A. Richard, R. W. Setzer, I. Shah, D. Villeneuve,

and E. Weber, "Computational Toxicology - A State of the Science Mini Review," Toxicological Sciences, vol. 103, no. 1, pp. 14–27, 2008.

[21] D. Ro, E. M. Paradise, M. Ouellet, K. J. Fisher, K. L. Newman, J. M. Ndungu, K. A. Ho, R. A. Eachus, T. S. Ham, J. Kirby, M. C. Y. Chang, S. T. Withers, Y. Shiba, R. Sarpong, and J. D. Keasling, "Production of the antimalarial drug precursor artemisinic acid in engineered yeast," Nature, vol. 440, no. 940-943, 2006.

[22] C. J. Musante, A. K. Lewis, and K. Hall, "Small- and large-scale biosimulation applied to drug discovery and development," Drug Discov. Today, vol. 7, pp. S192–S196, 2002.

[23] A. K. Lewis, T. Paterson, C. C. Leong, N. Defranoux, S. T. Holgate, and C. L. Stokes, "The Roles of Cells and Mediators in a Computer Model of Chronic Asthma," Allergy Immunology, vol. 124, pp. 282–286, 2001.

[24] M. J. Leckie, A. Brinke, J. Khan, Z. Diamant, B. O'Connor, C. Walls, A. Mathur, H. Cowley, D. Chung, and R. Djukanovic, "Effect of interleukin-5 blocking monoclonal antibody on eosinophils, airway hyper-responsiveness, and the late asthmatic response," The Lancet, vol. 356, no. 9248, pp. 2144–2148, 2000.

[25] A. R. Kansal, "Modeling approaches to type 2 diabetes," Diabetes Technol. Ther., vol. 6, pp. 39–47, 2004.

[26] O. T. Avery, C. M. MacLeod, and M. McCarty, "Studies of the chemical nature of the substance inducing transformation of pneumococcal type," J. Exp. Med., vol. 79, pp. 137–158, 1944.

[27] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, et al., "Initial sequencing and anlaysis of the human genome," Nature, vol. 409, pp. 860–921, 2001.

[28] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, et al., "The sequence of the human genome," Science, vol. 291, pp. 1304–1351, 2001.

[29] "BLAST: Basic Local Alignment and Search Tool." [Online]. Available: http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

[30] "Multiple Sequence Alignment by CLUSTALW." [Online]. Available: http://align.genome.jp/

[31] J. M. Walker, Medical Biomethods Handbook. Humana Press, 2005.

[32] B. Ø. Palsson, Systems Biology: Properties of Reconstructed Networks. Cambridge University Press, 2006.

[33] G. N. Stephanopoulos, A. A. Aristidou, and J. Nielsen, Metabolic Engineering: Principles and Methodologies. London, UK: Academic Press, 1998.

[34] R. Sharan and R. Shamir, "CLICK: A clustering algorithm with applications to gene expression analysis," in Proc. 8th International Conference on Intelligent Systems for Molecular Biology. AAAI Press, 2000, pp. 307–316.

[35] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," Proc. Natl. Acad. Sci., vol. 96, pp. 2907–2912, 1999.

[36] J. D. Helmann, M. Wu, P. A. Kobel, F. Gamo, M. Wilson, M. M. Morshedi, M. Navre, and C. Paddon, "Global transcriptional response of *Bacillus subtilis* to heat shock," J. Bacteriol., vol. 183, pp. 7318–7328, 2001.

[37] S. Raychaudhuri, J. M. Stuart, and R. B. Altman, "Principal component analysis to summarize microarray experiments: application to sporulation time series," Pac. Symp. Biocomput., vol. 5, pp. 455–466, 2000.

[38] W. Liebermeister, "Linear modes of gene expression determined by independent component analysis," Bioinformatics, vol. 18, pp. 51–60, 2002.

[39] O. Alter, P. O. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling," Proc. Natl. Acad. Sci., vol. 97, no. 18, pp. 10 101–10 106, 2000.

[40] J. Liao, R. Boscolo, Y. L. Yang, L. M. Tran, C. Sabatti, and V. Roychowdhury, "Network component analysis: Reconstruction of regulatory signals in biological systems," Proc. Nat. Acad. Sci., vol. 100, pp. 15 522–15 527, 2003.

[41] I. Shmulevich, E. R. Dougherty, and W. Zhang, "From Boolean to probabilistic Boolean networks as models of genetic regulatory networks," Proceeding of the IEEE, vol. 90, no. 11, pp. 1778–1792, 2002.

[42] T. Akutsu, S. Miyano, and S. Kuhara, "Identification of genetic networks from a small number of gene expression patterns under the Boolean network model," in Pacific Symposium on Biocomputing, 1999, pp. 17–28.

[43] S. A. Kauffman, "Metabolic stability and eqigenesis in randomly constructed genetic nets," J. Theor. Biol., vol. 22, no. 3, pp. 437–467, 1969.

[44] S. Huang, "Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery," J. Mol. Med., vol. 77, no. 6, pp. 469–480, 1999.

[45] S. A. Kauffman, The Origins of Order: Self-Organization and Selection in Evolution. New York: Oxford University Press, Inc., 1993.

[46] R. Somogyi and C. Sniegoski, "Modeling the complexity of gene networks: understanding multigenic and pleiotropic regulation," Complexity, vol. 1, pp. 45–63, 1996.

[47] M. Aldana, S. Coppersmith, and L. P. Kadanoff, "Boolean dynamics with random couplings," in Perspective and Problems in Nonlinear Science, ser. Applied Mathematical Sciences, E. Kaplan, J. E. Marsden, and K. R. Sreenivasan, Eds. New York, NY, USA: Springer-Verlag, 2003.

[48] I. Shmulevich and E. R. Dougherty, "Modeling genetic regulatory networks with probabilistic Boolean networks," in Genomic Signal Processing and Statistics, E. R. Dougherty, I. Shmulevich, J. Chen, and Z. J. Wang, Eds. New York: Hindawi Publishing Corporation, 2005.

[49] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang, "Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks," Bioinformatics, vol. 18, no. 2, pp. 261–274, 2002.

[50] H. de Jong and J. Geiselmann, "Modeling and simulation of genetic regulatory networks by ordinary differential equations," in

Genomic Signal Processing and Statistics, E. R. Dougherty, I. Shmulevich, J. Chen, and Z. J. Wang, Eds., 2005, pp. 203–239.

[51] C. Chassagnole, N. Noisommit-Rizzi, J. W. Schmid, K. Mauch, and M. Reuss, "Dynamic Modeling of the Central Carbon Metabolism of *Escherichia coli*," Biotechnol Bioeng, vol. 79, no. 1, pp. 53–73, 2002.

[52] M. Fussenegger, J. E. Bailey, and J. Varner, "A mathematical model of caspase function in apoptosis," Nat Biotechnol, vol. 18, pp. 768–774, 2000.

[53] R. A. Filter, X. Xia, and C. M. Gray, "Dynamic HIV/AIDS Parameter Estimation With Application to a Vaccine Readiness Study in Southern Africa," IEEE Trans. on Biomed. Eng., vol. 52, no. 5, pp. 784–791, 2005.

[54] F. Horn and R. Jackson, "General mass action kinetics," Archive for Rational Mechanics and Analysis, vol. 47, no. 2, pp. 81–116, 1972.

[55] L. Michaelis and M. L. Menten, "Die kinetic der invertinwirkung," Biochem. Zeitschrift, vol. 49, pp. 333–369, 1913.

[56] J. B. S. Haldane, Enzymes. Cambridge, MA: MIT Press, 1965.

[57] M. A. Savageau, "Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions," J. Theor. Biol., vol. 25, no. 3, pp. 365–369, 1969.

[58] ——, "Biochemical systems analysis. II. The steady-state solutions for an n-pool system using a power-law approximation," J. Theor. Biol., vol. 25, no. 3, pp. 370–379, 1969.

[59] ——, "Biochemical systems analysis. 3. Dynamic solutions using a power-law approximation," J. Theor. Biol., vol. 26, no. 2, pp. 215–226, 1970.

[60] A. Narang, A. Konopka, and D. Ramkrishna, "Dynamic Analysis of the Cybernetic Model for Diauxic Growth," Chemical Engineering Science, vol. 52, no. 15, pp. 2567–2578, 1997.

[61] A. A. Namjoshi and D. Ramkrishna, "A cybernetic modeling framework for analysis of metabolic systems," Computers and Chemical Engineering, vol. 29, pp. 487–498, 2005.

[62] ——, "Multiplicty and Stability of Steady States in Continuous Bioreactors: Dissection of Cybernetic Models," Chemical Engineering Science, vol. 56, pp. 5593–5607, 2001.

[63] I. Zevedei-Oancea and S. Schuster, "Topological analysis of metabolic networks based on Petri net theory," In Silico Biology, vol. 3, 2003. [Online]. Available: http://www.bioinfo.de/isb/2003/03/0029/

[64] K. Voss, M. Heiner, and I. Koch, "Steady-state analysis of metabolic pathways using Petri nets," In Silico Biology, vol. 3, 2003. [Online]. Available: http://www.bioinfo.de/isb/2003/03/0031/

[65] J. S. Oliveira, C. G. Bailey, J. B. Jones-Oliveira, D. A. Dixon, D. W. Gull, and M. L. Chandler, "A Computational model for the identification of biochemical pathways in the Krebs cycle," J. Comput. Biol., vol. 10, no. 1, pp. 57–82, 2003.

[66] M. Peleg, D. Rubin, and R. B. Altman, "Using Petri Net Tools to Study Properties and Dynamics of Biological Systems," Journal of American Medical Informatics Association, vol. 12, no. 2, pp. 181–199, 2005.

[67] M. Heiner, I. Koch, and J. Will, "Model validation of biological pathways using Petri nets - demonstrated for apoptosis," Bio Systems, vol. 75, pp. 15–28, 2004.

[68] A. Doi, S. Fujita, H. Matsuno, M. Nagasaki, and S. Miyano, "Constructing biological pathway models with hybrid functional Petri nets," In Silico Biology, vol. 4, 2004. [Online]. Available: http://www.bioinfo.de/isb/2004/04/0023/

[69] M. Nagasaki, A. Doi, H. Matsuno, and S. Miyano, "Genomic Object Net: a platform for modeling and simulating biopathways," Applied Bioinformatics, vol. 2, pp. 181–184, 2004.

[70] M. Chen, A. Freier, J. Köhler, and A. Rüegg, "The biology Petri net markup language," in Proceedings of Promise2002, vol. 21, 2002, pp. 150–161.

[71] K. G. Gadkar, J. Varner, and F. J. Doyle III, "Model identification of signal transduction networks from data using a state regulator problem," Syst. Biol., vol. 2, no. 1, pp. 17–30, 2005.

[72] C. Williams, W. Alexander, and W. Edmonson, "Estimating the Unmeasured Dynamics of Biological Systems using a Constrained Real-Coded Genetic Algorithm," in Conference Record of the Fortieth Asilomar Conference on Signals, Systems, and Computers, October 29 - November 1 2006.

[73] G. Bastin and D. Dochain, Eds., On-line estimation and adaptive control of bioreactors. Amsterdam, Netherlands: Elsevier, 1990.

[74] N. A. W. van Riel and E. D. Sontag, "Parameter estimation in models combining signal transduction and metabolic pathways: the dependent input approach," IEE Proc. on Syst. Biol., vol. 153, no. 4, pp. 263–274, 2006.

[75] Y. Zheng and A. Rundell, "Comparative study of parameter sensitivity analyses of the TCR-activated Erk-MAPK signalling pathway," IEE Proc. on Syst. Biol., vol. 153, no. 4, pp. 201–211, 2006.

[76] W. Tucker, Z. Kutalik, and V. Moulton, "Estimating parameters for generalized mass action models using constraint propagation," Mathematical Biosciences, vol. 208, pp. 607–620, 2007.

[77] P. K. Polisetty, E. O. Voit, and E. P. Gatzke, "Identification of metabolic system parameters using global optimization methods," Theoretical Biology and Medical Modelling, vol. 3, no. 4, 2006.

[78] H. U. Voss, J. Timmer, and J. Kurths, "Nonlinear Dynamical System Identification from Uncertain and Indirect Measurements," International Journal of Bifurcation and Chaos, vol. 14, no. 6, pp. 1905–1933, 2004.

[79] K. Schittkowski, "Parameter estimation in systems of nonlinear equations," Numerische Mathematik, vol. 68, pp. 129–142, 1994.

[80] J. Timmer, H. Rust, W. Horbelt, and H. U. Voss, "Parametric, nonparametric and parametric modelling of chaotic circuit time series," Physics Letters A, vol. 274, pp. 123–134, 2000.

[81] J. Almeida and E. O. Voit, "Neural network-based parameter estimation in complex biochemical systems," Genome Informatics, vol. 14, pp. 114–123, 2003.

[82] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," Nucleic Acids Res., vol. 28, pp. 27–30, 2000.

[83] "National Center for Biotechnology Information," Available at: http://www.ncbi.nlm.nih.gov.

[84] P. O. Brown and D. Botstein, "Exploring the new world of the genome with DNA microarrays," Nat. Genet., vol. 21, pp. 33–37, 1999.

[85] M. K. Kerr, M. Martin, and G. A. Churchill, "Analysis of Variance for Gene Expression Microarray Data," Journal of Computational Biology, vol. 7, no. 6, pp. 819–837, 2000.

[86] D. C. Weaver, C. T. Workman, and G. D. Stormo, "Modeling Regulatory Networks with Weight Matrices," Pac. Symp. Biocomput., vol. 4, pp. 112–123, 1999.

[87] M. A. Shea and G. K. Ackers, "The O_R Control System of Bacteriophage Lambda: A Physical-Chemical Model for Gene Regulation," J. Mol. Biol., vol. 181, pp. 211–230, 1985.

[88] L. Ljung, System Identification: Theory for the User. Prentice Hall, Inc., 1987.

[89] O. Nelles, Nonlinear System Identification. Springer-Verlag Berlin Heidelberg New York, 2001.

[90] S. Haykin, Neural Networks: A Comprehensive Foundation. Prentice Hall, Inc., 1999.

[91] A. Hyvärinen, J. Karhunen, and E. Oja, Independent Component Analysis. John Wiley & Sons, Inc., 2001.

[92] J. V. Stone, Independent Component Analysis: A Tutorial Introduction. The M.I.T. Press, 2004.

[93] I. T. Jolliffe, Principal Component Analysis. Springer-Verlag, 1986.

[94] G. H. Golub and C. F. V. Loan, Matrix Computations. JHU Press, 1996.

[95] J. C. Loehlin, Latent variable models: An Introduction to Factor, Path, and Structural Analysis. Lawrence Erlbaum Associates, 2004.

[96] S. Makeig, A. J. Bell, T. P. Jung, and T. J. Sejnowski, "Independent component analysis of electroencephalographic data," in Proceedings of the Advances in Neural Information Processing Systems, D. Touretzky, M. Mozer, and M. Hasselmo, Eds. MIT Press, 1996, pp. 145–151.

[97] R. Vigario, V. Jousmaki, M. Hamalainen, R. Hari, and E. Oja, "Independent component analysis for identification of artifacts in magnetoencephalographic recordings," in Proceedings of the Advances in Neural Information Processing Systems, M. Jordan, M. Kearns, and S. Solla, Eds. MIT Press, 1997, pp. 229–235.

[98] A. D. Back and A. S. Weigend, "A first application of independent component analysis to extracting structure from stock returns," Int. J. Neural Syst., vol. 8, pp. 473–484, 1997.

[99] K. Kiviluoto and E. Oja, "Independent component analysis for parallel financial time series," in Proceedings of the Fifth International Conference on Neural Information Processing, M. Kearns, S. Solla, and D. Cohn, Eds. MIT Press, 1999, pp. 895–898.

[100] R. F. Li and X. Z. Wang, "Dimension reduction of process dynamic trends using independent component analysis," Computers and Chemical Engineering, vol. 26, no. 3, pp. 467–473, 2002.

[101] P. Hoyer and A. Hyvärinen, "Independent Component Analysis Applied to Feature Extraction from Colour and Stereo Images," Network: Computation in Neural Systems, vol. 11, no. 3, pp. 191–210, 2000.

[102] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," IEEE Trans. on Neural Networks, vol. 13, no. 6, pp. 1450–1464, 2002.

[103] C. Williams, W. Alexander, and T. Pompey, "Independent and Principle Component Analyses of the Global Stress Response of *Bacillus subtilis*," in Proc. IEEE International Workshop on Genomic Signal Processing and Statistics, Newport, Rhode Island, May 2005.

[104] A. Papoulis and S. U. Pillai, Probability, Random Variables and Stochastic Processes. McGraw Hill, 2001.

[105] A. Hyvärinen, "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis," IEEE Trans. on Neural Networks, vol. 10, no. 3, pp. 626–634, 1999.

[106] "FastICA Toolbox for Matlab." [Online]. Available: http://www.cis.hut.fi/projects/ica/fastica/

[107] J. Himberg and A. Hyvärinen, "Icasso: software for investigating the reliability of ICA estimates by clustering and visualization," IEEE 13th Workshop on Neural Networks for Signal Processing, pp. 259–268, 17-19 Sept. 2003.

[108] J. Himberg, A. Hyvärinen, and F. Esposito, "Validating the independent components of neuroimaging time-series via clustering and visualization," NeuroImage, vol. 22, no. 3, pp. 1214–1222, 2004.

[109] B. Efron and R. Tibshirani, An Introduction to the Bootstrap. New York: Chapman and Hall, 1993.

[110] "Icasso Toolbox for MATLAB." [Online]. Available: http://www.cis.hut.fi/projects/ica/icasso/

[111] J. D. Helmann, M. Wu, A. Gaballa, P. A. Kobel, M. M. Morshedi, P. Fawcett, and C. Paddon, "The global transcriptional response of *Bacillus subtilis* to peroxide stress is coordinated by three transcription factors," J. Bacteriol., vol. 185, pp. 243–253, 2003.

[112] C. A. Ball, I. A. Awad, J. Demeter, J. Gollub, J. M. Hebert, T. Hernandez-Boussard, H. Jin, J. C. Matese, M. Nitzberg, F. Wymore, Z. K. Zachariah, P. O. Brown, and G. Sherlock, "The Stanford Microarray Database accommodates additional microarray platforms and data forms," Nucleic Acids Res, vol. 33, no. 1, pp. D580–2, 2005.

[113] M. Hecker and S. Engelmann, "Proteomics, DNA arrays and the analysis of still unknown regulons and unknown proteins of *Bacillus subtilis* and pathogenic gram-positive bacteria," Intl. J. Med. Microbiol., vol. 290, pp. 123–134, 2000.

[114] C. W. Price, "Protective function and regulation of the general stress response in *Bacillus subtilis* and related gram-positive bacteria," in Bacterial stress responses, G. Storz and R. Hengge-Aronis, Eds. Washington D.C.: ASM Press, 2000, pp. 179–197.

[115] C. W. Price, P. Fawcett, H. Ceremonine, Y. Su, C. K. Murphy, and P. Youngman, "Genome-wide analysis of the general stress response in *Bacillus subtilis*," Mol. Microbiol., vol. 41, pp. 757–774, 2001.

[116] A. Petersohn, H. Antelmann, U. Gerth, and M. Hecker, "Identification and transcriptional analysis of new members of the σ^{b} regulon in *Bacillus subtilis*," Microbiology, vol. 145, pp. 869–880, 1999.

[117] C. Haass, "Apoptosis. Dead End for Neurodegeneration," Nature, vol. 399, pp. 204–207, 1999.

[118] C. Wellbrock, M. Karasarides, and R. Marais, "The RAF proteins take centre stage," Nat. Rev. Mol. Cell. Biol., vol. 5, no. 11, pp. 875–885, 2004.

[119] C. Rodriguez-Caso, R. M. nez, M. Cascante, F. Sánchez-Jiménez, and M. A. Medina, "Mathematical Modeling of Polyamine Metabolism in Mammals," Journal of Biological Chemistry, vol. 281, no. 31, p. 21799, 21812 2006.

[120] P. M. Bramley, "Is lycopene beneficial to human health?" Phytochemistry, vol. 54, no. 3, pp. 233–236, 2000.

[121] M. Sticklen, "Plant genetic engineering to improve biomass characteristic for biofuels," Current Opinion in Biotechnology, vol. 17, no. 3, pp. 315–319, 2006.

[122] Z. Kutalik, K. H. Cho, and O. Wolkenhauer, "Optimal sampling time selection for parameter estimation in dynamic pathway modeling," Bio Systems, vol. 75, pp. 43–55, 2004.

[123] D. Gilbert, H. Fuß, X. Gu, R. Orton, S. Robinson, V. Vyshemirsky, M. Kurth, C. Stephen Downes, and W. Dubitzky, "Computational methodologies for modelling, analysis, and simulation of signalling networks," Briefings in Bioinformatics, vol. 7, no. 4, pp. 339–353, 2006.

[124] N. A. W. van Riel, "Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments," Briefings in Bioinformatics, vol. 7, no. 4, pp. 364–374, 2006.

[125] W. L. Brogan, Modern Control Theory. Prentice Hall, Inc., 1991.

[126] A. Gelb, J. F. Kasper, R. A. Nash, C. F. Price, and A. A. Sutherland, Applied Optimal Estimation. The M.I.T. Press, 1974.

[127] F. L. Lewis, Optimal Control. John Wiley & Sons, Inc., 1986.

[128] D. E. Kirk, Optimal Control Theory: An Introduction. Dover Publications, Inc., 2004.

[129] A. K. Griffith and N. K. Nichols, "Data Assimilation Using Optimal Control Theory," The University of Reading, Tech. Rep., 1994.

[130] A. Quarteroni, R. Sacco, and F. Saleri, Numerical Mathematics. New York, NY: Springer-Verlag, 2000.

[131] A. S. Torralba, K. Yu, P. Shen, P. J. Oefner, and J. Ross, "Experimental test of a method for determining casual connectivities of species in reactions," Proc. Natl. Acad. Sci., vol. 100, pp. 1494–1498, 2003.

[132] B. D. Plikaytis, D. Goldblatt, C. E. Frasch, C. Blondeau, M. J. Bybel, G. S. Giebink, I. Jonsdottir, H. Käyhty, H. B. Konradsen, D. V. Madore, M. H. Nahm, C. A. Schulman, P. F. Holder, T. Lezhava, C. M. Elie, and G. M. Carlone, "An Analytical Model Applied to a Multicenter Pneumoncoccal Enzyme-Linked Immunosorbent Assay Study," Journal of Clinical Microbiology, vol. 38, no. 6, pp. 2043–2050, 2000.